Sekvenciranje RNA in primer uporabe v diagnostiki

RNA sequencing and example of diagnostic use

Nika Breznik[1], Klementina Črepinšek[2,3], Maruša Debeljak[2]
[1]Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biokemijo in molekularno genetiko
[2]Univerzitetni klinični center Ljubljana, Pediatrična klinika, Klinični inštitut za specialno laboratorijsko diagnostiko
[3]Univerza v Ljubljani, Medicinska fakulteta

Avtor za korespondenco:
asist. Nika Breznik, mag. lab. biomed.
Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biokemijo in molekularno genetiko, Vrazov trg 2, 1000 Ljubljana
e-pošta: nika.breznik@mf.uni-lj.si

POVZETEK

Sekvenciranje RNA je napredna molekularna tehnika, ki omogoča preučevanje transkriptov. Prvi korak pri sekvenciranju RNA je izolacija RNA iz biološkega vzorca, ki ji sledijo priprava knjižnice RNA, sekvenciranje in bioinformatska analiza. Glavna cilja sekvenciranja sta kvantifikacija in primerjava izražanja genov med različnimi pogoji, kar nam omogoča vpogled v biološko funkcijo analiziranih genov. Kljub številnim prednostim ima sekvenciranje RNA tudi nekatere omejitve in pomanjkljivosti. Eni izmed glavnih pomanjkljivosti sta nestabilnost RNA in občutljivost na razgradnjo z RNAzami, ki se jih težko znebimo, zato je potrebna skrbna priprava vzorca. Klinična uporaba te tehnologije je prikazana na primeru bolnika z B-celično akutno limfoblastno levkemijo.

Ključne besede: sekvenciranje RNA, analiza različnega izražanja genov

ABSTRACT

RNA sequencing is an advanced molecular technique that allows the transcriptomes to be studied in detail, representing a remarkable advance in transcriptome studies. The first step in RNA sequencing is the isolation of RNA from a biological sample, followed by RNA library preparation, sequencing, and bioinformatics analysis. The main objective is to quantify and compare gene expression between different conditions, providing insights into the biological function of the genes analyzed. Despite its numerous advantages, RNA sequencing also has certain limitations and drawbacks. These include the instability of RNA and its susceptibility to degradation by RNAses, which are difficult to remove and require careful sample preparation. The clinical application of this technology is demonstrated using the example of a patient with B-cell acute lymphoblastic leukaemia.

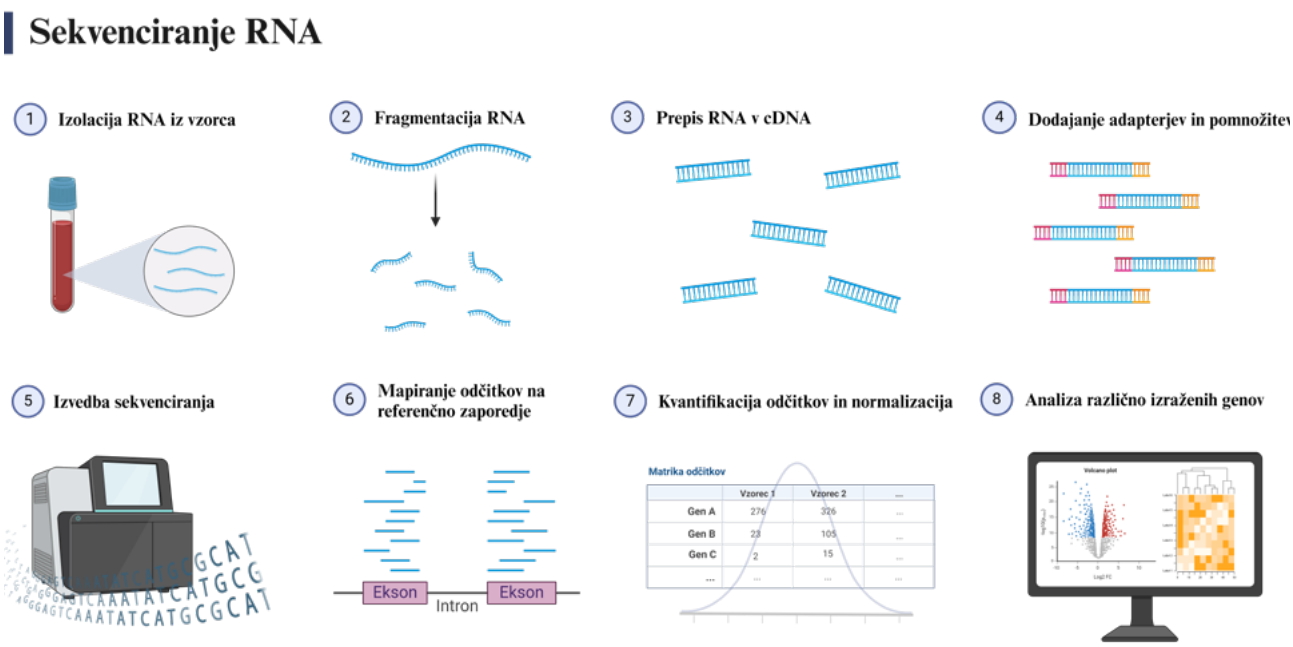
Key words: RNA sequencing, differential gene expression analysis

»

UVOD

Sekvenciranje RNA je sodobna molekularna tehnika, ki omogoča vpogled v transkriptome številnih celic in tkiv. Prinesla je izjemen napredek na področju transkriptomskih študij. V primerjavi z mikromrežami je sekvenciranje RNA bolj občutljivo in ponuja številne aplikacije zaradi širokega dinamičnega razpona, možnosti prepoznavanja novih transkriptov ter ločljivosti na ravni posameznega nukleotida (1). Glavni cilj analize transkriptoma je običajno kvantificirati in primerjati izražanje genov med različnimi stanji in na podlagi teh podatkov sklepati o biološki funkciji analiziranih genov (2). Podatki, pridobljeni s sekvenciranjem RNA, se lahko uporabljajo pri anotaciji – procesu določanja lokacije in zaporedja genomskih elementov ter določanja njihove biološke funkcije. Medtem ko se za anotacijo uporabljajo tudi genomski podatki, transkriptomski podatki omogočajo pridobitev informacij o nekaterih elementih, ki jih zgolj z uporabo genomskih podatkov ni mogoče zaznati (npr. neprevedene regije UTR, nekodirajoče RNA in posttranskripcijski dogodki) (3).

Sekvenciranje RNA omogoča tudi odkrivanje različnega alternativnega spajanja – razlik v vzorcih spajanja eksonov med različnimi biološkimi pogoji. Alternativno spajanje eksonov je zelo razširjen mehanizem uravnavanja izražanja genov (4), pri čemer lahko napake v tem sistemu povzročajo številne bolezni (5). Nadaljnji napredek v zmogljivostih in dolžinah odčitkov tehnologij NGS je olajšal tudi odkrivanje fuzijskih transkriptov, ki jih pogosto najdemo v tumorskih celicah in lahko služijo kot biooznačevalci ali terapevtske tarče (6), poleg tega pa lahko analiziramo tudi prisotnost variant v kodirajočih delih genoma (7). V nadaljevanju bomo najprej opisali potek sekvenciranja RNA, od izolacije RNA iz biološkega vzorca do priprave knjižnic RNA in bioinformatske analize (Slika 1), nato bomo izpostavili nekaj pomanjkljivosti in omejitev sekvenciranja RNA. Na koncu bomo na kliničnem primeru predstavili še opis poteka sekvenciranja RNA pri bolniku z B-celično akutno limfoblastno levkemijo (B-ALL).



Slika 1: Shema poteka sekvenciranja RNA, od izolacije RNA do analize različno izraženih genov. Posamezni koraki, ki so prikazani na shemi, so podrobneje razloženi v besedilu. Shema je bila pripravljena z BioRenderjem.

Figure 1: Image of RNA sequencing, from RNA isolation to the analysis of differentially expressed genes. The individual steps shown in the image are explained in more detail in the review. The image was created using BioRender.

IZOLACIJA RNA

Prvi korak pri sekvenciranju RNA je izolacija RNA iz biološkega vzorca, ki jo lahko izvedemo iz gojenih celic, periferne krvi, kostnega mozga, plazme, seruma in drugih telesnih tekočin ali tkiv. Izolacija je možna tako iz svežega kot tudi zamrznjenega tkiva ali celo iz s formalinom fiksiranih vzorcev, vklopljenih v parafin (8). Za izolacijo RNA se najpogosteje uporabljajo pripravljeni reagenčni kompleti, ki omogočajo učinkovito in kakovostno izolacijo (9). Pred začetkom priprave knjižnice je treba izolirani RNA določiti koncentracijo in kakovost (10). Za ocenjevanje kakovosti oz. degradiranosti RNA se uporablja parameter RIN (angl. *RNA integrity number*), katerega vrednosti so od 1 (popolnoma degradirana RNA) do 10 (intaktna RNA) (11).

PRIPRAVA KNJIŽNIC RNA

Obogatitev mRNA ali odstranitev rRNA

V celici najdemo več različnih zvrsti RNA, kar od 80 do 90 % vseh molekul RNA predstavlja ribosomalna RNA (rRNA), ki nas v procesu sekvenciranja ne zanima in jo je treba v prvem koraku priprave knjižnice odstraniti (12). To lahko naredimo s selekcijo poli-A koncev ali z deplecijo rRNA. Pri prvem pristopu uporabimo oligo-dT sonde, ki se povežejo s poli-A repi na zreli mRNA, in jih osamimo s pomočjo magnetnih kroglic. Vendar degradirani vzorci in številne nekodirajoče RNA, kot so mikro RNA (miRNA), ne vsebujejo poli-A koncev in jih z uporabo tega pristopa izgubimo. Za sekvenciranje celotnega transkriptoma je tako primernejši pristop deplecija rRNA, pri kateri uporabimo oligonukleotidne sonde s specifičnim zaporedjem, komplementarnim citoplazemskim in mitohondirjskim rRNA (13). Nastale oligo-DNA:RNA hibride nato odstranimo s pomočjo magnetnih kroglic (14) ali jih razgradimo z RNazo H (15).

Fragmentacija RNA in dodajanje adapterjev

Odstranitvi rRNA sledi fragmentacija RNA, potrebna zaradi velikostne omejitve večine tehnologij sekvenciranja. Mogoči sta kemijska ali encimska fragmentacija. Kemijska fragmentacija poteka z uporabo alkalnih raztopin ali raztopin z dvovalentinimi kationi (npr. Mg^{2+} ali Zn^{2+}) pri povišani temperaturi, običajno pri 70 °C (16). Encimska fragmentacija poteka z različnimi encimi, kot je RNaza III (17). Fragmentacija ni povsem naključna in je lahko vir povečane zastopanosti določenih regij RNA. Fragmentirano RNA nato z naključnimi heksameri prepišemo v komplementarno DNA (cDNA). Redkeje se najprej izvede reverzna transkripcija RNA v cDNA in nato fragmentacija cDNA (18). Ta se običajno izvaja z ultrazvočnimi valovi ali z DNazami. Fragmentirani cDNA v nadaljevanju dodamo adapterje, ki omogočajo klonalno pomnožitev knjižnice in njeno sekvenciranje. Adapterje lahko dodamo na več različnih načinov, vendar je treba paziti, da pri tem ohranimo informacijo o smeri RNA (19). Z uporabo metode dUTP je mogoče ohraniti informacijo o smeri RNA. S to metodo se med sintezo cDNA pri pripravi knjižnice namesto dTTP uporabljajo dUTP, ki se vgradijo v drugo verigo. Pred pomnoževanjem PCR se druga veriga, ki vsebuje uracile, razgradi z uracil-N-glikozilazo, tako da se pomnoži samo prva veriga (20). Takšna priprava knjižnice omogoča določitev izražanja prekrivajočih genov, torej tistih genov, ki imajo vsaj delno prekrivajoče genomske koordinate, a se prepisujejo iz različnih verig (21). Posameznim vzorcem lahko v procesu priprave knjižnice dodamo tudi indekse oz. molekularne črtne kode, ki omogočajo identifikacijo posameznega vzorca po sekvenciranju. Z uporabo indeksov lahko pripravljene knjižnice združimo in s tem povečamo učinkovitost ter zmanjšamo stroške sekvenciranja.

Klonalna amplifikacija knjižnice

Pripravljene knjižnice, označene z adapterji, je treba pred sekvenciranjem pomnožiti s PCR. Razlike v velikosti in sestavi cDNA lahko kljub le majhnemu številu ciklov povzročijo neenakomerno pomnoževanje. Za popravljanje pristranskosti PCR se lahko uporabljajo molekularne oznake, imenovane edinstveni molekularni identifikatorji (UMI, angl. *unique molecular identifiers*), ki omogočajo odstranitev PCR duplikatov (22). UMI so običajno vgrajeni v adaptersko zaporedje in se dodajo cDNA pred pomnoževanjem. Razlikujejo se po velikosti (številu baz) in kompleksnosti. Lahko so sestavljeni iz določenega ali naključnega zaporedja. Molekularno označevanje je še posebej uporabno pri manjših količinah vhodne RNA, kjer je potrebno večje število ciklov pomnoževanja (23).

»

Sekvenciranje

Sekvenciranje RNA se izvaja na enakih platformah kot sekvenciranje celotnega eksoma ali celotnega genoma. Najpogosteje se uporablja tehnologija Illumina, ki omogoča sekvenciranje z visoko natančnostjo in zmogljivostjo. Določimo lahko zaporedje enega ali obeh koncev fragmenta DNA. Sekvenciranje s parnimi konci omogoča natančnejšo mapiranje odčitkov na referenčni genom (24).

BIOINFORMATSKA ANALIZA

Po končanem sekvenciranju sledi bioinformatska analiza. Svetlobne ali električne signale, pridobljene med sekvenciranjem, najprej pretvorimo v nukleotidno zaporedje, med tem pa s pomočjo uporabljenih indeksov tudi določimo, kateri signali pripadajo kateremu vzorcu. Za vsak vzorec dobimo datoteko formata FASTQ, ki vsebuje podatke o sekvenciranju, surova sekvenčna zaporedja in oceno kakovosti za posamezno zaporedje. Najprej je treba preveriti kakovost odčitkov, kar izvedemo z orodjem FastQC. V tem koraku odstranimo baze z nizko kakovostjo, ki se običajno nahajajo na 3'-koncu, in adapterska zaporedja. V primeru, da smo uporabljali UMI, je treba tudi te pred poravnavo odstraniti. Med bioinformatsko analizo je potrebno upoštevati priporočila, ki jih predvideva konzorcij za standardizacijo ENCODE (25). S tem zagotavljamo primerljivost in reproducibilnost podatkov.

Poravnava zaporedji na referenčni genom ali transkriptom

V nadaljevanju je treba za vsak odčitek najti mesto, kjer se najbolje ujema z referenčnim zaporedjem, kar imenujemo poravnava oz. mapiranje odčitkov na referenčni genom ali transkriptom. Pri tem je treba upoštevati, da lahko odčitki vsebujejo polimorfizem posameznega nukleotida (SNP, angl. *single-nucleotide polymorphism*), delecije, insercije ali napake, nastale pri sekvenciranju, in se zato ne ujemajo popolnoma z referenčnim zaporedjem. Nekateri odčitki se lahko ujemajo z več lokacijami v referenčnem zaporedju. Takšne odčitke lahko algoritmi zavržejo (26), naključno mapirajo (27) ali mapirajo na podlagi povprečne pokritosti (28). Z uporabo sekvenciranja s parnimi konci se oba konca fragmentov nahajata blizu skupaj, kar v nekaterih primerih lahko odpravi dvoumnost pri mapiranju. Treba je upoštevati, da odčitki izhajajo iz transkriptoma in ne iz genoma. Enostaven pristop je uporaba samega genoma kot reference, vendar odčitki, ki segajo preko meje eksonov, ne bodo mapirani. Transkripti z manj eksoni so tako bolje pokriti od daljših odčitkov pri enaki ravni izražanja (29). Za poravnavo na referenčno zaporedje se najpogosteje uporabljajo orodja BWA (30), bowtie (31) in STAR (32). Enostavnejši pristop je uporaba "psevdo-poravnave", pri kateri se odčitki ne poravnajo na referenčni genom na običajen način. Namesto tega se odčitki samo klasificirajo glede na to, iz katerega gena ali transkripta izvirajo. Gre za hitrejšo metodo, ki potrebuje manj računalniških virov, saj ne vključuje natančne poravnave vsakega odčitka na specifično mesto v genomu. Namesto tega se prepoznajo vzorci v odčitkih, ki jih povezujejo z znanimi geni ali transkripti, kar omogoča hitro in učinkovito "poravnavo". Orodji, ki omogočata "psevdo-poravnavo", sta Salmon (33) in Kallisto (34). Mapirani odčitki so shranjeni v standardnem formatu SAM (angl. *sequence alignment map*), ki ga lahko pretvorimo v binarno obliko – format BAM (angl. *binary alignment map*). Transkriptom pa je mogoče sestaviti tudi *de novo*, kar pomeni, da transkriptom sestavimo iz sekvenciranih odčitkov brez uporabe referenčnega genoma. Ta pristop se uporablja predvsem za organizme, pri katerih referenčni genom ni na voljo. V postopku sestavljanja transkriptoma *de novo* se kratki odčitki, ki jih dobimo pri sekvenciranju, na podlagi prekrivanja med njimi združujejo v daljše sekvence. Te daljše sekvence predstavljajo transkripte, ki jih je treba na koncu identificirati.

Kvantifikacija genov oz. transkriptov

Po pridobitvi genomskih lokacij za čim več odčitkov sledi kvantifikacija odčitkov na biološko pomembne enote. Kvantifikacija je mogoča le na anotirane biološke enote; to pomeni, da imajo določene genomske koordinate ime in druge funkcionalne informacije. Običajno izvajamo kvantifikacijo na gene, mogoča pa je tudi kvantifikacija na eksone ali transkripte. V procesu kvantifikacije preštejemo, koliko odčitkov se prilega na določeno biološko pomembno enoto. Kot rezultat dobimo tabelo s številom odčitkov za vsako posamezno enoto (gen, transkript) pri vsakem vzorcu (35), ki jo imenujemo matrika odčitkov in je prikazana pri koraku 7 na Sliki 1.

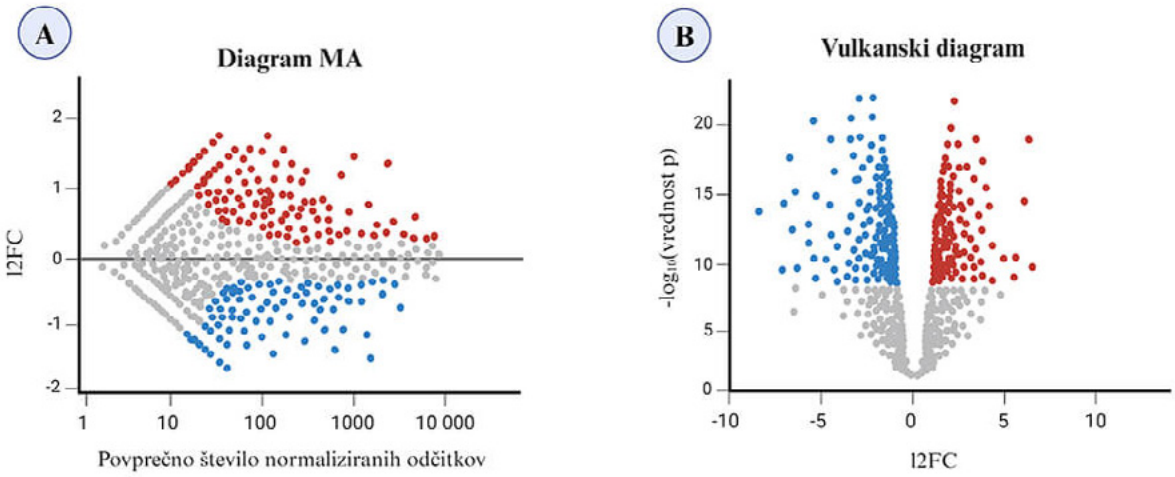
»

Normalizacija

Normalizacija omogoča primerjavo izražanja genov med vzorci in znotraj vzorca. Pri primerjavi izražanja znotraj vzorca uporabljamo FPKM (angl. *fragments per kilobase per million mapped reads*) ali TPM (angl. *transcripts per million mapped reads*), ki omogočata primerjavo izražanja vsakega gena glede na druge gene v vzorcu in upoštevata tako velikost knjižnic kot tudi dolžino genov (36). Kadar delamo primerjavo med različnimi vzorci, se tehnične napake večinoma izničijo, saj med seboj primerjamo iste gene. Ker pa gre za različne knjižnice, je normalizacija še zmeraj potrebna. Najpreprostejša in pogosto uporabljena normalizacija je prilagoditev glede na celokupno število odčitkov v knjižnici (37).

Analiza različnega izražanja genov oz. transkriptov

Cilj analize je prepoznati gene, katerih izraženost se med skupinami vzorcev statistično značilno razlikuje. Najpogosteje se za ta namen uporabljata parametrični metodi DESeq2 in EdgeR (38). DESeq2 se uporablja predvsem za podatke z manjšo variabilnostjo in natančno opredeljenimi skupinami (39), EdgeR pa je primeren za manjše, a kompleksnejše podatke z visoko variabilnostjo (40). Kot kriterij za določanje različne izraženosti genov uporabimo vrednost p in dvojiški logaritem spremembe (l2FC, angl. *log2 fold change*) (41), pri čemer vrednost p pove, ali je rezultat statistično značilen, l2FC pa pove, kolikokrat višje oziroma nižje je izražanje gena v testni v primerjavi s kontrolno skupino. Zaradi večkratnega testiranja je treba vrednost p popraviti – običajno uporabimo Benjamini-Hochbergov popravek (42). Izbira mejnih vrednosti p in l2FC, s katerimi identificiramo različno izražene gene, je odvisna od poskusa (43). Rezultate statistične analize običajno grafično predstavimo na diagramu MA (44) ali z vulkanskim diagramom (45). Diagram MA podatke pretvori v lestvico M (logaritemsko razmerje) in lestvico A (povprečna vrednost); prvotno se je uporabljal za prikazovanje podatkov, pridobljenih z mikromrežami. Prikazuje vrednosti l2FC, odvisne od povprečnega normaliziranega števila odčitkov. Na vulkanskem diagramu pa je prikazana odvisnost vrednosti p od l2FC (Slika 2).



Slika 2: Primer grafične predstavitve analize različnega izražanja genov. (A) Diagram MA. Na osi x imamo povprečno število normaliziranih odčitkov, na osi y pa l2FC. Z rdečo so označeni geni s povečanim izražanjem, z modro geni z zmanjšanim izražanjem, s sivo so predstavljeni geni, ki niso statistično značilni. (B) Vulkanski diagram. Na osi x imamo l2FC, na osi y pa -$\log_{10}$(popravljene vrednosti p). Z rdečo so označeni geni s povečanim izražanjem, z modro geni z zmanjšanim izražanjem, s sivo so predstavljeni geni, ki niso statistično značilni. Slika je bila pripravljena z BioRenderjem. l2FC, dvojiški logaritem spremembe.

Figure 2: Example of a graphical representation of differentially expressed genes. (A) MA plot. The x-axis represents mean expression values of genes, the y-axis represents l2FC. Red represents upregulated genes, blue represents downregulated genes and gray represents genes that are not statistically significant. (B) Volcano plot. The x-axis represents l2FC, the y-axis represents -$\log_{10}$(adjusted p-value). Red represents upregulated genes, blue represents downregulated genes and gray represents genes that are not statistically significant. The image was created using BioRender. l2FC, log2 fold change.

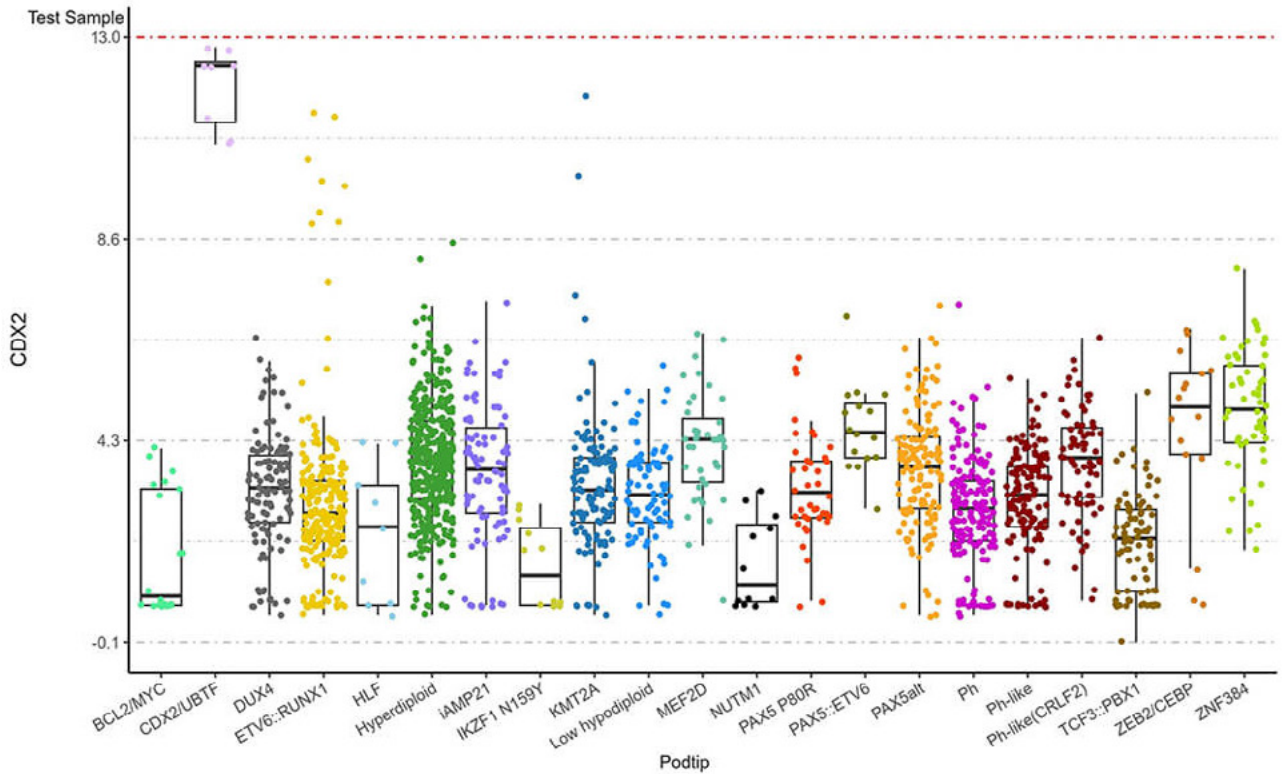
POMANJKLJIVOSTI IN OMEJITVE

Sekvenciranje RNA prinaša mnogo priložnosti, kot vsaka metoda pa ima tudi svoje omejitve in pomanjkljivosti. V primerjavi z DNA je RNA bolj podvržena razgradnji zaradi vseprisotnih RNaz, kar zahteva poseben transport vzorcev (čim hitrejši in na ledu) in previdnost pri celotnem procesu, od izolacije RNA do koraka reverzne transkripcije v postopku priprave knjižnice. V primeru slabše kakovosti izolirane RNA (nižja vrednost RIN) je treba postopek priprave knjižnice prilagoditi z izpustitvijo dodatne fragmentacije (46). Zelo pomembna je tudi pravilna izbira vzorca glede na biološko vprašanje, na katerega želimo odgovoriti. Medtem ko je DNA načeloma enaka v vsaki celici določenega organizma, pa je izražanje genov tkivno specifično (47), torej je kri manj oz. neprimerna, če želimo analizirati transkriptom pri bolezni, ki se na primer izraža v mišičnem tkivu. Prav tako je lahko izražanje genov različno v vsaki posamezni celici, vendar s sekvenciranjem celotne RNA iz tkiva (angl. *bulk RNA sequencing*) izgubimo te podatke in dobimo samo sliko povprečnega izražanja. Za naslavljanje tega problema se že uporabljajo tehnike sekvenciranja posameznih celic (angl. *single-cell sequencing*) (48). Previdni moramo biti tudi pri analiziranju variant, saj lahko pogosto spregledamo variante z nizko frekvenco. Prav tako smo omejeni na variante v kodirajočih delih genoma in na variante v tistih genih, ki so dejansko izraženi (49).

KLINIČNI PRIMER

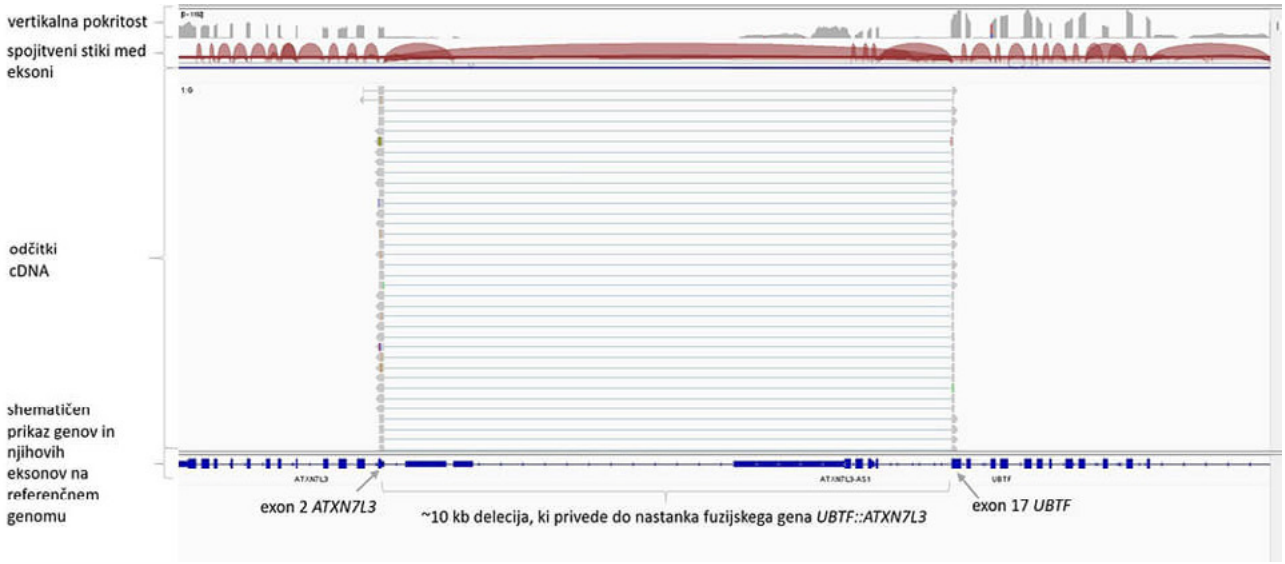
Za prikaz uporabe sekvenciranja RNA v diagnostiki navajamo primer bolnika z B-ALL, pri katerem s standardnimi diagnostičnimi preiskavami (kariotipizacija, fluorescentna *in situ* hibridizacija in multipleksni PCR) nismo našli nobenih ponavljajočih se genetskih sprememb. Raziskava je bila opravljena v skladu z načeli Helsniško-Tokijske deklaracije; preiskovanci so se strinjali z vključitvijo v raziskavo in so podpisali izjavo o zavestni in svobodni privolitvi k sodelovanju v raziskavi po poučitvi. Pri bolniku je prišlo do zgodnjega ekstramedularnega relapsa bolezni z nizko infiltracijo v kostnem mozgu (3 % blastnih celic). Takrat je bilo naročeno sekvenciranje RNA iz vzorca kostnega mozga, odvzetega ob diagnozi, in tistega, odvzetega ob ponovitvi bolezni. Iz vzorcev smo izolirali RNA in pripravili knjižnico za sekvenciranje celotnega transkriptoma. Najprej smo odstranili rRNA, ki predstavlja večino RNA v celicah, vendar nas pri analizi ne zanima. Nato smo preostalo RNA fragmentirali in prepisali v cDNA. Temu je sledila priprava knjižnic, podobna postopku priprave knjižnic za sekvenciranje celotnega eksoma in genoma. Na oba konca cDNA smo ligirali adapterska zaporedja, s pomočjo katerih se fragmenti cDNA lahko vežejo na pretočno celico sekvenatorja. Ta zaporedja vsebujejo tudi zaporedja za unikatno označevanje vzorcev ("molekularna črtna koda"). Po pripravi knjižnice RNA smo izvedli sekvenciranje s sekvenatorjem Illumina NovaSeq6000. Po končanem sekvenciranju smo z bioinformatsko analizo pretvorili svetlobne signale v zaporedje baz in tako določili nukleotidno zaporedje vseh fragmentov cDNA, s pomočjo "molekularne črtne kode" pa smo določili odčitke, ki so pripadali preiskovanima vzorcema. S specifičnim bioinformatskim orodjem smo odčitke nalegali na referenčni genom, temu pa je sledilo štetje količine odčitkov, ki se nalegajo na posamezne gene. Nato smo uporabili program MD-ALL, ki deluje na podlagi strojnega učenja in lahko iz profila izraženih genov vzorce razvrsti v 26 različnih genetskih podtipov B-ALL. Oba vzorca preiskovanega bolnika sta bila uvrščena v podtip CDX2/UBTF, za katerega sta značilna visoko izražanje gena *CDX2* in prisotnost fuzijskega gena *UBTF::ATXN7L3*. Program MD-ALL omogoča tudi vizualizacijo količine izraženega gena v primerjavi z drugimi vzorci v bazi podatkov. Pri obeh vzorcih smo videli povišano izražanje *CDX2* (Slika 3). Potem smo uporabili še bioinformatski cevovod nf-core rnafusion (v. 2.3.4) (50), ki s pomočjo petih različnih orodij (Arriba, FusionCatcher, STAR-Fusion, Squid in Pizzly) določi prisotnost fuzijskih transkriptov, vendar analiza ni pokazala prisotnosti *UBTF::ATXN7L3*. Podatke smo nato pregledali še v interaktivnem genomskem pregledovalniku IGV (angl. *Integrative Genomics Viewer*). Ob pregledu regije na dolgi ročici kromosoma 17 smo našli približno 10 kilobaz veliko delecijo, ki vodi v nastanek iskanega fuzijskega gena (Slika 4). S tem smo pri preiskovanem bolniku potrdili prisotnost genetskega podtipa CDX2/UBTF ob diagnozi in relapsu. Ta podtip je bil prvič opisan šele leta 2022, gre pa za podtip z visokim tveganjem za ponovitev bolezni in odpornost na zdravljenje, pri katerem se priporoča intenzivnejše zdravljenje (51). Zanj je značilen tudi specifičen aberanten imunofenotip z »

odsotnostjo CD10 in prisotnostjo IgM (52), ki je bil viden tudi pri našem bolniku. S standardnimi diagnostičnimi preiskavami te fuzije nismo odkrili, najverjetneje zaradi omejitve algortimov orodij, ki kličejo fuzije, saj imajo lahko ta orodja več problemov z iskanjem intrakromosomskih fuzij, prav tako pa lahko imajo določene kriterije za filtriranje lažno pozitivnih rezultatov. Na detekcijo lahko vplivajo nepopolne anotacije genov, kadar pa imajo geni homologne sekvence, se lahko odčitki nalegajo nepravilno in zato zgrešimo fuzijo. Primer tega bolnika nazorno prikazuje uporabnost sekvenciranja RNA pri bolnikih z B-ALL, sploh pri tistih, pri katerih standardne diagnostične preiskave ne pokažejo nobenih posebnosti (okoli 25 % vseh bolnikov z B-ALL).



Slika 3: Prikaz izražanja *CDX2* pri analiziranem vzorcu, odvzetem ob postavitvi diagnoze ("Test sample") v primerjavi z drugimi vzorci B-ALL v bazi podatkov. Na osi x so navedeni različni genetski podtipi B-ALL, na osi y pa je prikazano izražanje *CDX2* za posamezni vzorec. Gen *CDX2* je v testnem vzorcu zelo visoko izražen (rdeča črtkana črta), kar je tudi značilnost genetskega podtipa CDX2/UBTF.

Figure 3: Showing the *CDX2* expression of the analysed sample, collected at the time of diagnosis ("Test sample") compared to other B-ALL samples in the database. The x-axis shows the different B-ALL genetic subtypes and the y-axis shows the *CDX2* expression for each sample. The *CDX2* gene is highly expressed in the test sample, which is also a characteristic of the CDX2/UBTF genetic subtype.



Slika 4: Slika spojitvenega mesta med eksonom 17 gena *UBTF* in eksonom 2 gena *ATXN7L3* na dolgi ročici kromosoma 17 iz interaktivnega genomskega pregledovalnika IGV (angl. *Integrative Genomics Viewer*). Prikazan je izbrani odsek na humanem genomu, kjer se nahaja omenjena genetska sprememba. Zgornje sivo področje prikazuje vertikalno pokritost, pod tem je z rdečo označeno, med katerimi eksoni je prišlo do spajanja. Spojitveni stik med eksoni prikazuje, da je prišlo do spajanja med genom *ATXN7L3*, ki se nahaja na eksonu 2, in genom *UBTF*, ki se nahaja na eksonu 17. Spodaj so odčitki cDNA, iz katerih prav tako vidimo, da sta bili mesti, ki sta sicer na referenčnem genomu med seboj oddaljeni 10 kilobaznih parov in ju ločuje več eksonov, pri tem bolniku prisotni na skupnem fragmentu RNA oziroma cDNA, ki smo ga sekvencirali (dolg ~300 baznih parov). Do tega je prišlo zaradi približno 10 kb delecije, ki vključuje eksone na 3'-koncu gena *UBTF* (eksoni 18–21) in večino intergenske regije med *UBTF* in *ATXN7L3*, posledično pa pride do nastanka fuzijskega gena *UBTF::ATXN7L3*.

Figure 4: Image of the junction site between exon 17 of the *UBTF* gene and exon 2 of the *ATXN7L3* gene on the long arm of chromosome 17 from the Integrative Genomics Viewer (IGV). The selected section of the human genome, where this genetic alteration is located, is shown. The upper gray area represents the vertical coverage, and below it, the junctions between exons are marked in red. From this section, we can see that splicing occurred between exon 2 of the *ATXN7L3* gene and exon 17 of the *UBTF* gene. Below are the cDNA reads, which also show that the sites, which are 10 kilobase pairs apart on the reference genome and separated by several exons, are present in this patient on a common RNA or cDNA fragment that we sequenced (approximately 300 base pairs long). This resulted from an approximately 10 kb deletion that includes exons at the 3'-end of the *UBTF* gene (exons 18–21) and most of the intergenic region between *UBTF* and *ATXN7L3*, leading to the formation of the fusion gene *UBTF::ATXN7L3*.

ZAKLJUČEK

Sekvenciranje RNA je molekularna tehnika, ki se uporablja tako pri osnovnih kot tudi kliničnih raziskavah. Z njeno pomočjo lahko analiziramo izražanje genov, kar omogoča odkrivanje novih bioloških označevalcev, prav tako pa ima pomembno vlogo pri izboljšanju razumevanja kompleksnejših boleznih, kar je bilo predstavljeno na opisanem kliničnem primeru. S konstantnim tehnološkim napredkom postaja sekvenciranje RNA vse bolj dostopno in uporabno. S povezovanjem sekvenciranja posameznih celic s prostorsko transkriptomiko lahko dobimo vpogled v zgradbo tkiv, heterogenost celičnih populacij in vzorce izražanja genov. Takšen napredek izboljša razumevanje zapletenih bioloških sistemov in prispeva k razumevanju poteka bolezni. Predvidevamo, da bo uporaba sekvenciranja RNA v kliničnih laboratorijih naraščala. Kot je bilo prikazano na kliničnem primeru, se sekvenciranje RNA že uporablja v primerih, ko s standardnimi diagnostičnimi preiskavami ne odkrijemo nobenih genetskih sprememb. Na podlagi določenega molekularnega profila posameznih bolnikov lahko pomaga pri diagnostiki in usmerja odločitve o zdravljenju. Ta pristop se lahko uporablja predvsem pri raku, nevroloških in nevrodegenerativnih, imunskih ter drugih kompleksnih bolezni, kjer uporaba sekvenciranja RNA v zadnjih letih močno narašča, sploh v primerih, pri katerih je izplen sekvenciranja celotnega humanega eksoma in genoma negativen. Z integracijo sekvenciranja RNA z drugimi metodami lahko preučujemo interakcije med RNA in proteini oz. med RNA in RNA. S tem se izboljša tudi naše razumevanje regulatornih mehanizmov, »

kar bo v prihodnosti odprlo nove možnosti, predvsem na terapevtskem področju. Z napredkom strojnega učenja in umetne inteligence se izboljšuje tudi sposobnost interpretacije velike količine kompleksnih podatkov, ki jih dobimo s sekvenciranja RNA. To bo v prihodnosti dodatno olajšalo odkrivanje novih bioloških označevalcev in terapevtskih možnosti.

LITERATURA

1. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
2. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20(11):631–56.
3. Chen G, Shi T, Shi L. Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci. 2017;60(2):116–25.
4. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
5. Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell. 2009;136(4):777–93.
6. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer. 2015;15(6):371–81.
7. Adetunji MO, Lamont SJ, Abasht B, Schmidt CJ. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLoS One. 2019;14(9):e0216838.
8. Liu X, Harada S. RNA Isolation from Mammalian Samples. Curr Protoc Mol Biol. 2013;103(1):4.16.1–16.
9. Scholes AN, Lewis JA. Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses. BMC Genomics. 2020;21(1):249.
10. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
11. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7:3.
12. O'Neil D, Glowatz H, Schlumpberge M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013;103(1):4–19.
13. Zhao S, Zhang Y, Gamini R, Zhang B, Von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Sci Rep. 2018;8(1):4781.
14. Chen Z, Duan X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods Mol Biol. 2011;733:93–103.
15. Morlan JD, Qu K, Sinicropi D V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One. 2012;7(8):e42882.
16. Wery M, Descrimes M, Thermes C, Gautheret D, Morillon A. Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. Methods. 2013;63(1):25–31.
17. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA. 2017;8(1):10.1002/wrna.1364.
18. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: A method for comprehensive transcriptome analysis. Curr Protoc Mol Biol. 2010;89(1):4.11.1–13.
19. Borodina T, Adjaye J, Sultan M. A strand-specific library preparation protocol for RNA sequencing. Methods Enzymol. 2011;500:79–98.
20. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37(18):e123.
21. Zhao S, Zhang Y, Gordon W, Quan J, Hualin X, Du S, et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genom. 2015;16(1):675.
22. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2011;9(1):72–4.
23. Fan HC, Fu GK, Fodor SPA. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015;347(6222):1258367.
24. Modi A, Vai S, Caramelli D, Lari M. The Illumina Sequencing Protocol and the NovaSeq 6000 System. Methods Mol Biol. 2021;2242:15–42.
25. Bulk RNA-seq Data Standards and Processing Pipeline [Internet]. [assessed on 29.7.2024]. Available from: https://www.encodeproject.org/data-standards/rna-seq/long-rnas/.
26. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
27. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851–8.
28. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
29. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008;321(5891):956–60.
30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
31. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
33. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.
34. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
35. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.
36. Johnson KA, Krishnan A. Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. Genome Biol. 2022;23(1):1.

37. Li X, Brock GN, Rouchka EC, Cooper NGF, Wu D, OToole TE, et al. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. PLoS One. 2017;12(5).
38. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. J Vis Exp. 2021;2021(175).
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
40. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
41. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res. 2011;21(12):2213.
42. Ghosh D. Incorporating the empirical null hypothesis into the Benjamini-Hochberg procedure. Stat Appl Genet Mol Biol. 2012;11(4):31.
43. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12(12):e0190152.
44. Zhao T, Wang Z. GraphBio: A shiny web app to easily perform popular visualization analysis for omics data. Front Genet. 2022;13:957317.
45. Wodrich MD, Sawatlon B, Busch M, Corminboeuf C. The genesis of molecular volcano plots. ACS Cent Sci. 2019;5(5):796–804.
46. Precautions for Handling of RNA [Internet]. [assessed on 12.6.2024]. Available from: https://lifescience.roche.com/global/en/article-listing/article/precautions-for-handling-of-rna.html.
47. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. 4th ed. New York: Garland Science; 2002.
48. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021;13(1):36.
49. Kaya C, Dorsaint P, Mercurio S, Campbell AM, Eng KW, Nikiforova MN, et al. Limitations of detecting genetic variants from the RNA sequencing data in tissue and fine-needle aspiration samples. Thyroid. 2021;31(4):589–95.
50. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–278.
51. Bastian L, Hartmann AM, Beder T, Hänzelmann S, Kässens J, Bultmann M, et al. UBTF::ATXN7L3 gene fusion defines novel B cell precursor ALL subtype with CDX2 expression and need for intensified treatment. Leukemia. 2022;36(6):1676–80.
52. Kimura S, Montefiori L, Iacobucci I, Zhao Y, Gao Q, Paietta EM, et al. Enhancer retargeting of CDX2 and UBTF::ATXN7L3 define a subtype of high-risk B-progenitor acute lymphoblastic leukemia. Blood. 2022;139(24):3519–31.