**Optimalizácia postupu bioinformatického spracovania sekvenačných dát pre spoľahlivejšie neinvazívne prenatálne testovanie na prítomnosť chromozómových porúch**

Mgr. Rami Saade[1]
(patologická anatómia a súdne lekárstvo)

Školiteľ: prof. RNDr. Vanda Repiská, PhD. [1]
[1]Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a UN Bratislava

## Úvod

V súčasnosti je stále aktuálnou témou problematika prenatálnej diagnostiky v súvislosti s rizikami, ktoré so sebou prináša. V dobe, kedy si mladí ľudia nemôžu dovoliť založiť rodinu vo veku, v ktorom ich už rodičia držali v náručí, je veľmi dôležité aktuálne vnímať biologické riziká spojené s trendom tehotenstiev v pokročilom veku. Práve vek matky je jedným z najzásadnejších faktorov zvyšujúcich pravdepodobnosť výskytu chromozomálnych aberácií u dieťaťa (1). Bohužiaľ, moderná medicína zatiaľ nepozná nijaký spôsob, ktorým by bolo možné takéto ochorenia vyliečiť, alebo aspoň ich efekt zmierniť a je veľmi nepravdepodobné, aby sa v dohľadnej dobe objavila metodika na riešenie danej problematiky. Jediným aktuálnym spôsobom ako predísť nečakanému narodeniu trizomického dieťaťa je včasná diagnostika. Negatívny výsledok nepochybne poteší každého nastávajúceho rodiča a zároveň navodí psychickú pohodu do rodiny očakávajúcej zdravý prírastok. Naopak pozitívny výsledok býva zdrvujúcou správou, avšak v konečnom dôsledku je stále na rozhodnutí rodičov, hlavne matky, či je pripravená podstúpiť interrupciu, alebo sa rozhodne dieťa donosiť s vedomím, že ju čaká nesmierne náročné obdobie spojené so starostlivosťou o postihnuté dieťa.

Odhliadnuc od spomínaných možných výsledkov prenatálnych vyšetrení, je najhoršou možnosťou, ak je výsledok takéhoto vyšetrenia nesprávny a teda sa nezhoduje so skutočnosťou. V tomto prípade je úplne jedno, či je výsledok falošne pozitívny alebo falošne negatívny, obidve možnosti sú neprijateľné a zanechajú nenapraviteľné následky. Našťastie zaužívané prenatálne diagnostické metódy sa vyznačujú takmer 100% senzitivitou a špecificitou a nesprávne vyhodnotenie takéhoto vyšetrenia je skôr raritou. Tu však nastáva iný problém, paradoxne vychádzajúci zo samotného vyšetrenia. Tieto vyšetrenia sú invazívne, čiže dochádza k zásahu do tela matky, ktoré je počas tehotenstva mimoriadne citlivé na akékoľvek nežiaduce vonkajšie podnety. Takýmto invazívnym zásahom či už pri amniocentéze, alebo pri odbere choriových klkov je tehotenstvo vystavené zvýšenému riziku komplikácií jeho priebehu, ktoré môže v hraničnom prípade, štatisticky zhruba v 1% prípadov, končiť potratom (2). Z tohto dôvodu bolo nutné ponúknuť tehotným ženám alternatívu, ktorá bude informovať dostatočne spoľahlivo o karyotype ich dieťaťa a zároveň neohrozí priebeh ich tehotenstva. Takouto alternatívou sú skríningové metódy, ktorých úlohou je poskytnúť dostatočne spoľahlivé a jednoznačné výsledky na to, aby nebola žena nútená podstúpiť invazívne prenatálne diagnostické metódy.

V súčasnosti je preukázateľne najspoľahlivejšia prenatálna skríningová metóda neinvazívneho prenatálneho testovania (NIPT), založená na masívnom paralelnom sekvenovaní fragmentov cirkulujúcej voľnej fetálnej DNA izolovanej z krvnej plazmy matky. Keďže sa jedná o pomerne krátko využívanú metódu (prvýkrát zavedená do klinickej praxe v roku 2011 v Hongkongu), stále je nutné zdokonaľovať jednotlivé kroky súvisiace s laboratórnym spracovaním vzorky, ako aj s *in silico* spracovaním sekvenačných dát (3).

Cieľom optimalizácie jednotlivých parametrov tejto metodiky je nielen poskytnúť alternatívu pre konvenčne zaužívané skríningové metódy, ale časom dosiahnuť potenciálnu konkurencieschopnosť v rámci diagnostických metód - amniocentézy a odberu choriových klkov, naproti ktorým NIPT nepredstavuje riziko ohrozenia priebehu tehotenstva. Pre dosiahnutie tohto cieľa je nevyhnuté, aby pri použití optimalizovaných podmienok vykazovali výsledky vyšetrenia (v ideálnom prípade) 100% senzitivitu a špecificitu.

**Materiál a metódy**

Krvné vzorky boli ambulantne odobrané 100 tehotným ženám rôzneho veku v rozličnom gestačnom týždni (všetky aspoň po 11. týždni) do skúmavky s EDTA. Zo vzoriek bola dvojstupňovou centrifugáciou oddelená krvná plazma, z ktorej bola izolovaná celková plazmatická DNA pomocou kitu QIAamp DNA Blood Mini Kit (Qiagen, DE). Izolovaná DNA bola spracovaná do fragmentovanej DNA knižnice s použitím kitu TruSeq Nano DNA Library Prep Kit (Illumina), ktorá bola následne sekvenovaná NGS sekvenátorom Illumina MiSeq. Dáta zo sekvenovania boli primárne spracované podľa štandardnej analytickej pipeline na detekciu fetálnych aneuploídií, aby sme výsledky našej optimalizácie vedeli spätne porovnať s postupom využívaným v klinickej praxi. Pracovali sme s 90 vzorkami pochádzajúcimi zo zdravých plodov a s 10 vzorkami od trizomických.

Na základe zhodnotenia kvality čítania pôvodných výstupných FASTQ súborov sme odfiltrovali čítania s kvalitou čítania nižšou ako Phred=30 zo začiatku aj z konca čítania a zároveň sme nastavili "posúvajúce okno", ktoré vyhodnocovalo kvalitu čítania po 3 susediacich bázach a ak klesla kvalita čítania pod Phred=30, tak takéto čítanie program zahodil. Zároveň sme nastavili minimálnu dĺžku zachovaného čítania na 30 bp.

Všetky vzorky sme následne namapovali na referenčné genómy hg19 a GRCh38 (pre prehľadnosť označujeme ako hg38) pomocou mapovacích algoritmov Bowtie2, BWA a SOAP2. Tým sme získali 6 rôznych kombinácii - mapovací algoritmus/referenčný genóm. Pomocou nástroja v programovacom jazyku python, ktorý bol vytvorený podľa našich požiadaviek, sme vykonali LOESS korekciu (napravuje prípadný GC-bias) a taktiež sme testovali vplyv dĺžky ponechaných fragmentov na výsledné z-skóre analýzy. Vyhodnotenie počtu namapovaných fragmentov na jednotlivé chromozómy a zároveň kontrolu kvality mapovania sme vykonali pomocou programu QualiMap.

Na výpočet z-skóre sme vyskúšali a porovnali 6 rôznych metód, kde sme ako trénovacie vzorky použili 80 vzoriek od zdravých plodov. Základný rozdiel spočíval v porovnaní percentuálneho zastúpenia fragmentov namapovaných na chromozóm 21 voči ostatným chromozómom, vrátane mitochondriálneho, alebo len voči autozómom. Ďalším rozdielom bolo zahrnutie rôznych štatistických metód, pričom:

$$p_1 = \frac{počet\ fragmentov\ chr21\ v\ trénovacích\ vzorkách}{počet\ všetkých\ fragmentov\ v\ trénovacích\ vzorkách}$$

$$p_2 = \frac{počet\ fragmentov\ chr21\ v\ trénovacích\ vzorkách}{počet\ fragmentov\ autozómov\ v\ trénovacích\ vzorkách}$$

$$t_1 = \frac{počet\ fragmentov\ chr21\ v\ testovanej\ vzorke}{počet\ všetkých\ fragmentov\ v\ testovanej\ vzorke}$$

$$t_2 = \frac{počet\ fragmentov\ chr21\ v\ testovanej\ vzorke}{počet\ fragmentov\ autozómov\ v\ testovanej\ vzorke}$$

Výpočty z-skóre:

$$z_{1a} = \frac{t_1 - aritmetický\ priemer(p_1)}{štandardná\ odchýlka(p_1)} \qquad (4)$$

$$z_{1b} = \frac{t_2 - aritmetický\ priemer(p_2)}{štandardná\ odchýlka(p_2)} \qquad \text{(upravené podľa 4)}$$

$$z_{2a} = \frac{t_1 - medián(p_1)}{priemerná\ absolútna\ odchýlka\ (p_1)}$$

$$z_{2b} = \frac{t_2 - medián(p_2)}{priemerná\ absolútna\ odchýlka\ (p_2)}$$

$$z_{3a} = \frac{t_1 - medián(p_1)}{absolútna\ odchýlka\ od\ mediánu(p_1)} \qquad (5)$$

$$z_{3b} = \frac{t_2 - medián(p_2)}{absolútna\ odchýlka\ od\ mediánu(p_2)} \qquad \text{(upravené podľa 5)}$$

**Výsledky**

Rozhodujúcimi parametrami pri selektovaní jednotlivých postupov analýzy bola podmienka, aby sa v danom prípade nevyskytovali žiadne falošne pozitívne ani falošne negatívne vzorky a aby parameter DIFFz (rozdiel medzi priemerným z-skóre zdravých a trizomických vzoriek) bol čo najvyšší. Na základe našich výsledkov sme ďalej nepracovali s metódami výpočtu z-skóre $z_{1a}$ a $z_{1b,}$ keďže DIFFz bol všeobecne nízky (tabuľka č.1 vľavo). Pre objektívnejšie zhodnotenie vhodnej kombinácie mapovacieho algoritmu a referenčného genómu sme sa rozhodli odfiltrovať čítania s kvalitou mapovania nižšou ako Phred=40 (pri SOAP2 Phred=30) a *in silico* sme vyselektovali fragmenty s dĺžkou 130-170 bp (tabuľka č.1 vpravo).

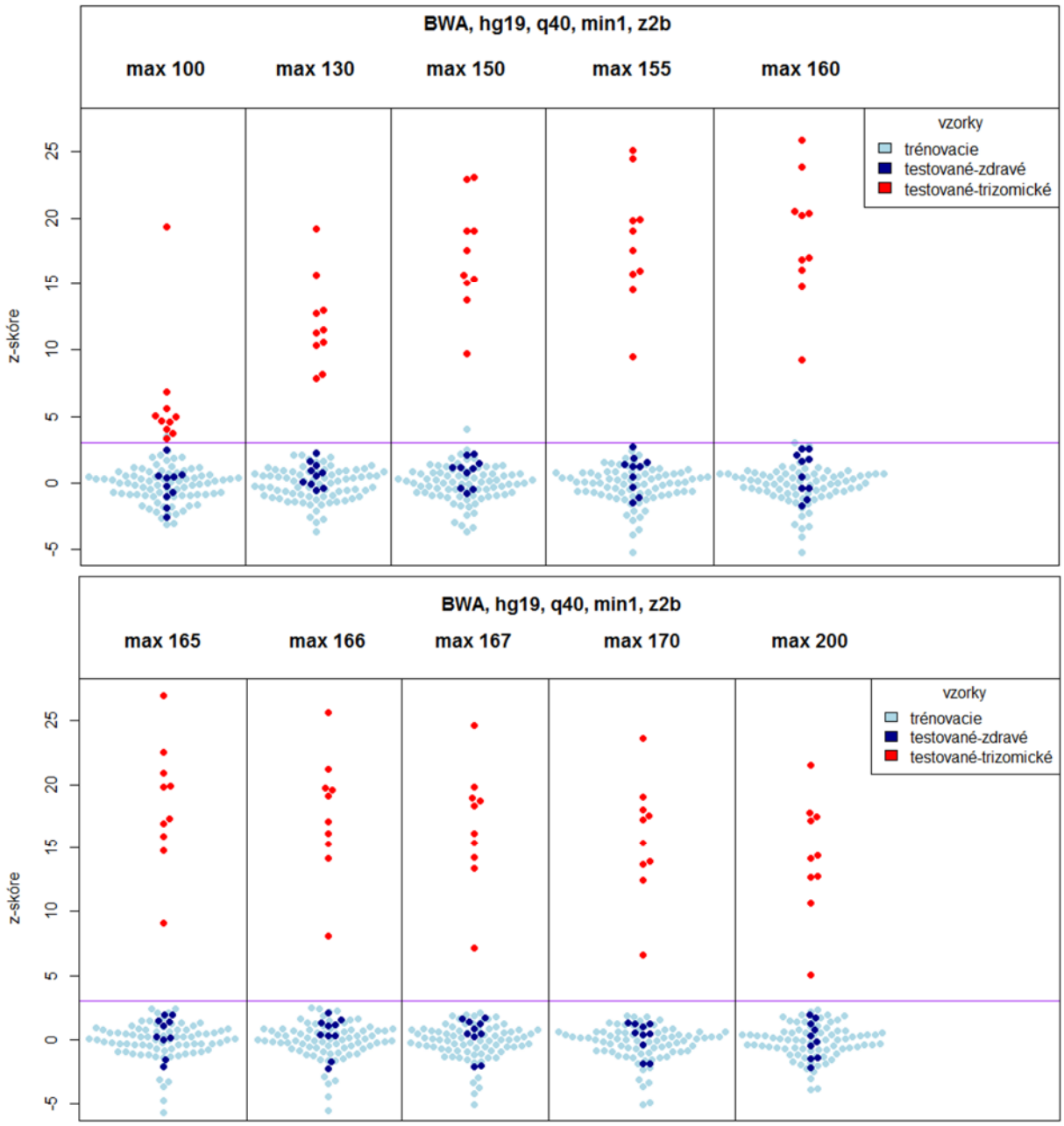
**Tabuľka č.1:** Vľavo: sumár výsledkov porovnania 6 metód na výpočet z-skóre. Vpravo: sumár výsledkov porovnania z-skóre kombinácií mapovacích algoritmov a referenčných genómov, vypočítaného 4 rôznymi metódami (vynechané $z_{1a}$ a $z_{1b}$) s minimálnou kvalitou mapovania Phred=40 (pri SOAP2 Phred=30) a *in silico* vyselektovanými fragmentmi s dĺžkou 130-170 bp .

| map | ref | met | $FP_{RH}$ | $FP_{TH}$ | $FN_{TT}$ | $\mu(z_{TH})$ | $\mu(z_{TT})$ | DIFFz |
|---|---|---|---|---|---|---|---|---|
| Bowtie2 | hg19 | $z_{1a}$ | - | - | 1 | 0,07 | 8,25 | 8,18 |
| Bowtie2 | hg19 | $z_{1b}$ | - | - | - | 0,01 | 8,52 | 8,51 |
| Bowtie2 | hg19 | $z_{2a}$ | - | - | - | 0,01 | 9,94 | 9,92 |
| Bowtie2 | hg19 | $z_{2b}$ | - | - | - | -0,15 | 10,35 | 10,49 |
| Bowtie2 | hg19 | $z_{3a}$ | - | - | - | 0,01 | 10,8 | 10,79 |
| Bowtie2 | hg19 | $z_{3b}$ | - | - | - | -0,16 | 11,42 | 11,58 |
| Bowtie2 | hg38 | $z_{1a}$ | - | - | 9 | -0,4 | 1,7 | 2,1 |
| Bowtie2 | hg38 | $z_{1b}$ | - | - | 9 | -0,42 | 1,75 | 2,17 |
| Bowtie2 | hg38 | $z_{2a}$ | - | - | 9 | -0,35 | 2,16 | 2,51 |
| Bowtie2 | hg38 | $z_{2b}$ | - | - | 9 | -0,35 | 2,23 | 2,58 |
| Bowtie2 | hg38 | $z_{3a}$ | 1 | - | 8 | -0,38 | 2,38 | 2,76 |
| Bowtie2 | hg38 | $z_{3b}$ | 2 | - | 6 | -0,41 | 2,56 | 2,97 |
| BWA | hg19 | $z_{1a}$ | - | - | - | 0,12 | 8,9 | 8,78 |
| BWA | hg19 | $z_{1b}$ | - | - | - | 0,03 | 9,36 | 9,33 |
| BWA | hg19 | $z_{2a}$ | - | - | - | -0,13 | 10,24 | 10,38 |
| BWA | hg19 | $z_{2b}$ | - | - | - | -0,08 | 11,12 | 11,2 |
| BWA | hg19 | $z_{3a}$ | - | - | - | -0,15 | 11,24 | 11,39 |
| BWA | hg19 | $z_{3b}$ | - | 1 | - | -0,1 | 13,08 | 13,18 |
| BWA | hg38 | $z_{1a}$ | - | - | 1 | 0,22 | 6,91 | 6,69 |
| BWA | hg38 | $z_{1b}$ | - | - | 1 | 0,16 | 7,27 | 7,11 |
| BWA | hg38 | $z_{2a}$ | - | - | 1 | 0,27 | 8,44 | 8,16 |
| BWA | hg38 | $z_{2b}$ | - | 1 | 1 | 0,2 | 8,78 | 8,58 |
| BWA | hg38 | $z_{3a}$ | 1 | 1 | 1 | 0,33 | 10,09 | 9,76 |
| BWA | hg38 | $z_{3b}$ | - | 1 | 1 | 0,24 | 10,24 | 10,01 |
| SOAP2 | hg19 | $z_{1a}$ | - | - | - | 0,22 | 9,4 | 9,17 |
| SOAP2 | hg19 | $z_{1b}$ | - | - | - | 0,14 | 9,88 | 9,74 |
| SOAP2 | hg19 | $z_{2a}$ | - | - | - | 0,21 | 11,4 | 11,19 |
| SOAP2 | hg19 | $z_{2b}$ | - | - | - | 0,16 | 12,34 | 12,18 |
| SOAP2 | hg19 | $z_{3a}$ | - | - | - | 0,23 | 12,88 | 12,65 |
| SOAP2 | hg19 | $z_{3b}$ | - | - | - | 0,2 | 14,62 | 14,42 |
| SOAP2 | hg38 | $z_{1a}$ | - | - | 1 | 0,29 | 7,8 | 7,51 |
| SOAP2 | hg38 | $z_{1b}$ | - | - | 1 | 0,22 | 8,17 | 7,95 |
| SOAP2 | hg38 | $z_{2a}$ | - | - | 1 | 0,41 | 9,38 | 8,97 |
| SOAP2 | hg38 | $z_{2b}$ | - | - | 1 | 0,33 | 9,85 | 9,52 |
| SOAP2 | hg38 | $z_{3a}$ | - | - | 1 | 0,43 | 9,93 | 9,5 |
| SOAP2 | hg38 | $z_{3b}$ | - | - | - | 0,35 | 10,63 | 10,27 |

| map | ref | met | q | fmin | fmax | $FP_{RH}$ | $FP_{TH}$ | $FN_{TT}$ | $\mu(z_{TH})$ | $\mu(z_{TT})$ | DIFFz |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bowtie2 | hg19 | $z_{2a}$ | 40 | 130 | 170 | - | - | - | -0,12 | 11,56 | 11,68 |
| Bowtie2 | hg19 | $z_{2b}$ | 40 | 130 | 170 | - | - | - | -0,19 | 12,5 | 12,69 |
| Bowtie2 | hg19 | $z_{3a}$ | 40 | 130 | 170 | - | - | - | -0,14 | 13,26 | 13,4 |
| Bowtie2 | hg19 | $z_{3b}$ | 40 | 130 | 170 | - | - | - | -0,24 | 15,77 | 16,01 |
| Bowtie2 | hg38 | $z_{2a}$ | 40 | 130 | 170 | - | - | - | -0,08 | 11,76 | 11,84 |
| Bowtie2 | hg38 | $z_{2b}$ | 40 | 130 | 170 | - | - | - | -0,12 | 12,77 | 12,9 |
| Bowtie2 | hg38 | $z_{3a}$ | 40 | 130 | 170 | - | - | - | -0,09 | 12,94 | 13,02 |
| Bowtie2 | hg38 | $z_{3b}$ | 40 | 130 | 170 | - | - | - | -0,16 | 16,57 | 16,73 |
| BWA | hg19 | $z_{2a}$ | 40 | 130 | 170 | - | - | - | 0,12 | 13,01 | 12,89 |
| BWA | hg19 | $z_{2b}$ | 40 | 130 | 170 | - | - | - | 0,15 | 14,11 | 13,96 |
| BWA | hg19 | $z_{3a}$ | 40 | 130 | 170 | 1 | - | - | 0,16 | 17,57 | 17,41 |
| BWA | hg19 | $z_{3b}$ | 40 | 130 | 170 | - | - | - | 0,18 | 17,08 | 16,9 |
| BWA | hg38 | $z_{2a}$ | 40 | 130 | 170 | 1 | - | - | 0,28 | 12,35 | 12,07 |
| BWA | hg38 | $z_{2b}$ | 40 | 130 | 170 | 1 | - | - | 0,06 | 12,87 | 12,8 |
| BWA | hg38 | $z_{3a}$ | 40 | 130 | 170 | 6 | - | - | 0,44 | 19,67 | 19,23 |
| BWA | hg38 | $z_{3b}$ | 40 | 130 | 170 | 3 | - | - | 0,09 | 18,68 | 18,59 |
| SOAP2 | hg19 | $z_{2a}$ | 30 | 130 | 170 | - | - | - | 0,33 | 11,88 | 11,56 |
| SOAP2 | hg19 | $z_{2b}$ | 30 | 130 | 170 | - | - | - | 0,15 | 12,2 | 12,04 |
| SOAP2 | hg19 | $z_{3a}$ | 30 | 130 | 170 | 3 | - | - | 0,4 | 14,23 | 13,83 |
| SOAP2 | hg19 | $z_{3b}$ | 30 | 130 | 170 | - | - | - | 0,18 | 14,42 | 14,24 |
| SOAP2 | hg38 | $z_{2a}$ | 30 | 130 | 170 | 1 | - | 1 | 0,24 | 9,3 | 9,06 |
| SOAP2 | hg38 | $z_{2b}$ | 30 | 130 | 170 | - | - | 1 | 0,19 | 9,59 | 9,4 |
| SOAP2 | hg38 | $z_{3a}$ | 30 | 130 | 170 | 1 | - | - | 0,28 | 10,85 | 10,57 |
| SOAP2 | hg38 | $z_{3b}$ | 30 | 130 | 170 | 1 | - | - | 0,21 | 10,83 | 10,62 |

map = mapovací algoritmus
ref = referenčný genóm
met = metóda na výpočet z-skóre
q = minimálna kvalita mapovania
fmin = minimálna dĺžka fragmentu
fmax = maximálna dĺžka fragmentu
$FP_{RH}$ = falošne pozitívne trénovacie vzorky - zdravé
$FP_{TH}$ = falošne pozitívne testované vzorky - zdravé
$FN_{TT}$ = falošne negatívne testované vzorky - trizomické
$\mu(z_{TH})$ = priemerné z-skóre zdravých testovaných vzoriek
$\mu(z_{TT})$ = priemerné z-skóre trizomických testovaných vzoriek
DIFFz = rozdiel medzi $\mu(z_{TT})$ a $\mu(z_{TH})$

V ďalšom kroku sme pracovali s čítaniami namapovanými pomocou BWA na referenčný genóm hg19 a z-skóre sme počítali metódou $z_{2b}$, keďže táto kombinácia sa ukázala pri predošlých analýzach ako spoľahlivá a otestovali sme vplyv rozdielnej maximálnej dĺžky fragmentov na z-skóre. Na výpočet z-skóre sme použili metódu $z_{2b}$, minimálnu kvalitu mapovania sme nastavili na Phred=40 a *in silico* sme vyselektovali fragmenty s minimálnou dĺžkou 1 bp, a testovali sme maximálne dĺžky fragmentov 100 bp, 130 bp, 150 bp, 155 bp, 160 bp, 165 bp, 166 bp, 167 bp, 170 bp a 200 bp (obrázok č.1).




**Obrázok č.1**: Porovnanie z-skóre čítaní s maximálnou dĺžkou fragmentov 100 bp, 130 bp, 150 bp, 155 bp, 160 bp 165 bp, 166 bp, 167 bp, 170 bp a 200 bp. Čítania boli namapované na referenčný genóm hg19 pomocou BWA, minimálna kvalita mapovania bola Phred=40, minimálna dĺžka fragmentov bola 1 bp a z-skóre bolo počítané metódou $z_{2b}$.

Optimalizované nastavenia kvality mapovania, *in silico* selekcie dĺžky fragmentov a metódy výpočtu z-skóre sme spätne otestovali pri kombináciách mapovacích algoritmov s referenčnými genómami a porovnali sme výsledné z-skóre, pričom sme potvrdili správnosť vhodného výberu kombinácie - mapovací algoritmus/referenčný genóm (tabuľka č.2).

**Tabuľka č.2:** Sumár výsledkov porovnania z-skóre kombinácií mapovacích algoritmov a referenčných genómov, pričom minimálna kvalita mapovania bola nastavená na Phred=40 (pre SOAP2 Phred=30), *in silico* boli vyselektované fragmenty s dĺžkou 1-165 bp a na výpočet z-skóre bola použitá metóda $z_{2b}$.

| map | ref | met | q | fmin | fmax | $FP_{RH}$ | $FP_{TH}$ | $FN_{TT}$ | $\mu(z_{TH})$ | $\mu(z_{TT})$ | DIFFz |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bowtie2 | hg19 | $z_{2b}$ | 40 | 1 | 165 | - | - | - | 0,11 | 16,99 | 16,87 |
| Bowtie2 | hg38 | $z_{2b}$ | 40 | 1 | 165 | - | - | - | 0,17 | 16,98 | 16,81 |
| BWA | hg19 | $z_{2b}$ | 40 | 1 | 165 | - | - | - | 0,44 | 18,42 | 17,98 |
| BWA | hg38 | $z_{2b}$ | 40 | 1 | 165 | - | - | - | 0,45 | 17,27 | 16,81 |
| SOAP2 | hg19 | $z_{2b}$ | 30 | 1 | 165 | - | - | - | 0,34 | 16,21 | 15,87 |
| SOAP2 | hg38 | $z_{2b}$ | 30 | 1 | 165 | - | - | - | 0,45 | 12,85 | 12,39 |

map = mapovací algoritmus
ref = referenčný genóm
met = metóda na výpočet z-skóre
q = minimálna kvalita mapovania
fmin = minimálna dĺžka fragmentu
fmax = maximálna dĺžka fragmentu
$FP_{RH}$ = falošne pozitívne trénovacie vzorky - zdravé
$FP_{TH}$ = falošne pozitívne testované vzorky - zdravé
$FN_{TT}$ = falošne negatívne testované vzorky - trizomické
$\mu(z_{TH})$ = priemerné z-skóre zdravých testovaných vzoriek
$\mu(z_{TT})$ = priemerné z-skóre trizomických testovaných vzoriek
DIFFz = rozdiel medzi $\mu(z_{TT})$ a $\mu(z_{TH})$

## Diskusia

Technológia sekvenovania budúcej generácie (NGS) založená na princípe masívneho paralelného sekvenovania priniesla množstvo nových možností aplikácie do oblasti vedy a výskumu, a zároveň aj do klinickej praxe, kde vykazuje veľký potenciál ako (zatiaľ) alternatívna metóda k bežne zaužívaným postupom. Potenciál klinického uplatnenia NGS sa preukázal pri celogenómovom sekvenovaní s nízkym pokrytím, ktoré sa pod názvom NIPT využíva na skríning genetických anomálií plodu. Pre tehotnú ženu nepredstavuje toto vyšetrenie o nič väčšie riziko než konvenčne zaužívaný prenatálny skríning biomarkerov, ktorý sa takisto vyšetruje z periférnej krvi. Cieľovými analyzovanými molekulami sú fragmenty voľnej cirkulujúcej fetálnej (cffDNA), ktoré sa prirodzene nachádzajú v krvnej plazme tehotnej ženy, kde sa dostávajú regulovanou bunkovou smrťou z buniek trofoblastu z placenty (6).

Najdôležitejším parametrom na vyhodnotenie štatistickej významnosti kvantitatívneho zastúpenia fragmentov sledovaných chromozómov vo vzorke je v prípade NIPT z-skóre. Na výpočet z-skóre sa často používa metóda, ktorú vo svojej publikácii navrhli Lau et. al, a ktorú označujeme ako $z_{1a}$. Takisto sme sa rozhodli porovnať metódu aplikovanú v klinickej praxi, ktorú v online príručke popisuje firma Eurofins LifeCodexx v rámci ponúkaného NIPT trizómii 13., 18. a 21. chromozómu pod názvom PrenaTest® (5) a túto metódu označujeme $z_{3a}$.

Principiálne pri výpočte z-skóre vypočítavame o koľko násobok štandardnej odchýlky (resp. priemernej odchýlky alebo odchýlky od mediánu) sa líši rozdiel pomeru zastúpenia

namapovaných fragmentov určitého chromozómu v testovanej vzorke a priemeru (mediánu) počtu fragmentov daného chromozómu v trénovacích vzorkách. Štandardne sa za hraničnú hodnotu považuje z-skóre = 3, pričom hodnoty menšie ako 3 sa považujú za normálne (zdravé) a hodnoty väčšie ako 3 znamenajú štatisticky významný rozdiel oproti hodnotám v zdravých vzorkách, čiže možno predpokladať, že sa bude jednať o trizomický plod (7).

Celkovo sme vyskúšali a porovnali 6 rôznych prístupov, pričom sme modifikovali existujúce metódy zamenením pomeru namapovaných fragmentov testovaného chromozómu (v našom prípade chromozómu 21) k fragmentom všetkých ostatných chromozómov len za fragmenty namapované na autozómy (v metódach $z_{1b}$ a $z_{3b}$). Zároveň sme navrhli alternatívnu metódu, ktorej použitie sme v čase písania práce nenašli v žiadnej odbornej publikácii. Túto metódu označujeme ako $z_{2a}$ a $z_{2b}$. Uvedené metódy sme otestovali pre všetky testované kombinácie mapovacích algoritmov a referenčných genómov, a výsledky uvádzame v tabuľke č. 1-vľavo.

Vo všetkých prípadoch je hodnota rozdielu (DIFFz) medzi priemerným z-skóre trizomických testovaných vzoriek $\mu(z_{TT})$ a priemerným z-skóre zdravých testovaných vzoriek $\mu(z_{TH})$ vyššia pri metóde, pri ktorej berieme do úvahy pomer fragmentov namapovaných na chromozóm 21 voči fragmentom namapovaných na ostatné autozómy. Tento jav mohol pravdepodobne nastať v dôsledku variabilných oblastí nachádzajúcich sa na chromozóme Y a na mitochondriálnom chromozóme, ktoré je náročné referenčne pokryť. Ďalšou možnou príčinou by mohol byť fakt, že vyše 95% ľudského mitochondriálneho genómu sa nachádza duplikovaný v jadrovom genóme vo forme pseudogénov, a preto je náročné pre mapovací algoritmus správne namapovať fragment obsahujúci takúto sekvenciu (8).

Všeobecne považujeme za dôležité faktory výberu vhodnej metódy výpočtu z-skóre hlavne vyššie spomenutú hodnotu DIFFz, pri ktorej platí, že čím je vyššia, tým je rozdiel medzi trizomickými a zdravými vzorkami signifikantnejší a zároveň by sa nemali pri danej metóde vyskytnúť žiadne falošne negatívne ani falošne pozitívne výsledky a takisto hodnota $\mu(z_{TH})$ by mala byť podľa možností čo najnižšia, resp. čo najbližšie k 0. Na základe týchto podmienok sme sa rozhodli v ďalších výpočtoch vynechať metódu $z_{1a}$ a $z_{1b}$, nakoľko, ako z výsledkov v tabuľke č.1-vľavo vyplýva, ju ostatné metódy v týchto zreteľoch prekonali.

Z neznámych príčin vykazuje kombinácia mapovacieho algoritmu Bowtie2 s referenčným genómom hg38 výrazne horšie výsledky oproti ostatným prípadom (tabuľka č.1). Túto kombináciu sme viacnásobne preverili v každom kroku spracovania, no dospeli sme k rovnakým výsledkom. Možnou príčinou môže byť väčšie množstvo neodfiltrovaných fragmentov s nízkou kvalitou mapovania, ktoré by mohli konečné výsledky výrazne skresliť tak, ako to v danom prípade máme možnosť pozorovať. Pre odstránenie podobných problémov ako vznikli v predošlom kroku pri kombinácii Bowtie2 s hg38 sme sa rozhodli odfiltrovať fragmenty s kvalitou mapovania nižšou ako Phred=40 (pre SOAP2 Phred=30) a zároveň sme *in silico* vyselektovali fragmenty dĺžky 130-170 bp. Pre tieto veľkosti fragmentov sme sa rozhodli kvôli predpokladu, že v danom intervale by sa mohla nachádzať väčšina fragmentov pochádzajúcich z plodu (9).

Cieľom tohto kroku bolo vybrať vhodnú kombináciu mapovacieho algoritmu, referenčného genómu a metódy výpočtu z-skóre pre nasledujúcu fázu optimalizácie. Preto sme otestovali všetky kombinácie (okrem metód výpočtu z-skóre $z_{1a}$ a $z_{1b}$ vyradených v predošlom kroku), ktorých výsledky sú uvedené v tabuľke č.1-vpravo.

Pri kombináciách mapovacích algoritmov BWA a SOAP2 s referenčným genómom hg38 sa objavil pri každej metóde výpočtu z-skóre minimálne 1 falošne pozitívny alebo falošne negatívny výsledok, preto sme ich dočasne vylúčili z ďalšej analýzy. Metódy výpočtu

z-skóre $z_{3a}$ a $z_{3b}$ síce mali vyššie hodnoty DIFFz a nižšie hodnoty $\mu(z_{TH})$ v porovnaní so $z_{2a}$ a $z_{2b}$, no v niektorých prípadoch to bolo na úkor senzitivity a špecificity a pozorovali sme výskyt falošne negatívnych (resp. pozitívnych) výsledkov (tabuľka č.1). To bolo dôvodom, prečo sme v tomto kroku vybrali ako vhodnú kombináciu mapovací algoritmus BWA s referenčným genómom hg19 a metódou výpočtu z-score $z_{2b}$, ktorá za daných podmienok predstavovala optimálny variant, čo sme nakoniec spätne overili (tabuľka č.2) .

Na základe predošlých výsledkov sme použili kombináciu mapovacieho algoritmu BWA, referenčného genómu hg19 a metódy výpočtu z-skóre $z_{2b}$ a parametre filtrácie namapovaných fragmentov sme nastavili na minimálnu kvalitu mapovania Phred=40 a minimálnu dĺžku fragmentov 1 bp. V tomto kroku sme optimalizovali maximálnu dĺžku fragmentov, pričom sme ako maximálne dĺžky vyskúšali nastaviť hodnoty 100 bp, 130 bp, 150 bp, 155 bp, 160 bp, 165 bp, 166 bp, 167 bp, 170 bp a 200 bp. Vo výsledkoch zosumarizovaných na obrázku č. 1 môžeme pozorovať istú postupnosť stúpania, resp. klesania hodnôt z-skóre trizomických vzoriek v závislosti od maximálnej dĺžky fragmentu. Hodnoty DIFFz evidentne postupne stúpajú až po dosiahnutie vrcholu pri dĺžke 165 bp. Už pri zvýšení maximálnej dĺžky fragmentu o 1 bp na hodnotu 166 bp môžeme pozorovať pokles DIFFz, a pri maximálnej dĺžke 167 bp klesá táto hodnota výraznejšie. Maximálne dĺžky fragmentov menšie ako 155 bp sa javia ako nevyhovujúce, keďže pri hodnotách 150 bp a 100 bp sa vyskytol 1 falošne pozitívny výsledok.

Medzi našimi výsledkami a výsledkami z odborných publikácii sme objavili pomerne výrazné rozdiely. Yu et al. publikovali v roku 2014 výsledky svojej štúdie, kde sa podobne ako my zamerali na optimalizáciu dĺžky namapovaných fragmentov pri detekcii aneuploidií. V ich prípade bola optimálna maximálna dĺžka namapovaných fragmentov 150 bp (10). Predpokladáme, že rozdielne výsledky oproti našim dosiahli v dôsledku použitia iného postupu spracovania dát, konkrétne použitím mapovacieho algoritmu SOAP2 a referenčného genómu hg18 a na výpočet z-skóre použili metódu (v našom ponímaní) $z_{1a}$.

Cieľom našej práce bolo optimalizovať podmienky *in silico* spracovania NGS dát s cieľom detegovať prípadnú prítomnosť štatisticky významnej odchýlky pomeru zastúpenia chromozómu 21 vo vzorke. Tento cieľ sa nám podarilo splniť, pričom naše výsledky sme podložili relevantnými štatistickými analýzami. Zároveň sa nám podarilo navrhnúť a preukázať potenciál využitia alternatívnej metódy výpočtu z-skóre, ktorá za istých okolností môže predstavovať optimálnu voľbu pri niektorých druhoch analýz, ako tomu bolo v našom prípade. Z kombinácií, ktoré sme testovali sme nakoniec vybrali kombináciu mapovacieho algoritmu BWA s referenčným genómom hg19 a metódou výpočtu z-skóre $z_{2b}$ spolu s nastaveniami parametrov filtrovania na minimálnu kvalitu mapovania Phred=40, minimálnu dĺžku namapovaného fragmentu 1 bp a maximálnu dĺžku 165 bp. Táto kombinácia vykazovala najlepšie výsledky v sledovaných hodnotách $\mu(zTH)$ a DIFFz a zároveň výsledky spracované týmito nastaveniami dosahovali 100% senzitivitu a 100% špecificitu.

Na naše výsledky by do budúcna bolo možné nadviazať súvisiacim výskumom, v ktorom odporúčame otestovať aj iné dostupné mapovacie algoritmy (Bowtie, Novoalign, GSNAP, SOAP3). Ideálne by bolo iniciovať spoluprácu s odborníkmi z oblasti informačných technológií a pokúsiť sa vyvinúť vlastný mapovací algoritmus, ktorý by spĺňal požadované nároky kladené pri celogenómovom sekvenovaní s nízkym pokrytím. Pri optimalizácii takéhoto algoritmu by mohli naše výsledky poslúžiť ako porovnávacie dáta a prispieť tým k ďalšiemu zdokonaleniu protokolu metódy NIPT, ako aj pri výskume potenciálu tekutej biopsie pri diagnostike a sledovaní terapie nádorov, z ktorých sa do krvného obehu taktiež uvoľňuje voľná cirkulujúca DNA, ktorej charakter a množstvo je špecifické pre jednotlivé typy a veľkosti nádorov.

**Poďakovanie**

Tento článok vznikol vďaka podpore v rámci OP Výskum a vývoj pre projekt: Dobudovanie multidisciplinárneho centra pre biomedicínsky výskum – BIOMEDIRES, ITMS 26210120041, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

**Zoznam použitej literatúry**

1. Grinshpun-Cohen J, Miron-Shatz T, Ries-Levavi L, Pras E: Factors that affect the decision to undergo amniocentesis in women with normal Down syndrome screening results: it is all about the age. Health Expectations 2015; 18: 2306-2317.

2. Ogilvie C, Akolekar R: Pregnancy Loss Following Amniocentesis or CVS Sampling—Time for a Reassessment of Risk. Journal of Clinical Medicine 2014; 3: 741-746.

3. Allyse M, Minear MA, Berson E, Sridhar S, Rote M, Hung A, Chandrasekharan S: Non-invasive prenatal testing: a review of international implementation and challenges. Journal of International Journal of Women's Health 2015; 7: 113-126.

4. Lau TK, Chen F, Pan X a spol.: Noninvasive prenatal diagnosis of common fetal chromosomal aneuploidies by maternal plasma DNA sequencing. The Journal of Maternal-Fetal & Neonatal Medicine 2012; 25: 1370-1374.

5.https://lifecodexx.com/wp-content/uploads/2015/03/PraenaTest_Poster_Non-invasive_prenatal_testing_NIPT_Laboratory_experiences_2014-02-13-gfh-2014_Essen1.pdf (navštívené 02.03.2020)

6. Alberry M, Maddocks D, Jones M, Abdel Hadi M, Abdel-Fattah S, Avent N, Soothill PW: Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. Prenatal Diagnosis 2007; 27: 415-418.

7. Bayindir B, Dehaspe L, Brison N a spol.: Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. European Journal of Human Genetics 2015; 23: 1286-1293.

8. Ajaz S, Czajka A, Malik A: Accurate measurement of circulating mitochondrial DNA content from human blood samples using real-time quantitative PCR. Methods in Molecular Biology 2015; 1264: 117-131.

9. Fan HCh, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR: Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. Clinical Chemistry 2010; 56: 1279-1286.

10. Yu SCh, Chan KC, Zheng YW a spol.: Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. Proceedings of the National Academy of Sciences of the United States of America 2014; 111: 8583-8588.