Rapídne celogenómové sekvenovanie a komplexná analýza s využitím AI Emedgene

RNDr. Martin Kašný, Ph.D.[1], Ing. Mgr. Přemysl Bartoš, Ph.D.[2], Mgr. Alžbeta Mičúchová[1], Bc. Veronika Bůžková[2], Mgr. Laura Ambrozová, Ph.D.[1], Mgr. Karolína Stivínová[1], Mgr. Irem Mertová[1], Mgr. Alžběta Hamplová[1], Mgr. Eliška Hladíková[1], Mgr. Nela Tomko[1], Mgr. Markéta Kovalová, Ph.D.[1], prof. MUDr. Milan Macek, DrSc.[3], RNDr. Petr Kvapil[2]

[1]Institute of Applied Biotechnologies, Prague, Czech Republic
[2]GeneTiCA s.r.o., Praha, Česká Republika
[3]Ústav biologie a lékařské genetiky 2. lékařské fakulty Univezity Karlovy a Fakultní nemocnice Motol, Praha, Česká Republika

Cieľ: Optimalizácia protokolu rýchleho WGS poskytujúceho vysokokvalitné dáta s následnou anotáciou prioritných variantov s využitím umelej inteligencie softvérového nástroja Emedgene.

Materiál a metódy: Z venóznej krvi 3 účastníkov (rodinné trio: matka, otec a proband-dcéra), proband s diagnózou Mayer-Rokitansky-Küster-Hauserovho syndrómu, bola izolovaná genómová DNA pomocou súpravy EZ1&2 DNA Blood 350 µl Kit. Celogenómové knižnice boli pripravené pomocou súpravy TruSeq DNA PCR-Free High Throughput Library Prep Kit. rWGS sekvenovanie bolo realizované na prístroji NovaSeq 6000 s použitím sekvenačnej chémie S1 flowcell. Na následnú sekundárnu analýzu bol použitý Dynamic Read Analysis for GENomics (DRAGEN) OnPremise Server v4.1 (Illumina) s hg38 Alignment, Variant Calling cloudový systém DRAGEN v4.0 a softvérový nástroj využívajúci umelú inteligenciu Emedgene (Illumina).

Výsledky: Vysokokvalitná genómová DNA bola úspešne izolovaná z venóznej krvi všetkých 3 účastníkov štúdie. Celogenómové knižnice boli úspešne sekvenované s generovaným objemom dát: 464,92 Gb, pri 92,57 % čítaní pokrytých viac ako 35× a pri 94,66 % čítaní so skóre kvality Phred nad 30 (Q30). Všetky tri celogenómové knižnice spĺňali požadované kritérium – 800 miliónov čítaní z oboch koncov. Sekundárnou analýzou s využitím cloudového systému DRAGEN v4.0 a softvérového nástroja Emedgene bolo vytipovaných 7 patologických variantov Mayer-Rokitansky-Küster-Hauserovho syndrómu, ktoré boli zhodnotené laboratórnym diagnostikom.

Záver: Optimalizovaný protokol umožňuje rýchlu prípravu vysokokvalitných indexovaných celogenómových sekvenačných knižníc, sekvenovanie a bioinformatickú analýzu za 39 hodín a 54 minút.

Kľúčové slová: genómová DNA, celogenómové sekvenovanie, Rapid WGS, NovaSeq6000, anotácia variantov, umelá inteligencia Emedgene

Rapid whole-genome sequencing and comprehensive analysis using AI Emedgene

Purpose: Protocol optimization of the rapid whole-genome sequencing providing highquality data with subsequent annotation of priority variants using the AI software tool Emedgene.

Materials and methods: Genomic DNA was isolated using the EZ1&2 DNA Blood 350 µl Kit from venous blood from 3 participants (family trio: mother, father and proband daughter); proband diagnosed with Mayer-Rokitansky-Küster-Hauser syndrome. Whole-genome libraries were prepared using the TruSeq DNA PCR-Free High Throughput Library Prep Kit. Sequencing was performed on a NovaSeq 6000 using S1 flowcell sequencing chemistry. Dynamic Read Analysis for GENomics (DRAGEN) On Premise Server v4.1 (Illumina) with hg38 Alignment and Variant Calling and DRAGEN v4.0 Cloud System and AI software tool Emedgene (Illumina) were used for subsequent secondary analysis.

Results: High-quality genomic DNA was successfully isolated from the venous blood of all 3 study participants. Whole-genome libraries were successfully sequenced with a generated volume of data: 464.92 Gb, with 92.57% of reads covered more than 35× and 94.66% of reads with a Phred quality score above 30 (Q30). All whole-genome libraries met the required criterion of 800 million of pair-end reads. Secondary analysis using the DRAGEN v4.0 cloud system and artificial intelligence Emedgene identified 7 pathological variants of Mayer-Rokitansky-Küster-Hauser syndrome, which were evaluated by a laboratory diagnostician.

Conclusion: The optimized protocol enables rapid preparation of high-quality indexed wholegenome sequencing libraries, sequencing and bioinformatic analysis in 39 hours and 54 minutes.

Key words: genomic DNA, whole genome sequencing, Rapid WGS, NovaSeq6000, variants´ annotation, Emedgene artificial intelligence

Lek. genet. diagn., 2024;1(2): 131-134

Úvod

Sekvenovanie kompletného genómu v urýchlenom pracovnom postupe (rWGS z anglic. rapid Whole Genome Sequencing) a jeho následná analýza s využitím automatizovaných vyhodnocovacích cloudových nástrojov na báze AI (z anglic. Artificial Intelligence), sa nedávno stali prelomovým prístupom najmä v prípadoch, keď je rýchlosť komplexnej genomickej analýzy kľúčová. Ilustračným príkladom môže byť testovanie pediatrických/novorodeneckých pacientov, a to hlavne z dôvodu až 40 % úmrtnosti novorodencov spojenej so zriedkavým genetickým ochorením s extrémne rýchlym priebehom (1, 2, 3, 4). V klinickej genetike môže

rWGS, v porovnaní s inými diagnostickými prístupmi (tabuľka 1), vďaka rýchlej identifikácii kauzálnych genetických variantov (vrátane komplexných štrukturálnych variantov), umožniť vylepšenie terapeutických prístupov liečby alebo definovanie prognózy priebehu ochorení (5). rWGS sa taktiež stáva zlatým štandardom v personalizovanej diagnostike, pretože v porovnaní s jeho prínosmi sa cena za analýzu pohybuje na nízkej úrovni, a to vďaka včasnej a presnej identifikácie kauzálnych variantov, ktorá má často za následok mimo iného aj zníženie nákladov na celkovú starostlivosť (6).

V rámci našej štúdie sme v spolupráci s 2. lekárskou fakultou Univerzity Karlovej a spoločnosťou GeneTiCA realizovali optimalizáciu protokolu rýchleho WGS poskytujúceho vysokokvalitné dáta, v ktorom sa spájajú technológie a nástroje NGS spoločností QIAGEN a Illumina.

Tabuľka 1. Porovnanie prístupov genetického testovania na základe schopnosti detekcie variantov. Chromozomálna mikroarray (CMA); zmeny v počte kópií (CNV); malé inzercie/delécie (Indel); sekvenovanie novej generácie (NGS); polymerázová reťazová reakcia (PCR); jednonukleotidový variant (SNV); sekvenovanie celého exómu (WES); sekvenovanie celého genómu (WGS). Prevzaté a upravené z Illumina (7).

	Sanger*	Cielené NGS*	PCR*	CMA*	WES*	WGS*
SNVs	✓✓	✓✓	✓✓		✓✓	✓✓
Indels	✓✓	✓✓	✓✓	✓	✓✓	✓✓
CNVs		✓	✓✓	✓✓	✓	✓✓
Expanzie repetícií			✓			✓✓
Štrukturálne varianty				✓	✓	✓✓
Mitochondrie	✓✓	✓✓			✓✓	✓✓
Paralógy	✓✓		✓✓			✓✓

**Detekcia variantov sa môže líšiť v závislosti od laboratória a ponuky testov, ✓✓ Detekcia, ✓ Obmedzená detekcia*

Materiál a metódy

Izolácia a purifikácia genómovej DNA

Vzorky krvi boli odobraté od každého z 3 účastníkov (rodinné trio: matka, otec a proband-dcéra); proband s diagnózou Mayer-Rokitansky-Küster-Hauserovho syndrómu – genéza maternice. Vzorky venóznej krvi (4 ml/účastník) boli odobraté do skúmaviek Vacuette K2EDTA (Becton Dickinson, USA), temperované na teplotu v miestnosti počas 30 minút a následne boli skladované pri 4 °C až do izolácie. Vysokokvalitná genómová DNA (gDNA) bola izolovaná pomocou súpravy EZ1&2 DNA Blood 350 µl Kit (QIAGEN) a automatického purifikátora nukleových kyselín EZ2 Connect (QIAGEN) fungujúceho na báze magnetických guličiek.

Kontrola kvality genómovej DNA

Čistota gDNA bola hodnotená na základe pomeru absorbancie A260/280 získaného spektrofotometricky na NanoPhotometri P300 (Implen, Nemecko) z 1 µl izolátov gDNA. Koncentrácie gDNA boli zmerané v duplikátoch (2 x 1 µl) pomocou súpravy Qubit 1 x dsDNA High Sensitivity (HS) kit (Thermo Fisher Scientific, USA). Integrita gDNA bola stanovená pomocou elektroforézy na 0,8 % agarózovom géli (GelRed 1 : 10 000, 100 V, 40 minút).

Príprava celogenómových knižníc

Celogenómové knižnice so vstupnou koncentráciou gDNA 1 050 ng boli pripravené pomocou súpravy TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina), podľa Kvapilová et al. (8).

Kontrola kvality celogenómových knižníc

Koncentrácia WGS knižníc bola meraná v duplikátoch (2 × 1 µl) pomocou súpravy Qubit 1 × dsDNA High Sensitivity (HS) Kit. Kvalita NGS knižníc bola analyzovaná pomocou automatizovanej elektroforézy na 2100 Bioanalyzer System s DNA High Sensitivity čipom (Agilent, USA).

Sekvenovanie dát

rWGS sekvenovanie bolo realizované na prístroji NovaSeq 6000 s použitím sekvenačnej chémie S1 flowcell (2 x 101 bp) (Illumina). Demultiplexovanie sekvenčných dát a konverzia CBCL na FASTQ boli uskutočnené prostredníctvom cloudového servera BaseSpace Sequence Hub (Illumina).

Analýza dát

Na následnú sekundárnu analýzu boli použité dva prístupy: 1) lokálny server Dynamic Read Analysis for GENomics (DRAGEN) OnPremise Server v4.1 (Illumina) s hg38 Alignment a Variant Calling. Na anotáciu identifikovaných variantov bol použitý anotačný nástroj Variant Effect Predictor (VEP) (Ensembl); 2) cloudový systém DRAGEN v4.0 a Emedgene (Illumina).

Emedgene bol použitý na doplnenie anotácie z databáz (GERP, REVEL, SpliceAI, VEP, dbNSFP a dbscSNV). Zároveň rozšíril naše znalosti o informácie v populačných databázach (DECIPHER, DGV, GnomAD, MITOMAP Polymorphisms, ThousandGenomeSTRAnnotation) a v databázach variantov (ClinGen Regions, Clinvar, Clinvar_SV, GnomAD STR, MITOMAP Known Variants, MitImpact, MitoTIP, Region Type Database, dbSNP. Emedgene pracoval s nasledujúcimi Knowleadbase zdrojmi (CGD, ClinGen_HI, DECIPHER_HI, EMEDGENE, Gnomad_constraint, HGNC, HPO, OMIM, Orphanet, Uniprot). Emedgene v rámci analýzy využil Explainable AI (XAI), ktorá slúži na priorizáciu variantov, zhrnutie informácií súvisiacich s variantom.

Výsledky a diskusia

Pre rapídne celogenómové sekvenovanie boli vybrané vzorky krvi od 3 účastníkov (rodinné trio: matka, otec a proband-dcéra s diagnózou Mayer-Rokitansky-Küster-Hauserovho syndrómu. Genómová DNA bola úspešne izolovaná pomocou súpravy EZ1&2 DNA Blood 350 µl Kit (QIAGEN) a automatického purifikátora nukleových kyselín EZ2 Connect (QIAGEN). Vyizolovaná gDNA bola bez známok kontaminácie proteínmi či reziduálnymi roztokmi použitými počas izolácie, v dostatočnej koncentrácii a s vysokou úrovňou integrity (tabuľka 2; výsledky analýzy integrity gDNA nie sú uvedené).

Tabuľka 2. Vybrané metriky kontroly kvality izolátov gDNA

Vzorky	Nanofotometer		Qubit	Objem vzorku po QC [µl]	Celkový obsah DNA [ng]
	c = [ng/µl]	260/280 [nm]	c = [ng/µl]		
Proband	125,0	1,9	85,8	38	3 260,4
Matka	110,0	2,0	61,6	43	2 648,8
Otec	168,0	1,9	83,3	59	4 914,7

Tabuľka 3. Vybrané metriky kontroly kvality WGS knižníc

Vzorky	Príprava knižníc		Qubit	qPCR	Bioanalyzer 2100		
	Input [µl]	Input [µg]	c = [ng/µl]	c = [ng/µl]	c = [ng/µl]	Ø veľkosť fragmentov [bp]	Diméry [%]
Proband	12,2	1 050	3,3	1,6	3,2	917	0
Matka	17	1 050	4,3	1,8	3,4	755	0
Otec	12,6	1 050	5,2	2,1	3,4	801	0

Tabuľka 4. Časový priebeh prípravy knižníc a kontroly kvality. Presný, reprodukovateľný harmonogram prípravy 3 celogenómových knižníc vrátane krokov kontroly pomocou súpravy TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina).

QC + TruSeq Príprava knižníc	Čas [min]	Poznámky
QC-Qubit	8	
Normalizácia	8	
Fragmentácia	10	
Purifikácia fragmentovanej DNA	22	
Úprava koncov DNA	37	Pipetovanie + cyklér (5 + 32)
Purifikácia veľkých + malých fragmentov	44	
Adenylácia	40	Pipetovanie + cyklér (5 + 32)
Ligácia adaptérov	18	Pipetovanie + cyklér + pipetovanie (5 + 10 + 3)
Purifikácia – 2-krát	44	(22 + 22)
QC-Qubit	8	
Poolovanie	10	
Riedenie a poolovanie knižníc	40	
Celkový čas	**289**	

Celogenómové knižnice boli pripravené pomocou súpravy TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina), ktorý predstavuje vynikajúce riešenie pre ťažko sekvenovateľné regióny DNA, ako sú oblasti bohaté na GC páry, promótory alebo repetície. Použitie mechanickej fragmentácie, ktorá je použitá v procese prípravy knižníc daným protokolom, navyše znižuje riziko vytvárania artefaktov a medzier v knižniciach. Na základe kvantitatívnych a kvalitatívnych meraní bola stanovená koncentrácia pripravených celogenómových knižníc. Veľkosť fragmentov dosahovala priemerne 824 bp, a nebola pozorovaná prítomnosť adaptorových dimérov, ktoré môžu negatívne ovplyvniť výstupy sekvenácie. Konkrétne metriky z kontroly kvantity a kvality celogenómových knižníc sú uvedené v tabuľke 3.

Keďže cieľom našej štúdie bolo vytvoriť protokol rýchleho WGS poskytujúceho vysokokvalitné dáta, bol zaznamenávaný čas trvania každého kroku od kontroly kvality izolátov gDNA až po nastavenie sekvenácie. Príprava celogenómových knižníc, vrátane všetkých krokov kontroly kvality, trvala 289 minút (tabuľka 4).

rWGS sekvenovanie bolo realizované na prístroji NovaSeq 6000 s použitím sekvenačnej chémie S1 flowcell (2 x 101 bp) (Illumina). Sekvenačný beh sa úspešne ukončil za 18 hodín s generovaným objemom dát: 464,92 Gb, pri 92,57 % čítaní pokrytých viac ako 35× a pri 94,66 % čítaní so skóre kvality Phred nad 30 (Q30). Všetky tri celogenómové knižnice spĺňali požadované kritérium – 800 miliónov čítaní z oboch koncov a bližšie informácie o výstupoch sekvenácie sú uvedené v tabuľke 5.

Demultiplexovanie sekvenčných dát a konverzia CBCL na FASTQ boli uskutočnené prostredníctvom cloudového servera BaseSpace Sequence Hub (Illumina) za 75 minút. Na následnú sekundárnu analýzu boli použité dva rôzne prístupy.

Prvý prístup využil lokálny server Dynamic Read Analysis for GENomics (DRAGEN) OnPremise Server v4.1 (Illumina) s hg38 Alignment a Variant Calling. Na anotáciu identifikovaných variantov bol použitý voľne dostupný anotačný nástroj Variant Effect Predictor (VEP) (Ensembl). Vďaka absencii limitov internetového pripojenia a plnej kontrole nad prioritami spracovania sekvenačných dát trvala analýza ~ 14 hodín. V druhom prístupe bol využitý cloudový systém DRAGEN v4.0 a Emedgene (Illumina). Vstupom do Emedgene boli VCF súbory spolu s HPO (Human Phenotype Ontology) odpovedajúce jednotlivým pacientom, konkrétne išlo o Aplasia of the uterus – HP:0000151 a Hypoplasia of the vagina – HP:0008726. Systém DRAGEN v4.0 a Emedgene dokončili anotáciu variantov za ~ 15,5 hodín. Dlhšie trvanie analýzy je však zanedbateľné, pretože klinickí špecialisti pracujú len s niekoľkými klinicky relevantnými variantmi vopred vybranými umelou inteligenciou Emedgene a nie so stovkou netriedených anotovaných variantov. Na hlbšie preskúmanie jednotlivých prípadov poskytuje Emedgene rýchle nastavenia filtrovania variantov, ktoré môžu laboratórni diagnostici alebo iní užívatelia aplikovať na zobrazenie exónových a intrónových variantov v kandidátnych génoch alebo sa zamerať na varianty s očakávanou segregáciou, vo svojich súkromných účtoch v Emedgene.

Na základe vstupných informácií, umelá inteligencia Emedgene do užšieho výberu zaradila 7 najpravdepodobnejších patologických variantov (tabuľka 6) a 63 ďalších kandidátnych variantov, ktoré boli zaslané na zhodnotenie laboratórnym diagnostikom z 2. lekárskej fakulty Univerzity Karlovej a FN Motol. Zo všetkých anotovaných variantov bolo definované, že varianty génov *COL7A1* a *PDGFRA* majú priamu súvislosť s fenotypovým prejavom hypoplazie vagíny, pričom variant génu *PDGFRA* bol zistený aj u matky probanda (tabuľka 6).

Tabuľka 5. Metriky sekvenovania. NovaSeq 6000 (2x101 bp) S1 prietoková komôrka s 3 ľudskými WGS knižnicami

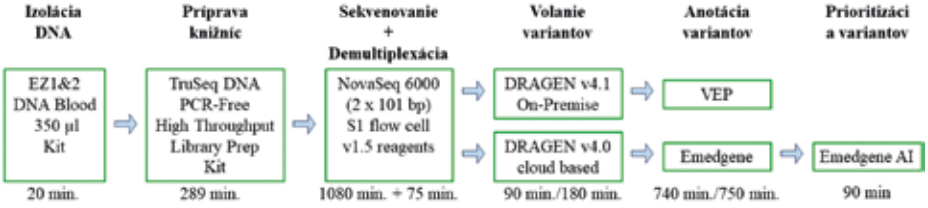
Metriky sekvenovania	Proband	Matka	Otec
Celkový počet čítaní (b)	1 391 117 692	1 419 897 506	1 393 989 054
Zarovnanie čítania (%)	99,38	99,47	99,48
Bázy ≥ Q30 (%)	94,23	94,82	94,80
Priemerné pokrytie (x)	35,79	36,65	35,67
Uniformita pokrytia (%)	93,78	93,80	94,36
Zarovnané čítania (b)	1 079 771 648	1 105 509 357	1 075 417 779
PCT genómu s pokrytím 20x – 50x (%)	86,88	85,62	83,87

Tabuľka 6. Kauzálne varianty genetických porúch. Anotácia cloudový systém DRAGEN v4.0 a AI softvérový nástroj Emedgene

Gén	Pozícia	Variant	Klasifikácia	Proband	Matka	Otec
NXN	chr17	819516 2.85(Kb) DEL	Aplázia maternice, Hypoplázia vagíny	✓	-	-
RNASEL	chr1	182586014C>A,rs74315364	Aplázia maternice, Hypoplázia vagíny	✓	-	✓
CNTN2	chr1	205058319C>T,	Aplázia maternice, Hypoplázia vagíny	✓	-	-
NAIP	chr5	71010576 3.392(Kb) DEL	-	✓	-	-
RP1L1	chr8	10608139G>A,rs200846354	Aplázia maternice, Hypoplázia vagíny	✓	✓	-
COL7A1	chr3	48583584G>A,rs79378857	Hypoplázia vagíny	✓	-	✓
PDGFRA	chr4	54265030G>A,	Hypoplázia vagíny	✓	✓	-

**Detekcia variantov, ✓ Detekcia, - Bez detekcie*

Obrázok 1. Pracovný postup rýchleho celogenómového sekvenovania (rWGS) na identifikáciu variantov. Pracovný postup od prípravy vzorky až po získanie výsledkov s použitím lokálneho servera Dynamic Read Analysis for GENomics (DRAGEN) verzie 4.1, po ktorom nasleduje anotácia variantov pomocou verejne dostupného nástroja Variant Effecta Predictor (VEP) za 38,2 hodín alebo cloudového servera DRAGEN verzie 4.0 integrujúceho umelú inteligenciu Emedgene na identifikáciu a prioritizáciu potenciálne relevantných variantov za 39 hodín a 54 minút.



Optimalizovaný NGS protokol, ktorý predkladáme v tejto štúdii, kombinuje najmodernejšie produkty a prístupy izolácie DNA, prípravy knižníc, sekvenovania a analýzy sekvenačných výstupov s následnou interpretáciou dát, vrátane anotácie variantov. Kompletnú diagnostiku probanda a jeho rodičov bolo možné uskutočniť už za 39 hodín a 54 minút (obrázok 1).

Záver

rWGS integruje všetky metodiky digitálnej genomiky vrátane umelej inteligencie do komplexného personalizovaného diagnostického prístupu. Hlavným cieľom práce bolo poskytnúť prioritné anotované varianty v efektívnom časovom rámci za menej ako 40 hodín od odberu vzorky. Inovatívny pracovný postup rWGS založený na technológiách spoločností QIAGEN a Illumina, posúva personalizovanú diagnostiku na vyššiu úroveň a otvára cestu širšiemu využitiu v klinickej praxi. Prezentované výsledky dokazujú, že rWGS a analýzou Emedgene predstavuje riešenie pre rýchlu a presnú diagnostiku, najmä vďaka rýchlej detekcii kauzálnych variantov genetických porúch, a to aj v náročných prípadoch, v ktorých iné diagnostické nástroje zaostávajú.

Autorom nie je známy potenciálny konflikt záujmov.

Literatúra

1. Weiner J, Sharma J, Lantos J, et al. How infants die in the neonatal intensive care unit: trends from 1999 through 2008. Arch Pediatr Adolesc Med. 2011;165(7):630-34.
2. Gunne E, McGarvey C, Hamilton K, et al. A retrospective review of the contribution of rare diseases to paediatric mortality in Ireland. Orphanet J Rare Dis. 2020;15(1):31.
3. Berry MA, Shah PS, Brouillette RT, et al. Predictors of mortality and length of stay for neonates admitted to children's hospital neonatal intensive care units. J Perinatol. 2008;28(4):297302.
4. Arth AC, Tinker SC, Simeone RM, et al. Inpatient Hospitalization Costs Associated with Birth Defects Among Persons of All Ages - United States, 2013. MMWR Morb Mortal Wkly Rep. 2017;66(2):41-6.
5. Petrikin JE, Cakici JA, Clark MM, et al. The NSIGHT1-randomized controlled trial: rapid whole-genome sequencing for accelerated etiologic diagnosis in critically ill infants. NPJ Genom Med. 2018;3:6.
6. Farnaes L, Hildreth A, Sweeney NM, et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. NPJ Genom Med. 2018;3:10.
7. Agrawal PB. Whole-Genome Sequencing [online]. Illumina 2022. Available from: <https://www.illumina.com/content/dam/illumina-marketing/documents/genetic-rare-disease/gdt-publication-booklet-mgl00605-v5-digital-final.pdf >. Accessed June 21, 2024.
8. Kvapilova K, Misenko P, Radvanszky J, et al. Validated WGS and WES protocols proved saliva-derived gDNA as an equivalent to blood-derived gDNA for clinical and population genomic analyses. BMC Genom. 2004;25:187.

RNDr. Martin Kašný, Ph.D.
Institute of Applied Biotechnologies,
Služeb 3056/4, 108 00 Prague, Czech Republic
kasny@iabio.eu