

Multiparametrické modelovanie mitochondriálnej DNA pre diagnostiku rakoviny

Prezentácia diplomovej práce

Bc. Ondrej Škornák

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Vedúci práce: Mgr. Jaroslav Budiš, PhD.

Február 2026

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky
- 7 Záver

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky
- 7 Záver

Prečo tekutá biopsia?

- **Obmedzenia tkanivovej biopsie:**

- Invazívny zákrok s nepohodlím a rizikami pre pacienta,
- Môže zachytiť len časť priestorovej heterogenity nádoru,
- Nie je vhodná pre opakovaný odber.

- **Výhody tekutej biopsie:**

- Minimálne invazívne profilovanie z krvnej plazmy,
- Umožňuje dlhodobé sledovanie.

- **Výzva:** Včasná detekcia rakoviny kvôli nízkej frakcii nádorovej DNA.

- **cfDNA fragmentomika:** dopĺňa mutačné analýzy o fyzikálne signály (distribúcie dĺžok, pozície koncov, vzory pokrytia).

Prečo mitochondriálna cfDNA (ccf-mtDNA)?

- **Charakteristiky mtDNA:**

- Kompaktný genóm ($\sim 16,5$ kb),
- Vysoký počet kópií na bunku ($\sim 100\text{--}10\,000$ kópií),
- Hojné ready za určitých experimentálnych podmienok.

- **Signály súvisiace s rakovinou v mtDNA:**

- Mutačné vzory,
- Fragmentačné profily,
- Genómová lokalizácia pozdĺž mitochondriálneho genómu.

- Fragmentomika mtDNA je **málo preskúmaná** v porovnaní s jadrovou cfDNA.

- Nedávne štúdie ukazujú, že ccf-mtDNA vykazuje odlišné fragmentomické správanie.

Centrálna hypotéza

Diskriminatívny signál sa nachádza v **koincidencii** atribútov fragmentov (dĺžka $\times$ pozícia $\times$ mutačné signály).

Problém s agregáciou na úrovni vzorky:

- Globálne histogramy a priemery môžu rozmazať združené vzory.
- Informatívna koincidencia sa stráca pri nezávislom sumarizovaní príznakov.

Ciele práce:

- 1 Charakterizovať ccf-mtDNA fragmenty (dĺžka, pozícia, mutácie).
- 2 Navrhnuť multiparametrický framework prirad'ujúci rizikové skóre fragmentom.
- 3 Vyhodnotiť oproti baseline na klasifikácii rakovina/kontrola.

- 1 Motivácia
- 2 Súčasný stav poznania**
- 3 Dáta
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky
- 7 Záver

Globálny prehľad tekutej biopsie (2020–2026)

- **Genetické analýzy:** mutácie ctDNA, CNV (cielené panely, WGS).
- **Epigenetické analýzy:** metylácia cfDNA pre včasnú detekciu viacerých typov rakoviny (MCED) + tkanivo pôvodu.
- **Fragmentomika:** veľkostné profily, koncové body, koncové motívy, genómovo-široké fragmentačné vzory (často ML).
- **Trend:** kombinácia viacerých slabých signálov $\Rightarrow$ lepšia senzitivita pri nízkej frakcii nádoru.

Vybrané ref.: Liu (MC) et al., Ann Oncol 2020; Klein et al., Ann Oncol 2021; Lo et al., Science 2021; Tsui et al., Cancer Cell 2025.

- **Veľkosti fragmentov:** obohatenie kratších fragmentov v ctDNA; chvostové správanie a pomery krátke/dlhé.
- **Genómovo-široké fragmentačné profily:** pomery krátke/dlhé na úrovni binov (DELFI-type príznaky).
- **Koncové body fragmentov:** preferované miesta rezu + **koncové motívy** odrážajú aktivitu nukleáz a tkanivo pôvodu.
- **Footprinting:** vzory ochrany nukleozómami / TF $\Rightarrow$ proxy pre chromatín a génovú expresiu.

Vybrané ref.: Cristiano et al., *Nature* 2019; Jiang et al., *Cancer Discov* 2020; Lo et al., *Science* 2021; Tsui et al., *Cancer Cell* 2025.

- **Multi-cancer (WGS):** DELFI dosiahlo AUC $\sim 0,86$ až $>0,99$ naprieč typmi rakoviny (vysoká špecificita).
- **Rakovina pľúc (prospektívna kohorta):** DELFI AUC $\approx 0,90$ (štádium I AUC $\approx 0,76$) + nezávislá validácia.
- **Praktický smer:** fragmentomické signály sa dajú extrahovať aj z **cielených klinických panelov**, nielen WGS.

Vybrané ref.: Cristiano et al., Nature 2019; Mathios et al., Nat Commun 2021; Helzer et al., Nat Commun 2025.

Fragmentomika ccf-mtDNA: multi-cancer detekcia (2024)

- ccf-mtDNA vykazuje **odlišnú fragmentomiku** oproti jadrovej cfDNA: veľkostný profil fragmentov, 5' bázová preferencia a diverzita motívov.
- **Regióno-závislá** fragmentácia pozdĺž mtDNA spojená s väzbou proteínov, zložením báz a štruktúrou genómu.
- Capture-based štúdia (**1607 plazmatických vzoriek**): multi-cancer modely dosiahli **AUC > 0,93** v internej/externej validácii a presnosť tkaniva pôvodu ~89%.

Vybraná ref.: Liu et al., EMBO Mol Med 2024.

- **Single-cancer štúdie:** včasná detekcia kolorektálneho karcinómu pomocou fragmentomických príznakov ccf-mtDNA.
- **Okrem fragmentomiky:** počet kópií cf-mtDNA a mtDNA mutácie sa aktívne skúmajú ako biomarkery.
- **Pre-analytická citlivosť:** extrakcia a príprava knižníc môže skresliť množstvo mtDNA a profily dĺžok fragmentov.
- **Špecifické confoundery mtDNA:** NUMTs, efekty hranice cirkulárneho genómu a nerovnomerné pokrytie.

Vybrané ref.: Peng et al., *Cancer Biol Med* 2023; Wang et al., 2025 (CRC); Liu et al., *EMBO Mol Med* 2024.

Kľúčová medzera

Mnoho prístupov stále spolieha na **agregáciu na úrovni vzorky** (marginály), zatiaľ čo diskriminatívny signál sa môže skrývať vo vzoroch **koincencií**.

- **Zameranie práce:** multiparametrická reprezentácia na úrovni fragmentov (dĺžka $\times$ pozícia $\times$ mutačné signály).
- **Cieľ robustnosti:** explicitne riešiť kohortový posun / batch efekty a testovať medzikohortový prenos.
- **Výstup:** end-to-end pipeline (rekonštrukcia $\rightarrow$ príznaky $\rightarrow$ skórovanie fragmentov $\rightarrow$ agregácia vzorky).

Toto priamo nadväzuje na dôraz zadania práce na koincencii dĺžky, mutácií a genómovej lokalizácie na úrovni fragmentov.

Vybrané referencie (súčasný stav)

- Cristiano et al. *Genome-wide cell-free DNA fragmentation in patients with cancer*. Nature (2019).
- Jiang et al. *Plasma DNA End-Motif Profiling as a Fragmentomic Marker*. Cancer Discovery (2020).
- Lo et al. *Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies*. Science (2021).
- Mathios et al. *Detection and characterization of lung cancer using cell-free DNA fragmentomes*. Nat Commun (2021).
- Liu (MC) et al. *Sensitive and specific multi-cancer detection and localization using methylation*. Ann Oncol (2020).
- Klein et al. *Clinical validation of a targeted methylation-based multi-cancer early detection test*. Ann Oncol (2021).
- Liu (Y) et al. *Aberrant fragmentomic features of ccf-mtDNA for multi-cancer detection*. EMBO Mol Med (2024).
- Helzer et al. *Fragmentomics metrics on commercial targeted sequencing panels*. Nat Commun (2025).
- Tsui et al. *Cell-free DNA fragmentomics in cancer*. Cancer Cell (2025).

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta**
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky
- 7 Záver

Kohorta PreveLynch:

- $n = 976$ vzoriek plazmatickej cfDNA
- 753 zdravých kontrol
- 223 pacientov s kolorektálnym karcinómom
- Silná triedna nevyváženosť ($\sim 3:1$ kontroly:rakovina).
- Niektorí pacienti prispievajú viacerými časovými bodmi $\Rightarrow$ potrebné **zoskupenie podľa pacienta**.
- Rôzne typy rakoviny $\Rightarrow$ potenciál pre typovo-špecifické efekty.

Kohorta GenoScan:

- $n = 453$ vzoriek plazmatickej cfDNA
- 380 zdravých kontrol
- 73 pacientov s karcinómom prostaty

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta
- 4 Predbežné baseline experimenty**
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky
- 7 Záver

Nastavenie baseline: príznaky z QualiMap

- Príznaky parsované z QualiMap BamQC reportov (mtDNA QC + fragmentačné súhrny).
- Modelovanie v `scikit-learn`: mediánová imputácia → štandardizácia → klasifikátor.
- **Evaluácia bez úniku dát**: 5-fold **GroupKFold** podľa ID pacienta.

Rodiny príznakových sád

$\mathcal{F}_{QC}$ (14)	QC súhrny (ready, pokrytie, chybovosť, MAPQ)
$\mathcal{F}_{Frag}$ (26)	Fragmentácia (percentily, šikmosť, špicatosť)
$\mathcal{F}_{Hist}$ (57)	Úplné deskriptory z histogramov
$\mathcal{F}_{Hybrid}$ (36)	QC + kurátorované fragmentačné príznaky

Výkon baseline v rámci štúdie (ROC-AUC)

Sada príznakov	# príz.	PreveLynch	GenoScan
$\mathcal{F}_{QC}$	14	0,760	0,630
$\mathcal{F}_{Frag}$	26	0,674	0,675
$\mathcal{F}_{Hist}$	57	0,664	0,703
$\mathcal{F}_{Hybrid}$	36	0,814	0,693

- Hybridná reprezentácia zlepšuje AUC (+0,054 na PreveLynch).
- **Obava:** top prediktory zahŕňajú technické QC premenné (error_rate, MAPQ).
- Možný confounding s batch efektmi.

Hybridná sada: porovnanie modelov na PreveLynch

Model	ROC-AUC	Presnosť	Recall	F1
Random Forest	0,791 $\pm$ 0,019	0,708	0,336	0,453
SVM (RBF)	0,814 $\pm$ 0,021	0,484	0,659	0,557
Gradient Boosting	0,809 $\pm$ 0,021	0,678	0,440	0,530

- SVM dosahuje najlepšie ROC-AUC, ale nižšiu presnosť (viac falošne pozitívnych).
- Random Forest: vysoká presnosť, ale slabý recall (minie veľa prípadov rakoviny).
- **Kľúčová otázka:** Prenesie sa výkon v rámci štúdie medzi kohortami?

Křížová validácia medzi štúdiami (dataset shift)

Tréning → Test	SVM AUC	RF AUC
PreveLynch → GenoScan	0,436	0,411
GenoScan → PreveLynch	0,382	0,386

Kritické zistenie

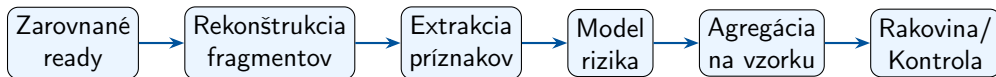
ROC-AUC $< 0,5$ znamená **horší výsledok ako náhodné hádanie**.

Modely sa naučia kohortovo-špecifické vzory, ktoré sa *invertujú* na druhom datasete.

Možné príčiny:

- Batch efekty a technické confoundery,
- Rozdiely v typoch rakoviny (kolorektálny vs. prostata),
- Kohortovo-špecifické protokoly a efekty pracoviska.

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov**
- 6 Obmedzenia a ďalšie kroky
- 7 Záver



- Cieľ: zachytiť **združené** vzory dĺžka/pozícia/mutácie.
- Prekonať agregáciu na úrovni vzorky, ktorá rozmazáva koincidenčné vzory.

- **Rekonštrukcia fragmentov:** z paired-end readov zarovnaných na chrM.
- **Kvalitatívne filtre:**
 - MAPQ prah na vylúčenie málo spoľahlivých zarovnaní,
 - iba správne páry (korektná orientácia, rozumná veľkosť insertu),
 - filtrovanie rozsahu veľkosti insertu (napr. 50–500 bp).
- **Špecifické výzvy pre mtDNA:**
 - NUMTs (nukleárne mitochondriálne DNA segmenty) $\Rightarrow$ prísne filtrovanie,
 - cirkulárny genóm $\Rightarrow$ ošetrenie breakpointov prechádzajúcich cez origin.

Fragmentačné príznaky:

- Dĺžka fragmentu
- Štart/end pozície (breakpointy)
- Pokrytie závislé od regiónu
- Priradenie genómového binu

Príznaky súvisiace s mutáciami:

- Počet mismatchov na fragment
- Typ mutácie (trieda substitúcie)
- Kvalita báz na variantných miestach
- Lokálny sekvenčný kontext

Výstupná reprezentácia

Tabuľkový dataset: **1 riadok = 1 fragment** s príznakmi fragmentu + metadáta vzorky.

- **Kľúčový insight:** diskriminatívny signál leží v *združených* distribúciách, nie v marginálnych.
- **Analytické prístupy:**
 - Heatmapy: dĺžka $\times$ pozícia, mutácia $\times$ pozícia,
 - Stratifikované distribúcie dĺžok podľa genómového regiónu,
 - Profily hustoty mutácií pozdĺž mtDNA.
- Exploratívna analýza na identifikáciu rakovina-špecifických vzorov pred modelovaním.

Dvojstupňový prístup:

- 1 **Skórovanie na úrovni fragmentov:** natrénovať model na priradenie rizikového skóre každému fragmentu.
 - Modely: logistická regresia, gradient boosting, random forest.
- 2 **Agregácia na úrovni vzorky:** konvertovať skóre fragmentov na predikciu vzorky.
 - Priemerné skóre, top- k priemer, podiel nad prahom.

Evaluácia

Krížová validácia zoskupená podľa pacientov; ROC-AUC a PR-AUC pri triednej nevyváženosti.

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky**
- 7 Záver

- **Batch efekty / confounding:**
 - Technické QC premenné (chybovosť, MAPQ) korelujú s labelmi,
 - Kohortovo-špecifické protokoly môžu spôsobiť falošné asociácie.
- **Špecifické problémy mapovania mtDNA:**
 - NUMTs môžu zaviesť falošné zarovnanie,
 - Low-complexity regióny vyžadujú opatrné MAPQ filtrovanie.
- **Heterogenita typov rakoviny:**
 - Pozorované opačné smery efektov (napr. frag_cv),
 - Univerzálny detektor môže vyžadovať diverzné multi-cancer tréningové dáta.

Analýza zlyhania krížovej validácie

Niekoľko príznakov vykazuje **opačné smery** medzi kohortami:

Príznak	PreveLynch	GenoScan
frag_cv	Nižší pri rakovine	Vyšší pri rakovine
frag_std	Nižší pri rakovine	Vyšší pri rakovine
frag_pct_gt250	Nižší pri rakovine	Vyšší pri rakovine

⇒ Motivuje reprezentácie na úrovni fragmentov, ktoré zachytia biologicky podloženejšie vzory.

Na základe baseline analýzy tri kľúčové princípy:

① Evaluácia bez úniku dát:

- Zoskupenie podľa pacientov v krížovej validácii,
- Žiadne vzorky od toho istého pacienta v trénovacej a testovacej sade.

② Robustnosť voči batch efektom:

- Explicitné hodnotenie efektov štúdie/pracoviska,
- Zvážiť batch-korekciu (napr. ComBat).

③ Reprezentácie na úrovni fragmentov:

- Zachytiť koincidenčné vzory stratené pri agregácii,
- Potenciálne lepšie prenositeľné medzi kohortami.

- **Rozšíriť diverzitu kohort:** multi-site, multi-cancer trénovacie dáta.
- **Doménová adaptácia:** naučiť sa štúdiovo-invariantné reprezentácie.
- **Pokročilé modely:**
 - Multiple Instance Learning (MIL) pre klasifikáciu na úrovni vzorky,
 - Attention-based agregácia skóre fragmentov.
- **Robustnosť príznakov:** vyberať príznaky na základe konzistencie medzi kohortami.
- **Analýza citlivosti:** kvantifikovať signál po odstránení technických QC premenných.

- 1 Motivácia
- 2 Súčasný stav poznania
- 3 Dáta
- 4 Predbežné baseline experimenty
- 5 Navrhovaný prístup na úrovni fragmentov
- 6 Obmedzenia a ďalšie kroky
- 7 Záver**

- **Problém:** Včasná detekcia rakoviny z tekutej biopsie je náročná kvôli nízkej frakcii nádoru.
- **Prístup:** Multiparametrické modelovanie mtDNA fragmentov, zachytávajúc združené vzory dĺžky, pozície a mutácií.
- **Zistenia z baseline:**
 - V rámci kohorty existuje signál (ROC-AUC až 0,81),
 - Prenos medzi kohortami zlyháva $\Rightarrow$ batch efekty a heterogenita typov rakoviny.
- **Motivácia pre prístup na úrovni fragmentov:**
 - Agregácia na úrovni vzorky môže zakryť diskriminatívne koincidenčné vzory,
 - Príznaky na úrovni fragmentov ponúkajú bohatšiu, potenciálne robustnejšiu reprezentáciu.

- ① Komplexná charakterizácia ccf-mtDNA fragmentomiky.
- ② Multiparametrický framework na úrovni fragmentov pre diskrimináciu rakovina/kontrola.
- ③ Systematické porovnanie s agregatívnymi baseline.
- ④ Poznatky o batch efektoch a generalizácii medzi kohortami.

Ďakujem za pozornosť!

Otázky?