Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Bioinformatika

Studijní obor: B-BINF

Lujza Milotová

Bioinformatické přístupy k vyhodnocení frekvenčních populačních dat založených na sekvenčních polymorfismech DNA

Bioinformatic approaches to the evaluation of frequency population data based on DNA sequence polymorphisms

Bakalářská práce

Školitel: Mgr. Vlastimil Stenzl

Konzultant: doc. RNDr. David Hoksza, Ph.D.

Praha 2024

Rada by som sa týmto poďakovala môjmu vedúcemu práce Mgr. Vlastimilovi Stenzlovi za trpezlivosť a ochotu zodpovedať všetky moje otázky a rýchlu spätnú väzbu počas celého procesu tvorby práce. Tiež ďakujem Ing. Michaele Nekardovej, Ph.D., ktorá mi umožnila sa k tejto práci dostať a doc. RNDr. Davidovi Hokszovi, Ph.D. za konzultácie. Moja vďaka patrí aj rodine a priateľom, ktorí mi poskytovali podporu počas celého štúdia.

Prohlášení:

Prohlašuji, že jsem závěrečnou práci vypracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze Podpis:

Abstrakt

Tato práce se zaměřuje na problém efektivní správy dat využitelných při DNA analýze, konkrétně dat získaných z krátkých tandemových repetic v lidské DNA, které mají klíčový význam ve forenzní analýze pro identifikaci osob. Práce představuje návrh a implementaci systému pro správu a vizualizaci těchto dat z české populace, získaných Kriminalistickým ústavem Policie České republiky. Hlavním cílem bylo transformovat stávající datové soubory do univerzálního a snadno spravovatelného formátu, který by byl kompatibilní s existujícími forenzními databázemi. Výsledkem návrhu je transformace dat do strukturovaného formátu a jejich následná vizualizace v tabulkovém uspořádání, což zvyšuje snadnost interpretace a manipulace s daty. Práce rovněž popisuje problémy zjištěné při zpracování dat a jejich řešení.

Klíčová slova: DNA polymorfismy, DNA markery, forenzní genetika, bioinformatická analýza

Abstract

This thesis focuses on the problem of efficient management of data useful in DNA analysis, specifically data obtained from short tandem repeats in human DNA, which are of key importance in forensic analysis for the identification of individuals. The thesis presents the design and implementation of a system for the management and visualization of these data from the Czech population, obtained by the Institute of Criminalistics of the Police of the Czech Republic. The main goal was to transform existing data sets into a universal and easy to manage format that would be compatible with existing forensic databases. As a result, the design involves transforming the data into a structured format and then visualising it in tabular layouts, enhancing the ease of data interpretation and manipulation. The thesis also describes the problems identified during data processing and their solutions.

Key words: DNA polymorphism, DNA markers, forensic genetics, bioinformatic analysis

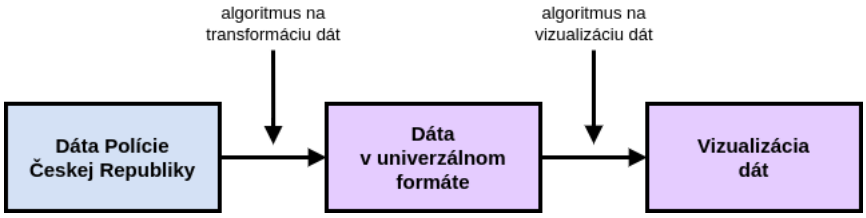
Obsah

Úvod

Analýza ľudskej DNA je základnou metódou forenznej analýzy, ktorá pomáha v riešení trestných prípadov. Jedinečnosť DNA každého človeka umožňuje presné určenie jeho identity, pretože genetický materiál každého jedinca sa líši a umožňuje tak jeho jednoznačnú identifikáciu. V analýze DNA sú kľučovými najmä polymorfizmy, teda rozdiely v sekvenciách DNA medzi jednotlivcami. V súčasnosti sú v analýze DNA významné najmä krátke tandemové repetície, teda krátke sekvencie DNA opakujúce sa bezprostredne za sebou. U rôznych osôb sa tieto repetície vyskytujú v rôznych počtoch opakovaní, čo dobre slúži na účely identifikácie. Niektoré krátke tandemové repetície sú pre analýzu DNA mimoriadne výpovedné, a preto sa využívajú vo forenznej analýze celosvetovo, teda slúžia ako genetické markery.

Zber a analýza dát o krátkych tandemových repetíciách z populácie má vo forenznej analýze veľký význam. Tieto dáta pomáhajú určiť, ako časté sú konkrétne varianty určitých DNA markerov, čo pomáha pri správnej interpretácii dôkazov. Preto je dôležité mať tieto populačné dáta uložené a organizované tak, aby boli jednoducho prístupné a kompatibilné s určitými štandardami.

Cieľom tejto práce je riešiť problém správy dát návrhom efektívnejšej metódy na organizáciu dát o sekvenčných polymorfizmoch krátkych tandemových repetícií získaných z českej populácie Kriminalistickým ústavom Polície Českej republiky. Nasledujúci diagram znázorňuje hlavné kroky vývoja programu, ktorý umožní prevedenie existujúcich populačných dát do formátu JSON, ktorý bude kompatibilný s existujúcimi forenznými databázami a následnú vizualizáciu týchto dát. Grafické znázornenie dát v jednoduchých tabuľkách formátu XLSX uľahčí ich analýzu, zrozumiteľnosť a intepretáciu.



Obr. 1: Diagram znázorňujúci návrh postupu práce s dátami

1. DNA profilovanie vo forenznej analýze

1.1 Ľudský genóm

Ľudský genóm pozostáva z deoxyribonukleovej kyseliny, ktorá obsahuje inštrukcie potrebné pre vývoj, rast, fungovanie a reprodukciu organizmu. DNA je pravotočivá šrúbovica tvorená dvoma dlhými vláknami. Tieto vlákna sú tvorené nukleotidmi, ktoré sú základnými stavebnými kameňmi DNA. Jeden nukleotid je tvorený molekulou cukru (v tomto prípade deoxyribózou), fosfátovou skupinou a dusíkovou bázou. Bázy v DNA sú štyroch typov: adenín (A), guanín (G), cytozín (C) a tymín (T). Konkrétne poradie a sekvencia nukleotidov pozdĺž vlákna DNA je to, čo kóduje genetickú informáciu.

Kľúčovým aspektom štruktúry DNA je princíp komplementarity, ktorý určuje vzájomné párovanie nukleotidov v špirálovitej štruktúre. Zvyčajne platí, že adenín páruje s tymínom a guanín s cytozínom. Tento mechanizmus párovania zabezpečuje, že vlákna DNA sú navzájom komplementárne, čo je základom procesu replikácie DNA, teda odovzdávania presnej genetickej informácie do nových buniek pri delení.

1.2 DNA polymorfizmy

DNA polymorfizmy sú rozdiely v sekvenciách DNA medzi jednotlivcami a sú základom genetickej rozmanitosti v ľudskej populácii. Až približne 99,9 % z celkových 3 miliónov bázových párov v ľudskom genóme je pre všetkých ľudí spoločných. [11] Tento vysoký stupeň sekvenčnej podobnosti sa zachováva vďaka našej spoločnej evolučnej histórii. Spoločná genetická výbava je potrebná pre základné biologické funkcie a štruktúry, ktoré definujú človeka ako druh. Tieto spoločné sekvencie kódujú proteíny, enzýmy a regulátory esenciálne pre základné procesy v bunke, fyziologické funkcie a anatomické vlastnosti. Práve táto variabilita v necelom 0,1 % genómu je využívaná v genetickej identifikácii biologického materiálu.

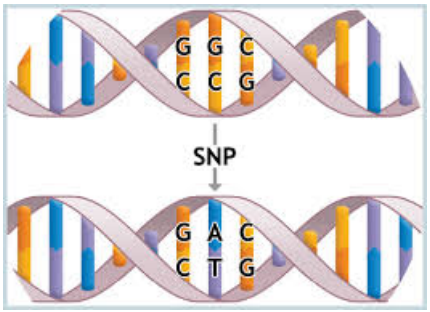
Polymorfné oblasti genómu sa vyskytujú v špecifických miestach, lokusoch, v celom genóme. Vo forenznej analýze sú tieto variabilné miesta kľúčové, pretože umožňujú rozlíšiť DNA rôznych jednotlivcov. Analýzou polymorfizmov vo viacerých lokusoch je možné vytvoriť profil DNA, unikátny pre každého jedinca, ktorý sa môže použiť na porovnanie osoby s biologickými stopami nájdenými na mieste činu alebo na určenie príbuzenských vzťahov medzi jednotlivcami.

Sekvencie v genóme, ktoré možno identifikovať a sledovať, sú nazývané DNA markery. [1] V kriminalistike sú markery špecifické lokusy DNA, kde sa analyzujú polymorfizmy na účely identifikácie. Je dôležité, akým spôsobom sú tieto markery zvolené. Musia byť vysoko polymorfné, aby sa zabezpečila dostatočná variabilita medzi jednotlivcami na ich presné rozlíšenie.

DNA polymorfizmy pozorujeme v genómoch v mnohých formách, ale v kontexte forenznej analýzy DNA sú najvýznamnejšími jednonukleotidové polymorfizmy (angl. single nucleotide polymorphism, SNP) a krátke tandemové repetície (angl. short tandem repeats, STR).

1.2.1 SNP markery

SNP predstavujú zmenu v jednom konkrétnom nukleotide. Napríklad na určitom mieste v sekvencii DNA môže mať jedna osoba nukleotid adenín, zatiaľ čo iná osoba môže mať cytozín. SNP patria k najbežnejším typom genetickej variabilily v ľudskom genóme. Priemerne sa vyskytujú každých 300 nukleotidov, čo znamená, že v ľudskom genóme je približne 10 miliónov SNP. [33] Aj tieto malé rozdiely teda môžu mať významný vplyv na genetický výskum a forenznú analýzu.



Obr. 1.1: Znázornenie jednonukleotidového polymorfizmu [27]

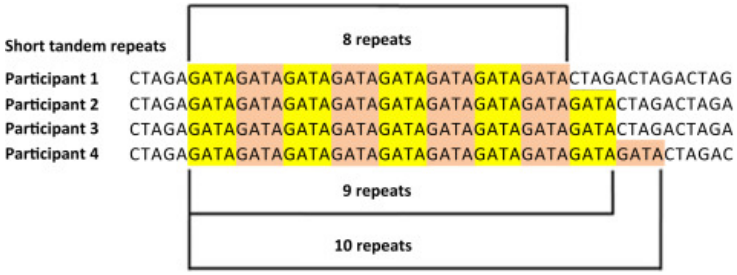
V kriminalistike sú SNP markery predmetom výskumu kvôli ich potenciálu v identifikácii jednotlivcov a analýze príbuzenských vzťahov, hoci v súčasnosti nie sú rutinne používané. Na rozdiel od STR markerov, pri ktorých je potrebné analyzovať dĺžku opakujúcich sa sekvencií, SNP markery vyžadujú len identifikáciu jednotlivých nukleotidových zmien a ich pozícií. SNP sa dajú analyzovať aj v najmenších fragmentmoch DNA, a preto sú výnimočne užitočné v situáciách, keď sú vzorky DNA degradované alebo veľmi malé, napríklad pri hromadných nešťastiach, prírodných katastrofách alebo v archeogenetike. [39] Okrem toho vysoká početnosť SNP v celom genóme zvyšuje rozlišovaciu schopnosť medzi jednotlivcami. SNP môžu tiež poskytnúť prehľad o biogeografickom pôvode a používajú sa pri fenotypizácii, teda predpovedaní fyzických znakov, ako je pigmentácia pleti, vlasov alebo očí či štruktúra tváre, čo môže byť užitočné pri vytváraní profilov neidentifikovaných podozrivých alebo obetí, keď nie sú k dispozícii iné informácie. [29]

S príchodom SNP čipov a vysoko výkonných sekvenačných technológií sa výrazne zvýšila schopnosť analyzovať viacero SNP markerov súčasne, čo umožňuje komplexnejšie a účinnejšie forenzné profilovanie. Rast databáz SNP markerov prispieva k zvyšovaniu presnosti a použiteľnosti biologických stôp ako dôkazov vo vyšetrovaní trestnej činnosti. [13]

1.2.2 STR markery

V ľudskom genóme existuje veľa repetitívnych oblastí, teda miest, kde sa určité sekvencie nukleotidov viackrát opakujú. Špeciálne významné sú STR, známe aj ako mikrosatelity, ktoré pozostávajú z krátkych sekvencií, zvyčajne dlhých 2 - 5 bázových párov, opakujúcich sa viackrát bezprostredne za sebou. [42] V každom STR lokuse môžeme u rôznych jedincov pozorovať rôzne dĺžkové varianty, teda sekvencie s rôznym počtom opakovaní krátkej repetície. Každá takáto dĺžková varianta ešte môže mať sekvenčné varianty, teda varianty, ktoré sa líšia v samotnej sekvencii, ktoré tiež odlišujú jednotlivých jedincov. Analýzou dĺžkových a sekvenčných variánt jednotlivých STR v genóme je možné vytvoriť unikátny genetický profil, ktorý je pre jednotlivca jedinečný. Tento profil sa potom porovnáva so vzorkami DNA odobratými z miest činov, obetí alebo podozrivých osôb

s cieľom zistiť genetickú zhodu alebo vylúčiť osoby z podozrenia.



Obr. 1.2: Znázornenie krátkych tandemových repetícií [21]

Forenzná analýza DNA sa zameriava na špecifické STR markery s vysokou mierou polymorfizmu a distribúciou v celom genóme. Medzi bežne používané STR markery, ktoré sa používajú celosvetovo, kvôli ich informatívnosti v rôznych populáciách, patria napríklad tieto markery:

- D1S1656
- TPOX
- D2S1338
- D2S441
- D3S1358
- FGA
- D5S818
- CSF1PO
- D7S820
- D8S1179
- D10S1248
- TH01
- vWA
- D12S391
- D13S317
- D16S539
- D18S51
- D19S433
- D21S11
- D22S1045

Databáza Combined DNA Index System (CODIS), ktorú vyvinul americký Federálny úrad pre vyšetrovanie (Federal Bureau of Investigation, FBI), využíva štandardný súbor 20 vyššie uvedených STR markerov, čím zvyšuje konzistentnosť a porovnateľnosť dôkazov DNA v rôznych jurisdikciách. [26]

Podobne ako využitie SNP markerov, aj využitie STR markerov v analýze DNA presahuje rámec vyšetrovania trestných činov. Zahŕňa aplikácie v oblasti identifikácie obetí hromadných nešťastí, katastrof, prípadov nezvestných osôb a testovania otcovstva. [3] Vysoká variabilita a rozsiahle rozšírenie lokusov STR v ľudskom genóme umožňujú citlivé a špecifické identifikačné procesy, a to aj

zo znehodnotených alebo obmedzených vzoriek DNA, ktoré sú vo forenznom kontexte bežné.

Pokroky v technikách molekulárnej biológie, najmä amplifikácia pomocou polymerázovej reťazovej reakcie, kapilárna elektroforéza a masívne paralelné sekvenovanie, zjednodušili analýzu STR markerov, čo umožňuje rýchle, presné a vysoko výkonné spracovanie forenzných vzoriek. Schopnosť súčasne analyzovať viacero STR lokusov z jednej vzorky DNA výrazne zvyšuje účinnosť a spoľahlivosť forenzného profilovania DNA. Tento vývoj pokročilých genetických markerov, rozvoj sekvenačných technoogií a vylepšenie bioinformatických nástrojov umožnili poskytovanie dôkazov aj v prípadoch, kde ide o analýzu aj inej než ľudskej DNA, akými sú napríklad útoky na zvieratá, obchodovanie s ohrozenými druhmi, bioterorizmus alebo potravinové podvody. [4]

1.3 Spôsoby získavania dát

Vo vedeckom výskume v oblasti genetiky sa na získavanie informácií o DNA využívajú rôzne techniky a metódy, ktoré prinášajú rôzne poznatky.

V prvom rade je dôležité spomenúť polymerázovú reťazovú reakciu (angl. polymerase chain reaction, PCR), ktorá slúži na amplifikáciu DNA. PCR beží v cykloch, kde opakovane dochádza k zahrievaniu a ochladzovaniu. Zahrievanie spôsobuje denaturáciu DNA, čiže oddelenie vláken od seba, a následné ochladenie umožňuje nasadnutie DNA primerov, teda krátkych úsekov DNA slúžiaciach ako počiatok replikácie, na tieto oddelené vlákna, ktoré potom predlžuje DNA polymeráza, a tak vytvára nové vlákna DNA. Takto sa zdvojnásobí množstvo DNA v každom cykle. Vďaka účinnosti pri amplifikácii malých množstiev DNA sa PCR stala kľúčovou metódou v mnohých oblastiach vedy. Táto metóda sa stále vyvíja a upravuje. K hlavným pokrokom v oblasti analýzy DNA patrí umožnenie amplifikácie viacerých DNA oblastí vrámci jednej reakcie, teda multiplexná PCR a presné načasovanie jednotlivých krokov PCR predstavené s príchodom súpravy Rapid DNA. [30]

Tradičné metódy gélovej elektroforézy sa využivajú na hrubé oddelenie DNA fragmentov podľa ich veľkosti, čo je dôležité pri posudzovaní kvality produktov

PCR a pri rozlišovaní zmiešaných vzoriek DNA.

DNA mikročipy (angl. microarrays) sa využívajú na detekciu a analýzu SNP, čo je významné pri výbere genetických markerov a pri odhaľovaní genetickej variability. Táto metóda využívajúca hybridizáciu krátkych úsekov DNA je pri analýze DNA účinná najmä pri vzorkách s krátkymi fragmentami DNA. Zároveň umožňuje analýzu niekoľkých markerov súčasne. [13]

Dnes už takmer nepoužívaná, ale v minulosti významná metóda analýzy polymorfizmu dĺžky restrikčných fragmentov (angl. restriction fragment length polymorphism, RFLP) umožňuje presné DNA profilovanie využiteľné vo forenznej analýze pri identifikácii osôb a v populačnej genetike. V RFLP sú sekvencie DNA štiepené na špecifické sekvencie pomocou restrikčnej endonukleázy a následne sú tieto fragmenty separované na gélovej elektroforéze. To vytvára unikátny DNA profil. [32]

V nasledujúcich podkapitolách budú podrobnejšie rozobrané metódy kapilárnej elektroforézy a masívneho paralelného sekvenovania, ktoré sú špeciálne dôležité pri získavaní dát týkajúcich sa STR markerov, a samotná príprava DNA na jej analýzu.

1.3.1 Príprava DNA na jej analýzu

Proces prípravy DNA profilu začína extrakciou samotnej DNA z buniek. Biologické vzorky musia prejsť procesmi, ktoré vyizolujú DNA z buniek a purifikujú ju. Medzi tieto procesy patrí lýza bunkovej membrány, rozrušenie jadrovej obálky, odseparovanie DNA molekuly a jej presun do mierne slaného pufru alebo vody. Aby DNA mohla byť ďalej spracovávaná, musí spĺňať isté kritériá - musí byť dostatočne čistá, zbavená látok akými sú napríklad proteíny, tuky či uhľovodíky, a musí jej byť dostatočné množstvo.

Po extrakcii a purifikácii DNA nasleduje kvantifikácia DNA, teda určenie presného množstva extrahovanej DNA, čím sa zabezpečí použitie vhodného objemu DNA na amplifikáciu STR lokusov pomocou PCR. Zvyčajne potrebné množstvo DNA je 1 ng.

Produkty aplifikácie STR lokusov sú detekovateľné vďaka fluorescencii indukovanej laserom, keďže fluorescenčné farbivo vie byť pripojené na primery použité

v PCR. Produkty sú navzájom odseparované napríklad pomocou kapilárnej elektroforézy a môžu byť sekvenované pomocou masívneho paralelného sekvenovania. [28]

1.3.2 Kapilárna elektroforéza

Kapilárna elektroforéza (angl. capillary electrophoresis, CE) je významným pokrokom v tradičných gélových elektroforetických technikách. Je aplikovateľná v mnohých oblastiach vedy vrátane chemickej, farmaceutickej, biomedicínskej a biochemickej analýzy. Hlavným účelom CE je oddeliť jednotlivé molekuly podľa ich náboja a veľkosti pomocou elektrického poľa v úzkej kapiláre. V kontexte STR markerov sa CE využíva na určenie počtu repetícií v jednotlivých alelických variantách na základe veľkosti molekúl.

Hlavné súčasti kapilárneho elektrografu sú injekčný systém, tenká kapilára, vysokonapäťový zdroj, elektródy a detektory. Tenká kapilára môže byť naplnená gélom, ktorý napodobňuje gélovu elektroforézu, alebo pufračným roztokom, ktorý slúži ako elektrolyt, v ktorom sa oddeľujú fragmenty. Po injekcii malých vzoriek DNA je napätie aplikované po celej dĺžke kapiláry, čo spôsobuje pohyb nabitých fragmentov pozdĺž kapiláry rôznou rýchlosťou v závislosti na pomere ich náboja a veľkosti. Menšie a viac nabité molekuly sa pohybujú rýchlejšie než väčšie a menej nabité. Na konci kapiláry prebieha detekcia absorbcie UV alebo viditeľného svetla, fluorescencie alebo hmotnostná spektometria. Detektor zaznamenáva elektrogram, ktorý vo výsledku vyzerá ako graf s vrcholmi tam, kde bola detekovaná molekula. Z elektrogramu je teda možné vyčítať charakteristiky jednotlivých DNA fragmentov. [44]

Existuje niekoľko modifikácií a variánt CE slúžiacich na rôzne typy analýz. Okrem analýzy biopolymérov je uplatniteľná aj napríklad v detekcii výbušnín či analýze drog. Vďaka tejto všestrannosti a účinnosti sa CE stala jednou z hlavných techník forenznej DNA analýzy.

STR profilovanie pomocou PCR a CE sa stalo najčastejšie používanou metódou na získanie DNA profilu. Prechod od tradičných kapilár naplnených gélom k roztokom zosieťovaných polymérov zlepšil rozlíšenie a reprodukovateľnosť separácií DNA. Vývoj detekčných techník, najmä implementácia detekcie viacerých

vlnových dĺžok (napríklad rôznych farbív naviazaných na špecifické STR markery), umožnil analýzu viacerých STR markerov súčasne. [8]

Vytváranie použiteľných profilov DNA je výzvou pri veľmi malých množstvách DNA (menej ako 200 pg templátovej DNA). Vtedy sa využíva analýza DNA s nízkym počtom kópií (angl. low copy number, LCN). Pri použití CE je LCN problémom vzhľadom na potenciálnu nízku citlivosť a špecifickosť analýzy. Nízkej citlivosti sa dá predísť zvýšením počtu cyklov PCR pred zavedením DNA vzoriek do kapiláry, no amplifikácia malého počtu vzoriek môže viesť k stochastickým javom, ako je napríklad náhodnosť molekúl, ktoré sa amplifikujú. To môže viesť ku vypadnutiu niektorých z prítomných alel (angl. allele drop-out) alebo k výskytu alely, ktorá v pôvodnej vzorke nebola prítomná (angl. allele drop-in). Zvýšená amplifikácia pomocou PCR tiež môže viesť k tvorbe artefaktov, akými sú napríklad falošné vrcholy v elektrograme (angl. stutter peaks), alebo ku zvýšenému šumu, ktorý je spôsobený napríklad častejším skĺzavaním (angl. stutter) polymerázy po templáte. [6]

Problém LCN dnes lepšie rieši masívne paralelné sekvenovanie popísané v nasledujúcej podkapitole. DNA analyzovaná pomocou CE je v súčasnosti často následne analyzovaná aj pomocou masívneho paralelného sekvenovania, pretože tento nový spôsob získavania dát môže rozšíriť informatívnu hodnotu analyzovaného markeru. [5]

1.3.3 Masívne paralelné sekvenovanie

Sekvenovanie DNA je v súčasnosti základom moderného biologického výskumu. Ide o presné určenie poradia nukleotidov v molekulách DNA alebo RNA. Sekvenačné technológie priniesli revolúciu v mnohých oblastiach vedy. V 70. rokoch 20. storočia, v počiatkoch bioinformatiky, dominovala Sangerova metóda sekvenovania, ktorú na začiatku 21. storočia nahradili metódy sekvenovania novej generácie (angl. next generation sequencing, NGS), nazývané tiež masívne paralelné sekvenovanie (angl. massive parallel sequencing, MPS). [41]

V MPS je kľúčová príprava knižníc z informatívnych lokusov. K týmto sekvenciám DNA sú najprv pripevnené adaptéry, teda krátke sekvencie syntetickej DNA. Sekvencie sú potom jednoducho amplifikované pomocou pridania primerov, ktoré

sú komplementárne ku adaptérom so známou sekvenciou. Takto pripravená knižnica sekvencií potom môže byť sekvenovaná. Vysoko výkonné technológie MPS umožňujú sekvenovať milióny fragmentov DNA súčasne, čím poskytujú obrovské množstvo genetických dát za oveľa menší čas a náklady, než technológie z minulosti. [24] Toto prináša vyššiu rozlišovaciu schopnosť v analýze DNA v prípadoch, kde sa vo vzorke nachádza DNA z viac než jedného jedinca (ak je podiel DNA od jedinca aspoň 10 %), alebo v prípadoch so znehodnotenými vzorkami. [46] Tento pokrok tiež umožnil komplexné genomické analýzy, ako napríklad celogenómové sekvenovanie, resekvenovanie alebo cielené sekvenovanie.

V súčasnosti nastupujú sekvenačné metódy vysokej priepustnosti druhej generácie (angl. high-throughput - next generation sequencing, HT-NGS), ktoré umožňujú sekvenovanie jedinej molekuly. [37] Ide napríklad o technológie Heliscope, SMRT alebo nanopórové sekvenovanie. Platformy HT-NGS dokážu osekvenovať 500 miliónov báz (technológia Roche) až miliardy báz (technológia Illumina) naraz. Platformy tretej generácie HT-NGS dokážu osekvenovať 28 miliárd báz (technológia Heliscope) až 100 miliárd báz (technológia SMRT) v rámci jedného sekvenovania s dĺžkou jednotlivých prečítaných úsekov (angl. reads) viac než 1000 báz. Ultimátnou sekvenačnou platformou sa stáva Oxfordské nanopórové sekvenovanie (angl. Oxford nanopore sequencing), kvôli jednoduchej manipulácii, umožneniu čítania jedinej molekuly, dlhým výsledným reads a nízkemu GC biasu, teda preferenčnej sekvenácii DNA oblastí s vysokým obsahom guanínu a cytozínu. [7]

Zavádzanie MPS vo forenzných vedách bolo spočiatku pomalé v dôsledku nedostatku akreditovaných sekvenátorov a vyššej chybovosti v porovnaní s tradičnou metódou Sangeroveho sekvenovania. S pokrokom v technológiách sa však podarilo mnohé z týchto problémov vyriešiť, čo umožnilo širšie uplatnenie MPS vo forenznej DNA analýze. V súčasnosti sa využívajú rôzne metódy, ako napríklad sekvenovanie syntézou (angl. sequencing by synthesis, SBS), a rôzne prístroje, ako napríklad platformy MiSeq FGx spoločnosti Illumina a Ion Torrent spoločnosti ThermoFisher. [5]

Pri probléme LCN sa v MPS často využíva okrem zvýšenej PCR aj celogenómová amplifikácia (angl. whole genome amplification, WGA). Tá sa zameriava

najmä na vzorky s malým obsahom DNA, ako sú napríklad vzorky získané z jednej bunky. WGA pomocou metódy aplifikácie s viacnásobným premiestnením (angl. multiple displacement amplification) účinne podporuje typizáciu pomocou STR a SNP markerov vo forenznej analýze. [10] Metóda WGA sa osvedčila pri použití CE aj MPS, MPS sa však v súčasnosti stáva čoraz populárnejšou a používanejšou metódou, pretože poskytuje komplexnejšiu analýzu DNA vzoriek.

1.4 Interpretácia dát

Po zíkaní dát nasleduje samotná interpretácia získaných dát o genotype, a to napríklad porovnanie profilu s profilom vytvoreným z DNA získanej z miesta činu, alebo s profilmi vo forenzných DNA databázách.

Metóda DNA profilovania priniesla do oblasti forenzných vied revolúciu vďaka jej presnosti pri vyšetrovaní trestných činov a indentifikácii osôb. Od jej počiatkov v 80. rokoch 20. storočia, ku ktorým prispel Alec Jeffreys, sa vyvinula od základných foriem DNA profilovania k sofistikovanej analýze STR markerov, či špecifických markerov na mitochondriálnej DNA alebo na Y chromozóme. [40]

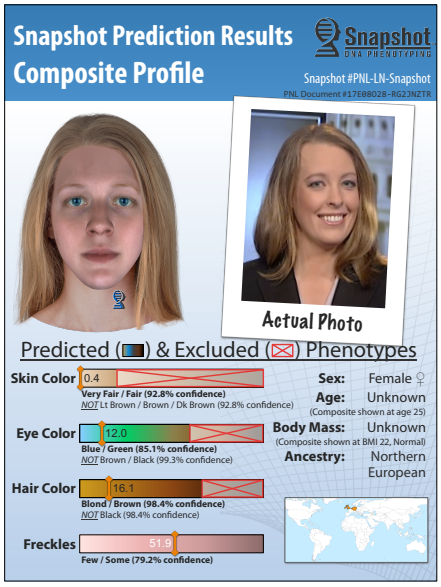
Pôvodné metódy klasického DNA profilovania boli založené na RFLP a spočívali v rádioaktivnom značení minisatelitov, ktoré hybridizovali s fragmentami DNA rôznych veľkostí odpovedajúcich rôznym počtom opakovaní. Tieto hybridizované fragmenty potom boli vizualizované a tak porovnané s iným, podobne vytvoreným DNA profilom.

V súčasnosti sa využíva analýza viacerých STR lokusov, ktoré majú podobnú štruktúru ako pôvodne používané minisatelity, ale repetitívna oblasť je kratšia, takže je jednoduchšie ich amplifikovať pomocou PCR. Pravdepodobnosť, že by dvaja jedinci mali zhodné všetky tieto lokusy, je približne jedna k miliónu, čo poskytuje istotu v usvedčovaní podozrivých pri zhode ich DNA profilov s DNA profilmi vytvorenými z DNA získanej z miesta činu.

Nové technológie zvýšili kvalitu profilovania a komplexitu DNA profilov v rôznych aspektoch, akými sú napríklad výška vrcholov na elektrograme, stutter polymerázy alebo alelický drop-in a drop-out. Najčastejšie používanými metódami na vyhodnotenie kvality DNA dôkazu sú pomer pravdepodobnosti (angl. like-

lihood ratio) a kombinovaná pravdepodobnosť zahrnutia/vylúčenia (angl. combined probability of inclusion/exclusion), ktoré väčšinou zahŕňajú prisudzovanie pravdepodobností rôznym možným genotypom a ich overovanie. [2]

Techniky DNA profilovania sa ďalej rozvinuli aj s príchodom DNA fenotypovania (angl. forensic DNA phenotyping, FDP), teda určenia fenotypu – vizuálnych znakov jedinca. FDP využíva analýzu DNA získanej z miesta činu na predikciu vizuálnych charakteristík (angl. externally visible characteristics), ako je napríklad farba pleti, vlasov či očí, plešatosť alebo črty tváre. [47] Táto metóda nevyžaduje referenčnú vzorku DNA získanú od podozrivého, čo je značnou výhodou. FDP spočíva v určení daných charakteristík na základe asociácie s konkretnými alelickými variantami STR či SNP lokusov. Napríklad za pigmentáciu očných dúhoviek je zodpovedných konkrétnych šesť SNP lokusov nachádzajúcich sa v rôznych génoch súvisejúcich s pigmentáciou. [29, 35] Výsledky FDP nemusia byť jednoznačne výpovedné ako forenzný dôkaz, ale môžu byť dôležitým faktorom určujúcim smer policajného vyšetrovania napríklad v dôsledku zúženia množiny podozrivých na malý počet.



Obr. 1.3: Príklad predikcie fenotypu pomocou technológie Snapshot™ DNA Phenotyping Service od spoločnosti Parabon NanoLabs [36]

2. Forenzné DNA databázy

2.1 Význam a cieľ

Forenzná DNA databáza je úložisko, v ktorom sa zhromažďujú a uchovávajú dáta týkajúce sa DNA profilovania, a to napríklad dáta o STR alebo SNP markeroch, na účely trestného súdnictva a presadzovania práva. Hlavným cieľom týchto databáz je identifikácia donorov neznámych biologických materiálov zanechaných na miestach súvisejúcich s trestným činom a identifikácia ľudských pozostatkov.

Rozvoj forenzných DNA databáz začal v roku 1994 s príchodom STR profilovania založeného na multiplexnej PCR. [45] Počnúc Spojeným kráľovstvom v roku 1995, po ktorom nasledovali ďalšie európske krajiny, boli tieto databázy formované odlišnými vnútroštátnymi právnymi orgánmi, čo viedlo k značným rozdielom. Kritériá pridávania, uchovávania alebo odstraňovania DNA profilov sa v jednotlivých krajinách líšili, čo ovplyvňovalo štruktúru a funkciu každej databázy.

Od týchto prvých forenzných DNA databáz sa kladie dôraz na ich štandardizáciu. V súčasnosti sa o to európske krajiny usilujú v záujme efektívnej cezhraničnej spolupráce v oblasti trestného súdnictva napríklad prijatím spoločných technologických noriem a európskeho štandardného súboru (angl. European Standard Set, ESS) markerov na DNA profilovanie. [12]

2.1.1 Referenčný genóm

V súčasnosti je v kontexte forenzných DNA databáz referenčný genóm kľúčový pri budovaní nových databáz založených na sekvenčných polymorfizmoch DNA (STR, SNP). Dnešné rutinne používané databázy sú založené na dĺžkových polymorfizmoch STR lokusov. Referenčný genóm poskytuje komplexnú informáciu, s ktorou možno porovnávať jednotlivé DNA profily. DNA zo vzoriek pochádzajúcich z miesta trestného činu by mohli byť porovnávané s referenčným genómom na určenie špecifických lokusov, ktoré sú výpovedné pre identifikáciu.

Vďaka štandardizovanému referenčnému genómu by tak forenzní analytici mohli zabezpečiť konzistentnosť a presnosť pri DNA profilovaní. Táto štandar-

dizácia je nevyhnutná na porovnávanie profilov DNA v rôznych jurisdikciách a na zachovanie integrity medzinárodných forenzných databáz. [14] Okrem toho je referenčný genóm kľúčový pri rozlišovaní medzi genetickými odchýlkami, ktoré sú relevantné pre overenie totožnosti, a tými, ktoré by mohli odhaliť citlivé lekárske alebo biologické informácie, čím sa zabezpečí, že na forenzné účely sa použijú len relevantné genetické markery.

V súčasnosti je v genomickom výskume a v mnohých aplikáciách, akou je napríklad aj pomenovávanie STR sekvencií, najpoužívanejším referenčným genómom GRCh38. [23, 16] V porovnaní so svojím predchodcom, GRCh37, ponúka množstvo vylepšení, akými sú napríklad oprava stoviek štrukturálnych chýb, vyplnenie mnohých medzier a komplexnejšie zastúpenie genetickej diverzity. Tieto aktualizácie výrazne zlepšili presnosť celého radu genomických analýz, umožnili presnejšie mapovanie sekvenčných dát a uľahčili štúdium komplexných genomických oblastí, ktoré predtým neboli dobre charakterizované. Získanie nových poznatkov o referenčnom genóme bolo umožnené najmä vďaka pokroku v sekvenačných technológiách, a to hlavne prostredníctvom technológie sekvenovania jedinej molekuly DNA v reálnom čase. [9] V dôsledku pokroku technológií bol identifikovaný veľký počet predtým nezistených delecií a inzercií, a to aj v STR lokusoch, čo poukazuje na zložitosť a dynamickú povahu ľudského genómu. Tieto nové zistenia o variantách STR lokusov vytvárajú potenciál pre presnejšie genetické analýzy vrátane forenznej DNA analýzy.

Konverzia SNP, teda proces prevodu pozície a anotácie SNP, medzi dvoma štandardnými verziami ľudského referenčného genómu, GRCh37 a GRChG38, nie je bezchybná, pričom pri konverzii z GRChG38 na GRCh37 je úspešnosť nižšia. Preto verzia GRChG38 môže byť efektívnejšia pri identifikácii SNP, a odporúča sa jej používanie na presnejšiu genetickú analýzu. [34]

2.1.2 Nomenklatúra STR markerov

Univerzálne pomenovávanie komplexných STR lokusov a ich sekvencií, ktoré sú používané vo forenznej DNA analýze, je zložité, a preto je potrebný štandardizovaný systém nomenklatúry. Samotné STR markery sú dnes štandardne označované jednoduchou nomenklatúrou, a to napríklad D3S1358, kde písmeno

D znamená DNA, číslo 3 znamená 3. chromozóm, písmeno S znamená, že ide o jednu kópiu (angl. single copy), nie viacero kópií, a číslo 1358 znamená, že sa lokus nachádza v oblasti 3. chromozómu, ktorá bola popísaná ako 1358. v poradí. [28] Markery, ktoré nespadajú do tejto nomenklatúry, napríklad známy marker CSF1PO, sú pomenované zvyčajne podľa génu, s ktorým sú asociované, v tomto prípade ide o gén kódujúci faktor stimulujúci kolónie makrofágov (angl. Colony Stimulating Factor 1, CSF1).

Jednotlivé sekvencie, ktoré sa môžu vyskytovať v týchto lokusoch, sú označované rôznymi spôsobmi. Stretnutie pracovnej skupiny STRAND (Short Tandem Repeat: Align, Name, Define), ktoré sa konalo v roku 2019 bolo zamerané na nomenklatúru týchto STR sekvencií. [18] Medzi účastníkmi bolo množstvo odborníkov z tejto oblasti, ktorí diskutovali o rôznych systémoch a prístupoch k nomenklatúre sekvencií. Cieľom stretnutia bolo podporiť informovanosť a spoluprácu medzi tými, ktorí vyvíjajú systémy nomenklatúry STR. Medzi hlavné výsledky patrila zhoda o potrebnej minimalite kódu, ktorý označuje konkrétnu sekvenciu. Došlo k návrhu rôznych metód generovania tohto unikátneho kódu, napríklad generovanie unikátneho 55-písmenového kódu z konkrétnej sekvencie pomocou hešovacej funkcie. Tento unikátny kód síce jednoznačne identifikuje sekvenciu, no sprostredkováva veľmi málo informácií o samotnej sekvencii a jej súvislosti s inými sekvenčnými variantami daného lokusu. [23]

Na skompaktnenie repetitívnej oblasti reťazca do opisného a človekom jednoducho čitateľného formátu slúžia tzv. zátvorkované opakovania (angl. bracketed repeats). V tomto formáte je repetícia opakujúca sa v repetitívnej oblasti sekvencie uvedená v hranatých zátvorkách, za ktorými nasleduje číslo určujúce počet opakovaní tejto repetície. Tento formát však môže byť neexaktný, čo môže viesť k nekompatibilným prístupom medzi rôznymi laboratóriami pri získaní dát, ktoré v databáze doposiaľ neboli uložené, a to hlavne v prípade komplexnejších lokusov. [18] Mnohé STR lokusy totiž nepozostávajú iba z jednoduchých repetícií, ale obsahujú aj nekonsenzuálne opakovania, či ešte komplexnejšie štruktúry, podľa ktorých je možné STR sekvencie kategorizovať tak, ako je uvedené nižšie. Ku každej kategórii je uvedený jeden príklad nominálnej alely lokusov, ktorými sa táto práca zaoberá. [19]

- Jednoduché repetície - TPOX:

  [AATG]8

- Jednoduché repetície s nekonsenzuálnymi sekvenciami - TH01:

  [AATG]6 ATG [AATG]3

- Zložené repetície s nekonsenzuálnymi repetíciami - vWA:

  [TAGA]9 [CAGA]4 TAGA

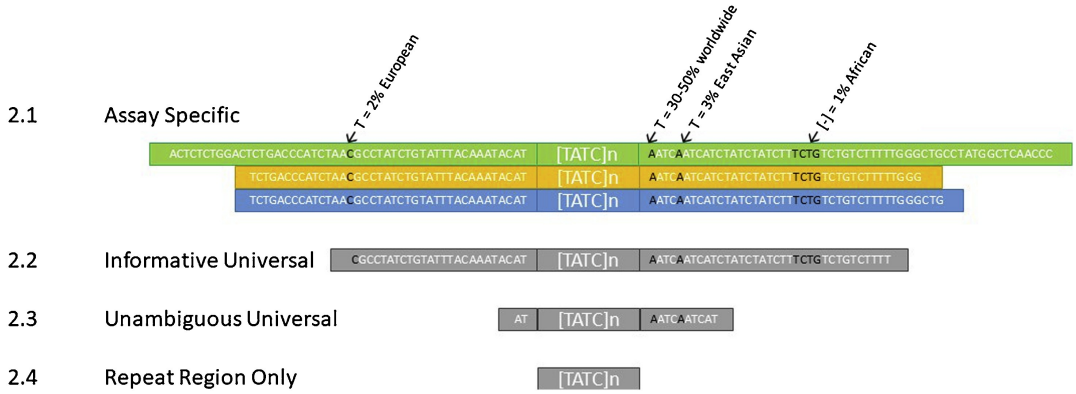
- Komplexné sekvencie - D21S11:

  [TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA

  TA [TCTA]4

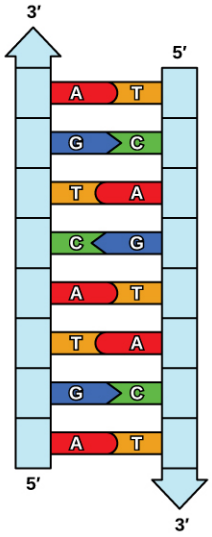
Ďalším významným výstupom zo stretnutia pracovnej skupiny STRAND bolo vyjadrenie potreby jednoznačne určiť súradnice počiatku a konca lokusu vzhľadom na referenčný genóm. Diskutovalo sa aj o potenciálnom referenčnom genóme špecifickom pre forenznú analýzu. V probléme určenia počiatočných a koncových súradníc STR lokusov sa rozoberali 4 možnosti, znázornené aj na obrázku uvedenom nižšie, a to brať do úvahy:

- súradnice maximalizujúce sekvenciu flanking oblastí vzhľadom na konkrétne potreby,
- súradnice, ktoré maximalizujú informatívnosť flanking oblastí, teda najmä výskyt SNP a inzercií či delecií v týchto oblastiach,
- minimálny rozsah súradníc z ohraničujúcich oblastí (angl. flanking regions), ktoré poskytujú jednoznačné ukončenie repetitívnej oblasti,
- alebo iba repetitívnu oblasť.



Obr. 2.1: Možnosti určenia súradníc začiatku a konca STR lokusu [18]

Pre jednotnú nomenklatúru STR sekvencií je dôležitý aj výber referenčného vlákna DNA. Sekvencie DNA sú štandardne čítané od 5'-konca k 3'-koncu, teda v prípade voľby opačného vlákna, než je štandardné, dôjde k čítaniu komplementárnej sekvencie. Obhajované je používanie najčastejšie akceptovanej sekvencie vo forenznej komunite. [19, 16]



Obr. 2.2: Znázornenie komplementarity vláken DNA [31]

V kontexte jednotnosti dát je v súčasnosti významným medzinárodným projektom projekt STRSeq, ktorý je zameraný na katalogizáciu sekvenčnej diverzity v STR lokusoch, ktoré sa bežne používajú pri forenznom profilovaní DNA na identifikáciu ľudí. [17] Cieľom tejto iniciatívy, ktorú schválila komisia ISFG (International Society for Forensic Genetics), je spomínaná štandardizácia nomenklatúry

alel STR lokusov pre zabezpečenie konzistentnosti a presnosti vo všetkých forenzných laboratóriách. Spojením dát od štyroch výskumných skupín (NIST, King's College London, University of North Texas Health Sciences Center a University of Santiago de Compostela) sa v rámci projektu zhromaždili počiatočné dáta od 4612 osôb. Tieto dáta sa uchovávajú v záznamoch GenBank, organizovaných v rámci projektu BioProject v spoločnosti National Center for Biotechnology Information (NCBI), čo uľahčuje prístup k informáciám o sekvenciách, anotáciám a ukazovateľom kvality. Projekt rieši potrebu forenznej komunity po štandardizovanej nomenklatúre sekvencií STR a jeho cieľom je podporiť budúce forenzné aplikácie technológie sekvenovania DNA vrátane kontroly kvality a analýzy frekvencie alel.

2.2 Austrian National DNA Database

Rakúska národná DNA databáza sekvenčných polymorfizmov STR lokusov je projekt, ktorý bol založený v Rakúsku v roku 1997 na výskumné účely. Novo vyvíjaná databáza sa totiž sústredí na sekvenčné polymorfizmy, ktoré v porovnaní s dĺžkovými polymorfizmami dnes zatiaľ nie sú rutinne využívané. Vrámci tohto projektu bol už v roku 1998 navrhnutý univerzálny principiálny model pre dizajn národnej DNA databázy. [38] Projekt bol zaintegrovaný do existujúceho forenzného laboratória, ale bol preň vyvinutý nový systém narábania so vzorkami a organizácie dát získaných z nich.

V súčasnosti databáza vyvinutá vrámci tohto projektu obsahuje referenčné vzorky od 248 osôb rakúskej štátnej príslušnosti, ktoré boli predtým analyzované pomocou konvenčných metód CE na genotypizáciu STR. Všetkých 248 bukálnych sterov bolo získaných v súlade s rakúskym režimom ochrany osobných údajov. Vzorky boli náhodne vybrané výkonnými orgánmi Rakúskeho federálneho ministerstva vnútra (Austrian Federal Ministry of the Interior) podľa kriterií, akými boli napríklad mužské pohlavie, rakúska národnosť a konkrétne miesta narodenia. Nové dáta získané z týchto vzoriek pomocou MPS boli generované za použitia sekvenačnej knižnice PowerSeq 46GY kit. [22]

Popis dát

Databáza zahŕňa dáta získané z analýzy 22 autozomálnych STR lokusov, 23 Y-STR lokusov, a amelogenínu. Je teda rozšírením systémov ESS či CODIS. Autozomálne STR lokusy zahrnuté v databáze sú konkrétne tieto:

- D1S1656
- TPOX
- D2S1338
- D2S441
- D3S1358
- FGA
- D5S818
- CSF1PO
- D7S820
- D8S1179
- D10S1248
- TH01
- vWA
- D12S391
- D13S317
- Penta E
- D16S539
- D18S51
- D19S433
- D21S11
- Penta D
- D22S1045

Do tejto databázy boli zaintegrované technológie MPS s cieľom zvýšiť rozlíšenie analýzy STR lokusov a zlepšiť tak presnosť genetickej identifikácie vo forenzných aplikáciách v rámci rakúskeho právneho systému. V rámci využitia MPS technológií bolo identifikovaných 25 nových sekvenčných variánt z 15 rôznych analyzovaných markerov. Tieto varianty predtým neboli popísané ani v spomínanom katalógu STRSeq. Napriek veľkej miere zhody medzi dátami získanymi pomocou CE a MPS, výsledky tejto štúdie preukázali potrebu jednotného systému nomenklatúry alel, ktorý by bol rovnako použiteľný pre obe technológie a zaručil tak spätnú kompatibilitu, ale zároveň by bol schopný využiť zvýšený informačný obsah MPS. [22]

V rámci tejto práce boli dáta Austrian National DNA Database inšpiráciou pre usporiadanie dát z českej databázy CODISeq. Práve rakúska databáza bola zvolená v dôsledku geografickej aj právnej blízkosti rakúskej a českej judikatúry, ktoré by v budúcnosti mohli tieto databázy využívať. Výstupom práce sú okrem iného aj tabuľky vizualizujúce dáta autozomálnych STR markerov, spoločných pre obe databázy, podobným spôsobom ako vrámci tohto rakúskeho projektu.

2.3 Databáza CODISeq

Databáza CODISeq bola vyvinutá Kriminalistickým ústavom Polície Českej republiky (KÚ PČR). CODISeq je názov používaný dočasne počas vývojového štádia databázy, keďže databáza zatiaľ nie je oficiálnou databázou a slúži na vnútorné výskumné účely. Poskytuje rozhranie na databázovanie STR polymorfizmov, ktoré sa využívajú na určenie príbuznosti osôb vrámci identifikácie neznámeho biologického materiálu. Databáza obsahuje dáta získané pomocou MPS, ktoré rozširujú spektrum bežne používaných identifikačných markerov v databáze CODIS. Aby teda mohla byť aj táto databáza plne využívaná, je potrebné zabezpečiť ich vzájomnú kompatibilitu.

Vstupné dáta do databázy CODISeq boli generované sekvenovaním za použitia sekvenačnej knižnice ForenSeq. Tieto dáta boli následne spracované softvérom UAS (Universal Analysis Software), ktorý umožňuje generovať dva typy výstupov, ktorými sú:

- analýza repetitívnej oblasti sekvencie
- analýza repetitívnej oblasti sekvencie a jej priľahlých oblastí (angl. flanking regions)

Väčšina dát z databázy CODISeq je validovaná porovnaním hodnoty nominálnej alely určitého STR získanej pomocou MPS s hodnotou tejto nominálnej alely získanej pomocou fragmentačnej analýzy na kapilárnej elektroforéze. V prípade kompatibility oboch typov dát boli sekvencie uložené do databázy.

Vrámci tejto práce bolo k týmto dátam pristupované pomocou rozhrania MySQL Workbench, pomocou ktorého bola databáza nainštalovaná prostredníctvom príslušných skriptov sprostredkovaných KÚ PČR spustených nad súbormi textového formátu comma separated value (CSV) obsahujúcimi samotné dáta. Tieto súbory predstavujúce jednotlivé tabuľky databázy obsahovali rôzne dáta týkajúce sa autozomálnych aj gonozomálnych markerov, táto práca sa však sústredila na tie autozomálne.

Popis dát

Databáza CODISeq predstavuje vzorku populačných dát získanú z približne 500 ľudí z českej populácie. Približne 400 profilov sa už nachádzalo v databáze CODIS a ďalších 100 bolo získaných analýzou biologického materiálu, konkrétne slín, dobrovoľných darcov. V databáze sa nachádzajú dáta týkajúce sa 28 STR markerov, pričom 1 marker je gonozomálny a 27 markerov je autozomálnych. Gonozomálnym markerom slúžiacim na určenie pohlavia je Amelogenin X/Y a autozomálne markery sú:

- D1S1656
- TPOX
- D2S1338
- D2S441
- D3S1358
- D4S2408
- FGA
- D5S818
- CSF1PO
- D6S1043
- D7S820
- D8S1179
- D9S1122
- D10S1248
- TH01
- vWA
- D12S391
- D13S317
- Penta E
- D16S539
- D17S1301
- D18S51
- D19S433
- D20S482
- D21S11
- Penta D
- D22S1045

Tabuľky spájajúce tieto dáta umožňujú analýzu frekvencií jednotlivých alelických variánt, či už dĺžkových alebo sekvenčných. Zahŕňajú dáta ako počty ľudí, u ktorých boli pozorované určité sekvenčné varianty, a z toho vypočítané samotné frekvencie jednotlivých sekvenčných variánt. Tieto údaje sú významné pri analýze pravdepodobnosti náhodnej zhody a v diskriminačnej analýze. Pravdepodobnosť náhodnej zhody určitého DNA profilu predstavuje pravdepodobnosť, že náhodne vybraný jedinec z populácie bude mať rovnaký DNA profil ako daný DNA profil. Nižšia frekvencia sekvenčnej varianty preto znamená vyššiu rozlišovaciu schopnosť, keďže je menej častá, a teda individuálne identifikovateľnejšia.

Cieľom práce je využiť tieto populačné dáta a zvýšiť tak úroveň práce s nimi

– poskytnúť lepšiu prístupnosť, a to konkrétne transformáciou príslušných dát do kompatibilnejšieho formátu a vizualizáciou týchto dát, ktorá je inšpirovaná rakúskym projektom Austrian National DNA Database a má poskytnúť jednoduchšiu interpretovateľnosť dát než pôvodné tabuľky.

2.4 Budúcnosť forenzných DNA databáz

Úspech a rýchle rozširovanie forenzných DNA databáz so sebou prináša výzvy, ako je napríklad riziko vyššieho počtu náhodných zhôd, teda falošných zhôd dvoch DNA profilov čisto náhodou, v dôsledku veľmi veľkého počtu profilov, potreba zlepšenia výkonu pre nové aplikácie, ako je identifikácia nezvestných osôb či hľadanie príbuzenstva, a zvýšenie snahy medzinárodného zdieľania údajov. V identifikácii nezvestných osôb pomocou medzinárodného porovnávania príbuzenských vzťahov na základe DNA je v súčasnosti priekopníkom databáza I-Familia a s hľadaním príbuzenstva vo forenznom kontexte pomáha projekt GEDmatch PRO. [25, 15]

Diskusia o vytvorení univerzálnej DNA databázy pre orgány činné v trestnom konaní tiež prináša niekoľko výhod a zároveň čelí problémom. V súčasnosti sú na riešenie trestných činov využívané rôzne štátne, federálne a súkromné genetické databázy, pričom tomuto prístupu chýba komplexná regulácia. Univerzálna databáza, ktorá by zahŕňala DNA osôb všetkých krajín, by mohla zefektívniť vyšetrovanie, zvýšiť spravodlivosť a potenciálne zabrániť trestnej činnosti. Tento návrh však vyvoláva závažné otázky týkajúce sa ochrany súkromia a etiky, keďže komplexný charakter takejto databázy by mohol viesť k potenciálnemu zneužitiu genetických informácií. [20]

3. Transformácia formátu dát

Táto kapitola popisuje transformáciu populačných dát Českej republiky o STR markeroch z CSV súborov do univerzálneho formátu JavaScript Object Notation (JSON) a ich následnú vizualizáciu v tabuľkách formátu Microsoft Excel Spreadsheet (XLSX). Táto praktická implementácia je inšpirovaná vizualizáciou obdobných dát vrámci rakúskeho projektu Austrian National DNA Database a zdôrazňuje tak význam kompatibility údajov naprieč databázami. Vizualizácia využíva dnes často používanú nomenklatúru s použitím bracketed repeats. Táto vizualizácia, ako aj formát JSON poskytuje prehľadnosť a efektívnosť analýzy a interpretácie týchto dát.

3.1 Nástroje

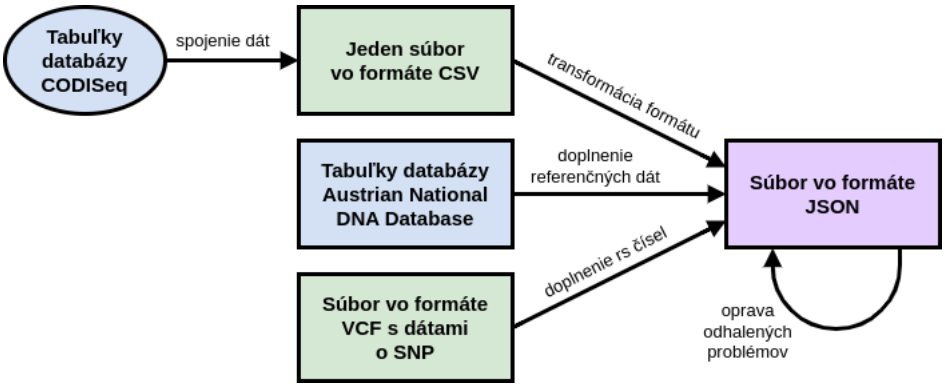
Nižšie popísané skripty boli implementované v programovacom jazyku Python s výnimkou prvého, ktorý bol napísaný v jazyku Structured Query Language (SQL). Boli použité nasledujúce Python knižnice:

- `pandas` – na základné operácie s dátami
- `json` – na prácu s JSON súborom
- `re` – na prácu s regulárnymi výrazmi
- `pysam` – na prácu so súborom formátu Variant Call Format (VCF)
- `xlsxwriter` – na zapisovanie dát do výstupných Excel tabuliek
- `math` – na zložitejšie matematické výpočty
- `csv` – na prácu s CSV súbormi

Pre správny beh programu je potrebné tieto knižnice nainštalovať, napríklad pomocou príkazu `pip install pandas` v príkazovom riadku. Podobne aj pre ostatné knižnice.

3.2 Proces transformácie dát

Uvedený diagram znázorňuje postup práce s dátami pri ich prevode z CSV tabuliek databázy CODISeq do jedného štrukturovaného JSON súboru. Na vykonanie jednotlivých krokov boli vyvinuté jednotlivé skripty popísané nižšie.



Obr. 3.1: Diagram znázorňujúci algoritmus transformácie dát

3.2.1 Súbor `0_join_tables.sql`

Jednoduchý SQL skript spája dáta z databázy CODISeq do jedného CSV súboru pre lepšiu prácu s nimi. Používa jednoduchý JOIN dvoch tabuliek. Tabuľka `marker_auto_strview` obsahuje dáta o jednotlivých sekvenčných variantách 27 autozomálnych markerov. Tabuľka `marker_auto_strview_flankingreg` obsahuje dáta o sekvenčných variantách markerov spolu s ich flanking oblasťami, teda oblasťami, ktoré ohraničujú repetitívnu oblasť markera. Skript spája tieto dve tabuľky, aby sa dalo s danými dátami pracovať naraz. Tieto dve tabuľky sú spojené prostredníctvom stĺpca, ktorý obsahuje samotnú sekvenciu (v prípade tabuľky `marker_auto_strview` ide iba o repetitívnu oblasť, v prípade tabuľky `marker_auto_strview_flankingreg` ide o repetitívnu oblasť spolu s flanking oblasťami). Sekvencia je reprezentovaná ako reťazec znakov, skript teda každej sekvencii z tabuľky `marker_auto_strview_flankingreg` priraďuje príslušný riadok z tabuľky `marker_auto_strview`, kde sa nachádza rovnaká repetitívna oblasť.

V niekoľko málo prípadoch však došlo k tomu, že sa zhodná repetitívna oblasť nenašla, takže nebolo na základe čoho vyčleniť príslušnú repetitívnu oblasť

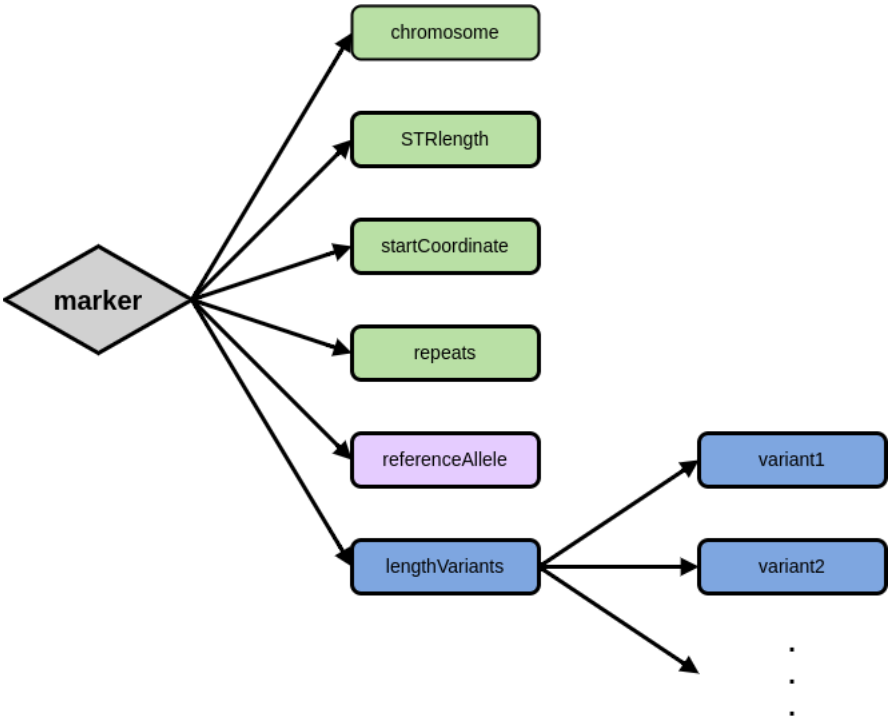
od flanking oblastí, a preto došlo k vypadnutiu niektorých riadkov z tabuľky `marker_auto_strview_flankingreg`. Spôsobilo to však len malé odchýlky, a to konkrétne, že namiesto celkového súčtu záznamov, ktorý by mal byť 420, je pri uvedených piatich markeroch tento súčet menší.

D1S1656	418
D2S1338	419
D12S391	415
D18S51	416
D21S11	419

3.2.2 Súbor `1_csv_to_json.py`

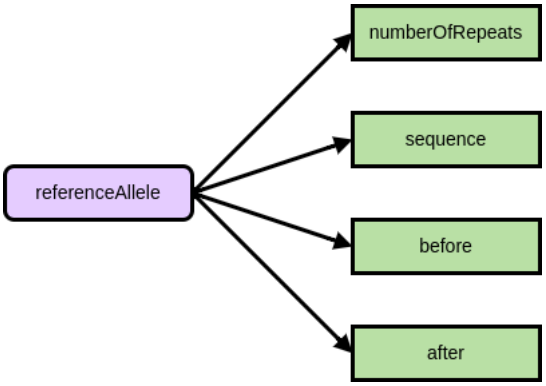
Jednoduchý skript napísaný v programovacom jazyku Python prevádza dáta z formátu CSV do formátu JSON, pre ich univerzalitu, štruktúrovateľnosť a lepšiu prácu s nimi. Skript iteruje všetkými riadkami CSV súboru a postupne z nich ukladá dáta do JSON súboru na príslušné miesta do JSON štruktúry. Štruktúra JSON súboru je nižšie popísaná pomocou diagramov.

Každý marker je v JSON súbore uložený ako samostatný objekt pod vlastným názvom, ktorý má potom ako atribúty číslo chromozómu `chromosome` (keďže sa jedná len o autozomálne markery), číselnú veľkosť repetície `STRlength` (v tomto prípade 3, 4 alebo 5), číselnú súradnicu (podľa referenčného genómu GRCh38) počiatku repetetívnej oblasti `startCoordinate`, zoznam repetícií `repeats`, ktoré sa v referenčnej sekvencii vyskytujú, referenčnú alelu `referenceAllele` a dĺžkové varianty `lengthVariants`.



Obr. 3.2: Štruktúra objektu reprezentujúceho jeden STR marker

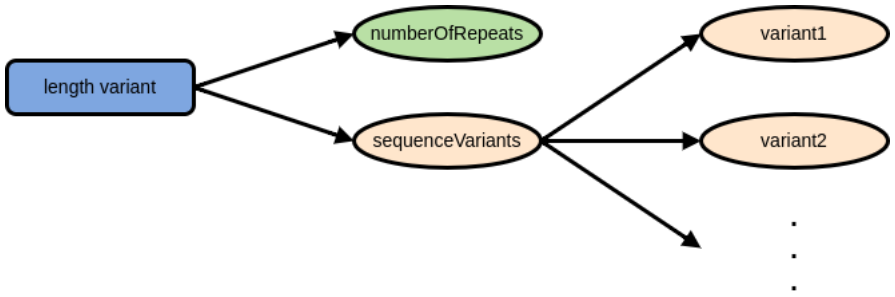
Referenčná alela `referenceAllele` je samostatným objektom, ktorý ako má atribúty počet opakovaní repetícií `numberOfRepeats`, sekvenciu `sequence` reprezentovanú ako reťazec a flanking oblasti `before` a `after` tiež reprezentované ako reťazce. Tieto dáta sú však doplnené až ďalším skriptom, ktorý je popísaný v ďalšej sekcii.



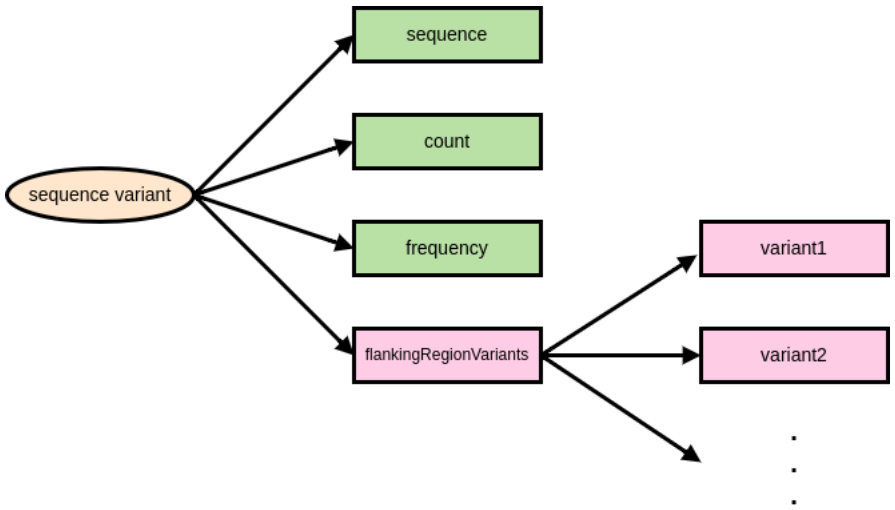
Obr. 3.3: Štruktúra objektu reprezentujúceho referenčnú alelu

Dĺžkové varianty `lengthVariants` sú tvorené zoznamom jednotlivých variánt, pričom každá varianta je samostatný objekt. Tento objekt má ako atribút opäť počet opakovaní repetícií `numberOfRepeats` a zoznam jednotlivých sekvenčných

variánt `sequenceVariants`. Každá sekvenčná varianta je opäť samostatný objekt, ktorý má ako atribúty samotnú sekvenciu `sequence` reprezentovanú reťazcom, počet ľudí `count`, u ktorých bola táto varianta prečítaná, z nej vypočítanú frekvenciu tejto varianty v populácii `frequency` a zoznam variánt flanking oblastí `flankingRegionsVariants`.

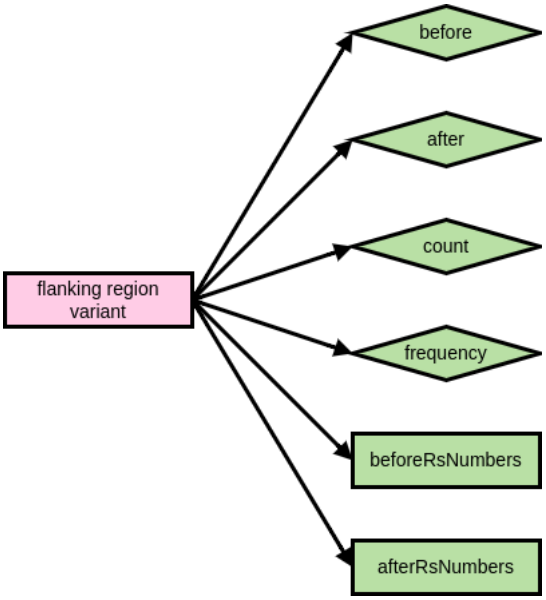


Obr. 3.4: Štruktúra objektu reprezentujúceho jednu dĺžkovú variantu



Obr. 3.5: Štruktúra objektu reprezentujúceho jednu sekvenčnú variantu

V zozname variánt flanking oblastí `flankingRegionsVariants` je opäť každá varianta samostatným objektom, ktorý má ako atribúty sekvencie `before` a `after` reprezentované ako reťazce, počet ľudí `count`, u ktorých bola táto varianta prečítaná a z tohto počtu vypočítanú frekvenciu tejto varianty v populácii `frequency`.



Obr. 3.6: Štruktúra objektu reprezentujúceho jednu variantu flanking oblastí

3.2.3 Súbor `2_xlsx_to_json.py`

Jednoduchý Python skript čerpá dáta z tabuľky `Table S6` referenčnej rakúskej databázy Austrian National DNA Database. Ide o atribúty týkajúce sa samotných markerov, a to konkrétne o atribúty `chromosome`, `STRlength` a `startCoordinate`, a atribúty ich referenčných alel, a to konkrétne o atribúty `numberOfRepeats`, `sequence`, `before` a `after`, ktoré sú do JSON súboru uložené na príslušné miesta.

Skript postupne prechádza tabuľku v XLSX formáte a číta z nej jednotlivé referenčné sekvencia, podľa toho, kde je v danej tabuľke uvedené, že sa nachádza, a to vždy v riadku označenom *Reference sequence*. Vyčlenenie repetitívnej oblasti od flanking oblastí sa deje na základe detekcie čísla 1 pri začiatku repetitívnej oblasti. Toto číslo označuje prvú repetíciu. Odtiaľ pochádza aj počiatočná súradnica repetitívnej oblasti. Dĺžka repetície je potom určená ako vzdialenosť tohto čísla 1 od čísla 2, teda od označenia najbližšej repetície.

Keďže v databáze CODISeq je o 5 markerov viac než v Austrian National DNA Database, v JSON súbore je referenčná alela uvedená pri 22 z 27 markerov.

3.2.4 Súbor `3_fix_problems.py`

Pri práci s dátami došlo k odhaleniu viacerých problémov, ktoré tento skript rieši. Okrem toho skript napĺňa zoznam `repeats`, ktorý je atribútom referenčnej alely každého markera. Deje sa to na tomto mieste, pretože v niektorých prípadoch bolo potrebné manuálne porovnanie repetitívnej sekvencie s databázami a pevná úprava tohto zoznamu. Napríklad marker D21S11 má výnimočne variabilnú sekvenciu, a tak sa pri automatickom vytváraní zoznamu `repeats` doň dostali aj repetície, ktoré nie sú považované za bežné v tomto lokuse. [43]

Pri práci s pôvodnými dátami databázy CODISeq došlo k zisteniu, že databáza Austrian National DNA Database sa s databázou CODISeq nezhoduje pri niektorých markeroch vo vlákne DNA, z ktorého je sekvencia čítaná. Skript zjednocuje vlákno, z ktorého je DNA čítaná, tak, aby pre každý marker bolo použité rovnaké vlákno v CODISeq aj Austrian National DNA Database. Prechádza jednotlivými sekvenčnými variantami alel markrov z CODISeq, pri ktorých bola manuálne zistená nezhoda, a mení ich sekvenciu na komplementárnu. Konkrétne išlo o týchto 9 markerov: D1S1656, D2S1338, FGA, D5S818, CSF1PO, D7S820, vWA, PentaE, D19S433.

Okrem toho došlo k zisteniu, že 9 markerov z databázy CODISeq má nesprávne oddelené flanking oblasti. O 5'-flanking oblasť ide v týchto 5 prípadoch: D1S1656, D5S818, D7S820, PentaD, vWA. O 3'-flanking oblasť ide v týchto 4 prípadoch: D13S317, D18S51, D19S433, D21S11. Pravdepodobne došlo k nesprávnemu vyčleneniu repetitívnej oblasti od flanking oblastí počas vytvárania databázy CODISeq. V týchto 9 prípadoch je repetitívna oblasť dlhšia než má byť, pretože je v nej zahrnutá aj časť flanking oblasti. Skript `3_fix_problems.py` oddeľuje časť, ktorá je v repetitívnej oblasti navyše a priraďuje ju k flanking oblasti kde, naopak, táto časť chýba. V prípade markera D19S433 sa nepodarilo vyčleniť repetitívnu oblasť správne v dôsledku chýb v pôvodných dátach. V týchto sekvenciách sa pravdepodobne nachádzajú inzercie či delecie, ktoré však v pôvodných dátach neboli nijakým spôsobom popísané, a tak sa vyčlenenie repetitívnej oblasti nemalo o čo oprieť. Tento problém bol odhalený pri vizualizácii dát, keď sa zarovnanie týchto sekvencií ukázalo ako chybné aj po snahe o správne vyčlenenie repetitívnej oblasti. V dôsledku tohto bol marker D19S433 z procesovania dát

vynechaný a vo vizualizácii je teda 21 markerov.

V prípade markeru PentaE nedošlo k správnemu prečítaniu jeho refenčnej sekvencie. Konkrétne nedošlo k vyčleneniu repetitívnej oblasti od flanking oblastí, pretože v referenčnej databáze Austrian National DNA Database chýbalo označenie prvej repetície číslom 1. Bolo teda potrebné manuálne identifikovať repetitívnu oblasť a vyčleniť ju od flanking oblastí rovnako ako napevno nastaviť atribúty `chromosome`, `STRlength` a `startCoordinate`.

Ďalej boli pomocou tohto skriptu poupravené rozsahy sekvencií flanking oblastí referenčných alel. Keďže získavanie dát uvedených v referenčnej rakúskej databáze Austrian National DNA Database sa pochopiteľne líšilo od získavania dát uvedených v českej databáze CODISeq, flanking oblasti boli prečítané vždy do rôznej dĺžky. Tento skript zjednocuje ich dĺžky, a to v tom zmysle, že ak referenčná alela mala flanking oblasti dlhšie, než alelické varianty z databázy CODISeq, tak boli referenčné sekvencie skrátené, pretože pre účely databázy CODISeq boli zbytočne pridlhé. Tento problém bol veľmi častý. Naopak, ak boli sekvencie flanking oblastí v dátach CODISeq dlhšie než referenčné, došlo k ich skráteniu, pretože by ich nebolo s čím porovnávať. V tomto prípade by pravdepodobne bolo lepšie rozšíriť referenčnú sekvenciu pomocou inej databázy, no keďže išlo o prípad veľmi výnimočný, tak bol zvolený tento prístup. Konkrétne bola skrátená iba sekvencia markeru D22S1045 o 6 nukleotidov a sekvencia markeru PentaD o 5 nukleotidov, čo je dokopy v porovnaní s dátami obsiahnutými v celej databáze zanedbateľné.

3.2.5 Súbor `4_rs_to_json.py`

Tento skript napísaný tiež v programovacom jazyku Python dopĺňa JSON súbor o ďalšie dôležité dáta, a to konkrétne o Reference SNP cluster ID (rs) čísla, teda označenia jednotlivých SNP vo flanking oblastiach. Tento skript prechádza všetky varianty flanking oblastí pre každý marker nachádzajúci sa v JSON súbore a porovnáva ich s referenčnou sekvenciou. Ak je na nejakom mieste odhalený SNP, podľa koordináty tohto nukleotidu v referenčnom genóme GRCh38 a príslušného čísla chromozómu je mu vyhľadané unikátne príslušné rs číslo.

Toto rs číslo je vyhľadané v súbore `00-common_all.vcf` formátu VCF, ktorý

bol stiahnutý z databázy NCBI prostredníctvom príkazového riadku. Tento súbor obsahuje všetky aktuálne existujúce rs čísla pre doteraz identifikované SNP v ľudskom genóme. Vzhľadom na veľkosť týchto dát by do budúcna bolo vhodné nahradiť sťahovanie tohto súboru priamym vyhľadávaním rs čísel v online databáze. Pri pokuse o tento prístup však niektorým SNP bolo priradené viac než jedno rs číslo (to mohlo byť spôsobené napríklad tým, že databázy obsahujú aj staršie označenia SNP), naproti čomu VCF súbor zabezpečuje unikátnosť rs čísel.

Identifikátory SNP sú do JSON súboru ukladané spolu s indexom príslušných SNP vrámci flanking oblastí pre ďalšiu prácu s týmito dátami. V niekoľkých prípadoch však nebolo rs číslo nájdené vôbec. V týchto prípadoch ide buď o nesprávne osekvenovanie sekvencií pri vytváraní primárnych dát (to je pravdepodobné najmä pri sekvenčných variantách, kde došlo k nízkemu počtu čítaní), alebo sa môže skutočne jednať o novo objavené SNP. To si však vyžaduje ďalšie preverenie a prípadné následné nahranie týchto pozícií do online databáz, ktoré tieto dáta spravujú, aby im bolo priradené nové rs číslo. Indexy týchto SNP boli do JSON súboru uložené aj napriek chýbajúcemu rs číslu, aby mohli byť neskôr vizualizované.

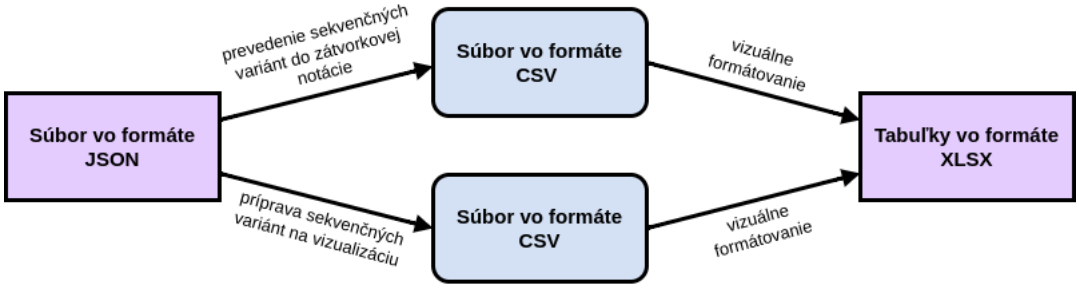
3.3 Vizualizácia dát

Ďalším cieľom práce bolo dáta vizualizovať dvoma spôsobmi. Prvý spôsob sa riadi vizualizáciou dát v tabuľke `Table S1` rakúskej databázy Austrian National DNA Database. Dôraz sa kladie na prehľadnosť jednotlivých sekvenčných variánt, a to hlavne v dvoch aspektoch:

- SNP polymorfizmy sú vyznačené červeno a príslušné rs čísla sú uvedené v samostatnom stĺpci.
- Sekvencia repetitívnej oblasti je uvedená v univerzálnej zátvorkovej notácii.

Druhá vizualizácia sa riadi vizualizáciou dát v tabuľke `Table S6` rakúskej databázy Austrian National DNA Database. Dôraz sa kladie na prehľadnosť jednotlivých sekvenčných variánt v súvislosti s referenčnou alelou. Sekvencie sú uvedené po jednotlivých nukleotidoch a tie sú farebne odlíšené pre lepšiu viditeľnosť sekvenčných variánt a SNP vo flanking oblastiach.

Diagram znázorňuje proces vizualizácie dát dvoma uvedenými spôsobmi s využitím prevodu dát do jednoduchých CSV súborov pred samotnou formátovanou vizualizáciou v tabuľkách. Na vykonanie jednotlivých krokov boli vyvinuté jednotlivé skripty popísané nižšie.



Obr. 3.7: Diagram znázorňujúci algoritmus vizualizácie dát

3.3.1 súbor `5_table1_vis_prep.py`

Skript prevádza dáta z JSON súboru do CSV súboru, ktorý slúži ako príprava na výslednú vizualizáciu. Už v tomto súbore sú však sekvencie prehľadne zapísané v notácii s použitím bracket repeats. Tento skript je od finálnej vizualizácie oddelený preto, aby bolo možné dáta zobraziť aj bez akéhokoľvek špeciálneho formátovania a potrebného softvéru, prípadne s nimi ďalej pracovať.

3.3.2 súbor `6_table2_vis_prep.py`

Skript, ktorý opäť slúži na prípravu dát pred samotnou vizualizáciou, tiež čerpá dáta z JSON súboru a ukladá ich do CSV súboru, aby mohli byť tieto dáta zobrazené bez akéhokoľvek formátovania. Sekvencie sú ukladané po jednotlivých nukleotidoch, pre lepšiu viditeľnosť polymorfizmov. Zobrazené sú tiež referenčné sekvencie, koordináty z referenčného genómu GRCh38 a číselné vzdialenosti jednotlivých nukleotidov vo flanking oblastiach od repetitívnej oblasti.

3.3.3 súbor `7_visualize.py`

Tento skript prevádza dáta z dvoch prípravných CSV súborov do finálnej Excel tabuľky, teda do súboru formátu XLSX, s dvoma hárkami. Ide prevažne o kopírovanie jednotlivých hodnôt a ich formátovanie do vizuálne príjemnejšej podoby.

V prvom hárku je dôležité najmä zvýraznenie SNP vo flanking oblastiach červeným podčiarknutým písmom, použitie fontu Courier New na zarovnanie jednotlivých sekvencií a nastavenie šírky jednotlivých stĺpcov tak, aby boli všetky údaje viditeľné.

V druhom hárku je dôležité formátovanie jednotlivých písmen reprezentujúcich nukleotidy. Tie sú podfarbované jednou z prislúchajúcich štyroch farieb. Toto poskytuje prehľadnosť dát a jednoduchú rozoznateľnosť SNP. Ďalej sú formátované stĺpce a riadky tak, aby boli dostatočne široké či vysoké, aby všetky hodnoty, ktoré obsahujú, boli viditeľné celé.

Pri vizualizácii dát boli odhalené ďalšie komplikácie, ktoré by bolo dobré v budúcnosti zlepšiť. V prvom rade ide o dve identifikované inzercie, ktoré v pôvodných dátach neboli nijakým spôsobom popísané. Prvá inzercia bola identifikovaná v 5'-flanking oblasti varianty nominálnej alely 9.1 markeru D7S820 a druhá v 3'-flanking oblasti jednej z variánt nominálnej alely 24 markeru FGA. V druhom rade ide o neoptimálne zarovnanie sekvencií, a to v prípade dvoch vysoko komplexných markerov FGA a D21S11.

3.4 Spustenie programu

Na spustenie skriptov je potrebné naklonovať GitHub repozitár, ku ktorému vedie QR kód uvedený v prílohe A. V tomto repozitári sa však nenachádzajú pôvodné dáta, pretože tie nemôžu byť verejne prístupné.

Pred samotnou transformáciou formátu dát a ich vizualizáciou je potrebné stiahnuť spomínaný VCF súbor obsahujúci rs čísla. Na to slúži skript `get_vcf.sh`, takže stačí v GitHub repozitári spustiť nasledujúci príkaz v príkazovom riadku:

```
$ ./code/get_vcf.sh
```

Na transformáciu formátu dát je možné spustiť štyri skripty popísané vyššie, a to priamo pomocou skriptu `transform_data.sh`, ktorý ich spúšťa v potrebnom poradí. Stačí teda v GitHub repozitári spustiť nasledujúci príkaz:

```
$ ./code/transform_data.sh
```

Na vizualizáciu dát je možné spustiť tri skripty popísané vyššie, a to priamo pomocou skriptu `visualize.sh`, ktorý ich spúšťa v potrebnom poradí. Stačí teda v GitHub repozitári spustiť nasledujúci príkaz:

```
$ ./code/visualize.sh
```

Tieto skripty sa starajú aj o prípadné mazanie existujúcich výstupných súborov pred spúšťaním Python skriptov. Výstupné súbory, teda súbor vo formáte JSON `transformed_data.json` a tabuľky vo formáte XLSX `Tables.xlsx` sa potom nachádzajú v adresári `data/output`.

3.5 Výsledky

Hlavným výstupom tejto práce sú súbory formátu JSON a XLSX a samotné skripty vyvinuté na prácu s týmito dátami.

Súbor `transformed_data.json` poskytuje univerzálne rozhranie pre budúcu štatistickú analýzy alebo ďalšie formy vizualizáci. Tento súbor je navrhnutý tak, aby bol jednoducho rozšíriteľný a kompatibilný s rôznymi analytickými nástrojmi a poskytoval flexibilitu v prístupe k dátam.

Súbor `Tables.xlsx` obsahujúci dva hárky s rozlišnými spôsobmi vizualizácie dát je zameraný na poskytnutie intuitívnej a prístupnej vizualizácie dát, ktorá uľahčuje analýzu a interpretáciu sekvenčných variánt STR lokusov v českej populácii. Je to praktický nástroj, ktorý by mohol slúžiť Kriminalistickému ústavu PČR na jednoduchšiu analýzu a intepretáciu dát.

Okrem toho sú skripty, ktoré boli vyvinuté a použité pre transformáciu dát, manipuláciu s nimi a ich vizualizáciu, dôležitou súčasťou výsledkov tejto práce. Tieto skripty predstavujú rozhranie, ktoré by mohlo slúžiť ako nástroj na efektívnu prácu s dátami o STR markeroch po ďalšom rozšírení či adaptácii na spracovanie konkrétnych podobných dátových súborov.

Záver

Cieľom tejto práce bolo navrhnúť efektívnejšiu organizáciu a jednoduchšiu správu dát DNA polymorfizmov, konkrétne STR markerov, ktoré sú kľúčové pre identifikáciu osôb vo forenznej DNA analýze. Práca sa sústredila na transformáciu existujúcich populačných dát Českej republiky do formátu, ktorý je dnes široko používaný a mohol by byť kompatibilný so súčasnými forenznými DNA databázami. Implementovaná transformácia dát zo súborov formátu CSV do súboru formátu JSON a následná vizualizácia dát inšpirovaná projektom Austrian National DNA Database, by mohli umožniť lepšiu interpretovateľnosť a analyzovateľnosť dát.

V práci bolo narábané so sekvenčnými variantami STR polymorfizmov, teda s dátami, ktoré dnes zatiaľ nie sú rutinne využívané. Tieto dáta boli získané na Kriminalistickom ústave Polície Českej republiky pomocou metód masívneho paralelného sekvenovania, ktoré sa v súčastnosti stáva dominantným pri akomkoľvek získavaní dát z DNA sekvencií. Práca sa teda opiera o moderné metódy súčasnej vedy a prispieva k nej jej výstupmi.

Pri práci s dátami bolo odhalených niekoľko problémov, ktoré pravdepodobne súviseli s generovaním primárnych dát pomocou masívneho paralelného sekvenovania a ich následnej správy. Prípadné rozšírenie tejto práce by malo spočívať v opravení chýb, ktoré vrámci práce opravené neboli. Ide hlavne o popísanie dvoch identifikovaných inzercií a o optimalizáciu zarovnania sekvencií vo vizualizácii v prípade vysoko komplexných markerov.

Ciele práce boli v zásade naplnené. Dáta boli úspešne pretransformované do prístupnejšieho formátu, vytvorená vizualizácia by mohla slúžiť na skutočnú dátovú analýzu a skripty by po menších úpravách či rozšíreniach mohli slúžiť na ďalšiu manipuláciu s podobnými dátami a potenciálne tak pomôcť v tejto oblasti výskumu na Kriminalistickom ústave Polície Českej republiky.

Zoznam použitej literatúry

[1] Al-Samarai, F. R. a Al-Kazaz, A. A. (2015). Molecular markers: An introduction and applications. *European journal of molecular biotechnology*, **9**(3), 118–130.

[2] Alamoudi, E., Mehmood, R., Albeshri, A. a Gojobori, T. (2018). DNA Profiling Methods and Tools: A Review. In *Smart Societies, Infrastructure, Technologies and Applications*, pages 216–231.

[3] Alonso, A., Martín, P., Albarrán, C., Garcí, P., De Simón, L. F., Iturralde, M. J., Fernández-Rodríguez, A., Atienza, I., Capilla, J., García-Hirschfeld, J. a kol. (2005). Challenges of DNA profiling in mass disaster investigations. *Croatian medical journal*, **46**(4).

[4] Arenas, M., Pereira, F., Oliveira, M., Pinto, N., Lopes, A. M., Gomes, V., Carracedo, A. a Amorim, A. (2017). Forensic genetics and genomics: Much more than just a human affair. *PLoS genetics*, **13**(9).

[5] Ballard, D., Winkler-Galicki, J. a Wesoły, J. (2020). Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *International Journal of Legal Medicine*, **134**(4), 1291–1303.

[6] Budowle, B., Eisenberg, A. J. a Daal, A. v. (2009). Validity of low copy number typing and applications to forensic science. *Croatian medical journal*, **50**(3), 207–217.

[7] Buermans, H. a den Dunnen, J. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1842**(10), 1932–1941.

[8] Butler, J. M., Buel, E., Crivellente, F. a McCord, B. R. (2004). Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *ELECTROPHORESIS*, **25** (10-11), 1397–1412.

[9] Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J. a Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**(7536), 608–611.

[10] Chen, M., Zhang, J., Zhao, J., Chen, T., Liu, Z., Cheng, F., Fan, Q. a Yan, J. (2020). Comparison of CE- and MPS-based analyses of forensic markers in a single cell after whole genome amplification. *Forensic Science International: Genetics*, **45**, 102211.

[11] Collins, F. S. a Mansoura, M. K. (2001). The human genome project. revealing the shared inheritance of all humankind. *Cancer*, **91**(1 Suppl), 221–225.

[12] Corte-Real, F. (2004). Forensic DNA databases. *Forensic Science International*, **146**, S143–S144.

[13] Divne, A.-M. a Allen, M. (2005). A DNA microarray system for forensic SNP analysis. *Forensic Science International*, **154**(2), 111–121.

[14] Fraser, J. (2020). 46C5DNA profiling and databases. In *Forensic Science: A Very Short Introduction*.

[15] GETmatch PRO (2024). About GEDmatch PRO. URL `https://pro.gedmatch.com/about`. Accessed: 28. 4. 2024.

[16] Gettings, K. B., Bodner, M., Borsuk, L. A., King, J. L., Ballard, D., Parson, W., Benschop, C. C., Børsting, C., Budowle, B., Butler, J. M., van der Gaag, K. J., Gill, P., Gusmão, L., Hares, D. R., Hoogenboom, J., Irwin, J., Prieto, L., Schneider, P. M., Vennemann, M. a Phillips, C. (2024). Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on short tandem repeat sequence nomenclature. *Forensic Science International: Genetics*, **68**, 102946.

[17] Gettings, K. B., Borsuk, L. A., Ballard, D., Bodner, M., Budowle, B., Devesse, L., King, J., Parson, W., Phillips, C. a Vallone, P. M. (2017). STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. *Forensic Science International: Genetics*, **31**, 111–117.

[18] Gettings, K. B., Ballard, D., Bodner, M., Borsuk, L. A., King, J. L., Parson, W. a Phillips, C. (2019). Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. *Forensic Science International: Genetics*, **43**, 102165.

[19] Gill, P., Brinkmann, B., d'Aloja, E., Andersen, J., Bar, W., Carracedo, A., Dupuy, B., Eriksen, B., Jangblad, M., Johnsson, V., Kloosterman, A., Lincoln, P., Morling, N., Rand, S., Sabatier, M., Scheithauer, R., Schneider, P. a Vide, M. (1997). Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. *Forensic Science International*, **87**(3), 185–192.

[20] Hazel, J. W., Clayton, E. W., Malin, B. A. a Slobogin, C. (2018). Is it time for a universal genetic forensic database? *Science*, **362**(6417), 898–900.

[21] Herrera, R. J. a Garcia-Bertrand, R. (2018). *Ancestral DNA, human origins, and migrations*. Academic press.

[22] Hölzl-Müller, P., Bodner, M., Berger, B. a Parson, W. (2021). Exploring STR sequencing for forensic DNA intelligence databasing using the Austrian National DNA Database as an example. *International Journal of Legal Medicine*, **135**(6), 2235–2246.

[23] Hoogenboom, J., Sijen, T. a van der Gaag, K. J. (2021). STRNaming: Generating simple, informative names for sequenced STR alleles in a standardised and automated manner. *Forensic Science International: Genetics*, **52**, 102473.

[24] Huszar, T. I., Gettings, K. B. a Vallone, P. M. (2021). An Introductory Overview of Open-Source and Commercial Software Options for the Analysis of Forensic Sequencing Data. *Genes*, **12**(11).

[25] I-Familia (2024). I-Familia. URL `https://www.interpol.int/How-we-work/Forensics/I-Familia`. Accessed: 28. 4. 2024.

[26] Katsanis, S. H. a Wagner, J. K. (2013). Characterization of the Standard and Recommended CODIS Markers. *Journal of Forensic Sciences*, **58** (s1), S169–S172.

[27] Lavebratt, C. a Sengul, S. (2006). Single nucleotide polymorphism (snp) allele frequency estimation in dna pools using pyrosequencing™. *Nature protocols*, **1**(6), 2573–2582.

[28] Linacre, A. a Templeton, J. E. (2014). Forensic DNA profiling: state of the art. *Research and Reports in Forensic Medical Science*, **4**, 25–36.

[29] Marano, L. A. a Fridman, C. (2019). DNA phenotyping: current application in forensic science. *Research and Reports in Forensic Medical Science*, **9**, 1–8.

[30] McDonald, C., Taylor, D. a Linacre, A. (2024). PCR in Forensic Science: A Critical Review. *Genes*, **15**(4).

[31] Molnar, C. a Gair, J. (2015). *Concepts of biology*.

[32] Moody, M. D. (1989). DNA Analysis in Forensic Science. *BioScience*, **39** (1), 31–36.

[33] Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R. a Braun, A. (2004). Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Research*, **14**, 1664–1668.

[34] Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., Sakkiah, S., Guo, W., Gong, P., Zhang, C., Ge, W., Shi, L., Tong, W. a Hong, H. (2019). Similarities and differences between variants called

with human reference genome HG19 or HG38. *BMC Bioinformatics*, **20**(2), 101.

[35] Parabon NanoLabs (2024). The Snapshot DNA Phenotyping Service. URL `https://snapshot.parabon-nanolabs.com/phenotyping`. Accessed: 28. 4. 2024.

[36] Parabon NanoLabs (2024). Parabon Snapshot™ Gives Crime Solvers a New Way to Use DNA. URL `https://parabon-nanolabs.com/news-events/2014/12/announcing-parabon-snapshot.html`. Accessed: 28. 4. 2024.

[37] Pareek, C. S., Smoczynski, R. a Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, **52**(4), 413–435.

[38] Parson, W., Steinlechner, M., Scheithauer, R. a Schneider, P. (1998). National DNA intelligence databases in Europe–report on the current situation. In *Proceedings of the 9th Symposium on Human Identification*, pages 7–10.

[39] Ringbauer, H., Huang, Y., Akbari, A., Mallick, S., Olalde, I., Patterson, N. a Reich, D. (2024). Accurate detection of identity-by-descent segments in human ancient dna. *Nature Genetics*, **56**(1), 143–151.

[40] Roewer, L. (2013). DNA fingerprinting in forensics: past, present, future. *Investigative Genetics*, **4**, 22.

[41] Saini, M. K., Gaurav, H., Kumar, J. a Sanu, K. (2023). DNA Sequencing techniques: Sanger to Next Generation Sequencing. *DNA*, **3**(09), 2378–2393.

[42] Schlötterer, C. a Pemberton, J. (1994). The use of microsatellites for genetic analysis of natural populations. In *Molecular Ecology and Evolution: Approaches and Applications*, pages 203–214.

[43] STRBase (2012). STRBase Fact Sheet – D21S11. URL `https://strbase-archive.nist.gov/str_D21S11.htm`. Accessed: 13. 4. 2024.

[44] Tagliaro, F., Manetto, G., Crivellente, F. a Smith, F. (1998). A brief introduction to capillary electrophoresis. *Forensic Science International*, **92**(2), 75–88.

[45] van der Beek, C. (2017). Past, present and future of forensic dna databases. In *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation*, pages 217–229.

[46] Van Neste, C., Van Nieuwerburgh, F., Van Hoofstat, D. a Deforce, D. (2012). Forensic STR analysis using massive parallel sequencing. *Forensic Science International: Genetics*, **6**(6), 810–818.

[47] White, J. D., Indencleef, K., Naqvi, S., Eller, R. J., Hoskens, H., Roosenboom, J., Lee, M. K., Li, J., Mohammed, J., Richmond, S., Quillen, E. E., Norton, H. L., Feingold, E., Swigut, T., Marazita, M. L., Peeters, H., Hens, G., Shaffer, J. R., Wysocka, J., Walsh, S., Weinberg, S. M., Shriver, M. D. a Claes, P. (2021). Insights into the genetic architecture of the human face. *Nature Genetics*, **53**(1), 45–53.

Zoznam obrázkov

Zoznam použitých skratiek

DNA - Deoxyribonucleic Acid (deoxyribonukleová kyselina)

SNP - Single Nucleotide Polymorphism (jednonukleotidové polymorfizmy)

STR - Short Tandem Repeats (krátke tandemové repetície)

CODIS - Combined DNA Index System

FBI - Federal Bureau of Investigation (Federálny úrad pre vyšetrovanie)

PCR - Polymerase Chain Reaction (polymerázová reťazová reakcia)

RFLP - Restriction Fragment Length Polymorphism (polymorfizmus dĺžky restrikčných fragmentov)

CE - Capillary Electrophoresis (kapilárna elektroforéza)

LCN - Low Copy Number (nízky počet kópií)

NGS - Next Generation Sequencing (sekvenovanie novej generácie)

MPS - Massive Parallel Sequencing (masívne paralelné sekvenovanie)

WGA - Whole Genome Amplification (celogenómová amplifikácia)

FDP - Forensic DNA Phenotyping (DNA fenotypovanie)

ESS - European Standard Set

KÚ PČR - Kriminalistický ústav Policie České republiky

NCBI - National Center for Biotechnology Information

CSV - Comma-separated Values

JSON - JavaScript Object Notation

XLSX - Microsoft Excel Spreadsheet

SQL - Structured Query Language

rs - Reference SNP cluster ID

VCF - Variant Call Format

A. Príloha A

A.1 Zdrojové kódy

V GitHub repozitári dostupnom pod QR kódom uvedenom nižšie alebo na adrese `https://github.com/Lujza44/str-markers-data-processing` sú dostupné zdrojové kódy potrebné pre procesovanie dát ako prevedenie formátu či vizualizácia.

Obr. A.1: QR kód vedúci ku GitHub repozitáru obsahujúcemu skripty potrebné pre prácu s dátami STR markerov