

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Nástroj pro automatickou identifikaci KIR alel

Místo této strany bude
zadání práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 6. srpna 2020

Kateřina Kratochvílová

Poděkování

Ráda bych poděkovala Ing. Lucii Houdové, Ph.D. za cenné rady, věcné připomínky, trpělivost a ochotu, kterou mi v průběhu zpracování této práce věnovala. Dále bych chtěla poděkovat panu Ing. Jiřímu Fatkovi za jeho rady a pomoc při vytváření praktické části.

Abstract

This masters thesis is focused on the identification of KIR alleles. The aim is to design and implement a tool for their automatic identification. First, the KIR genes and methods used to obtain the genomic data using DNA sequencing - next-generation sequencing (NGS) - were introduced. Secondly, the possible bioinformatic tools were analyzed. The identification tool was developed on synthetic reads and then tested and verified by commercial DNA data acquired from FN Plzeň / BC LF UK Plzeň. The creation of synthetic reads was done with a help of ART tool and referencial DNA sequence alligning was done by Bowtie2. During the development, several approaches were proposed and evaluated based on their possible application.

Abstrakt

Diplomová práce se zabývá identifikací KIR alel. Cílem práce je návrh a implementace nástroje pro jejich automatickou identifikaci. V práci jsou představeny KIR geny a metody získávání genomických dat s využitím DNA sekvenace, konkrétně next-generation sequencing (NGS). Dále byly analyzovány využitelné bioinformatické nástroje. Samotný identifikační nástroj byl vyvíjen na syntetických readech a nakonec testován a verifikován na datech komerčních linií DNA získaných z FN Plzeň/BC LF UK Plzeň. Vytváření syntetických readů probíhalo pomocí nástroje ART, pro zarovnávání readů na referenční DNA sekvenci byl využit nástroj Bowtie2. V rámci vývoje bylo navrženo několik možných přístupů, které byly poté vyhodnoceny s ohledem na jejich možné využití.

Obsah

| | | |
|----------|---|-----------|
| 1 | Úvod | 8 |
| 2 | Imunitní systém a jeho spojitost s geny | 10 |
| 2.1 | Geny | 10 |
| 2.2 | HLA a non-HLA geny | 11 |
| 2.2.1 | Alela a gen | 13 |
| 2.3 | Natural killer a jeho receptory | 13 |
| 2.3.1 | NKG2D receptor | 14 |
| 2.3.2 | KIR receptor | 15 |
| 3 | Sekvenační metody získávání DNA dat | 21 |
| 3.1 | Sanger sequencing | 21 |
| 3.2 | NGS next-generation sekvenování | 22 |
| 3.2.1 | Single-end, paired-end a mate-pair | 23 |
| 3.2.2 | 454 sekvenování a Ion Torrent | 23 |
| 3.2.3 | Illumina | 24 |
| 3.2.4 | SOLiD | 25 |
| 3.3 | Metody třetí generace | 25 |
| 4 | Analýza dostupných bioinformatických nástrojů pro zpracování NGS dat | 26 |
| 4.1 | Simulační nástroje pro generování syntetických readů | 26 |
| 4.1.1 | ART | 27 |
| 4.2 | Nástroje pro zarovnávání readů | 28 |
| 4.2.1 | Bowtie2 | 29 |
| 4.2.2 | Burrows-Wheeler transformace | 32 |
| 4.3 | Další pomocné metody | 34 |
| 4.3.1 | Levenshteinova vzdálenost | 34 |
| 5 | Implementace | 35 |
| 5.1 | Popis problému | 35 |
| 5.2 | Referenční geny | 35 |
| 5.3 | Testovací KIR genomy | 36 |
| 5.4 | Návrh systému | 37 |
| 5.4.1 | Použité programové prostředky | 38 |
| 5.5 | Modulové jednotky programu | 39 |

| | | |
|-----------|--|-----------|
| 5.5.1 | Config | 39 |
| 5.5.2 | Simulování dat | 39 |
| 5.5.3 | Zarovnání vzhledem k referenčním genům | 40 |
| 5.5.4 | Přístupy k identifikaci alel | 40 |
| 6 | Porovnání přístupů k identifikaci a parametrů | 52 |
| 7 | Verifikace na reálných datech | 53 |
| 8 | Zhodnocení z hlediska úrovně rozlišení | 57 |
| 9 | Závěr | 59 |
| 10 | Výkladový slovník pojmů a zkratk | 61 |
| | Literatura | 62 |
| A | Uživatelská dokumentace | 66 |
| A.1 | Získání potřebného softwaru | 66 |
| A.2 | Získání referenčních sekvencí | 67 |
| A.3 | Nastavení programu | 67 |
| A.4 | Doporučená adresářová struktura pro data | 69 |
| A.5 | Spuštění programu | 70 |
| A.6 | Výstupy programu | 70 |
| A.6.1 | Výsledky | 70 |
| A.7 | Analýza referenčních genů | 70 |
| A.7.1 | Spuštění | 71 |
| A.7.2 | Výstupy programu | 71 |
| A.8 | Analýzování experimentů | 73 |
| A.8.1 | Spuštění | 73 |
| A.8.2 | Výstupy programu | 73 |
| A.9 | Používané soubory | 74 |
| B | Testovací genomy | 75 |
| C | Detailní výsledky | 77 |
| C.1 | Experiment1 | 78 |
| C.2 | Experiment2 | 84 |
| C.3 | Experiment3 | 86 |
| C.4 | Výsledky verifikace | 96 |

1 Úvod

Transplantace krvetvorných buněk se využívá jako terapeutická procedura pro mnoho vážných hematologických poruch mezi které patří například akutní myeloidní leukemie. Transplantace je proces, při kterém jsou dárci odebrány krvetvorné buňky, které jsou následně vpraveny do těla pacienta trpícího hematologickou poruchou. Jednou z komplikací, která může nastat, je reakce imunitního systému na nově vložené dárcovské buňky resp. štěp. V případě, že si štěp s imunitním systémem nebudou rozumět, může dojít k silné zánětlivé reakci, která může skončit až smrtí pacienta. V neposlední řadě může dojít také k relapsu (návratu) onemocnění.

K potlačení odmítnutí dárcovského štěpu se primárně vybírají dárci podle shody v HLA znacích následovaných sekundárními znaky, kterými mohou být například věk či pohlaví. Shoda v HLA znacích se určuje podle shody v alelách genů HLA -A, -B, -C, -DRB1, -DQB1. Alela je konkrétní forma genu. Každý jedinec má tyto HLA geny dvakrát (jednu pěticí od matky a druhou od otce), a proto se úplná shoda označuje jako 10/10. Nověji je možné se setkat s označením 12/12. Znamená to 10/10 navýšené o gen HLA -DPB1. Tento gen, ale na rozdíl od standardních HLA znaků, nevyžaduje přesnou shodu. Klíčové je zde zda patří do skupiny permissivních (tolerančních) alel, které by měly snížit možnost relapsu (návrat nemoci) a rizika transplantace. Oproti tomu některé skupiny alel naopak mohou rizika zvýšit. V poslední době se objevují studie, které prokazují vliv i takzvaných non-HLA genů. Jedním z nich může být i skupina genů Killer-cell immunoglobulin-like receptor (KIR). Jednou z výhod je, že určité seskupení KIR genů snižuje riziko návratu nemoci. V případě, kdy by se rozhodovalo mezi více dárci shodných v HLA znacích, by se mohl ten vhodnější vybrat právě na základě KIR. Pro zjištění jak HLA znaků tak KIR genů se využívají sekvenační metody. [31] [20] [22]

Cílem práce je navrhnout a implementovat nástroj pro automatickou identifikaci KIR alel. Vstupní data, tzv. ready, jsou krátké kusy DNA (posloupnost písmen A, C, G a T) a jsou výstupem ze sekvenačních technik. Může to být gen, část genu nebo několik různých genů. Tyto ready se zarovnávají vůči referenční alelám. Díky tomu je možné zjistit, o které alely se jedná. Ready budou pro vývoj nástroje simulovány a v konečné fázi testování budou vyměněny za data z FN Plzeň/BC LF UK Plzeň. V poslední fázi bude

vyhodnocena shoda readů a referenčních genů.

2 Imunitní systém a jeho spojitost s geny

Imunitní systém chrání organismus před škodlivinami. Skládá se ze dvou hlavních částí, z vrozené imunity a získané imunity. Reakce imunitního systému je vždy komplexní reakcí organismu, přičemž tato reakce je tvořena jednotlivými buňkami imunitního systému reagující na přítomnost specifických antigenů. Antigeny jsou látky, které imunitní systém rozpozná a zareaguje na ně. V podstatě to může být jakákoli bílkovinná sloučenina. Antigen se obvykle nachází na povrchu buňky jako vyjádření genu. Imunitní systém následně zjistí o jaký antigen se jedná, respektive o jakou buňku se jedná, zda tělu vlastní (např. zdravá buňka) nebo buňku tělu cizí (např. nádorová buňka), tedy jedná-li se o expresi lidského genu nebo například viru. Jedná-li se o buňku tělu cizí, imunitní systém reaguje snahou ji zničit.

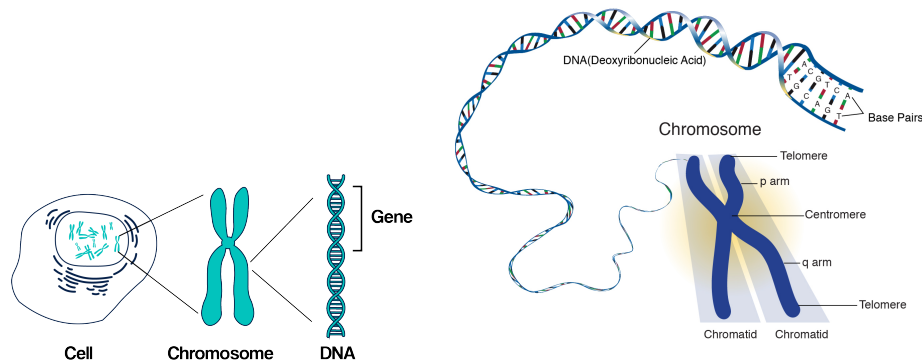
Vrozená imunita, též označována přirozená, neadaptivní, antigeně nespecifická je neměnně zapsána v DNA. To znamená, že při každém setkání s antigenem odpoví stejnou reakcí. Buňky nesoucí vrozenou imunitu jsou stále přítomny v krvi, takže jejich případná aktivace je takřka okamžitá (minuty až hodiny). Do této imunity patří i natural killer buňky s KIR receptory, kterým bude věnována pozornost dále v textu.

Získaná imunita, též označována specifická či adaptivní, má oproti specifické v genomu zapsány pouze své základy. V průběhu lidského života se vyvíjí a mění. Změna může nastat například očkováním nebo proděláním patřičné choroby. Tato změna ovšem nemusí být trvalá. Z těchto důvodů může být odpověď získané imunity při setkání se stejnou chorobou rozdílná. Fungování získané imunity zajišťují T- a B- lymfocyty. Získaná imunita nefunguje samostatně. Při zabíjení patogenů spolupracuje s vrozenou imunitou.

2.1 Geny

V každé buňce lidského organismu, konkrétně v buněčném jádře, se nachází 46 chromozomů. Jeden chromozom představuje stočenou dlouhou molekulu DNA (Deoxyribonukleovou kyselinu). Všechny 46 chromozomů obsahuje okolo 100 000 genů. Genem se rozumí drobný segment DNA, který řídí buněčnou

funkci. Konkrétní forma genu je alela. [38]



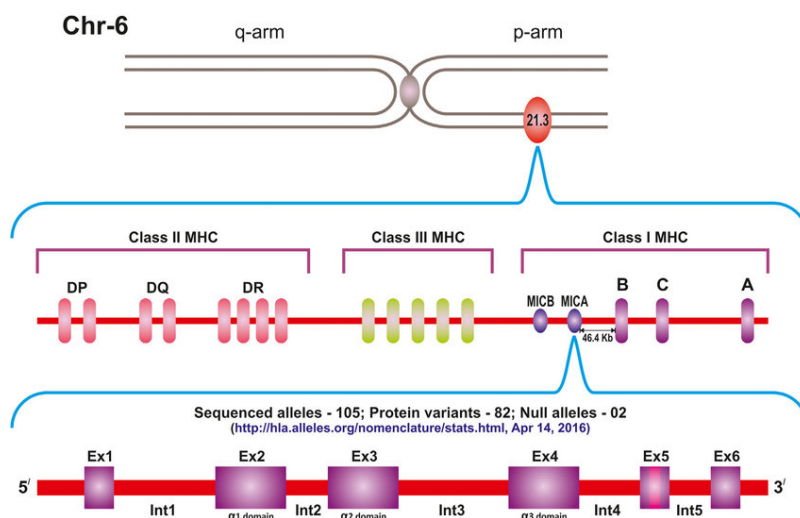
Obrázek 2.1: Převzato z [6] a [3]

Uvnitř buňky je obsažen celý genom, který se ovšem nemusí projevit na povrchu buňky. Pokud se vlastnost, kterou gen přenáší, projeví na povrchu buňky označujeme to jako exprese genu (jeho sebevyjádření). Od toho se odvíjí i konkrétní názvosloví typu gen, receptor či molekula.

2.2 HLA a non-HLA geny

Human leucocyte antigen (HLA) je genetický systém, který je primárně zodpovědný za rozeznávání vlastního od cizorodého. Někdy je termín HLA zaměňován s MHC. MHC (Major histocompatibility complex) je souhrnný termín pro všechny komplexy, kdy podskupinou jsou právě HLA (H - Human), který se vyskytuje u lidí. Stejně tak existuje DLA (D - Dog), který je pro psy. Z funkčního i biologického hlediska jde však u všech savců o stejnou skupinu genů. [31]

Přesná definice rozdílu mezi HLA a non-HLA geny neexistuje. Mimo jiné ani jejich rozdělení není v literatuře sjednocené. Jak je vidět na obrázku 2.2, je možné geny rozdělit do tří tříd. V některé literatuře je možné nalézt označení non-HLA genů jako geny III. třídy, v jiné, že jsou to všechny geny III. třídy a některé geny třídy I. Tato práce bude terminologicky vycházet v označení gen za non-HLA či HLA z definice uvedené v [36]. Zjednodušeně tedy můžeme říci, že geny, které nejsou řazeny k HLA skupinám, jsou non-HLA. Je-li gen označen za non-HLA, neznamená to, že by neměl souvislost s funkcí imunitního systému. Naopak má, jen ne výlučně s HLA systémem. Non-HLA geny kódují produkty spojené s imunitními procesy. Mezi non-HLA geny mimo jiné patří MICA, MICB a KIR. [36]



Obrázek 2.2: Šestý chromozom zobrazující HLA(-A, -B, -C, -DR, -DQ, -DP) i non-HLA(MICA, MICB) geny. Protein vzniklý expresí genu je definován exony, které definují transkripci(přepis) do RNA. Introny při translaci(překladu) nehrají roli a často jsou sekvenovány jen exony. [15]

HLA a některé non-HLA geny se nacházejí na krátkém raménku 6. chromozomu, konkrétně 6p21.3, a zauímají úsek přibližně jedné tisíciný genomu. Tento region je nejvíce komplexní a polymorfní v lidském genomu s více než 220 geny. Oproti tomu jedna ze skupin non-HLA genů, konkrétně KIR geny, se nachází na 19. chromozomu. Rozsáhlá diverzita těchto genů vznikala snahou eliminovat neustále se měnící spektrum patogenů. Produkty těchto genů na povrchu buňky významně ovlivňují odpověď na infekční choroby a výsledky buněčné či orgánové transplantace. [36]

Při určování shody dárce a pacienta se rozhoduje na základě shody alel u genů HLA -A, -B, -C, -DRB1, -DQB1. Díky velké diverzitě HLA genů je počet možných kombinací několik miliard. Některé kombinace genů se vyskytují na základě oblasti či národnosti častěji nebo mohou být naopak vzácné. HLA geny se obvykle dědí jako blok (celý haplotyp), avšak ve výjimečných případech může dojít k rekombinaci. Z tohoto důvodu je nejsnadnější nalézt shodu v pokrevním příbuzenstvu.

Jelikož každý jedinec má dvakrát geny na pozicích HLA -A, -B, -C, -DRB1 a -DQB1 (jednu pěticí od otce, druhou pěticí od matky), je maximální shoda 10/10 (shoda obou alel v lokusech) popř. DPB1 a shoda 12/12. Čím je shoda menší, tím větší je riziko nepřijetí štěpu. U nepříbuzných jedinců lze tolerovat shodu 9/10 či 8/10. [20] [31]

V posledních letech se objevuje Haploidentická transplantace, kdy je možné použít krvetvorné buňky příbuzného se shodou pouze jednoho haplotypu (5/10) například jedinec může využít transplantace od vlastního rodičů nebo sourozenců. Umožňuje to podávání chemoterapie pár dní po transplantaci, která zničí všechny buňky, které tělo nepřijme. Tímto způsobem může být celý proces transplantace výrazně urychlen, zejména v případech, kdy není čas hledat dárce v registrech [17]

2.2.1 Alela a gen

Alelu můžeme definovat jako variantu genu s nepatrným rozdílem v sekvenci nukleotidů DNA oproti jiné alele stejného genu. Jinak řečeno, alely jsou varianty genu na molekulární úrovni. Geny se zpravidla vyskytují minimálně ve dvou formách (dvou alelách). Gen určuje výskyt nějakého znaku, například oči konkrétního živočicha budou mít barvu. Alela pak určuje, jaká barva to bude. Jinak řečeno, alela zajišťuje konkrétní fenotypový projev genu.

V případě genu *KIR2DL1* mohou být jeho alely 0010101 a 0010102. Zápis genů tak, jak s nimi budeme pracovat, může vypadat způsobem zobrazeným v 2.2.1.

$$\begin{aligned} > KIR : KIR00001 KIR2DL1 * 0010101 14738 bp \\ & \quad GTTCGGGAGGTTGGATCTCAGACGTG... \end{aligned} \quad (2.2.1)$$

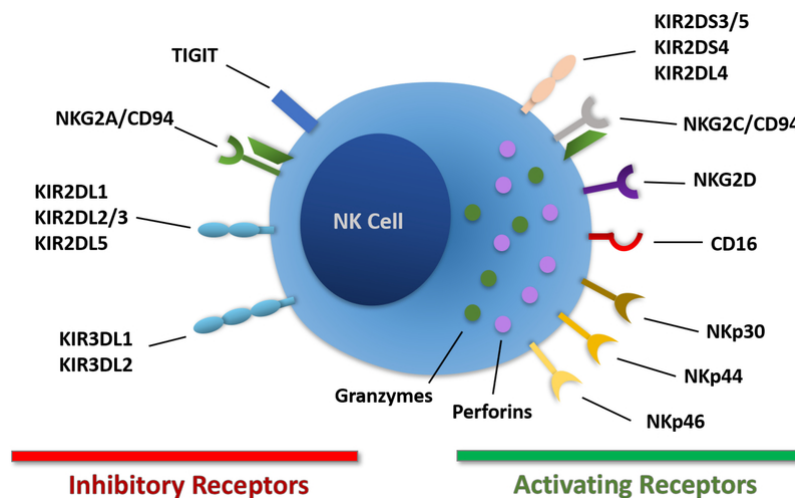
Označení *KIR* : *KIR0001* označuje pořadové číslo, kdy alela byla nalezena. Oproti tomu *KIR2DL1* * 0010101 je označení genu podle jeho vlastností.

K zjištění konkrétních alelických variant se pro tzv. typizaci využívají sekvenační metody, typicky s polymerázovou řetězovou reakcí.

2.3 Natural killer a jeho receptory

Natural killer buňky (NK buňky) jsou velké granulární lymfocyty vrozeného imunitního systému. V krevním oběhu lidského těla je jich možné nalézt 10 – 15%. Klíčovou vlastností NK buněk je nejenom schopnost rozlišit poškozené buňky od zdravých, ale také poškozené buňky rychle a efektivně likvidovat. Poškozené buňky mohou být buňky infikované virem či buňky transformované v nádorové. Na povrchu NK buňky se nachází receptory, které

jsou zobrazeny na obrázku 2.3, regulující odpověď imunitního systému. Natural killer buňky oproti B- a T- lymfocytům (buňkám získané imunity) nemají antigenně specifické receptory. Jedním ze způsobů jak NK buňky rozpoznávají a zabíjejí poškozené buňky, je na základě interakce mezi KIR receptorem a HLA molekulou na povrchu zkoumané buňky (podrobněji viz sekce KIR). Stejně tak mohou zabíjet na základě receptoru NKG2D, který aktivuje cytotoxickou reakci při setkání s ligandem MICA a MICB. Ligandem označujeme malou molekulu, která se váže na vazebné místo cílového proteinu(receptoru) a vyvolává fyziologickou odpověď, která může mít inhibiční či aktivační charakter.



Obrázek 2.3: Natural killer buňka a její receptory, rozděleny na aktivační a inhibiční.[16]

2.3.1 NKG2D receptor

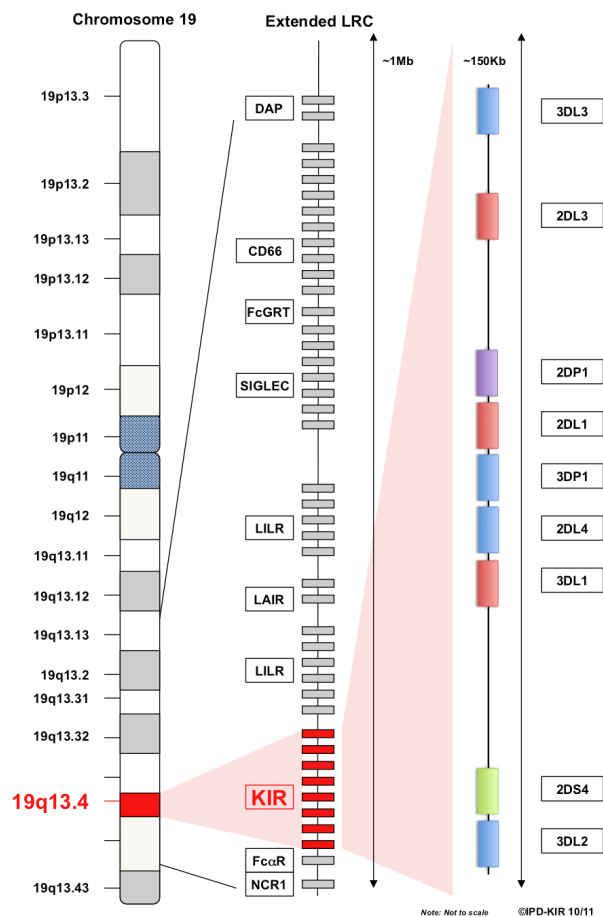
NKG2D je jeden z nejvýznamnějších aktivačních receptorů na NK buňce rozpoznávající především buněčný stres, který může spustit cytotoxicitu (schopnost ničit buňky), i když se na povrchu buňky nachází inhibiční HLA-I ligandy.

Geny skupiny MICA a MICB jsou označovány jako class I chain-related gene. To znamená, že se běžně neřadí do I. třídy MHC. Takto označované geny mají souvislost s MHC I. třídy, ale na rozdíl od nich neváží peptidy. Oproti HLA genům, které mají svoje produkty na lymfocytech, se produkty MICA a MICB nachází na epitelových buňkách. Nejedná se tedy o standardní HLA geny, proto jsou v novějších zdrojích označovány jako non-HLA. Jejich expresí na povrch buňky jsou ligandy, které se váží na receptor NKG2D. Buňky

s ligandy MICA a MICB se množí při nádorovém onemocnění, zanětu nebo pod vlivem různých forem buněčného stresu, a díky navázání na receptor může být spuštěna imunitní reakce. [34] [21] [16] [36]

2.3.2 KIR receptor

Killer immunoglobulin-like receptor (KIR) je skupina genů řazených mezi non-HLA geny. Jejich zvláštností je fakt, že se nenachází na 6. chromozomu, ale na 19. a tak se dárce se schodnými HLA znaky nemusí schodovat v KIR znacích. Expresí KIR genů jsou receptory na povrchu natural killer buněk. Dnes je známo 15 genů a 2 pseudogeny rozlišující se na inhibiční a aktivační na základě cytoplasmatického ocásku a počtu imunoglobulinových domén. [31]



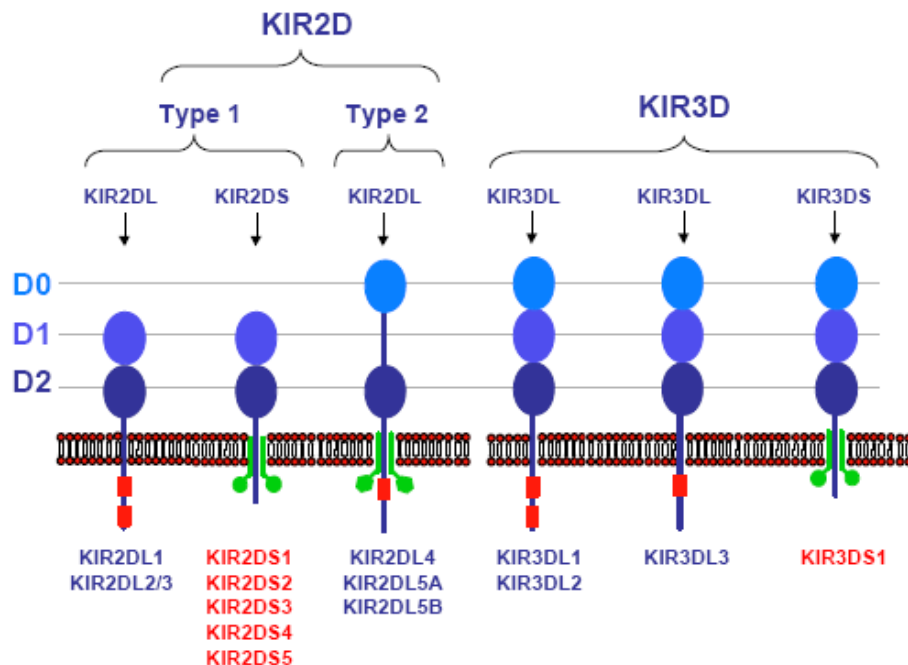
Obrázek 2.4: KIR se nachází na 19. chromozomu v oblasti jménem leukocyte receptor complex (LRC). [36]

Nomenklatura KIR genů

KIR geny (na obrázku 2.5) se liší různou délkou cytoplasmatického ocásku (tail) a různým počtem imunoglobulin-like domén (Ig-like). Na základě této rozmanitosti byla založena nomenklatura KIR genů, tedy jejich pojmenování.

Jak je vidět na obrázku 2.5, cytoplasmatický ocásek může být dlouhý (long - L) nebo krátký (short - S). Je možné se setkat i s označením P, které slouží pro pseudogeny. Oproti tomu imunoglobulinové domény se mohou vyskytovat 2 (2D) nebo 3 (3D). Právě z těchto vlastností vychází základ pojmenování KIR genů.

Příkladem může být KIR2DL1*010101, kde 2D označuje dvě imunoglobulinové domény, L značí dlouhý ocásek, 1 značí že je to první 2DL protein. Numerická definice alely je poté oddělena hvězdičkou. První tři čísla označují alely, které se liší v sekvencích jejich kódovaných proteinů, další dvě číslice se používají k rozlišení alel, které se liší synonymními rozdíly v kódující sekvenci. Konečné dvě cifry rozlišují alely na základě substituce v intronu, promotoru nebo jiné nekódující oblasti. [36]

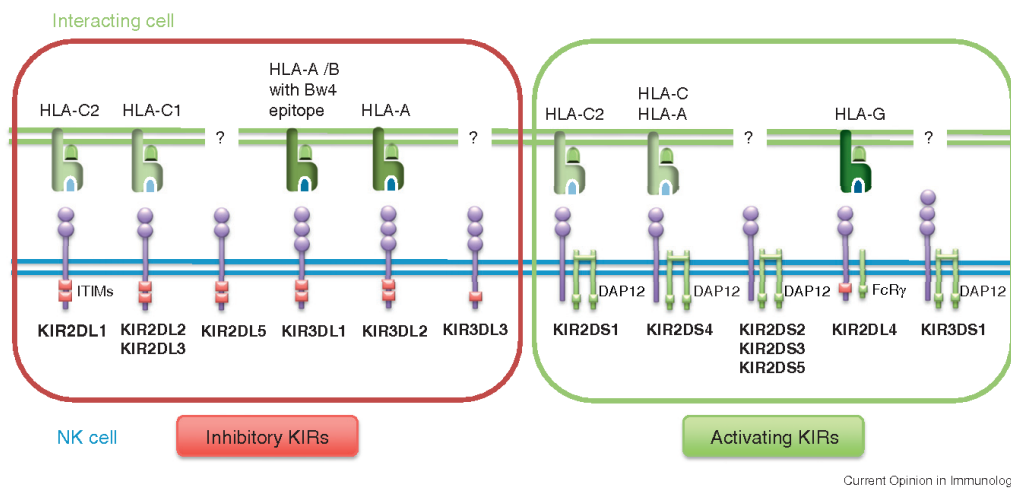


Obrázek 2.5: Nomenklatura KIR genů. [31]

Další rozdělení KIR genů je na již výše zmíněné inhibiční a aktivační. Dle obrázků 2.3 a 2.5 je možné si povšimnout detailu, že až na KIR2DL4 jsou aktivační KIR s krátkým ocáskem, zatímco inhibiční jsou s dlouhým ocáskem.

Aktivace NK buněk pomocí KIR

Jak již bylo výše zmíněno, KIR receptory můžeme rozdělit na inhibiční a aktivační. O tom, zda dojde k aktivaci NK buňky, rozhoduje právě jejich rovnováha na zkoumané buňce. Obrázek 2.6 ukazuje vazebné ligandy pro jednotlivé KIR receptory.

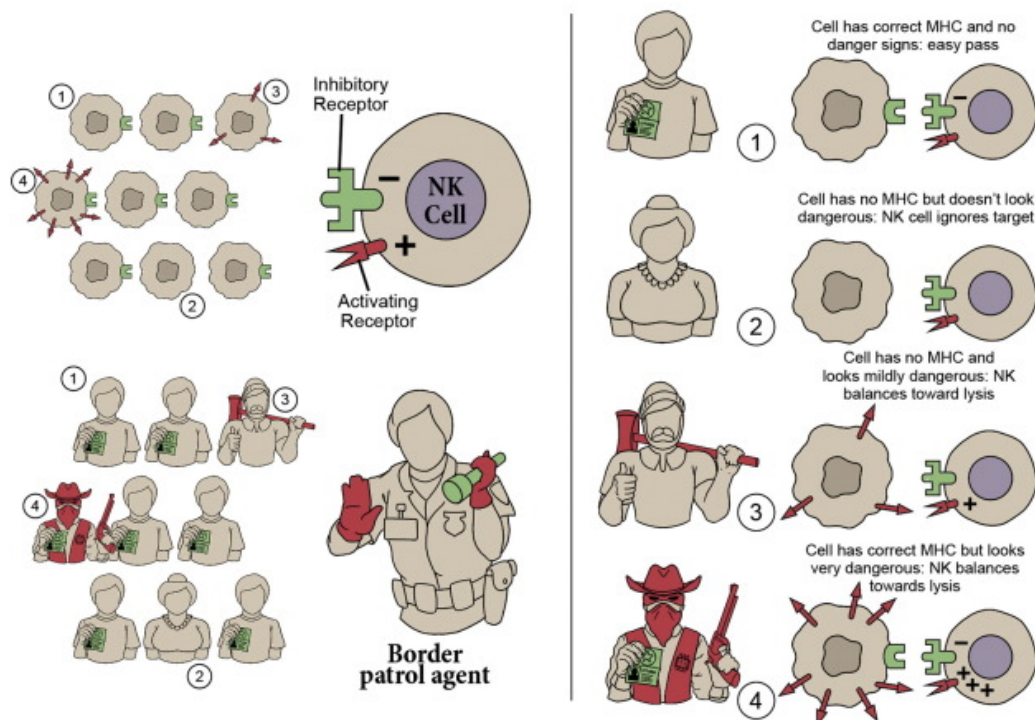


Obrázek 2.6: KIR geny a jejich vazebné ligandy. Pokud je v obrázku ?, značí to, že pro daný receptor není znám vazebný ligand. [39]

NK buňky ustavičně prohledávají své okolí a testují přítomnost příslušných HLA ligand pro své KIR receptory. Pokud je příslušný HLA ligand přítomen, naváže se na NK buňku (2.7 případ 1). Tímto systémem jsou ochráněny vlastní buňky. Pokud přítomen není, je spuštěna cytotoxická reakce a zkoumaná buňka je zničena.

Některé virem napadené buňky potlačují propsání HLA ligandu na povrch buňky a tím se brání cytotoxicitě vůči T lymfocytům. Zároveň se ale stávají citlivějšími na cytotoxicitu proti NK buňkám, jak je zobrazeno na obrázku 2.7 případ 3.

The NK Cell is like a border patrol agent

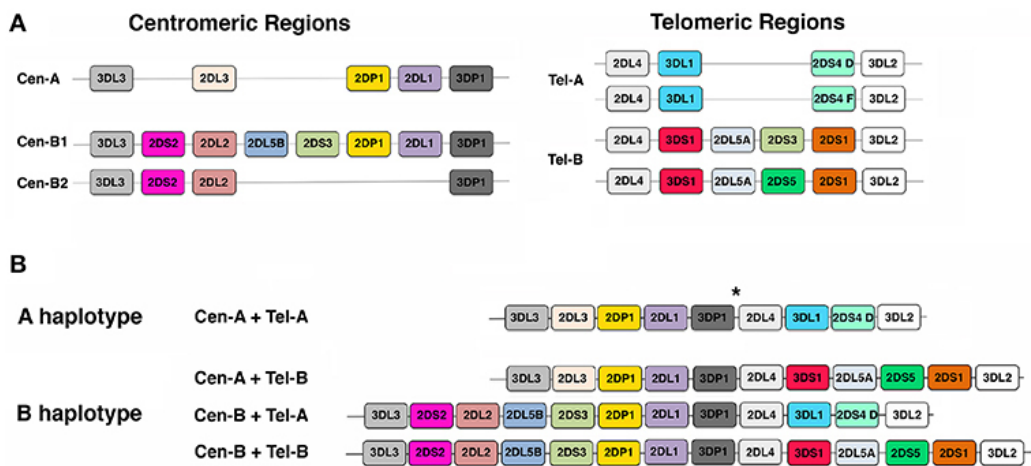


Obrázek 2.7: Přirovnání fungování natural killer buňky k pasové kontrole. V pravé části jsou zobrazené případy, které mohou nastat, když natural killer buňka potká jinou buňku. V 1. případě je tělu vlastní zdravá buňka, kde se KIR receptor naváže na HLA ligand a k cytotoxické reakci nedojde. Druhým případem je červená krvinka. K reakci NK buňky opět nedojde, protože na zkoumané buňce nepřevažují aktivační receptory. V 3. případě je to nádorová buňka, která schová HLA ligand, (může nastat po transplantaci kostní dřeně), a tím se "schová" proti T- lymfocytům. Avšak aktivační receptory převažují, a tak k cytotoxicitě dojde. Ve 4. příkladu je nádorová buňka nebo virem nakažená buňka (stresové ligandy). Aktivační receptory převažují, a tudíž k cytotoxické reakci dojde.[37]

KIR genotyp a haplotyp

KIR genotyp je vyjádření, jaké konkrétní KIR geny genom obsahuje. Genotyp je možné rozdělit na dvě části, takzvané haplotypy. Jeden haplotyp je od otce, druhý je od matky. Na základě kombinací všech genů je možné vytvořit velký počet KIR genotypů. Proto byl díky shromážděným haplotypům sestaven model, který toto množství mírně redukuje, přičemž samozřejmě nepokrývá všechny možné varianty. Haplotyp se rozděluje na dvě části, centromerickou a telomerickou v závislosti na tom, zda je blíže k centroměře či

telomeře (viz. obrázek 2.1 pravá část). Jednotlivé části mezi sebou mohou být kombinovány. Centromerická i telomerická část může být zařazena do jedné ze dvou skupin A či B na základě genů, které obsahuje (viz. obrázek 2.8 část A). Celý haplotyp je následně přiřazen do jedné ze dvou skupin podle kombinace centromerické a telomerické části. V případě, kdy jsou obě části A/A, je haplotyp označen za A, u ostatních kombinací (A/B, B/A, B/B) je haplotyp B (viz. obrázek 2.8 část B). Jiná definice pro rozdělení haplotypů uvádí, že skupina B musí obsahovat alespoň jeden z genů KIR2DL5, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS5 a KIR3DS1. Naopak skupina A neobsahuje ani jeden z těchto genů. Je třeba si zde uvědomit, že každý jedinec má 2 KIR haplotypy. [33]



Obrázek 2.8: Rozdělení KIR genů na centromerickou a telomerickou část, pojmenování je na základě toho, zda je úsek blíže k centromeru nebo k telomeru (viz obrázek 2.1). Je možné si zde povšimnout, že je možné některé geny najít jak v centromerické části, tak v telomerické části. Upraveno z [33]

Podle některých studií zabývajících se vlivem KIR haplotypů na výsledky transplantace bylo zjištěno, že KIR haplotypy ovlivňují výsledky u akutní myeloidní leukemie či lymfoblastické leukemie. Ve srovnání s haplotypem A měl haplotyp B, především jeho centromerická část, ochranný účinek před návratem nemoci a zároveň zvýšil pravděpodobnost přežití pacienta. Na základě této skutečnosti se mohou dárci řadit do tří skupin: best, better a neutral. Rozřazení do třídy se vyhodnocuje jako počet B a jejich umístění v centromerické oblasti či telomerické oblasti. Mimo jiné je možné se setkat s pojmem B-skóre. Toto číslo udává počet B, které se v daném haplotypu

nachází. Best je definován B-skórem alespoň 2, přičemž dvě B se musejí nacházet v centromerické oblasti Cen-B/B a Tel-x/x. Better je definován B-skórem alespoň 2, aby nebyl haplotyp zařazen do best, musí být logicky alespoň jedna z centromerických oblastí A - Cen-A/x a Tel-B/x. Neutral je v případě jedné B části nebo žádné. [18]

KIR geny se stejně jako HLA dědí jako celý blok. Jelikož HLA se nachází na 6. chromozomu a KIR na 19., tak shodní dárce v HLA znacích se jen menšinou shodují v KIR genech. V případě příbuzného dárce shodujícího se v HLA znacích je pouze 25 % shodných také v KIR. [18]

| Hapl Group | Genotype ID ¹ | 3DL1 | 2DL1 | 2DL3 | 2DS4 | 2DL2 | 2DL5 | 3DS1 | 2DS1 | 2DS2 | 2DS3 | 2DS5 | 2DL4 | 3DL2 | 3DL3 | 2DP1 | 3DP1 | Populations | Individuals |
|------------|--------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------------|-------------|
| AA | 1 | | | | | | | | | | | | | | | | | 190 | 7,540 |
| Bx | 2 | | | | | | | | | | | | | | | | | 178 | 2,522 |
| Bx | 4 | | | | | | | | | | | | | | | | | 178 | 2,096 |
| Bx | 3 | | | | | | | | | | | | | | | | | 167 | 1,157 |
| Bx | 5 | | | | | | | | | | | | | | | | | 161 | 1,536 |
| Bx | 6 | | | | | | | | | | | | | | | | | 155 | 899 |
| Bx | 7 | | | | | | | | | | | | | | | | | 134 | 583 |
| Bx | 8 | | | | | | | | | | | | | | | | | 130 | 635 |
| Bx | 9 | | | | | | | | | | | | | | | | | 120 | 395 |
| Bx | 71 | | | | | | | | | | | | | | | | | 112 | 443 |

Obrázek 2.9: Deset nejčastějších KIR haplotypů. Šedý obdélník značí přítomnost genu, bílý jeho nepřítomnost. [8]

3 Sekvenační metody získávání DNA dat

Po pojmem sekvence DNA se skrývá posloupnost písmen představujících primární strukturu reálné nebo hypotetické molekuly či vlákna DNA, které nese nějakou informaci. Jednotlivá písmena jsou označována jako nukleotidy nebo nukleové báze. Nukleové báze mohou být A - adenin, C - cytosin, G - guanin a T - thymín. [5]

Příkladem může být následující úsek sekvence na základě obrázku 2.1

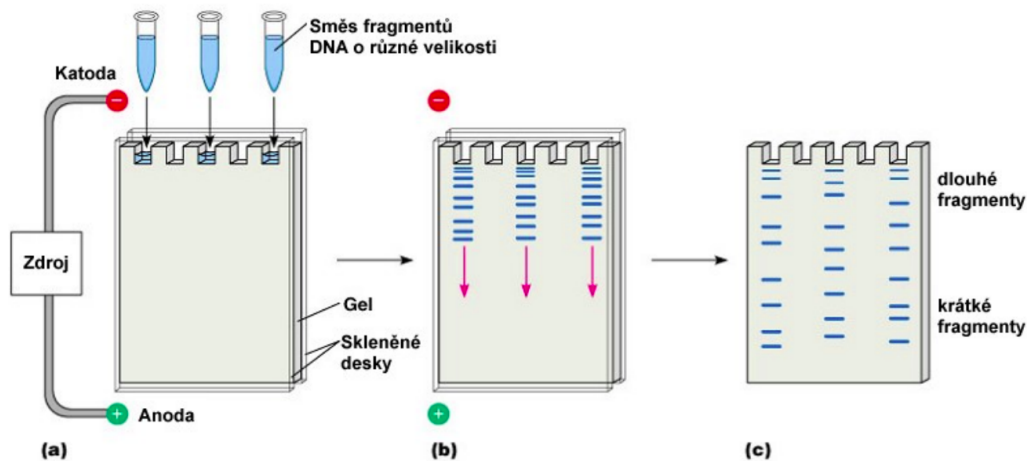
$$ACGTCA \quad (3.0.1)$$

Sekvenování DNA, někdy pouze sekvenování, jsou biochemické metody, kterými se zjišťuje pořadí nukleotidů (A, C, G, T) v sekvenci DNA. Díky tomu je možné zjistit typizaci konkrétního člověka. Sekvenační metody se liší zejména délkou řetězce, kterou dokáží zpracovat, cenou a rychlostí sekvenace. Pro porovnání, sekvenování celého genomu Sangerovo metodou by stálo několik milionů dolarů a trvalo zhruba 10 let. Při použití dnešních metod by cena byla zhruba tisíc dolarů. Většina sekvenačních metod využívá vlastnosti přitahování báze do páru pouze jednou konkrétní bází. To znamená že se adenin vždy páruje s thymínem a cytosin se vždy páruje s guaninem. Z těchto párů vzniká již známá dvojitá šroubovice DNA. V praxi není neobvyklé, že se sekvenuje jen konkrétní kus DNA, který je zrovna výzkumně či jinak potřebný. Největším problémem u sekvenování je, že úseky DNA vzniklé ze sekvenátoru (označovány jako ready) jsou jen kousky, které je třeba poskládat zpět. K tomu slouží zarovnávání. [25]

3.1 Sanger sequencing

Sanger sekvenování využívá možnosti namnožení řetězce díky vzájemnému přitahování konkrétních bází. V prvním kroku replikace jsou nastříhané řetězce rozděleny na dvě vlákna. Lze si představit, že tato dvě oddělená vlákna jsou umístěna do směsi, kde plavou jednotlivé nukleotidy spolu s upravenými nukleotidy, které nesou specifickou fluorescenční barvu a na které není možné nic navázat. Následně za pomoci střídání teploty volně plující nukleotidy tvoří postupně páry s řetězcem, který chceme namnožit. Pokud se

povede celý řetězec namnožit, je odtržen a může se dále množit. Postupně ale bude docházet k navazování nukleotidů s fluorescenční barvou. Tím se vytvoří několik různě dlouhých sekvencí zakončených označeným nukleotidem. Podle jeho barvy je možné poznat o jaký nukleotid se jedná. Následně jsou za pomoci elektroforézy seřazeny v gelu podle délky. Elektroforéza rozděluje různě dlouhé sekvence na základě odlišnosti pohybu v elektrickém poli. Kratší doputují dále než delší. Pomocí Sanger metody je možné sekvenovat řetěče dlouhé až 1000 bází.



Obrázek 3.1: Elektroforéza. [32]

3.2 NGS next-generation sekvenování

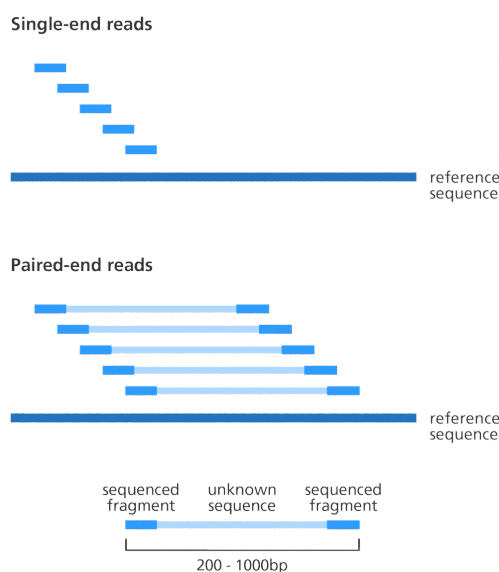
Next-generation sekvenování, někdy označováno jako metody druhé generace, je v porovnání se Sanger sekvenováním rychlejší a levnější, na druhou stranu, ale dokáže zpracovávat jen řetězce dlouhé 100 až 500 bází, má menší přesnost a častěji chybí. Jeho rychlost spočívá především ve schopnosti detekovat přidávání bází jednu po druhé a zároveň sekvenovat tisíce až miliony rozdílných molekul DNA najednou.

Všechny tyto metody si předpřipraví řetězce nastříháním DNA na krátké části a připevněním takzvaného adaptéru na jejich konec. Adaptér je krátká molekula DNA, která slouží k uchycení sekvenovaného úseku na pevný povrch. Řetězce DNA jsou namnoženy, díky čemuž vznikají klastry identických molekul koncentrovaných v jednom místě. Díky tomu je posílen signál, který by z pouhé jedné molekuly nebyl dostatečně silný. Tento signál je zachycen kamerou. Jeden z důvodů popularity NGS metod jsou i cenově dostupné

stolní sekvenátory.

3.2.1 Single-end, paired-end a mate-pair

Single-end je sekvenování pouze jednoho konce molekuly. Nevýhoda tohoto způsobu se projevuje především na krátkých readech, kde se zvyšuje problém jejich správného umístění. Oproti tomu v případě paired-end se sekvenuje z obou konců daného úseku. Vzniklé dva ready jsou označeny a zároveň je známá jejich vzdálenost, která se pohybuje od 200 do 400 bp (base pair). Mate-pair je v podstatě paired-end s tím rozdílem, že je mezi ready větší vzdálenosti od 2 do 5 kb (kilobase) - takže přibližně 2000 - 5000 bp. [7]

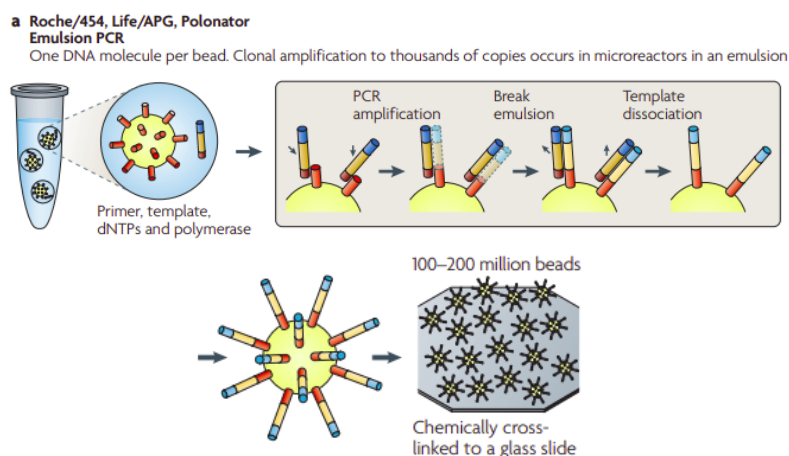


Obrázek 3.2: Single-end a paired-end read. [14]

3.2.2 454 sekvenování a Ion Torrent

Pomocí 454 sekvenování je možné analyzovat více než milion molekul DNA najednou a délka každé jednotlivé sekvence se pohybuje okolo 700 až 1000 bází. V prvním kroku sekvenování je fragment DNA přichycen na malou "kuličku" na jejíž povrchu se postupně namnoží až kuličku zcela pokryjí identické fragmenty DNA. Následuje vložení kuličky i s DNA do jedné z milionů komůrek na destičce s reakční směsí. Postup je znázorněn na obrázku 3.3. V určitém momentu je do této směsi přidán vždy jen jeden typ báze. Mezi jednotlivými fázemi přidávání určité báze jsou přebytečné nukleotidy z předešlého kroku odstraněny. To znamená, že v reakční směsi je vždy jen jeden typ nukleotidů. Během vložení každé nové báze do rostoucího řetězce

DNA je uvolněna molekula zvaná pyrofosfát, která spustí několik chemických reakcí. V poslední fázi enzym luciferáza vydá světelný záblesk, který je možné zachytit citlivou kamerou. Tento postup se nazývá pyrosekvenování. V případě, kdy je do řetězce přidáno několik stejných bází za sebou, například gen obsahuje podřetězec AAA, je vyzářeno, v tomto případě třikrát více světla než v případě jedné přiřazené báze. Kamera snímá celou destičku a na základě toho, která komůrka se rozsvítí, pozná, kde proběhlo přidání báze. Intenzita světla pak určuje kolik bází bylo přidáno najednou.



Obrázek 3.3: 454 sekvenování. [30]

Sekvenování Ion Torrent funguje na podobném principu sekvenování s tím rozdílem, že místo světla se měří změna pH v reakční směsi. Podle intenzity změny pH lze pak poznat kolik nukleotidů bylo přidáno do rostoucího řetězce.

Hlavní slabinou těchto dvou metod je značná chybovost při přidání mnoha stejných nukleotidů do řetězce za sebou. Například při přidání 10 A, nebude možné jednoznačně určit, zda se jedná o 10 A nebo 9 A.

3.2.3 Illumina

Při sekvenování pomocí Illumina jsou páry dvoušrobovice rozděleny na dva řetězce. Jednotlivé řetězce jsou následně přichyceny na malou destičku pomocí adaptéru. Každý řetězec se následně opakovaně množí až na destičce vznikne několik shluků. Přidání jedné molekuly ke druhé probíhá obdobně jako u Sanger sekvenování. Každý shluk tvoří jednu skupinu vzájemně identických řetězců. Mezi volné nukleotidy jsou opět zahrnuty nukleotidy označené fluorescenční barvou, za které nelze nic navázat. Oproti Sangerovu sek-

venování je ale tato blokace vratná a po přečtení citlivou kamerou dojde k odstranění blokující části molekuly. Počítač si pak následně zpětně spočítá, o který nukleotid se jednalo (na základě barvy). [7] [25]

3.2.4 SOLiD

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) se spoléhá na enzym ligáza. Enzym je bílkovina, která určuje rychlost chemických reakcí. Enzym ligáza konkrétně umožňuje připojení jednořetězcových molekul k stávajícím řetězcům. K řetězcům jsou přidávány takzvané sondy, což jsou kousky DNA. Sondy začínají všemi možnými dvojkombinacemi čtyř základních nukleotidů. V součtu je 16 sond. Na každé sondě je jedna ze čtyř fluorescenčních barev. V jednotlivých krocích jsou sondy připojeny k rostoucímu řetězci. Následně je přečtena fluorescenční barva, která je pak odstraněna, a může se tak navázat další sonda. Z výsledného signálu lze pak odvodit sekvenci DNA.

3.3 Metody třetí generace

Velkým rozdílem oproti druhé generaci je, že DNA řetězec není před sekvenováním namnožen a je čten pouze z jedné původní molekuly. Existuje například PacBio od Pacific Bioscience, který k detekci využívá fluorescenčně značené nukleotidy. Díky jeho vysoké citlivosti je možné v reálném čase zachytit přidání i jediného nukleotidu do jediného řetězce DNA. Dalším zástupcem těchto metod je Oxford Nanopore, jehož výhodou je jeho velikost. Oxford využívá odlišného tvaru bází. Obě metody jsou schopné přečíst přes 10 tisíc bází v rámci jedné analyzované molekuly DNA.

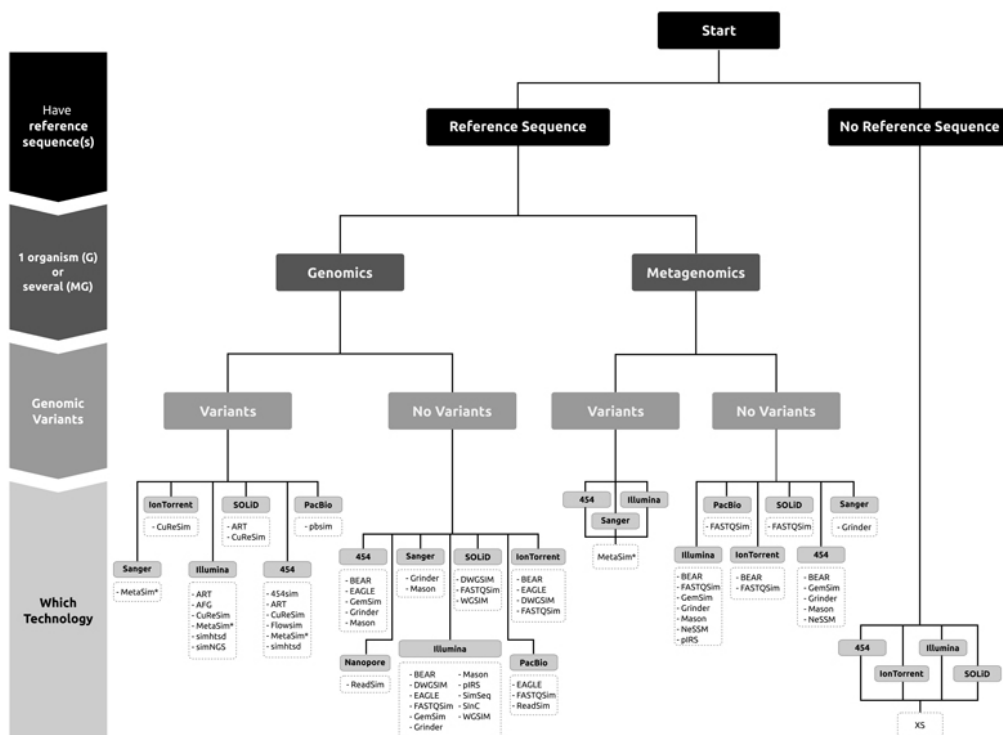
4 Analýza dostupných bioinformatických nástrojů pro zpracování NGS dat

NGS metody snižují náklady a zrychlují proces sekvenování za cenu kratších readů a menší přesnosti, což vedlo k mnoha bioinformatickým výzvám jako je vytvoření nástrojů pro analýzu readů. Nástroje je možné mezi sebou porovnat pomocí reálných nebo simulovaných dat. Přestože je validace na reálných datech nezbytná, skutečné hodnoty na kterých jsou data založena jsou obvykle neznámé, což komplikuje jejich použití pro posouzení přesnosti (tj. jak blízko je odhadovaná hodnota ke skutečné hodnotě). Díky tomu je simulování dat čím dál více populární pro hodnocení, validaci či nastavování optimálních parametrů nástroje. [19]

4.1 Simulační nástroje pro generování syntetických readů

Dále uvedené nástroje byly vybrány na základě těchto parametrů: simulování DNA, udržitelnost a volná dostupnost. Následující informace vychází z článku [19], pokud nebude uvedeno jinak.

Většina simulátorů NGS vyžaduje referenční sekvenci, ze které budou generovat ready. Tato referenční sekvence může být konkrétní genomická oblast, více zřetězených genomických oblastí, chromozom či celý genom. Některé simulátory vytváří zarovnání readů přímo do referenčního souboru (soubory SAM/BAM). Při používání simulátorů může být pro uživatele obtížné se rozhodnout, kterou konkrétní hodnotu pro daný parametr určit nebo který vlastní profil vytvořit, proto některé simulátory poskytují výchozí profily. Jedním z nich může být i generování chyb či modelu kvality. Nástroje jako jsou ART nebo SInC generují tyto profily na základě extrahovaných modelů ze skutečných dat. Nejčastější chyby jsou způsobeny substitucí, vložním či smazáním (INDEL - insert-deletion).



Obrázek 4.1: Strom pro usnadnění výběru generátoru syntetických readů. [19]

Data získaná z FN Plzeň/BC LF UK Plzeň jsou primárně sekvenována nástrojem Illumina, proto je podle [19] na výběr z ART, AFG, CuReSim, Flowsim, MetaSim, simhtsd a simNGS. U simulátoru AFG je třeba chybové profily definovat ručně, CuReSim a MetaSim nejsou open source, simhtsd podporuje jen operační systém Linux, simNGS podporuje jen operační systémy Linux a MacOS. Proto byl vybrán simulační nástroj ART který podporuje operační systémy Linux, Windows a MacOS. Tento nástroj dále generuje podle chybových profilů a profilů kvality, které byly vytvořeny pomocí extrakce chyb získaných ze skutečných dat.

4.1.1 ART

ART (next-generation sequencing read simulator) je sada simulačních nástrojů, které generují syntetické ready, jako kdyby byly získány sekvenováním pomocí NGS. Nástroj ART dokáže simulovat single-end a paired-end ready ze sekvenátorů Illumina, 454 společnosti Roch a SOLid od společnosti Applied biosystem. Ready vytvořené nástrojem ART, jsou používány pro testování a analýzu nástrojů zpracovávajících právě NGS sekvence, jako například nástroje Bowtie (zarovnávání). Při použití nástroje ART je vstupním

souborem sekvence genů, na základě kterých jsou vygenerovány ready. [23]

Illumina je sekvenování založené na vratném umístění báze označené barvou do rostoucího řetězce. Jeho nejčastější chybou je substituce. Pravděpodobnost chyby substituce je určena na základě skóre kvality dané báze, které je závislé na pozici v rostoucím řetězci. Průměrné skóre kvality klesá v závislosti na zvyšování pozice báze. ART simuluje substituční chybu na základě zvyšování kvality báze a modelu pravděpodobnosti chyby získaného z reálných datasetů. INDEL chyba je simulována jen na základě modelu z reálných dat a u Illuminy se vyskytuje jen zřídka. Pro paired-end simulaci využívá ART dvou rozdílných skóre kvality pro každý pár readů jiný.

ART je implementován v jazyce C++ a je dostupný s licencí GPL verze 3 pro operační systémy Linux, MacOS a Windows. Je možné ho použít i jako C++ package. Pro jeho spuštění je nutné mít nainstalovaný compiler GNU g++ 4.0 nebo vyšší a knihovnu GNU gsl. Výstupy se čtou ve formátu FASTQ a zarovnání ve formátu ALN. ART může generovat zarovnávání také ve formátu SAM nebo UCS BED. Paired end ready jsou označeny stejným názvem souboru s 1 či 2 na jeho konci.

4.2 Nástroje pro zarovnávání readů

Zarovnávání bývá prvním krokem v mnoha genomických pipelinech. Často je to jejich nejpomalejší část, protože pro každý read musí zarovnávač vyřešit obtížný výpočetní problém. Určit pravděpodobné umístění v referenčním genomu. V současnosti je na výběr více než 90 nástrojů pro zarovnávání NGS readů. Nástroje jsou mezi sebou obvykle porovnávány na základě přesnosti a rychlosti mapování. V článku [10] bylo porováno 5 nástrojů pro zarovnávání DNA readů. Nástroj STAR měl narozdíl od ostatních nástrojů menší přesnost, nástroj segemehl byl zase náročný na paměť (podle článku až 70 GB), což se na stolním počítači těžko dosahuje. Ze zbývajících nástrojů byl vybrán nástroj Bowtie2 díky jeho rychlosti v porovnání s nástrojem BWA, která vzhledem k množství požadovaných zarovnání bude přínosem.

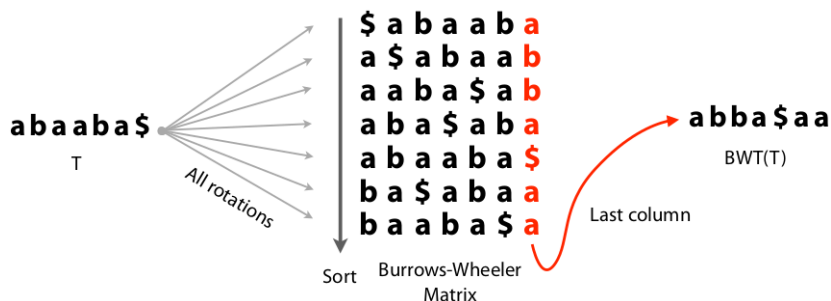
Mnoho zarovnávačů používá indexy k rychlému snižování kandidátů pro umístění zarovnávaného readu. Bowtie permanentní indexy referenčních genů a lze je tak použít napříč běhy. Algoritmus FM indexu obvykle funguje na vyhledání přesně shody. V případě hledání umístění readů na referenční gen není toto řešení použitelné, protože ready mohou obsahovat chyby vzniklé sekvenováním, případně genové mutace. Proto Bowtie každé zarovnání zakládá na kvalitě znaku báze v daném readu. Bowtie postupně vytváří dlouhý sufix. Pokud se sufix nevyskytuje v textu, pak se může algoritmus vrátit a v již vytvořeném sufixu nahradit bázi za jinou. Dále pokračuje obdobným způsobem. Tento způsob změny báze je dále označován jako backtracking. Pokud by měl algoritmus na výběr substituovat za více bází, vybere tu s nejnižší kvalitou znaku v readu. Protože algoritmus Bowtie v základu bere první přijatelné řešení, je možné, že jeho nalezené řešení není to nejlepší. Pro nalezení toho nejlepšího řešení je třeba použít přepínač – *best*, jehož funkčnost je ale na úkor rychlosti, která může být 2x či 3x nižší. Bohužel nemůže být tento přepínač použit u paired-end readů. Zároveň je možné nastavit maximální počet nahrazených bází v readu. [28]

V případě, že backtracking mechanismus není úspěšný, může docházet k jeho nadměrnému výskytu. Bowtie2 se tento jev snaží zmírnit dvojitým indexováním. První index obsahuje BWT genomu a je označován jako dopředný index. Druhý obsahuje opět BWT genomu, ale se znaky v sekvenci v opačném pořadí, označován jako zrcadlový index. Read je pak v půlce rozdělen na dvě části a jejich zarovnávání probíhá odděleně tak, že je vždy backtracking povolen jen v dané části, která je zrovna zarovnávaná. Pravá část je zarovnávaná podle dopředného indexu a levá část je zarovnávaná podle zrcadlového indexu. Předchozí algoritmus funguje dobře pouze v případě, kdy reference nebo read neobsahují mezery (báze chybí nebo naopak přebývá). Proto byl algoritmus rozšířen, jak je popsáno dále.

4.2.2 Burrows-Wheeler transformace

Burrows-Wheelerova transformace (BWT) je reverzibilní permutace řetězců v textu. Původně byla používána pro kompresi dat. Indexace založená na BWT umožňuje efektivní vyhledávání ve velkém textu s malou pamětovou náročností.

BW transformace řetězce T , $BWT(T)$, je zobrazena na obrázku 4.4. Znak $\$$ je připojen na konec řetězce a zároveň musí platit, že se tento znak v řetězci nevyskytuje. Burrows-Wheeler matice řetězce T je konstruovaná jako všechny cyklické rotace řetězce T , které byly seřazeny podle abecedy, znak $\$$ je řazen na začátek abecedy. Výstup, $BWT(T)$ pak představuje poslední sloupec matice. Tento řetězec má stejnou délku jako původní řetězec T . [28]



Obrázek 4.4: Burrows-Wheeler transformace řetězce T . [26]

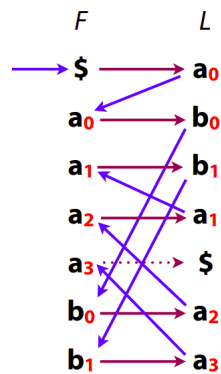
Burrows-Wheeler matice má vlastnost, která se nazývá last first mapping (LF). To znamená, že i -tý výskyt znaku X v prvním sloupci je i -tý výskyt znaku X v posledním sloupci. V případě přidání indexu do řetězce T je toto pravidlo pro znak a zobrazeno na obrázku 4.5. Obdobně to platí i pro ostatní znaky v řetězci.

$$T = a_0 b_0 a_1 a_2 b_1 a_3 \$ \tag{4.2.1}$$

| F | L |
|-------|---------------------------|
| $\$$ | $a_0 b_0 a_1 a_2 b_1 a_3$ |
| a_3 | $\$ a_0 b_0 a_1 a_2 b_1$ |
| a_1 | $a_2 b_1 a_3 \$ a_0 b_0$ |
| a_2 | $b_1 a_3 \$ a_0 b_0 a_1$ |
| a_0 | $b_0 a_1 a_2 b_1 a_3 \$$ |
| b_1 | $a_3 \$ a_0 b_0 a_1 a_2$ |
| b_0 | $a_1 a_2 b_1 a_3 \$ a_0$ |

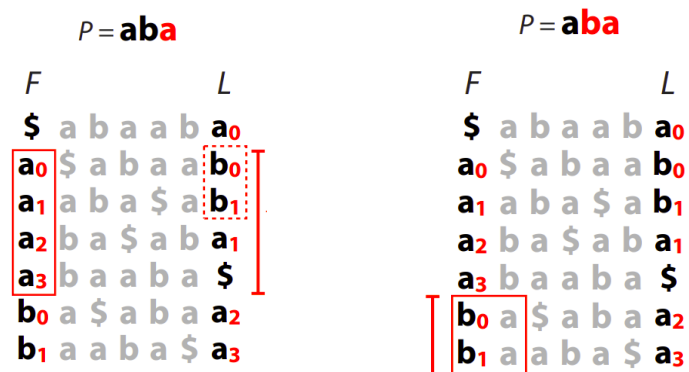
Obrázek 4.5: Burrows-Wheeler transformace last first mapping (LF). [26]

Zpětné získání řetězce je znázorněno na obrázku 4.6. L sloupec je řetězec, který je výstupem BW transformace. F sloupec, je na základě L sloupce, snadné odvodit. Jelikož platí pravidlo, že počet jednotlivých znaků je stejný, stačí je pouze přemístit do F sloupce a seřadit podle abecedy. Dále s využitím LF je řetězec získán zpět. Jako první se vezme přidávaný znak \$. Ve stejném řádku ve sloupci L se nachází a_0 . To znamená, že řetězec začíná \$. Algoritmus pokračuje s a_0 v F sloupci. Ve stejném řádku v L sloupci je b_0 . b_0 je přidáno do řetězce a pokračuje až do doby, než je opět na řadě znak \$.



Obrázek 4.6: Burrows-Wheeler transformace zpětné získání původního řetězce. [26]

Díky vztahu mezi F a L sloupcem je možné vyhledávat daný řetězec (zobrazeno na obrázku 4.7). Například vyhledávaný řetězec bude $P = aba$. Při pohledu do F sloupce jsou nalezeny všechny sloupce začínající a , následně v L sloupci ve stejných řádcích jsou nalezeny dva výskyty b . Již je získán sufix ba , který existuje. Pokračuje se dále na řádky, které začínají právě nalezenými b . V sloupci L pro dané řádky je nalezena a . Řetězec $P = aba$ se v textu vyskytuje.



Obrázek 4.7: FM index - získání prefixu. [26]

4.3 Další pomocné metody

4.3.1 Levenshteinova vzdálenost

Levenshteinova vzdálenost zjišťuje rozdílnost dvou textů na základě počtu změn, které je třeba udělat, aby byl z jednoho řetězce získán druhý řetzec. Za úpravy se považuje vložení, smazání a nahrazení. Algoritmus funguje tak, že se snaží ze slova, které bylo předáno jako první v argumentu vytvořit slovo předané jako druhý argument. Příkladem může být vzdálenost mezi řetězcí *SPAM* a *PARK*. Vzdálenost těchto slov je 3. Výstup v případě python knihovny je možné vidět následovně. Výstup 4.3.1 je v případě *SPAM*, *PARK*. Výstup 4.3.2 je v případě *PARK*, *SPAM*. Změny jsou definovány: o jakou změnu jde, index znaku v prvním řetězci a index znaku v druhém řetězci. Je možné si všimnout závislosti mezi těmito dvěma postupy. [41]

$$\begin{aligned} & ('delete', 0, 0), ('insert', 3, 2), ('replace', 3, 3) & (4.3.1) \\ SPAM - & \rightarrow _PAM - & \rightarrow _PARM - & \rightarrow PARK \end{aligned}$$

$$\begin{aligned} & ('insert', 0, 0), ('delete', 2, 3), ('replace', 3, 3) & (4.3.2) \\ PARK - & \rightarrow SPARK - & \rightarrow SPAR_ - & \rightarrow SPAM \end{aligned}$$

5 Implementace

5.1 Popis problému

Hlavním úkolem navrhnutého nástroje je určit alely, které představují ready ze sekvenátoru. Jednou z výzev je, že alela může být dlouhá až téměř 16 000 bp. Získané ready jsou ale dlouhé jen 250 bp a navíc mohou obsahovat chyby (záměna bází, chybějící báze nebo báze navíc).

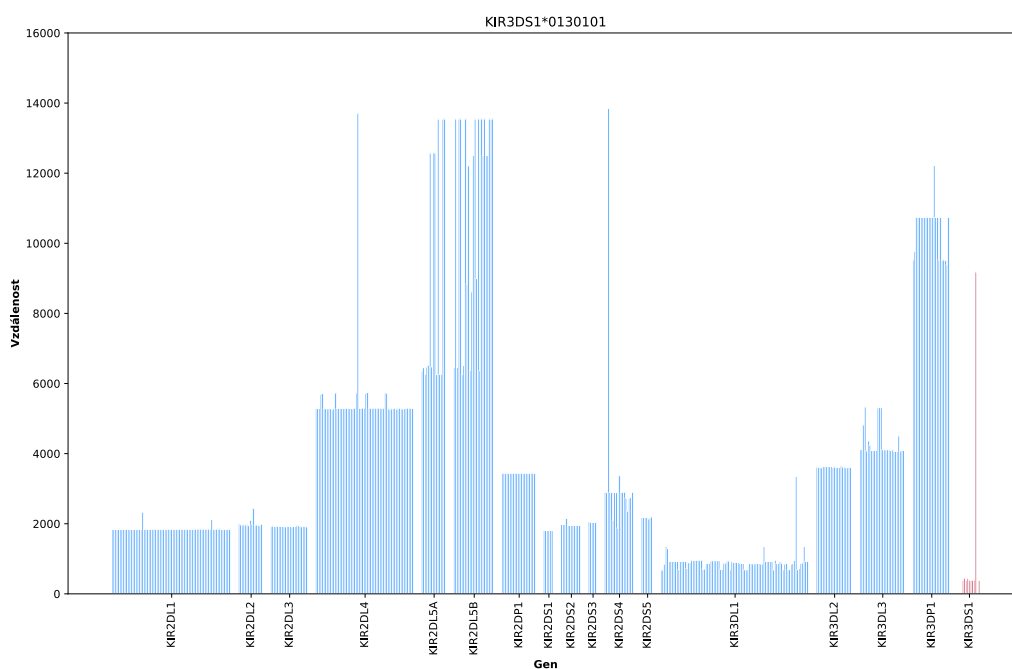
5.2 Referenční geny

Referenční geny byly převzaty z IPD-KIR [36], konkrétně soubory ve formátu *fasta* uloženy ve stejnojmenné složce. Jednotlivé soubory jsou pojmenovány genem, který obsahují, např. *KIR2DL1_gen.fasta*. Každý soubor představuje všechny dostupné alely konkrétního genu. Jedinou výjimku tvoří soubory *KIR_gen.fasta* či *KIR_nuc.fasta*, které obsahují všechny geny, a navíc i pseudogeny.

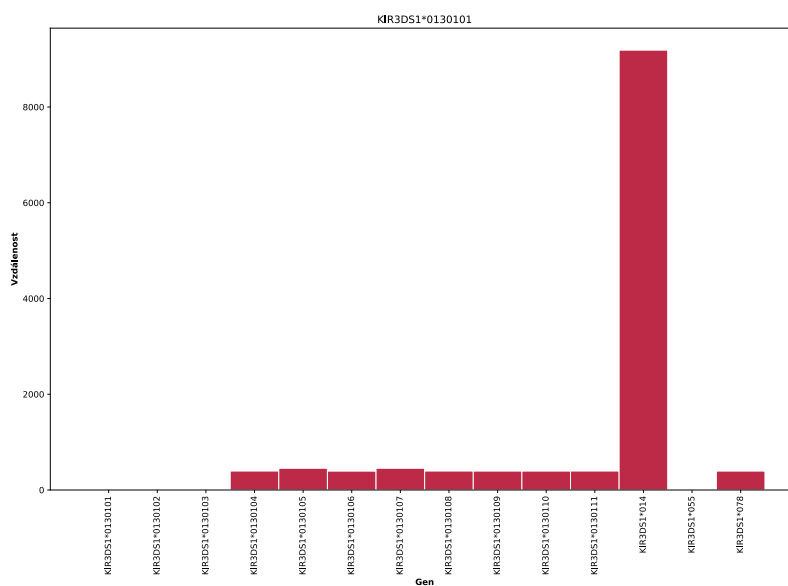
Kromě souborů **_gen.fasta* obsahuje složka *fasta* také soubory **_prot.fast* a **_nuc.fasta*. Soubor *_gen.fasta* obsahuje informace o celých genech. Oproti tomu *_nuc.fasta* obsahuje nukleotidy, tedy pouze exony bez intronů. Soubor **_prot.fast* obsahuje sekvence proteinů, které vznikly z RNA. Data získaná z nemocnice budou odpovídat alelám uvedených v *_gen.fasta*. V práci bude jako reference použit soubor *_gen*.

Při analýze porovnávání souboru *nuc* a *gen* bylo zjištěno, že v souboru *nuc* je více alel než v souboru *gen*. Konkrétně v souboru *gen* je 461 alel a v souboru *nuc* je 1109 alel. Tento údaj mimo jiné dokazuje, že nejsou přístupné referenční sekvence ke všem reálným alelám. Nejmenší Levenshteinova vzdálenost mezi alelami je 1, největší 15 943 a průměrná 4768.98.

Na následujícím obrázku je možné vidět všechny vzdálenosti ostatních alel ve skupinkách podle genů vzhledem k alele 3DS1*0130101. Je možné si povšimnout, že nejbližší má k alelám ze stejného genu.



Následující obrázek zobrazuje vzdálenosti pouze v rámci genu 3DS1 opět k alele 3DS1*0130101.



5.3 Testovací KIR genomy

Genomy, na kterých byl nástroj testován, byly dodány vedoucí práce a vychází ze známých genotypů KIR - Genotype Reference List uvedených v [8]. Genomy test1 - test11 byly sestaveny podle definovaných genotypů a umý-

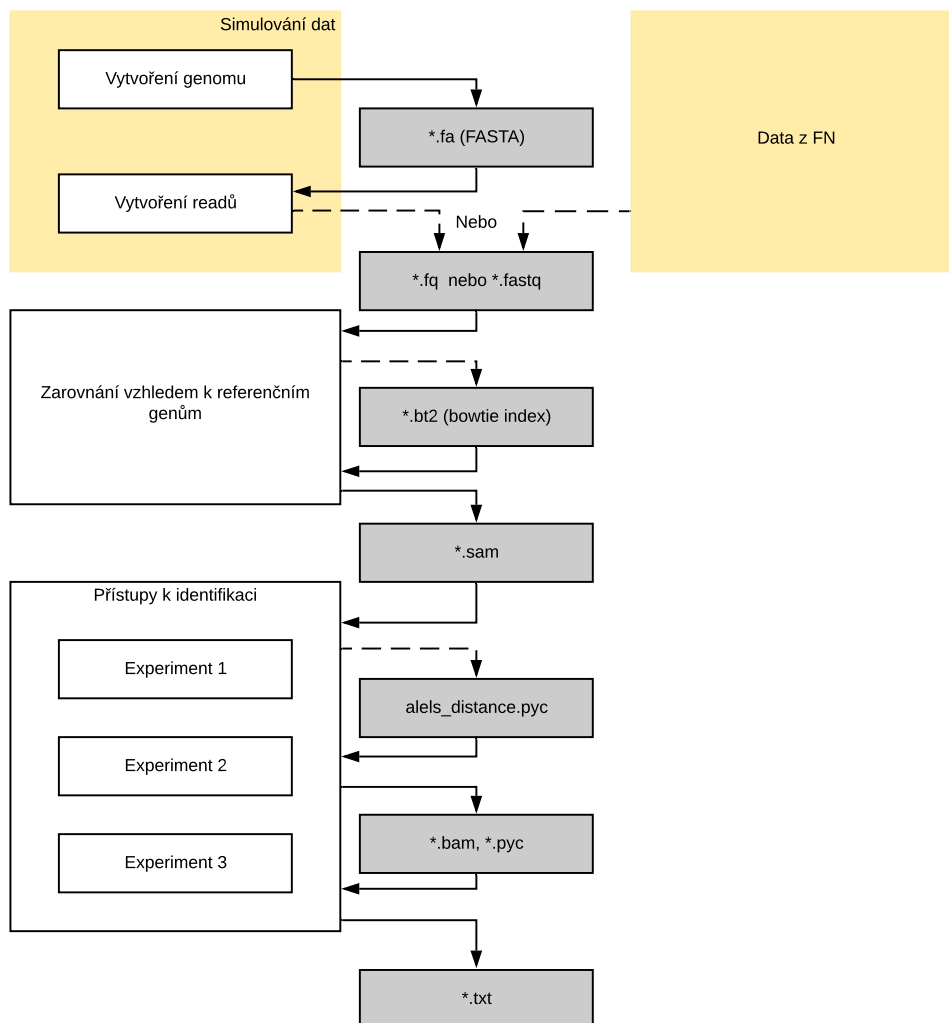
slně vybrány, ty které představují určitým způsobem výzvu pro navrhnutý nástroj. Přesný obsah testovacích genomů je uveden v příloze.

5.4 Návrh systému

Systém byl navržen jako modulární, díky čemuž je možná jednoduchá náhrada jakékoliv jeho části.

Vše začíná získáním readů, pro něž má být vyhodnoceno, jaké KIR alely představují. První možností je použít data přímo z FN Plzeň/BC LF UK Plzeň. Tyto data představují ready, na kterých bude prováděna verifikace nástroje. Druhou možností je ready simulovat. Na těchto datech byl nástroj vyvíjen a laděn. Genom může být vyroben manuálně nebo ho lze vyrobit za pomoci skriptu. V dalším kroku musí být vytvořený genom "rozbit" do podoby, jako by vyšel ze sekvenátoru. Rozdělí se na ready a vytvoří se v něm chyby, aby data co nejvíce odpovídala reálným. To se provádí za pomoci nástroje ART.

V následující části jsou získaná data, tedy ready, zarovnána na referenční genom pomocí nástroje Bowtie2. Nakonec je zarovnání vyhodnoceno a rozpoznáno, o jaké alely genů se pravděpodobně jedná. Vyhodnocení je rozděleno do několika experimentů.



Obrázek 5.1: Návrh systému.

5.4.1 Použité programové prostředky

Program byl navržen a implementován na operačním systému Linux za použití programovacího jazyka Python. Tento jazyk byl zvolen z důvodu jednoduchého použití, díky čemuž nebyl vývoj prototypů časově náročný. Pro spuštění programu je nutné mít nainstalovaný Python ve verzi 3.8, nástroj ART ve verzi MountRainier, Bowtie 2 ve verzi 2.4.1. Dále byly použity Python knihovny pysam ve verzi 0.14, python-Levenshtein, Matplotlib a Numpy.

Pysam je modul do Pythonu, který pomáhá číst a manipulovat s výstupními

soubory ze zarovnávačů. Python-Levenshtein je modul do pythonu který obsahuje funkce na výpočet Levenhsteinovy vzdálenosti a zároveň dokáže vrátit pozice rozdílných znaků v řetězci. Matplotlib je rozsáhlá knihovna pro vytváření animací a interaktivních vizualizací v Pythonu. V práci byla použita pro vykreslování grafů.

5.5 Modulové jednotky programu

Vše potřebné pro samotný běh programu obstarává skript *run.py* spolu s nastavením v souboru *config.py*. Skript *run.py* postupně pošle jednotlivé moduly. Díky nastavení v *config.py* je možné si zvolit spuštění jen některých modulů. Například pouhé vytvoření testovacích dat nebo pouze jejich zarovnání, a v neposlední řadě spuštění vyhodnocování zarovnávaných dat.

5.5.1 Config

S konfiguračním souborem *config.py* jsou spojeny všechny skripty a obsahuje jejich veškerá nastavení. Jak již bylo zmíněno, je možné pomocí tohoto nastavení spustit jednotlivé moduly. Jedná se především o položky *CREATE_READS*, *ALIGN* a *IDENTIFY*. *CREATE_READS* udržuje informaci o spuštění vytvoření syntetických readů. *ALIGN* se stará o spuštění zarovnání. V neposlední řadě *IDENTIFY*, která řídí spuštění vyhodnocení zarovnání. V případě nastavené hodnoty *IDENTIFY* na *True* je možné řídit spuštění experimentů (položky *EXP1*, *EXP2* a *EXP3*). Důležitou položkou v configu jsou cesty ke zdrojovým a výstupním složkám. Dalším nastavením je obsah genomů při případném vygenerování testovacích dat.

5.5.2 Simulování dat

O simulování dat se stará skript *create_syntetic_reads.py*, přičemž toto simulování je rozděleno na dva kroky: vytvoření genomů a vytvoření readů. Mezi těmito dvěma fázemi vznikají soubory s příponou *.fa*. Každý tento soubor obsahuje právě jeden KIR genom. Tyto genomy jsou následně použity jako vstupní soubor pro vytvoření readů, které se provádí za pomoci nástroje ART. Pro vytvoření genomů je volán skript *create_genome.py*, který vytvoří genomy na základě nastavení v configu pod položkou *GENOMES* a za pomoci referenčních genů v souboru pod položkou v configu *REFERENCE_KIR_GENS_FILE*. Vytvoření probíhá obdobně jako je popsáno níže u manuálního vytvoření testovacího genomu. Výsledné genomy jsou

uloženy do adresáře z configu *GENOME_FOLDER*. Výstupem modulu *create_syntetic_reads* je soubor s příponou *.fq*, který by měl odpovídat formátu reálných dat, které byly dodány. Výstupní soubory jsou uloženy do složky pod proměnnou *READS_FOLDER*.

Manuální vytvoření testovacího genomu lze provést následujícím způsobem. V prvním kroku je v referenčním souboru vybrána konkrétní alela genu. Někdy je možné najít takovou shodu, kdy se alely liší jen v konečné části jejich označení a v genomu je pouze prvních 5 čísel. V tomto případě může být vložena jakákoli z těchto alel. Vkládaná alela musí být vložena včetně její hlavičky tedy: *>KIR:KIR00138 KIR3DL3*0040201 12390 bp*.

5.5.3 Zarovnání vzhledem k referenčním genům

Zarovnávání obstarává skript *alignment_reads_to_reference.py* s pomocí nástroje Bowtie2. V souboru *config.py* je nutné vyplnit položku *BOWTIE_HOME_DIRECTORY* podle umístění nástroje Bowtie na konkrétním počítači. V prvním kroku jsou vytvořeny Bowtie2 indexy, které je možné použít napříč běhy, proto je v nastavení položka *BOWTIE_BUIL_INDEX*, díky které je možné toto vytvoření povolit nebo zakázat. Bez vytvořených indexů, tedy indexů z minulých běhů, ale Bowtie2 nebude zarovnávat. Bowtie2 vytváří indexy na základě obsahu souboru *REFERENCE_KIR_GENS_FILE*. V dalším kroku jsou načteny všechny ready ze složky v *READS_FOLDER* a následně jsou zarovnány nástrojem Bowtie2. Tady je nutné, aby byly ready paired-end, tedy aby se vyskytovaly dvakrát, přičemž jednou s *1* na konci a podruhé s *2*. Tento předpoklad zajistí správné nastavení ARTU, který je takto nastaven. Výstupní soubory jsou ve formátu *.sam* a jsou umístěny dle položky *ALIGNMENT_FOLDER* v nastavení.

5.5.4 Přístupy k identifikaci alel

Detailní výsledky pro všechny genomy je možné najít v příloze. V následující teoretické části jsou uvedeny vždy simulované genomy.

Experiment 1

Jako první pokus o určení alel, které by mohly být obsaženy v genomu, bylo pouhé oříznutí alel, které by měly větší procentuální šířku zarovnání (například 90 %). V případě buněčné linie AMALA, která byla simulována

nástrojem ART a která obsahuje 19 alel, je výsledek zobrazen níže. Červeně jsou označeny alely, které se v genomu skutečně nacházejí a číslo v závorce udává jejich procentuální pokrytí. Pravděpodobným důvodem proč mají alely, které do genomu nepatří tak vysoké zarovnání, je jejich podobnost. Například alely 2DS4*0010101 a 2DS4*0010102 jejíž vzdálenost je 2. Z toho lze logicky usoudit, že na alelu 2DS4*0010102 se zarovnal některé z readů alely 2DS4*0010101.

- 3DL2*0070102 (99.64%)
- 2DS4*0010101 (99.64%)
- 2DS4*0010102 (99.55%)
- 2DS5*0020101 (99.50%)
- 3DL2*0020105 (99.48%)
- 2DS2*0010101 (99.43%)
- 3DL3*00802 (99.34%)
- 3DL2*0070103 (99.29%)
- 3DL3*0040201 (99.22%)
- 3DL2*0020106 (99.15%)
- 3DP1*007 (99.04%)
- 2DS4*0010107 (98.99%)
- 2DS5*0020103 (98.98%)
- 3DS1*055 (98.70%)
- 3DL2*0020102 (98.40%)
- 3DL2*0020104 (98.19%)
- 3DL2*018 (98.07%)
- 2DL2*0030101 (98.07%)
- 2DP1*0020105 (97.99%)

Následně jsou uvedeny alely, které do genomu patří, mají pokrytí více než 90 %, ale nejsou v prvních 19.

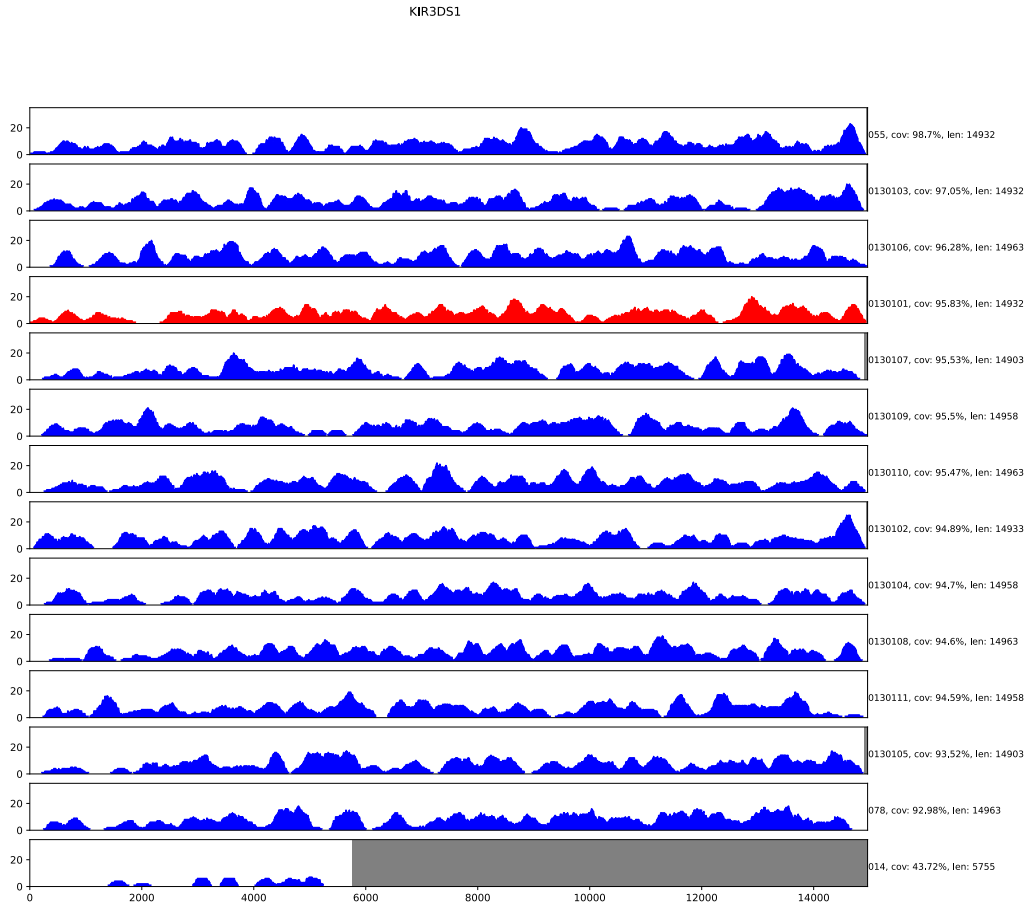
- 2DL2*0030102 (97.29%)
- 2DL3*0010109 (94.16%)
- 2DL5A*00102 (96.6%)
- 2DS1*0020106 (96.57%)
- 3DP1*0090101 (96.46%)
- 3DS1*130101 (95.83%)

Nakonec jsou uvedené alely, které do genomu patří, ale mají pokrytí menší než 90 %. Po tomto zjištění se nabízí pokusit se snížit požadované zarovnání na 70 % a přidat další upřesňující kroky.

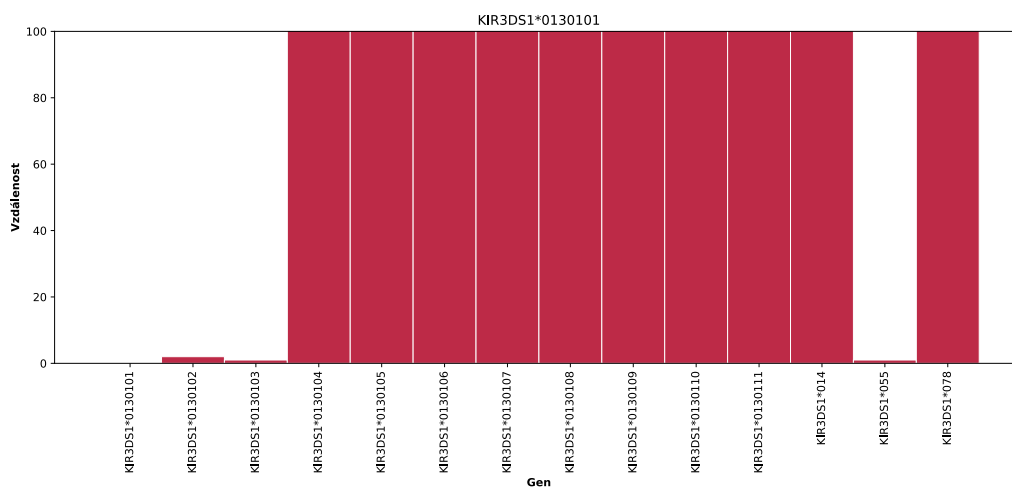
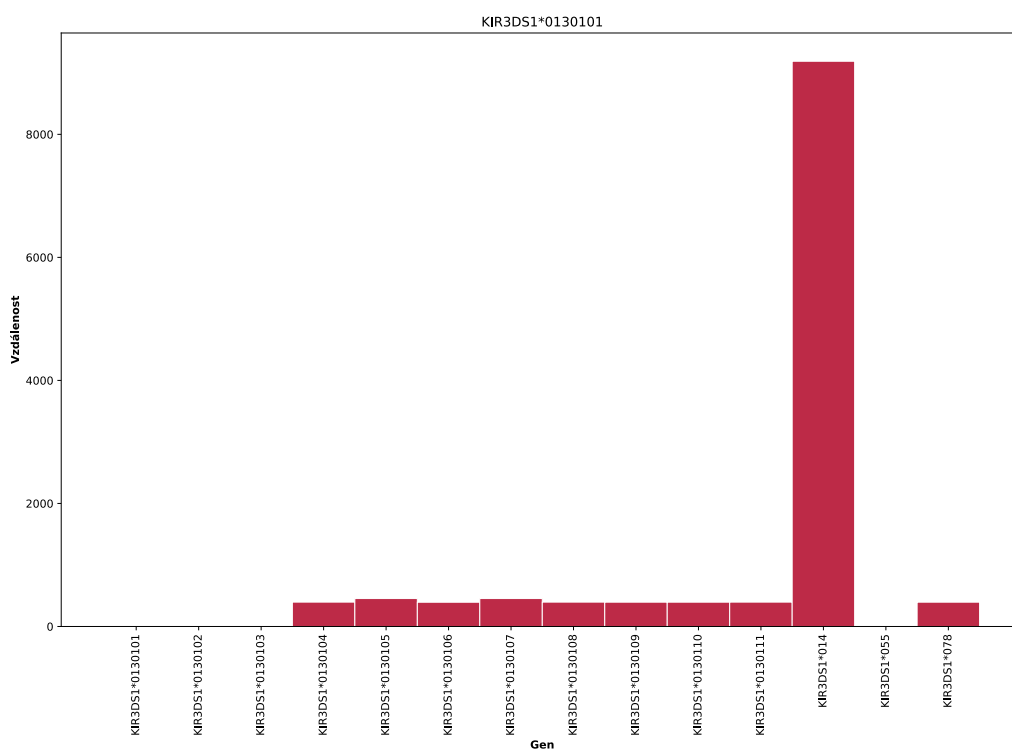
- 2DL1*0030201 (81.8%)
- 3DL1*0150201 (70.71%)

Ná následujícím obrázku jsou zobrazeny všechny alely genu KIR3DS1 a jejich zarovnání. Červeně je označena alela, která do genomu AMALA patří. Je možné si povšimnout, že alela s největším pokrytím do genomu nepatří. Navíc pokrytí větší než 90 % mají všechny alely až na jednu. Neméně důležité je, že poslední alela má značně kratší délku než všechny ostatní alely stejného genu. Poslední věcí je hloubka pokrytí, která by díky nastavení sekvenátoru měla odpovídat hladině 100 u alely, která do genu patří. Jak je vidět

z obrázku, hloubka pokrytí se u většiny alel pohybuje v extrémech kolem 20. Z toho lze odvodit, že ready alely 3DS1*0130101 byly pravděpodobně rozděleny mezi všechny alely tohoto genu.



Všechny výše uvedené analýzy vybízí k bližšímu prozkoumání podobnosti alel. Níže jsou uvedeny grafy Levenshteinovy vzdálenosti ostatních alel stejného genu vzhledem k alele 3DS1*0130101, tedy té, která je obsažena v genomu.



Z předchozích obrázků je patrné, že 3 alely jsou vzdálené v rámci jednotek, 9 alel je vzdálených kolem 500 a jen jedna alela je vzdálená více, což pravděpodobně způsobuje to, že je kratší. Přichází tedy na řadu otázka, jak určit, které alely jsou si až moc podobné, a pak rozhodnout, kterou z nich odstranit. Vzdálenost, kdy jsou si alely blízké je možné nastavit v configu pod položkou *CLOSE_DISTANCE*. Dále je porovnáváno jejich pokrytí v místech, kde jsou rozdílné. Pokud je hloubka pokrytí jedné alely 2krát větší než hloubka pokrytí druhé alely, je druhá alela odstřižena.

V tabulce níže je uvedeno porovnání výsledků požadovaného pokrytí 90 % a 70 % u genomu AMALA. Za blízké byly alely považovány v případě vzdálenosti menší než 100. Genom AMALA obsahuje celkem 19 alel a žádná z nich se neopakuje. Krok 1 je pouze statistika prvního zarovnání. Obsahuje všechny alely, které jsou v referenci. Celkově je v referenci 461 alel. Pouze dva geny ze všech genů uvedených v referenční sekvenci se nenacházejí v tomto genomu. Krok 2 je po odstrižení alel, které mají pokrytí menší než 90 %, případně 70 %. V posledním kroku 3 jsou uvedené výsledky po odstrižení podobných.

| | Krok 1 Zbývá alel | Zbývá alel | Krok 2 Ztraceno | Geny navíc | Zbývá alel | Krok 3 Ztraceno | Geny navíc |
|----------------------------|-------------------------|---------------|--------------------------------|---------------|---------------|--------------------------------|---------------|
| amala pokrytí 90 % | 461 | 113 | 2 2DL1*0030201 3DL1*0150201 | 0 - | 23 | 4 3DP1*0090101 2DL4*0010201 | 0 - |
| amala pokrytí % 70 % | 461 | 193 | 0 - | 0 - | 41 | 2 2DL4*0010201 3DP1*0090101 | 0 - |

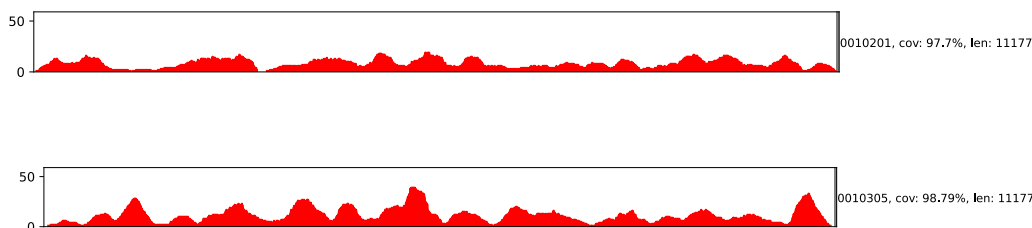
Tabulka 5.1: Výsledky experimentu 1 u genomu AMALA. Odřezány byly alely, které měly pokrytí menší než 90 %. *Alel* u genomu značí počet v daném genomu. Číslo v závorkách udává, kolik alel je v daném genomu dvakrát. V každém kroku *zbývá alel* je kolik alel ještě zůstalo ve výběru, *ztraceno* určuje kolik alel má být v genomu, ale algoritmus je vyřadil. Za tímto číslem jsou vypsané alely, které byly ztraceny. V dalších krocích jsou vypsané alely bez těch, které už byly ztraceny v předchozích krocích. Obdobně je to s *geny navíc*, které udávají počet a druh genů již neobsahujících žádnou z alel, která náleží do daného genomu.

Ve výsledků v příloze jistě zaujme genom Test7, který ztratil 7 alel v rámci třetího kroku. Tedy při odřezávání podobných alel.

| Zbývá alel | Krok 1 | | Geny navíc | Zbývá alel | Krok 2 | | Geny navíc | Zbývá alel | Krok 3 | | Geny navíc | | |
|------------|----------|---|--|------------|----------|---|------------|------------|----------|---|--|---|---|
| | Ztraceno | | | | Ztraceno | | | | Ztraceno | | | | |
| 461 | 0 | - | 2DS1 3DS1 2DL2 2DL5B 2DS2 2DS3 2DL5A 2DS5 | 188 | 0 | - | 0 | - | 24 | 7 | 2DL4*0010201 3DL1*0150203 2DL3*0010103 2DP1*0020106 2DS4*0010107 3DL3*0090103 2DL1*0030205 | 0 | - |

Tabulka 5.2: Výsledky experimentu 1 u genomu Test7. Odřezány byly alely, které měly pokrytí menší než 70 %. *Alel* u genomu značí počet v daném genomu. Číslo v závorkách udává kolik alel je v daném genomu dvakrát. V každém kroku *zbývá alel* je kolik alel ještě zůstalo ve výběru, *ztraceno* určuje kolik alel má být v genomu, ale algoritmus je vyřadil. Za tímto číslem jsou vypsané alely, které byly ztraceny. V dalších krocích jsou vypsané alely bez těch, které už byly ztraceny v předchozích krocích. Obdobně je to s *geny navíc*, které udávají počet a druh genů již neobsahující žádnou z alel, která náleží do daného genomu.

Jednou ze smazaných alel je i 2DL4*0010201, která byla smazána kvůli alele KIR2DL4*0010305, taktéž patřící do genomu. Vzdálenost mezi těmito alelami je 10. Suma hloubky pokrytí alely KIR2DL4*0010201 v místě kde se alely neshodují je 63. U alely KIR2DL4*0010305 je to 154.



Po analyzování všech smazaných alel se ukázalo, že všechny byly smazány kvůli alele, která do genomu také patří.

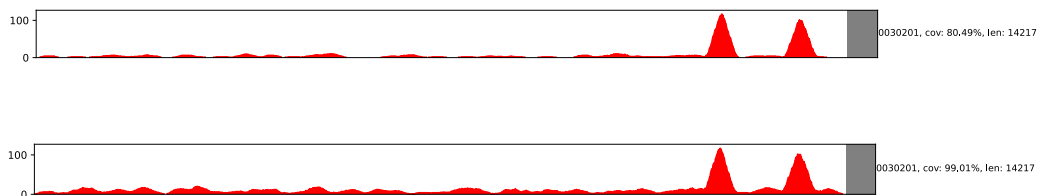
Experiment 2

Experiment 2 byl navržen tak, aby zarovnávání probíhalo ve více iteracích. Mělo by pak být v dalších krocích jasnější, která alela do genomu skutečně patří a která ne, díky tomu, že ready zarovnané na odstřižené alely by se měly zarovnat na alely, které zbývají, a tak podpořit alely patřící do genomu.

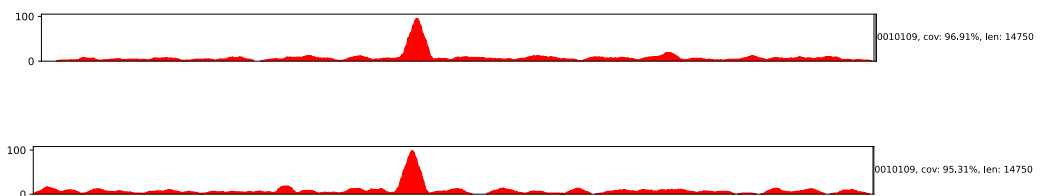
V první fázi se provede zarovnání na celou KIR referenci. V kroku 1 je vytvořena statistika obdobně jako v prvním experimentu a jsou odstřiženy alely, které mají menší pokrytí než *CUT_COVERAGE_ALELS*, které

je uvedené v *configu*. Ze zbylých alel je vytvořena nová reference, na kterou je provedeno nové zarovnání readů. Opět je vytvořena statistika. Výsledky této statistiky nalezneme v kroku 2. Následuje krok 3, v němž soutěží podobné alely v rámci jednoho genu. Podobnost alel je určována na základě Levenshteinovy vzdálenosti. Alely jsou si podobné v případě, kdy je jejich vzdálenost menší než vzdálenost uvedená v *configu* pod parametrem *CLOSE_DISTANCE*. Následuje vytvoření nové reference a nové zarovnání.

Na následujících obrázcích je vidět alela 2DL1*0030201 genomu AMALA. První obrázek značí její zarovnání po prvním kroku. Druhý obrázek ukazuje její zarovnání po druhém kroku. V tomto případě bylo zarovnání jednoznačně ku prospěchu.



Ovšem je možné se potkat i s alelami, kde pokrytí zarovnání klesne, jako v případě 2DL3*0010109 genomu AMALA.

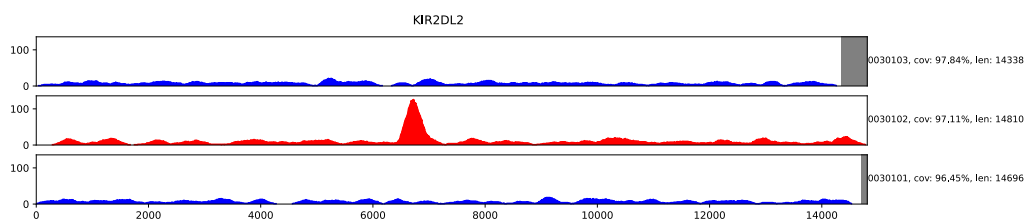


Následující tabulka zobrazuje základní porovnání mezi výsledky experimentu 1 a experimentu 2. Je patrné, že genomu AMALA více zarovnání prospělo. Bylo vyřazeno více alel, které do genomu nepatří. Naopak u genomu RSH došlo ke zhoršení výsledků.

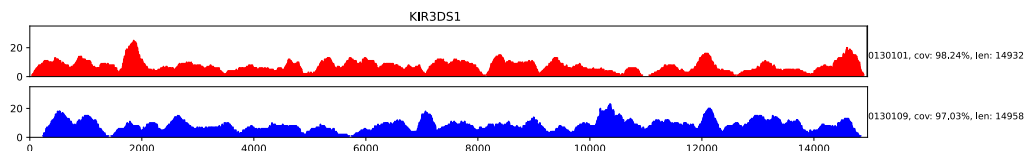
| genom | | Krok 2 | | | Geny navíc | Krok 3 | | |
|-------------|--------|------------|----------|---------|------------|------------|------------------------------|------------|
| | | Zbývá alel | Ztraceno | | | Zbývá alel | Ztraceno | Geny navíc |
| amala exp 1 | 19 (0) | 193 | 0 - | 0 - | 41 | 2 | 2DL4*0010201 3DP1*0090101 | 0 - |
| amala exp 2 | 19 (0) | 193 | 0 - | 0 - | 38 | 2 | 2DL4*0010201 3DP1*0090101 | 0 - |
| rsh exp 1 | 20 (0) | 228 | 0 - | 1 2DL5A | 40 | 2 | 2DP1*0020110 2DL1*0030205 | 1 - |
| rsh exp 2 | 20 (0) | 228 | 0 - | 1 2DL5A | 44 | 2 | 2DP1*0020110 2DL1*0030205 | 1 - |

Tabulka 5.3: Porovnání experimentu 1 a experimentu 2 při stejném nastavení. Odřezány byly alely, které měly pokrytí menší než 70 %. Za podobné byly alely považovány v případě, kdy byla jejich vzdálenost nižší než 100. *Alel* u genomu značí počet v daném genomu. Číslo v závorkách udává kolik alel se v daném genomu vyskytuje dvakrát. V každém kroku *zbývá alel* je kolik alel ještě zůstalo ve výběru, *ztraceno* určuje kolik alel má být v genomu, ale algoritmus je vyřadil. Za tímto číslem jsou vypsané alely, které byly ztraceny. V dalších krocích jsou vypsané alely bez těch, které už byly ztraceny v předchozích krocích. Obdobně je to s *geny navíc*, které udávají počet a druh genů již neobsahující žádnou z alel, která náleží do daného genomu.

Na následujícím obrázku je výsledek genomu AMALA genu 2DL2 po třetím kroku. Alela, která do genomu patří, má jasný vrchol. Tento vrchol označuje místo, kde bylo zarovnáno mnoho readů.



Oproti tomu u genu 3DS1 po kroku 3 jsou zbývající alely k nerozeznání.

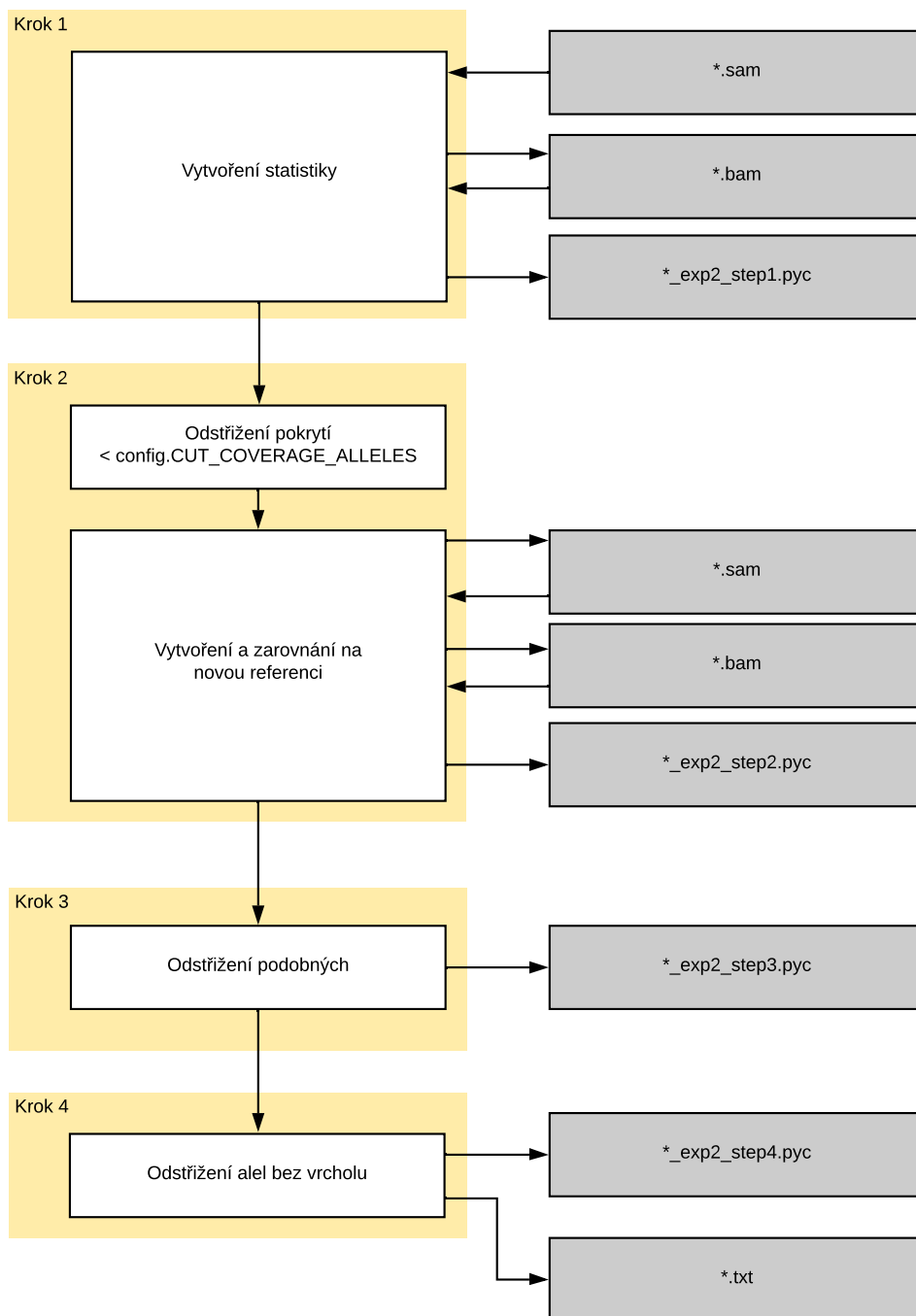


Výsledkem kroku 4 je odstranění alel v případě, kdy sdílejí gen s alelami s jednoznačným vrcholem, ale u genů, kde to není jasné, nic odstraňovat nebudou. Pro každý gen je spočítána hloubka pokrytí pro jednu alelu. Z jednoho genu jsou porovnávány každá alela s každou. Pokud alela1 má 2krát větší maximální hloubku pokrytí než alela2 a zároveň je maximální histogram alely1 2krát větší než průměrná hloubka pokrytí pro všechny alely daného genu, je alela2 odstraněna.

V následující tabulce je porovnání genomu AMALA a kroku 3 a 4. Díky kroku 4 byl algoritmus zpřesněn a odstraněno 19 alel, které do genomu nepatří za cenu ztracení jedné alely, která do genomu patří. Jelikož v praxi je mnohem lepší alelu ztratit než určit že se v genomu nachází alela, která tam

skutečně není, dá se říci, že u tohoto genomu byl krok 4 prospěšný, stejně jako u ostatních genomů.

| Genom | Alel | Krok 3 | | | Krok 4 | | | |
|-------|--------|------------|-----------------------------------|------------|------------|-------------------|------------|--|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | |
| amala | 19 (0) | 38 | 2 2DL4*0010201 3DP1*0090101 | 0 - | 18 | 3 3DL2*0020105 | 0 - | |



Obrázek 5.2: Postup algoritmu experimentu 2.

Experiment 3

Další experiment byl založen na omezení stavového prostoru vytvořením clusterů. Alely jsou seřazeny od nejvíce pokryté po nejméně pokrytou. První alela s největším pokrytím je vložena do prvního clusteru. Další alely projdou všechny clusteru a pokud je vzdálenost alespoň jedné alely z clusteru a právě procházející alely menší než *CLUSTER_DISTANCE* uvedená v *configu*, je alela přiřazena do tohoto clusteru. Pokud ne, pokračuje v porovnávání dokud nekontroluje všechny clusteru. V případě, kde nezapadá ani do jednoho z clusterů, vytvoří nový. Následně jsou porovnány všechny alely z clusteru. Pokud alespoň jedna alela v daném clusteru má pokrytí větší než *CUT_COVERAGE_ALLELES* uvedené v *configu*, není žádná alela v daném clusteru smazána. V opačném případě jsou všechny alely z clusteru smazány. Další postup je stejný s experimentem 2.

Výhodou tohoto přístupu by mělo být zamezení smazání alel, které do genomu patří, ale jejichž ready se zarovnal na alelu jí podobnou.

Při vzdálenosti 5 bylo vytvořeno kolem 224 clusterů, přičemž jich může být o pár jednotek více nebo méně z toho důvodu, že alely nejdou za sebou ve stejném pořadí, ale v pořadí od nejvíce pokrytých po nejméně pokryté. Při vzdálenosti 10 vznikalo kolem 171 clusterů, při vzdálenosti 15 vznikalo 143, a nakonec při vzdálenosti 30 vznikalo kolem 122 clusterů. Takto navržené clusterování nepřineslo očekávaná zlepšení. Je třeba se nad clustrováním alel zamyslet jinak. Na příkladech clusterů uvedených v příloze si lze povšimnout, že většinou alely pocházely ze stejných genů.

Návrh na experiment 4

Clusteru se vytváří zejména z toho důvodu, že si alely mezi sebou "kradou" ready. Možným vylepšením clusterů by mohlo být dát dohromady ty alely, které mají minimální rozdíl v 250 dlouhém useku oproti jiné (protože ready jsou dlouhé 250bp). Následně si pak mohlo vyhledat kolik takových úseků má jedna alela společných s jinou. Problémem tohoto návrhu bude výpočetní náročnost.

Zhodnocení experimentů

Mezi problematické geny lze řadit 2DL5B a 2DL5A, které často měly vysoké pokrytí bez toho aby jakákoliv z jejich alel byla v genomu. Častou ztratovost vykazoval gen 3DL1. Pro zajímavost v případě genomu Test1 bylo po prvním

kroku experimentu 2 maximální pokrytí jeho alel 72 %. V kroku následujícím toto číslo poskočilo na 93 %.

6 Porovnání přístupů k identifikaci a parametrů

Porovnání jednotlivých přístupů probíhalo pomocí přesnosti (precision) 6.0.1 a úplnosti (recall) 6.0.2.

$$Prec = \frac{TP}{TP + FP} \quad (6.0.1)$$

$$Rec = \frac{TP}{TP + FN} \quad (6.0.2)$$

Díky přesnosti je možné odhadnout jak moc jsou výsledky relevantní. Naopak pomocí úplnosti je možné odhadnout kolik skutečně relevantních výsledků bylo přiřazeno. Při převyšující úplnosti bylo získáno mnoho alel, ale mnoho jich do genomu nepatří. Naopak pokud přesnost převyšuje úplnost, tak většina vybraných alel do genomu patří, ale mnoho jich bylo ztraceno. V případě klasických klasifikátorů je snaha balancovat přístup tak, aby obě hodnoty dosahovaly co nejvyšších procent. V tomto případě je ovšem mnohem horší varianta identifikovat alelu, která do genomu nepatří, než neurčit alelu, která do genomu patří. Proto je přesnost důležitějším parametrem než úplnost. V každém případě je jistě nápomocné pokud je o možných výskytech alespoň nějaká informace i kdyby to znamenalo snížit úroveň rozlišení. To by znamenalo v případě nerozhodnosti mezi 3DS1*0130101 a 3DS1*0130109 říci, že tam je jedna z alel na úrovni 3DS1*01301.

| | Konečný krok | Parametry | Přesnost (%) | Úplnost (%) |
|------|--------------|--|--------------|-------------|
| exp1 | 3 | CUT_COVERAGE_ALLELES = 90 CLOSE_DISTANCE = 100 | 65 | 84 |
| | 3 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 | 49 | 90 |
| exp2 | 3 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 | 48 | 91 |
| | 4 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 | 78 | 87 |
| exp3 | 4 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 CLUSTER_DISTANCE = 5 | 78 | 87 |
| | 4 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 CLUSTER_DISTANCE = 10 | 80 | 87 |
| | 4 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 CLUSTER_DISTANCE = 20 | 78 | 88 |
| | 4 | CUT_COVERAGE_ALLELES = 70 CLOSE_DISTANCE = 100 CLUSTER_DISTANCE = 30 | 79 | 87 |

7 Verifikace na reálných datech

Pro validaci algoritmu na reálných datech byl získán kompletní KIR genom u 9-ti komerčních linií ve spolupráci FN Plzeň – LFP UK spolu s definováním konkrétních alel v těchto liniích. Reads byly získány dle protokolu uvedeného v [29], konkrétně byly amplifikovány z izolované DNA pomocí long-range PCR za použití směsi 6 primerů. Sekvenační knihovna je připravena modifikovaným protokolem NEBNext[®] Ultra TM II FS DNA Library Prep Kit for Illumina (New England Biolabs) a knihovny jsou sekvenovány v pair-end režimu na přístroji Illumina Miseq s pokrytím 100.

Prvním krokem při verifikaci na reálných datech je zarovnání. Na následujících obrázcích je porovnání logu Bowtie2 genomu BOB na syntetických readech 7.1 a readech přímo ze sekvenátoru Illumina 7.2. Je možné si zde povšimnout, že v případě syntetických readů byly zarovnány všechny reads. Naopak při zarovnávání reálných dat bylo zarovnáno pouze necelých 70 % readů. Pojem concordantly v tomto případě znamená, že oba páry readů byly zarovnány ve vzájemném souladu. Oproti tomu discordantly znamená, že každý jeden z páru readů byl zarovnán samostatně.

```
align: bob KIR_gen
49000 reads; of these:
  49000 (100.00%) were paired; of these:
    0 (0.00%) aligned concordantly 0 times
    0 (0.00%) aligned concordantly exactly 1 time
    49000 (100.00%) aligned concordantly >1 times
  ----
  0 pairs aligned concordantly 0 times; of these:
    0 (0.00%) aligned discordantly 1 time
  ----
  0 pairs aligned 0 times concordantly or discordantly; of these:
    0 mates make up the pairs; of these:
      0 (0.00%) aligned 0 times
      0 (0.00%) aligned exactly 1 time
      0 (0.00%) aligned >1 times
100.00% overall alignment rate
```

Obrázek 7.1: Výpis zarovnání syntetických readů genomu BOB z Bowtie2.

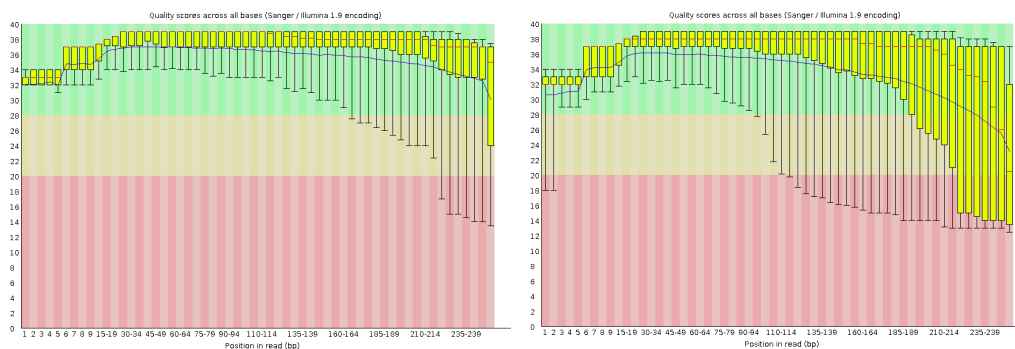
```

align: bob KIR_gen
436363 reads; of these:
  436363 (100.00%) were paired; of these:
    158219 (36.26%) aligned concordantly 0 times
    293 (0.07%) aligned concordantly exactly 1 time
    277851 (63.67%) aligned concordantly >1 times
  ----
    158219 pairs aligned concordantly 0 times; of these:
      10 (0.01%) aligned discordantly 1 time
  ----
    158209 pairs aligned 0 times concordantly or discordantly; of these:
      316418 mates make up the pairs; of these:
        266084 (84.09%) aligned 0 times
        353 (0.11%) aligned exactly 1 time
        49981 (15.80%) aligned >1 times
69.51% overall alignment rate

```

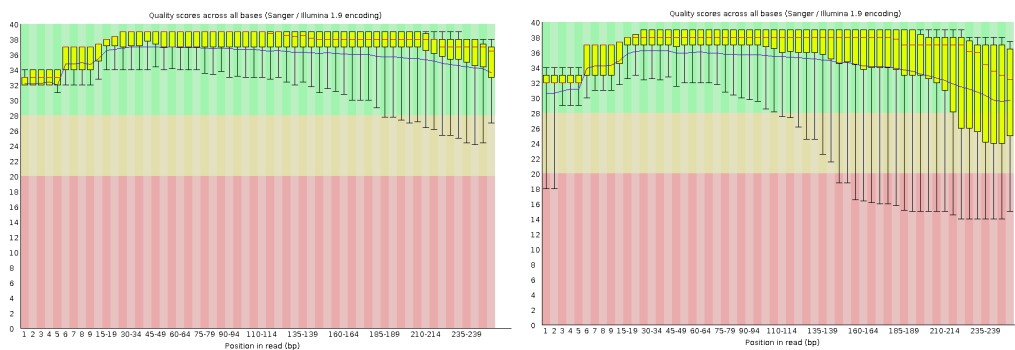
Obrázek 7.2: Výpis zarovnání reálných readů genomu BOB z Bowtie2.

Jedním z důvodů, proč se to děje, mohou být chybné báze v reálných readech, které vznikly sekvenováním. V souboru obsahujícím ready je možné nalézt informaci o kvalitě sekvenace jednotlivých bází. V podstatě se jedná o vyjádření toho, jak moc si je sekvenátor jistý danou bází. Kvalita se označuje jako ASCII znak a je rovna ASCII kódu - 33 (přeskočení netisknutelných znaků). Čím vyšší je ASCII kód, tím je báze kvalitnější. Nejčastěji se hodnota kvality pohybuje mezi 0 a 40, zřídka překročí hodnotu 60. Na následujícím obrázku je zobrazen graf kvality bází vzhledem k pozici v readu.



Obrázek 7.3: Zobrazení kvality dat genomu BOB, v levé části pro první z páru readů a v pravé části pro druhé části readů. V grafu se směrem nahoru zvyšuje kvalita báze. Zleva doprava se pak zvyšuje pozice báze v readu. Je patrné, že čím je pozice v readu vyšší, tím nižší je kvalita báze, a zároveň v druhých párech readů je kvalita znatelně horší.

Nabízí se tedy řešení nekvalitní báze na konci readů odstranit. Na odstranění nekvalitních bází byl použit nástroj Trimmomatic 0.38.0 na serveru Galaxy [4]. Na odstranění bylo použito posuvné okénko velikosti 8 s průměrnými kvalitami 15.



Obrázek 7.4: Zobrazení kvality dat genomu BOB po ořezu nekvalitních bází, v levé části pro první z páru readů a v pravé části pro druhé části readů. V grafu se směrem nahoru zvyšuje kvalita báze. Zleva do prava se pak zvyšuje pozice báze v readu. Je patrné, že čím je pozice v readu vyšší, tím nižší je kvalita báze, a zároveň v druhých párech readů je kvalita znatelně horší.

Následující tabulka zobrazuje porovnání výsledků bez ořezu nekvalitních bází a s jejich ořezem. Jak je vidět, nelze jednoznačně říci, že by se výsledky zlepšily.

| Genom | Alel | Krok 4 | | |
|----------------------------|--------|------------|---|----------------------------|
| | | Zbývá alel | Ztraceno | Geny navíc |
| amala bez ořezu kvality | 19 (0) | 60 | 5 2DL2*00301 2DL3*001 3DL2*0070102 3DL2*0020105 3DP1*00901 | 2 2DL5B 3DL2 |
| amala s ořezem kvality | 19 (0) | 52 | 4 2DL3*001 2DL2*00301 3DL2*0070102 3DP1*00901 | 1 2DL5B |
| bob bez ořezu kvality | 19 (0) | 52 | 7 3DL3*01303 2DL2*00301 2DL3*00201 3DL2*0070102 3DP1*00302 2DL4*001 2DL1*00302 | 2 2DL5B 2DL1 |
| bob s ořezem kvality | 19 (0) | 51 | 7 2DL2*00301 2DL3*00201 3DL3*01303 3DL2*0070102 3DP1*00302 2DL4*001 2DL1*00302 | 3 2DL5B 2DS3 2DL1 |

Tabulka 7.1: Porovnání výsledků při odstrižení nekvalitních bází na experimentu 2. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. *Alel* u genomu značí počet alel v daném genomu. Číslo v závorkách udává, kolik alel se v daném genomu vyskytuje dvakrát. *Zbývá alel* určuje, kolik alel ještě zůstalo ve výběru, *ztraceno* značí kolik alel má být v genomu, ale algoritmus je vyřadil. Za tímto číslem jsou vypsané alely, které byly ztraceny, *geny navíc* udávají počet a jaké geny již neobsahují žádnou z alel, která náleží do daného genomu.

V následujícím textu přibude k referenční alele a nástrojem identifikované alele pojem biologická reference. Myslí se tím identifikace alel provedená v laboratoři, odkud pochází komerční linie.

Dodaná reálná data nemají biologickou referenci až na úroveň non-coding regionu. Příkladem může být biologická reference 2DL1*00201 v níže uvedené ukázce biologické reference genomu KAS011. K této alele je možné přiřadit až 16 alel z referenčních sekvencí. Díky tomuto problému není možné plnohodnotně posoudit přesnost a úplnost nástroje. Znak + značí přítomnost genu bez bližšího určení alely, tzn. že v některých případech není alela určena vůbec.

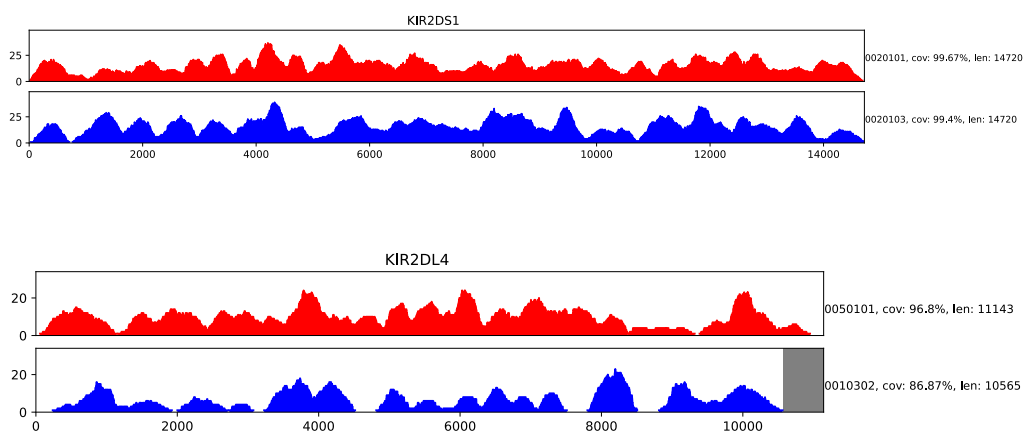
Ukázka dodaných identifikací KAS011

- 2DL1: 00201, 00302
- 2DL2:
- 2DL3: +
- 2DL4: 00103, 005
- 2DL5A: 00101
- 2DL5B:
- 2DP1: 002, 00301
- 2DS1: 00201
- 2DS2:
- 2DS3:
- 2DS4: 00301
- 2DS5: 00201
- 3DL1: 008
- 3DL2: 01001, 019
- 3DL3: 00901, 01302
- 3DP1: 00302, 006
- 3DS1: 01301

Podrobnější zhodnocení všech verifikačních linií vycházející z identifikace dané úrovně rozlišení je uvedeno v následující kapitole.

8 Zhodnocení z hlediska úrovně rozlišení

Na následujících dvou grafech jsou zobrazeny všechny zbývající alely genů 2DS1 a 2DL4 u genomu BOB. Jak je vidět z obrázku, u genu 2DS1 je možné určit coding region tzn. prvních pět čísel. Oproti tomu v případě genu 2DL4 to možné není.



Následující tabulka zobrazuje možnost rozlišení alel u syntetických readů. V případě genu 3DL3 lze jednoznačně prohlásit, že alely od sebe rozlišit nelze. Oproti tomu v případě genu 2DS2 a 3DL3 alely rozlišit lze.

| | 2DL1 | 2DL4 | 2DS1 | 2DS2 | 2DS4 | 2DS5 | 3DL1 | 3DL2 | 3DL3 | 3DP1 | 3DS1 |
|--------|-------|-------|-------|-------|-------|------|-------|-------|------|------|-------|
| amala | | | | | | | | | N | | A - 5 |
| bob | | N | A - 5 | A - 5 | | | N | | | | |
| cox | | | | | | | A - 5 | A - 5 | N | | |
| ho301 | | A - 3 | | | | | N | | | | |
| jvm | | A - 5 | | | N | | | | | | |
| kas011 | | A - 5 | | | A - 5 | | | | N | | |
| olga | A - 5 | | | | | | N | A - 5 | | | |
| rsh | | | | A - 5 | | | | | | | |
| wt51 | | | | | | N | | | | | |
| test1 | | | | A - 5 | | | N | | | | |
| test2 | | | | | | | A - 5 | | | | |
| test3 | | | | | | | N | | | | |
| test4 | | | A - 5 | | | | A - 5 | | | | |
| test5 | | | | | | | | | N | N | |
| test6 | N | | | | | | | | N | | |
| test7 | A - 5 | A - 5 | | | | | | | N | | |
| test8 | | | | A - 5 | | | | | N | | |
| test9 | | | | A - 5 | A - 5 | | | | N | | |
| test10 | | | | A - 5 | | | A - 5 | | | | |
| test11 | | | | A - 5 | | | | | | | |

Tabulka 8.1: Zhodnocení zbylých alel v experimentu 2 po 4. kroku z hlediska úrovně rozlišení. Tabulka obsahuje jen alely, které měly ve výsledcích nějakou informaci o úrovni rozlišení. V případě, kdy se v políčku nachází znak *N*, znamená to, že zbylo více alel, přičemž některá z nich nepatřila do genomu. V případě znaku "A" alely rozlišit lze. Za pomlčkou se pak nachází číslo určující úroveň rozlišení, tedy na kolik čísel mají společný základ.

V případě reálných dat není možné úroveň rozlišení jednoduše určit, protože biologická reference není určena až na nejnižší úroveň, jak již bylo uvedeno v předchozí kapitole. Níže je zobrazená tabulka zjednodušeně popisující úroveň rozlišení u reálných dat. U genů 2DL5A, 2DS1, 2DS2 a 2DS3 bylo určení alely minimálně na 3 místa v coding regionu nebo v souladu s biologickou referencí. V případě ostatních genů se vždy vyskytoval alespoň jeden případ, kdy bylo možné alely rozlišit.

| | amala | bob | cox | ho301 | jvm | kas011 | olga | rsh | wt51 |
|-------|-------|-----|-----|-------|-----|--------|------|-----|------|
| 2DL1 | NE | NE | 5R | 3R | 5R | 5R | NE | 5R | 0R |
| 2DL4 | 3 | 3R | NE | 3 | 5R | 3R | NE | 7R | 5R |
| 2DL5A | 3R | 3 | 5R | | | 5R | 5R | | 5R |
| 2DL5B | | | | NE | | | | 4R | |
| 2DP1 | 5R | 5R | 5R | 5R | NE | 5R | 5R | 3R | 3R |
| 2DS1 | 5R | 5R | 5R | | | 5R | 3R | | 3R |
| 2DS2 | 5R | 3R | | 5R | 5R | | | 5R | 5R |
| 2DS3 | | | | 5R | | | | | 0R |
| 2DS4 | NE | NE | 3R | 3R | 6R | 5R | 3R | 0R | |
| 2DS5 | NE | NE | 5R | | | NE | NE | 6R | NE |
| 3DL1 | NE | NE | 5 | 3R | 3R | 3R | NE | 3R | |
| 3DL2 | NE | NE | NE | 5R | 3R | NE | NE | 0R | 0R |
| 3DL3 | NE | NE | NE | NE | NE | NE | NE | NE | 3R |
| 3DP1 | 3R | NE | 3R | 5R | 3R | 3R | NE | 3R | 0R |
| 3DS1 | NE | 5R | 3R | | | NE | NE | | NE |

Tabulka 8.2: Zhodnocení zbylých alel u reálných dat v experimentu 2 po 4. kroku z hlediska úrovně rozlišení. Tabulka obsahuje jen alely, které měly ve výsledcích nějakou informaci o úrovni rozlišení. V případě, kdy se v políčku nachází znak *NE* nebylo možné alelu rozpoznat, ale v biologické referenci byla uvedena. V případě čísla značí, na jak velkou přesnost byla alela určena. R za tímto číslem značí, že tato úroveň rozlišení byla uvedena i v biologické referenci. Pokud je políčko prázdné, gen se v genomu nenachází nebo byl ztracen.

Ať už z důvodu nezarovnání všech reálných readů, možné přítomnosti alely, pro kterou není referenční sekvence, či kvůli jiné z výše uvedených příčin, jsou tyto výsledky očekávané.

9 Závěr

Práce se zabývá návrhem nástroje pro automatickou identifikaci KIR alel. Identifikace KIR alel je využitelná při transplantaci krvetvorných buněk, kdy se rozhoduje mezi více dárci shodnými v HLA znacích. Zvolení dárce s vhodnějšími KIR geny snižuje riziko relapsu (návratu nemoci).

Toto téma je stále aktuální. Zatím je prokázán vliv jen KIR genů, resp. jejich haplotypových kombinací. V současné době se zkoumá také vliv alelických variant, jelikož doposud bylo prokázáno, že některé formy genů se neexprimují, tzn. gen se sice v genomu nachází, ale není z něj vytvořen žádný produkt na povrchu buňky a tak nemá jak vyvolat či ovlivnit jakoukoliv reakci v organismu [40]. Z biologického hlediska je důležitá právě expimace, která je závislá jen na coding-regionu. Tudíž z praktického hlediska stačí, aby nástroj byl schopný určit alelu na coding úroveň, tedy prvních pět čísel alely.

V teoretické části byly popsány a rozbrány geny, především non-HLA gen Killer-cell immunoglobulin-like receptor (KIR), a jejich vliv na transplantaci krvetvorných buněk. Dále byly představeny natural killer buňky a jejich funkce v rámci imunitního systému. Pro zjištění typizace HLA a KIR slouží sekvenční metody, které byly shrnuty a popsány. Konkrétněji Sanger sekvenování a next-generation sequencing (NGS). V neposlední řadě byly analyzovány bioinformatické nástroje pro zpracování NGS dat se zaměřením na generátor syntetických readů ART a zarovnávač NGS readů Bowtie2.

V realizační části bylo navrženo a implementováno prototypové řešení v jazyce Python za pomoci knihovny pysam pro identifikaci KIR alel. Jazyk Python byl zvolen z důvodu jednoduchého použití, díky čemuž nebyl vývoj časově náročný. Parametry nástroje ART byly určeny tak, aby výstupní ready co nejvíce odpovídaly datům získaným z FN Plzeň/BC LF UK Plzeň. Během práce bylo implementováno a vyzkoušeno několik přístupů, které byly v práci zhodnoceny. Největší překážkou je podobnost alel. Alely mohou být dlouhé až 16 000 bp (nukleových bází). Ready ze sekvenátoru jsou ovšem dlouhé 250 bp. Vzhledem k podobnosti alel (u některých dokonce jen jedna nukleová báze) není žádnou výjimkou, že read je zarovnan na alelu, ke které nepatří.

Nástroj byl implementován a laděn na syntetických readech. Následná verifi-

kace byla provedena na komerčních buněčných liniích sekvenovaných ve FN Plzeň/BC LF UK Plzeň. Verifikace nástroje na reálných datech odpovídala očekávaným předpokladům. Rozdíl reálných readů od syntetických byl už v jejich zarovnání, kdy v případě reálných dat bylo zarovnáno kolem 70 % readů, oproti tomu v případě syntetických readů byly zarovnány všechny. V rámci návrhu nástroje a provedených experimentů byly identifikovány základní slabiny automatické identifikace KIR alel, a sice že reálná data nejsou vždy identifikována až na úroveň non-coding region, protože z biologického hlediska není v některých případech možné určit alelu na nejnižší známou úroveň, či označení, že se zde gen nachází, ale ani zmínka o tom, o jakou alelu by se mohlo jednat. Tyto problémy souvisejí i se zhodnocením z hlediska úrovně rozlišení, kde u reálných dat v případě biologické reference pouze na coding region vyjde více alel na non-coding region se stejným základem codingu regionu. Kromě toho je v reálných datech možné narazit na alelu, která se mezi referenčními sekvencemi nenachází. Nyní je známo 1109 alel, přičemž existuje referenční sekvence pouze pro 461 alel.

Do budoucna se nabízí zamyslet se více nad efektivním vytvořením clusterů, kde by se zohledňovala délka readů. V případě verifikace by bylo vhodné více probádat kvalitu dat a odšťihávání nekvalitníchází.

10 Výkladový slovník pojmů a zkratek

| | |
|--------|---|
| WHO | World health organization, světová zdravotnická organizace |
| ČNRDD | Český národní registr dárců kostní dřeně |
| MHC | Major histocompatibility complex, genetický systém |
| HLA | Human leucocyte antigen, podskupina MHC |
| KIR | Killer immunoglobulin like-receptor, skupina genů |
| NK | Natural killer, buňka imunitního systému |
| DNA | Deoxyribonukleová kyselina; dvoušroubovice, která obsahuje páry bází C, G, A, T |
| RNA | Ribonuklové kyselina; obsahuje báze C, G, A, U; šablona přímo pro vytvoření proteinů; hlavní funkcí zajištění překladu DNA do struktury proteinů (DNA -> mRNA -> rRNA -> tRNA -> RNA) |
| Báze | nukleové báze; A - Adenin, C - Cytosin, G - Guanin, T - Thymin |
| bp | base pair; jeden z párů A - T nebo C - G |
| kb | kilobase 1 kb = 1000 bp |
| ART | nástroj na vytváření syntetických readů |
| Bowtie | nástroj na zarovnání readů proti referenčním genům |
| SAM | Sequence Alignment/Map; Formát souboru na uložení zarovnání |
| BAM | Binární verze souboru SAM |

Literatura

- [1] *National Institute of Environmental Health Sciences* [online]. Dostupné z: <https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm>.
- [2] *Bowtie2* [online]. Dostupné z: <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>.
- [3] *Chromosome* [online]. [cit. 2020/12/3]. Dostupné z: <https://www.genome.gov/genetics-glossary/Chromosome>.
- [4] *Galaxy* [online]. Dostupné z: <https://galaxyproject.org/>.
- [5] *DNA sequencing Fact Sheet* [online]. [cit. 2019/03/1]. Dostupné z: <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet>.
- [6] *Basic genetics* [online]. [cit. 2020/12/3]. Dostupné z: <https://kintalk.org/genetics-101/>.
- [7] *Illumina* [online]. [cit. 2019/03/1]. Dostupné z: <https://www.illumina.com/>.
- [8] *KIR genotypes* [online]. Dostupné z: <http://www.allelefrequencies.net/kir6001a.asp>.
- [9] *Matplotlib* [online]. Dostupné z: <https://matplotlib.org/>.
- [10] *What is the best NGS alignment software?* [online]. Dostupné z: <https://www.ecseq.com/support/ngs/what-is-the-best-ngs-alignment-software>.
- [11] *Numpy* [online]. Dostupné z: <https://numpy.org/>.
- [12] *Pysam* [online]. Dostupné z: <https://pypi.org/project/pysam/>.
- [13] *Python-Levenshtein* [online]. Dostupné z: <https://pypi.org/project/python-Levenshtein/>.
- [14] *How do you put a genome back together after sequencing?* [online]. Dostupné z: <https://www.yourgenome.org/facts/how-do-you-put-a-genome-back-together-after-sequencing>.

- [15] BARANWAL, A. – MEHRA, N. Major Histocompatibility Complex Class I Chain-Related A (MICA) Molecules: Relevance in Solid Organ Transplantation. *Frontiers in Immunology*. 02 2017, 8. doi: 10.3389/fimmu.2017.00182.
- [16] BERNAREGGI, D. – POUYANFARD, S. – KAUFMAN, D. S. Development of innate immune cells from human pluripotent stem cells. 2019. Dostupné z: <https://www.sciencedirect.com/science/article/pii/S0301472X19300037?via%3Dihub>.
- [17] CIMBÁLNÍKOVÁ, B. E. Transplantace kostní dřeně, její historický vývoj a současné trendy, 2019.
- [18] COOLEY, S. – WISDORF, D. J. – GUETHLEIN, L. A. Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia. 2010. Dostupné z: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2953880/>.
- [19] ESCALONA, M. – ROCHA, S. – POSADA1, D. A comparison of tools for the simulation of genomic next-generation sequencing data. 2016. Dostupné z: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5224698/#>.
- [20] FRYČOVÁ, M. Lze u pacientů s AML indikovaných k nepřibuzenské transplantaci provádět v klinické praxi výběr nepřibuzných dárců na základě KIR genotypů, JCU - BP, 2016.
- [21] HERNYCHOVÁ, L. Receptory NK buněk. 2012. Dostupné z: <https://is.cuni.cz/webapps/zzp/detail/118114>.
- [22] HOUDOVÁ, L. et al. Optimal Donor Selection During Verification Process: Which Factors Are Worth Knowing? 2018. doi: 10.1016/j.transproceed.2018.08.043.
- [23] HUANG, W. et al. ART: a next-generation sequencing read simulator. 2012. Dostupné z: <https://academic.oup.com/bioinformatics/article/28/4/593/213322>.
- [24] J, R. et al. *Nomenclature* [online]. Nucleic Acids Research, 2015. [cit. 2019/10/1]. 43:D423-431. Dostupné z: <http://hla.alleles.org/misc/citing.html>.
- [25] KOLÍSKO, M. Moderní metody sekvenování DNA. 2017. Dostupné z: <https://ziva.avcr.cz/files/ziva/pdf/moderni-metody-sekvenovani-dna.pdf>.
- [26] LANGMEAD, B. [online]. [cit. 2019/03/1]. Dostupné z: <http://www.langmead-lab.org/>.

- [27] LANGMEAD, B. – SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. 2012. Dostupné z: https://www.researchgate.net/publication/221886241_Langmead_B_Salzberg_SL_Fast_gapped-read_alignment_with_Bowtie_2_Nat_Methods_9_357-359.
- [28] LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. 2009. Dostupné z: <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25>.
- [29] MANIANGOU, B. et al. Killer Immunoglobulin-Like Receptor Allele Determination Using Next-Generation Sequencing Technology. *Frontiers in Immunology*. 2017, 8, s. 547. ISSN 1664-3224. doi: 10.3389/fimmu.2017.00547. Dostupné z: <https://www.frontiersin.org/article/10.3389/fimmu.2017.00547>.
- [30] MERZKER, M. L. Sequencing technologies-the next generation. 2013. doi: 10.1038/nrg2626.
- [31] MUDR. PAVEL JINDRA, P. D. *Imunopatologické a imunogenetické aspekty transplantací krevetvorných buněk a solidních orgánů*. PhD thesis, Universita Karlova v Praze, 2011.
- [32] PAPOUŠEK, I. Elektroforéza nukleových kyselin. 2017. Dostupné z: https://fvhe.vfu.cz/files/mbhp_2018_02.pdf.
- [33] PENDE, D. et al. *Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation* [online]. 2019. Dostupné z: <https://www.frontiersin.org/articles/10.3389/fimmu.2019.01179/full>.
- [34] PENKA, M. – KOLEKTIV, E. T. *Hematologie a transfuzní lékařství II*. 2012. ISBN 978-80-247-3460-6.
- [35] ROBINSON, J. et al. IPD—the Immuno Polymorphism Database. 2013. Dostupné z: <https://www.ebi.ac.uk/ipd/index.html>.
- [36] ROBINSON, J. et al. The IPD and IMGT/HLA Database:allele variant databases. 2015. Dostupné z: <https://www.ebi.ac.uk/ipd/index.html>.
- [37] S.KANNANA, G. – ARIANEXYSAQUINO-LOPEZ – A.LEED, D. Natural killer cells in malignant hematology: A primer for the non-immunologist. 2017. Dostupné z: <https://www.sciencedirect.com/science/article/pii/S0268960X16300704>.
- [38] SMITH, D. T. *Encyklopedie lidského těla*. 2005. ISBN 80-7321-156-4.

- [39] THIELENS, A. – VIVIER, E. – ROMAGNÉ, F. NK cell MHC class I specific receptors (KIR): from biology to clinical intervention. *Current opinion in immunology*. 2012, 24 2, s. 239–45.
- [40] VILCHES, C. – GARDINER, C. – PARHAM, P. Gene Structure and Promoter Variation of Expressed and Nonexpressed Variants of the KIR2DL5 Gene. *Journal of immunology (Baltimore, Md. : 1950)*. 01 2001, 165, s. 6416–21. doi: 10.4049/jimmunol.165.11.6416.
- [41] VODSLOŇ, F. Vyhodnocování podobnosti zdrojových textů, 2007.
Dostupné z: <http://www1.cuni.cz/~obo/vyuka/projekty/vodslon-ukazkova-bc.pdf>.

A Uživatelská dokumentace

Ke spuštění programu je třeba mít nainstalovaný:

- ART ve verzi MountRainier
- Bowtie2 ve verzi 2.4.1
- Python ve verzi 3.8
- Python knihovny
 - pysam
 - python-Levenshtein
 - matplotlib
 - numpy

Doporučený operační systém je Linux, na Windows se mohou vyskytnout problémy s knihovnou htslib.

A.1 Získání potřebného softwaru

Python knihovny a moduly

Knihovny a moduly je možné získat pomocí PyPi.

- `pip install pysam`
- `pip install python-Levenshtein`
- `pip install matplotlib`
- `pip install numpy`

Jinou možností je použít oficiální návody uvedené pro knihovnu pysam na [12], knihovnu python-Levenshtein [13], matplotlib [9] a numpy [11].

ART

Nástroj ART slouží ke generování syntetických readů. Odkaz na stažení nástroje u oficiálního článku [23] je bohužel nefunkční. Je možné využít tento zdroj [1]. V případě, že bude i tento link nefunkční je možné nástroj vyhledat na stejné stránce po zadání textu *ART* do vyhledávacího pole, typicky první

výsledek *ART - Set of simulation tools ...*

Práce je odzkoušená s verzí MountRainer. Ve staženém archivu se nachází několik souborů *README*. Pro tuto práci je důležitý soubor *art_illumina_README.txt*, který obsahuje návod na instalaci nástroje a vysvětlení parametrů.

Parametry shodující se s parametry dat získaných z FN Plzeň/BC LF UK Plzeň jsou paired end ready dlouhé 250 bp, profil sekvenátoru MSv1, s nako-pírování sekvencí 100. Všechny tyto parametry jsou již nastavené ve skriptu.

Bowtie2

Bowtie2 slouží k zarovnání readů na referenční sekvenci a je dostupný na [2]. Práce je odzkoušená s verzí 2.4.1. Ve stažené verzi je manuál s potřebnými instrukcemi. Stejně tak je na stránkách Bowtie2 sekce Getting started.

A.2 Získání referenčních sekvencí

Referenční soubory *_gen.fa* a *_nuc.fa* použité v práci jsou k nalezení na CD ve složce *prilozena data/referencni sekvence*. Případně nejnovější verzi je možné získat na IPD-KIR [36] v záložce *KIR*, dále v pravém menu *GitRepos* složka *fasta*.

A.3 Nastavení programu

Veškeré nastavení aplikace probíhá pomocí souboru *config.py*

Parametry configu:

- `CREATE_READS` - Značí, zda má být spuštěn modul vytvoření syntetických readů. Očekávaná hodnota je `True` nebo `False`.
- `ALIGN` - Značí, zda má být spuštěn modul pro zarovnání readů vzhledem k referenčním genům. Očekávaná hodnota je `True` nebo `False`.
- `IDENTIFY` - Značí, zda má být spuštěna identifikace alel. Očekávaná hodnota je `True` nebo `False`.
- `PRECOMPUTATION_DISTANCE` - Předpočítání vzájemné vzdálenosti mezi jednotlivými alelami do `pyc` souboru.

- EXP1 - Značí, zda má být spuštěn experiment 1 pro identifikaci.
- EXP2 - Značí, zda má být spuštěn experiment 2 pro identifikaci.
- EXP3 - Značí, zda má být spuštěn experiment 3 pro identifikaci.
- PATH_TO_DATA_FOLDER - Značí cestu ke složce s daty. Dá se využít v případě, kdy adresářová struktura odpovídá struktuře navržené níže.
- REFERENCE_KIR_GENS_FILE - Soubor se všemi referenčními geny.
- GENOME_FOLDER - Označuje cestu složky, do které jsou ukládány vytvořené genomy.
- GENOMES - Slovník, který definuje genomy podle obsahů genů. Na základě toho budou vytvořeny genomy.
- BOWTIE_HOME_DIRECTORY - Označuje cestu ke nástroji Bowtie.
- READS_FOLDER - Označuje složku, do které budou ukládány ready. Případně z které budou načítány.
- BOWTIE_INDEX_FOLDER - Označuje složku, do které budou ukládány indexy z Bowtie. Případně z které budou načítány.
- BOWTIE_BUILD_INDEX - Značí zda mají být vytvořeny Bowtie indexy. Pokud bude hodnota nastavena na False, musí být přítomny indexy z minulého běhu, jinak zarovnávání nebude fungovat. Očekávaná hodnota je True nebo False. (Pro první běh je typická hodnota True)
- BOWTIE_THREADS - Počet vláken na která má být Bowtie spuštěn.
- ALIGNMENT_FOLDER - Označuje složku, do které budou ukládány zarovnané ready. Případně z které budou načítány.
- BAM_FOLDER - Označuje složku na uložení BAM souborů.
- RESULT_FOLDER - Označuje složku, do které budou uloženy výsledky vyhodnocení.
- ALELS_STATISTICS_FOLDER - Označuje složku, do které budou uloženy statistiky mezivýsledků.

- TEMP_FOLDER - Složka, kam se budou ukládat soubory, které jsou potřeba jen pro mezizpracování, jako jsou bowtie indexy při mnohonásobném zarovnávání u experimentu 2 a experimentu 3.
- REFERENCE_FOLDER - Složka kam se ukládají nově vytvořené reference z experimentů.
- ALELS_DISTANCE_FILE_PYC - Soubor do kterého se budou ukládat předpočítané vzdálenosti mezi jednotlivými alelami.
- LEVENSHTAIN_DISTANCE_CUT - Tato hodnota udává, do jak vysoké hodnoty bude věnována pozornost vzdálenosti mezi jednotlivými alelami a které vzdálenosti tedy budou ukládány. Je to především z důvodu paměťové náročnosti.
- CUT_COVERAGE_ALLELES - Hodnota určující minimální procentuální pokrytí alel, které mají být zachovány ve výsledcích.
- CLOSE_DISTANCE - Hodnota určující do jaké vzdálenosti jsou dvě alely považovány za blízké.
- CLUSTER_DISTANCE - Hodnota určující maximální vzdálenost alel, které spolu budou sloučeny do clusteru.

A.4 Doporučená adresářová struktura pro data

- data
 - alignments
 - bam
 - bowtie_index
 - genome
 - reads
 - reference
 - result
 - statistics
 - temp

A.5 Spuštění programu

Program je možné spustit z příkazové řádky pomocí příkazu:

- `python run.py`

Podmínkou fungování tohoto postupu je, že je třeba se nacházet v umístění skriptu.

A.6 Výstupy programu

V případě tvorby vlastních genomů s doporučenou adresářovou strukturou najdeme ve složce *genome* soubory s příponou *.fa*. Každý soubor obsahuje jeden genom. Vytvořené ready se budou nacházet ve složce *reads*. Protože genomy jsou paired-end, náleží každému genomu dva soubory s příponou *.fq*. Jeden s *1* na konci a druhý s *2* na konci. V případě zarovnání mohou být výstupní soubory indexy ve složce *bowtie_index*. Kdy pro každý referenční soubor je vytvořeno 6 souborů s příponou *bt2*. Výsledné zarovnání se pak nachází ve složce *alignments* ve formátu *.sam*.

A.6.1 Výsledky

Výsledky samotné identifikace je možné nalézt ve složce *result* ve formátu *.txt*. Mezivýsledky jednotlivých kroků experimentů jsou uloženy ve složce *statistics* a je možné si je prohlédnout pomocí přiloženého skriptu *analysis_after_align.py*, který bude popsán níže. Soubor mezivýsledků může vypadat následovně *amala_KIR_GEN_exp2_step1.pyc*. Označení začíná názvem genomu - *amala*, *KIR_GEN* značí referenční sekvenci, *exp2* udává o který experiment se jedná, *step1* určuje ke kterému kroku mezi výsledky patří.

A.7 Analýza referenčních genů

K analyzování referenčních genů slouží skript *analysis_alels.py* nacházející se na CD ve složce *software/analysis*. V horní části skriptu je třeba nastavit příslušné cesty.

- `NUC_FILE` - Soubor z referencí *_nuc.fa*. Slouží pro porovnání se souborem *_gen.fa*
- `GEN_FILE` - Soubor s referenčními sekvencemi.

- `DISTANCE_FILE_PYC` - Soubor obsahující všechny vzdálenosti a který bude vytvořen na začátku skriptu.
- `PLOT_OUTPUT_FOLDER` - Složka udávající kam se budou ukládat výsledné grafy.
- `COUNT_LEVENHSTEIN` - Předpočítá vzdálenosti mezi alelami. Využitelné především při opakovaných bězích se stejnou referencí. Při nastavení hodnoty na `True` je výpočet velmi pomalý.

Získání referenčních souborů `__gen.fa` a `__nuc.fa`, které byly použity v práci, je popsáno v sekci Získání referenčních sekvencí.

A.7.1 Spuštění

Analýzu referenčních genů je možné spustit z příkazové řádky pomocí příkazu:

- `python analysis_alels.py`

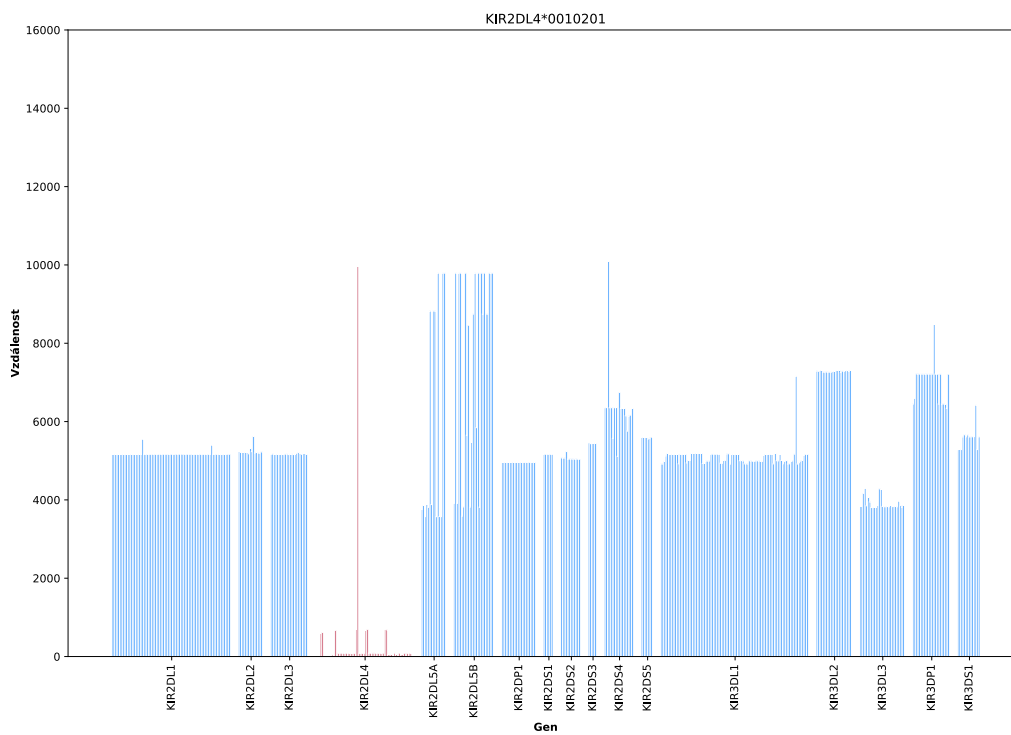
Podmínkou fungování tohoto postupu je, že je třeba se nacházet v umístění skriptu.

A.7.2 Výstupy programu

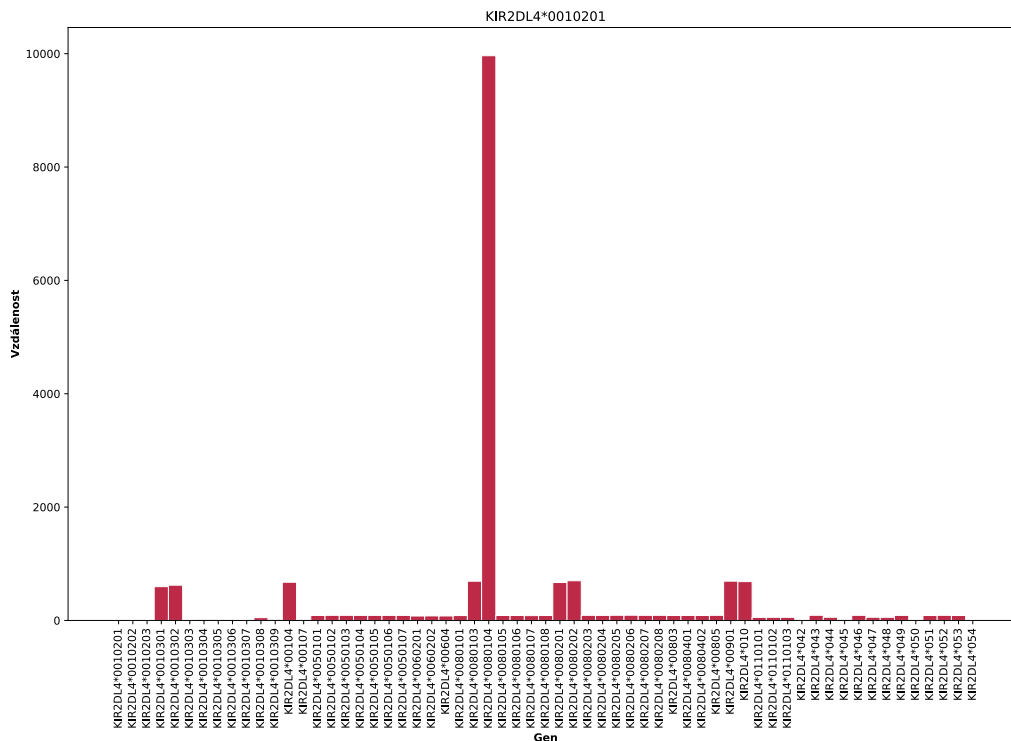
Výstupem analýzy jsou grafy vzdáleností mezi jednotlivými alelami uložených ve složce `PLOT_OUTPUT_FOLDER` a jednoduchá statistika, která je zobrazena na následujícím obrázku. `count_gen` označuje celkový počet alel v souboru `GEN_FILE`, `count_nuc` označuje počet alel v souboru `NUC_FILE`, `match` značí kolik alel je shodných v souboru `GEN_FILE` a `NUC_FILE`, `key not found` značí kolik alel se vyskytuje jen v jenom ze souborů `GEN_FILE` a `NUC_FILE`, `something_wrong` značí nesoulad mezi označením alel podle obsahu a podle pořadového čísla nalezení mezi soubory `NUC_FILE` a `GEN_FILE`. Na druhém řádku `max` značí největší vzdálenost mezi alelami, `min` značí nejmenší vzdálenost mezi alelami, `average_distance` značí průměrnou vzdálenost mezi alelami, `median` značí median vzdáleností mezi alelami.

```
count_gen: 461 , count_nuc: 1109 , match: 461 , key not found: 648 , something wrong: 0
max: 15943 , min: 1 , average_distance: 4768.984212015467 , median: 4251.0
```

Obrázek A.1: Zobrazení výstupu ze skriptu `analysis_alels.py`



Obrázek A.2: Ukázka výstupního grafu z analýzy referenčních genů



Obrázek A.3: Ukázka výstupního grafu z analýzy referenčních genů

A.8 Analyzování experimentů

Pro analyzování kroků jednotlivých experimentů je možné využít skript *analysis_after_align_auto.py* nacházející se na CD ve složce *software/analysis*. V horní části skriptu je třeba nastavit příslušné cesty.

- GEN_FILE - Soubor s referenčními alelami.
- ALELS_STATISTICS_PYC_FOLDER - Složka s pyc soubory obsahující statistiku jednotlivých kroků. Typickým příkladem tohoto souboru může být *amala_KIR_gen_exp1_step1.pyc*.
- ALELS_STATISTICS_PYC_REFERENCE_NAME - Jméno reference, toto je jeden z parametrů, ze kterých se skládá název souborů ze statistics.
- ALELS_STATISTICS_PYC_EXPERIMENT - Jméno experimentu, typické hodnoty: exp1, exp2 a exp3
- STEPS - Kroky, které se mají analyzovat, typické hodnoty: step1, step2, step3 a step4
- STEPS_LATECH_TABLE - Určuje, které kroky se mají vygenerovat do latex tabulky
- PLOT_OUTPUT_FOLDER - Složka udávající kam se budou ukládat výsledné grafy.
- GENOMES_LIST - Výpis genomů, které mají být analyzovány.
- GENOMES_ALLELES - Slovník referenčních odpovědí.

A.8.1 Spuštění

Analýzu je možné spustit z příkazové řádky pomocí příkazu:

- `python analysis_after_align_auto.py`

Podmínkou fungování tohoto postupu je, že je třeba se nacházet v umístění skriptu.

A.8.2 Výstupy programu

Výstupem jsou grafy pokrytí jednotlivých alel uložených ve složce *PLOT_OUTPUT_FOLDER*, rozříděných podle genů, kroku a genomu.

A.9 Používané soubory

Fastq a fq

Soubory s příponou *.fastq* nebo *.fq* obsahují ready a jsou typickým výstupním souborem ze sekvenátoru. V těchto souborech je mimo jiné pro každou bázi vyznačena i kvalita dané báze. Dá se říci, že to vyjadřuje jak si je sekvenátor jistý danou bází. Kvalita se označuje jako ASCII znak. Kvalita je pak rovna ASCII kodu -33 (přeskočení netisknutelných znaků). Čím vyšší je ASCII kód, tím je báze kvalitnější. Nejčastější kvalita je od 0 do 40, zřídka překročí hodnotu 60. Pro paired end ready jsou vytvořeny vždy dva soubory, jeden s prvními z párů readu a druhý soubor s druhými z párů readů.

SAM

Soubor s příponou *.sam* obsahuje výsledky zarovnání readů.

B Testovací genomy

| Test1 | Test2 | Test3 | Test4 |
|---|---|---|--|
| 3DL3: 0030101, 0140201 2DS2: 0010104 2DL2: 0030105 2DL3: 0020101 2DL5B: 01301 2DS3: 0020102 (2x) 2DP1: 0030101, 0010203 2DL1: 0030203, 007 3DP1: 004 (2x) 2DL4: 0080104, 010 3DL1: 0150101 3DS1: 014 2DL5A: 00102 2DS5: 2DS1: 0020104 2DS4: 0010103 3DL2: 00501 (2x) | 3DL3: 0090102, 019 2DS2: 2DL2: 0010101 2DL3: 0010111 2DL5B: 0020101 2DS3: 2DP1: 0020107, 0030102 2DL1: 0020102, 0030210 3DP1: 004, 01001 2DL4: 0080104 (2x) 3DL1: 0070101 3DS1: 078 2DL5A: 0050101 2DS5: 010 2DS1: 0020102 2DS4: 0010103 3DL2: 0020101, 00501 | 3DL3: 005, 0140201 2DS2: 2DL2: 2DL3: 0010101, 0020103 2DL5B: 2DS3: 2DP1: 0020108 (2x) 2DL1: 0040101, 008 3DP1: 0030102, 00902 2DL4: 0010306, 0050104 3DL1: 002, 0040101 3DS1: 2DL5A: 2DS5: 2DS1: 2DS4: 3DL2: 0010301, 008 | 3DL3: 0030104, 007 2DS2: 0010105 2DL2: 0030101 2DL3: 0010102 2DL5B: 2DS3: 2DP1: 008 2DL1: 007 3DP1: 007, 00902 2DL4: 0010307, 0080104 3DL1: 0150202 3DS1: 055 2DL5A: 2DS5: 007 2DS1: 0020101 2DS4: 0060101 3DL2: 0020101, 00903 |
| Test5 | Test6 | Test7 | Test8 |
| 3DL3: 0140202, 036 2DS2: 2DL2: 2DL3: 0010109, 006 2DL5B: 2DS3: 0010301 2DP1: 0030102, 009 2DL1: 0030208, 00303 3DP1: 001, 002 2DL4: 0010202 (2x) 3DL1: 0200101 3DS1: 0130102 2DL5A: 0010102 2DS5: 2DS1: 0020105 2DS4: 3DL2: 00202, 018 | 3DL3: 0090102, 0140203 2DS2: 0010111 2DL2: 0010105 2DL3: 0010102 2DL5B: 0080101 2DS3: 0010302 2DP1: 0020103, 010 2DL1: 0030203, 0040102 3DP1: 0030202, 0030402 2DL4: 0010303, 00901 3DL1: 0050102, 0250102 3DS1: 2DL5A: 2DS5: 2DS1: 2DS4: 0010104, 010 3DL2: 0010302, 01001 | 3DL3: 00802,0090103 2DS2: 2DL2: 2DL3: 0010103,0010108 2DL5B: 2DS3: 2DP1: 0020106, 004 2DL1: 0030204, 0030205 3DP1: 0030202 (2x) 2DL4: 0010201, 0010305 3DL1: 008, 0150203 3DS1: 2DL5A: 2DS5: 2DS1: 2DS4: 0010107, 0030104 3DL2: 0020105, 00901 | 3DL3: 0030103, 00601 2DS2: 0010103, 0010112 2DL2: 0010102, 0030101 2DL3: 2DL5B: 0070101 2DS3: 0020101 (2x), 0010302 2DP1: 0030102 2DL1: 00402 3DP1: 0030101, 005 2DL4: 00104, 0080104 3DL1: 3DS1: 0130104, 055 2DL5A: 0050102 (2x) 2DS5: 2DS1: 0020102, 0020105 2DS4: 3DL2: 0010102, 0070102 |
| Test9 | Test10 | Test11 | |
| 3DL3: 0030103,00601 2DS2: 0010103, 0010112 2DL2: 0010102, 0030101 2DL3: 2DL5B: 0070101 2DS3: 0020101 (2x) 2DP1: 0030102 2DL1: 00402 3DP1: 0030101, 005 2DL4: 00104, 0080104 3DL1: 0150208 3DS1: 0130104 2DL5A: 01201 2DS5: 2DS1: 0020102 2DS4: 0040101 3DL2: 0010102, 0070102 | 3DL3: 0030101, 0140201 2DS2: 0010104 2DL2: 0030105 2DL3: 0020101 2DL5B: 01301 2DS3: 0020102 (2x) 2DP1: 0010203, 0030101 2DL1: 0030203, 007 3DP1: 004 (2x) 2DL4: 0080104, 010 3DL1: 0150101 3DS1: 014 2DL5A: 2DS5: 2DS1: 0020104 2DS4: 3DL2: 00501 (2x) | 3DL3: 0030103, 00601 2DS2: 0010103, 0010112 2DL2: 0010102, 0030101 2DL3: 2DL5B: 2DS3: 2DP1: 0030102, 008 2DL1: 00402 (2x) 3DP1: 0030101, 005 2DL4: 00104, 0080104 3DL1: 3DS1: 0130104, 055 2DL5A: 0050102 (2x) 2DS5: 0020102, 0020103 2DS1: 0020102, 0020105 2DS4: 3DL2: 0010102, 0070102 | |

| AMALA | BOB | COX | HO301 |
|---|---|---|--|
| 3DL3: 0040201, 00802 2DS2: 0010101 2DL2: 0030102 2DL3: 0010109 2DL5B: 2DS3: 2DP1: 0020108 2DL1: 0030201 3DP1: 007, 0090101 2DL4: 0010201, 0050106 3DL1: 0150201 3DS1: 0130101 2DL5A: 00102 2DS5: 0020101 2DS1: 0020106 2DS4: 0010101 3DL2: 0020105, 0070102 | 3DL3: 00101, 019 2DS2: 0010104 2DL2: 0030101 2DL3: 0020102 2DL5B: 2DS3: 2DP1: 0030101 2DL1: 0030210 3DP1: 002, 0030203 2DL4: 0010202, 0050101 3DL1: 002 3DS1: 0130105 2DL5A: 0010101 2DS5: 0020104 2DS1: 0020101 2DS4: 0010105 3DL2: 0020101, 0070102 | 3DL3: 00102, 0090101 2DS2: 2DL2: 2DL3: 0020101, 006 2DL5B: 2DS3: 2DP1: 0030102 (2x) 2DL1: 0020102 (2x) 3DP1: 005, 006 2DL4: 0050102, 00901 3DL1: 0050103 3DS1: 055 2DL5A: 2DS5: 0020102 2DS1: 0020105 2DS4: 010 3DL2: 0010301, 0070103 | 3DL3: 00102, 0090101 2DS2: 2DL2: 2DL3: 0020101, 006 2DL5B: 2DS3: 2DP1: 0030102 (2x) 2DL1: 0020102 (2x) 3DP1: 005, 006 2DL4: 00501, 00901 3DL1: 0050103 3DS1: 055 2DL5A: 2DS5: 0020102 2DS1: 0020105 2DS4: 010 3DL2: 0010301, 0070103 |
| JVM | KAS011 | OLGA | RSH |
| 3DL3: 00801, 0140201 2DS2: 0010110 2DL2: 0030102 2DL3: 010 2DL5B: 2DS3: 2DP1: 004 2DL1: 0030203 3DP1: 001, 0030202 2DL4: 0010304, 0080101 3DL1: 0010104, 008 3DS1: 2DL5A: 2DS5: 2DS1: 2DS4: 0030103 (2x) 3DL2: 0010101, 018 | 3DL3: 0090101, 0140203 2DS2: 2DL2: 2DL3: 0020103 (2x) 2DL5B: 2DS3: 2DP1: 0020104, 0030101 2DL1: 0020101, 0030209 3DP1: 0030206, 009 2DL4: 0010301, 0050107 3DL1: 008 3DS1: 013011 2DL5A: 0010102 2DS5: 0020101 2DS1: 0020101 2DS4: 0030101 3DL2: 01001, 018 | 3DL3: 00201, 00202 2DS2: 2DL2: 2DL3: 0010105 (2x) 2DL5B: 2DS3: 2DP1: 0020105, 006 2DL1: 0030204 (2x) 3DP1: 0030201 (2x) 2DL4: 0050103, 00901 3DL1: 0010102, 0050101 3DS1: 0130107 2DL5A: 00103 2DS5: 0020103 2DS1: 0020101 2DS4: 010 3DL2: 0070101, 0070102 | 3DL3: 00202, 0040202 2DS2: 0010108 2DL2: 0030104 2DL3: 0010107 2DL5B: 004 2DS3: 2DP1: 0020110, 009 2DL1: 0030205, 01201 3DP1: 0030401, 008 2DL4: 0010307, 00901 3DL1: 0050101, 01701 3DS1: 2DL5A: 2DS5: 006 2DS1: 2DS4: 0060102 3DL2: 023, 056 |
| WT51 | | | |
| 3DL3: 0090101, 036 2DS2: 0010103 2DL2: 0010107 2DL3: 006 2DL5B: 0020103 2DS3: 0020103, 0010302 2DP1: 0010202, 004 2DL1: 01201 (2x) 3DP1: 00303, 007 2DL4: 0050105, 0050103 3DL1: 3DS1: 0130102 (2x) 2DL5A: 0010103, 0050104 2DS5: 0020101 2DS1: 0020103 (2x) 2DS4: 3DL2: 00202, 00903 | | | |

C Detailní výsledky

V další části jsou uvedeny tabulky, přičemž nyní následuje vysvětlení významu jejich jednotlivých sloupců. *Alel* u genomu značí počet v daném genomu. Číslo v závorkách udává kolik alel se v daném genomu vyskytuje dvakrát. V každém kroku *zbývá alel* udává kolik alel ještě zůstalo ve výběru, *ztraceno* určuje kolik alel má být v genomu, ale algoritmus je vyřadil. Za tímto číslem jsou vypsané alely, které byly ztraceny. V dalších krocích jsou vypsané alely bez těch, které už byly ztraceny v předchozích krocích. Podobně je to s *geny navíc*, které udávají počet a druh genů již neobsahujících žádnou z alel, která náleží do daného genomu.

C.1 Experiment1

| Genom | Alel | Krok 1 | | | Krok 2 | | | Krok 3 | | |
|--------|--------|------------|----------|---|------------|---|------------|------------|--|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 461 | 0 - | 2 2DL5B 2DS3 | 113 | 2 2DL1*0030201 3DL1*0150201 | 0 - | 23 | 4 3DP1*0090101 2DL4*0010201 | 0 - |
| bob | 19 (0) | 461 | 0 - | 2 2DL5B 2DS3 | 100 | 2 3DL1*002 2DL1*0030210 | 0 - | 28 | 2 - | 0 - |
| cox | 19 (2) | 461 | 0 - | 5 2DS3 2DL2 2DL5B 2DS2 2DL5A | 73 | 1 2DL4*00901 | 0 - | 20 | 3 3DL3*0090101 3DP1*006 | 0 - |
| ho301 | 24 (6) | 461 | 0 - | 5 2DL5A 2DS1 2DS5 3DS1 2DL3 | 80 | 0 - | 0 - | 20 | 4 3DL2*0020106 2DL1*00402 2DS3*0020103 3DL1*002 | 1 3DL1 |
| jvm | 17 (1) | 461 | 0 - | 6 2DL5B 2DS1 3DS1 2DL5A 2DS5 2DS3 | 76 | 2 2DL4*0080101 2DL1*0030203 | 0 - | 26 | 2 - | 0 - |
| kas011 | 20 (1) | 461 | 0 - | 4 2DL5B 2DS3 2DL2 2DS2 | 94 | 2 2DL4*0050107 3DL1*008 | 0 - | 29 | 3 3DL2*01001 | 0 - |
| olga | 21 (3) | 461 | 0 - | 4 2DL5B 2DS3 2DL2 2DS2 | 99 | 1 3DL1*0010102 | 0 - | 24 | 2 2DP1*0020105 | 0 - |
| rsh | 20 (0) | 461 | 0 - | 4 2DL5A 2DS1 2DS3 3DS1 | 98 | 3 2DL1*0030205 3DL1*01701 2DL4*0010307 | 0 - | 25 | 4 2DP1*0020110 | 0 - |
| wt51 | 24 (2) | 461 | 0 - | 2 3DL1 2DS4 | 118 | 0 - | 0 - | 34 | 4 3DL3*0090101 2DL5A*0010103 2DS3*0020103 2DS1*0020103 | 1 2DS1 |
| test1 | 23 (3) | 461 | 0 - | 1 2DS5 | 73 | 2 3DL1*0150101 2DL1*0030203 | 0 - | 27 | 2 - | 0 - |
| test2 | 21 (1) | 461 | 0 - | 2 2DS3 2DS2 | 82 | 1 3DL1*0070101 | 1 3DL1 | 23 | 3 2DP1*0020107 2DL1*0020102 | 1 - |

| | | | | | | | | | | | | | | | | |
|--------|--------|-----|---|---|---|--|-----|---|--|---|------|----|---|--|---|---|
| test3 | 16 (1) | 461 | 0 | - | 9 | 2DS1 2DL5A 2DL2 3DS1 2DL5B 2DS2 2DS4 2DS5 2DS3 | 69 | 3 | 3DL1*002 2DL1*008 3DL1*0040101 | 1 | 3DL1 | 17 | 4 | 2DL4*0010306 | 1 | - |
| test4 | 18 (0) | 461 | 0 | - | 3 | 2DL5B 2DS3 2DL5A | 89 | 2 | 3DL1*0150202 2DL4*0010307 | 0 | - | 22 | 3 | 3DL2*0020101 | 0 | - |
| test5 | 19 (1) | 461 | 0 | - | 5 | 2DL5B 2DL2 2DS4 2DS5 2DS2 | 83 | 1 | 3DL1*0200101 | 0 | - | 21 | 3 | 2DL3*0010109 2DL1*0030208 | 0 | - |
| test6 | 21 (0) | 461 | 0 | - | 4 | 2DL5A 2DS1 3DS1 2DS5 | 100 | 3 | 2DL1*0030203 3DL1*0250102 2DL4*00901 | 0 | - | 29 | 4 | 3DP1*0030202 | 0 | - |
| test7 | 18 (1) | 461 | 0 | - | 8 | 2DS1 3DS1 2DL2 2DL5B 2DS2 2DS3 2DL5A 2DS5 | 94 | 2 | 3DL1*0150203 2DL1*0030205 | 0 | - | 15 | 7 | 2DS4*0010107 2DL4*0010201 2DP1*0020106 3DL3*0090103 2DL3*0010103 | 0 | - |
| test8 | 24 (2) | 461 | 0 | - | 4 | 3DL1 2DS5 2DS4 2DL3 | 100 | 0 | - | 0 | - | 26 | 2 | 2DS1*0020102 3DL3*0030103 | 0 | - |
| test9 | 22 (1) | 461 | 0 | - | 2 | 2DS5 2DL3 | 99 | 2 | 3DL1*0150208 2DL4*00104 | 0 | - | 26 | 3 | 3DL2*0070102 | 0 | - |
| test10 | 21 (3) | 461 | 0 | - | 3 | 2DL5A 2DS5 2DS4 | 69 | 2 | 3DL1*0150101 2DL1*0030203 | 0 | - | 24 | 2 | - | 0 | - |
| test11 | 24 (2) | 461 | 0 | - | 5 | 2DL5B 3DL1 2DS3 2DS4 2DL3 | 110 | 0 | - | 0 | - | 28 | 0 | - | 0 | - |

Tabulka C.2: Výsledky experimentu 1 na syntetických readech. Odřezány byly alely, které měly pokrytí menší než 90 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

| Genom | Alel | Krok 1 | | | Krok 2 | | | Krok 3 | | | TP | FP | FN |
|--------|--------|------------|----------|---|------------|-------------------|------------|------------|--|------------|----|----|----|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | | | |
| amala | 19 (0) | 461 | 0 - | 2 2DL5B 2DS3 | 193 | 0 - | 0 - | 41 | 2 2DL4*0010201 3DP1*0090101 | 0 - | 17 | 24 | 2 |
| bob | 19 (0) | 461 | 0 - | 2 2DL5B 2DS3 | 207 | 0 - | 1 - | 43 | 1 3DL1*002 | 2 3DL1 | 18 | 25 | 1 |
| cox | 19 (2) | 461 | 0 - | 5 2DS3 2DL2 2DL5B 2DS2 2DL5A | 156 | 0 - | 0 - | 24 | 2 3DL3*0090101 3DP1*006 | 0 - | 15 | 9 | 2 |
| ho301 | 24 (6) | 461 | 0 - | 5 2DL5A 2DS1 2DS5 3DS1 2DL3 | 151 | 0 - | 1 - | 29 | 3 3DL2*0020106 2DL1*00402 2DS3*0020103 | 1 - | 15 | 14 | 3 |
| jvm | 17 (1) | 461 | 0 - | 6 2DL5B 2DS1 3DS1 2DL5A 2DS5 2DS3 | 177 | 0 - | 0 - | 40 | 0 - | 0 - | 16 | 24 | 0 |
| kas011 | 20 (1) | 461 | 0 - | 4 2DL5B 2DS3 2DL2 2DS2 | 204 | 0 - | 1 - | 43 | 1 3DL2*01001 | 1 - | 18 | 25 | 1 |
| olga | 21 (3) | 461 | 0 - | 4 2DL5B 2DS3 2DL2 2DS2 | 194 | 0 - | 1 - | 34 | 1 2DP1*0020105 | 1 - | 17 | 17 | 1 |
| rsh | 20 (0) | 461 | 0 - | 4 2DL5A 2DS1 2DS3 3DS1 | 228 | 0 - | 1 - | 40 | 2 2DP1*0020110 2DL1*0030205 | 1 - | 18 | 22 | 2 |
| wt51 | 24 (2) | 461 | 0 - | 2 3DL1 2DS4 | 178 | 0 - | 0 - | 42 | 4 3DL3*0090101 2DL5A*0010103 2DS3*0020103 2DS1*0020103 | 1 2DS1 | 18 | 24 | 4 |
| test1 | 23 (3) | 461 | 0 - | 1 2DS5 | 152 | 1 3DL1*0150101 | 1 3DL1 | 33 | 1 - | 1 - | 19 | 14 | 1 |
| test2 | 21 (1) | 461 | 0 - | 2 2DS3 2DS2 | 204 | 0 - | 0 - | 38 | 2 2DL1*0020102 2DP1*0020107 | 0 - | 18 | 20 | 2 |

| | | | | | | | | | | | | | | | | | | | |
|--------|--------|-----|---|---|---|--|-----|---|--------------|---|------|----|---|--|---|---|----|----|---|
| test3 | 16 (1) | 461 | 0 | - | 9 | 2DS1 2DL5A 2DL2 3DS1 2DL5B 2DS2 2DS4 2DS5 2DS3 | 163 | 0 | - | 0 | - | 28 | 1 | 2DL4*0010306 | 0 | - | 14 | 14 | 1 |
| test4 | 18 (0) | 461 | 0 | - | 3 | 2DL5B 2DS3 2DL5A | 163 | 0 | - | 0 | - | 34 | 1 | 3DL2*0020101 | 0 | - | 17 | 17 | 1 |
| test5 | 19 (1) | 461 | 0 | - | 5 | 2DL5B 2DL2 2DS4 2DS5 2DS2 | 187 | 0 | - | 1 | - | 31 | 2 | 2DL1*0030208 2DL3*0010109 | 1 | - | 16 | 15 | 2 |
| test6 | 21 (0) | 461 | 0 | - | 4 | 2DL5A 2DS1 3DS1 2DS5 | 228 | 1 | 2DL1*0030203 | 1 | - | 46 | 2 | 3DP1*0030202 | 1 | - | 19 | 27 | 2 |
| test7 | 18 (1) | 461 | 0 | - | 8 | 2DS1 3DS1 2DL2 2DL5B 2DS2 2DS3 2DL5A 2DS5 | 188 | 0 | - | 0 | - | 24 | 7 | 2DL4*0010201 3DL1*0150203 2DL3*0010103 2DP1*0020106 2DS4*0010107 3DL3*0090103 2DL1*0030205 | 0 | - | 10 | 14 | 7 |
| test8 | 24 (2) | 461 | 0 | - | 4 | 3DL1 2DS5 2DS4 2DL3 | 138 | 0 | - | 0 | - | 33 | 2 | 3DL3*0030103 2DS1*0020102 | 0 | - | 20 | 13 | 2 |
| test9 | 22 (1) | 461 | 0 | - | 2 | 2DS5 2DL3 | 145 | 1 | 3DL1*0150208 | 1 | 3DL1 | 37 | 2 | 3DL2*0070102 | 1 | - | 19 | 18 | 2 |
| test10 | 21 (3) | 461 | 0 | - | 3 | 2DL5A 2DS5 2DS4 | 145 | 1 | 3DL1*0150101 | 2 | 3DL1 | 32 | 1 | - | 2 | - | 17 | 15 | 1 |
| test11 | 24 (2) | 461 | 0 | - | 5 | 2DL5B 3DL1 2DS3 2DS4 2DL3 | 148 | 0 | - | 1 | - | 35 | 0 | - | 1 | - | 22 | 13 | 0 |

Tabulka C.4: Výsledky experimentu 1 na syntetických readech. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

C.2 Experiment2

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | | TP | FP | FN |
|--------|--------|------------|----------------|-----------------|------------|--|------------|------------|--|------------|----|----|----|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | | | |
| amala | 19 (0) | 193 | 0 - | 0 - | 38 | 2 2DL4*0010201 3DP1*0090101 | 0 - | 18 | 3 3DL2*0020105 | 0 - | 16 | 2 | 3 |
| bob | 19 (0) | 207 | 0 - | 1 2DL5B | 43 | 2 2DL4*0010202 3DL1*002 | 2 3DL1 | 25 | 2 - | 2 - | 17 | 8 | 2 |
| cox | 19 (2) | 156 | 0 - | 0 - | 24 | 1 3DL3*0090101 | 0 - | 22 | 1 - | 0 - | 16 | 6 | 1 |
| ho301 | 24 (6) | 151 | 0 - | 1 2DL5A | 29 | 2 2DL1*00402 3DL2*0020106 | 1 - | 18 | 3 2DS2*0010104 | 1 - | 15 | 3 | 3 |
| jvm | 17 (1) | 177 | 0 - | 0 - | 40 | 0 - | 0 - | 17 | 3 3DL3*00801 3DP1*0030202 2DL4*0080101 | 0 - | 13 | 4 | 3 |
| kas011 | 20 (1) | 204 | 0 - | 1 2DL5B | 45 | 0 - | 1 - | 26 | 1 3DL3*0090101 | 1 - | 18 | 8 | 1 |
| olga | 21 (3) | 194 | 0 - | 1 2DL5B | 36 | 1 2DP1*0020105 | 1 - | 20 | 2 3DL1*0050101 | 1 - | 16 | 4 | 2 |
| rsh | 20 (0) | 228 | 0 - | 1 2DL5A | 44 | 2 2DP1*0020110 2DL1*0030205 | 1 - | 20 | 4 3DL3*0040202 3DL1*0050101 | 1 - | 16 | 4 | 4 |
| wt51 | 24 (2) | 178 | 0 - | 0 - | 43 | 2 2DS3*0020103 3DL3*0090101 | 0 - | 24 | 2 - | 0 - | 20 | 4 | 2 |
| test1 | 23 (3) | 152 | 1 3DL1*0150101 | 1 3DL1 | 33 | 1 - | 1 - | 23 | 1 - | 1 - | 19 | 4 | 1 |
| test2 | 21 (1) | 204 | 0 - | 0 - | 38 | 2 2DL1*0020102 2DP1*0020107 | 0 - | 23 | 3 3DL3*019 | 0 - | 17 | 6 | 3 |
| test3 | 16 (1) | 163 | 0 - | 0 - | 28 | 2 2DL1*0040101 2DL4*0010306 | 0 - | 15 | 2 - | 0 - | 13 | 2 | 2 |
| test4 | 18 (0) | 163 | 0 - | 0 - | 35 | 1 3DL2*0020101 2DP1*0030102 | 0 - | 18 | 2 3DL3*007 | 0 - | 16 | 2 | 2 |
| test5 | 19 (1) | 187 | 0 - | 1 2DL5B | 31 | 3 2DL3*0010109 2DL1*0030208 | 1 - | 20 | 3 - | 1 - | 15 | 5 | 3 |
| test6 | 21 (0) | 228 | 1 2DL1*0030203 | 1 2DL5A | 48 | 2 3DP1*0030202 | 1 - | 26 | 4 3DL3*0140203 2DP1*0020103 | 1 - | 17 | 9 | 4 |
| test7 | 18 (1) | 188 | 0 - | 0 - | 24 | 7 2DL1*0030205 2DL4*0010201 3DL1*0150203 2DP1*0020106 2DL3*0010103 2DS4*0010107 3DL3*0090103 | 0 - | 13 | 8 3DL2*0020105 | 0 - | 9 | 4 | 8 |
| test8 | 24 (2) | 138 | 0 - | 0 - | 35 | 1 2DS1*0020102 | 0 - | 24 | 2 3DL2*0070102 | 0 - | 20 | 4 | 2 |
| test9 | 22 (1) | 145 | 1 3DL1*0150208 | 1 3DL1 | 37 | 1 - | 1 - | 24 | 1 - | 1 - | 20 | 4 | 1 |
| test10 | 21 (3) | 145 | 1 3DL1*0150101 | 2 3DL1 2DL5A | 33 | 1 - | 2 - | 22 | 1 - | 2 - | 17 | 5 | 1 |

| | | | | | | | | | | | | | | | | | | | |
|--------|--------|-----|---|---|---|-------|----|---|--------------|---|---|----|---|---|---|---|----|---|---|
| test11 | 24 (2) | 148 | 0 | - | 1 | 2DL5B | 33 | 2 | 2DS5*0020103 | 1 | - | 24 | 2 | - | 1 | - | 20 | 4 | 2 |
| | | | | | | | | | 2DS2*0010103 | | | | | | | | | | |

Tabulka C.6: Výsledky experimentu 2. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

C.3 Experiment3

86

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|----------|------------|------------|--|------------|------------|--|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 231 | 0 - | 0 - | 41 | 2 3DP1*0090101 2DL4*0010201 | 0 - | 20 | 2 - | 0 - |
| bob | 19 (0) | 228 | 0 - | 1 2DL5B | 41 | 1 3DL1*002 | 2 3DL1 | 24 | 2 2DL4*0050101 | 2 - |
| cox | 19 (2) | 165 | 0 - | 0 - | 25 | 2 3DP1*006 3DL3*0090101 | 0 - | 20 | 2 - | 0 - |
| ho301 | 24 (6) | 162 | 0 - | 1 2DL5A | 29 | 2 2DL1*00402 3DL2*0020106 | 1 - | 18 | 3 2DS2*0010104 | 1 - |
| jvm | 17 (1) | 199 | 0 - | 0 - | 38 | 0 - | 0 - | 17 | 3 2DL4*0080101 3DL3*00801 3DP1*0030202 | 0 - |
| kas011 | 20 (1) | 213 | 0 - | 1 2DL5B | 43 | 1 2DP1*0020104 | 1 - | 24 | 2 3DL3*0090101 | 1 - |
| olga | 21 (3) | 205 | 0 - | 1 2DL5B | 34 | 2 3DL2*0070102 2DP1*0020105 | 1 - | 19 | 2 - | 1 - |
| rsh | 20 (0) | 250 | 0 - | 1 2DL5A | 42 | 2 2DP1*0020110 2DL1*0030205 | 1 - | 20 | 4 3DL1*0050101 3DL3*0040202 | 1 - |
| wt51 | 24 (2) | 208 | 0 - | 0 - | 45 | 3 2DL5A*0010103 2DS3*0020103 3DL3*0090101 | 0 - | 26 | 3 - | 0 - |
| test1 | 23 (3) | 174 | 0 - | 0 - | 35 | 0 - | 0 - | 24 | 0 - | 0 - |
| test2 | 21 (1) | 208 | 0 - | 0 - | 37 | 2 2DL1*0020102 2DP1*0020107 | 0 - | 23 | 3 3DL3*019 | 0 - |
| test3 | 16 (1) | 181 | 0 - | 0 - | 28 | 1 2DL4*0010306 | 0 - | 15 | 2 2DL1*0040101 | 0 - |
| test4 | 18 (0) | 183 | 0 - | 0 - | 35 | 2 2DS1*0020101 3DL2*0020101 | 1 2DS1 | 17 | 2 - | 1 - |
| test5 | 19 (1) | 198 | 0 - | 1 2DL5B | 32 | 3 2DL1*0030208 2DL3*0010109 2DP1*0030102 | 1 - | 20 | 3 - | 1 - |
| test6 | 21 (0) | 279 | 0 - | 1 2DL5A | 51 | 1 3DP1*0030202 | 1 - | 27 | 3 3DL3*0140203 2DP1*0020103 | 1 - |
| test7 | 18 (1) | 193 | 0 - | 0 - | 23 | 7 3DL3*0090103 2DS4*0010107 2DL1*0030205 2DL3*0010103 2DP1*0020106 2DL4*0010201 3DL1*0150203 | 0 - | 13 | 8 3DL2*0020105 | 0 - |
| test8 | 24 (2) | 150 | 0 - | 0 - | 36 | 1 2DS1*0020102 | 0 - | 25 | 2 3DL2*0070102 | 0 - |
| test9 | 22 (1) | 174 | 0 - | 0 - | 37 | 0 - | 0 - | 22 | 1 3DL2*0070102 | 0 - |

| | | | | | | | | | | | | | | | | |
|--------|--------|-----|---|--------------|---|---------------|----|---|------------------------------|---|---|----|---|---|---|---|
| test10 | 21 (3) | 162 | 1 | 3DL1*0150101 | 2 | 3DL1 2DL5A | 34 | 1 | - | 2 | - | 23 | 1 | - | 2 | - |
| test11 | 24 (2) | 148 | 0 | - | 1 | 2DL5B | 33 | 2 | 2DS5*0020103 2DS2*0010103 | 1 | - | 24 | 2 | - | 1 | - |

Tabulka C.8: Výsledky experimentu 3. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. Shluky vytvářely alely, které od sebe měly vzdálenost maximálně 5.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|----------|------------|------------|--|------------|------------|--|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 237 | 0 - | 0 - | 37 | 2 3DP1*0090101 2DL4*0010201 | 0 - | 16 | 3 3DL2*0020105 | 0 - |
| bob | 19 (0) | 242 | 0 - | 1 2DL5B | 43 | 0 - | 1 - | 24 | 1 2DL4*0050101 | 1 - |
| cox | 19 (2) | 175 | 0 - | 0 - | 23 | 2 3DP1*006 3DL3*0090101 | 0 - | 19 | 3 3DL2*0070103 | 0 - |
| ho301 | 24 (6) | 174 | 0 - | 1 2DL5A | 28 | 2 3DL2*0020106 2DL1*00402 | 1 - | 18 | 3 2DS2*0010104 | 1 - |
| jvm | 17 (1) | 210 | 0 - | 0 - | 36 | 0 - | 0 - | 17 | 3 2DL4*0080101 3DL3*00801 3DP1*0030202 | 0 - |
| kas011 | 20 (1) | 220 | 0 - | 1 2DL5B | 44 | 0 - | 1 - | 25 | 1 3DL3*0090101 | 1 - |
| olga | 21 (3) | 212 | 0 - | 1 2DL5B | 35 | 1 2DP1*0020105 | 1 - | 20 | 2 3DL1*0050101 | 1 - |
| rsh | 20 (0) | 262 | 0 - | 1 2DL5A | 43 | 2 2DL1*0030205 2DP1*0020110 | 1 - | 20 | 4 3DL3*0040202 3DL1*0050101 | 1 - |
| wt51 | 24 (2) | 222 | 0 - | 0 - | 45 | 3 2DS3*0020103 2DL5A*0010103 3DL3*0090101 | 0 - | 26 | 3 - | 0 - |
| test1 | 23 (3) | 188 | 0 - | 0 - | 34 | 0 - | 0 - | 23 | 1 3DL3*0030101 | 0 - |
| test2 | 21 (1) | 209 | 0 - | 0 - | 37 | 2 2DL1*0020102 2DP1*0020107 | 0 - | 24 | 3 3DL3*019 | 0 - |
| test3 | 16 (1) | 191 | 0 - | 0 - | 28 | 1 2DL4*0010306 | 0 - | 13 | 3 3DL1*002 2DL1*0040101 | 0 - |
| test4 | 18 (0) | 192 | 0 - | 0 - | 39 | 1 3DL2*0020101 2DL1*0030208 | 0 - | 17 | 2 3DL3*007 | 0 - |
| test5 | 19 (1) | 208 | 0 - | 1 2DL5B | 32 | 3 2DP1*0030102 2DL3*0010109 | 1 - | 19 | 3 - | 1 - |
| test6 | 21 (0) | 290 | 0 - | 1 2DL5A | 52 | 1 3DP1*0030202 | 1 - | 25 | 3 2DP1*0020103 3DL3*0140203 | 1 - |
| test7 | 18 (1) | 196 | 0 - | 0 - | 26 | 7 3DL1*0150203 2DL1*0030205 3DL3*0090103 2DP1*0020106 2DS4*0010107 2DL4*0010201 2DL3*0010103 | 0 - | 14 | 8 3DL2*0020105 | 0 - |
| test8 | 24 (2) | 158 | 0 - | 0 - | 36 | 1 2DS1*0020102 | 0 - | 25 | 2 3DL2*0070102 | 0 - |
| test9 | 22 (1) | 189 | 0 - | 0 - | 39 | 0 - | 0 - | 24 | 0 - | 0 - |
| test10 | 21 (3) | 174 | 0 - | 1 2DL5A | 31 | 0 - | 1 - | 21 | 0 - | 1 - |
| test11 | 24 (2) | 156 | 0 - | 1 2DL5B | 33 | 2 2DS5*0020103 2DS2*0010103 | 1 - | 24 | 2 - | 1 - |

Tabulka C.10: Výsledky experimentu 3. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. Shluky vytvářely alely, které od sebe měly vzdálenost maximálně 10.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | | TP | FP | FN |
|--------|--------|------------|----------|------------|------------|---|------------|------------|---|------------|----|----|----|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | | | |
| amala | 19 (0) | 258 | 0 - | 0 - | 40 | 2 2DL4*0010201 3DP1*0090101 | 0 - | 20 | 3 3DL2*0020105 | 0 - | 16 | 4 | 3 |
| bob | 19 (0) | 252 | 0 - | 1 2DL5B | 45 | 0 - | 1 - | 25 | 1 2DL4*0050101 | 1 - | 18 | 7 | 1 |
| cox | 19 (2) | 177 | 0 - | 0 - | 23 | 2 3DP1*006 3DL3*0090101 | 0 - | 21 | 2 - | 0 - | 15 | 6 | 2 |
| ho301 | 24 (6) | 180 | 0 - | 1 2DL5A | 30 | 2 2DL1*00402 3DL2*0020106 | 1 - | 18 | 3 2DS2*0010104 | 1 - | 15 | 3 | 3 |
| jvm | 17 (1) | 227 | 0 - | 0 - | 36 | 0 - | 0 - | 17 | 3 3DP1*0030202 3DL3*00801 2DL4*0080101 | 0 - | 13 | 4 | 3 |
| kas011 | 20 (1) | 242 | 0 - | 1 2DL5B | 44 | 0 - | 1 - | 26 | 1 3DL3*0090101 | 1 - | 18 | 8 | 1 |
| olga | 21 (3) | 213 | 0 - | 1 2DL5B | 38 | 1 2DP1*0020105 | 1 - | 20 | 2 3DL1*0050101 | 1 - | 16 | 4 | 2 |
| rsh | 20 (0) | 270 | 0 - | 1 2DL5A | 44 | 2 2DL1*0030205 2DP1*0020110 | 1 - | 20 | 4 3DL1*0050101 3DL3*0040202 | 1 - | 16 | 4 | 4 |
| wt51 | 24 (2) | 223 | 0 - | 0 - | 47 | 2 3DL3*0090101 2DS3*0020103 | 0 - | 29 | 2 - | 0 - | 20 | 9 | 2 |
| test1 | 23 (3) | 200 | 0 - | 0 - | 35 | 0 - | 0 - | 24 | 1 3DL3*0030101 | 0 - | 19 | 5 | 1 |
| test2 | 21 (1) | 212 | 0 - | 0 - | 39 | 2 2DL1*0020102 2DP1*0020107 | 0 - | 26 | 2 - | 0 - | 18 | 8 | 2 |
| test3 | 16 (1) | 199 | 0 - | 0 - | 30 | 1 2DL4*0010306 | 0 - | 15 | 2 2DL1*0040101 | 0 - | 13 | 2 | 2 |
| test4 | 18 (0) | 215 | 0 - | 0 - | 34 | 2 2DS1*0020101 3DL2*0020101 | 1 2DS1 | 17 | 2 - | 1 - | 16 | 1 | 2 |
| test5 | 19 (1) | 218 | 0 - | 1 2DL5B | 30 | 3 2DL3*0010109 2DL1*0030208 2DP1*0030102 | 1 - | 19 | 3 - | 1 - | 15 | 4 | 3 |
| test6 | 21 (0) | 311 | 0 - | 1 2DL5A | 48 | 1 3DP1*0030202 2DL1*0030205 3DL1*0150203 2DP1*0020106 | 1 - | 26 | 2 2DP1*0020103 | 1 - | 19 | 7 | 2 |
| test7 | 18 (1) | 201 | 0 - | 0 - | 24 | 7 2DS4*0010107 2DL3*0010103 2DL4*0010201 3DL3*0090103 | 0 - | 14 | 8 3DL2*0020105 | 0 - | 9 | 5 | 8 |
| test8 | 24 (2) | 160 | 0 - | 0 - | 37 | 1 2DS1*0020102 | 0 - | 24 | 2 3DL2*0070102 | 0 - | 20 | 4 | 2 |
| test9 | 22 (1) | 204 | 0 - | 0 - | 38 | 0 - | 0 - | 23 | 0 - | 0 - | 21 | 2 | 0 |
| test10 | 21 (3) | 184 | 0 - | 1 2DL5A | 31 | 0 - | 1 - | 20 | 1 3DL3*0030101 | 1 - | 17 | 3 | 1 |
| test11 | 24 (2) | 156 | 0 - | 1 2DL5B | 33 | 2 2DS2*0010103 2DS5*0020103 | 1 - | 24 | 2 - | 1 - | 20 | 4 | 2 |

Tabulka C.12: Výsledky experimentu 3. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. Shluky vytvářely alely, které od sebe měly vzdálenost maximálně 20.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|----------|------------|------------|---|------------|------------|---|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 281 | 0 - | 0 - | 39 | 2 2DL4*0010201 3DP1*0090101 | 0 - | 19 | 3 3DL2*0020105 | 0 - |
| bob | 19 (0) | 276 | 0 - | 1 2DL5B | 41 | 0 - | 1 - | 23 | 2 3DL1*002 2DL4*0050101 | 2 3DL1 |
| cox | 19 (2) | 223 | 0 - | 0 - | 26 | 2 3DP1*006 3DL3*0090101 | 0 - | 22 | 3 3DL2*0070103 | 0 - |
| ho301 | 24 (6) | 181 | 0 - | 1 2DL5A | 28 | 3 2DL2*0010103 3DL2*0020106 2DL1*00402 | 1 - | 17 | 4 2DS2*0010104 | 1 - |
| jvm | 17 (1) | 268 | 0 - | 0 - | 36 | 0 - | 0 - | 17 | 3 3DP1*0030202 2DL4*0080101 3DL3*00801 | 0 - |
| kas011 | 20 (1) | 259 | 0 - | 1 2DL5B | 43 | 0 - | 1 - | 26 | 1 3DL3*0090101 | 1 - |
| olga | 21 (3) | 249 | 0 - | 1 2DL5B | 35 | 1 2DP1*0020105 | 1 - | 19 | 2 3DL1*0050101 | 1 - |
| rsh | 20 (0) | 285 | 0 - | 1 2DL5A | 45 | 2 2DL1*0030205 2DP1*0020110 | 1 - | 19 | 5 3DL1*0050101 3DL3*0040202 3DL2*023 | 1 - |
| wt51 | 24 (2) | 223 | 0 - | 0 - | 47 | 2 2DS3*0020103 3DL3*0090101 | 0 - | 29 | 2 - | 0 - |
| test1 | 23 (3) | 227 | 0 - | 0 - | 35 | 0 - | 0 - | 24 | 1 3DL3*0030101 | 0 - |
| test2 | 21 (1) | 218 | 0 - | 0 - | 39 | 2 2DP1*0020107 2DL1*0020102 | 0 - | 23 | 3 3DL3*019 | 0 - |
| test3 | 16 (1) | 238 | 0 - | 0 - | 28 | 1 2DL4*0010306 | 0 - | 12 | 3 3DL1*002 2DL1*0040101 | 0 - |
| test4 | 18 (0) | 218 | 0 - | 0 - | 35 | 2 3DL2*0020101 2DS1*0020101 | 1 2DS1 | 17 | 2 - | 1 - |
| test5 | 19 (1) | 219 | 0 - | 1 2DL5B | 32 | 3 2DP1*0030102 2DL3*0010109 2DL1*0030208 | 1 - | 19 | 3 - | 1 - |
| test6 | 21 (0) | 337 | 0 - | 1 2DL5A | 47 | 1 3DP1*0030202 | 1 - | 27 | 2 2DP1*0020103 | 1 - |
| test7 | 18 (1) | 232 | 0 - | 0 - | 24 | 7 2DP1*0020106 2DL1*0030205 2DL3*0010103 3DL1*0150203 2DL4*0010201 2DS4*0010107 3DL3*0090103 | 0 - | 14 | 8 3DL2*0020105 | 0 - |
| test8 | 24 (2) | 163 | 0 - | 0 - | 36 | 1 2DS1*0020102 | 0 - | 24 | 2 3DL2*0070102 | 0 - |
| test9 | 22 (1) | 210 | 0 - | 0 - | 38 | 0 - | 0 - | 23 | 0 - | 0 - |
| test10 | 21 (3) | 216 | 0 - | 1 2DL5A | 32 | 0 - | 1 - | 20 | 1 3DL3*0030101 | 1 - |
| test11 | 24 (2) | 159 | 0 - | 1 2DL5B | 35 | 0 - | 1 - | 26 | 0 - | 1 - |

Tabulka C.14: Výsledky experimentu 3. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. Shluky vytvářely alely, které od sebe měly vzdálenost maximálně 30.

Ukázky vzniklých clusterů

Níže jsou uvedeny největší clustery.

Při vzdálenosti 5 bylo vytvořeno kolem 224 clusterů.

| | | | | |
|--|---|---|---|---|
| 28 2DL1*0030226 2DL1*0030219 2DL1*0030230 2DL1*0030212 2DL1*0030229 2DL1*037 2DL1*0030210 2DL1*0030208 2DL1*0030211 2DL1*0030205 2DL1*025 2DL1*032N 2DL1*0030214 2DL1*0030221 2DL1*0030216 2DL1*0030213 2DL1*0030228 2DL1*0030218 2DL1*0030209 2DL1*0030204 2DL1*0030227 2DL1*0030215 2DL1*0030217 2DL1*0030203 2DL1*0030224 2DL1*0030202 2DL1*0030231 2DL1*0030223 | 17 2DL1*0020112 2DL1*0020108 2DL1*0020114 2DL1*0020113 2DL1*0020109 2DL1*0020110 2DL1*0020103 2DL1*0020102 2DL1*0020115 2DL1*0020106 2DL1*0020101 2DL1*0020104 2DL1*0020107 2DL1*0020111 2DL1*0020116 2DL1*008 2DL1*0020105 | 15 2DL1*0040114 2DL1*00402 2DL1*0040107 2DL1*0040106 2DL1*0040109 2DL1*0040110 2DL1*0040101 2DL1*0040113 2DL1*0040105 2DL1*0040111 2DL1*0040104 2DL1*0040103 2DL1*0040108 2DL1*007 2DL1*0040115 | 11 2DP1*0020107 2DP1*0020105 2DP1*0020103 2DP1*0020109 2DP1*0020106 2DP1*0020108 2DP1*0020104 2DP1*0020102 2DP1*0020110 2DP1*0020101 2DP1*008 | 11 3DL1*0020103 3DL1*0020102 3DL1*0150215 3DL1*0020104 3DL1*0150214 3DL1*0020105 3DL1*0150216 3DL1*1190101 3DL1*1190102 3DL1*0150217 3DL1*0150218 |
| 10 2DL4*043 2DL4*046 2DL4*0080203 2DL4*0080206 2DL4*0080207 2DL4*052 2DL4*00805 2DL4*0080204 2DL4*0080208 2DL4*0080205 | 10 2DL4*0080101 2DL4*0080107 2DL4*00803 2DL4*0080402 2DL4*0080401 2DL4*0080108 2DL4*051 2DL4*053 2DL4*0080106 2DL4*0080105 | 9 3DP1*0030203 3DP1*0030202 3DP1*0030401 3DP1*00303 3DP1*0030206 3DP1*0030201 3DP1*008 3DP1*0030205 3DP1*0030204 | 9 2DL4*050 2DL4*0010202 2DL4*00107 2DL4*0010203 2DL4*054 2DL4*0010303 2DL4*0010201 2DL4*042 2DL4*0010306 | 8 2DL4*0050105 2DL4*0050101 2DL4*049 2DL4*0050106 2DL4*0050107 2DL4*0050102 2DL4*0050103 2DL4*0050104 |

Při vzdálenosti 10 bylo vytvořeno kolem 1771 clusterů.

| | | | | |
|--|---|---|---|---|
| 33 2DL1*0030206 2DL1*0030214 2DL1*0030203 2DL1*0030219 2DL1*0030202 2DL1*0030210 2DL1*032N 2DL1*0030212 2DL1*0030208 2DL1*0030211 2DL1*0030216 2DL1*0030221 2DL1*0030207 2DL1*0030222 2DL1*0030229 2DL1*025 2DL1*0030218 2DL1*0030205 2DL1*0030230 2DL1*0030220 2DL1*0030231 2DL1*0030227 2DL1*0030225 2DL1*0030215 2DL1*0030217 2DL1*037 2DL1*0030228 2DL1*0030226 2DL1*0030224 2DL1*0030223 2DL1*0030204 2DL1*0030213 2DL1*0030209 | 17 2DL1*0020115 2DL1*0020103 2DL1*0020104 2DL1*0020105 2DL1*008 2DL1*0020111 2DL1*0020112 2DL1*0020106 2DL1*0020102 2DL1*0020107 2DL1*0020116 2DL1*0020108 2DL1*0020109 2DL1*0020113 2DL1*0020101 2DL1*0020114 2DL1*0020110 | 15 2DL1*0040103 2DL1*0040110 2DL1*0040109 2DL1*0040101 2DL1*0040111 2DL1*0040106 2DL1*0040114 2DL1*0040113 2DL1*0040105 2DL1*0040108 2DL1*007 2DL1*00402 2DL1*0040104 2DL1*0040107 2DL1*0040115 | 14 2DL5A*0050102 2DL5A*01201 2DL5A*01202 2DL5B*0020102 2DL5B*0020104 2DL5B*0020106 2DL5B*0020201 2DL5B*00603 2DL5B*0070102 2DL5B*0080102 2DL5B*009 2DL5B*01301 2DL5B*01302 2DL5B*01303 | 13 3DP1*0030202 3DP1*0030203 3DP1*0030206 3DP1*0030201 3DP1*0030204 3DP1*00303 3DP1*008 3DP1*0030205 3DP1*0030401 3DP1*01001 3DP1*0030402 3DP1*005 3DP1*006 |
| 11 2DL4*0010309 2DL4*0010305 2DL4*045 2DL4*0010307 2DL4*042 2DL4*050 2DL4*0010306 2DL4*0010203 2DL4*0010202 2DL4*0010303 2DL4*00107 | 11 3DL1*0150103 3DL1*0150101 3DL1*0200101 3DL1*0150102 3DL1*0150204 3DL1*077 3DL1*0200102 3DL1*0250103 3DL1*0150208 3DL1*0150210 3DL1*0150205 | 11 3DL1*0150216 3DL1*0150218 3DL1*0150215 3DL1*0020102 3DL1*0020103 3DL1*0020104 3DL1*1190101 3DL1*1190102 3DL1*0020105 3DL1*0150217 3DL1*0150214 | 11 2DP1*0020104 2DP1*0020109 2DP1*0020103 2DP1*0020107 2DP1*0020106 2DP1*0020101 2DP1*0020105 2DP1*0020102 2DP1*008 2DP1*0020108 2DP1*0020110 | 10 2DL4*0080402 2DL4*0080105 2DL4*0080401 2DL4*00803 2DL4*053 2DL4*0080106 2DL4*0080107 2DL4*051 2DL4*0080101 2DL4*0080108 |

Při vzdálenosti 30 bylo vytvořeno 122 clusterů. Zde je uveden jen jeden ze zajímavějších, jelikož obsahuje jak alely genu 2DL5A, tak 2DL5B.

| |
|---|
| 14 2DL5A*01201 2DL5B*0080102 2DL5A*01202 2DL5B*00603 2DL5B*0070102 2DL5B*0020104 2DL5B*01303 2DL5B*01301 2DL5B*0020201 2DL5B*0020106 2DL5B*0020102 2DL5A*0050102 2DL5B*01302 2DL5B*009 |
|---|

C.4 Výsledky verifikace

| Genom | Alel | Krok 2 | | | Krok 3 | | |
|--------|--------|------------|--|--|------------|---|-------------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 344 | 2 2DL3*001 2DL2*00301 | 1 2DL5B | 78 | 5 3DP1*00901 3DL2*0020105 3DL2*0070102 | 2 3DL2 |
| bob | 19 (0) | 366 | 3 3DL3*01303 2DL3*00201 2DL2*00301 | 1 2DL5B | 86 | 5 3DP1*00302 3DL2*0070102 | 1 - |
| cox | 18 (0) | 329 | 3 3DL3*00103 2DL3*007 2DL3*00201 | 1 2DL5B | 68 | 4 3DP1*006 | 1 - |
| ho301 | 19 (2) | 296 | 4 2DL5B*010 2DL2*00101 2DL2*00301 2DS2*002 | 4 2DL5B 2DL5A 2DS1 2DS5 | 65 | 6 2DL1*010 2DS3*00201 | 4 - |
| jvm | 15 (0) | 315 | 2 2DP1*005 2DL2*00301 | 1 2DP1 | 52 | 3 3DL3*00801 | 1 - |
| kas011 | 18 (0) | 311 | 2 3DL3*01302 3DL2*019 | 1 2DL5B | 70 | 5 2DP1*002 3DL3*00901 3DP1*00302 | 2 3DL3 |
| olga | 17 (0) | 328 | 2 2DL3*00101 3DL3*00902 | 1 2DL5B | 74 | 7 3DL2*00701 2DL4*011 3DL1*001 2DP1*006 3DP1*00302 | 3 3DL2 3DP1 |
| rsh | 15 (0) | 361 | 0 - | 5 2DS4 3DL2 2DL5A 2DS1 2DS3 | 73 | 1 2DP1*00201 | 5 - |
| wt51 | 12 (0) | 297 | 1 3DL3*00103 | 4 3DP1 2DS3 2DL1 3DL2 | 66 | 2 2DL5A*00501 | 4 - |

Tabulka C.16: Výsledky experimentu 1 na reálných datech. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

| Genom | Alel | Krok 2 | | | Krok 3 | | |
|--------|--------|------------|--|-------------------------------------|------------|---|----------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 345 | 2 2DL3*001 2DL2*00301 | 1 2DL5B | 79 | 4 3DP1*00901 3DL2*0070102 | 1 - |
| bob | 19 (0) | 369 | 3 3DL3*01303 2DL3*00201 2DL2*00301 | 2 2DL5B 2DS3 | 87 | 5 3DP1*00302 3DL2*0070102 | 2 - |
| cox | 18 (0) | 334 | 3 2DL3*007 3DL3*00103 2DL3*00201 | 1 2DL5B | 69 | 4 3DP1*006 | 1 - |
| ho301 | 19 (2) | 298 | 4 2DS2*002 2DL5B*010 2DL2*00101 2DL2*00301 | 4 2DL5B 2DL5A 2DS1 2DS5 | 65 | 6 2DL1*010 2DS3*00201 | 4 - |
| jvm | 15 (0) | 316 | 2 2DP1*005 2DL2*00301 | 1 2DP1 | 49 | 3 3DL3*00801 | 1 - |
| kas011 | 18 (0) | 306 | 2 3DL3*01302 3DL2*019 | 1 2DL5B | 64 | 5 3DL3*00901 2DP1*002 3DP1*00302 | 2 3DL3 |
| olga | 17 (0) | 332 | 2 2DL3*00101 3DL3*00902 | 1 2DL5B | 77 | 7 3DL2*00701 3DL1*001 2DP1*006 2DL4*011 3DP1*00302 | 3 3DL2 3DP1 |
| rsh | 15 (0) | 361 | 0 - | 4 2DS4 3DL2 2DL5A 2DS1 | 73 | 1 2DP1*00201 | 4 - |
| wt51 | 12 (0) | 300 | 1 3DL3*00103 | 4 3DP1 2DS3 2DL1 3DL2 | 69 | 1 - | 4 - |

Tabulka C.18: Výsledky experimentu 1 na reálných datech po odstržení nekvalitních bází. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|--|--|------------|---|-------------------|------------|---|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 344 | 2 2DL2*00301 2DL3*001 | 1 2DL5B | 80 | 5 3DL2*0070102 3DL2*0020105 3DP1*00901 | 2 3DL2 | 60 | 5 - | 2 - |
| bob | 19 (0) | 366 | 3 3DL3*01303 2DL2*00301 2DL3*00201 | 1 2DL5B | 87 | 5 3DL2*0070102 3DP1*00302 | 1 - | 52 | 7 2DL4*001 2DL1*00302 | 2 2DL1 |
| cox | 18 (0) | 329 | 3 2DL3*007 2DL3*00201 3DL3*00103 | 1 2DL5B | 65 | 4 3DP1*006 | 1 - | 30 | 4 - | 1 - |
| ho301 | 19 (2) | 296 | 4 2DL2*00101 2DL5B*010 2DL2*00301 2DS2*002 | 4 2DL5B 2DL5A 2DS1 2DS5 | 66 | 6 2DL1*010 2DS3*00201 | 4 - | 37 | 8 3DL3*014 3DP1*004 | 5 3DL3 |
| jvm | 15 (0) | 315 | 2 2DL2*00301 2DP1*005 | 1 2DP1 | 54 | 3 3DL3*00801 | 1 - | 19 | 6 2DL4*00801 3DP1*00302 3DL1*00101 | 1 - |
| kas011 | 18 (0) | 311 | 2 3DL2*019 3DL3*01302 | 1 2DL5B | 68 | 6 3DL3*00901 3DP1*00302 2DP1*002 3DL2*01001 | 3 3DL3 3DL2 | 37 | 7 2DL1*00302 | 3 - |
| olga | 17 (0) | 328 | 2 2DL3*00101 3DL3*00902 | 1 2DL5B | 74 | 7 3DL1*001 2DL4*011 2DP1*006 3DP1*00302 3DL2*00701 | 3 3DL2 3DP1 | 40 | 8 3DL3*00201 | 4 3DL3 |
| rsh | 15 (0) | 361 | 0 - | 5 2DS4 3DL2 2DL5A 2DS1 2DS3 | 74 | 1 2DP1*00201 | 5 - | 34 | 3 2DL1*00302 2DL4*011 | 5 - |
| wt51 | 12 (0) | 297 | 1 3DL3*00103 | 4 2DS3 2DL1 3DP1 3DL2 | 68 | 1 - | 4 - | 29 | 3 2DL5A*00501 2DP1*001 | 4 - |

Tabulka C.20: Výsledky experimentu 2 na reálných datech. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|--|-------------------------------------|------------|---|-------------------|------------|---|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 345 | 2 2DL3*001 2DL2*00301 | 1 2DL5B | 81 | 4 3DL2*0070102 3DP1*00901 | 1 - | 52 | 4 - | 1 - |
| bob | 19 (0) | 369 | 3 2DL2*00301 2DL3*00201 3DL3*01303 | 2 2DL5B 2DS3 | 88 | 5 3DL2*0070102 3DP1*00302 | 2 - | 51 | 7 2DL4*001 2DL1*00302 | 3 2DL1 |
| cox | 18 (0) | 334 | 3 2DL3*007 2DL3*00201 3DL3*00103 | 1 2DL5B | 71 | 4 3DP1*006 | 1 - | 33 | 4 - | 1 - |
| ho301 | 19 (2) | 298 | 4 2DL2*00301 2DS2*002 2DL5B*010 2DL2*00101 | 4 2DL5B 2DL5A 2DS1 2DS5 | 68 | 6 2DS3*00201 2DL1*010 | 4 - | 39 | 8 3DP1*004 3DL3*014 | 5 3DL3 |
| jvm | 15 (0) | 316 | 2 2DL2*00301 2DP1*005 | 1 2DP1 | 52 | 3 3DL3*00801 | 1 - | 22 | 6 2DL4*00801 3DP1*00302 3DL1*00101 | 1 - |
| kas011 | 18 (0) | 306 | 2 3DL3*01302 3DL2*019 | 1 2DL5B | 66 | 5 3DL3*00901 3DP1*00302 2DP1*002 | 2 3DL3 | 35 | 8 2DL4*00103 2DL1*00302 3DL2*01001 | 3 3DL2 |
| olga | 17 (0) | 332 | 2 3DL3*00902 2DL3*00101 | 1 2DL5B | 76 | 7 3DL1*001 2DL4*011 3DL2*00701 3DP1*00302 2DP1*006 | 3 3DL2 3DP1 | 47 | 8 3DL3*00201 | 4 3DL3 |
| rsh | 15 (0) | 361 | 0 - | 4 2DL5A 2DS4 3DL2 2DS1 | 71 | 1 2DP1*00201 | 4 - | 33 | 3 2DL1*00302 2DL4*011 | 4 - |
| wt51 | 12 (0) | 300 | 1 3DL3*00103 | 4 3DP1 2DL1 2DS3 3DL2 | 69 | 1 - | 4 - | 30 | 3 2DL5A*00501 2DP1*001 | 4 - |

Tabulka C.22: Výsledky experimentu 2 na reálných datech po odřezání nekvalitních bází. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|--|--|------------|---|-------------------|------------|---|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 374 | 2 2DL3*001 2DL2*00301 | 1 2DL5B | 80 | 5 3DP1*00901 3DL2*0070102 3DL2*0020105 | 2 3DL2 | 58 | 5 - | 2 - |
| bob | 19 (0) | 386 | 3 2DL3*00201 3DL3*01303 2DL2*00301 | 1 2DL5B | 86 | 5 3DP1*00302 3DL2*0070102 | 1 - | 45 | 7 2DL1*00302 2DL4*001 | 2 2DL1 |
| cox | 18 (0) | 335 | 3 3DL3*00103 2DL3*00201 2DL3*007 | 1 2DL5B | 68 | 4 3DP1*006 | 1 - | 32 | 4 - | 1 - |
| ho301 | 19 (2) | 297 | 4 2DL2*00101 2DL2*00301 2DL5B*010 2DS2*002 | 4 2DL5B 2DL5A 2DS1 2DS5 | 63 | 6 2DS3*00201 2DL1*010 | 4 - | 36 | 8 3DP1*004 3DL3*014 | 5 3DL3 |
| jvm | 15 (0) | 329 | 2 2DL2*00301 2DP1*005 | 1 2DP1 | 49 | 3 3DL3*00801 | 1 - | 18 | 6 3DP1*00302 2DL4*00801 3DL1*00101 | 1 - |
| kas011 | 18 (0) | 362 | 2 3DL2*019 3DL3*01302 | 1 2DL5B | 68 | 5 3DP1*00302 3DL3*00901 2DP1*002 | 2 3DL3 | 34 | 6 3DL2*01001 | 3 3DL2 |
| olga | 17 (0) | 346 | 2 2DL3*00101 3DL3*00902 | 1 2DL5B | 74 | 7 3DL2*00701 3DP1*00302 2DP1*006 3DL1*001 2DL4*011 | 3 3DL2 3DP1 | 39 | 8 3DL3*00201 | 4 3DL3 |
| rsh | 15 (0) | 365 | 0 - | 5 2DS4 3DL2 2DL5A 2DS3 2DS1 | 75 | 1 2DP1*00201 | 5 - | 35 | 3 2DL4*011 2DL1*00302 | 5 - |
| wt51 | 12 (0) | 311 | 1 3DL3*00103 | 4 3DP1 2DL1 2DS3 3DL2 | 75 | 1 - | 4 - | 31 | 3 2DL5A*00501 2DP1*001 | 4 - |

Tabulka C.24: Výsledky experimentu 3 na reálných datech. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. Shluky vytvářely alely, které od sebe měly vzdálenost maximálně 20.

| Genom | Alel | Krok 2 | | | Krok 3 | | | Krok 4 | | |
|--------|--------|------------|--|-------------------------------------|------------|---|-------------------|------------|---|------------|
| | | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc | Zbývá alel | Ztraceno | Geny navíc |
| amala | 19 (0) | 372 | 2 2DL2*00301 2DL3*001 | 1 2DL5B | 83 | 4 3DP1*00901 3DL2*0070102 | 1 - | 57 | 5 2DL1*00302 | 2 2DL1 |
| bob | 19 (0) | 390 | 3 2DL2*00301 2DL3*00201 3DL3*01303 | 2 2DL5B 2DS3 | 92 | 5 3DL2*0070102 3DP1*00302 | 2 - | 51 | 7 2DL4*001 2DL1*00302 | 3 2DL1 |
| cox | 18 (0) | 354 | 3 2DL3*00201 2DL3*007 3DL3*00103 | 1 2DL5B | 71 | 4 3DP1*006 | 1 - | 31 | 4 - | 1 - |
| ho301 | 19 (2) | 302 | 4 2DS2*002 2DL5B*010 2DL2*00301 2DL2*00101 | 4 2DL5B 2DL5A 2DS1 2DS5 | 66 | 6 2DS3*00201 2DL1*010 | 4 - | 37 | 8 3DL3*014 3DP1*004 | 5 3DL3 |
| jvm | 15 (0) | 327 | 2 2DL2*00301 2DP1*005 | 1 2DP1 | 49 | 3 3DL3*00801 | 1 - | 20 | 6 2DL4*00801 3DL1*00101 3DP1*00302 | 1 - |
| kas011 | 18 (0) | 345 | 2 3DL3*01302 3DL2*019 | 1 2DL5B | 64 | 5 3DP1*00302 2DP1*002 3DL3*00901 | 2 3DL3 | 37 | 7 3DL2*01001 2DL4*00103 | 3 3DL2 |
| olga | 17 (0) | 348 | 2 3DL3*00902 2DL3*00101 | 1 2DL5B | 72 | 7 2DP1*006 3DL2*00701 3DL1*001 2DL4*011 3DP1*00302 | 3 3DL2 3DP1 | 41 | 8 3DL3*00201 | 4 3DL3 |
| rsh | 15 (0) | 365 | 0 - | 4 2DS4 3DL2 2DL5A 2DS1 | 80 | 1 2DP1*00201 | 4 - | 35 | 3 2DL4*011 2DL1*00302 | 4 - |
| wt51 | 12 (0) | 313 | 1 3DL3*00103 | 4 2DL1 2DS3 3DP1 3DL2 | 77 | 1 - | 4 - | 32 | 3 2DP1*001 2DL5A*00501 | 4 - |

Tabulka C.26: Výsledky experimentu 3 na reálných datech po odřezání nekvalitních bází. Odřezány byly alely, které měly pokrytí menší než 70 %. Za blízké byly považovány v případě, kdy byla jejich vzdálenost mezi sebou menší než 100. Shluky vytvářely alely, které od sebe měly vzdálenost maximálně 20.