VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

LOKALIZACE METYLAČNÍCH MÍST TRANSPOSONŮ

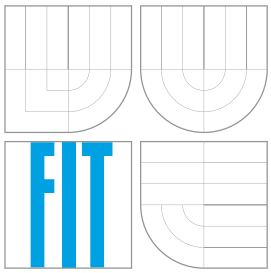
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. MIROSLAV KMEŤ

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

LOKALIZACE METYLAČNÍCH MÍST TRANSPOSONŮ

LOCALIZATION OF METHYLATION SITES IN TRANSPOSONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. MIROSLAV KMEŤ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. IVAN VOGEL

BRNO 2015

Abstrakt

Tato diplomová práce se zabývá realizací nástroje na extrakci metylační úrovně transpozonových sekvencí. Transpozony jsou prvky DNA schopné množení nebo přemístňování a jejich aktivita je regulována mechanismem metylace DNA. Informace o metylovaných pozicích jednotlivých sekvencí je uložená v bisulfitových datech a pro jejich zpracování jsou využívány části existujících nástrojů v kombinaci s vytvořenými moduly. Vzniklý nástroj zohledňuje unikátní výzvy, které transpozónové sekvence přináší do procesu analýzy metylace a jeho funkcionalita je prezentováná na sadě experimentů se simulačními a reálnými daty.

Abstract

This master's thesis deals with the creation of a tool for the extraction of methylation level from transposon sequences. Transposons are DNA elements with ability to move or copy themselves and their activity is regulated by DNA methylation. Sequence methylation information is stored in the bisulfite data and their processing is done with parts of two existing tools in a combination with implemented modules. Created tool takes into consideration unique challenges brought in the methylation calling process by transposable elements and it's functionality is presented on a set of experiments with simulated and real data.

Klíčová slova

transpozony, metylace, DNA, bisulfitové sekvenování, next-generation sekvenování, 3-písmenové mapování, extrakce metylace.

Keywords

transposons, methylation, DNA, bisulfite sequencing, next-generation sequencing, 3-letter mapping, methylation calling.

Citace

Miroslav Kmeť: Lokalizace metylačních míst transposonů, diplomová práce, Brno, FIT VUT v Brně, 2015

Lokalizace metylačních míst transposonů

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Ivana Vogela. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Miroslav Kmeť
24. května 2015

Poděkování

Děkuji svému vedoucímu Ing. Ivanu Vogelovi za odbornou pomoc a rady a Mgr. Zdeňkovi Kubátovi, PhD. z Biofyzikálního ústavu AVČR za odbornou pomoc a konzultace, které mi poskytli během vytváření práce.

© Miroslav Kmeť, 2015.
Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

Kapitola 1

Úvod

Molekula DNA nachádzajúca sa v každej bunke živých organizmov obsahuje okrem kódujúcich oblastí aj sekvencie, ktoré sa po dlhú dobu zdali byť nepodstatné, bez akejkoľvek významnej funkcie. Tieto sekvencie sa spoločne označujú ako transpozóny a ide o úseky DNA schopné kopírovania alebo zmeny pozície v rámci genómu. Táto transpozícia môže vyústiť k ovplyvneniu génovej expresie alebo génovým mutáciám, čo má väčšinou negatívny dopad na organizmus. Na druhú stranu, transpozóny prispievali významne z evolučného pohľadu k rozširovaniu genómu a bunky vyšších organizmov sa naučili využívať ich funkcionalitu vo svoj prospech. Regulácia ich aktivity je teda veľmi dôležitou súčasťou tohto stabilného systému a jednu z najvýznamnejších úloh hrá v tomto ohľade metylácia.

Metylácia je nástroj regulácie génovej expresie v eukaryotických bunkách siahajúci ďaleko za umlčovanie transpozónov a jej vývin sa prikladá práve nutnosti potlačovať parazitické transponovateľné elementy. Nádorové bunky sa často vyznačujú stratou kontroly nad týmito parazitmi práve v dôsledku neadekvátnej metylácie. Jej aktivita sa líši v rôznych štádiách vývoja organizmu a sledovanie týchto zmien môže poskytnúť unikátny náhľad na niektoré procesy, ktoré sú nimi ovplyvňované. Popis mechanizmu metylácie, transponovateľných elementov a ich vzťahu ku genómu bude predmetom 2. kapitoly.

Pre skúmanie a analýzu úrovne metylácie DNA sekvencií sa najčastejšie používajú metódy next-generation sekvenovania (NGS), predovšetkým ide o metódu bisulfitového sekvenovania využívajúcu aplikáciu hydrogénsiričitanu sodného, ktorá dokáže rozlíšiť metylované bázy od nemetylovaných. Postup týchto metód je súčasťou 3. kapitoly.

V súčasnosti existuje mnoho nástrojov venujúcich sa problému spracovania bisulfitových dát pre odhad metylačných hodnôt a postupu výpočtu ich jednotlivých fáz je venovaná 4. kapitola. Je v nej analyzovaných niekoľko najčastejšie používaných nástrojov využívajúcich odlišné princípy a tieto nástroje sú medzi sebou porovnávané.

Problém lokalizácie metylačných miest transpozónov spolu s návrhom jeho riešenia je súčasťou 5. kapitoly. Obsahuje taktiež detailný popis každej časti procesu od spracovania vstupných dát až po získanie výstupov. Tento proces pozostáva z dvoch celkov, jeden sa venuje fáze zarovnania vstupných sekvencií a druhý popisuje princíp extrakcie metylačných miest a celkovej metylačnej úrovne transpozónových rodín.

V 6. kapitole sú popísané realizované experimenty s generovanými a reálnymi dátami a ich výsledky sú v nej podrobne diskutované. V závere je zhrnutie týchto experimentov spolu s možnosťami využitia vzniknutého nástroja a návrhmi na jeho vylepšenie.

Kapitola 2

DNA

Deoxyribonukleová kyselina (DNA) je molekula nachádzajúca sa vo všetkých bunkách organizmov a slúži na uchovanie genetickej informácie ako podklad pre tvorbu proteínov. U eukaryotických buniek sa nachádza v bunečnom jadre obklopenom membránou, zatiaľ čo u prokaryot je uložená voľne v cytoplazme.

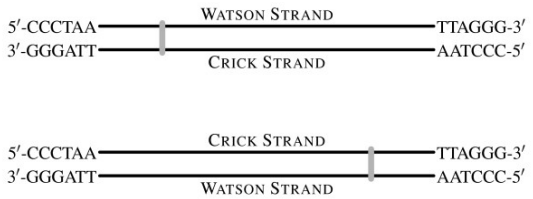
Jej objav siaha až do 19. storočia, konkrétne rok 1869. Významným objavom v tejto oblasti bolo odhalenie jej trojrozmernej štruktúry Watsonom a Crickom v roku 1953, ktorí predstavili jej dvojzávitnicový model [51].

2.1 Štruktúra DNA

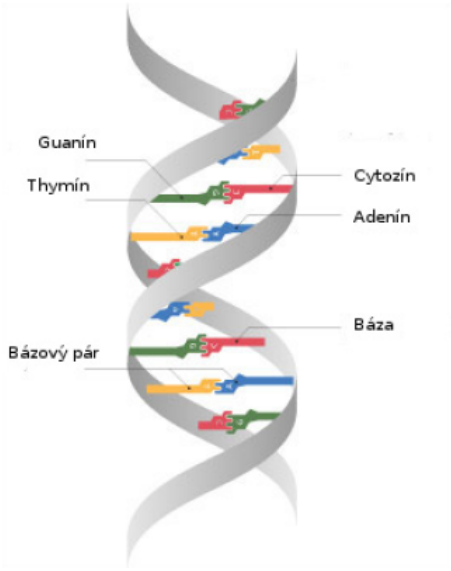
Nukleové kyseliny sú makromolekuly zložené z opakujúcich sa podjednotiek nazývaných nukleotidy. Každý jeden nukleotid pozostáva z (1) fosfátovej skupiny; (2) päťuhlíkatého cukru - pentózy a (3) dusíkatej zložky - bázy. V DNA sa bežne vyskytujú štyri bázy: adenín (A), guanín (G), thymín (T) a cytozín (C), kde prvé dva sa označujú ako puriny a druhé dva ako pyrimidiny. V molekule RNA sa namiesto T nachádza uracil (U). V obidvoch polynukleotidoch (DNA aj RNA) sú jednotlivé bázy prepojené do dlhých reťazcov.

Watson a Crick zistili, že takéto dva polynukleotidové reťazce sa v bunkách nachádzajú ako pravotočivá dvojitá závitnica, točiace sa navzájom okolo seba v špirále (obr. 2.2). Vlákna DNA držia pri sebe pomocou vodíkových väzieb medzi nukleotidami, konkrétne medzi purínom a pyrimidínom (A s T, G s C). Na základe tejto vlastnosti, tzv. komplementarity, je možné z jedného reťazca odvodiť druhý a vice versa [51].

Pokiaľ tieto dve vlákna DNA zapíšeme pod seba ako dva komplementárne reťazce, ktoré sú tvorené abecedou pozostávajúcou zo symbolov A, T, G, C, dostaneme sekvencie vlákna 5'-3' (Watsonove vlákno) a zároveň jeho komplementárne vlákno 3'-5' (Crickove vlákno) obr. 2.1 [15].



Obrázok 2.1: Watsonove a Crickove vlákna molekuly DNA [15].



Obrázok 2.2: Dvojzávitnicový model molekuly DNA s nukleotidmi A, G, T a C [2].

2.2 Význam

Hlavnou úlohou DNA je kódovanie informácií pre syntézu polypeptidových reťazcov alebo v niektorých prípadoch molekuly RNA. Takéto informácie sú zakódované vo funkčných jednotkách - génoch. Štruktúra génu je tvorená regulačnými oblasťami na jeho začiatku, kódujúcou oblasťou, ktorá je pri eukaryotických génoch prerušovaná nekódujúcimi sekvenciami (intrónmi) a signálnou sekvenciou pre ukončenie génovej expresie.

2.2.1 Génová expresia

Proces génovej expresie je dvojkrokový mechanizmus, kde v prvom kroku označovanom ako transkripcia dochádza k prenosu genetickej informácie z DNA do vlákna RNA - génového transkriptu. Naviazaním transkripčných faktorov na promotor génu sa vytvoria v DNA podmienky pre pripojenie RNA polymerázy, ktorej úlohou je syntéza mRNA. Takto vytvorená molekula je vstupom druhej fázy - translácie. Z mRNA sa v ribozómových iniciačných komplexoch postupne vytvára cieľový polypeptid, ktorý následne môže vykonávať rôzne funkcie v rámci organizmu.

2.2.2 CpG ostrovy

Regulačné oblasti génov sa taktiež označujú ako jeho promotor - miesto, od ktorého začína transkripcia. Nukleotidové sekvencie v tejto oblasti sú u väčšiny génov bohaté predovšetkým na C a G bázy tvoriace Cytozín-fosfát-Guanín dinukleotidy. V literatúre sú tieto úseky označované ako CpG ostrovy (angl. *CpG islands* - CGI) a existuje veľká korelácia medzi CGI a inicializáciou transkripcie. Dôvodom tejto funkcionality môže byť zvýšená pravdepodobnosť naviazania transkripčných faktorov na GC bohaté regióny. Ich funkcia však siaha ešte oveľa ďalej a iné mechanizmy, v ktorých sa CGI objavujú, sa stále len skúmajú. Už teraz sú ale známe ako dôležité regulačné štruktúry, ktoré sa vyvinuli selekciou v genómoch, kde DNA metylácia hrá dôležitú úlohu (viď sekcia 2.4) [18, 11].

2.3 Transpozóny

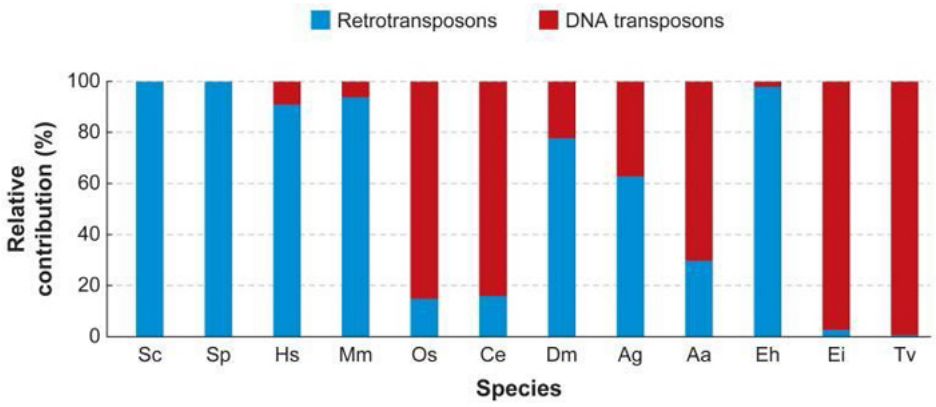
Molekula DNA nachádzajúca sa v každej bunke živých organizmov neobsahuje výlučne len génové sekvencie kódujúce informácie pre vytváranie proteínov. Z veľkej časti je tvorená transponovateľnými elementami (TE) označujúcimi sa ako transpozóny. U človeka tvoria približne 45% a u niektorých rastlín dokonca až viac ako 80% celkového genómu [41].

Transpozóny sú mobilné genetické elementy - DNA sekvencie, ktoré sa pohybujú v rámci genómu používaním odlišných mechanizmov špecifických pre danú skupinu týchto prvkov a toto premiestňovanie sa označuje ako transpozícia.

2.3.1 Rozdelenie

V rámci organizmov a ich DNA sekvencií existuje veľké množstvo TE, takisto ako aj veľa rôznych metód ich kategorizovania. Najznámejšou metrikou je však spôsob ich množenia a začleňovania do molekuly DNA. Na základe tohto sa rozdeľujú medzi transpozóny vyžadujúce reverznú transkripciu (transkripcia RNA do DNA) - retrotranspozóny, a TE, ktoré sa šíria bez nej - DNA transpozóny.

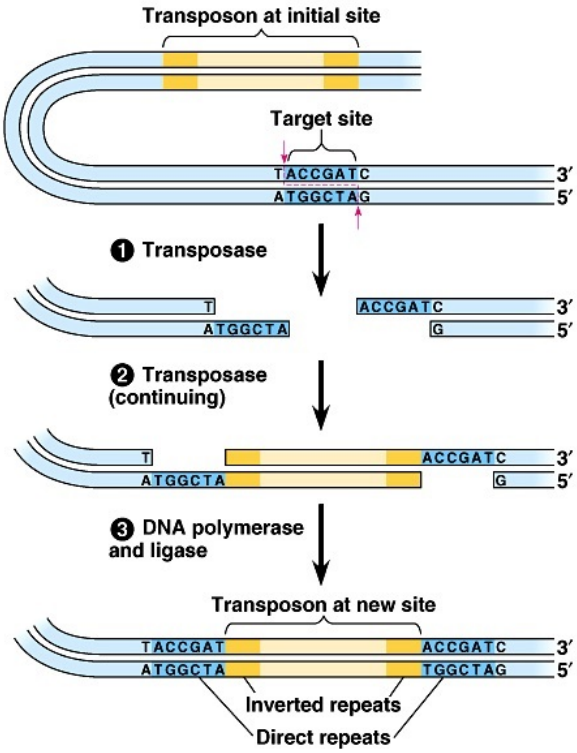
Ich zastúpenie v organizmoch sa líši a je zobrazené na vybraných druhoch na obrázku 2.3.



Obrázok 2.3: Relatívne množstvo retrotranspozónov a DNA transpozónov v rôznych eukaryotických genómoch (Sc: Saccharomyces cerevisiae; Sp: Schizosaccharomyces pombe; Hs: Homo sapiens; Mm: Mus musculus; Os: Oryza sativa; Ce: Caenorhabditis elegans; Dm: Drosophila melanogaster; Ag: Anopheles gambiae, malaria mosquito; Aa: Aedes aegypti, yellow fever mosquito; Eh: Entamoeba histolytica; Ei: Entamoeba invadens; Tv: Trichomonas vaginalis) [21].

DNA transpozóny

DNA transpozóny označované aj ako Typ II sú mobilné sekvencie DNA, ktoré sa väčšinou šíria prostredníctvom *„cut and paste“* mechanizmu. Princíp ich množenia spočíva v činnosti proteínu transponázy kódovaného v géne transpozónov, ktorý rozozná obrátené koncové repetície obklopujúce daný TE, vystrihne ho z tejto pozície a vykoná jeho vloženie na nové miesto v genóme. Pri vkladaní sa súčasne vytvoria duplikácie cieľového miesta ako dôsledok tvorby posunutých zlomov, ktorými je dvojreťazcová molekula rozdelená (obr. 2.4) [51].



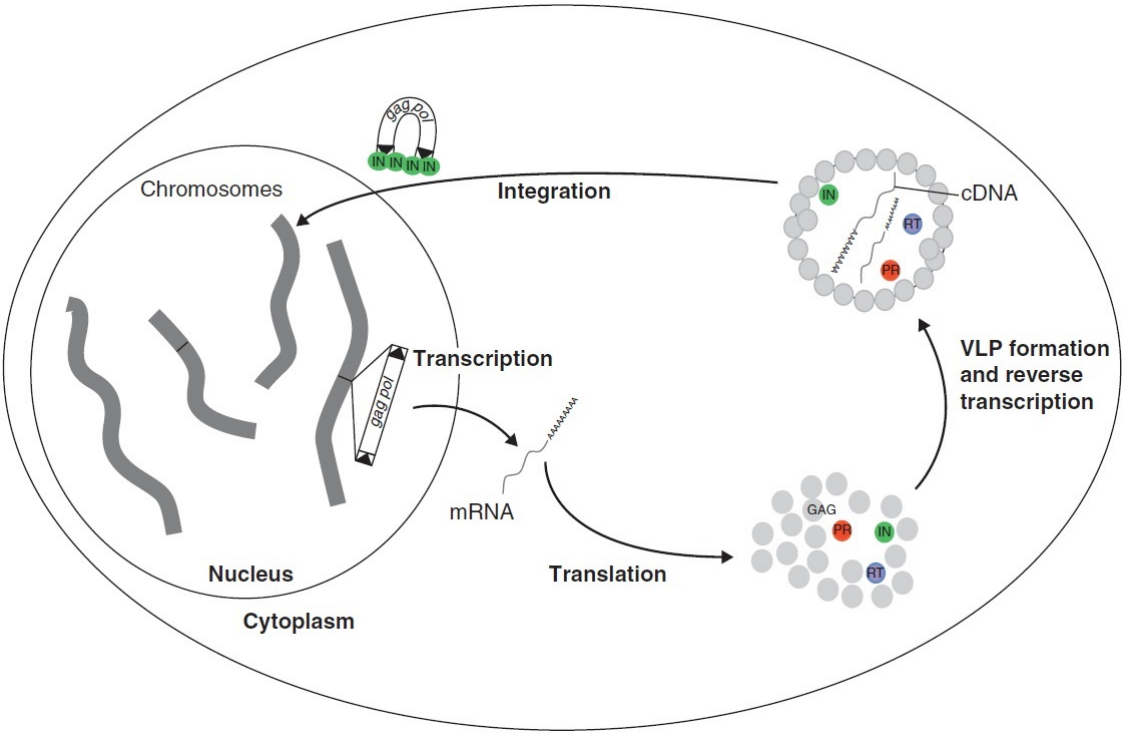
Obrázok 2.4: Proces šírenia DNA transpozónov mechanizmom *cut and paste* [3].

Retrotranspozóny

Replikácia retrotranspozónov je realizovaná reverznou transkripciou ich RNA a výsledná cDNA je integrovaná do odlišného lokusu. V eukaryotických genómoch sa vyskytuje prevažne jedna skupina retroelementov, tzv. retrotranspozóny s dlhými koncovými repetíciami (LTR retrotranspozóny). LTR sú repetície, ktoré obklopujú vnútorné kódovacie regióny - gény zodpovedné za syntézu štruktúrnych a enzymatických proteínov.

Štruktúrne proteíny sú kódované *gag* génom a ich úlohou je vytvoriť časticu podobnú vírom (*virus-like particle*, VLP), vnútri ktorej prebieha reverzná transkripcia. Proteíny s enzymatickou funkciou ako je proteáza pre Pol polyproteín, reverzná transkriptáza kopírujúca RNA retrotranspozónu do cDNA a integráza, ktorá integruje cDNA do genómu sú kódované génom *pol*.

Proces začleňovania retrotranspozónov do DNA začína transkripciou jeho RNA molekuly s pomocou RNA polymerázy II od promotoru nachádzajúcom sa v 5' LTR. RNA je následne vo fáze translácie preložená na proteíny, ktoré formujú VLP a realizujú reverznú transkripciu a integráciu. V jednej VLP sa typicky nachádzajú dve RNA molekuly, z ktorých sa vytvorí dlhá kópia DNA. Najprv sa tRNA naviaže na sekvenciu blízko 5' LTR, výsledná čiastočná cDNA sa premiestni z 5' LTR na 3' LTR, kde reverzná transkripcia pokračuje. Potom v blízkosti 3' LTR dôjde na cDNA k ďalšiemu prenosu vlákna za vzniku dvojvláknovej cDNA molekuly. V poslednom kroku sa cDNA integruje naspäť do hostiteľskej DNA a vznikne tým ďalšia kópia tohto retrotranspozónu [25]. Celý tento proces je prehľadne zobrazený na obrázku 2.5.



Obrázok 2.5: Životný cyklus LTR retrotranspozónov. IN - integráza, PR - proteáza, RT - reverzná transkriptáza, VLP - častica podobná vírusu, čierne trojuholníky reprezentujú LTR [25].

Rodiny

Okrem rozlišovania transpozónov do vyššie popísaných tried na základe ich mechanizmu šírenia v molekule DNA sa transpozóny radia do skupín v rámci týchto tried označovaných ako rodiny. Náležitosť do konkrétnej rodiny je závislá od ich štruktúrnych vlastností, zdielanej podobnosti sekvencií (predovšetkým príbuznosť transponáz) a veľkosť a sekvencia cieľových miest duplikácie generovaných pri vložení [12].

Medzi bežné rodiny DNA transpozónov patrí napríklad Tc1/mariner, hAT, P element, CACTA, PiggyBac, Merlin, Helitron, Maverick a iné [21].

2.3.2 Funkcia

Transpozóny, často označované ako sebecká (angl. *selfish*) alebo odpadová (angl. *junk*) DNA, zdanlivo neprispievajú k fyziologickým procesom organizmu, pretože väčšina z nich je neaktívnych. Napriek tomuto označeniu sa však predpokladá, že TE výrazne prispievali v procese evolúcie k zväčšeniu veľkosti genómu [47].

Aktívne TE sú schopné produkovať rôzne genetické zmeny v procese ich začleňovania do DNA (inzercia, vyčlenenie, duplikácia alebo premiestnenie). Ide predovšetkým o deaktiváciu génov alebo ovplyvnenie génovej expresie pri inzercii do intrónov, exónov alebo regulačných oblastí. Taktiež sa môžu podielať na reorganizácii genómu prostredntíctvom mobilizácie netranspozónovej DNA [41].

Transpozóny po začlenení do génov môžu spôsobiť mutáciu vyúsťujúcu v choroby ako hemofília [26] alebo rakovina hrubého čreva [40].

2.3.3 Regulácia

Ovplyvňovanie génovej expresie ako hlavný dôsledok existencie transpozónov v genóme organizmov má negatívny vplyv na fitnes daného jedinca. TE sú touto skutočnosťou taktiež ovplyvňované, pretože prežitie hostiteľa je kritické pre šírenie transpozónov. Z tohto dôvodu boli vyvinuté ako hostiteľom, tak aj transpozónmi rôzne stratégie pre minimalizovanie dopadu transpozície na fitnes daného organizmu [41].

Z pohľadu TE ide napríklad o vkladanie do neesenciálnych oblastí s veľmi malým vplyvom na fyziológiu [27] alebo aktivitu predovšetkým v embryonálnom štádiu, kde len mierne škodlivé inzercie môžu byť prenesené na nasledujúce generácie. Hostiteľský organizmus si vytvoril vlastné mechanizmy ako redukovať vysokú úroveň aktivity transpozónov. Medzi najvýznamnejší z nich patrí **metylácia DNA** [19, 55]. Výsledkom týchto mechanizmov je umlčanie aktivity väčšiny TE nachádzajúcich sa v genóme organizmu.

Schopnosť transpozónov zvyšovať genetickú rozmanitosť spolu so schopnosťou genómu obmedzovania aktivity väčšiny TE vyúsťuje v rovnováhu, ktorá robí transpozóny dôležitou súčasťou evolúcie a génovej regulácie u všetkých organizmov vyznačujúcich sa týmito sekvenciami [45].

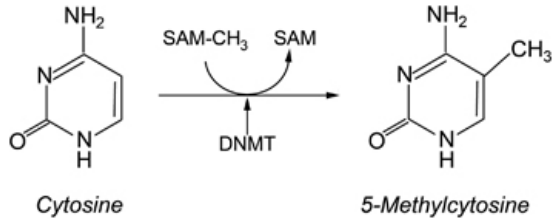
2.4 Metylácia DNA

Conrad Waddington v roku 1946 predstavil pojem „epigenetika“ ako súbor interakcií medzi génmi a okolitým prostredím, ktoré určujú fenotyp alebo fyzické črty v organizme. Počiatočný výskum sa sústredil na genómové regióny ako *heterochromatin* a *euchromatin*, pretože obsahujú aktívne aj neaktívne gény. Neskôr však bola objavená rola modifikácie histónov a predovšetkým proces metylácie DNA ako kľúčové faktory regulácie génovej expresie. Kontrola génovej expresie je nesmierne dôležitá v životnom cykle organizmov, ale nie je jediným dôvodom existencie metylácie. Tento mechanizmus slúži z veľkej časti pre ochranu genómovej DNA pred cudzími DNA inzerciami vznikajúcimi napríklad v životnom cykle transpozónov alebo vírov - spôsobuje ich umlčanie [52].

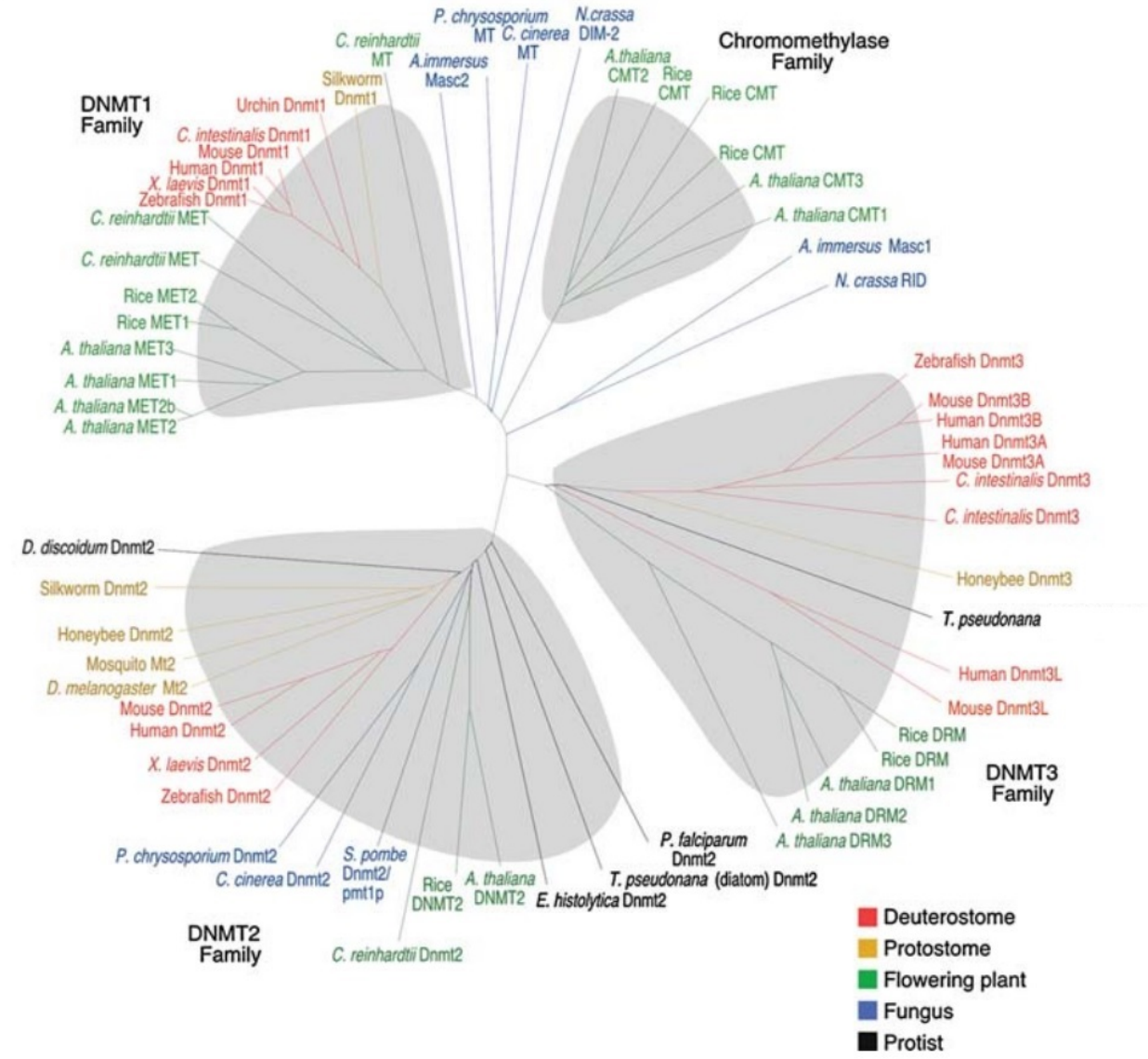
2.4.1 Popis

DNA metylácia je proces pridania metylovej skupiny (-CH_3) cez kovalentnú väzbu na cytozínovú bázu kostry DNA a metylovaný cytozín sa označuje mC (obr. 2.6). Typicky k nej dochádza na CpG dinukleotidoch. Táto modifikácia môže zabrániť asociácii niektorých faktorov viažúcich sa na DNA (napr. transkripčné faktory) s ich príbuznými DNA rozpoznateľnými sekvenciami alebo umožní proteínom, ktoré rozpoznajú mCpG, vyvolať potlačovací potenciál metylovanej DNA.

Metylácia je realizovaná a udržovaná prostredníctvom enzýmov označovaných ako *DNA Metyl Transferázy* (DNMT). Táto veľká skupina enzýmov sa naviac rozdeľuje do niekoľko rodín podľa ich sekvenčnej homológie, pričom niektoré z nich sú zdielané naprieč rôznymi druhmi a ríšami organizmov a naopak, niektoré sú špecifické len pre určité typy (napr. chromometylázy pre rastlinnú ríšu) [23]. Prehľad týchto rodín je možné vidieť na obrázku 2.7.



Obrázok 2.6: Metylácia cytozínu na 5-metylcytozín katalyzovaná DNA metyltransferázami (DNMT) [42].



Obrázok 2.7: Prehľad jednotlivých enzýmov a ich rodín podieľajúcich sa na metylácii DNA v rôznych organizmoch [23].

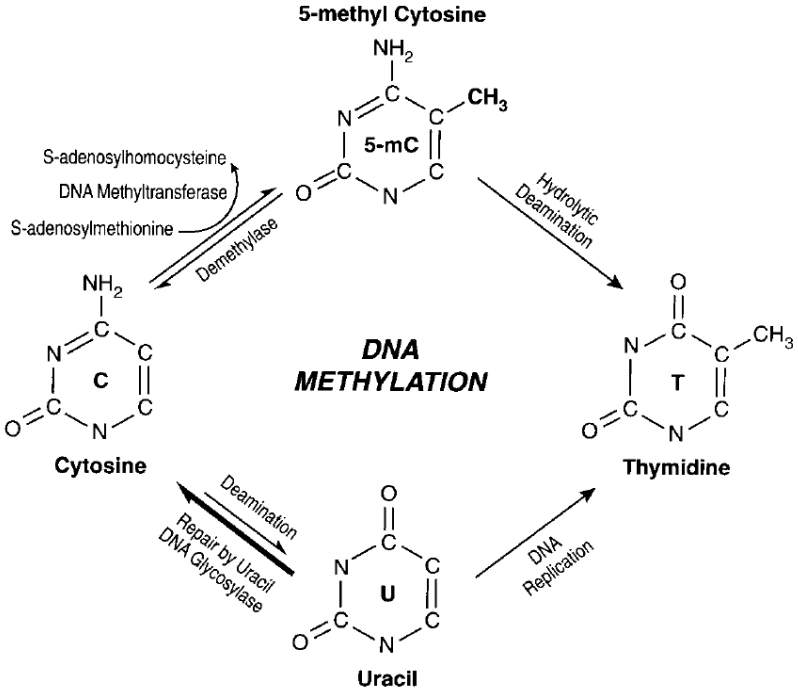
2.4.2 DNA metylácia v rastlinách

V cicavčích DNA prebieha metylácia väčšinou v kontexte CG (úseky bohaté na nukleotidy C a G). Nejedná sa však o jediné možné kontexty, v ktorých môže dochádzať k tomuto procesu. Okrem CG existujú ešte aj CHG a CHH DNA kontexty, kde H reprezentuje jeden z nukleotidov A, C alebo T. U živočíchov je v týchto ďalších kontextoch pravdepodobnosť výskytu metylácie rádovo menšia ako u CG, avšak u rastlín je situácia odlišná. Úroveň metylácie môže pri rastlinách dosahovať v CG kontexte približne rovnakej úrovne ako pri cicavčích genómoch (80-90%), ale kontexty CHG, respektíve CHH sa môžu výrazne líšiť v závislosti od druhu rastliny a štádia vývoja pohybujúce sa od 20% do 90%, respektíve 5% až 80%.

Aktivita DNMT bude taktiež odlišná naprieč rôznymi kontextami a medzi najvýznamnejšie enzýmy podieľajúce sa na metylácii u rastlín patrí DNA metyltransferáza1 (MET1) zabezpečujúca zachovanie CG metylácií pri DNA replikácii. Pre rastliny špecifické Chrometyláza2 (CMT2) a CMT3, kde CMT2 metyluje obidva kontexty CHG a CHH *de novo* a CMT3 udržuje CHG metyláciu a zabezpečuje jej propagáciu cez delenie buniek [29].

2.4.3 Demetylácia a deaminácia

Metyláciou sa potláča aktivita daného úseku sekvencie DNA, ale pre úpravu už stanoveného metylačného vzoru musí existovať samostatný mechanizmus. V súčasnosti sa vie o dvoch procesoch vykonávajúcich túto funkciu a to pasívna demetylácia, ku ktorej dochádza, keď DNMT zlyhajú pri udržiavaní existujúcej metylácii. Druhým je aktívna demetylácia realizovaná špeciálnym enzýmom - demetylázou.



Obrázok 2.8: Schématická reprezentácia biochemických ciest pre metyláciu cytozínu, demetyláciu a deamináciu cytozínu a metylcytozínu [49].

Metylované C môžu naviac prejsť na molekuly T alebo U v procese deaminácie - odstránení amino skupiny z molekuly. Tento princíp je využívaný pri sekvenovaní DNA pre získanie úrovní metylácie a bude popísaný v kapitole 3. Výsledkom deaminácie C v DNA je redukcia množstva CpG ostrovov a trvalá deaktivácia prislúchajúcich sekvencií. Počas evolúcie boli práve procesom deaminácie dinukleotidy CpG postupne eliminované z genómu vyšších eukaryot a sú prítomné iba z 5% až 10% ich predpokladanej frekvencie [49].

Obrázok 2.8 zobrazuje štádia, ktorými môže cytozín prechádzať v molekule DNA v kontexte metylačných zmien.

Kapitola 3

Sekvenovanie DNA

DNA sekvencia reprezentuje formát, z ktorého je možné extrahovať široké spektrum biologických informácií. Získava sa prostredníctvom sekvenačných metód, ktorých história siaha až desiatky rokov dozadu. Väčšina z nich je založená na nejakej variante Sangerovej biochémie a tvoria metódy prvej generácie používané až približne 25 rokov [48, 46].

3.1 Next-generation sekvenovanie

Potreba zlepšenia procesu sekvenovania viedla k vývoju metód druhej generácie, označovaných aj ako *next-generation sequencing* (NGS). Hlavným cieľom využitia týchto metód je zrýchlenie sekvenačného procesu, možnosť paralelnej sekvenácie a zníženie nákladov, ktoré boli v prípade Sangerovej metódy veľmi vysoké. NGS produkuje obrovské množstvo sekvenčných čítaní (*readov*), čo sú krátke nasekvenované úseky DNA. U niektorých prístrojov je uvedená dokonca miliarda *readov* v jednom behu. Schopnosť takto rozsiahlej sekvenácie celého genómu rôznych organizmov prispela k rozvoju výskumu, ktorý bol predtým nerealizovateľný [46].

Mnohé výhody NGS metód sú v súčasnosti sprevádzané aj niekoľkými nevýhodami. Najvýznamnejšou z nich je dĺžka *readov* (kratšia ako u konvenčných sekvenácií) a presnosť (aspoň desaťkrát menej presné ako Sangerova metóda). Ide však prevažne o relatívne nové metódy a je možné očakávať, že ich výkonnosť sa bude len zlepšovať s ohľadom na tieto parametre, takisto ako to bolo u metód prvej generácie.

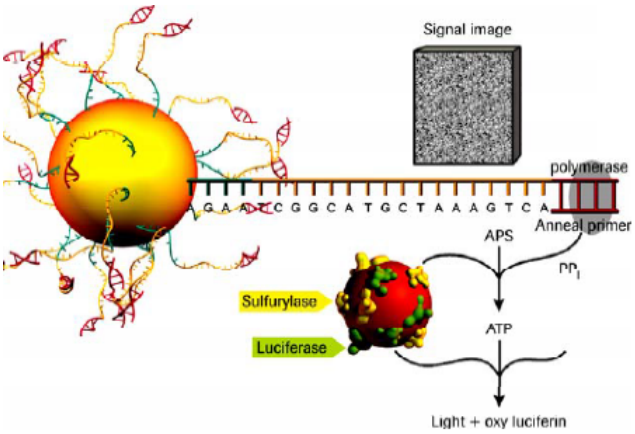
Rôzne platformy NGS sa od seba líšia biochemickými postupmi, ich proces vykonávania je však principiálne podobný. Príprava knižnice - fragmentov DNA dĺžky rádovo stovky bázových párov - sa dosiahne náhodnou fragmentáciou DNA nasledované ligáciou adapterových sekvencií (krátkych oligonukleotidov). V ďalšej fáze dochádza ku klonovaniu sekvencií a tieto molekuly skončia priestorovo oddelené v zhlukoch. Samotné sekvenovanie pozostáva zo striedajúcich sa cyklov enzymatických reakcií (prostredníctvom polymerázy alebo ligázy) a detekciou toho, aká báza sa začlenila do vznikajúceho reťazca v aktuálnom cykle [48].

3.1.1 454 pyrosequencing

454 bol prvou NGS sekvenačnou platformou dostupnou ako komerčný produkt. Knižnica môže byť zostrojená akoukoľvek metódou, ktorej výsledkom sú krátke, adapterom označené fragmenty. Proces klonovania je realizovaný emulziou PCR na mikroguličkách (*beads*), kde sa na ich povrch naviažu jednotlivé amplikóny. Na mikroguličkách obsahujúcich ampli-

kóny sa sekvenčný primer hybridizuje k adapteru na pozícii priľahlej k začiatku neznámej sekvencie.

Sekvenovanie je realizované metódou pyrosekvenovania. Začlenenie nukleotidu do vznikajúceho reťazca je rozpoznateľné vďaka luminiscencii vyvolanej sériou reakcií enzýmov DNA polymerázy, ATP sulfurylázy, luciferázy a apyrázy. K mikroguličkám sa pridá roztok konkrétneho nukleotidu a pokiaľ dôjde k začleneniu do vytváraného reťazca prostredníctvom DNA polymerázy, uvoľní sa pyrofosfát reagujúci s ATP sulfurylázou. Vzniknutá ATP poháňa činnosť enzýmu luciferázy, pri ktorej sa uvoľní svetelné žiarenie zachytené snímacou kamerou (CCD) - informácia o inkorporovanom nukleotide (obr. 3.1) [48, 28].



Obrázok 3.1: Sekvenačná reakcia pri pyrosekvenovaní. Jedna mikroguličkа obsahuje milióny kópií DNA molekúl [1].

3.1.2 Illumina Genome Analyzer (Solexa)

Podobne ako u predchádzajúcej platformy, aj tu je možné vytvoriť knižnicu akoukoľvek metódou, ktorá produkuje fragmenty označené adapterom dĺžky niekoľko stovák bázových párov. K amplifikácii sekvencií dochádza tzv. mostíkovou PCR (*bridge PCR*) za vzniku niekoľko miliónov zhlukov amplikónov podľa templátovej molekuly. Každý zhluk ich obsahuje približne 1000. Amplikóny sú následne linearizované na jednovláknové molekuly a na konce sekvencií sa naviaže primer.

Samotná sekvenácia prebieha v prístroji Illumina Genome Analyzer, kde jednotlivé templáty sú sekvenované pomocou štyroch odlišných fluorescenčných farbív. Postupne dochádza k naviazaniu fluorescenčne značených báz na templát a po každom kole syntézy sú jednotlivé bázy rozpoznané laserom podľa daného fluorescenčného farbiva [48].

3.1.3 Applied Biosystems (AB) SOLiD

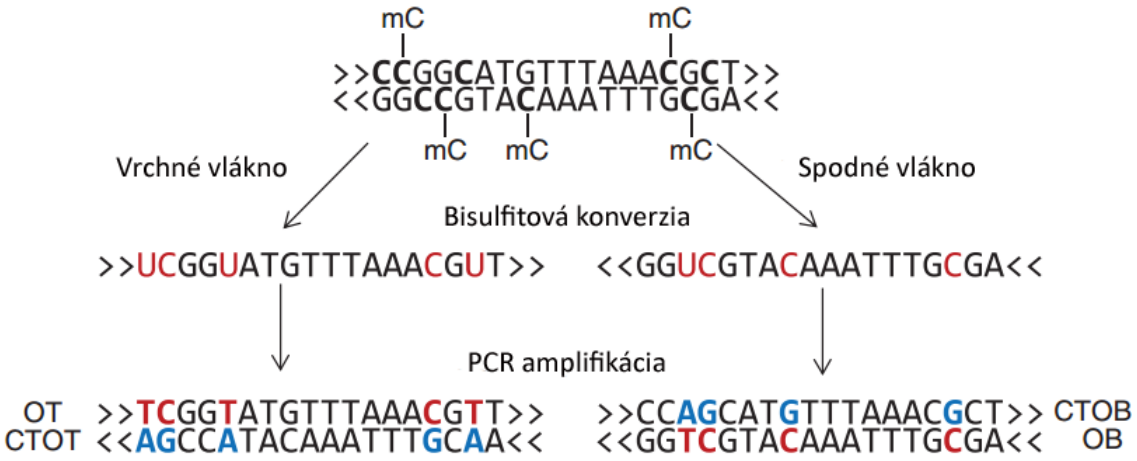
Knižnica je vytvorená pomocou fragmentov dlžky 60-90bp, na ktorých obidva konce sa naviažu adapterové sekvencie. Klonovanie, podobne ako u 454, je sprostredkované PCR emulziou s amplikónmi zachytenými na mikroguličkách. Mikroguličky sú selektívne vybrané a umiestnené na pevný rovinný povrch.

Sekvenovanie metódou SOLiD sa od predchádzajúcich líši. Dochádza tu totiž k sekvenovaniu ligáciou (*sequencing by ligation*) a nie polymerázou. Univerzálny primer komplementárny k adapterovej sekvencii je hybridizovaný k mikroguličkám obsahujúcim amplikóny. Potom dochádza k naviazaniu oligonukleotidových oktamerov, ktoré sú štrukturované tak,

aby špecifická pozícia v rámci oktameru (napr. báza 5) korelovala s fluorescenčným značením. Po ligácii sú údaje zozbierané pre jednu rovnakú pozíciu zo všetkých templátov mikroguličiek a celý tento proces sa opakuje pre všetky nukleotidy. V druhom kole sa použije primer o bázu kratší a celkovo prebehne päť takýchto cyklov [48].

3.2 Bisulfitové sekvenovanie

Vďaka schopnosti NGS metód rozlišovať jednotlivé nukleotidy sa stalo bisulfitové genómové sekvenovanie (BS-Seq) jednou z najvýznamnejších metód detekcie DNA metylácie, pretože poskytuje kvantitatívny, kvalitatívny a veľmi efektívny prístup k identifikácii metylcytozínov u jednotlivých bázových párov. Je založená na zistení, že reakcia amínácie na C a mC prebehne s odlišným výsledkom po ošetrení sekvencií hydrogénsiričitanom sodným (bisulfitom). Cytozíny budú v jednovláknovej DNA po tomto procese konvertované na uracil a rozozonané ako thymín v následnom rozmnožení pomocou PCR a sekvenovaní. Metylované cytozíny sú však voči tejto konverzii imúnne, a teda zostávajú ako cytozíny, čo umožňuje ich odlíšenie od nemetylovaných cytozínov (obr. 3.2) [36].



Obrázok 3.2: Efekt aplikácie bisulfitu na DNA. Výsledkom bisulfitovej konverzie genómovej DNA a následnej PCR amplifikácie sú dva PCR produkty a až štyri potenciálne odlišné DNA fragmenty pre každý lokus. Metylované cytozíny sú odolné voči bisulfitovej konverzii a môžu byť použité pre určenie metylačného stavu DNA. OT, originálne vrchné vlákno; CTOT, komplementárne vlákno k originálnemu vrchnému vláknu; OB, originálne spodné vlákno; a CTOB, komplementárne vlákno k originálnemu spodnému vláknu.

PCR je v tejto fáze nutná pre určenie metylačného stavu špecifického lokusu s použitím špeciálnych metylačných primerov po aplikácii bisulfitu. Skutočný metylačný stav môže byť zistený buď pomocou priameho PCR sekvenovania (detekcia priemerného metylačného stavu) alebo tzv. „*sub-cloning*“ sekvenovaním (detekcia metylácie jednotlivých molekúl). Analýzou dát metódy bisulfitového sekvenovania môžeme získať metylačné vzory ako na jednovláknovej DNA molekule, tak aj na oboch DNA vláknach, pretože konvertované DNA vlákna už nie sú k sebe navzájom komplementárne a výsledky amplifikácie môžu byť merané

individuálne. Existujú dva bežné, ale rozdielne protokoly, ktoré sa k tomuto účelu používajú: *directional* [37] a *non-directional* [17] protokol.

Directional protokol

Prvý z nich je menej komplexný. Genomická DNA je najprv náhodne fragmentovaná, adaptéry sú ligované a gélová elektroforéza je aplikovaná na vybrané fragmenty požadovanej dĺžky. Potom je DNA ošetrená bisulfitom a amplifikovaná s PCR. Ďalšia sada adaptérov je ligovaná pre finálnu časť sekvenačného procesu. Metylačné vzory nie sú symetrické na dvoch DNA vláknach, a preto je nutné rozlišovať medzi vrchným a spodným vláknom. Pri amplifikácii dôjde k syntéze reverzných komplementárnych vlákien k pôvodnému vrchnému a pôvodnému spodnému (obr. 3.2). *Directional* protokol zabezpečuje s pomocou špeciálnych adaptérov, že v prípade *single-end readov* sú sekvenované iba fragmenty z originálneho vrchného a spodného vlákna. Nemetylované C budú reprezentované ako T vo všetkých *readoch*. U *pair-end readov* sú naviac sekvenované aj reverzné komplementy k originálnym vláknam a znovu, vďaka adaptérom, je zabezpečené, že ľavé *ready* budú vždy pochádzať z originálneho vlákna a pravé *ready* z komplementárneho. Z ich orientácie je možné určiť odpovedajúce vlákno. V tomto prípade sú u pravých *readov* C>T bisulfitové konverzie zobrazené ako G>A konverzie, pretože pochádzajú z reverzného komplementu a pre ľavé *ready* platí to isté ako pre *single-end ready* [32].

Non-directional protokol

Non-directional protokol sa odlišuje tým, že po odstránení prvej adapterovej sady nasledovanej ligáciou ďalšej adapterovej sady je vykonaná ešte PCR amplifikácia. Vzniknú štyri rôzne typy *readov* ako pre *single-end*, tak aj pre *pair-end ready*. Dva z nich obsahujú informáciu o originálnych vláknach, kde sa T nachádza na mieste nemetylovaných C a dva z reverzných komplementov, ktoré môžu obsahovať G>A konverzie. Hlavnou nevýhodou tohto protokolu je práve strata orientácie - každý *read* môže pochádzať z ľubovoľného vlákna. Táto situácia je ale riešiteľná, ak má *read* mapovanie na danú orientáciu a určitý počet pozorovateľných konverzií buď typu C>T alebo G>A. Najpoužívanejším je *directional* protokol [32].

Existuje niekoľko techník na základe princípu bisulfitovej modifikácie ako *Methylation Specific PCR* (MSP), *Combined Bisulfite Restriction Analysis* (COBRA), *Methylation-sensitive Single Nucleotide Primer Extension* (Ms-SNuPE) a ďalšie v závislosti od rôznych aplikácií. V porovnaní s inými prístupmi zisťovania DNA metylácie (enzýmy špecificky rozpoznávajúce metylované cytozíny) sú metódy analýzy metylácie DNA založené na bisulfite presnejšie, citlivejšie, efektivnejšie a ich používanie je viac rozšírené [36].

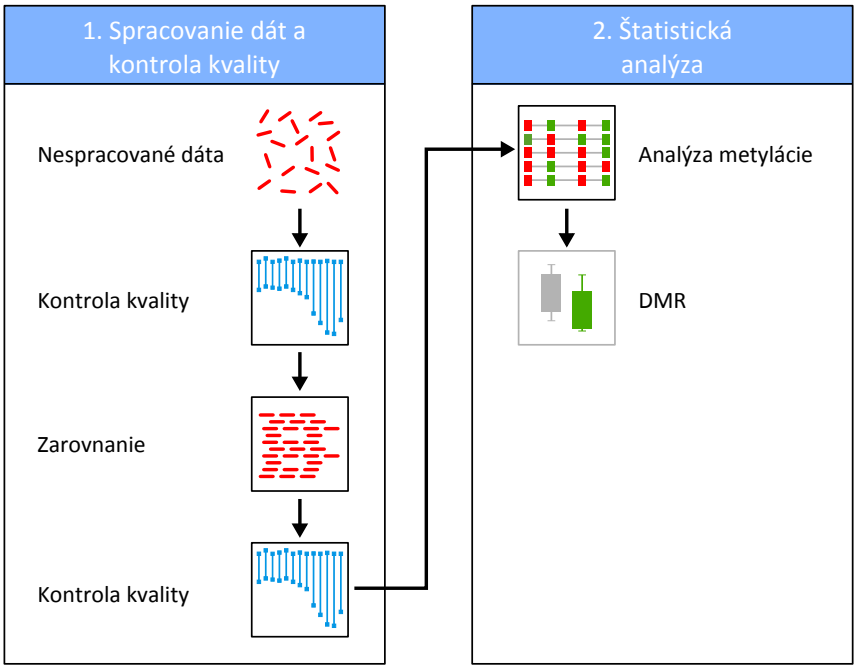
Kapitola 4

Bioinformatické spracovanie bisulfitových dát

V súčasnosti neexistujú kompletné bioinformatické nástroje pre efektívnu extrakciu metylačných miest transpozónov a celý tento proces je realizovaný čiste len s pomocou vyššie popísaných metód bisulfitového sekvenovania. Ide o zdĺhavý, nákladný a nie úplne praktický spôsob získavania tejto informácie pre veľké množstvo sekvencií. Z tohto dôvodu je vytvorenie nástroja, ktorý pracuje priamo s bisulfitovými *readmi* a extrahuje z nich metylačnú informáciu pre jej následnú analýzu, dôležitým krokom pri skúmaní vplyvu transpozónov a ich úrovne metylácie na organizmy.

4.1 Profilovanie DNA metylácie

Získanie údajov o úrovni metylácie z vytvorených sekvencií je niekoľkokrokový proces rozdelený do dvoch hlavných častí. V prvej z nich sa kontroluje kvalita nespracovaných dát. Tento proces sa označuje ako predspracovanie. Bisulfitové dáta sú následne zarovnávané voči referenčnému genómu a po zarovnaní môže dochádzať k dodatočnej kontrole kvality a filtrovaniu nájdených výsledkov. Z namapovaných sekvencií sú vo fáze štatistickej analýzy najprv extrahované metylačné miesta jednotlivých cytozínov a z týchto údajov je ďalej možné v štatistickom spracovaní zisťovať rozdiely medzi odlišne metylovanými regiónmi (*Differentially methylated regions* - DMR). Táto práca je zameraná na celý tento proces s výnimkou analýzy DMR, a preto nebude táto časť ďalej rozoberaná (obr. 4.1).



Obrázok 4.1: Proces profilovania DNA metylácie pozostávajúci zo spracovania dát a štatistickej analýzy výsledkov mapovania.

4.1.1 Predspracovanie sekvencií

Predspracovanie sekvencií je predovšetkým proces eliminácie alebo určitej manipulácie sekvencií s nízkou kvalitou a príprava pre nasledujúci krok - zarovnanie. Vytvorené krátke sekvencie metódou BS-Seq môžu strácať na kvalite k 3' koncu, špeciálne u dlhších sekvencií, kedy kvalita nukleotidu na konkrétnej osekvenovanej pozícii (*base-call*) klesá s narastajúcou dĺžkou *readov*. Výsledkom tohto procesu je narušená kompozícia báz, čo by mohlo neskôr viesť k nekorektnej extrakcii metylačných miest. Existujú dva spôsoby ako sa vysporiadať s týmto problémom. Jedným z nich je nastavenie striktnejších parametrov vo fáze zarovnania za cenu redukcie efektivity mapovania jednotlivých *readov*. Druhým - lepším spôsobom je orezať sekvencie od prvého výskytu *base-call* s nízkou kvalitou ešte pred zarovnaním [33, 52].

Zlepšenie výsledkov zarovnania je ďalej možné dosiahnuť dodatočným odstránením adapterových sekvencií, ktoré môžu byť taktiež osekvenované, ak je veľkosť DNA fragmentov menšia ako dĺžka *readov*. Ponechaním týchto sekvencií by sa narušilo mapovanie reťazcov a u niektorých nástrojov pre zarovnanie je tento krok povinný. Okrem zníženej efektivity mapovania sú adapterové sekvencie úzko spojené s chybami na úrovni extrakcie metylácie. Obecne platí, že korektné predspracovanie sekvencií vedie k podstatne kvalitnejšiemu mapovaniu a presnejšiemu určovaniu úrovne metylácie [33, 52].

4.1.2 Zarovnanie

Po ošetrení genómovej DNA hydrogénsiričitanom sodným je komplexita sekvencií výrazne redukovaná, keďže všetky cytozíny, s výnimkou metylovaných, budú prevedené na thymíny. Táto redukcia výrazne komplikuje proces zarovnania (detekcia genómového lokusu odkiaľ *read* pochádza). Mnoho existujúcich programov akceptuje iba unikátne zarovnanie, avšak

jeden *read* s univerzálnym zarovnaním v 4-písmenovej abecede sa môže mapovať na viac miest v 3-písmenovej abecede. Naviac dochádza k zväčšeniu prehľadávacieho priestoru, pretože Watsonove a Crickove vlákno nie sú po ošetrení hydrogénsiričitanom sodným k sebe navzájom komplementárne (k zmene dochádza len u cytozínov), a teda obidve vlákna budú mať svoje vlastné reverzné komplementy.

Na základe týchto skutočností nemôžu byť jednotlivé *ready* zarovnávané iba voči referenčnej sekvencii. V súčasnosti existujú dva prístupy, ktoré sa používajú pre riešenie tejto fázy spracovania bisulfitových sekvencií. Prvý prístup využíva maticu s rovnakými váhami pre zhody a pre C/T nezhody (cytozín v referenčnej sekvencii a T v *reade*). Druhý pracuje s 3-písmenovou abecedou, pomocou ktorej prispôsobí referenčnú sekvenciu k redukovanej komplexite bisulfitových *readov* [52].

4.1.3 Analýza metylačných miest

Zo zarovnaných sekvencií voči referenčnému genómu je možné extrahovať metylačnú informáciu jednotlivých cytozínov. Najprv je však nutný prevod naspäť do 4-písmenovej abecedy. Metylované cytozíny sú dané C/C zhodami, zatiaľ čo nemetylované cytozíny prostredníctvom T/C nezhôd zarovnania. Metylačná úroveň danej pozície cytozínu v genóme je určená zo všetkých *readov*, ktoré prekrývajú túto pozíciu ako podiel počtu mC a celkového počtu takýchto *readov*. Úroveň metylácie je teda číslo od 0 (kompletne nemetylované) do 1 (kompletne metylované) a určuje sa vo všetkých kontextoch CG, CHG a CHH [52].

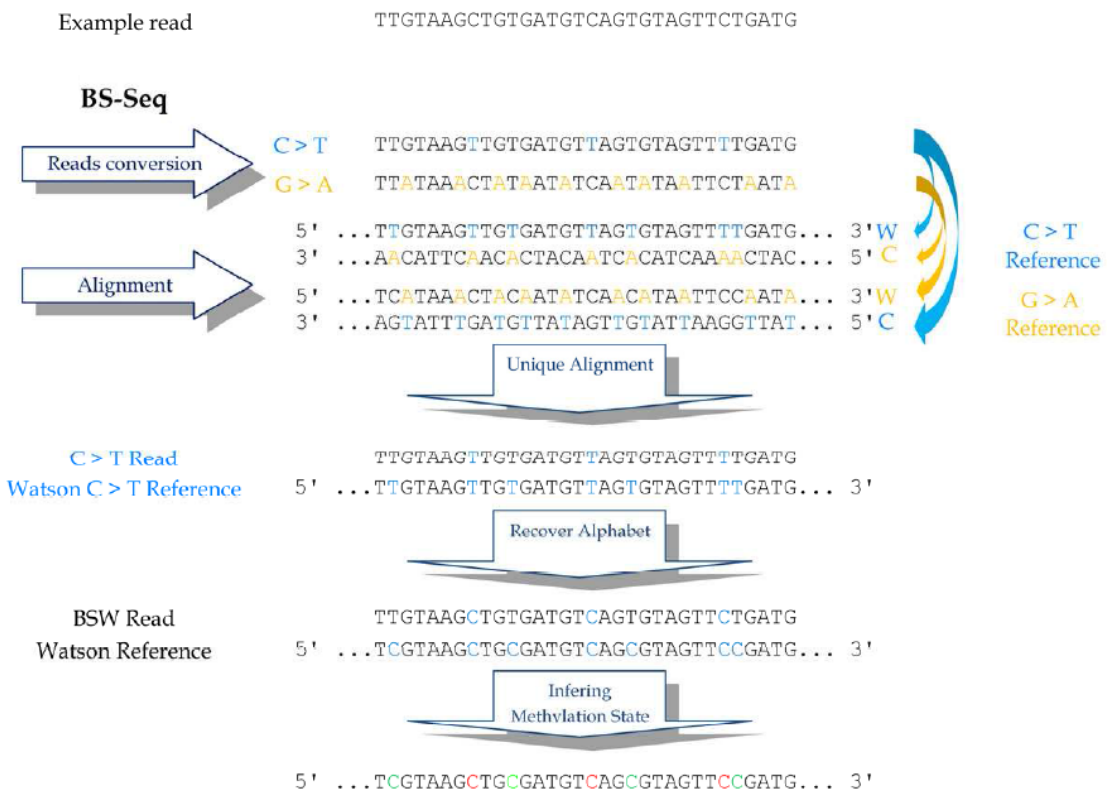
Postup procesu zarovnania a extrakcie metylácie zo sekvencii je zobrazený na obrázku 4.2.

4.1.4 Kontrola kvality

Kvalitu metylačných máp môže ovplyvniť niekoľko zdrojov chýb, ku ktorým predovšetkým patrí (1) nesprávne zarovnanie *readov*, (2) existencia jednonukleotidových variant (angl. *single-nucleotide variants* - SNV), (3) sekvenovacie chyby a (4) bisulfitové zlyhanie.

Nesprávne zarovnanie

Redukovaná komplexita sekvencií bisulfitových *readov* a existencia sekvenovacích chýb môžu viesť k zarovnaniu sekvencie na genómovú pozíciu, z ktorej nepochádza. Toto je častý prípad pre vysoko repetitívne DNA sekvencie, ktoré sú väčšinou bohaté na CGI (ako napríklad transpozóny) a zároveň metylované. Momentálne je najlepším spôsobom pre obmedzenie počtu nesprávnych zarovnaní vhodná voľba parametrov pre daný nástroj. Každý z nich má svoje vlastné nastavenia, ktoré sú špecifické pre rôzne algoritmy. Dva konkrétne parametre budú však najdôležitejšie pre náš problém. Ide o počet povolených nezhôd a minimálna dĺžka zarovnania. Vyšší počet povolených nezhôd a malá dĺžka zarovnávacieho semienka bude viesť k mapovaniu viacerých *readov* na genóm, avšak počet nesprávnych zarovnaní taktiež vzrastie. Na druhú stranu, striktnejšie nastavenie parametrov zabráni zarovnaniu mnohých odpovedajúcich *readov* a veľa genómových regiónov tak zostane neprofilovaných [52].



Obrázok 4.2: Profilovanie metylácie z bisulfitom ošetrených *readov*. Aby bolo možné získať túto informáciu, bisulfitové *ready* musia byť najprv namapované na referenčný genóm. V tomto procese sú nemetylované cytozíny prekonvertované na thymíny. Dochádza tak k redukcii sekvenčnej komplexity, čo je možné vyriešiť prostredníctvom konverzie referenčného genómu taktiež do 3-písmenovej abecedy. *Ready* sa môžu mapovať na Watsonove a Crickove C na T konvertované referencie a súčasne aj na Watsonove a Crickove G na A konvertované referencie (reverzné komplementy). Preto musia byť *ready* konvertované v dvoch abecedách: C na T zmenené *ready* budú mapované na C na T konvertované referencie (modré šípky) a G na A zmenené *ready*, ktoré budú mapované na G na A konvertované referencie (žlté šípky). Ak existuje unikátne zarovnanie, je možné priamo vyčítať metylačnú informáciu: C/T nezhoda udáva nemetylovaný cytozín (zelená farba) a cytozín v referenčnej sekvencii aj v *reade* udáva metyláciu (červená farba). Pri reverznom komplemente guanín v referenčnej sekvencii a adenín v *reade* značí nemetylovaný cytozín, pričom guanínová zhoda udáva metyláciu [52].

Detekcia SNV

V prírode majú metylované cytozíny väčšiu šancu zmeniť svoju štruktúru na thymín alebo uracil (viď sekcia 2.4.3) ako nemetylované C. Kvôli tejto reakcii je C>T najčastejšou variáciou medzi referenčným a sekvenovaným genómom (C u referenčného sa často objavuje na rovnakej pozícii ako T pri *reade*). Takéto variácie sa označujú ako *Single Nukleotide Variants* (SNV). Špeciálne pri bisulfitových dátach, kde dochádza k meraniu C>T kon-

vertovaných báz, môže mať táto skutočnosť veľmi výrazný vplyv, pretože takáto variácia môže byť ľahko chybne interpretovaná ako bisulfitová konverzia. Ignorovaním prítomnosti podobnej sekvenčnej variácie by viedlo k označeniu cytozínu v referenčnom genóme ako nemetylovaný. Správne by však nemal byť nájdený žiadny cytozín a nemal by byť detekovaný žiadny metylačný stav. Tento problém je ešte významnejší u transpozónov, u ktorých sa C>T mutácia vyskytuje s vyššou frekvenciou [52].

Sekvenovacie chyby

Ďalšou možnou chybou pri získavaní metylačných máp je chybné označenie bázy ako thymín namiesto cytozínu. Táto chyba by bola nesprávne interpretovaná ako nemetylovaný cytozín. Sekvenovacie chyby vznikajú behom sekvenačného procesu a môžu sa prejaviť ako zle označená báza alebo naviac pridané alebo odobrané bázy. Dohromady je reálnych dvanásť rôznych substitučných chýb, ale analýza eukaryotických datasetov ukazuje iba 2% pre C>G substitúcie, zatiaľ čo T>C substitučná chyba sa objavuje vo viac ako 15% prípadov. C>T substitúcie sa objavujú znova len s nízkou 4% frekvenciou [32]. Z dôvodu C>T konverzie bisulfitových dát obsahujú sekvenované dáta veľké množstvo T a sekvenačné chyby by tak mohli spôsobiť odchýlku pri extrakcii metylácie, a preto musia byť adresované. Napríklad C>T alebo T>C (na konvertovaných cytozínoch) sekvenačné chyby by boli nesprávne interpretované ako nemetylované respektíve metylované, a tým ovplyvňovali úroveň metylácie k nižším alebo vyšším hodnotám. Našťastie technológie pre NGS často poskytujú pre každú bázu *readov* kvalitu sekvenovania vo forme tzv. PHRED skóre, ktoré sa interpretuje ako pravdepodobnosť, že daná báza bola nesprávne určená [52]. Bázové kvality sú dané podľa:

$$Q = -10 \log_{10} P(basecall\_chyba) \tag{4.1}$$

Táto informácia je uložená spolu so sekvenciou *readu* vo FASTQ formáte (viď sekcia 5.3.1) a môže byť ďalej použitá pre zabránenie ovplyvňovania výsledku chybnými bázami.

Bisulfitové zlyhanie

Chybná bisulfitová konverzia môže byť spôsobená nekompletnou denaturáciou pred aplikovaním hydrogénsiričitanu sodného a niektoré cytozíny tak zostanú nekonvertované nezávisle na ich metylačnom stave. Ak nie je táto situácia detekovaná, všetky takéto cytozíny budú označené ako metylované [52].

4.2 Bioinformatické nástroje

Extrakcia metylačných máp je viackrokový proces, kde každá jednotlivá fáza vyžaduje odlišnú funkcionalitu. V súčasnosti existuje mnoho rôznych nástrojov realizujúcich jednu alebo viac častí tohto procesu a obecne ich môžeme rozdeliť do troch kategórií:

1. Bisulfitové nástroje vykonávajúce len zarovnanie, ktoré nedokážu extrahovať metylačné mapy.
2. Nástroje umožňujúce len extrakciu metylácie na špecifických pozíciách.
3. Nástroje realizujúce celý vyššie popísaný proces (niektoré aj s chybovou kontrolou).

Neexistuje však nástroj, ktorý by bol celý priamo zameraný na analýzu metylómu rastlinnej DNA s vyšším podielom transpozónových sekvencií a nukleotidových mutácií. Na základe týchto skutočností je dôležité vytvorenie nástroja zohľadňujúceho tieto vlastnosti.

Využívanie NGS metód s bisulfitovou konverziou umožňuje meranie úrovne DNA metylácie celého genómu, ale zároveň predstavuje unikátne výzvy pre spracovanie a interpretáciu metylačných údajov z dôvodu ich veľkého objemu a dôsledkov bisulfitovej modifikácie. Pre tieto účely vzniklo veľa programov, ktoré zvládajú spracovanie jednotlivých fáz. Ich výkon, použiteľnost a presnosť sa však často odlišujú. Hlavné rozdiely, ktoré je možné pri nich pozorovať sú predovšetkým v kontrole kvality sekvenovaných *readov* a výsledkov jednotlivých procesov, spôsob zarovnania sekvencií k referenčnému genómu a v princípe extrakcie metylačných miest.

Zoznam dostupných nástrojov pre spracovanie a analýzu bisulfitových dát je možné nájsť v [20, 14]. Analýza ich výkonu je väčšinou zameraná na relevantnosť mapovania readov v rôznych sekvenčných kontextoch a vytváranie presných metylačných máp (unikátne mapovanie, globálna úroveň metylácie), výpočetné prostriedky a doba behu algoritmu. Ide predovšetkým o výsledky fázy zarovnania sekvencií a extrakcie metylačných miest, pričom kontrola kvality v predspracovaní je väčšinou realizovaná špeciálnymi nástrojmi pre určenie skóre kvality (napr. FastQC [10]) a správnej dĺžky *readov* (napr. Cutadapt [39]).

4.2.1 Bisulfitové mapovanie

Z hľadiska mapovania bisulfitových *readov* rozlišujeme celkovo dva odlišné prístupy pri spracovaní bisulfitových dát: (1) mapovanie s divokou kartou (*wild-card*) a (2) 3-písmenové mapovanie.

Wild-card mapovanie

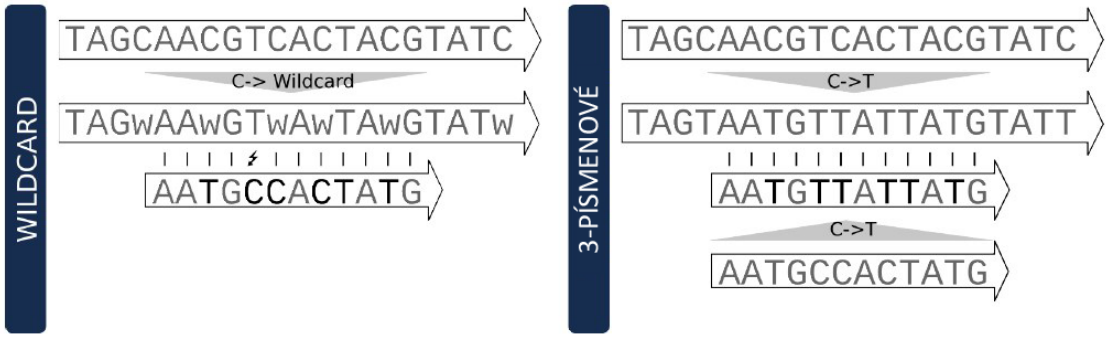
Prvý z nich využíva všetky dostupné informácie a mapovanie genomického C a T *readu* počíta ako zhodu, zatiaľ čo mapovanie genomického T a C *readu* ako nezhodu. Toto je realizované s použitím pridaného *wild-card* písmena pre referenčné C zhodujúce sa s C a T. Tento typ mapovania sa vyznačuje relatívne veľkou senzitivitou. Jeho nevýhodou je však možnosť ovplyvnenia výsledku, pretože *ready* s vyšším pomerom C môžu byť mapované s vyššou účinnosťou [32].

BSMAP

Najznámejším nástrojom v tejto kategórii je BSMAP. Využíva pre zarovnanie program SOAP (*Short Oligonucleotide Alignment Program*) a indexuje referenčný genóm pre všetky možné sekvencie určitej dĺžky. BSMAP maskuje T v bisulfitových *readoch* ako C iba na C pozíciách referenčného genómu a ostatné T sú v bisulfitových *readoch* nezmenené. Potom mapuje takto maskované *ready* priamo na referenčný genóm. Je veľmi citlivý, ale má oveľa dlhší beh ako iné nástroje [53].

LAST

Ďalším veľmi kvalitným programom je LAST [22], originálne vytvorený pre mapovanie klasických (nie bisulfitových) *readov*. Bol modifikovaný pre bisulfitové dáta zavedením nových skóre hodnotení s ohľadom na C/T konverzie. Autori tohto programu dokonca ukázali, že je citlivejší a oveľa rýchlejší ako BSMAP. LAST je využívaný v nástroji Bisulfighter [57] a ten používa jeho schopnosť určovať pravdepodobnosť úspešnosti mapovania k filtrovaniu zarovnaní a určovaniu metylačnej úrovne.



Obrázok 4.3: Rozdiel *wild-card* a 3-písmenového zarovnania. *Wild-card* umožňuje cytozínu a thymínu v sekvenčnom čítaní zarovnanie voči referenčnému C, pričom C čítania sa pri mapovaní na referenčné T počíta za nezhodu. U 3-písmenového mapovania je táto informácia stratená a skutočné C/T nezhody nie sú viditeľné [32].

3-písmenové mapovanie

V druhom prístupe sú referenčné sekvencie a *ready* konvertované do 3-písmenovej abecedy s C a T vystupujúcimi ako jedno písmeno. Dôsledkom je mapovanie genomického T oproti C *readu* ako zhoda. Mapovanie v 3-písmenovej abecede vedie k mierne zníženej efektivite zarovnania hlavne kvôli zväčšeniu prehľadávacieho priestoru a viac chybných pozitívnych nálezov. Výhodou je naopak to, že môžeme predpokladať nijak neovplyvnený proces extrakcie metylačnej úrovne. Nástroje tejto kategórie majú ešte možnosť využívať externých, kvalitných a odskúšaných aplikácii pre zarovnávanie len s pridaním fáz predspracovania a spracovania. Ukážka rozdielu medzi týmito dvomi metódami je na obrázku 4.3.

Bismark
: Bismark je pravdepodobne najpoužívanejším nástrojom v tejto kategórii, ktorý využíva pre mapovanie nástroj Bowtie (prípadne Bowtie2). V predspracovaní prevedie *ready* a referenčné sekvencie do 3-písmenovej abecedy. V prípade *directional single-end readov* sú všetky C>T konvertované. Tie sú potom mapované na C>T a G>A konvertované referenčné sekvencie reprezentujúce bisulfitové zmeny na vrchnom a spodnom vlákne. Dve inštancie Bowtie algoritmu teda bežia súčasne. U *pair-end readov* Bismark naviac prevádza *ready* na G>A verziu aby pokryl *ready* z reverzných komplementov originálnych vlákien. V spracovaní vyfiltruje *ready* s veľkým množstvom skutočných C/T nezhôd [53].

BS-Seeker
: BS-Seeker používa veľmi podobný prístup, ale naviac zahadzuje *ready* mapujúce sa na obidve vlákna a tie, ktoré majú veľa reálnych C/T nezhôd. Obidva podporujú *directional* aj *non-directional* protokoly a BS-Seeker tiež využíva Bowtie2 ako vnútorný mapovací algoritmus. BS-Seeker pracuje iba so *single-end readmi*, zatiaľ čo Bismark aj s *pair-end readmi*. Ich výkonnosť sa taktiež príliš nelíši pre doménu, v ktorej obidva fungujú [53].

MethylCoder
: MethylCoder [44] je flexibilnejší vo svojej činnosti. Pre mapovanie využíva buď Bow-

tie2 alebo GSNAP a zvyšné spracovanie je veľmi podobné predchádzajúcim dvom aplikáciám.

4.2.2 Extrakcia metylácie

Niektoré z predchádzajúcich popísaných programov, ako Bismark, MethylCoder alebo Bisulfighter, naviac vykonávajú odhad úrovne metylácie pre genomické C pozície na základe C/T pomeru mapovaných *readov*. Ide však o dosť nepresnú metódu, pretože dáta môžu byť podstatne ovplyvnené sekvenačnými chybami, genomickými variáciami alebo bisulfitovou chybou a tieto programy nepracujú so skoro žiadnym z týchto (a vyššie popísaných) zdrojov chýb. Veľa z nich je avšak veľmi presných, predovšetkým Bismark a Bisulfighter [57], ani jeden ale nedokáže rozoznávať SNV, na ktoré sú transpozóny špeciálne bohaté.

Bis-SNP

Jedným z prvých nástrojov adekvátne adresujúcich problém SNV je Bis-SNP [54]. Ide o sadu (celkovo 10) skriptov pre realizovanie celého procesu spracovania bisulfitových dát od mapovania až po extrakciu metylácie spolu s kontrolou kvality báz a je postavený na *Genome Analysis Toolkit* (GATK). Využíva Bayesovský prístup a na základe informácií z vrchného a spodného DNA vlákna rozlišuje SNV a bisulfitové konverzie. C>T variácie tak nebudú interpretované ako nemetylovaný cytozín. Keďže u *non-directional* protokolu nie je možné určiť, z ktorého vlákna *ready* pochádzajú, Bis-SNP podporuje iba *directional* BS-Seq protokol [32].

MethylExtract

MethylExtract je druhým z nástrojov schopných generovať vysoko kvalitné metylačné mapy na celom genóme a súčasne rozpoznáva aj sekvenčné variácie (SNV). Metyláciu kontroluje vo všetkých možných kontextoch (CG, CHG a CHH) a implementuje najviac metód odpovedajúcich kontrole kvality. Narozdiel od Bis-SNP je jeho beh realizovaný len prostredníctvom jedného skriptu, a preto je z uživateľského pohľadu jednoduchší na používanie [13].

4.2.3 Porovnanie nástrojov

Programov pre spracovanie bisulfitových dát je mnoho a vybrať spomedzi nich ten najvhodnejší nie je ľahká úloha. Každý z dostupných nástrojov môže byť vhodný pre jeden typ úloh a nevhodný pre iný. Ideálne by bolo nájsť nástroj, ktorý vykazuje vynikajúce ako časové, tak aj kvalitatívne (z hľadiska extrakcie metylácie) výsledky pre mapovanie repetitívnych sekvencií alebo aspoň zvláda čo najväčšie genómové pokrytie.

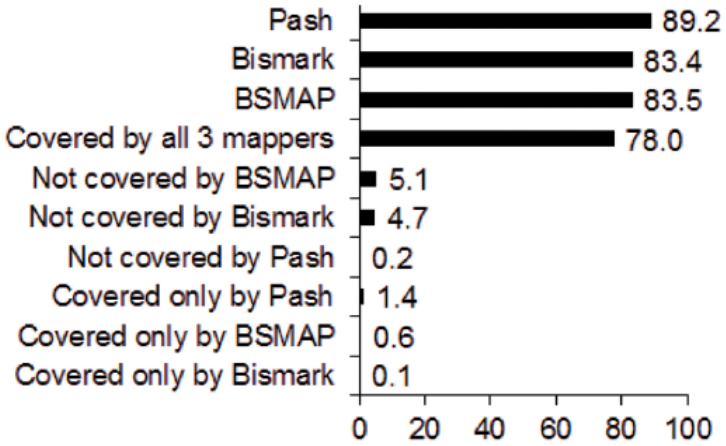
V [34] autori porovnávali niekoľko z vyššie popísaných nástrojov, konkrétne Bismark, BSMAP, Pash, BS-Seeker a BatMeth na reálnych údajoch ľudskej DNA dvoch rôznych tkanív a na simulovaných sekvenciách. Vyhodnocovali ich výpočetnú rýchlosť a pokrytie genómu, a overovali percentuálne odhady metylácie. Zameriavali sa predovšetkým na prvé tri z vyššie menovaných s tým, že BS-Seeker dosahoval veľmi podobné výpočetné výkony ako Bismark pre dané sekvencie z hľadiska mapovania. Čas výpočtu jednotlivých programov sa výrazne líšil vo fázach mapovania reťazcov a analýze metylácie, kde BatMeth bol z nich najrýchlejší a za ním Bismark. Pash a BS-Seeker boli výrazne pomalšie (tab. 4.1) [34].

Všetky tri hlavné algoritmy poskytovali výborné pokrytie a súhlasné odhady CpG metylácie na celom genóme. Každý z nich pokryl viac ako 80% CpG miest, pričom Pash si viedol najlepšie ako z hľadiska pokrytia, tak aj z hľadiska miest, ktoré pokryté neboli (obr.

Algoritmus	Mapovanie	Analýza metylácie	Celkovo
Bismark	1514	81	1595
BSMAP	800	1081	1881
Pash	3486	1504	4990
BS-Seeker	1324	3867	5191
BatMeth	904	70	974

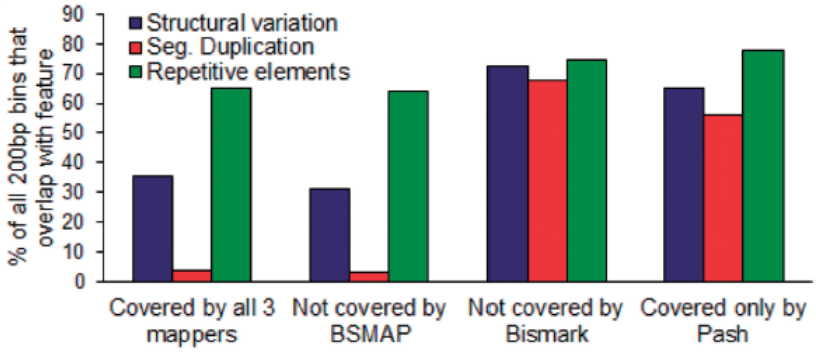
Tabulka 4.1: Porovnanie času (v sekundách) mapovania a analýzy metylácie pre 1 milión *readov*. Analýza metylácie zahŕňa čas potrebný pre odhad percentuálnej metylácie každého metylovaného cytozínu [34].

4.4). BatMeth, i keď veľmi rýchly, unikátne namapoval len približne 50% *readov* a BS-Seeker mapoval približne 80% *readov*, viac než 98% z toho na rovnaké pozície ako Bismark.



Obrázok 4.4: Porovnanie percentuálneho pokrytia CpG miest na celom genóme [34].

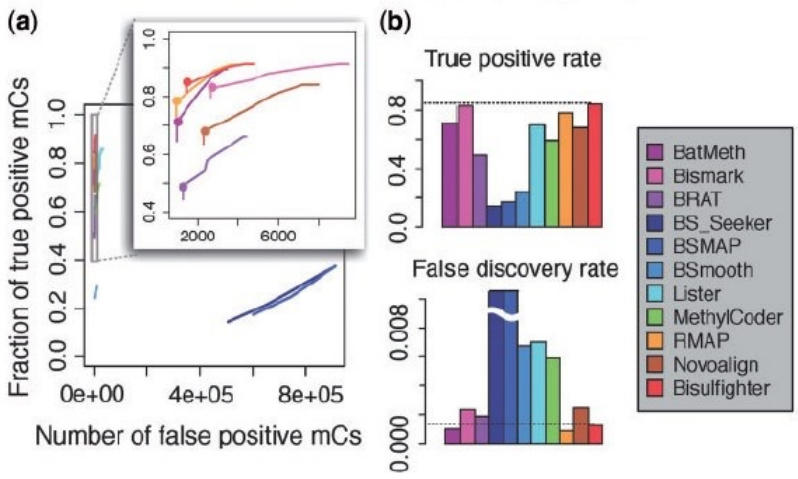
V ďalších testoch BSMAP nepokrýval najväčšiu časť testovaných sekvencií (8-12%), Bismark podával skoro o polovicu lepšie výsledky (5-6%) a Pash bol v týchto experimentoch najlepší s nepokrytím <1%. Väčšina regiónov pokrytých programom BSMAP a Pash, ale nie Bismark, boli regióny so štruktúrnymi variáciami a segmentovými duplikáciami (obr. 4.5). Táto vlastnosť sa v simuláciách potvrdila len pre Pash. Regióny unikátne namapované Bismarkom a Pashom, ale nie BSMAPom, boli chudobné na G a bohaté na T. Po bisulfitovej konverzii sa ešte zväčší počet T, a tým sa podstatne znižuje komplexita sekvencie, čo robí toto mapovanie náročným.



Obrázok 4.5: Percentuálny pomer pokrytia 200-bp úsekov obsahujúcich CpG prekrývajúcich sa s odlišnými genómovými rysmi - štruktúrnymi variáciami, segmentovými duplikáciami a repetitívnymi elementami [34].

Bismark predstavoval nástroj poskytujúci výbornú kombináciu rýchlosti spracovania dát, pokrytia genómu a presnosti mapovania, zatiaľ čo Pash bol o niečo presnejší, predovšetkým v regiónoch so štruktúrnymi variáciami, ale výpočetne pomalší. BS-Seeker poskytoval rovnakú presnosť ako Bismark s dlhším výpočtom, vyznačuje sa však detailnejšími informáciami vo fáze analýzy metylácie [34].

Autori nástroja Bisulfighter porovnávali z hľadiska úspešnosti extrakcie metylácie jeho výkonnosť s ostatnými voľne dostupnými nástrojmi, konkrétne BatMeth, Bismark, BRAT-BW, BS Seeker, BSMAP, BSmooth [24], Lister [38], MethylCoder, RMAP [50] a Novoalign. Konkrétne ich zaujímala úroveň pravdivých pozitívnych nálezov a chybných pozitívnych nálezov. Na obrázku 4.6 je vidieť, že Bisulfighter aj Bismark boli jedné z najlepších nástrojov a viedli si veľmi podobne v lokalizácii pravdivých pozitívnych zhôd. Čo sa týka chybných pozitívnych nálezov, tak ako Bisulfighter, aj Bismark podával relatívne dobré výsledky (i keď už nie najlepšie) [57].

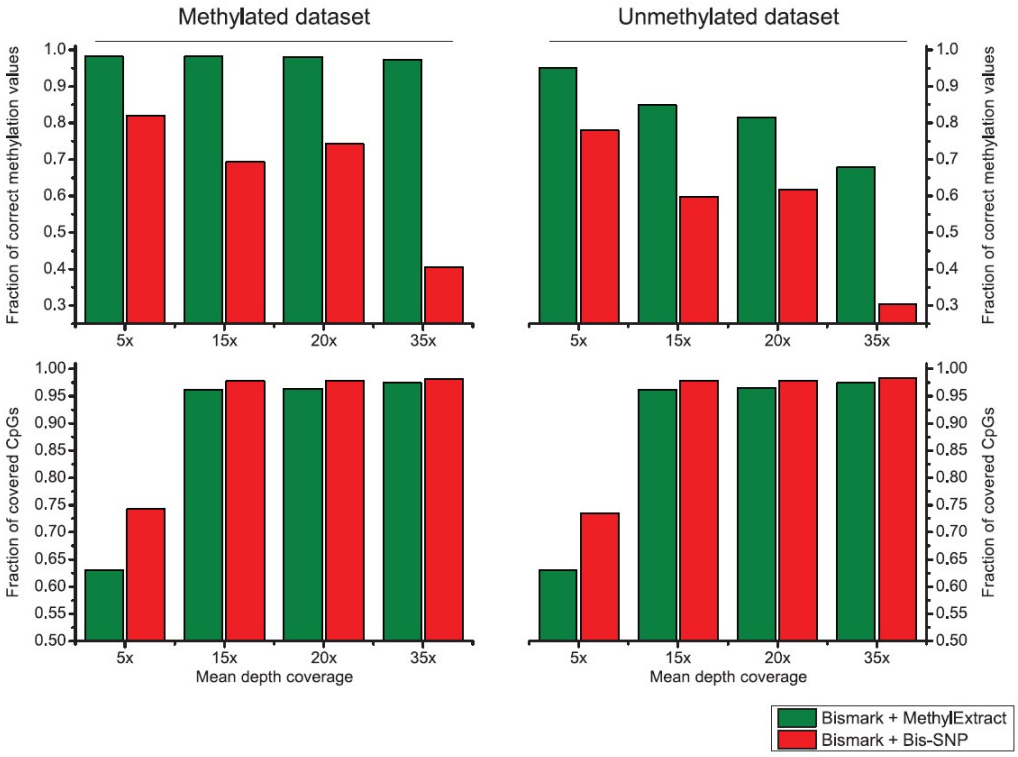


Obrázok 4.6: Porovnanie výkonu extrakcie metylácie. (a) Kompromis medzi úrovňou pravdivých pozitívnych zhôd a chybných pozitívnych zhôd pre rôzne sekvenčné hĺbky. (b) Úroveň pravdivých pozitívnych zhôd a úroveň chyby [57].

Výkon týchto nástrojov je výrazne závislý od konkrétneho genómu a cieľových regiónov,

obecne ale Bismark poskytuje veľmi kvalitné mapovanie s relatívne rýchlym výpočtom v porovnaní s inými, bežne používanými aplikáciami. Pre zarovnanie využíva dve alebo štyri inštancie Bowtie algoritmu v závislosti od typu *readov*. Je jedným z najpoužívanejších nástrojov pre spracovanie bisulfitových dát.

Vyššie popísané porovnania sa síce týkali z veľkej časti extrakcii metylácie, veľmi dôležitou časťou tohto procesu je však samotné zarovnanie sekvencií a môžeme predpokladať, že nástroje s korektnými odhadmi metylácií budú taktiež excelovať aj v mapovaní. Práve časť zarovnania je z pohľadu tejto práce u popísaných nástrojov relevantná, pretože žiadny z nich nedokáže detekovať SNV vo fáze extrakcie metylácie, čo je u transpozónov nevyhnutné. Boli uvedené dva programy, ktoré toto dokážu, Bis-SNP a MethylExtract, a pri ich porovnaní [13] bol Bis-SNP mierne viac špecifickejší (medzi 1.9% a 3.9% vyššia pozitívna predikčná hodnota) a MethylExtract zas viac citlivý (medzi 1% a 3.1% vyššia senzitivita). Porovnaním časti korektne získaných metylačných hodnôt sa však medzi nimi objavili podstatné rozdiely (obr. 4.7). MethylExtract dokáže aj pri volnejšom nastavení parametrov získať väčšie podiely aj napriek tomu, že Bis-SNP vie pokryť viac pozícií (časť pokrytých CpG). Celkovo dokázal MethylExtract získať až o 20% viac korektne profilovaných metylácií.



Obrázok 4.7: Porovnanie metylačných hodnôt na CpG medzi MethylExtractom a Bis-SNP. Obidve metódy sú porovnané z hľadiska časti korektne profilovanej CpG metylácie (vrchná časť) a časti pokrytých CpG pozícií (spodná časť) pri kompletne metylovanom a kompletne nemetylovanom datasete [13].

Kapitola 5

Nástroj na lokalizáciu metylácie

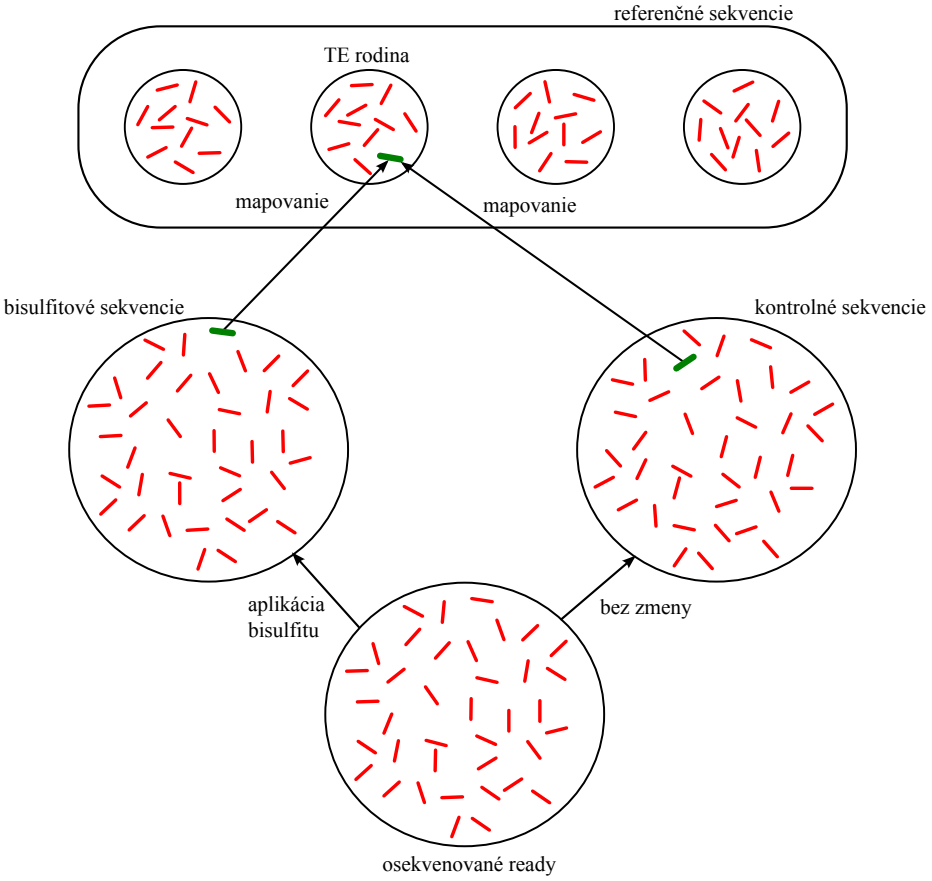
V predchádzajúcej kapitole bolo popísané z akých častí proces spracovania bisulfitových dát pozostáva a aké nástroje sú v nich využívané. V tejto kapitole bude podrobne analyzovaný problém lokalizácie metylačných miest transpozónov a rozobratá celá fáza bisulfitového mapovania a extrakcie metylácie vrátane všetkých krokov v nich realizovaných.

5.1 Analýza lokalizácie metyláčných miest transpozónov

Získanie metylačných máp jednotlivých TE umožní analyzovať úroveň metylácie pre rôzne rodiny transpozónov (reprezentované klastrami sekvencií) v odlišných fázach vývoja organizmu. Pre tieto účely sa ako referenčné sekvencie použijú skupiny príbuzných DNA fragmentov reprezentujúcich rodiny transpozónov obsahujúcich jednotlivé TE danej rodiny. Voči každej z týchto sekvencií sa bude aplikácia snažiť zarovnať niektorý zo sekvenovaných *readov*. Tieto sekvenčné čítania budú rozdelené celkovo do dvoch častí. Na jednu z nich bude aplikovaný bisulfit, všetky metylované cytozíny ostanú nezmenené, pričom ostatné sa zmenia procesom deaminácie na uracil (sekvenované ako thymín). Druhú časť tvoria kontrolné sekvencie, čo sú rovnaké sekvencie ako v bisulfitovej skupine, ale bez aplikácie bisulfitu. Ich účelom bude verifikácia korektnosti mapovania bisulfitových *readov* na referenčné sekvencie. Z vybraných korektne namapovaných *readov* bude vo fáze 2 extrahovaná metylačná úroveň každého nájdeného cytozínu referenčných transpozónových sekvencií. Popísaný problém je graficky zobrazený na obrázku 5.1.

Lokalizácia metylačných miest transpozónov je teda niekoľkokrokový proces:

1. Mapovanie bisulfitových *readov* na referečné sekvencie klastru.
2. Filtrovanie a verifikácia namapovaných *readov* s pomocou kontrolných *readov*.
3. Extrakcia metylácie cytozínov vzhľadom k referenčným sekvenciám.
4. Spracovanie dát na úrovni klastrov transpozónových rodín.



Obrázok 5.1: Lokalizácia úrovne metylácie transpozónov z bisulfitových dát voči referenčným TE. Kontrolné sekvencie sa používajú pre verifikáciu správnosti mapovania bisulfitových *readov* na referenčné sekvencie.

5.1.1 Mapovanie a filtrácia

V predchádzajúcej kapitole boli popísané a porovnané prístupy a nástroje používané k mapovaniu bisulfitových *readov*. Pre mapovanie *readov* k transpozónovým sekvenciám je dôležité, aby bolo možné nastavovať citlivosť zarovnania, pretože aj mapovanie s niekoľkými nezhodami môže byť korektné kvôli vyššiemu pomeru mutácií na jednom nukleotide (SNV). Výsledok zarovnania musí obsahovať dostatočné množstvo informácií pre filtrovanie chybných pozitívnych nálezov. Rýchlosť a uživateľská použiteľnosť hrajú taktiež významnú úlohu pri obsiahlych datasetoch bisulfitových dát. Na základe týchto špecifikácií bolo možné buď použiť existujúci nástroj splňujúci dané podmienky alebo vytvoriť nový nástroj s týmito vlastnosťami.

Spôsob zarovnania

U dvoch možných prístupov spracovania bisulfitových dát (*wild-card* a 3-písmenové mapovanie) sa ich rozdiely najviac prejavia práve v genomických regiónoch s vysokou úrovňou sekvenčnej identity s inými časťami genómu, akými sú napríklad retrotranspozóny. *Wild-card* nástroje majú tendenciu ovplyvňovať metylačnú úroveň k vyšším hodnotám, pretože extra metylované C môžu zvýšiť sekvenčnú komplexitu tak, že bude dostatočná pre unikátne zarovnanie, zatiaľ čo odpovedajúci *read* obsahujúci T (nemetylovaný C) bude vyradený.

Navyše, vo vyššie popísanom porovnaní bolo vidieť, že BSMAP (najpoužívanejší zástupca *wild-card* metódy) má problém s mapovaním *readov* bohatých na T, ktoré sa vo vyššej miere vyskytujú v transpozónoch. Na druhú stranu nástroje s 3-písmenovou stratégiou mapovania znižujú sekvenčnú komplexitu, čím môže dôjsť k zahodeniu väčšieho počtu *readov*. Na základe týchto skutočností bol pre zarovnanie vybraný 3-písmenový prístup s tým, že jeho citlivosť bude zvýšená pre lokalizáciu väčšieho počtu transpozónových mapovaní.

Nástroj na zarovnanie

Z nástrojov pracujúcich s 3-písmenovou abecedou vykazoval v porovnaní nástrojov najlepší výkon Bismark, ktorý využíva k mapovaniu rýchly a presný program Bowtie, kde je možné zvyšovanie jeho citlivosti prostredníctvom menenia minimálneho skóre potrebného k akceptovaniu daného zarovnania. Bowtie je program kombinujúci presnosť a rýchlosť, v presnosti nepatrí ale medzi tie najkvalitnejšie, a preto som testoval vhodnosť použitia presnejšieho zarovnávacieho nástroja na tomto mieste workflow. Konkrétne išlo o BLAT [30], ktorý je síce pomalší, ale dokáže nájsť viac zarovnaní ako Bowtie a s väčšou presnosťou v určitých typoch problémov. Pri zarovnávaní simulovaných bisulfitových *readov* s 15% výskytom SNV k referenčným sekvenciám organizmu *Silene latifolia* skutočne dosiahol nálezu väčšieho počtu pravdivých pozitívnych zhôd - *ready* namapované korektne tvorili 46% *readov*, ktoré sa mali namapovať korektne. U Bowtie s predvoleným nastavením parametrov predstavovala táto časť len 31%. BLAT našiel ale rádovo väčšie množstvo chybných pozitívnych nálezov a pri voľnejšom nastavení (znížení minimálneho skóre pre akceptovanie zarovnania) citlivosti, Bowtie algoritmus dosahoval porovnateľné výsledky pravdivých pozitívnych zhôd ako BLAT (42,3%) a stále držal chybné pozitívne nálezy v rovnakom ráde ako predtým.

Na základe týchto skutočností bol pre realizáciu fázy mapovania vybraný nástroj Bismark vybavený mapovacím nástrojom Bowtie. Bismark naviac realizuje filtrovanie nájdených mapovaní takým spôsobom, že vo výsledku ostanú len unikátne zarovnané sekvencie. Pri striktnejšom nastavení parametrov vie tento spôsob veľmi efektívne odstrániť väčšinu chybne pozitívnych nálezov, u transpozónových sekvencií je však nutná väčšia citlivosť, čo vedie k väčšiemu množstvu chybne pozitívnych nálezov. Pre filtráciu a verifikáciu správnych pozitívnych nálezov bude vytvorený špeciálny modul, ktorý bude hľadať zarovnanie kontrolných sekvencií k referenčným v klasickej 4-písmenovej abecede, čím sa redukuje rozšírený prehľadávací priestor spôsobený bisulfitovou konverziou a prevodom do 3-písmenovej abecedy.

5.1.2 Extrakcia metylácie a spracovanie dát

Priama extrakcia metylácie je ďalším problémom, ktorý je adresovaný širokým spektrom dlhodobo vyvíjaných a odladených nástrojov. U transpozónov je najväčší problém s rozoznávaním nukleotidových variant. Schopnosť získať úroveň metylácie vo všetkých troch kontextoch (CG, CHG a CHH) na jednotlivých pozíciách tak, aby tieto hodnoty neboli ovplyvnené vzniknutými mutáciami je základnou funkcionalitou, ktorú nástroj v tejto fáze musí mať implementovanú. Naviac musí byť schopný pracovať s výstupom predchádzajúcej fázy a jeho výsledky by mali byť čo najmenej ovplyvnené možnými chybami, ku ktorým dochádza pri sekvenácii. Všetky tieto parametre spĺňajú (do určitej, prijateľnej miery) v súčasnosti existujúce dva programy - Bis-SNP a MethylExtract. V predchádzajúcej sekcii boli tieto dva nástroje porovnané a na základe tohto porovnania, uživateľskej použiteľnosti a implementovanej funkcionality bol vybraný dobre zavedený nástroj MethylExtract pre realizovanie fázy extrakcie metylácie z nájdených zarovnaní bisulfitových readov.

MethylExtract lokalizuje pre každú referečnú sekvenciu klastru transpozónovej rodiny úroveň metylácie cytozínov na jednotlivých pozíciách. Pre naše účely je však nutné určiť celkovú metylačnú úroveň danej rodiny pre všetky kontexty tak, ako aj metylačnú mapu celej konsenzuálnej sekvencie vytvorenej z kontigov klastru, ktoré tu vystupujú ako referenčné sekvencie. Ide teda o spracovanie metylačných dát nástroja MethylExtract do takej podoby, aby z nich bolo možné vyvodiť potrebné závery. K tomuto účelu bude slúžiť špeciálny modul pre spracovanie a analýzu metylačných údajov.

V tabuľke 5.1 je prehľad najpoužívanejších existujúcich nástrojov spolu s relevantnými vlastnosťami pre zarovnanie a extrakciu metylácie rastlinných transponovateľných elementov. Zvýraznená časť tabuľky predstavuje kombináciu vlastností, ktoré budú dostupné pri výbere nástroja Bismark a MethylExtract.

	Zarovnanie				Extrakcia				
	metóda	nástroj	využitie PHRED	výstup	vstup	bisulfitové zlyhanie	minimálna hĺbka	sekvenačné chyby	SNV detekcia
Bismark	3L	Bowtie/ Bowtie2	✓	SAM/ BAM	SAM/ BAM	×	×	×	×
BS-Seeker	3L	Bowtie/ Bowtie2	×	SAM/ BAM	SAM/ BAM	✓	×	×	×
MethylCoder	3L/ WC	Bowtie/ GSNAP	✓/ ×	SAM/ *	*	×	×	×	×
BRAT-BW	3L	BRAT	×	SAM/ BAM	SAM/ BAM	×	×	×	×
BSMAP	WC	SOAP	×	SAM/ BAM	SAM	×	✓	×	×
Bisulfighter	WC	LAST	✓	MAF/ SAM	MAF/ SAM	×	×	×	×
Bis-SNP					BAM	×	×	✓	✓
MethylExtract					SAM/ BAM	✓	✓	✓	✓

Tabulka 5.1: Prehľad nástrojov realizujúcich bisulfitové mapovanie a extrakciu metylácie. 3L znamená 3-písmenové mapovanie a WC je skratka pre wild-card. Znak hviezdy reprezentuje neštandardizovaný formát.

5.2 Konštrukcia nástroja na lokalizáciu metylačných miest transpozónov

Na základe definovaných požiadavkov a možnosti využitia určitých existujúcich programov bol vytvorený nástroj, ktorý dokáže lokalizovať metylačné miesta vstupných sekvencií

a extrahovať z nich celkovú úroveň metylácie. Jednotlivé časti a funkcie sú zobrazené na workflow v obrázku 5.2. Počiatočným vstupom sú:

- Súbor obsahujúci všetky referenčné sekvencie transpozónovej rodiny získané prostredníctvom nástroja RepeatExplorer [43] z genómu daného organizmu.
- Kontrolné *ready* osekvenované pomocou NGS metódy.
- Bisulfitom ošetrené osekvenované *ready*.

V prvej časti fázy bisulfitového mapovania prevedie program Bismark referenčné sekvencie a bisulfitové *ready* do 3-písmenovej abecedy. Pre každý typ konverzie vznikne jeden súbor s prevedenými nukleotidmi pôvodných sekvencií a v nasledujúcom kroku 3-písmenového mapovania je konvertovaná verzia bisulfitových *readov* zarovnávaná v Bismarku k obidvom prevedeným verziám referenčných sekvencií. Bismark spojí výsledok tohto mapovania do jedného výstupného súboru. Súčasťou tejto fázy je naviac implementovaný modul, ktorý v 4-písmenovej abecede zarovná kontrolné osekvenované ready k pôvodným referenčným sekvenciám a na základe tohto mapovania prefiltruje výstup Bismarku. Týmto procesom sa získajú korektné zarovnania.

Vo fáze extrakcie úrovni metylácie sú z verifikovaných mapovaní pomocou nástroja MethylExtract získané pozície metylovaných cytozínov referenčných klastrových sekvencií. Implementovaný modul pre spracovanie metylačného pokrytia následne vypočíta podľa navrhnutej metriky celkovú úroveň metylácie v kontextoch CG, CHG a CHH pre daný klaster, čo je jedným z výstupov tohto workflow. Potom vyberie z referenčných sekvencií tie, u ktorých sa našla metylácia a pomocou programu BLAT ich zarovná ku konsenzuálnej sekvencii reprezentujúcej celý daný klaster. Informácie o metylovaných cytozínoch na špecifických pozíciách ďalej využije k tomu, aby získal výstupné metylačné pokrytie celej konsenzuálnej sekvencie pre každý kontext.

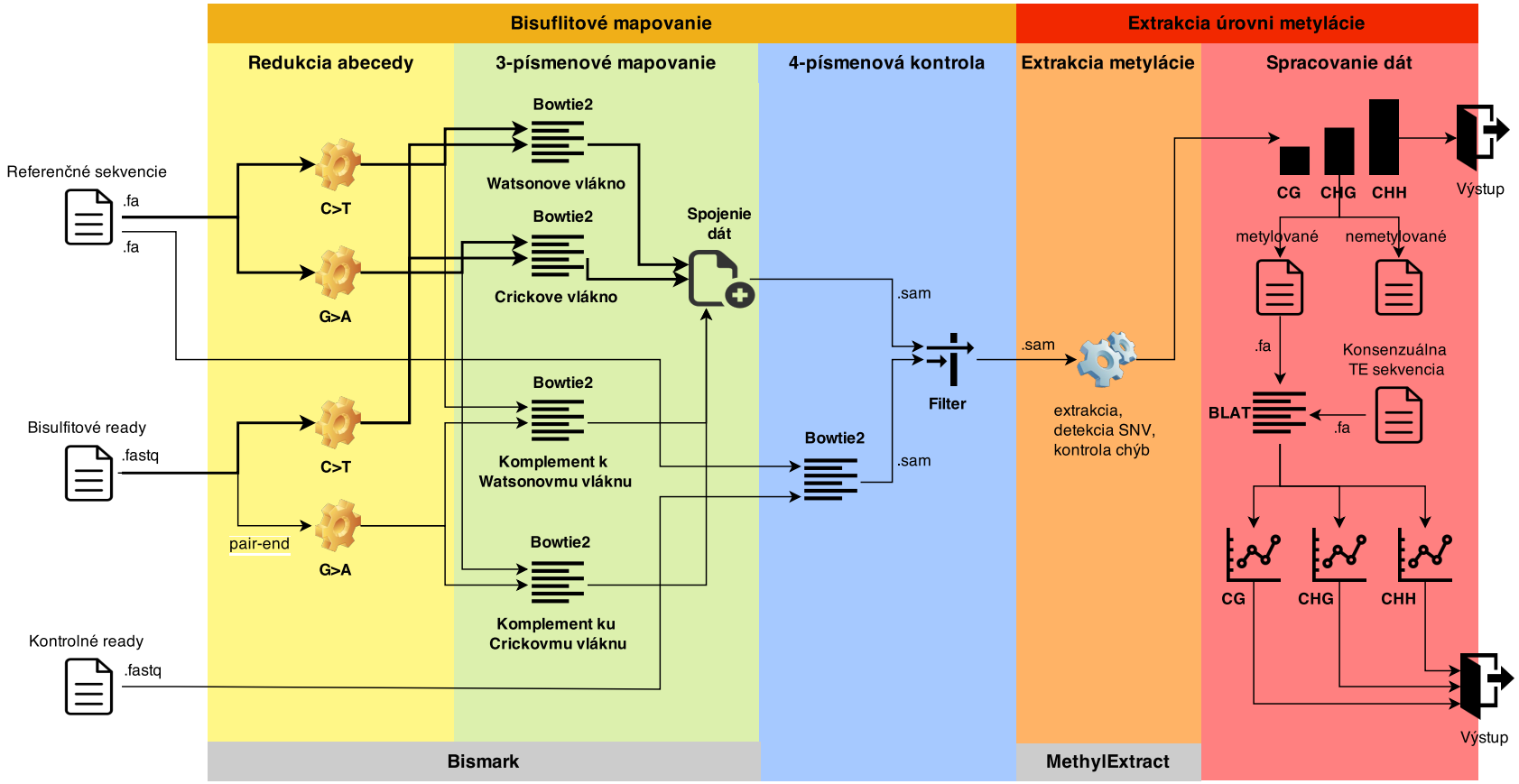
5.3 Mapovanie bisulfitových readov

Správne namapovanie bisulfitových *readov* je základom korektnosti celého procesu, ktorý bude viesť k výsledným metylačným mapám a je jedno z najdôležitejších častí celého procesu tejto analýzy. Nesprávne namapované *ready* by spôsobili v nasledujúcom kroku nesprávne metylačné pokrytie. Naviac, vďaka redukcii abecedy sa výrazne zvýši počet namapovaní *readov* s väčším počtom nesprávnych zhôd. Takúto situáciu je nutné riešiť a výsledok mapovania musia tvoriť predovšetkým *ready* namapované na správne referenčné sekvencie. Rozsiahla kontrola kvality a verifikácia výsledkov tohto procesu je nutnosťou pri spracovávaní bisulfitových dát, a preto je táto fáza rozdelená celkovo do troch častí:

1. Redukcia abecedy vstupných sekvencií a *readov*;
2. Mapovanie *readov* na referenčné sekvencie;
3. Filtrovanie nekorektných mapovaní a verifikácia mapovania.

5.3.1 Redukcia abecedy

Predtým, ako je možné vykonávať samotné mapovanie v 3-písmenovej abecede, osekvenované bisulfitové *ready* a referenčné sekvencie musia byť vhodne prevedené z 4-písmenovej



Obrázok 5.2: Workflow procesu mapovania bisulfitových *readov*, filtrácii výsledkov v 4-písmenovej abecede, extrakcii metylácie a určenia metylačnej úrovni a metylačných máp transpozónových rodín.

abecedy do 3-písmenenovej abecedy. U všetkých referenčných sekvencií sú preto prevedené C na T, čo pokryje všetky možné bisulfitové konverzie spôsobené nemetylovanými C na originálnom vrchnom vlákne a súčasne všetky G na A, čo pokryje možné konverzie na originálnom spodnom vlákne. Tým vzniknú dva rôzne súbory s redukovanou 3-písmenovou abecedou reprezentujúce referenčné sekvencie.

Single-end bisulfitové *ready* majú počiatok buď v pôvodnom vrchnom alebo spodnom vlákne, a teda môžu obsahovať len C>T konverzie. U *pair-end readov* je situácia odlišná a zatiaľ čo jedna časť sa chová ako *single-end ready*, druhá pochádza z reverzného komplementárneho vlákna a môže obsahovať G>A konverzie. Na základe tejto skutočnosti sa *single-end ready* prevedú do C>T konvertovanej verzii a sú mapované voči C>T a G>A referenčným sekvenciám. U *pair-end readov* je nutné C>T skonvertovať jednu (ľavé *ready*) časť a G>A skonvertovať druhú časť (pravé *ready*) a obidva typy mapovať voči C>T a G>A referenčným sekvenciám.

FASTA

Referenčné sekvencie predstavujúce určitý klaster transpozónovej rodiny vstupujú do tejto časti aplikácie vo FASTA formáte ako jeden multi-FASTA súbor obsahujúci všetky sekvencie.

FASTA formát je textový formát slúžiaci pre reprezentáciu buď nukleotidových sekvencií alebo peptidových sekvencií. Sekvencia v tomto formáte začína s jednoriadkovým popisom nasledovaná riadkami so samotnými dátami sekvencie. Popisný riadok je oddelený od sekvenačných dát symbolom „väčší ako“ (>) v prvom stĺpci [6].

Sekvencie musia byť reprezentované v štandardných IUB/IUPAC[4] kódoch s niekoľkými výnimkami:

- malé písmená sú akceptované a sú mapované na veľké písamená
- pomlčka môže reprezentovať medzeru
- akékoľvek číselné znaky v sekvencii by mali byť odstránené alebo nahradené vhodným znakom

Príklad sekvencie vo FASTA formáte vyzerá následovne:

```
>AB000263|acc=AB000263|descr=Homo sapiens mRNA.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

FASTQ

Sekvenované bisulfitové *ready*, ktorých mapovanie voči referenčným sekvenciám sa bude hľadať, vstupujú do tejto časti aplikácie vo formáte FASTQ. FASTQ ukladá sekvencie a PHRED skóre ich jednotlivých bází v jednom súbore. Vďaka uchovanému PHRED skóre sa v neskoršej fáze bude zohľadňovať kvalita *readov*, čo bude relevatné pri určovaní úrovne metylácie.

Každý *read* je tvorený hlavičkou začínajúcou znakom @ s názvom a popisom *readu*. V ďalších riadkoch nasleduje jeho sekvencia vo FASTA formáte a platia pre ňu rovnaké pravidlá. Ak niektorý z nasledujúcich riadkov začína znakom +, udáva to koniec sekvencie a môže byť v ňom zopakovaný riadok s názvom. Za ním nasledujú riadky so zakódovanou kvalitou [16].

Príklad sekvencie vo FASTQ formáte je následovný:

```
@SRR001666.1 071112SLXA-EAS1s7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112SLXA-EAS1s7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

5.3.2 3-písmenové mapovanie

Mapovanie v 3-písmenovej abecede je možné realizovať akýmkoľvek programom, ktorý má na výstupe valídny SAM súbor (viď sekcia 5.3.3). Niektoré z takýchto nástrojov už boli popísané v sekcii 4.2. Kvôli redukovanej abecede sa väčšie množstvo *readov* namapuje na viacero genomických pozícií v porovnaní s klasickým prístupom, kde nie sú prítomné bisulfitové dáta. Naviac je ešte k tomu nutné brať ohľad na pravdivé *read*-C a referenčné-T nezhody, ktoré môžu byť skryté za C>T konverziami a tak prispievať k ďalším chybám. Všetky tieto problémy spojené s bisulfitovou modifikáciou zvyšujú čas behu procesu zarovnávania a kladú väčšie nároky na verifikáciu kvality mapovania.

5.3.3 Bismark

Redukcia abecedy a zarovnanie bisulfitových *readov* k referenčným klastrom reprezentujúcich transpozónové rodiny sú realizované externým programom - Bismarkom. Dôvody tejto voľby boli bližšie popísané v sekcii 5.1.

Predspracovanie

Predspracovanie sekvencií nepatrí implicitne do tohto procesu, ide však o veľmi užitočný krok vzhľadom ku kvalite datasetu, ktorý tvoria bisulfitové *ready*. Kontrolou kvality osekvenovaných *readov* ešte predtým, ako sa spustí zarovnanie je možné získať informácie o možných problémoch vo vstupných dátach, ktoré by sa mohli negatívne prejaviť v samotnom behu ako pri mapovaní, tak aj pri určovaní metylácie. Pre tieto účely je možné napríklad použiť voľne dostupnú aplikáciu FastQC [7].

Medzi kontroly relevantné pre bisulfitové sekvenovanie patrí zloženie báz v sekvenciách, z čoho je možné určiť ako efektívne prebehla bisulfitová konverzia a taktiež aj možnú adepterovú kontamináciu. Úroveň sekvenčných duplikácií umožní odhaliť prípadné zamorenie datasetu PCR duplikátmi. Obsah báz GC tiež môže predstavovať náhľad do kvality experimentu.

Pre odstránenie týchto problémov by mali byť *ready* podrobené adaptérovému a bázovému orezávaniu. Jednoduchým skriptom pre tieto základné kroky je napríklad Trim Galore [8] využívajúci funkcionality nástroja Cutadapt a pri jeho základnom spustení vykoná:

- odstránenie báz s PHRED skóre menším ako 20
- odstránenie akýchkoľvek známok adapterovej Ilumina sekvencie z 3' konca

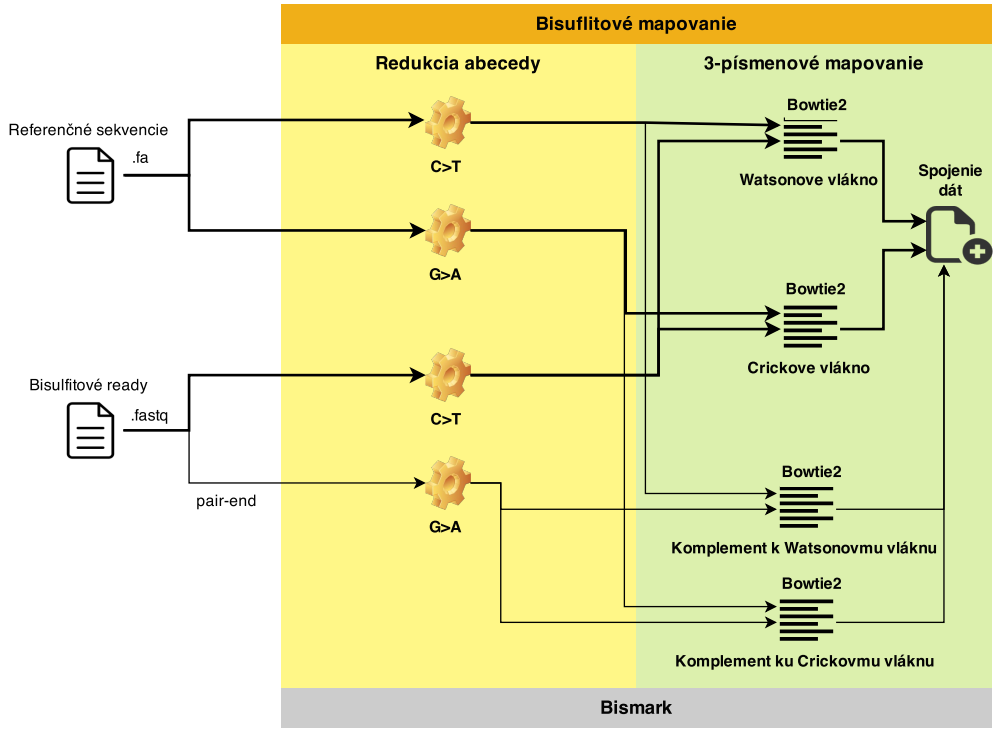
- odstránenie sekvencií kratších ako 20bp

Niektoré z týchto kontrol kvality sú zakomponované v ďalších častiach (napr. kontrola PHRED skóre pri získavaní metylácie), neplatí to ale pre všetky a je vhodnejšie pracovať s validnými dátami hneď od začiatku.

Zarovnanie

Dáta vysokej kvality je potom možné priviesť na vstup nástroja Bismark, ktorý v prvom kroku prevedie referenčné sekvencie na dva rozličné súbory redukovanej abecedy. Prvý z nich obsahuje C>T (Watsonove vlákno) a druhý G>A (Crickove vlákno) konverzie. Z týchto súborov sa následne vytvorí index pre externý zarovnávací program Bowtie. Táto prvá časť sa nazýva príprava genómu.

V druhej časti Bismark najprv konvertuje vstupné bisulfitové *ready* na C>T verziu (ak by sa jednalo o *pair-end ready*, vytvorí naviac aj G>A konvertovaný súbor) a pomocou dvoch paralelne bežiacich inštancií Bowtie [35] algoritmu mapuje C>T konvertované *ready* na C>T referečné sekvencie (odpovedá mapovaniu na vrchnom vlákne) a C>T *ready* na G>A referenčné sekvencie (odpovedá mapovaniu na spodnom vlákne). V prípade *pair-end readov* by sa naviac mapovali G>A *ready* na obidve konvertované verzie referenčných sekvencií a paralelne by bežali štyri inštancie Bowtie. Ukážka tohto procesu je zobrazená na obrázku 5.3.



Obrázok 5.3: Zarovnanie bisulfitových *readov* k referenčným sekvenciám pomocou nástroja Bismark. Hrubo zvýraznené šípky reprezentujú mapovanie *single-end readov*.

Pri zarovnávaní transponovateľných elementov je podstatné brať do úvahy vyšší podiel mutácií (hlavne na cytozínoch) prejavujúcich sa ako SNV než je to u klasických génových

sekvencií. Táto skutočnosť môže robiť mapovací proces náročnejší, pretože pri striktnom výpočte nemusí dôjsť k nájdeniu dostatočného počtu namapovaných sekvencií.

Externý mapovací program Bowtie integrovaný v Bismarku je rýchly a pamäťovo efektívny nástroj pre zarovnávanie *readov* dlhších ako 50bp. Využíva FM-index pre indexovanie genómu a udržiava malú pamäťovú stopu, preto je vhodné ho používať aj na väčšie datasety (akými bisulfitové dáta sú).

Každé jednotlivé zarovnanie má uvedené skóre odvíjajúce sa následovne. Nezhoda v bázy na vysokokvalitnej pozícii obdrží postih -6. Medzera v *reade* dĺžky (počet báz)-2 dostane postih -11 (-5 pre otvorenie medzery, -3 pre prvé predĺženie a -3 pre druhé predĺženie). Toto platí pre *end-to-end* zarovnávací mód a u *readu* dĺžky 50bp, u ktorého sa nájde mapovanie len s jednou nezhodou na vysokokvalitnej pozícii a jednou (počet báz-2) medzerou, bude výsledné skóre $-(6 + 11) = -17$. Najlepšie skóre je teda 0 a je možné ho dosiahnúť ak je medzi *readom* a referenčnou sekvenciou totálna zhoda. Parametrické nastavenie behu programu umožňuje meniť minimálne skóre, ktoré musí dané zarovnanie nadobudnúť, aby bolo akceptované. Minimálne skóre je vyjadrené ako funkcia dĺžky *readu* a nastavuje sa podľa troch parametrov: (a) typ funkcie F; (b) konštanta B a (c) koeficient A. Dostupné typy funkcií sú konštantná (C), lineárna (L), odmocnina (S) a prirodzený logaritmus (G). Parametre sú špecifikované ako F, B, A a pre hodnoty $L, -0.4, -0.6$ bude funkcia definovaná ako:

$$f(x) = -0.4 + -0.6 * x \tag{5.1}$$

kde pre *read* dĺžky 50bp by v tomto prípade bolo minimálne skóre -30.4. Voľnejšie nastavenie minimálneho skóre tak, aby počítalo s niekoľkými nezhodami na danú dĺžku *readu* umožní Bismarku akceptovať aj *ready* s menej kvalitným mapovaním (bežné pri transpozónoch), čo zvýši počet pravdivých pozitívnych nálezov. Na druhú stranu sa zvýši aj počet chybných pozitívnych nálezov, tieto je však možné neskôr redukovať do takej miery, aby spôsobovali čo najmenšiu odchýlku.

Bismark podporuje Bowtie verziu 1 alebo 2. Vyššie popísané nastavenie minimálneho skóre je aplikovateľné iba v Bowtie 2 a naviac má verzia 2 niekoľko ďalších dôležitých výhod oproti 1, ktoré sú pre zarovnanie transpozónov relevantné:

- Pre *ready* dlhšie ako 50bp je Bowtie 2 rýchlejší, citlivejší a využíva menej pamäti
- Bowtie 2 podporuje medzery v zarovnaní
- Nemá žiadne limity na dĺžku *readov* (Bowtie 1 zarovnáva *ready* dĺžky maximálne 1000bp)
- Efektívnejšie *pair-end* mapovanie

Filtrovanie

Bismark pri filtrovaní nájdených výsledkov pracuje s dvoma typmi skóre, ktoré Bowtie uvádza u každého mapovania. Jedno reprezentuje samotné skóre nájdeného zarovnania a druhé je skóre pre najlepšie nájdené zarovnanie, ktoré je iné ako to zaznamenané. Tieto dva údaje sa použijú pre nájdenie zarovnania s najvyšším skóre. Vo všetkých mapovacích inštanciách sa následne hľadá unikátne najlepšie zarovnanie. Ak sa nájde, a teda existuje len jedna najlepšia pozícia, toto zarovnanie sa uloží. Ak sa však nájde viac zhôd s rovnakým (najvyšším) skóre, daná sekvencia nebude vôbec akceptovaná. Odmietnuté mapovania môžu

byť explicitne exportované do samostatného súboru (ako nenamapované alebo viacznačné). Pri voľbe *directional* protokolu sa ešte naviac odstraňujú všetky nálezy na iné ako originálne vrchné alebo spodné vlákno.

SAM

Výsledkom zarovnania sekvencií v 3-písmenovej abecede musí byť súbor vo formáte SAM (*Sequence Alignment/Map*) očakávaný v ďalších fázach tohto workflow. Ide o textový formát oddelený medzerami pozostávajúci z dvoch hlavných častí. Riadky hlavičky začínajúce znakom @ (obsahujúce názvy sekvencií) a jednotlivé mapovania. Každý riadok s mapovaním má 11 povinných polí pre uchovanie dôležitých informácií relevantných vzhľadom k mapovaniu ako je napríklad namapovaná pozícia a premenlivý počet volitelných polí pre aplikačne špecifické informácie. V tomto prípade ide predovšetkým o skóre využívané pri filtrácii.

Každý riadok s mapovaním reprezentuje jedno mapovanie *readu* na konkrétnu referenčnú sekvenciu. Jednotlivé jeho stĺpce predstavujú konkrétne informácie, s pomocou ktorých budú vytvorené metylačné mapy vstupných sekvencií. Informácie obsiahnuté v týchto stĺpcoch sú uvedené v tabuľke 5.2.

Stĺpec	Pole	Typ	Reg. výraz/Rozsah	Popis
1	QNAME	String	$[!-?\texttt{A}-\sim]\{1,255\}$	Názov mapovaného *readu*
2	FLAG	Int	$[0, 2^{16}-1]$	bitový FLAG
3	RNAME	String	$\backslash * \mid [!-()+-<>-\sim][!-\sim]*$	Názov referenčnej sekvencie
4	POS	Int	$[0, 2^{31}-1]$	Najľavejšia mapovacia pozícia
5	MAPQ	Int	$[0, 2^{8}-1]$	Kvalita mapovania
6	CIGAR	String	$\backslash * \mid ([0-9]+[\texttt{MIDNSHPX}=])+$	reťazec CIGAR
7	RNEXT	String	$\backslash * \mid = \mid [!-()+-<>-\sim][!-\sim]*$	Názov ďalšieho *readu*
8	PNEXT	Int	$[0, 2^{31}-1]$	Pozícia ďalšieho *readu*
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	Dĺžka templátu
10	SEQ	String	$*\mid [\texttt{A}-\texttt{Za}-\texttt{z}=.]+$	Namapovaný read
11	QUAL	String	$[!-\sim]+$	ASCII bázovej kvality

Tabulka 5.2: Popis hlavných polí SAM formátu [5]

Detailnejší popis jednotlivých polí je možné nájsť v špecifikácii SAM formátu [5]. Súbor v tomto formáte (a teda aj výstup časti mapovania *readov*) môže vyzerať následovne:

```
@HD VN:1.0 SO:unsorted
@SQ SN:fETM21CH01C4COJ LN:181
@SQ SN:fETM21CH01CYS84 LN:212
@SQ SN:fETM21CH01CTH25 LN:216
read1 16 fETM21CH01C4COJ 213 3 10M * 0 0 CCACACAAAA ffffffffff NM:i:21 MD:Z:14
read2 16 fETM21CH01CYS84 212 4 10M * 0 0 ACTCACTCAA ffffffffff NM:i:14 MD:Z:15
read3  0 fETM21CH01CTH25 211 8 10M * 0 0 AATCCTACCA ffffffffff NM:i:20 MD:Z:16
```

5.3.4 4-písmenová kontrola kvality mapovania

Súbor mapovania bisulfitových *readov* už obsahuje tie najkvalitnejšie unikátne nálezy. Z dôvodu nutnosti vyššej senzitivity pri zarovnávaní na transpozónové referenčné sekvencie a redukcii komplexity vyplývajúcej z bisulfitovej konverzie sa v nich však budú do určitej miery

nachádzať aj nekorektné mapovania, ktoré môžu ovplyvniť výslednú metylačnú úroveň. Pre kontrolu správnosti mapovania bol preto vytvorený samostatný modul realizujúci zarovnanie kontrolných *readov* (viď sekcia 5.1) k referenčným sekvenciám v plnej 4-písmenovej abecede, ktorého cieľom je čo najviac zvýšiť pozitívnu predikčnú hodnotu. Pracuje v nasledujúcich krokoch:

1. Pomocou Bowtie2 [9] programu zarovná kontrolné *ready* k referenčným sekvenciám v 4-písmenovej abecede.
2. Akceptuje len nálezy s parametricky definovanou minimálnou kvalitou mapovania ('minQ').
3. Záznamy rozdelí podľa Watsonovho a Crickovho vlákna.
4. Do výsledného SAM súboru budú zapísané všetky záznamy 3-písmenového mapovania, ktoré sa namapujú na rovnaké pozície daného vlákna referenčných sekvencií ako výsledky 4-písmenového mapovania.

Vďaka tomuto modulu dosiahneme vo výsledku to, že sa výrazne redukuje komplexita zavedená redukciou na 3-písmenovú abecedu a pre extrakciu metylácie ostanú len tie *ready*, ktoré by sa namapovali v 4-písmenovej abecede.

Niektoré nástroje pre spracovanie bisulfitových dát majú v tejto fáze ešte naviac implementovaný modul pre 4-písmenové zarovnanie sekvenčných čítaní, ktoré sa predtým nenamapovali alebo namapovali viacznačne. Bismark poskytuje na výstupe 3-písmenového mapovania súbory s týmito dvoma typmi *readov*, a preto som v tomto mieste naviac implementoval aj túto techniku. Jej vplyv sa však neprezentoval pozitívne a vo výslednom nástroji sa táto implementácia z tohto dôvodu neobjavuje.

5.4 Extrakcia úrovni metylácie

V predchádzajúcej kapitole sme si popísali postup mapovania bisulfitových *readov* na referenčné sekvencie spolu s jeho verifikáciou. Výstupom tejto časti je mapovanie *readov* na špecifické pozície. V tejto kapitole popíšeme, ako využiť získané informácie pri zisťovaní úrovne metylácie.

Cieľom fázy extrakcie metylácií je zobrať do úvahy všetky informácie o mapovaných bázach pre každú jednu genomickú pozíciu a zároveň minimalizovať skreslenie, ktoré do tohto procesu môže zanášať nukleotidový polymorfizmus, sekvenačné chyby alebo bisulfitové zlyhanie. Výstupom bude celková metylačná úroveň v kontexte CG, CHG a CHH a metylačné pokrytie konsenzuálnej sekvencie reprezentujúcej daný klaster.

5.4.1 MethylExtract

Jadrom tohto modulu je nástroj určujúci metylačné mapy pre jednotlivé sekvencie predstavujúce vstupný klaster transpozónovej rodiny. Musí byť dostatočne citlivý s veľkou hodnotou pozitívnej predikcie pri určovaní metylácie, čo úzko súvisí so správnym rozlišovaním zdrojov chýb, ktoré sú pre túto časť procesu typické. V kapitole 4 boli všetky známe chyby popísané a boli v nej predstavené existujúce aplikácie spolu s ich schopnosťou riešenia možných zdrojov odchýlok akými sú, predovšetkým u transpozónov, SNV. Z dvoch v súčasnosti existujúcich programov, o ktorých si je autor vedomý, pre adresovanie SNV bol ako súčasť modulu extrakcie metylácie vybraný MethylExtract.

MethylExtract bol popísaný v sekcii 4.2.2 a následne bol porovnaný s ostatnými existujúcimi nástrojmi. Pri lokalizácii metylačných miest transpozónov je predovšetkým dôležité:

1. Je možné presne lokalizovať metyláciu na špecifických pozíciach referenčných sekvencií z výstupu predchádzajúcej fázy mapovania;
2. SNV, na ktoré sú TE špeciálne bohaté, nepredstavujú významný zdoj chýb pri odhadovaní metylácie cytozínov;
3. Existuje rozlišovanie medzi metyláciou v kontexte CG, CHG a CHH;
4. Relatívne rýchly beh a jednoduché používanie;

MethylExtract spĺňa všetky tieto podmienky, naviac sa jedná o vyladený nástroj vyvíjaný už niekoľko rokov, ktorý autori neustále zlepšujú a momentálne implementuje najviac metód pre kontrolu kvality a redukovanie odchýlky spomedzi všetkých ostatných voľne dostupných a používaných aplikácií.

Získanie metylácie

SAM súbor získaný z mapovania obsahuje informáciu o zarovnaní readov k referenčným sekvenciám od určitej pozície. Pre každú referenčnú sekvenciu, ku ktorej boli *ready* zarovnané, hľadá MethylExtract tie pozície s C, kde nájde minimálny počet ('minDepthMeth') *readov* mapovaných na daný cytozín. Rozlišuje medzi počtom *readov* obsahujúcich metylovaný cytozín na odpovedajúcej pozícii (C/C zhoda, v rovnici reprezentované ako C) a počtom *readov* obsahujúcich thymín na tejto pozícii (C/T nezhoda, v rovnici reprezentované ako T). Na základe týchto dvoch údajov je možné určiť úroveň metylácie na danej pozícii ako:

$$meth\_lvl = 100 * (C/(C+T))(\%) \tag{5.2}$$

SNV

V analýze bisulfitových sekvenčných dát sú SNV najviac zanedbávaným zdrojom chýb. Väčšina nástrojov interpretuje C>T konverziu ako nemetylovaný cytozín aj napriek tomu, že určitý počet sú SNV, a preto by bol tento záver chybný. C/T SNV sa manifestuje na komplementárnom DNA vlákne ako adenín, zatiaľ čo bisulfitová deaminácia neovplyvní guanín na komplementárnom vlákne. Algoritmus na detekciu SNV implementovaný v MethylExtrakte je adaptáciou široko používaného algoritmu *VarScan* [31]. Princíp jeho fungovania je nasledovný:

1. Usporiada a ohodnotí všetky mapovania každého *readu* a zahodí tie záznamy, ktoré majú nízku identitu alebo sa mapujú na viac pozícií v referenčnom genóme.
2. Každé najlepšie unikátne zarovnanie pre *read* je skúmané voči možným sekvenačným zmenám.
3. Varianty detekované vo viacerých *readoch* sú potom kombinované do unikátnych SNP.
4. Pre každú predpovedanú variantu určuje VarScan celkové pokrytie, počet *readov* s výskytom tohto SNP a priemernú kvalitu báz. Výsledné hodnoty sú dané na základe hodnôt prahu pre pokrytie, kvalitu, frekvenciu variant a počet *readov* nutných pre označenie bázy ako varianty.

Hlavný rozdiel oproti DNA bez bisulfitovej aplikácie je v redukovanom množstve sekvenčných informácií, ktoré môžu byť použité pre detekciu SNV. Nemetylované cytozíny sú prevedené na thymín, a z tohto dôvodu nemôžu byť na C pozíciách použité nukleotidy, ktoré mohli vzniknúť bisulfitovou konverziou, pre detekciu variácií. Algoritmus pracuje následovne:

1. Odfiltruje pozície, ktoré sú pokryté menej *readmi* ako je minimálna hĺbka *readov* ('minDepthSNV'). Predvolená hodnota je 1, a tým skúma všetky pozície pokryté aspoň jedným *readom*.
2. Vypočíta nukleotidové frekvencie vrátane všetkých báz nad minimálnym prahom PHRED skóre ('minQ').
3. Zahodí všetky nukleotidy s frekvenciami pod definovaný prah ('varFraction').
4. Vypočíta *p-value* pre variantné pozície (viac ako dva nukleotidy nad 'varFraction') podľa Fisherovho exaktného testu.
5. Iba pozície s *p-value* pod určitým prahom sú považované za SNV ('maxPval').
6. Dva nukleotidy s najvyššími frekvenciami sú určené ako údajný genotyp vzorky na tejto pozícii.

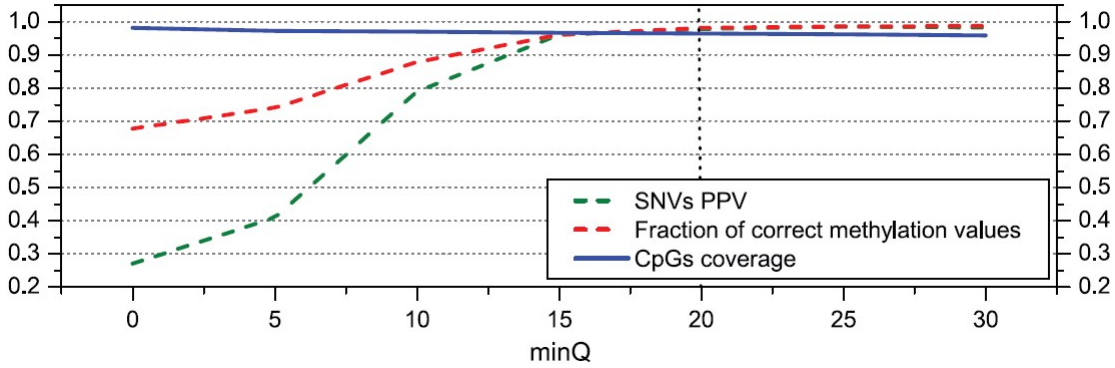
Detekované SNV sú taktiež výstupom tohto programu a môžu slúžiť napríklad pre verifikáciu korektnosti detekcie nukleotidových variant.

Kontrola chýb

MethylExtract implementuje chybové kontroly pre všetky najväčšie známe zdroje chýb, a tým poskytuje možnosť veľmi presného odhadu metylačnej úrovne.

Bisulfitové zlyhanie Pravdepodobnosť chyby pri bisulfitovej konverzii u nemetylovaných cytozínov je pri moderných protokoloch zvyčajne menej ako 1%. Aj napriek tomu však môžu byť niektoré pozície nesprávne profilované. S predvolenými nastaveniami MethylExtract eliminuje všetky *ready* s aspoň 90% (pravdepodobne) nekonvertovanými cytozínmi v inom ako CpG kontexte. Táto práca je však zameraná na analýzu rastlinnej metylácie, kde sa vyskytuje metylácia vo veľkej miere aj v kontextoch CHG a CHH, a preto je nutné túto kontrolu preskočiť, inak by mohla spôsobiť nežiadané odchýlky. Z dôvodu relatívne veľmi nízkej pravdepodobnosti výskytu tejto chyby by jej zanedbanie nemalo viesť k významným nepresnostiam oproti situácii, kedy by táto kontrola ostala zapnutá [13]. Jej vplyv je najlepšie redukovať ešte v predspracovaní vstupných *readov* predtým, ako celý proces získavania metylačnej úrovne začne.

Sekvenačné chyby Vplyv jednotlivých báz na odvodenie metylačných stavov môže byť kontrolovaný prostredníctvom PHRED skóre. Pri jeho nastavení na PHRED ≥ 20, a teda akceptovaním báz s pravdepodobnosťou < 0.01 nesprávneho určenia, prispenie tejto chyby k celkovej chybe bude menej ako 1%. Hodnota 20 pre túto kontrolu je predvoleným nastavením MethylExtractu u odpovedajúceho parametru ('minQ') a je zvolená na základe testov vykonaných autormi tohto nástroja (obr. 5.4) [13]. Aby prebehla kontrola sekvenačných chýb v poriadku, je nutné správne nastavenie kódovania FASTQ súboru ('qscore').



Obrázok 5.4: Graf znázorňuje pozitívnu predikčnú hodnotu (PPV) pre SNV a časť korektne profilovanej metylácie v CpG kontexte ako funkciu minimálnej kvality bázy ('minQ'). Y os reprezentuje SNV PPV, časť korektných metylačných hodnôt a pokrytie CpG [13].

Výsledné metylačné pokrytie sekvencií

Výsledkom extracie metylácie z bisulfitových sekvencií je pokrytie referenčných sekvencií na špecifických pozíciách v danom kontexte. Toto pokrytie je dostupné ako pre Watsonove (vrchné) vlákno, tak aj pre Crickove (spodné) vlákno a je uložené v súbore {kontext}.output. Nižšie je ukážka tohto súboru a vysvetlenie významu jednotlivých polí.

CHROM názov chromozómu (alebo danej sekvencie)

POS 5' pozícia v sekvenčnom kontexte Watsonovho vlákna

CONTEXT sekvenčný kontext s SNV anotovanými v IUPAC kóde

W METH počet metyl-cytozínov (referované k Watsonovmu vláknu)

W COVERAGE počet *readov* pokrývajúcich cytozín v danom kontexte (referované k Watsonovmu vláknu)

W QUAL priemerné PHRED skóre *readov* pokrývajúcich cytozín (referované k Watsonovmu vláknu)

C METH počet metyl-cytozínov (referované k Watsonovmu vláknu)

C COVERAGE počet *readov* pokrývajúcich guanín v danom kontexte (referované k Watsonovmu vláknu)

C QUAL priemerné PHRED skóre *readov* pokrývajúcich cytozín (referované k Watsonovmu vláknu)

CHROM	POS	CONTEXT	W METH	W COVERAGE	W QUAL	C METH	C COVERAGE	C QUAL
fETM21CH01C4COJ	12	CG	4	4	69	.	.	.
fETM21CH01C4COJ	15	CG	5	5	69	.	.	.
fETM21CH01C4COJ	18	CG	0	6	69	5	5	69
fETM21CH01CTH25	156	CG	5	5	69	9	9	69
fETM21CH01CTH25	201	CG	.	.	.	5	8	69
fETM21CH01CTH25	204	CG	.	.	.	2	4	69

```
fETM21CH01AFBYS     26      CG      3       3       69      .       .       .
fETM21CH01AFBYS     31      CG      3       9       69      .       .       .
fETM21CH01AFBYS     53      CG      3       3       69      4       4       69
fETM21CH01CYS84     160     CG      11      11      69      7       11      69
fETM21CH01CYS84     163     CG      11      18      69      11      11      69
fETM21CH01CYS84     187     CG      6       6       69      5       5       69
fETM21CH01CYS84     192     CG      4       4       69      5       5       69
```

Vyššie uvedené výsledky je možné interpretovať následovne (príklad pre tretí riadok). U sekvencii s názvom fETM21CH01C4COJ bolo nájdených na pozícii 18 (prvá báza je pozícia 1) v kontexte CG šesť *readov* pokrývajúcich cytozín, ale ani jeden z nich nebol metylovaný (C/T nezhoda) a päť *readov* pokrývajúcich guanín na pozícii 19 Watsonovho vlákna, čo odpovedá cytozínu na Crickovom vlákne, z ktorých bolo všetkých päť metylovaných (C/C zhoda).

5.4.2 Spracovanie metylačných údajov

Metylačné pokrytie transpozónových sekvencií rozličných kontextov je v časti spracovania dát využité k určeniu celkovej metylácie danej transpozónovej rodiny a k vytvoreniu metylačnej mapy pozdĺž báz celej konsenzuálnej sekvencie reprezentujúcej túto rodinu.

Metylácia transpozónovej rodiny

Každá sekvencia s nájdenými metylovanými pozíciami reprezentuje časť klastru, ktorá bude určitým spôsobom vplývať na jeho výslednú metylačnú úroveň. Niektoré sekvencie sú však pokryté väčším počtom *readov* ako iné, a teda ich metylačná úroveň ma vyššiu váhu pri určovaní výslednej hodnoty. Pre daný klaster a jeden z kontextov sa jeho úroveň vypočíta podľa nasledujúcej metriky:

$$Meth_{avg} = \sum_{seq \in Cl_m} meth\_lvl_{seq} * \left(\frac{cov_{seq}}{cov_{total}} \right) \tag{5.3}$$

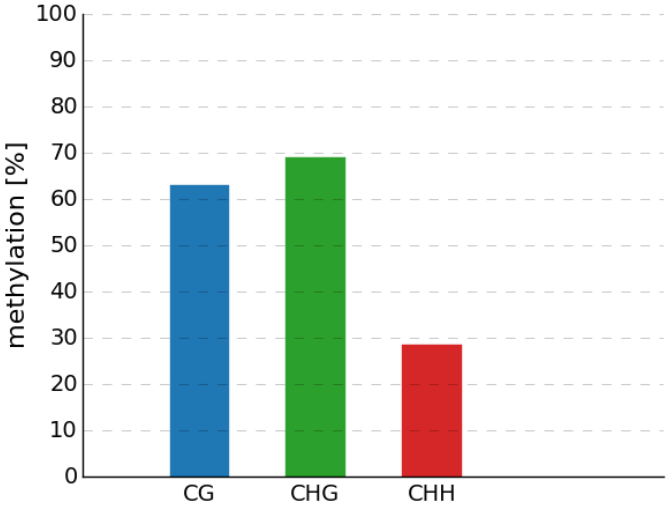
kde Cl_m je množina metylovaných sekvencií daného klastru, $meth\_lvl_{seq}$ je metylačná úroveň celej sekvencie získaná pomocou 5.2, cov_{seq} je pokrytie sekvencie namapovanými *readmi* a cov_{total} je počet namapovaných *readov* na všetky sekvencie.

Týmto výpočtom sa získa metylačná úroveň klastru z jedného behu. Pri reálnom používaní aplikácie je však dôležité mať štatistiku z viacerých experimentov (napr. tri sekvenácie DNA) a výsledok získať na základe výstupov každého jedného z nich. Môžeme predpokladať, že odlišné datasety budú produkovať odlišné výsledky mapovania a pokrytie referenčných sekvencií, a preto je nutné aj v tomto prípade brať do úvahy relevantnosť vypočítanej úrovne daného klastru oproti ostatným výpočtom. Tá sa bude takisto ako v predchádzajúcom prípade odvíjať od pokrytia metylovaných pozícií klastru a celkového pokrytia všetkých klastrov:

$$Meth_{Cl\_avg} = \sum_{Cl \in Res} meth\_lvl_{Cl} * \left(\frac{cov_{Cl}}{cov_{total}} \right) \tag{5.4}$$

kde Res je množina záznamov pre každý realizovaný experiment, $meth\_lvl_{Cl}$ je metylačná úroveň celého klastru získaná pomocou 5.3, cov_{Cl} je pokrytie klastru namapovanými *readmi* a cov_{total} je počet namapovaných *readov* všetkych experimenetov daného klastru.

Tento výpočet realizuje špeciálna utilita, ktorá nie je súčasťou workflow, ale slúži pre dodatočnú analýzu. Jej výstupom je stĺpcový graf s metylačnými úrovňami v jednotlivých kontextoch (obr. 5.5).



Obrázok 5.5: Stĺpcový graf metylačnej úrovne klastru v kontextoch CG, CHG a CHH získaný z viacerých experimentov.

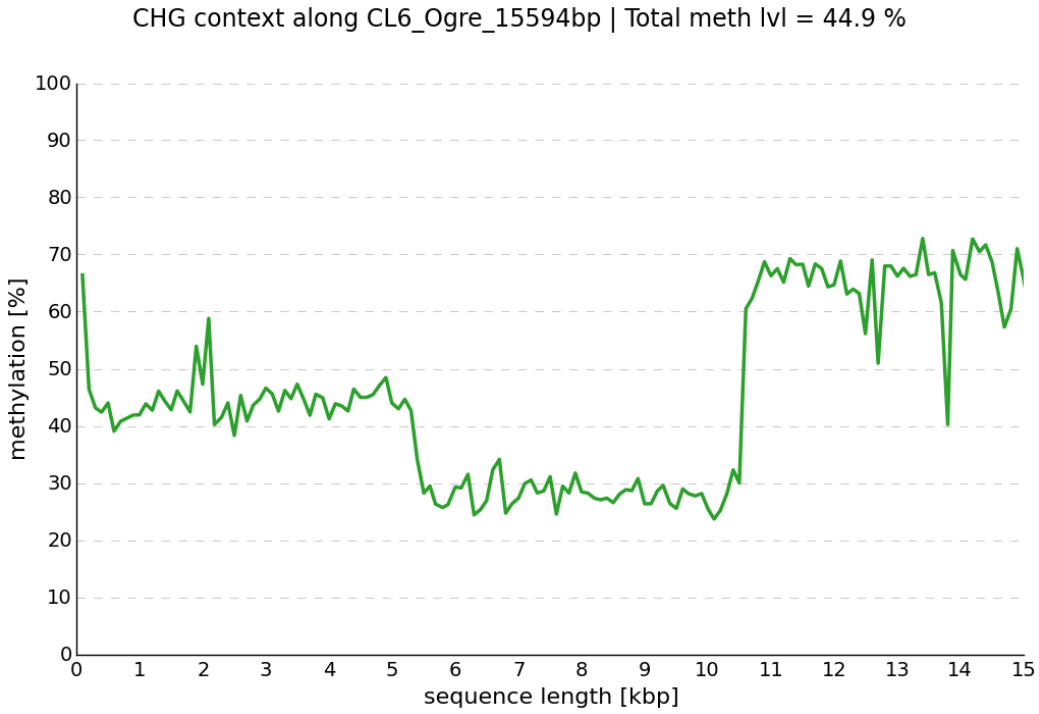
Metylačné mapy

Klastre vytvorené pomocou RepeatExploreru obsahujú okrem sekvencií, ktoré ich tvoria aj jednu konsenzuálnu sekvenciu vytvorenú zo všetkých týchto sekvencií. Vďaka tomu, že dokážeme získať metylačné pokrytie každej jednotlivej sekvencie klastru, u ktorej sa našla metylácia, môžeme taktiež určiť metylačnú mapu celej konsenzuálnej sekvencie. Táto funkcionalita je implementovaná v module pre spracovanie metylačných dát a jej výpočet je realizovaný následovne:

1. Rozdelí referenčné sekvencie na sekvencie s lokalizovanými metyl-cytozínmi a bez metylovaných cytozínov.
2. Metylované sekvencie vo FASTA formáte zarovná voči konsenzuálnej sekvencii takisto vo FASTA formáte prostredníctvom nástroja BLAT.
3. Akceptované sú všetky záznamy s aspoň 80% dĺžkou mapovania a 80% identitou.
4. Pre každé akceptované zarovnanie (sekvenciu) sa v kontextoch CG, CHG a CHH prechádzajú všetky metylované pozície tejto sekvencie.
5. Na pozíciu (začiatok mapovania)+(metylovaná pozícia) konsenzuálnej sekvencie sa uloží alebo inkrementuje počet metylovaných cytozínov a počet pokrytí z metylovanej sekvencie.

Hodnoty metylovaných pozícií sa počítajú v predvolených nastaveniach pre Watsonove a Crickove vlákno dohromady, čo umožní presnejšie výsledky vďaka dostupnosti väčšieho množstva dát. Parametricky je však možné toto nastavenie zmeniť na generovanie dát aj pre jednotlivé konkrétne vlákna.

Príkladom tohto procesu by mohlo byť napríklad mapovanie vyššie popísanej sekvencie fETM21CH01C4COJ na pozíciu 100 konsenzuálnej sekvencie. fETM21CH01C4COJ obsahuje metylované cytozíny na pozíciách 12, 15 a 18, a teda v dátovej štruktúre reprezentujúcej pokrytie konsenzuálnej sekvencie sa zvýší počet metylovaných cytozínov o 4 a pokrytie o 4 na pozícii 112, na pozícii 115 počet metylovaných cytozínov o 5 a pokrytie o 5 a na pozícii 118 počet metylovaných cytozínov o 5 a počet pokrytí o 11. Podobným spôsobom sa naplní dátová štruktúra z hodnôt pre všetky metylované pozície. Ako výsledok tohto procesu sú súbory s počtom metylovaných C a počtom pokrytí na danej pozícii a graf metylačnej mapy pre každý kontext alebo pre všetky kontexty dohromady. Hodnoty grafu sa počítajú na základe parametru 'rate', ktorý udáva z koľkých báz konsenzuálnej sekvencie sa má počítať výsledná hodnota v grafe. Príklad takéhoto výstupu pre CHG kontext je na obrázku 5.6.



Obrázok 5.6: Metylačná mapa konsenzuálnej sekvencie klastru CL6 pre cytozínový kontext CHG a nastavenie parametru 'rate' na hodnotu 100.

Kapitola 6

Výsledky

V tejto kapitole sú popísané výsledky získané vytvoreným nástrojom na lokalizáciu metylačných miest transpozónov. Experimentálne pokusy sú pre dosiahnutie presných výsledkov realizované so simulovanými sekvenčnými knižnicami. Vďaka tomuto vieme presne určiť pravé genomické pozície, odkiaľ *ready* pochádzajú a simulovanú metyláciu spolu s nukleotidovými mutáciami na konkrétnych pozíciách. Pri týchto experimentoch bola predovšetkým skúmaná citlivosť a presnosť vytvorenej metódy. Okrem simulačných experimentov sú naviac realizované aj pokusy na reálnych, voľne dostupných dátach dvoch rastlinných organizmov.

6.1 Metacentrum

Proces mapovania bisulfitových sekvenčných čítaní je náročný na výpočet hlavne kvôli robustnosti bisulfitových dát a nutnosti niekoľkonásobného zarovnania. Vďaka možnosti paralelizácie ako vo fáze zarovnania, tak aj pri extrakcii metylácie sa tento proces výrazne urýchli, ale pri tom spotrebuje veľké množstvo výpočetných prostriedkov. Z týchto dôvodov bola pre experimentovanie využitá infraštruktúra virtuálnej organizácie MetaCentra [1].

MetaCentrum je systém umožňujúci rezerváciu výpočetných zdrojov pre vykonávanie náročných výpočtov. Jednotlivé časti aplikácie využívajú len moduly voľne dostupné v MetaCentre, a preto je ich spúšťanie bezkonfliktné. Pred samotným spustením je nutné najprv nahrať všetky časti aplikácie spolu so sekvenčnými databázami na úložisko rezervovaných počítačov, pridať požadované moduly a následne spustiť jednotlivé skripty s definovanými atribútmi. Pre spustenie výpočtu v prostredí MetaCentra je dostupný samostatný skript, ktorý realizuje celý tento proces. Nastavenie požadovaných výpočetných zdrojov (počet jadier CPU, veľkosť RAM a HDD) by však mali byť volené vždy na základe veľkosti knižníc sekvenčných čítaní a veľkosti klastru transpozónovej rodiny.

6.2 Simulácie

Experimenty s vytvoreným nástrojom boli realizované s využitím generovaných dát, u ktorých je možné presne nastaviť a kontrolovať jednotlivé relevantné faktory. Ich cieľom je zistiť, ako presne odpovedajú namerané údaje referenčným v každej fáze tohto workflow, a aké faktory tieto výsledky výrazne ovplyvňujú.

[1] http://metavo.metacentrum.cz/cs

6.2.1 Simulačný nástroj BSSim

Dáta pre simulačné experimenty boli generované prostredníctvom programu BSSim [56]. Ide o nástroj implementovaný v jazyku Python schopný simulovať bisulfitovú konverziu spolu s nastavením metylačnej úrovne cytozínov v kontextoch CG, CHG a CHH, takže je vhodný aj pre rastlinné genómy. Dokáže taktiež simulovať genetické variácie (SNV) a sekvenačné chyby. Výstupy môžu byť generované pre *directional* aj *non-directional* protokol, *single* aj *pair-end ready* a rôzne dĺžky *readov*. Súčasťou jeho výstupu je súbor s referenčným rozložením metylovaných cytozínov pre jednotlivé sekvencie, ktorý bude slúžiť pre overenie správnosti určovania metylovaných cytozínov.

Pre účely tejto práce bol BSSim modifikovaný takým spôsobom, že spolu s bisulfitovými *readmi* generuje naviac kontrolné *ready*, čo sú rovnaké *ready* ako bisulfitové (spoločné rozloženie SNV aj sekvenačných chýb), ale bez bisulfitovej konverzie.

Bisulfitové *ready* sú vytvárané z referenčných sekvencií tak, že najprv dôjde k pridaniu nukleotidových variácií podľa definovanej pravdepodobnosti alebo podľa danej šablóny. U šablónového nastavenia nukleotidov, ktoré sa využilo pri generovaní dát, bola objavená chyba v kóde simulátoru BSSim, a preto som celú túto časť kódu prerobil. U vzniknutej sekvencii sú jendotlivé bázy modifikované podľa princípov bisulfitovej konverzie s ohľadom na definovanú úroveň metylácie vo všetkých možných kontextoch. Tento krok je vynechaný pri súčasnom vytváraní kontrolných *readov*. Z takto modifikovaných sekvencií sú náhodne vybrané fragmenty určitej dĺžky, z ktorých sú generované krátke *ready* danej dĺžky. Sekvenačná chyba može byť následne zavedená na jednotlivé bázy. Z tohto procesu vznikne FASTQ súbor s bisulfitovými readmi, FASTQ súbor s kontrolnými *readmi*, súbor obsahujúci referenčnú metyláciu cytozínov a súbory obsahujúce referenčné zarovnanie pre obidve DNA vlákna. Postup vytvárania BS-Seq *readov* je zobrazený na diagrame v obrázku 6.1.

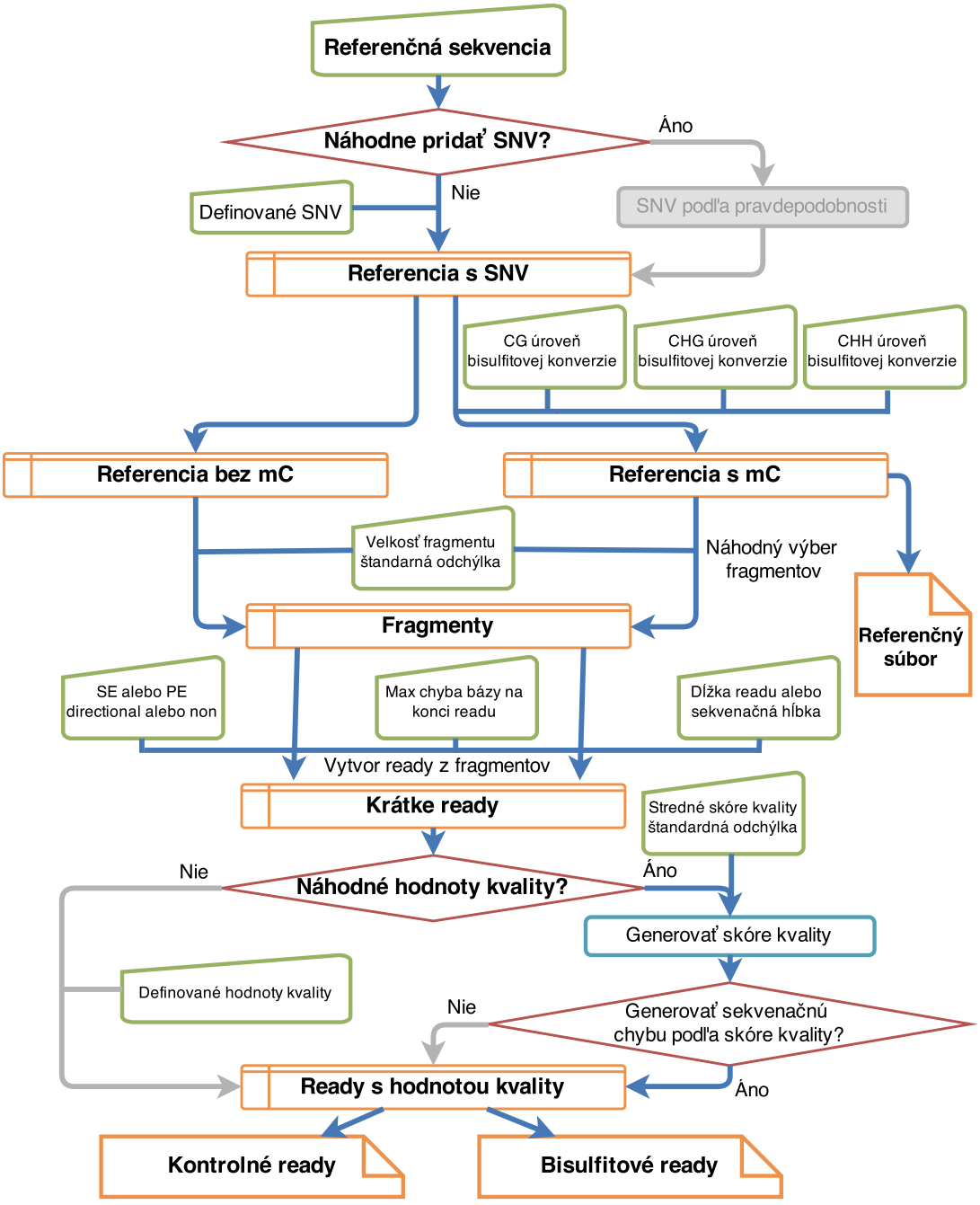
6.2.2 Simulované dáta

Za referenčné sekvencie, ktoré slúžili ako predloha pre vytvorenie generovaných sekvenčných čítaní bol pre všetky experimenty zvolený klaster transpozónovej rodiny organizmu *Silene latifolia* získaný z RepeatExploreru. Keďže sa jedná o zhluk transpozónov, jednotlivé sekvencie sú repetitívne. Parametre týchto simulovaných dát boli zvolené tak, aby čo najviac odpovedali reálnej situácii a pokiaľ nie je uvedené inak, nastavenie ich hodnôt odpovedá nasledujúcemu popisu.

Ako sekvenačná platforma bola zvolená Solexa s dĺžkou *readov* 90bp. Sekvenčnú hĺbku (pokrytie klastru) som zvolil 30X pre väčšie pokrytie transpozónových sekvencií, aby nedochádzalo k nepresnostiam z dôvodu malej veľkosti bisulfitovej knižnice. Všetky čítania sú *single-end* pochádzajúce z *directional* protokolu. Simulovaná úroveň metylácie odpovedá hodnotám bežne sa vyskytujúcim v rastlinných genómoch. V kontexte CG je metylovaných 81% cytozínov, CHG 63% a CHH 40%. Štandardná odchýlka metylovaných cytozínov bola ponechaná na predvolených nastaveniach nástroja BSSim.

Pre zanesenie SNV do generovaných sekvencií som vytvoril samostatný skript, ktorý vytvára súbor s presne definovanými variantami na konkrétnych pozíciach tak, že ich frekvencia je vyššia na nukleotide C (C>T konverzia na Watsonovom a G>A na Crickovom vlákne) oproti ostatným nukleotidom a celkový podiel SNV sa pohybuje v rozmedzí 8-12%. Úroveň sekvenačnej chyby bola ponechaná na predvolených nastaveniach programu.

Vo výsledku tak vznikli repetitívne dáta s vysokým podielom SNV hlavne na cytozínových bázach, odlišnou metyláciou pre každý cytozínový kontext, veľkým pokrytím a sekvenačnou chybou odpovedajúcou danej metóde.



Obrázok 6.1: Diagram generovania BS-Seq *readov* upraveným simulátorom BSSim. Modrou farbou je zvýraznená cesta, ktorou sú sekvenčné čítania vytvárané pre naše experimenty [56].

6.2.3 Experimentálna metóda

Mapovanie sekvenčných čítaní

Pre analýzu výsledkov fázy mapovania bol vytvorený špeciálny skript, ktorý prechádza jeden SAM súbor obsahujúci referenčné zarovnanie simulovaných *readov* a jeden s nájdeným zarovnaním, k vyhodnoteniu mapovacích štatistík.

Vo vykonaných experimentoch je ohodnocovaná citlivosť (angl. *recall*) a pozitívna predikčná hodnota (angl. *precision*). Ako citlivosť je definovaná časť sekvenčných čítaní, u ktorej sa našla simulovaná mapovacia pozícia. Pozitívna predikčná hodnota je definovaná ako časť čítaní mapovaná korektne zo všetkých mapovaných čítaní.

Táto metóda dokáže ohodnotiť iba namapovanú pozíciu v referenčných sekvenciách, ale nevyhodnocuje zarovnanie na jednotlivých bázach. Z tohto dôvodu bola vytvorená ešte ďalšia metóda pre analýzu presnosti na jednotlivých pozíciách.

Extrakcia metylácie

Z výsledkov extrakcie je možné získať až dvojitú analytickú hodnotu. Môžu slúžiť pre porovnávanie výstupu fázy extrakcie metylácie, a taktiež, keďže tento výstup je závislý na stĺpcovom zarovnaní, môžu určovať pozičnú presnosť mapovania.

BSSim poskytuje referenčný výstup s metylovanými konkrétnymi pozíciami v daných kontextoch, a s pomocou vytvoreného skriptu na základe porovnania tohto súboru s výsledkami extrakcie metylácie môže dôjsť k niektorej zo sledovaných situácií:

Chybne negatívna (FN) Simulovaná metylácia nebola extrahovaná

Chybne pozitívna (FP) Extrahovaná metylácia nebola simulovaná

Pravdivo negatívna (TN) Metylácia nebola simulovaná a ani extrahovaná

Pravdivo pozitívna (TP) Metylácia bola simulovaná a aj extrahovaná

Počty jednotlivých výskytov daných prípadov budú priamo dostupné a naviac sa znovu využijú pre výpočet citlivosti, pozitívnej predikčnej hodnoty a úrovne chyby. Citlivosť je definovaná ako:

$$recall = \frac{TP}{TP + FN} \tag{6.1}$$

kde každá hodnota reprezentuje počet výskytov danej kategórie. Je nutné podotknúť, že táto hodnota sa bude odvíjať od citlivosti zarovnania, pretože ak sa nenamapuje určitá pozícia, nemôže v nej byť ani lokalizovaná metylácia. Jej hodnota môže teda indikovať samotné stĺpcové zarovnanie a z pohľadu extrakcii metylácie je ju nutné brať vždy v kontexte mapovania. Pozitívna predikčná hodnota je definovaná ako:

$$precision = \frac{TP}{TP + FP} \tag{6.2}$$

Úroveň chyby, alebo tiež úroveň chybného nálezu je definovaná ako:

$$error\_rate = \frac{FP}{TP + FP} \tag{6.3}$$

K vyhodnocovaniu týchto štatistík je ešte pridané porovnanie celkovej úrovni metylácie pre daný klaster so simulovanými hodnotami.

Keďže SNV tvoria významný zdroj chýb pri transpozónoch, ich detekcia je taktiež súčasťou týchto experimentov a jej vyhodnotenie je realizované samostatným skriptom. Podobne ako v prípade metylácie, aj tu môže dochádzať k niekoľkým scenárom, ale naviac je možné rozlišovať korektné určenie SNV od nesprávneho. Detekcia SNV teda môže byť:

Chybne negatívna (FN) Simulované SNV nebolo detekované

Chybne pozitívna (FP) Detekované SNV nebolo simulované

Správna detekcia (RC) Simulované SNV bolo detekované správne

Nesprávna detekcia (WC) Simulované SNV bolo detekované, ale nesprávne

Výpočet citlivosti a pozitívnej predikčnej hodnoty je daný vzťahmi:

$$recall = \frac{RC}{RC + WC + FN} \tag{6.4}$$

$$precision = \frac{RC}{RC + WC + FP} \tag{6.5}$$

6.2.4 Nastavenie parametrov

Nastavenie hodnôt parametrov pre realizované experimentálne metódy sa riadi nasledujúcim popisom, ak nie je špecifikované inak. Zarovnanie nástrojom Bismark je vykonané s predvoleným nastavením parametrov a zmenená je len jeho hodnota minimálneho skóre pre akceptovanie záznamu (viď ďalej).

4-písmenové filtrovanie využívajúce Bowtie2 má nastavené rovnaké minimálne skóre ako v predchádzajúcom prípade a minimálna kvalita mapovania záznamu, aby bol relevantný pre filtrovanie, je 0.

Pri exktrakcii metylácie je vypnutá bisulfitová kontrola a ostatné parametre majú predvolenú hodnotu, a teda minimálny počet *readov* pokrývajúcich cytozín aby bola zaznamenaná metylačná pozícia je 3, minimálna hodnota PHRED skóre je 20 a minimálna hĺbka pre kontrolu SNV má hodnotu 1. Pri generovaní grafu metylačnej mapy konsenzuálnej transpozónovej sekvencie sa budú jednotlivé hodnoty počítať zo 100 báz a vo výstupe bude desať hodnôt na jednu kilobázu sekvencie.

6.2.5 Experimentálne výsledky

Pre otestovanie funkcionality vzniknutého nástroja som realizoval niekoľko experimentov popisujúcich vplyv jednotlivých častí aplikácie a ich modifikácie na výsledné hodnoty metylačnej úrovne.

Nastavenie skóre zarovnania

Kvalitné mapovanie je základom korektnosti celého procesu lokalizácie metylácie a pri repetitívnych sekvenciách nie je zatiaľ možné dosiahnuť tak kvalitných výsledkov ako u iných. Kvôli vyššiemu pomeru mutácií (SNV) bude pri mapovaní dochádzať k častejším nezhodám alebo medzerám, a preto je nastavenie optimálneho minimálneho skóre zarovnania pre

akceptovanie mapovania, odvíjajúce sa práve od týchto dvoch hodnôt, dôležitým krokom v tomto procese.

Bismark má ako predvolené nastavenie *score_min* $L, 0, -0.2$ a to pre dĺžku generovaných sekvenčných čítaní 90bp znamená, že mapovanie musí mať skóre minimálne -18. Znižovaním tretieho parametru tejto funkcie sa bude znižovať aj hodnota prahu, a tým sa povolí väčší počet nezhôd a medzier vo výslednom zarovnaní. Cieľom tohto experimentu bolo nájsť také nastavenie minimálneho skóre, ktoré umožní najcitlivejšie a najpresnejšie výsledky.

V tabuľke 6.1 je zobrazaná citlivosť a pozitívna predikčná hodnota (PPV) pre rôzne nastavenie funkcie minimálneho skóre v 3-písmenovej abecede.

	$L, 0, -0.2$ (predvolené)	$L, 0, -0.5$	$L, 0, -0.7$	$L, 0, -0.9$	$L, 0, -1.5$
citlivosť	25.41%	44.78%	45.76%	45.84%	45.87%
PPV	93.80%	90.69%	89.28%	88.52%	87.87%

Tabulka 6.1: Citlivosť a pozitívna predikčná hodnota (PPV) pre rôzne nastavenie minimálneho skóre pre akceptovanie zarovnania.

Je zreteľné, že predvolené nastavenie na takto upravené dáta nestačia a znižovaním minimálneho skóre sa citlivosť zvyšuje. Výrazné zvýšenie je však len po skóre dané funkciou $L, 0, -0.5$, respektíve $L, 0, -0.7$. S každým zvýšením citlivosti na druhú stranu klesá PPV a od skóre nižšie ako s $L, 0, -0.7$ už dochádza viac k znižovaniu PPV ako ku zvyšovaniu citlivosti. Z tohto vyplýva, že je výhodné znížiť minimálne skóre, ale len do určitej miery, ktorá v tomto prípade predstavovala skóre -63 dané predpisom $L, 0, -0.7$ a toto nastavenie je použité v nasledujúcich experimentoch. Znižovanie minimálneho skóre má ešte naviac nevýhodu v predĺžení doby výpočtu.

Vplyv filtrovacieho modulu

Pre získanie dojmu o dôležitosti dodatočného filtrovania zarovnania s pomocou kontrolných sekvencií som porovnával výsledky získané z 3-písmenového mapovania (3L) s konečnými výstupnými údajmi s aplikovanou 4-písmenovou (4L) filtráciou. K tomu bol ešte naviac na tieto rozličné výsledky zavolaný modul pre extrakciu metylácie a testoval som, aký vplyv má filtrovanie na výslednú metyláciu a detekciu SNV. Filtrovanie bolo spúšťané v základnom nastavení s minimálnou kvalitou mapovania 0 (všetky nájdené zarovnania sa uvažovali pre filtráciu) a v druhom prípade s minimálnou kvalitou 1.

V tabuľke 6.2 sú prezentované výsledky citlivosti a PPV pre fázu mapovania a pre extrakciu metylácie a detekciu SNV.

Z výsledkov je vidieť, že 4-písmenová kontrola vedie k výrazne vyššej pozitívnej predikčnej hodnote pre všetky tri sledované prípady. Zatiaľ čo výsledok je presnejší, došlo súčasne aj k významnému zníženiu citlivosti. Z tohto vyplýva, že aj keď sa podstatne znížil počet chybne pozitívnych nálezov, pravdivo pozitívne nálezy taktiež klesli.

Veľký rozdiel je medzi filtráciou s minimálnou kvalitou 0 a 1, a to vo všetkých pozorovaných prípadoch. Len zavedenie 4-písmenového zarovnania ako kontroly má výrazný vplyv na spresnenie výsledkov, ktorých komplexita bola výrazne redukovaná konverziou do 3-písmenovej abecedy. Zvyšovaním minimálnej kvality mapovania nedochádzalo k podstatnému zvýšeniu PPV, ale k zníženiu citlivosti áno, preto je jeho predvolená hodnota nastavená na 0.

		3L mapovanie	4L filtrovanie ('minQ=0')	4L filtrovanie ('minQ=1')
mapovanie	citlivosť	45.76%	39.02%	24.85%
	PPV	89.28%	94.92%	95.72%
extr. metylácie	citlivosť	38.84%	33.91%	26.49%
	PPV	94.23%	97.49%	97.51%
detekcia SNV	citlivosť	52.29%	46.40%	32.80%
	PPV	78.36%	91.39%	90.50%

Tabulka 6.2: Citlivosť a pozitívna predikčná hodnota s a bez dodatočného 4-písmenového filtrovania.

V tabuľke 6.3 je zobrazený rozdiel v simulovaných a lokalizovaných metylačných úrovniach pre všetky kontexty bez filtrovania a s filtrovaním.

	simulované	3L mapovanie	4L filtrovanie ('minQ=0')	4L filtrovanie ('minQ=1')
CG	81.00%	75.01%	77.17%	77.63%
CHG	63.00%	59.51%	61.08%	61.37%
CHH	40.00%	39.84%	40.64%	40.52%

Tabulka 6.3: Úroveň metylácie extrahovanej priamo z 3-písmenového mapovania a z filtrovaných dát v porovnaní so simulovanými hodnotami.

Aj z týchto výsledkov je vidieť zlepšenie presnosti odhadu metylačnej úrovne zavedením dodatočného filtrovania, čo potvrdzuje nutnosť používania kontrolných sekvencií pri zarovnaní sekvencií s nízkou komplexitou, akými bisulfitom konvertované transpozóny sú.

Vplyv detekcie chýb a SNV

Zanesenie SNV do generovaných dát simuluje ich výskyt v reálnych údajoch a spolu so sekvenačnými chybami sú hlavným zdrojom možnej odchýlky. Túto situáciu som testoval v experimente, kde bol pre extrakciu metylácie pre rovnaký výstup zarovnania použitý nástroj neschopný detekcie týchto chýb a vzniknutý nástroj, ktorý ich rozoznáva.

V tabuľke 6.4 sú simulované hodnoty metylácie pre každý z kontextov a úroveň metylácie nájdená s detekciou zdrojov chýb a bez nej.

Môžeme vidieť, že detekcia chýb má významný vplyv na výslednú úroveň metylácie a táto skutočnosť je očakávaná, pretože SNV tvoria veľký podiel v simulovaných sekvenciách. Tento rozdiel by mohol byť dokonca ešte aj väčší, pretože mnoho sekvenčných čítaní sa nenamapovalo z časti práve z dôvodu zvýšeného podielu SNV, a tým sa vplyv nukleotidových variácií vo fáze extrakcie redukoval. Zlepšením citlivosti zarovnania by tak táto dodatočná kontrola chýb nabrala ešte väčšiu dôležitosť.

Porovnaním týchto výsledkov a výsledkov predchádzajúceho experimentu pre 3-písmenové mapovanie môžeme dôjsť k záveru, že detekcia SNV je ešte dôležitejšia pre určenie výslednej metylačnej úrovne ako dodatočná 4-písmenová kontrola.

	simulované	bez detekcie	s detekciou
CG	81.00%	74.30%	77.20%
CHG	63.00%	57.90%	61.06%
CHH	40.00%	36.90%	40.36%

Tabulka 6.4: Úroveň metylácie extrahovanej priamo z filtrovaných dát nástrojom Bismark (bez chybovej kontroly) a vytvoreným nástrojom s kontrolou sekvenačných chýb a nukleotidových variant.

Porovnanie s wild-card metódou

Pre fázu zarovnania bisulfitových *readov* je používaný nástroj Bismark pracujúci s 3-písmenovým mapovaním. Cieľom tohto experimentu bolo zistiť, aký bude rozdiel vo výsledných hodnotách mapovania a metylácie, ak by sa pre zarovnanie využila *wild-card* metóda. Spomedzi dostupných nástrojov (tab. 5.1) využívajúcich tento princíp som pre porovnanie vybral nástroj LAST používaný ako súčasť programu Bisulfighter predovšetkým kvôli jeho výpočetnému výkonu a kompatibilite (SAM výstup) s fázou extrakcie metylácie.

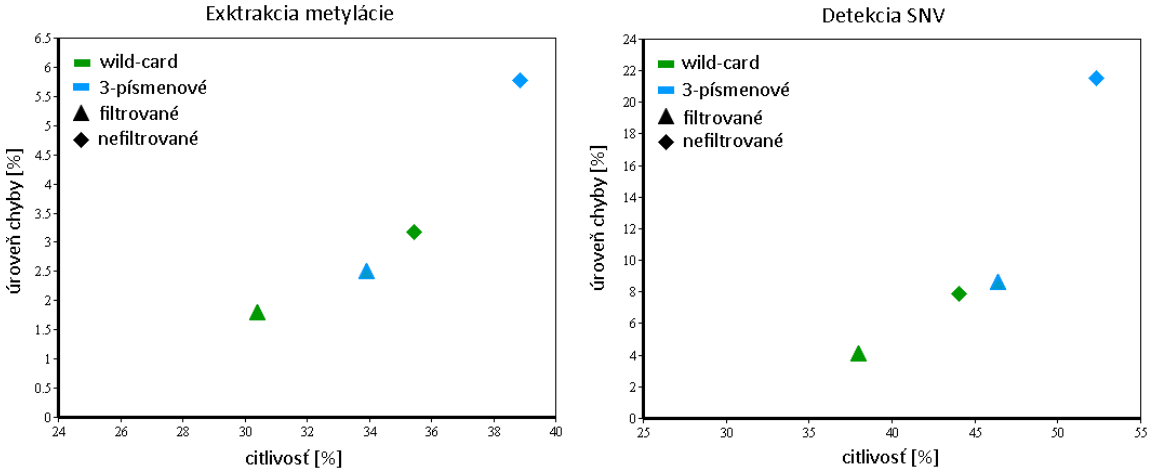
Tabuľka 6.5 zobrazuje výsledky mapovania *wild-card* metódou v porovnaní s 3-písmenovou metódou.

	Bismark		LAST	
	3L	3L+filter	Wild-card	Wild-card+filter
chybne negatívne	418329	493629	434824	551862
chybne pozitívne	14162	5401	7714	3217
chybné mapovanie	31393	11904	87335	21655
korektné mapovanie	379336	323525	306899	255541
citlivosť	45.76%	39.02%	37.02%	30.82%
PPV	89.28%	94.92%	76.35%	91.13%

Tabulka 6.5: Štatistika zarovnania pre 3-písmenovú metódu využívanú Bismarkom a *wild-card* metódu využívanú v nástroji LAST.

3-písmenovou metódou bolo možné nájsť podstatne väčšie množstvo mapovaní, čo vyplýva z hodnoty citlivosti a dokonca aj ich presnosť je významne lepšia. *Wild-card* metóda bola pre tento problém menej efektívna hlavne čo sa týka korektnosti namapovaných *readov* (veľa z nich sa namapovalo, ale na nesprávne pozície). Menšia citlivosť a vyššia úroveň chyby na rastlinných transpozónových sekvenciách s väčším počtom výskytu nukleotidu T odpovedá predpokladom o *wild-card* metóde a voľba 3-písmenového zarovnania sa javí na základe týchto výsledkov ako prijateľnejšia varianta pre fázu mapovania.

Na výslednom súbore zarovnania som následne realizoval extrakciu metylácie rovnakým spôsobom ako pre 3-písmenovú metódu s rovnakým nastavením parametrov. Cieľom bolo zistiť, ako ovplyvní táto zmena v mapovaní výslednú metylačnú úroveň a detekciu SNV. Na obrázku 6.2 je zobrazená citlivosť lokalizácie metylácie a detekcie SNV spolu s odpovedajúcou úrovňou chyby.



Obrázok 6.2: Citlivosť a úroveň chyby filtrovaných a nefiltrovaných dát získané našou 3-písmenovou metódou a *wild-card* metódou implementovanou v nástroji LAST.

Obrázok extrakcie metylácie ukazuje, že *wild-card* metóda má v obidvoch prípadoch nižšiu úroveň chyby (väčšiu pozitívnu predikčnú hodnotu) ako je tomu u 3-písmenovej metódy na úkor menšej citlivosti. Znamená to, že síce našla menej výsledkov, ale tieto výsledky vedú k presnejšej extrakcii. U detekcii SNV je tento fakt podobný, nefiltrovaná detekcia pri *wild-card* metóde je dokonca ešte presnejšia ako filtrovaná 3-písmenová metóda. Z týchto výsledkov vyplýva, že aj keď LAST nájde menej záznamov, jeho mapovanie z hľadiska jednotlivých nukleotidov je presnejšie na čom sa odrazí aj kvalitnejšia extrakcia metylácie (tab. 6.6) a detekcia SNV.

	Bismark		LAST	
	3L	3L+filter	Wild-card	Wild-card+filter
CG	75.01%	77.17%	76.42%	77.26%
CHG	59.51%	61.08%	60.51%	61.11%
CHH	39.84%	40.64%	40.31%	40.54%

Tabulka 6.6: Metylácia v kontextoch CG, CHG a CHH pre 3-písmenovú a *wild-card* metódu zarovnania.

Použitie *wild-card* metódy namiesto 3-písmenovej by teda mohlo viesť k lepším výsledkom, u generovaných dát je však jednoduchšie získať metylačnú úroveň odpovedajúcu parametrom simulácie, pretože redukcia citlivosti tu hrá menšiu úlohu ako by to bolo v prípade reálnych údajov. V nich by mohlo 3-písmenové zarovnanie nájsť viac korektných mapovaní, ktoré by vplývali výraznejšie na výslednú metyláciu než je to u simulovaných dát, kde nie je metylácia tak rôznorodá.

Metylačná mapa pri rozdielne simulovanom klastre

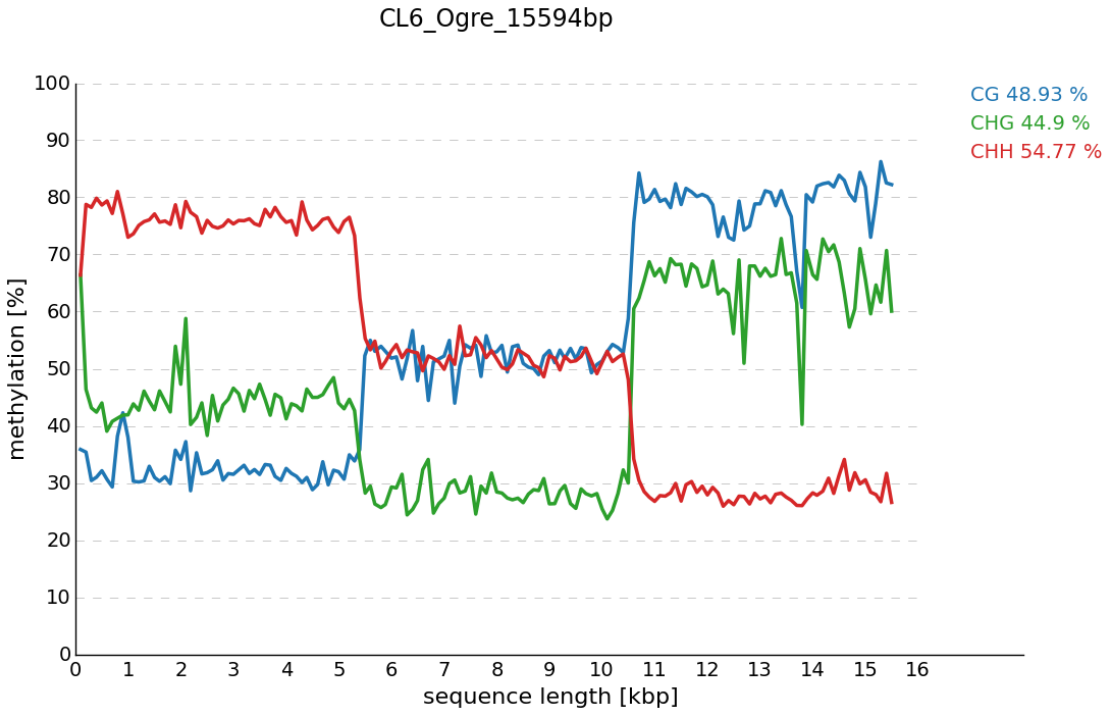
Predchádzajúce experimenty využívali dáta s rovnakou metylačnou úrovňou pre všetky referenčné sekvencie celého klastru. V reálnych situáciách však dochádza k odlišnej mety-

lácii rôznych transpozónových sekvencií danej transpozónovej rodiny. Pre tento experiment som najprv všetky sekvencie vstupného klastru rozdelil do troch segmentov tak, že prvý segment tvorili sekvencie mapované na prvú tretinu konsenzuálnej sekvencie s aspoň 80% dĺžkou a 80% identitou. Druhý segment sekvencie mapované do druhej tretiny a tretí do poslednej. Vznikli tri samostatné FASTA súbory, ktorých úroveň metylácie bola simulovaná s odlišnými hodnotami. Prehľad týchto hodnôt pre každý kontext a pre každý segment je v tabuľke 6.7.

	CG	CHG	CHH
segment 1	31.50%	45.00%	81.00%
segment 2	54.00%	27.00%	54.00%
segment 3	85.50%	72.00%	27.00%

Tabulka 6.7: Simulovaná úroveň metylácie pre jednotlivé časti konsenzuálnej sekvencie klastru.

Simulované *ready* z každého segmentu som spojil do jedného FASTQ súboru, ktorý predstavoval bisulfitovú knižnicu na vstupe procesu lokalizácie metylačných miest vzniknutým nástrojom. Výsledný graf extrahovaných pozícií je zobrazený na obrázku 6.3.



Obrázok 6.3: Metylačná mapa konsenzuálnej sekvencie klastru CL6 s rozdielnymi hodnotami metylácie. Cytozínové kontexty sú farebne odlíšené a veľkosť jedného segmentu predstavuje tretinu dĺžky celej sekvencie (5198bp).

Z metylačnej mapy konsenzuálnej sekvencie klastru je vidieť, že metylačná úroveň jednotlivých segmentov pre každý kontext odpovedá predpokladanému vývoju podľa generovanej metylačnej úrovne. Získané metylačné hodnoty teda pokrývajú celú túto sekvenciu

aj napriek redukovanej citlivosti zarovnania.

6.3 Reálne dáta

Experimenty s reálnymi dátami mali byť pôvodne realizované s bisulfitovými knižnicami osekvenovanej DNA organizmu *Silene latifolia* a zarovnávané na sekvencie klastrov použitých v simulačných experimentoch. Táto rastlina sa vyznačuje veľkým podielom transpozónových sekvencií a vytvorené klastre majú dostatočnú veľkosť potrebnú pre takúto formu experimentov. V časovom harmonograme realizovania tejto práce však bisulfitové sekvenčné čítania neboli dostupné, a preto som zvolil iné organizmy, konkrétne *Zea mays* a *Oryza sativa* s tým, že na bisulfitových dátach *Silene latifolia* to bude testované až budú tieto údaje dostupné a vzniknutý nástroj bude súčasťou výskumu metylácie rastlinných transpozónov.

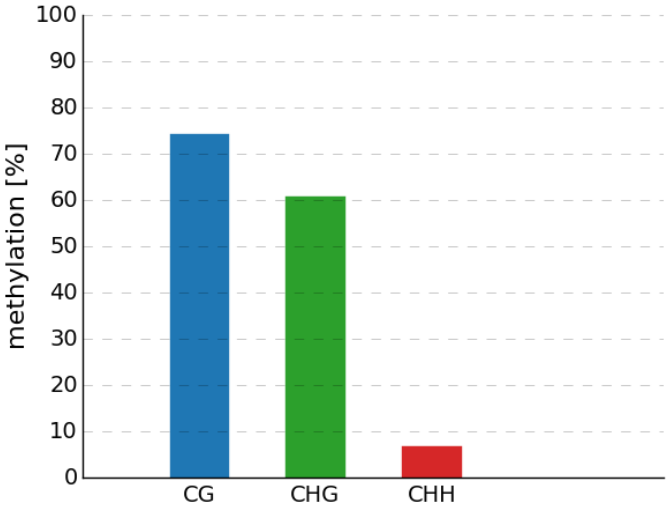
Pre *Oryza sativa* som mal dostupné simulované klastre TE a v prípade *Zea mays* som prostredníctvom RepeatExploreru vytvoril nové klastre z orezanej knižnice genómových sekvencií. V obidvoch prípadoch neobsahovali jednotlivé klastre ani zďaleka také množstvo sekvencií ako to je u organizmu *Silene latifolia* (napr. 690 sekvencií oproti 42110 pre najväčší klaster) a dĺžka sekvencií je taktiež porovnateľne menšia. Z týchto dôvodov bude veľmi malá šanca akéhokoľvek mapovania do týchto klastrov.

Bisulfitové knižnice pre obidva organizmy som stiahol z Európskeho Nukleotidového Archívu (*European Nucleotide Archive* - ENA). Ide o *single-end ready* získané platformou Illumina. Jednotlivé sekvenčné čítania boli kvalitatívne upravené nástrojom Trim Galore podľa popisu predspracovania v sekcii 5.3.3. Vzniknuté *ready* sú taktiež porovnateľne kratšie (približne 50bp) s dátami *Silene latifolia*, s ktorými mali byť tieto experimenty pôvodne realizované (90bp). Ani u jedného z týchto organizmov neboli dostupné kontrolné sekvencie, preto bolo filtrovanie vypnuté.

Kvôli tejto charakteristike dát som očakával, že zarovnanie transpozónových sekvencií a extrakcia metylácie budú pre takéto datasety veľmi neefektívne.

Oryza sativa

U ryži som bisulfitové sekvenčné čítania zarovnával na simulovaný klaster veľkosti 690 sekvencií. Z knižnici obsahujúcej 10677048 *readov* sa do tohto klastru namapovalo len 12497 čítaní. Celkovo bola nájdená metylácia v 570 sekvenciách (83%) klastru a aj napriek takémuto malému počtu mapovaní bolo možné extrahovať metylačnú úroveň. Hodnoty pre jednotlivé kontexty sú zobrazené na obrázku 6.4.



Obrázok 6.4: Metylačná úroveň repetitívnych sekvencií organizmu *Oryza sativa* v kontexte CG (74.44%), CHG (60.88%) a CHH (6.78%).

Tento nález relatívne odpovedá úrovni metylácie TE v rastlinných organizmoch, jeho validita je však silne ovplyvnená nedostačujúcimi vstupnými dátami.

Zea mays

Pri kukurici bola situácia ešte horšia. Použitá bisulfitová knižnica bola síce raz tak veľká, *ready* v nej mali ale menšiu dĺžku (platilo pre všetky skúmané knižnice) a sekvencie v klastroch boli taktiež menšie v porovnaní s *Oryza sativa*. Obidva tieto faktory tak sťažili mapovanie a z celkových 27653233 *readov* sa do zvoleného klastru namapovalo len 17894 sekvencií. Lokalizovaná metylácia predstavovala v kontexte CG 65.02%, CHG 47.55% a CHH 10.06% na 2262 sekvenciách klastru veľkosti 11852. Takisto ako v predchádzajúcom prípade, ani toto nie sú hodnoty, ktoré by sa príliš odlišovali od toho, čo je možné nájsť v rastlinách, ich relevantnosť je však minimálna. Metylačné mapy konsenzuálnych sekvencií u týchto organizmov nemohli byť získané, pretože neexistujú rekonštruované elementy, ktoré by slúžili ako šablóna pre tento proces.

Kapitola 7

Záver

Cieľom tejto práce bolo vytvoriť nástroj na lokalizáciu a extrakciu metylačných miest transponovateľných elementov - transpozónov. Transpozóny sú prvky molekuly DNA, ktoré sú schopné sa premiestňovať z jedného miesta DNA na druhé. Boli objavené v roku 1953 a dlho sa považovali za nečinné, odpadové elementy so žiadnou funkciou. Väčšina z nich tomuto popisu odpovedá, ale existujú aj aktívne transponovateľné elementy produkujúce genetické zmeny pri ich začleňovaní do DNA molekuly. Medzi tieto zmeny patrí predovšetkým ovplyvňovanie génovej expresie a génové mutácie vyúsťujúce v rôzne choroby. Ich genómový vplyv môže byť aj pozitívneho charakteru, pretože existujú dôkazy o ich prispeniu k zväčšovaniu genómu. Organizmy obsahujúce transponovateľné elementy naviac disponujú mechanizmom, ktorý im pomáha s potlačovaním ich aktivity označovaný ako metylácia. Metylácia je doležitá epigenetická modifikácia na nukleotidových bázach, ktorou sa vyznačujú všetky vyššie organizmy a ovplyvňuje celý genóm nie len transponovateľné elementy. Niektoré experimenty dokonca ukazujú, že metylácia by mohla byť výsledkom obrany buniek pred parazatickými transpozónmi, a tým by sa stali transpozóny dôležitou súčasťou evolúcie vyšších organizmov.

Monitorovanie aktivity metylácie týchto mobilných elementov nám preto môže poskytnúť unikátny náhľad do toho, v akom štádiu vývoja sa organizmus najviac bráni pred vplyvom transpozónov a v akom je ich aktivita naopak prospešná. V dnešnej dobe je tieto informácie možné získať vďaka NGS metódam sekvenovania v kombinácii s aplikáciou hydrogénsiričitanu sodného (bisulfitu), výsledkom ktorých sú krátke sekvenčné čítania nesúce informáciu o metylovaných cytozínových bázach. Pre spracovanie bisulfitových dát bolo už v predchádzajúcich rokoch vyvinutých mnoho nástrojov a táto oblasť bioinformatiky venujúca sa analýze epigenetických faktorov je v dnešnej dobe veľmi aktívna. Aj napriek tejto skutočnosti neexistuje jeden nástroj riešiaci unikátne výzvy, ktoré vnášajú transpozónové sekvencie do analýzy metylácie. Ide predovšetkým o nukleotidové mutácie a vysokú úroveň repetície.

V rámci tejto práce vznikol nástroj, ktorý využíva časti už existujúcich odladených programov relevantných pre riešenie určitých fáz celého procesu lokalizácie metylácie a implementuje vlastnú doplnkovú funkcionalitu s adresovaním nedostatkov analýzy metylácie bežne dostupných nástrojov. Nástroj bol rozdelený do piatich hlavných častí, kde každá z nich plní vlastnú úlohu so zarovnaním a extrakciou metylácie vystupujúcich ako tie najdôležitejšie. Z informácií o konkrétnych metylovaných pozíciách jednotlivých sekvencií je získaná podľa navrhnutej metriky celková metylácia pre každý kontext a metylačná mapa pozdĺž celej sekvencie reprezentujúcej jednu rodinu transponovateľných elementov. Tieto informácie je možné využiť ako priamy výstup výsledku experimentu alebo ako vstup pre

ďalšie štatistické spracovanie.

Funkcionalita vzniknutého nástroja bola prezentovaná v experimentoch so simulačnými dátami a bolo v nich ukázané, ako veľmi vplývajú jednotlivé špecifické vlastnosti transpozónov na výslednú kvalitu úrovne metylácie a ako túto kvalitu zlepšujú zavedené techniky. Dodatočné filtrovanie namapovaných sekvencií prinieslo určité zlepšenie v pozitívnej predikčnej hodnote ako pri zarovnaní (z 89.28% na 94.92%), tak aj v extrakcii metylácie (zmena z 94.23% na 97.49%), ale detekcia nukleotidových variant predstavovala rovnako dôležitú, ak nie mierne dôležitejšiu časť v tomto procese. Ďalšie experimenty ukázali, že voľba 3-písmenového mapovania bola vhodnou variantou pre časť zarovnania, ale v celkovom procese extrakcie metylačnej úrovne si viedla porovnateľne s ďalšou metódou pracujúcou s divokou kartou. Pri experimentoch s reálnymi dátami sa sila tohto nástroja príliš neprezentovala, čo s veľkou pravdepodobnosťou mohlo byť spôsobené nekvalitnými zdrojovými knižnicami. Vzniknutý nástroj sa však bude využívať na kvalitnejších dátach pri analýze metylačných miest transpozónov organizmu *Silene latifolia* v prebiehajúcom výskume Akademie věd České republiky, kde bude jeho funkcionalita s reálnymi dátami testovaná podrobnejšie.

Vzniknutý nástroj je možné využívať ako kompletnú *pipeline* pre analýzu bisulfitových dát nižšej kvality, kde je potrebná prísnejšia kontrola kvality. Vďaka tomu, že nástroj pracuje nad klastrami sekvencií získanými pomocou aplikácie RepeatExplorer, v budúcnosti by mohol predstavovať ďalší krok v analýze repetitívnych sekvencií.

Najdôležitejšie časti aplikácie - zarovnanie sekvencií a extrakcia presných metylačných pozícií sú realizované externými nástrojmi so štandardizovanými vstupmi a výstupmi, a preto je možné tieto časti priebežne nahradiť za nové verzie s rozšírenou funkcionalitou alebo kvalitnejšími výsledkami a tým prispievať k ich údržbe. Hlavne časti zarovnania by sa v budúcnosti mala venovať pozornosť, pretože v súčasnosti je veľkým problémom zarovnanie repetitívnych sekvencií, ktoré sa mapujú na viac genomických pozícií. Výsledkom toho sú viacznačné mapovania, ktoré sa neuvažujú pre následnú analýzu, ale je možné ich získať v samostatnom súbore, a preto by mohli byť súčasťou ďalšieho spracovania.

Filtrovanie momentálne poskytuje najlepšie výsledky pre jeho základnú variantu - akceptovanie nálezov s mapovaním na rovnakú pozíciu v 3-písmenovej aj v 4-písmenovej abecede. V tejto časti by mohol byť v budúcnosti testovaný iný nástroj pre zarovnanie ako aktuálny Bowtie2, ktorý by eventuelne mohol poskytovať presnejšie výsledky.

Ako doplnok k fáze spracovania metylačných dát by mohla byť naviac implementovaná štatistická analýza odlišne metylovaných regiónov (DMR) využívajúca textové výstupy tohto nástroja a generujúca prehľadné štatistiky vo forme grafov.

Literatura

[1] 454 pyrosequencing [online]. [cit. 2015-01-15].
URL http://454.com

[2] DNA helix [online]. [cit. 2015-01-15].
URL http://www.wellcome.ac.uk/en/fourplus/DNA.html

[3] Microbial Models: The Genetics of Viruses and Bacteria [online]. [cit. 2015-01-16].
URL http://www.bio.utexas.edu/faculty/sjasper/bio212/microbial.html

[4] Nomenclature and Symbolism for Amino Acids and Peptides [online]. [cit. 2015-04-16].
URL http://chem.qmul.ac.uk/iupac/AminoAcid

[5] Sequence Alignment/Map Format Specification [online]. [cit. 2015-04-16].
URL https://samtools.github.io/hts-specs/SAMv1.pdf

[6] What is FASTA format? [online]. [cit. 2015-04-16].
URL http://zhanglab.ccmb.med.umich.edu/FASTA

[7] Quality Control, trimming and alignment of Bisulfite-Seq data (Prot 57) [online]. [cit. 2015-04-24].
URL http://www.epigenesys.eu/en/protocols/bio-informatics/483-quality-control-trimming-and-alignment-of-bisulfite-seq-data-prot-57

[8] Trim Galore! [online]. [cit. 2015-04-24].
URL http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

[9] Bowtie2: Fast and sensitive read alignment [online]. [cit. 2015-04-25].
URL http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

[10] Andrews, S.: FastQC [online]. [cit. 2014-12-27].
URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[11] Ashikawa, I.: Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *The Plant Journal*, 2001, 26(6), 617-625.

[12] Bao, W.; Jurka, M.; Kapitonov, V.; aj.: New Superfamilies of Eukaryotic DNA Transposons and Their Internal Divisions. *Molecular Biology and Evolution*, 2009, 26(5), 983-993.

[13] Barturen, G.; Rueda, A.; Oliver, J.; aj.: MethyExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*, 2013, 2:217.

[14] Bock, C.: Analysing and interpreting DNA methylation data. *Nature*, 2012, 13(10)705-19.

[15] Cartwright, R.; Graur, D.: The multiple personalities of Watson and Crick strands. *Biology Direct*, 2011, 6:7.

[16] Cock, P.; Fields, C.; Goto, N.; aj.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 2010, 38(6), 1767-1771.

[17] Cokus, S.; Feng, S.; Zhang, X.; aj.: Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 2008, 452(7184): 215-219.

[18] Deaton, A.; Bird, A.: CpG islands and the regulation of transcription. *Genes and Development*, 2011, 25, 1010-1022.

[19] Dennis, E.; Berttell, R.: DNA methylation of maize transposable elements is correlated with activity. *Philosophical transactions of the Royal Society of London*, 1990, 326(1235), 217-229.

[20] Desfeux, A.: Bisulfite-seq analysis [online]. [cit. 2014-12-27].
URL http://www.omictools.com/bisulfite-seq-c1217-p1.html

[21] Feschotte, C.; Pritham, E.: DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual review of genetics*, 2007, 41, 331-368.

[22] Firth, M.; Mori, R.; Asai, K.: A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic acids research*, 2012, 40(13).

[23] Goll, M.; Bestor, T.: Eukaryotic Cytosine Methyltransferases. *Annual review of biochemistry*, 2005, 74, 481-514.

[24] Hansen, K.; Langmead, B.; Irizarry, R.: BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 2012, 13, R83.

[25] Havecker, E.; Gao, X.; Voytas, D.: The diversity of LTR retrotransposons. *Annual review of genetics*, 2004, 5:225.

[26] H.H. Kazazian, e. a.: Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, 1988, 332(10), 164-166.

[27] Ikeda, R.; Kokubu, C.; K. Yusa, e. a.: Sleeping Beauty Transposase Has an Affinity for Heterochromatin Conformation. *Molecular and Cellular Biology*, 2007, 27(5), 1665-1676.

[28] Jarvie, T.: Next generation sequencing technologies. *Drug Discrovery Today: Technologies*, 2005, 2(3), 255-260.

[29] Kawashima, T.; Berger, F.: Epigenteic reprogramming in plant sexual reproduction. *Nature*, 2014, 15, 613-624.

[30] Kent, W.: BLAT–the BLAST-like alignment tool. *Genome Research*, 2002, 12(4)656-64.

[31] Koboldt, D.; Chen, K.; Wylie, T.; aj.: VarScan: variant detection in massively parallel sequencing if indivdual and pooled samples. *Bioinformatics*, 2009, 25(17), 2283-2285.

[32] Krakau, S.: *Developing a BS-Seq Analysis Workflow for Genomic Variation and Methylation Level Calling*. Master thesis, Bioinformatik, Freie Universität Berlin, 2013.

[33] Krueger, F.; Kreck, B.; Franke, A.; aj.: DNA methylome analysis using short bisulfite sequencing data. *The Plant Journal*, 2001, 26(6), 617-625.

[34] Kunde-Ramamoorthy, G.; Coarfa, C.; Laritsky, E.; aj.: Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Research*, 2014, 42:e43.

[35] Langmead, B.; Trapnell, C.; Pop, M.; aj.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 2009, 10:R25.

[36] Li, Y.; Tollefsbol, T.: DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in molecular biology*, 2011, 791:11-21.

[37] Lister, R.; Pelizzola, M.; Dowen, R.; aj.: Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 2009, 462(7271): 315-322.

[38] Lister, R.; Pelizzola, M.; Kida, Y.; aj.: Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 2011, 471, 68-73.

[39] Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, 2011, 17(1), 10-12.

[40] Miki, Y.; Nishisho, I.; A. Horii, e. a.: Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research*, 1992, 52(3), 643-645.

[41] Munoz-López, M.; García-Pérez, J. L.: DNA Transposons: Nature and Applications in Genomics. *Current Genomics*, 2010, 11(2), 115-128.

[42] Murrell, A.: DNA methylation and Genome Stability [online]. [cit. 2015-01-16]. URL http://www-medchem.ch.cam.ac.uk/lab_rotations/murrell.php

[43] Novák, P.; Neumann, P.; Pech, J.; aj.: RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics*, 2013, 29(6)792-3.

[44] Pedersen, B.; Hsieh, T.; Ibarra, C.; aj.: MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, 2011, 27(17), 2435-2436.

[45] Pray, L.: Transposons: The Jumping Genes [online]. [cit. 2014-11-30]. URL http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518

[46] Salanda, V.: *Optimization of the next-generation sequencing data alignment.* Diplomová práce, Faculty of information technology, Brno University of technology, 2013.

[47] SanMiguel, P.; Gaut, B.; Tikhonov, A.; aj.: The paleontology of intergene retrotransposons of maize. *Nature Genetics*, 1998, 20(1), 43-45.

[48] Shendure, J.; Ji, H.: Next-generation DNA sequencing. *Nature biotechnology*, 2008, 26, 1135-1145.

[49] Singal, R.; Ginder, G.: DNA Methylation. *Blood Journal*, 1999, 93(12).

[50] Smith, A.; Chung, W.; Hodges, E.; aj.: Updates to the RMAP short-read mapping software. *Bioinformatics*, 2009, 25, 2841-2842.

[51] Snustad, D.; Simmons, M.: *Genetika*. Munipress, 2009, iSBN 978-80-210-4852-2.

[52] Tatarinova, T.; Kerton, O.: *DNA Methylation - From Genomics to Technology*. InTech, 2012, iSBN 978-953-51-0320-2.

[53] Tran, H.; Porter, J.; Sun, M.; aj.: Objective and Comprehensive Evaluation of Bisulfite Short Read Mapping Tools. *Advances in Bioinformatics*, 2014, 2014:472045.

[54] Yaping, L.; Siegmund, K.; Laird, P.; aj.: Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, 2012, 13:R61.

[55] Yoder, J.; Walsh, C.; Bestor, T.: Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics*, 1997, 13(8), 335-340.

[56] You, J.: BSSim: Bisulfite sequencing simulator for next-generation sequencing [online]. [cit. 2015-05-02].
URL 122.228.158.106/BSSim

[57] Yutaka, S.; Junko, T.; Toutai, M.: Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic acids research*, 2014, 42(6): e45.