INTRASTRAIN VARIABILITY IN TREPONEMAL STRAINS

Vojtěch Bartoň
Master Degree Programme (2), FEEC BUT
E-mail: xbarto80@vutbr.cz

Supervised by: Denisa Maděránková
E-mail: maderankova@vutbr.cz

Abstract: This paper deals with identification and analysis of genetic variability in treponemal strains. We include thirteen strands of three subspecies of *Treponema pallidum*. The proposed workflow identify variable spots in resequenced genomes and proposed a comparision by using whole-genome aligning.

Keywords: genetic variability, *Treponema pallidum*

1 ÚVOD

Využití sekvenačních dat nespočívá pouze v určení dané sekvence zkoumaného organismu. Díky mnohanásobnému čtení jediné pozice, jsme schopni odhalovat i místa v genomu, která se u dané sekvence svou bází liší. V případě, že jsme schopni odfiltrovat změny vzniklé sekvenační chybou a náhodnou mutací, pak se bavíme o identifikaci variabilních míst genomu. Výskyt alternativních alel je důsledkem evolučního adaptačního tlaku vyvíjeného na daný organismus. Alternativní alela může měnit strukturu i funkci daného exprimovaného produktu a tím měnit chování organismu. [1]

V naší práci se zaměříme na organismus *Treponema pallidum* a jeho poddruhy *pallidum* (TPA), *pertenue* (TPE) a *endemicum* (TEN). V rámci několika genomů provedeme identifikaci variabilních pozic a především jejich srovnání napříč všemi zpracovávanými genomy.

2 METODOLOGIE IDENTIFIKACE VARIABILNÍCH MÍST

K identifikaci variabilních míst využijeme jejich osekvenované soubory. Sekvenace probíhala prostřednictvím technologie Illumina paired-end. Zvolili jsme inovativní metodiku umožňující nám filtraci skutečně variabilních míst od míst, kde je variabilita způsobena čistě náhodnou mutací, či jde o chybu způsobenou použitou technologií. Identifikace variability probíhá zvlášť u každého zahrnutého genomu a následně je variabilita porovnávána napříč genomy pomocí celogenomového zarovnání sekvencí.

2.1 SESTAVENÍ A FILTRACE

Jednotlivé soubory čtení jsou předem zkontrolovány pomocí programu FastQC [2]. K samotnému sestavení použijeme program BWA s algoritmem mem [3]. Následně pomocí Samtools [4] odstraníme nenamapovaná a duplikovaná čtení. Ze souboru rovněž odstraníme čtení, u kterých není namapován celý pár. Soubor si převedeme do pozičně orientovaného **.vcf* (variant call format) souboru. Zahrneme však pouze čtení s vysokým Phred skóre, kde máme jistotu, že je báze na dané pozici určená správně.

Z analyzovaného souboru vyřadíme všechny pozice, jež mají nízkou hloubku čtení nebo je referenční báze ve více jak 99 % čtení, takovýto úsek považujeme za vysoce konzervovaný. Alternativní báze

musí být podpořena alespoň 8 čteními, abychom zamezili výskytu falešně pozitivní identifikace variability. Poměr čtení přímého a reverzního vlákna se musí pohybovat mezi 0,4–2,3 (maximální poměr je 30/70), v případě, že tomu tak není považujeme úsek za vysoce chybově čtený. Ze souboru vyřadíme také veškeré pozice nacházející se v homopolymerních úsecích (>5 bp) nebo v jejich okolí ($\pm$ 5 bp), kde je chybovost technologie mnohem vyšší. Pro každý genom zvlášť navíc určíme práh variability jako odhad sekvenační chyby a odstraníme pozice, na kterých je výskyt alternativní alely pod tímto prahem. Práh stanovíme na takovou hodnotu, kdy by jeho další snižování způsobilo více jak 10% nárůst 1. diference počtu identifikovaných variabilních míst v okně o délce 4, při zvoleném vzorkování prahu 0,01. Při těchto hodnotách jsme schopni oddělit málo variabilní místa, kde nejsme schopni rozlišit variabilitu od náhodné mutace, či chyby způsobené použitou technologií. Ostatní pozice označíme za variabilní. Přehled prahů a nalezených pozic je ukázán v tabulce 1.

2.2 ZAROVNÁNÍ SEKVENCÍ

Všechny použité sekvence treponemálních kmenů je třeba zarovnat. Jelikož jde o sekvence genomů stejného druhu (*Treponema pallidum*) jsou si všechny velmi podobné a nejsou od sebe fylogeneticky příliš vzdálené. Pro zarovnání jsme využili algoritmu MUSCLE [5], především kvůli jeho vyšší rychlosti a nižší výpočetní náročnosti. Zarovnání probíhá v posuvném okně o délce 5 000 bp s překryvem 2 000 bp. Před posunem okna jsou odstraněny veškeré mezery, které jsou přidány na konec aktuálně zpracovávaného okna. Nastavené hodnoty jsou zvoleny tak, aby byl celý postup rychlý a zároveň dokázal zohlednit i poměrně dlouhé indely. Celý zarovnaný soubor je poté ještě manuálně zkontrolován.

3 VYHODNOCENÍ VARIABILITY

Identifikovaná variabilní místa každého genomu promítneme do zarovnaného souboru a převedeme pozici v genomu na pozici v zarovnání.

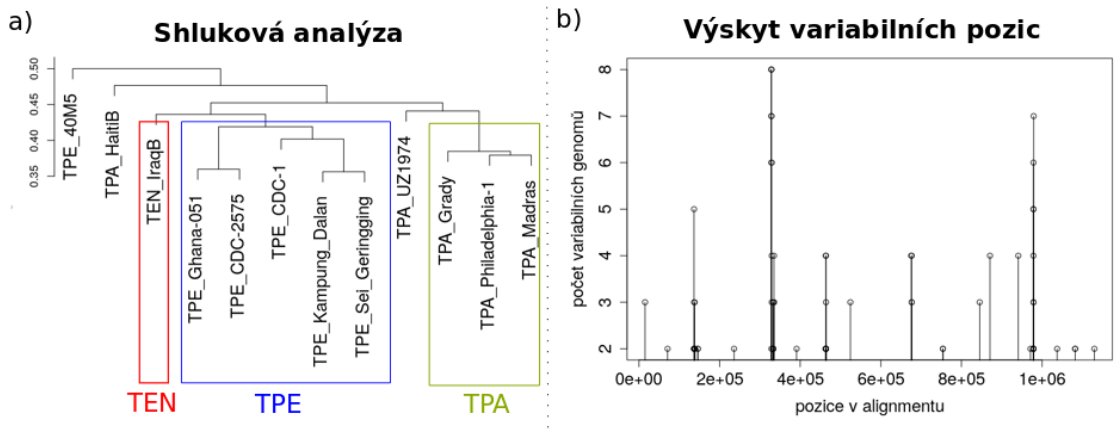
Celkem jsme identifikovali 35 shodných pozic, kde se vyskytují variabilní místa v alespoň dvou genomech, jak je vidět na obrázku 1b. Z toho 30 spadá do kódujících úseků. Ve všech genomech na stejné variabilní pozici alternují stejné báze. V 77% jde o substituce typu tranzice.

Provedením shlukové analýzy (UPGMA) (obrázek 1a) na základě korelace výskytu variabilních míst v genomech dochází k vytvoření shluků podle poddruhu. Do shluků nezapadají především genomy s vyšším prahem identifikované variability, tedy s nízkou kvalitou sekvenace. Obecně se tedy výskyt variabilních míst jeví jako druhově specifická charakteristika jednotlivých genomů.

Tabulka 1: Analyzované genomy

	Genom	Reference*	Práh variability	Variabilních míst
TEN	IraqB	CP007548.1 (BosniaA)	4,20 %	99
TPA	Grady	CP004011.1 (SS14)	5,90 %	13
	HaitiB	LF BIO	8,80%	15
	Madras	CP004010.2 (Nichols)	2,70 %	117
	Philadelpia-1	LF BIO	2,90 %	71
	UZ1974	LF BIO (Philadelphia-1)	3,20 %	77
TPE	CDC-1	LF BIO	8,30 %	18
	CDC-2575	CP020366.1	5,20 %	53
	Ghana-051	CP020365.1	7,50 %	47
	Kampung Dalan	LF BIO	5,40 %	56
	M540	LF BIO	9,00 %	15
	Sei Geringging	LF BIO	3,40 %	70

* GenBank identifikátor použité sekvence, v závorce je uveden referenční genom liší-li se od zpracovávaného. LF BIO = sekvence poskytnutá Biologickým ústavem Lékařské fakulty Masarykovy univerzity.



Obrázek 1: a) Shluková analýza metodou UPGMA na základě korelací výskytu variabilních míst. b) Výskyt vícečetných variabilních pozic podle počtu genomů s výskytem.

4 ZÁVĚR

V práci jsme představili nový a ucelený postup identifikace variabilních míst v genomech. Nalezené pozice dále porovnáváme na základě celogenomového zarovnání. Identifikovali jsme celkem 30 variabilních pozic v kódujících úsecích, jejichž výskyt je shodný ve více zpracovávaných genomech. Shlukovou analýzou bylo ukázáno, že výskyt variabilních pozic je druhově závislou charakteristikou.

Variabilní místa přispívají k adaptabilitě daného organismu a svým výskytem mohou ovlivnit mnohé vnitřní procesy organismu jako je třeba rezistence vůči vnějším vlivům. Analýza variability přispívá ke studiu infekčních mechanismů, či k identifikaci kmenových subpopulací.

Pro účely sestavení a filtrace sekvenačních dat vznikl skript v Bashi a dále soubor funkcí v jazyce R, pro zarovnání sekvencí a další analýzu variabilních míst pro potřeby Biologického ústavu LF MU.

PODĚKOVÁNÍ

Rád bych zde poděkoval týmu prof. MUDr. Davida Šmajse, Ph.D., z Biologického ústavu Lékařské fakulty Masarykovy univerzity za poskytnutí dat a četné konzultace.

REFERENCE

[1] ČEJKOVÁ, D., M. STROUHAL, et al. (2015) A Retrospective Study on Genetic Heterogeneity within Treponema Strains: Subpopulations Are Genetically Distinct in a Limited Number of Positions. *PLOS Neglected Tropical Diseases*. DOI: 10.1371/journal.pntd.0004110.

[2] ANDREWS, S. (2010) FastQC: a quality control tool for high throughput sequence data. Dostupné z: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[3] LI, H., R. DURBIN. (2009) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754-1760. DOI: 10.1093/bioinformatics/btp698.

[4] LI, H., B. HANDSAKER, A. WYSOKER, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079. DOI: 10.1093/bioinformatics/btp352.

[5] EDGAR, R. C. MUSCLE. (2004) multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Research*, **32**, 1792-1797. DOI: 10.1093/nar/gkh340.