IDENTIFICATION OF GENETIC HETEROGENEITY IN SEQUENCING DATA

Vojtěch Bartoň

Master Degree Programme (1), FEEC BUT

E-mail: xbarto80@stud.feec.vutbr.cz

Supervised by: Denisa Maděránková

E-mail: maderankova@feec.vutbr.cz

Abstract: This paper describes the preparation of data obtained from Illumina sequencing platform to analyze low frequency genetic changes within population of *Treponema* strains, i.e. DNA heterogeneity. A method of characterization of raw sequencing data is proposed. The method includes data filtering to achieve better quality of the collected data and calculation of parameters determining which ones of the DNA changes should be considered as heterogeneity.

Keywords: genetic heterogeneity, sequencing, reads quality

1. ÚVOD

Výskyt genomové heterogenity (zahrnující bodové mutace, indely a mobilní elementy jako plasmidy a fágy) je známý u kmenů mnoha patogenních bakterií. Heterogenní oblasti mohou přispívat k obraně vůči imunitní reakci hostitelského organismu, případně reprezentují adaptivní změny jako reakci na různorodé prostředí infikovaného organismu a jeho části. Identifikace heterogenních oblastí je důležitým krokem ke studiu infekčních mechanismů, šíření infekce, imunitních reakcí a identifikaci kmenových subpopulací [1].

V této práci popisuji způsob identifikace heterogenity ze sekvenčních dat genomů bakterií rodu *Treponema*. Tento rod zahrnuje několik nekultivovatelných lidských a zvířecích patogenů způsobujících syfilis, endemický syfilis nebo frambézii. Ze vzorků nakažených hostitelů byla odhalena přítomnost různých subpopulací u jednoho hostitele, projevujících se například jako rezistentní vůči fagocytóze. Genetická diverzita uvnitř jednotlivých kmenů byla objevena v genech z *tpr* rodiny. Předmětem výzkumu je hledání dalších heterogenních úseků.

2. PŘÍPRAVA DAT

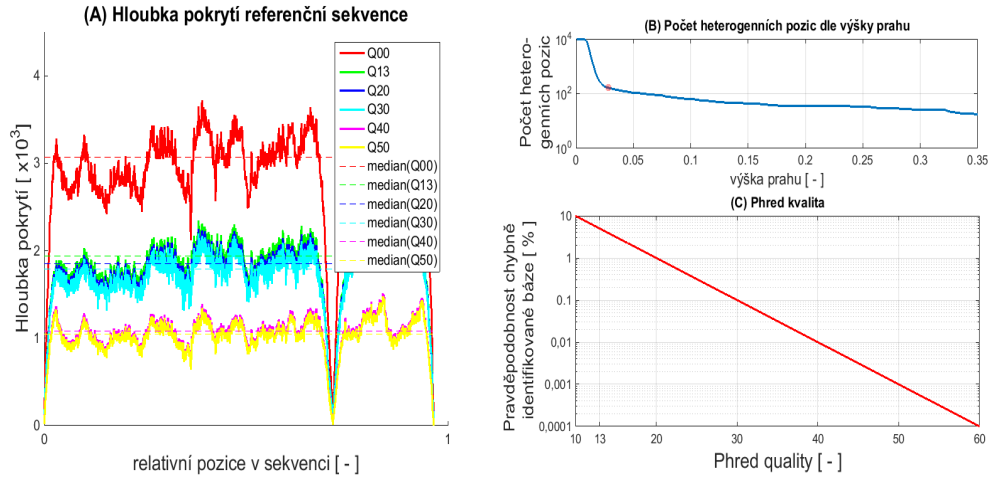
Genom *Treponmy pallidum subsp. pallidum* byl osekvenován metodou pooled sequencing, kdy je celý genom rozdělen do několika menších částí – poolů tak, aby vzájemně podobné geny ležely v jiných sekvenovaných částech. Takovéto rozdělení na pooly odstraní nejednoznačnost identifikace původu sekvence čtení (jednotlivé osekvenované subsekvence). Pooly jsou výhodné i z hlediska rozdělení velmi objemného datového souboru (desítky GB textových souborů) do několika menších, lépe zpracovatelných souborů.

Jednotlivé pooly se vzájemně překrývají, protože pokrytí genomu jednotlivými čteními bývá výrazně nižší na začátku a na konci sekvenovaného řetězce. Při sestavení celého genomu se pak použijí čtení z obou překrývajících se částí poolů a tím dojde k navýšení počtu čtení v oblasti.

2.1. SESTAVENÍ GENOMU

K sestavení sekvenačních dat zahrnujících přes 20 miliónů čtení je používán BWA aligner [2] s algoritmem MEM, který je vhodný pro použití na data z platformy Illumina. Genom se sestavuje vůči referenční sekvenci. Po namapovaní čtení k referenci jsou ze souboru odstraněny veškerá nepou-

žitá (nenamapovaná) čtení. Dále jsou identifikovány potenciální PCR duplikovaná čtení a tyto kopie jsou také odstraněny. Data z jednotlivých poolů je třeba sloučit do jednoho souboru a konsolidovat oblasti spadající do překrývajících se částí jednotlivých poolů. Po prvotní filtraci dostáváme soubor asi 19 miliónů subsekvencí, čítajících přes 2 miliardy čtení jednotlivých pozic.



Obrázek 1: Hloubka pokrytí sekvence (A), počet nalezených pozic dle výšky prahu (B), Phred kvalita (C).

2.2. SKÓRE PHRED KVALITY

Phred skóre je vyjádřením kvality identifikace nukleobází z automatického sekvenátoru. Původně byl tento systém měření kvality vytvořen pro Human Genome Project, poté se však široce rozšířil a je používán většinou automatických sekvenačních technik, pro stanovení kvality získaných sekvencí.

Phred kvalita Q je logaritmickou závislostí pravděpodobnosti P, že měřená nukleobáze je určená chybně. Závislost ukazuje obrázek 1(C).

$$Q = -10 \cdot \log_{10} P \qquad (1)$$

2.3. ANALÝZA KVALITY SOUBORU

Kvalitu sestaveného souboru hodnotíme z několika hledisek. Prvním hodnoceným parametrem je délka osekvenované části kompletního genomu. V ideálním případě 100 %, vzhledem k "N" (neurčitým) bázím v referenci (z důvodu vysoce variabilní *tprK* oblasti genomu, pro kterou nebyla získána referenční sekvence) nelze hodnoty 100 % dosáhnout. Snažíme se však k této hodnotě přiblížit, aby bylo možné nalézt v ideálním případě veškeré heterogenní oblasti genomu. V našem analyzovaném genomu dosahujeme hodnoty 99,98 %.

Dalším hodnoceným parametrem je kvalita určení původu (namapování) jednotlivých sekvencí čtení. Kvalita je udávána v Phred skóre. Vzhledem k poměrně dlouhým čtením (průměrně 118±12nt) je jejich původ určen poměrně přesně a jejich Phred kvalita se pohybuje nad hodnotou Q50, což považujeme za více než dostačující, neboť máme jistotu, že je dané čtení namapováno správně s pravděpodobností vyšší jak 99,999 %.

Kromě délky osekvenované části genomu je důležitá i hloubka jeho pokrytí. Pojmem hloubka pokrytí zde rozumíme počet čtení, které pokrývají danou oblast osekvenovaného genomu. V našem případě sledujeme počet čtení jednotlivých nukleotidů referenčního řetězce. Čím vyšší hloubka pokrytí, tím lepší statistická podpora pro stanovení heterogenních oblastí.

Dalším hodnoceným parametrem je kvalita čtení jedné referenční báze v jednotlivých subsekvencích [3]. Závislost hloubky pokrytí na kvalitě čtení této pozice u části analyzovaného genomu je zobrazena na obrázku 1(A). Vzhledem k povaze heterogenity, která se vyskytuje i ve velmi malém procentu populace, jsou do další analýzy zahrnuty jen ta čtení pozice, která mají chybovost mini-

málně o řád nižší, tedy minimálně Q30 (chybovost 0,1 %), lépe však až Q50 (chybovost 0,001 %). Tímto dojde k vyřazení až 65 % čtení a hloubka pokrytí se sníží v mediánu z 1 745 na 613. Z původních 2 miliard čtení jednotlivých pozic pracujeme dále jen s 800 milióny vysoce kvalitními čteními jednotlivých pozic.

3. STANOVENÍ HETEROGENNÍCH OBLASTÍ

Pro označení oblasti jako heterogenní jsme stanovili následující podmínky. Za prvé, pokrytí oblasti musí být aspoň 50 bází, aby byla zaručena dostatečná statistická podpora pro vyslovení závěru. Za druhé, výskyt dominantní báze je menší nebo roven 99 % pokrytí pozice. Od hodnoty 99 % a výše považujeme danou pozici za silně konzervovanou. Za další, alternativní báze je podpořena alespoň 8 bázemi, abychom mohli vyloučit falešně pozitivní čtení alternativní báze na dané pozici, způsobené chybou v sekvenování. A nakonec výskyt alternativní alely je vyšší jak stanovený práh. Závislost počtu pozic označených jako heterogenní na zvolené výšce prahu ukazuje obrázek 1(B). Prudký nárůst počtu pozic v levé části grafu je zřejmě způsoben vyšším zastoupením náhodných mutací nebo chybou sekvenovaní. Práh určíme buď okometricky tak, aby nebylo zahrnuto prudké stoupání v levé části grafu, nebo jej určujeme výpočetně jako hodnotu, kdy jeho následující snižování o jednu tisícinu způsobí nárůst počtu pozic označených jako heterogenní o více než 5 %.

Zjištěné pozice označené jako heterogenní srovnáváme s referenčním genomem a zjišťujeme, zda jde o synonymní, či nesynonymní záměnu a také zda jde o transici, či transverzi, a které proteiny jsou případnou změnou zasaženy.

4. ZÁVĚR

Byl představen navržený způsob předzpracování a filtrace surových sekvenačních dat pro identifikaci heterogenních oblastí genomu *Treponemy*. Pro práci se sekvenačními daty bylo použito několik volně dostupných programů (BWA [2], Samtools [3]) a pro další analýzu bylo zvoleno programové prostředí R.

Byly vytvořeny funkce pro extrakci počtu a obsahu jednotlivých čtení referenční sekvence z vcf formátu (variant call format), sloučení jednotlivých poolů, určení dominantní báze, výpočet poziční frekvence výskytu alel a funkce pro filtraci souboru dle zadaných podmínek pro heterogenní oblasti. Dále vznikl skript pro určení zasaženého genu v heterogenních oblastech a určení o jaký druh změny se jedná. Uvedený postup vznikl ve spolupráci s Biologickým ústavem Masarykovy univerzity a lze jej aplikovat pro další analýzu heterogenních oblastí v genomech *treponemálních* kmenů a jejich srovnání.

PODĚKOVÁNÍ

Rád bych zde poděkoval prof. MUDr. Davidu Šmajsovi, Ph.D., z Biologického ústavu Lékařské fakulty Masarykovy univerzity za odborné vedení, četné konzultace a poskytnutí sekvenačních dat.

REFERENCE

[1] ČEJKOVÁ, D., M. STROUHAL, S. J. NORRIS, G. M. WEINSTOCK, D. ŠMAJS and M. PICARDEAU. A Retrospective Study on Genetic Heterogeneity within Treponema Strains: Subpopulations Are Genetically Distinct in a Limited Number of Positions. *PLOS Neglected Tropical Diseases* [online]. 2015-10-5, **9**(10). DOI: 10.1371/journal.pntd.0004110.

[2] LI, H. and R. DURBIN. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* [online]. 2010, **26**(5), 589-595. DOI: 10.1093/bioinformatics/btp698.

[3] LI, H., B. HANDSAKER, A. WYSOKER, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [online]. 2009, **25**(16), 2078-2079. DOI: 10.1093/bioinformatics/btp352.