DETECTION OF GENOME VARIATIONS

Tomáš Beluský
Master Degree Programme (2), FIT BUT
E-mail: xbelus01@stud.fit.vutbr.cz

Supervised by: Tomáš Martínek
E-mail: martinto@fit.vutbr.cz

Abstract: The influence of variations in human genome is huge at first glance on people's appearance showing several differences between individuals and populations. Also a behavior and probability of occurrence of certain disease are influenced in large way by differences between genomes. Due to inability of whole genome sequencing of a human individual, the existing methods work with reads produced by next-generation machines. Each method has its advantages and limitations and each detects a subset of variations. Appropriate combination of these methods can lead to increase of accuracy and sensibility in genome variation detection.

Keywords: structural variations, alignment, reference genome, read pair, split read, depth of coverage

1 ÚVOD

Zmeny medzi ľuďmi môžeme nájsť už na najnižšej úrovni – v kompletnej genetickej výbave, v genóme. Okrem rozmanitosti medzi ľuďmi ako jedincami sú študované i zmeny v rôznych populáciách. Variácie sú takisto vyhľadávané i v rámci určitého regiónu, ktorý napríklad spôsobuje určitú chorobu, a po získaní variácii práve v tomto regióne sa môžeme dozvedieť viac o samotnej príčine choroby. Príchod výpočtových metód pre detekciu genómových variácii umožnil samotný príchod sekvenovania druhej generácie (*angl.* next-generation sequencing, NGS). Tieto metódy vznikajú kvôli nemožnému získaniu kompletnej sekvencie ľudského genómu. Dnešné technológie rozdelia genóm na menšie časti a tie sa nasekvenujú z jedného alebo oboch koncov, ktoré nazývame ready. Tieto ready sú následne zarovnané k referenčnému genómu pre nájdenie jednotlivých variácii. Referenčný genóm je zostavený z DNA niekoľkých darcov.

2 ROZDELENIE VARIÁCIÍ

Štruktúrne variácie je možné rozdeliť na [1]: inzercie (pridanie novej sekvencie oproti referenčnej), delécie (odstránenie určitej sekvencie), inverzie (reverzia sekvencie), duplikácie (tandemové a rozptýlené skopírovanie sekvencie) a translokácie (presunutie sekvencie vrámci chromozómu alebo intrachromozomálne). Väčšina nástrojov deteguje práve tieto variácie, resp. ich podmnožinu. Okrem spomenutých štruktúrnych variácii existujú aj omnoho komplexnejšie variácie. Takisto existuje najjednoduchšia odchýlka medzi dvoma genómami a to práve zmena jedného nukleotidu na iný, čo nazývame jednonukleotidovými variáciami (*angl.* single nucleotide polymorphism, SNP).

3 EXISTUJÚCE METÓDY

Metódy sa delia podľa toho, akú informáciu využívajú pri detekcii variácii [1, 2, 3]. Metóda *read pair* už podľa názvu používa párové ready, ktoré sú zarovnané nesúhlasne. To znamená, že sú zarovnané v nesprávnej orientácii, poradí alebo sú vzájomne prehodené. Táto metóda sa ďalej delí na klastrovacie a distribučné metódy. Zo všetkých metód deteguje najširšiu škálu typov a veľkosti štruktúrnych

variácii. Avšak na druhú stranu je limitovaná v detekcii väčších inzercií ako je vzdialenosť medzi readmi. Taktiež vzdialenosť medzi readmi sa môže líšiť, čo zhoršuje rozlíšenie. Novšie metódy kombinujú prístupy klastrovacích a distribučných metód pre odstránenie nedostatkov.

Ďalšia metóda, *split read*, podobne pracuje s párovými readmi. Avšak oproti predošlej metóde pracuje s readmi, z ktorých sa iba jeden správne zarovnal a druhý nie je buď zarovnaný vôbec, alebo sa rozdelil na viaceré časti, ktoré môžu mať rôznu orientáciu aj poradie. Split read metódy sú veľmi vhodné pre detekciu hraníc variácii, čo dokážu s presnosťou na jednu bázu. Kvôli nejednoznačnému zarovnaniu krátkych readov sú v súčasnosti efektívne len v unikátnych regiónoch genómu. Väčšie a komplexnejšie variácie môžu byť detegované s použitím väčšej dĺžky readov.

Tretia metóda, *read depth*, deteguje variácie na základe hĺbky pokrytia, ktorá udáva počet, koľkokrát bol daný genóm nasekvenovaný. Pomocou tejto metódy sme schopní odhaliť zmeny v počte kópii v genóme (*angl.* copy number variations, CNV). Väčší počet pokrytia v regióne oproti pokrytiu v celom genóme znamená duplikáciu, menší naopak deléciu. Ostatné variácie nie je táto metóda schopná odhaliť. Takisto nie je schopná určiť lokalizáciu detegovaných CNV. Väčšina týchto metód odstraňuje rôzne šumy a mapuje ready pomocou rozšírených štatistických modelov pre zvýšenie presnosti a rozlíšenia.

Posledná metóda, *sequence assembly*, sa môže rozdeliť na *de novo* a *comparative assembly*. Prvá metóda nevyžaduje zarovnanie readov k referenčnému genómu, ale naopak snaží sa zostaviť genóm kúsok po kúsku. Druhá metóda pri zostavovaní používa aj referenčný genóm. Najvhodnejšie použitie je pre menšie genómy, napríklad pre genóm baktérie. Pre zostavenie ľudského genómu sa používa menej, pretože zostavenie krátkych readov do sekvencie v opakujúcich sa regiónoch je náročné. Z technických príčin sa sústreďuje len na neopakujúce sa regióny. Je schopná detegovať len limitované množstvo štruktúrnych variácii.

Kvôli limitáciám spôsobeným NGS je odhadom stále 50 % variácii v ľudskom genóme nedetegovaných. Krátke ready a veľké chyby vo väčších readoch limitujú súčasné metódy v detekcii variácii v opakujúcich sa a duplikovaných regiónoch. Kombinácia predošlých metód sa zdá byť sľubnou a touto cestou sa už aj niektoré práce vybrali.

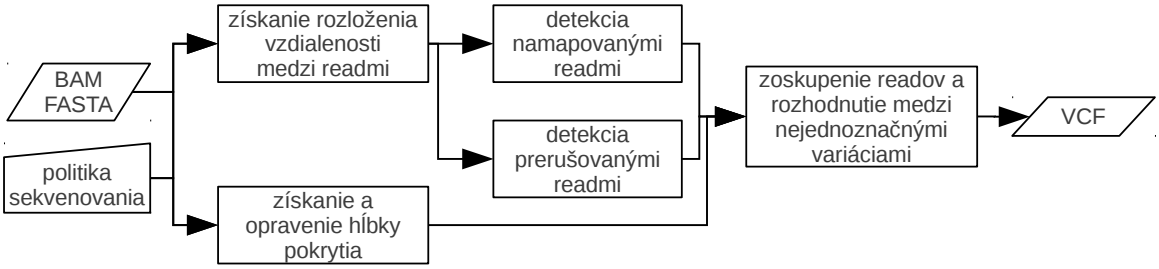
4 NÁVRH NÁSTROJA

Vo vlastnom nástroji budú použité prvé dve opísané metódy, ktoré sa použijú podľa toho ako sa daný párový read podarilo namapovať. Čiže v prípade namapovania oboch readov sa detegujú variácie metódou read pair a v druhom prípade, keď sa jeden read nenamapoval alebo sa namapoval prerušovane, sa použije split read. Okrem toho pre zvýšenie spoľahlivosti detekcie bude získaná hĺbka pokrytia, ktorá musí byť najskôr kvôli obsahu GC opravená. Obsah GC uvádza podiel páru G-C v rámci celého genómu, ktorý býva typicky väčší ako obsah AT. V regiónoch s veľkým množstvom AT stúpa hĺbka pokrytia, ak stúpa obsah GC a v regiónoch s veľkým množstvom GC naopak klesá hĺbka pokrytia so stúpajúcim množstvom GC. Túto korekciu používa nástroj Rdxplorer [4] spoločne s ďalším množstvom nástrojov založených na hĺbke pokrytia. V každom okne i o veľkosti 100 bp bude získaný počet readov r_i na základe ich počiatočnej pozície. Budú pripočítané iba ready označené ako správne namapované. Následne bude vykonaná korekcia podľa vzorca $r_i = r_i * \frac{m}{m_{GC}}$, kde m je medián zo všetkých okien a m_{GC} je medián zo všetkých okien s rovnakým percentuálnym obsahom GC ako okno i. Percentuálny obsah GC je vypočítaný ako $c_{GC} = \frac{G+C}{A+T+G+C} * 100$, kde jednotlivé písmená uvádzajú počet odpovedajúcich báz v rámci daného okna.

Pred samotným vyhodnotením variácii budú ready detegujúce rovnakú variáciu pridané do klastrov pre zvýšenie spoľahlivosti. Okrem zvýšenia spoľahlivosti dôjde k rozhodnutiu medzi nejednoznačnými variáciami. Vytvorené klastre môžu takisto bližšie určiť veľkosť, hranice a presnú sekvenciu danej variácie. K tomuto prispejú najmä prerušované ready, ktoré už zo svojej podstaty detegujú kom-

pletné sekvencie variácii alebo aspoň jednu hranicu sekvencie.

Vstupom systému budú namapované ready na referenčný genóm uložené v BAM súbore a takisto bude vstupom samotný referenčný genóm vo FASTA súbore. Okrem vstupných súborov bude nutné zadať, aká politika bola použitá pri sekvenovaní párových readov, resp. z ktorého vlákna je získaný ktorý read. Výstupom systému bude VCF súbor, ktorý bude obsahovať všetky nájdené variácie. Celý tento proces zobrazuje obrázok 1.



Obrázek 1: Dátový tok návrhu vlastného nástroja

Výsledná implementácia bude vytvorená ako balíček pre skriptovací jazyk Python, ktorý je medzi bioinformatikmi veľmi používaný, spoločne so štatistickým jazykom R a jeho balíkom Bioconductor. Vytvorený balíček môže byť súčasťou väčšieho zreťazeného spracovania, keď po namapovaní readov na referenčný genóm, príde práve na rad tento balík a jeho výstup môže byť takisto následne ďalej spracovaný. Spomínaný jazyk R môže takisto využiť vytvorený balík, napr. prostredníctvom vlastného balíka rJython, ktorý práve dokáže spúšťať programy napísané v jazyku Python. Rovnako môže byť využitý i v iných jazykoch, ak toto prepojenie podporujú.

5 ZÁVER

Skúmanie ľudských genómov je silne závislé na sekvenačných technológiách a je možné očakávať, že s príchodom sekvenátorov tretej generácie sa spoľahlivosť súčasných metód zvýši, prípadne vzniknú nové metódy. Avšak kým nenastane doba, keď bude možné získať kompletnú sekvenciu genómu akéhokoľvek človeka, budú súčasné metódy a ich obmeny stále používané a preto je vhodné sa nimi zaoberať. Prezentovaný návrh sa snaží skombinovať tri metódy pre zvýšenie spoľahlivosti a citlivosti detekcie variácii. Výsledný nástroj bude porovnaný s vybranými vhodnými implementáciami pre porovnanie výkonnosti.

REFERENCE

[1] Alkan, C., Coe, B. P., Eichler, E. E.: Genome structural variation discovery and genotyping, Nature Reviews Genetics, ročník 12, č. 5, 2011, s. 363–376

[2] Hoogendoorn, E.: Computational methods for the detection of structural variation in the human genome, diplomová práce, Utrecht Graduate School of Life Sciences, 2012

[3] Teo, S. M., Pawitan, Y., Ku, C. S., aj.: Statistical challenges associated with detecting copy number variations with next-generation sequencing, Bioinformatics, ročník 28, č. 21, 2012, s. 2711–2718, doi:10.1093/bioinformatics/bts535

[4] Yoon, S., Xuan, Z., Makarov, V., aj.: Sensitive and accurate detection of copy number variants using read depth of coverage, Genome Research, ročník 19, august 2009: s. 1586–1592, doi:10.1101/gr.092981.109