

Ćwiczenie nr 3 z MBI, Resekwencjonowanie genomu człowieka

Kinga Kimnes, Jakub Skałeczki

11 grudnia 2019

W ramach przeprowadzanego ćwiczenia wykorzystano plik FASTA zawierający fragment ludzkiego genomu referencyjnego (GRCh37) uzyskanego przez Human Genome Consortium w 2009 roku. Plik ten zawiera sekwencję chromosomu 1.

1 Mapowanie

Początkowo dokonano indeksowania pliku FASTA poprzez zastosowanie programu BWA (Burrows-Wheeler Alignment Tool). Jest on stosowany do mapowania sekwencji o małej rozbieżności względem dużego genomu referencyjnego, takiego jak genom ludzki. Pakiet BWA składa się z trzech algorytmów: BWA-backtrack, BWA-SW oraz BWA-MEM, przy czym dla zapytań o wysokiej jakości zalecany jest ostatni z nich z uwagi na wydajność i dokładność. Następnie, za pomocą ostatniego ze wskazanych algorytmów - BWA-MEM, wygenerowano plik SAM (Sequence Alignment/Map) zawierający mapowania kolejnych sekwencji z pliku coriell_chr1.fq na sekwencję referencyjną, którą stanowił plik chr1.fa (sekwencja chromosomu 1). Typowa (średnia) długość odczytu to 75.5bp, o odchyleniu standardowym równym około 1.17. W dalszej kolejności dokonano sortowania mapowań oraz wygenerowano plik BAM (Binary Alignment/Map - stanowiący binarny odpowiednik pliku SAM). Rozmiary plików wynosiły kolejno: FASTQ - 57MB, SAM - 77MB, BAM - 14MB. Plik FASTQ zawiera sekwencje (odczyty) oraz zapisaną symbolicznie ocenę jakości dla każdego nukleotydu (ilość równa długości sekwencji). Plik SAM zawiera ponadto informacje o mapowaniu sekwencji na genom referencyjny, co czyni go większym.

2 Wizualizacja w IGV oraz przykładowy wariant w genie IQGAP3

3 Wykrywanie wariantów

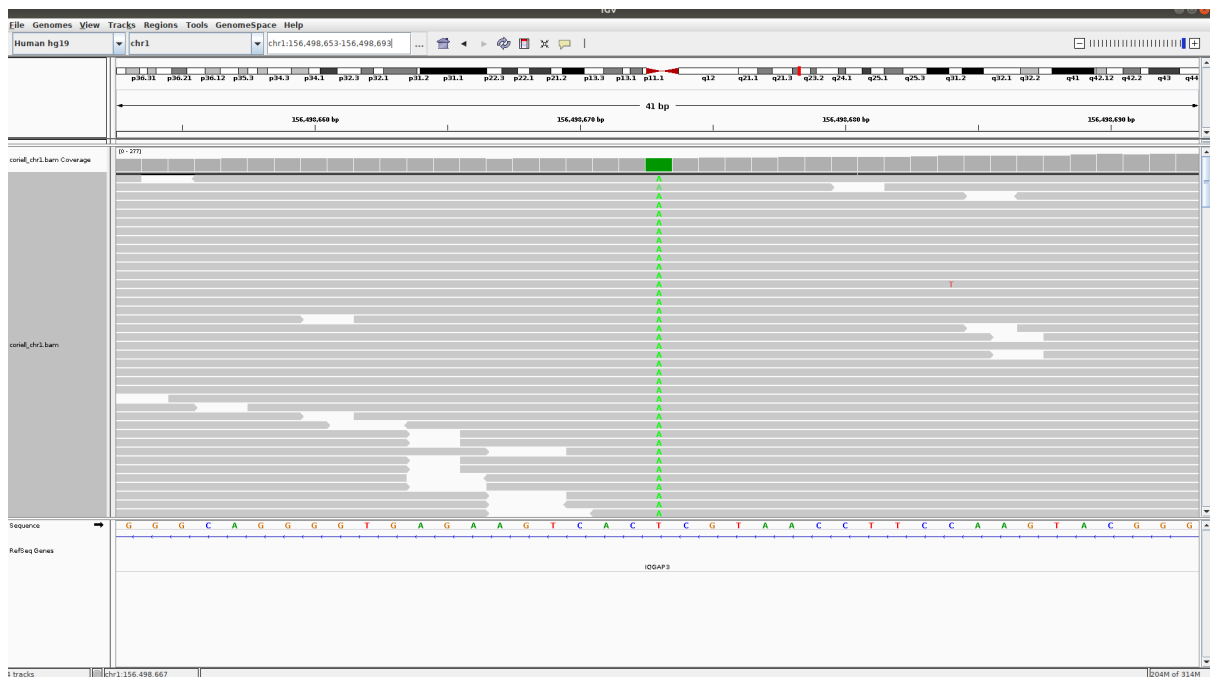
Pozycja wybranego wariantu wynosi 156.498.673. Genom referencyjny wskazuje nukleotyd T, natomiast odczyt A. Zrzut ekranu z programu IGV przedstawiono na rysunku 2.

4 Adnotacje wariantów

W wyniku przeprowadzonej za pomocą narzędzia Variant Effect Predictor adnotacji przefiltrowanego pliku VCF uzyskano wykres typu "pie chart" przedstawiający statystyki wariantów. Największy udział procentowy posiada intron_variant (36%). Odpowiadające mu wiersze zamieszczono w tabeli 1.

Tabela 1:

. 1:156498673-156498673 A intron_variant MODIFIER IQGAP3 ENSG00000183856 Transcript ENST00000361170.2 protein_coding 35/37 -----rs1171564 --1 -HGNC 20669 ----0.5998 -----
. 1:156498673-156498673 A upstream_gene_variant MODIFIER snoU13 ENSG00000238843 Transcript ENST00000458777.1 snoRNA -----rs1171564 449 1 -RFAM ----0.5998 -----
. 1:156498673-156498673 A intron_variant,NMD_transcript_variant MODIFIER IQGAP3 ENSG00000183856 Transcript ENST00000491900.1 nonsense_mediated_decay 34/37 -----rs1171564 --1 -HGNC 20669 ----0.5998 ----



Rysunek 1: Przykładowy wariant w oknie IGV

Jest to wariant w sekwencji kodującej (świadczy o tym wartość protein_coding w kolumnie Exon) w pierwszym wierszu powyższego listingu.

5 Wnioski

Ćwiczenie polegało na analizie danych pozyskanych w trakcie resekwencjonowania fragmentu genomu człowieka. Przedmiotem badań był fragment sekwencji, dotyczący chromosomu pierwszego (chr1). Mapowanie danej sekwencji do sekwencji referencyjnej dostarczyło nam dodatkowych danych w postaci wykrytych wariantów. Dane z pliku SAM w sposób czytelny dla człowieka wizualizuje narzędzie IGV. Pozwala ono na sprawdzenie nie tylko faktu różnic w stosunku do genomu referencyjnego, ale także zmierzenie tej różnicy (liczba oznaczonych nukleotydów innych niż referencyjne, zsumowana dla wszystkich pokryć). Raport wygenerowany z filtrowanego pliku VCF poprzez narzędzie Variant Effect Predictor pozwolił dowiedzieć się o typie adnotowanych wariantów. Wiele z nich występuje w sekwencjach niekodujących - intronach (36%) oraz transkryptach niekodujących (10%). Spośród części kodujących, najczęściej jest wariantów synonimicznych (57%), czyli takich, które kodują ten sam aminokwas. Jest to możliwe dlatego, że DNA jest "zdegenerowane", to znaczy istnieje kilka sekwencji (kodonów), które kodują ten sam aminokwas.