Laboratorium 7 i 8

Czas na oddanie finalnego raportu:	6 tygodni
Nazwa pliku:	imie_nazwisko_7_8_bio.pdf
Typ ćwiczenia:	dwutygodniowe

1 Wejdź na stronę serwera Galaxy. Dalsza praca wymaga założenia konta.

2 Załaduj plik `NA12878_subset_R1.fq.gz` do narzędzia Galaxy poprzez **Upload data -> Paste/Fetch Data** ("Choose local files"). Koniecznie zaznacz w Genome(set all) "hg19Patch10". Dane z pliku `NA12878_subset_R1.fq.gz` pochodzą z sekwencjonowania całego genomu próbki oznaczonej jako NA12878 pochodzącej od kobiety z USA. Uzyskano je w ramach "1000 genome project". Żeby analiza odbywała się szybciej plik zawiera tylko częściowe dane tzn. tylko z chromosomu 10. Na podstawie znajomości baz danych z poprzednich zajęć odpowiedz jaki ma rozmiar ten chromosom u człowieka, znajdź jego sekwencję referencyjną w bazach danych i podlinkuj (jak to zrobić znajdziesz w metodyce do ćwiczeń z baz danych) [**task 1**].

3 Oryginalnie sekwencjonowanie wykonano w technologii "double-end" ("paired-end"), ale dla celów demonstracyjnych plik jest w formacie "single-end". Jaka jest różnica między nimi [**task 2**]?

4 Kiedy już plik się załaduje otwórz podgląd pliku (ikona oka na liście zadań z prawej strony obok pliku) i odpowiedz na pytanie: jakiego systemu sekwencjonowania użyto (nazwa widoczna jest po symbolu "@" w pierwszej linii pliku i powtarza się w kolejnych, pierwsze pięć liter po "@") [**task 3**].

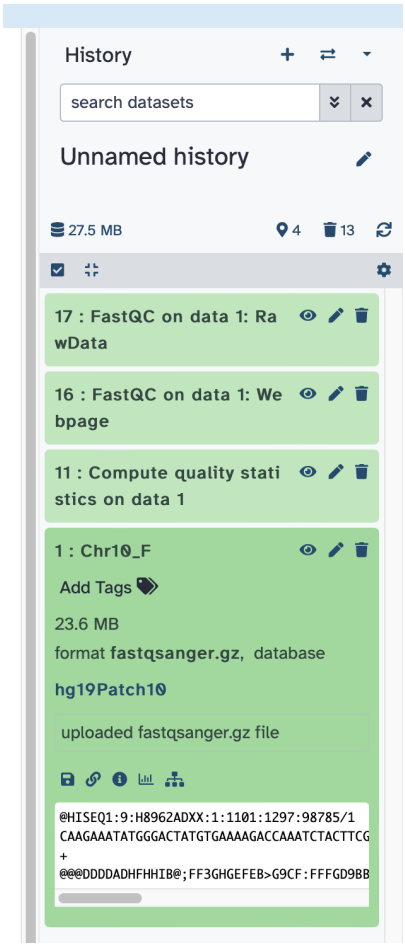
5 Dane prezentowane są w formacie FASTQ, charakterystycznego dla sekwenatorów Illumina, jeden odczyt reprezentowany jest przez 4 linijki, które odpowiednio zawierają (1) opis, (2) sekwencję, (3) znak "+", (4) symbole ASCII reprezentujące dane dotyczące jakości danego odczytu. Skopiuj 4 linijki z pierwszego odczytu jako wynik realizacji tego zadania [**task 4**].

6 Zmień nazwę załadowanego pliku (ołówek przy nazwie pliku) na `Chr10_F` (chromosom 10, kobieta(female))

7 Żeby skontrolować jakość danych użyj narzędzia FastQ Quallity Control (menu po lewej stronie, sekcja GENOMIC FILE MANIPULATION). Załaduj dane do "FastaQC Read Quality reports" i kliknij "Run task".

Alternatywnie, żeby było szybciej, możesz zainstalować lokalnie FastQC i przeprowadzić taką analizę u siebie na komputerze FastaQC.

1. W trakcie trwania analizy spójrz do dokumentacji FastQC i odpowiedz na pytanie czym się różnią dane dobrej jakości (Good Illumina Data) od tych złej jakości (Bad Illumina Data) w raportach wynikowych FastQC [**task 5**]
2. Wyświetl raport wygenerowany przez FastQC na serwerze Galaxy lub lokalnie. W wyniku analizy otrzymasz dwa raporty:



Rysunek 1: Screen 1

- pierwszy zawiera wyniki w postaci pliku tekstowego “Raw Data”
- drugi prezentuje wyniki w postaci graficznej “Webpage”

3. Jaką liczbę odczytów sekwencji zawiera plik (patrz na “Total Sequences”) [**task 6**]. Czy zawiera sekwencje zakwalifikowane przez algorytm jako złej jakości, jeśli tak to w jakiej ilości (patrz na “Sequences flagged as poor quality”) [**task 7**], jakiej długości są odczyty (patrz na “Sequence length”) [**task 8**]

4. Obejrzyj wykres Per base sequence quality (w “Webpage”). Na podstawie dokumentacji programu wyjaśnij co i w jaki sposób reprezentuje ten wykres [**task 9**]. Odpowiedz na pytanie: czy analizowane dane są dobrej jakości [**task 10**]?

5. Obejrzyj wykres “Per sequence quality scores”. Na podstawie dokumentacji programu wyjaśnij w jakiej sytuacji tzw. error rate wynosi 0,2%, w jakiej sytuacji 1% i co oznacza error rate w sekwencjonowaniu (wpisz w google “error rate sequencing”) [**task 11**].

6. Obejrzyj wykres “Per base sequence content”. Dlaczego ilość AT oraz CG powinna się zgadzać [**task 12**]?

7. Per base N content - co to jest N w sekwencji DNA pochodzącej z sekwencjonowania [**task 13**]?

8 Po kontroli jakości uzyskanych w sekwencjonowaniu danych przechodzimy do mapowania do genomu:

1. W Galaxy w menu z lewej strony znajdź odnośnik “Mapping”, następnie “Map with BWA-MEM”. Do czego służy to narzędzie [**task 14**]? Jakiej długości odczyty są optymalne dla jego użycia [**task 15**]?

2. W “Using reference genome” zaznacz “Human (Homo sapiens) (b37): hg19”, w “Single or Paired-end reads” wybierz “Single”. Upewnij się, że w “Select fastq dataset” znajduje się Twój zbiór danych. Uruchom proces mapowania klikając na “Run Tool”. Wygenerujesz plik BAM.

3. W Galaxy w menu z lewej strony znajdź “SAM/BAM”, następnie “Samtools flagstat tabulate descriptive stats for BAM datset”. Upewnij się, że Twój plik BAM jest wybrany do analizy (sprawdź numer porządkowy z historii i jego nazwę pod śródtytułem “BAM File to report statistics of”) i uruchom narzędzie. Co ciekawe, możesz to narzędzie uruchomić nawet jeśli plik BAM jeszcze nie jest wygenerowany.

4. Poczekaj na zakończenie działania poprzednich skryptów. Następnie w historii przy rekordzie “Samtools flagstat on data” kliknij w ikonę z okiem. Narzędzie to również wykonuje kontrolę jakości, z linijki zawierającej “QC-passed reads + QC-failed reads” odczytaj ile odczytów przeszło QC (quality control) [**task 16**]?

5. Z linijki nr. 7 (zawiera słowo “mapped”) odczytaj jaka część sekwencji została zmapowana do genomu (w procentach) [**task 17**]? Co to znaczy mapować sekwencje do genomu (wpisz w google “przyrównywanie (ang. alignment) odczytów do sekwencji wzorcowej” żeby znaleźć odpowiedź) [**task 18**]?

6. W Galaxy w menu z lewej strony znajdź odnośnik “BAM-to-SAM convert..” i przekształć uzyskany plik BAM w plik SAM (aby zobaczyć podgląd pliku), gdyż plik BAM jest binarny.

7. W historii przy rekordzie "BAM-to-SAM on data.." kliknij w ikonę z okiem. Z nagłówka pliku i pierwszej linijki zaczynającej się od "@SQ" odczytaj do jakiego chromosomu mapuje większość odczytów ("chr...") [**task 19**]

8. Można również sprawdzić to ręcznie kopiując odczyt

   ```
   ATCAAGGAAATAAAAGAGGATACAAACAAATGGAAGAACATTCCATGCCCATGGGTCGG
   AAGAATCCATATTGCGAAAATCGCCATACTGCCCCAGGCCTTTTCCAGATTCAATGCCA
   TCCCCATCAAGCTACCAATGACTTTCTTCA
   ```

   (sekwencję widniejącą pod nagłówkiem w linijce zaczynającej się od `HISEQ1:13:H8G92ADXX:2:1102:2234:` i sprawdź za pomocą BLAST (Database: refseq*representative_genomes, organism: _Homo sapiens*). Podaj Accession Number rekordu z bazy o największym podobieństwie [**task 20**]. Zauważ, że wprawdzie pierwszy rekord należy do chromosomu, który odpowiada temu z poprzedniego zadania, jednakże w innych chromosomach sekwencje również są bardzo podobne. Dlatego mapowanie DNA do genomu referencyjnego nie jest takie proste, szczególnie przy krótkich odczytach.

9 Po zmapowaniu danych przejdziemy do ich wizualizacji:

1. Pod rekordem w Historii o nazwie "Map with BWA-MEM on data.." znajduje się opcja wizualizacji w postaci ikony wykresu. Po kliknięciu w tą ikonę widnieje kilka możliwości. Wybierz "display at UCSC (main)". W narzędziu, które się otworzy przejdź do obszaru `chr10:96,425,415-96,614,984` i kliknij "Go". Kliknij w śródtytuł "Map with BWA-MEM on data.." i wybierz z manu podręcznego "squish" i zobaczysz wizualizację wszystkich odczytów (dostosuj powiększenie tak aby zobaczyć pojedyncze "read'y"). Jak widzisz obszar ten pokryty jest wieloma odczytami w Twoich danych. Znajdź nazwy dwóch genów, które znajdują się w tym obszarze (ich nazwy zaczynają się od CYP2...) [**task 21**].
2. Chcąc bliżej przyjrzeć się fragmentowi tej sekwencji wpisz `chr10:96541600-96541630` i kliknij "Go". W jednej z pozycji w genomie referencyjnym jest "G", natomiast w Twoich danych w 50% odczytów jest "A" (co za tym idzie w 50% pozostałych jest "G"). Czy potrafisz wyjaśnić co to oznacza (tzn. jaki genotyp ma osoba, której dane analizujesz GG, GA, czy AA) [**task 22**]? Spróbuj ogólnie odpowiedzieć jakie konsekwencje może mieć ten genotyp dla tej osoby (zobacz tutaj) [**task 23**]

10 Po wizualizacji zmapowanych danych przeprowadzimy tzw. "Variant calling":

1. W Galaxy w menu z lewej strony znajdź odnośnik "VCF/BCF", następnie "bcftools mpileup". Do czego służy to narzędzie [**task 24**]?
2. Upewnij się, że Input BAM/CRAM znajduje się plik BAM, który został wygenerowany poprzednio, a w pozycji "Select reference genome" wybierz "Human (Homo sapiens): hg19". Żeby móc obejrzeć wygenerowany plik w pozycji "output_type" wybierz "uncompressed VCF". Kliknij "Run tool"
3. Po tym jak narzędzie zakończy działanie wejdź do pliku VCF klikając symbol oka obok rekordu z Historii o nazwie "bcftools mpileup on data..". Jako efekt tego zadania skopiuj pierwszą linijkę znajdującą się pod nagłówkiem pliku (linie nagłówka zaczynają się od "##") czyli pierwszą linijkę rozpoczynającą się od "chr10" [**task 25**]
4. W Galaxy w menu z lewej strony znajdź odnośnik "BCF/VCF", następnie "bcftools call". Dane wejściowe to wynik poprzedniego programu mpileup (czyli dopilnuj żeby plik VCF

znalazł się jako input dla tego programu). W "Input/output Options" zaznacz "Yes" przy "Output variant sites only". Żeby móc obejrzeć wygenerowany plik w pozycji "output_type" wybierz "uncompressed VCF". Kliknij "Run tool"

5. Sprawdź i wpisz jako odpowiedź do zadania - ile linii posiada plik wygenerowany uprzednio narzędziem mpileup (widnieje w historii jak rozwiniesz kafelek z napisem "bcftools mpileup on data..") oraz ile linii wygenerowało ostatnie narzędzie tj. "bcftools call" (analogicznie jak poprzednio w historii przy odpowiednim kafelku) [**task 26**]. Skąd może wynikać ta różnica [**task 27**]?
6. Otwórz pogląd pliku VCF ("bcftools call.." –> ikona oka z prawej strony). W pierwszym rekordzie reference allel to G, an alternate allel A. Czasami podawany jest też drugi alternate allel. W kolumnie "Map_with_BWA_MEM.." znajduje się oznaczenie pokazujące obecność alleli w naszym genomie badanym, gdzie 0 to reference allel, 1 to alternate allel, a 2 to second alternate allel czyli zapis "0/1" oznaczałby w naszym przypadku GA (to jest genotyp). Odczytaj genotyp dla kolejnych trzech oznaczonych zmienności genetycznych (trzech kolejnych linijek) [**task 28**]. Pamiętaj, genotyp to nie zawsze tylko dwa nukleotydy - w linijce nr. 3 jest to trochę bardziej skomplikowane.
7. W kolumnie INFO widnieje pole DP, który ma różną wartość w przypadku każdego wariantu. Wiedząc, że jest to tzw. sequence depth odpowiedz na pytanie co ten parametr oznacza [**task 29**]
8. Żeby lepiej podsumować znalezione warianty genetyczne znajdź wśród narzędzi w Galaxy, narzędzie o nazwie: bcftools counts. Do czego służy [**task 30**]? Kliknij "Run tool".
9. Wejdź do poglądu ("bcftools counts.."–>ikona oka) i odpowiedz w jakiej ilości i jakie zmienności genetyczne zostały wykryte [**task 31**]. Wyjaśnij jednym zdaniem co oznaczają: SNP, INDELs, MNPs [**task 32**].
10. W Galaxy w menu z lewej strony znajdź odnośnik "VCFfilter" i spróbuj odpowiedzieć na pytanie - co robi ten program [**task 33**]. Input dla tego narzędzia to plik VCF uzyskany z użyciem "bcftools call". W polu "Specify filtering value" umieść "QUAL > 200". Odpowiedz na pytanie: co oznacza "QUAL > 200" [**task 34**]? Kliknij "Run tool".
11. Rozwijając kafelek w historii z nagłówkiem "VCFfilter: on data. . . " widzisz ile zmienności genetycznych spełniło zadane przez Ciebie warunki (number of lines). Jaki jest to % wszystkich zmienności, które wcześniej wykazano w pliku VCF (liczba ta widnieje w narzędziu "bcftools counts..") [**task 35**]?
12. Wejdź do wygenerowanego teraz pliku VCF ("VCFfilter: on data. . . "–>ikona oka) i spróbuj znaleźć marker genetyczny, który poprzednio był analizowany w narzędziu UCSC. Znajdź ten marker wyszukując pozycję na chromosomie 10 "96541616". Odczytaj tym razem z pliku VCF ponownie jaki genotyp znajduje się w tej pozycji (genotyp to np. "CT") [**task 36**]

11 Pobierz wszystkie pliki wygenerowane przez Ciebie w narzędziu Galaxy i dołącz je do raportu końcowego w MS Teams jako dodatkowe załączniki oprócz pliku pdf.