

PODSTAWY BIOINFORMATYKI
LISTA ZADAŃ 3

DANE NGS - KONTROLA JAKOŚCI

1. Czym różnią się dane single-end od danych paired-end? Wyjaśnij jakie korzyści płyną z sekwencjonowania typu paired-end?
2. Korzystając z opisu kodowania jakości w formacie FASTQ, określ jaki zakres wartości przyjmują znaki w formacie Illumina 1.5+?
3. Określ jakie kodowanie zostało użyte do zdefiniowania jakości nukleotydów w następujących sekwencjach:
 - a. #####\$%&00035689 DFFGHGHHFDDDCDEFGHIJ
 - b. @@<FFDFFGHGHHFDDDGHHHDDDDHIIJJDDIIGDabcdef
4. Jakość nukleotydu wyrażona jest wzorem $Q = -10 \log_{10} P$. Jakie jest prawdopodobieństwo wystąpienia błędu przy jakości równej 20?
5. Program FastQC jest bardzo popularnym programem do kontroli jakości surowych danych NGS. Jest on dostępny w wersji online (Galaxy), jak i w wersji stand-alone. W celu skorzystania z platformy <https://usegalaxy.org/> należy utworzyć tam konto oraz zalogować się. Załaduj dane w formacie fastq (SeqA.fastq oraz SeqB.fastq) na serwer (Upload Data). Po wczytaniu danych zapoznaj się z opcjami znajdującymi się obok nich („oko”, „ołówek” oraz „krzyżyk”), a także z narzędziami („Tools”) dostępnymi w panelu po lewej stronie ekranu. Przeprowadź kontrolę jakości za pomocą opcji FastQC Read Quality reports i odpowiedz na pytania:
 - a. Dla których parametrów danych zostały zwrócone ostrzeżenia i awarie? Które przeszły kontrolę bez zastrzeżeń?
 - b. Jaka jest najczęstsza średnia jakość sekwencji w obu plikach?
 - c. Ile sekwencji reprezentuje najwyższą średnią jakość?
 - d. Jaka jest średnia jakość danych dla pozycji 36?
 - e. Który z plików jest lepszej jakości?
6. Korzystając z platformy Galaxy, odfiltruj te sekwencje, w których jakość nukleotydów spada poniżej 20 (Filter FASTQ). Następnie, ponownie wygeneruj raport dotyczący jakości danych. Czy jakość uległa poprawie? Na podstawie nowego raportu, odpowiedz na pytania z podpunktów od (a) do (e) z poprzedniego zadania.

DANE NGS – PIPELINE

1. Zaprojektuj pipeline, którego wynikiem będą polimorfizmy genetyczne typu SNP. Podaj wszystkie kroki i korespondujące z nimi programy.
2. Do czego służy program BWA? Jakie pliki wsadowe należy mu zapewnić?
3. Jakiego rodzaju informacje są przechowywane w formacie BAM? Na podstawie poniższego fragmentu danych wyjaśnij zawartość kolumn 1,3,4 oraz pola CIGAR.

HWI-1KL157:58:D2FVAACXX:2:2313:3871:71331 147 Chr15 33794413 50 101M =
33794252 -262
GCTCAGCTTTCTTCACAGTCCAACTCTCACATCCATACATGACCACTGGAAAAACCATA
GCCTTGACTGGACGGACCT
TTGTTAGAGGTTGCTAAAGACTG
DBACCDDCDDDDCC>DECDDC@;?3?3HAHGEJHHEFHDAF@CF=EDEEGDGG@IGF@
HEHGCGIIJIG
GDIHFGIJIJIEIIJIGG?FHHHFDDDD@@@ NM:i:2 AS:i:91 XS:i:83 RG:Z:D2FVAACXX_2