**Zadanie: Dla jednej dowolnej próbki zinterpretuj jakość danych. Załącz dwa wybrane wykresy które uzasadniają interpretację.**

**A screenshot of a graph

Description automatically generated**

Dane mają wysoką jakość na początku odczytu (pozycje 1-50) i jakość spada w dalszych pozycjach. Jest to normalne zjawisko w sekwencjonowaniu, ponieważ dokładność zwykle maleje wraz z długością odczytu. Warto rozważyć przycięcie końców odczytów, aby uzyskać lepszą ogólną jakość, zwłaszcza jeśli końcowe fragmenty mają być używane do dokładnych analiz.

**A graph with a line

Description automatically generated**

Dane sekwencyjne mają wysoką jakość, a większość odczytów osiąga średnie wartości jakości w zakresie 35–39, co jest typowym wskaźnikiem dobrej jakości sekwencjonowania. Taki rozkład sugeruje, że dane mogą być użyteczne w analizach wymagających wysokiej precyzji, gdyż prawdopodobieństwo błędów jest bardzo niskie dla większości odczytów.

**Zadanie: Wybierz pierwszy wiersz opisujący mapowanie inne niż i) pełne dopasowanie odczytu do referencji (100M) lub ii) brak mapowania. Wklej kod CIGAR i zinterpretuj go**

Wiersz:

@PG ID:hisat2 PN:hisat2 VN:2.2.1 CL:"/usr/local/bin/hisat2-align-s --wrapper basic-0 -p 2 -x /cvmfs/data.galaxyproject.org/managed/hisat2\_index/mm10/mm10 --read-lengths 100 -U /tmp/2916798.unp"

Kod cigar:

-x /cvmfs/data.galaxyproject.org/managed/hisat2\_index/mm10/mm10

CIGAR zapisuje informacje o dopasowaniach, wstawieniach, usunięciach itp.

* **-x**: opcja -x służy do wskazania ścieżki do pliku lub katalogu, który zawiera zindeksowany genom (w tym przypadku genom myszy wersja mm10). Indeks ten jest używany podczas mapowania sekwencji RNA do odpowiednich miejsc w genomie.
* **/cvmfs/data.galaxyproject.org/managed/hisat2\_index/mm10/mm10**: To jest pełna ścieżka do pliku indeksu genomowego. W tym przypadku jest to plik indeksu genomu mm10 (mysz) przechowywany na serwerze Galaxy (platforma do analizy danych w bioinformatyce). Jest to miejsce, w którym przechowywany jest zindeksowany genom, którego HISAT2 używa do mapowania sekwencji.

**Zadanie: Podaj statystyki zliczonych cech zawarte w jednym z outputów narzędzia**

**A screenshot of a computer

Description automatically generated**

Statystyki z narzędzia featureCounts pokazane na załączonym obrazie przedstawiają liczby przypisanych i nieprzypisanych odczytów w pliku BAM dla próbki MCL1-DL. FeatureCounts to narzędzie do liczenia odczytów w genomie lub transkryptomie, które przypisuje odczyty do konkretnych genów lub innych funkcji genomowych na podstawie pliku BAM.

Oto szczegóły poszczególnych kategorii:

1. **Assigned**: 765
   * Liczba odczytów poprawnie przypisanych do funkcji (genów). Są to odczyty, które spełniają kryteria przypisania i zostały zliczone do określonych funkcji.
2. **Unassigned\_Unmapped**: 40
   * Odczyty, które nie zostały przypisane, ponieważ nie udało się ich zmapować na genom lub transkryptom. Mogą to być np. odczyty o niskiej jakości mapowania.
3. **Unassigned\_MultiMapping**: 192
   * Odczyty, które nie zostały przypisane, ponieważ zmapowały się w więcej niż jednym miejscu (wielokrotne mapowanie). Zazwyczaj odczyty takie są pomijane, aby uniknąć niepewności w przypisaniu.
4. **Unassigned\_NoFeatures**: 91
   * Odczyty, które nie zostały przypisane, ponieważ nie pokrywają żadnych funkcji (genów) w pliku referencyjnym. Mogły one zmapować się do regionów niekodujących lub regionów międzygenowych.
5. **Unassigned\_Ambiguity**: 36
   * Odczyty, które nie zostały przypisane, ponieważ były ambiwalentnie przypisane do więcej niż jednej funkcji. Narzędzie featureCounts nie przypisuje takich odczytów, aby uniknąć błędów wynikających z niejednoznacznego przypisania.

**Podsumowanie**

* Łącznie przypisanych odczytów: **765**
* Łącznie nieprzypisanych odczytów: suma odczytów we wszystkich kategoriach "Unassigned" (40 + 192 + 91 + 36 = 359)

**Zadanie: Opisz co znajduje się w pliku factordata**

**A screenshot of a computer

Description automatically generated**

Id próbki i do jakiej grupy należy (CellTypeStatus)

**Zadanie: Opisz co znajduje się w pliku annotations**

**A screenshot of a computer

Description automatically generated**

Zawierają enterez id, symbol genu i nazwę genu.

**Zadanie: Opisz co oznacza sekcja “Contrast of Interest”**

Termin "contrast of interest" odnosi się do porównania, które jest kluczowe dla danej analizy danych ekspresji genów.

"Contrast" w kontekście narzędzia limma oznacza różnicę między dwiema grupami lub warunkami, które chcemy porównać. Kontrast może obejmować porównania między różnymi typami próbek (np. chory vs. zdrowy, przed leczeniem vs. po leczeniu) lub różnymi warunkami eksperymentalnymi.

"Contrast of interest" to konkretny kontrast, który jest przedmiotem zainteresowania w analizie, czyli porównanie, które ma dostarczyć istotnych biologicznie informacji. Użytkownik określa ten kontrast w limma, aby dowiedzieć się, które geny są różnie wyrażane między grupami (np. które geny są istotnie zmienione w grupie chorych w porównaniu do grupy zdrowych).

**Zadanie: Opisz uzyskane wyniki, wskaż geny wyraźnie różnicujące próbki, załącz dwa dowolne wykresy popierające interpretację**

Spośród genów różnicujących próbki, szczególnie wyróżniają się:

1. **Pclaf** (PCNA clamp associated factor) - logFC: 4.58, adj.P.Val: 0.037
2. **C3** (complement component 3) - logFC: 4.02, adj.P.Val: 0.048
3. **Mpeg1** (macrophage expressed gene 1) - logFC: 3.85, adj.P.Val: 0.035
4. **Sfrp2** (secreted frizzled-related protein 2) - logFC: 3.38, adj.P.Val: 0.027
5. **Prrg4** - logFC: 3.29, adj.P.Val: 0.027

Te geny charakteryzują się zarówno dużą wartością logFC, jak i niską wartością skorygowaną p, co sugeruje, że mogą pełnić kluczową rolę w różnicach między analizowanymi próbkami.

**Volcano Plot** (z pliku volplot\_basalpregnant-basallactate.pdf) pokazuje rozkład genów na podstawie ich wartości log2 Fold Change i -log10(p-wartość). Wykres ten wyraźnie wskazuje geny o dużej różnicy ekspresji i niskiej wartości p, co wspiera wnioski z analizy. Widać tu wyraźnie geny takie jak **Tnni2**, **Naaa**, **Wdr43**, **Itprid1**, a także **Sfrp2** – zidentyfikowane jako różnicujące próbki​(volplot\_basalpregnant-b…).