Naujos kartos sekoskaitos (NGS) duomenų analizė

Buvo paimtas pasėlis nuo tualeto rankenos. Bakterijos užaugintos ir išskirta DNR. Naudojant naujos kartos sekoskaitos metodas gauta daug DNR sekos fragmentų (apie 25 tūkst.), kurių kiekvienas yra ~ 150 bp ilgio. Sekos pateikiamos FASTQ formatu. Failas pavadinimu reads\_for\_analysis.fastq.

**1. Apibūdinkite fastq formatą. (https://en.wikipedia.org/wiki/FASTQ\_format). Kokia papildoma informacija pateikiam lyginant su FASTA formatu?**

FASTQ yra tekstinio pagrindo (text-based) formatas, skirtas saugoti biologinėms sekoms. Lyginant su įprastu FASTA formatu, galima pastebėti, jog papildomai yra saugomi ir sekos kokybės įverčiai, užkoduoti ASCII simboliais. Formato sudėtis:

* Pirma eilutė prasideda ‘@’ simboliu, toliau seka sekos identifikatorius ir neprivalomas aprašymas, pvz.: @SEQ\_ID
* Antroje eilutėje yra pati užkoduota biologinė seka
* Trečia eilutė prasideda ‘+’ ženklu, toliau eina neprivalomi tie patys sekos identifikatorius ir aprašymas
* 4 eilutė yra užkoduotos sekos kokybės reikšmės ASCII simboliais. Simbolių skaičius turi sutapti su 2 eilutėje esančios sekos simbolių skaičiumi.

**2. Kurią mėnesio dieną Jūs gimėte? Prie dienos pridėkite 33. Koks ASCII simbolis atitinka šį skaičių?**

Gimiau spalio 5d., 5+33=38. ASCII lentelėje skaičių 38 atitinka simbolis &.

**3. Kodėl pirmi 32 ASCII kodai negali būti naudojami sekos kokybei koduoti?**

Pirmi 32 kodai ASCII lentelėje yra nespausdinami. Jie skirti įvairių signalų apie veiksmus siuntimui į terminalą. Pavyzdžiai: DLE, ESC, BEL, BS, NUL, SOH ir panašiai.

**4. Parašykite skriptą, kuris:**

**a) nustatytų koks kokybės kodavimas yra naudojamas pateiktame faile. Galimos**

**koduotės:**

**i. Sanger Phred+33**

**ii. Solexa Solexa+64**

**iii. Illumina 1.3+ Phred+64**

**iv. Illumina 1.5+ Phred+64**

**v. Illumina 1.8+ Phred+33**

**Parašykite, kokią koduotę nustatėte ir kuo remiantis?**

Pateiktame faile naudojamas Sanger Phred+33 kokybės kodavimas. Ši koduotė buvo nustatyta remiantis „bioinfokit“ biblioteka, analys.format.fq\_qual\_var(file) funkcija, jai grąžinant failo kokybės koduotę. Koduotes galima nustatyti pagal šią lentelę <https://en.wikipedia.org/wiki/FASTQ_format#Encoding>. Galima atskirti ir pagal naudojamus simbolius.

Sanger – Phred+33 (0, 40) ASCII koduotės simboliai.

Solexa – Solexa+64 (-5, 40) ASCII koduotės simboliai.

Illumina 1.3+ – Phred+64 (0, 40) ASCII koduotės simboliai.

Illumina 1.5+ – Phred+64 (3, 41) ASCII koduotės simboliai.

Illumina 1.8+ – Phred+33 (0, 41) ASCII koduotės simboliai.

**b) analizuotų C/G nukleotidų pasiskirstymą read’uose. Pateikite grafiką, kurio y ašyje būtų read’ų skaičius, x ašyje - C/G nukletidų dalis read’o sekoje (100 proc. Reikštų, kad visi simboliai read’o sekoje yra G ir C)**

**Parašykite, koks „stambių“ pikų skaičius yra gautame grafike? (tikrai mažiau nei 6)**

Chart, histogram

Description automatically generated

Iš grafiko matome, jog „stambių“ pikų skaičius yra 3. Šie pikai yra [20;40], [40;60] ir [60;80] GC % rėžiuose.

**c) paimtų po 5 kiekvieno piko viršūnės sekų ir atliktų blast’o paieškas. Naudokite nr/nt duombazę, paiešką apribokite taip, kad ieškotų atitikmenų tik bakterinės sekose (organizmas “bacteria”). Analizei naudokite tik patį pirmą atitikmenį.**

**Pateikite lentelę, kurioje būtų read’o id ir rasto mikroorganizmo rūšis**

Table

Description automatically generated

**5. Kokių rūšių bakterijų buvo mėginyje?**

 Staphylococcus aureus

 Escherichia coli

 Thermus thermophilus