Antras laboratorinis darbas

Tomas Sirutavičius

NGS duomenų analizė

Užduotys

1. Apibūdinkite FASTQ formatą.

FASTQ – standartizuotas FASTA failo formatas, kuris dažniau naudojamas trumpoms sekoms aprašyti. Formato pradžioje pateikiama sekos, užkoduota ASCII formatu, skirta apibūdinti pateiktos sekos kokybei (tikslumą).

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

   13 + 33 = 46

   ASCII simbolis būtų taškas ‚.' .

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

Pirmi 32 ASCII kodai negali būti naudojami sekos kokybei koduoti, nes šie kodai užima daugiau nei 1 bit'ą todėl neatitinka FASTAQ kokybės kodavimo taisyklių. Taip pat, tai yra simboliai, skirti apibūdinti tam tikrą funkcinę klaviatūros mygtukų įvesti.

4. Parašykite skriptą, kuris:

   a) Nustatytu koks kokybės kodavimas yra naudojamas pateiktame faile. Parašykite, kokią koduotę nustatėte ir kuo remiantis?

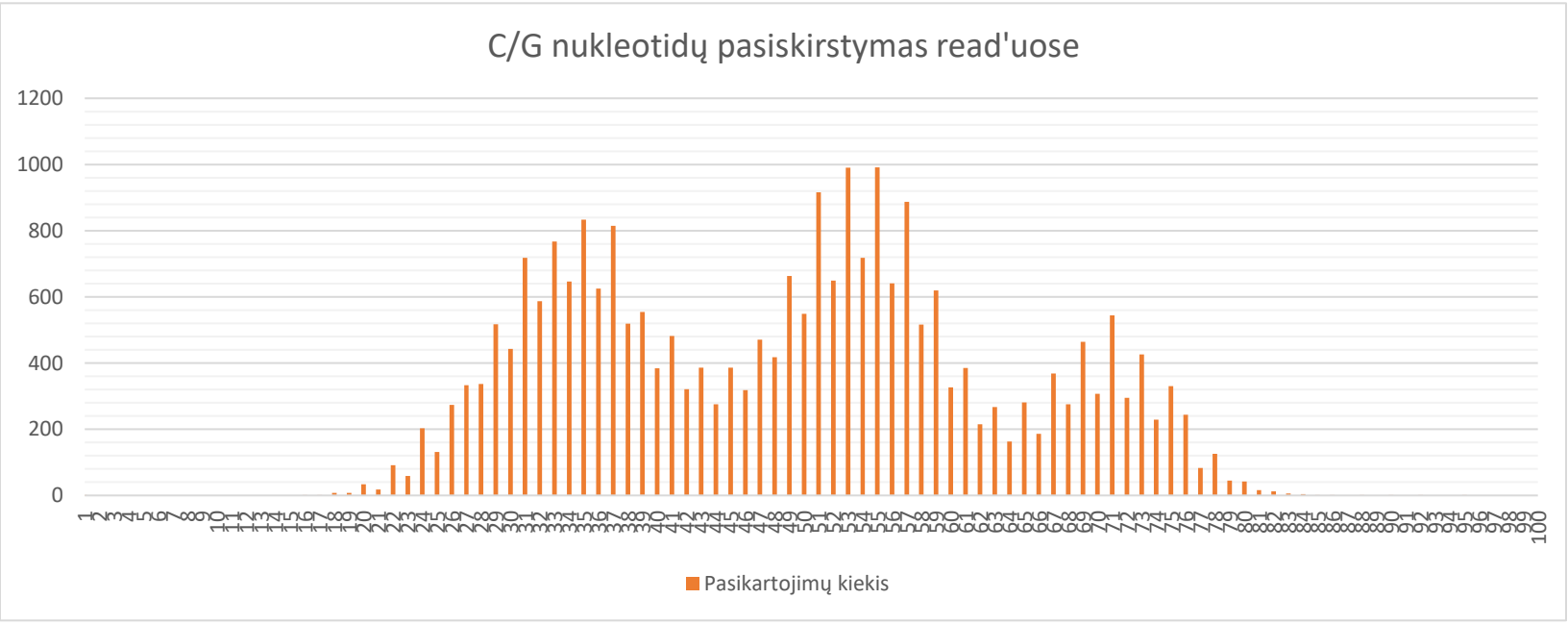
      Failo kokybės kodavimas: Sanger Phred+33
      Nustatytas naudojant „bioinfokit" bilbioteką. Pasinaudojus „analys.format.fq_qual_var(file)" funkciją, kuri gražina failo kokybės kodavimo fomratą.

      Koduotes taip pat galima atskirti 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



Nukleotidų sekų pasiskirstyme matome 3 stambius pikus.

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

Programa suranda kiekvieno piko 5 sekas ir atliko online BLAST paieškas. Apačioje pateikta lentelė, su sekų ID ir paieškos rezultatais. Iš paieškų rezultatų imta, kaip nurodyta užduotyje, pats pirmas rezultatas, tačiau sutampantis su kitais rezultatas būdavo pateikiamas antras su tokiu pačiu atitikimo procentu.

ID	Mikroorganizmo rūšis
M00827:12:000000000-AEUNW:1:1101:12898:3746	Staphylococcus aureus subsp. aureus NCTC 8325
M00827:12:000000000-AEUNW:1:1101:15734:4405	Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:14559:5316	Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:10754:8831	Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:21560:11525	Staphylococcus argenteus strain BN75 chromosome
M00827:12:000000000-AEUNW:1:1101:14568:2958	Escherichia coli str. K-12 substr. MG1655, complete genome
M00827:12:000000000-AEUNW:1:1101:11742:3950	Shigella sonnei strain ECH+12 133-HLP106_NODE_14.ctg_1, whole genome shotgun sequence
M00827:12:000000000-AEUNW:1:1101:9837:4027	Escherichia coli str. K-12 substr. MG1655, complete genome
M00827:12:000000000-AEUNW:1:1101:12227:4848	Escherichia coli str. K-12 substr. MG1655, complete genome
M00827:12:000000000-AEUNW:1:1101:18054:6383	Escherichia coli str. K-12 substr. MG1655, complete genome
M00827:12:000000000-AEUNW:1:1101:18070:3392	Thermus thermophilus HB8 chromosome 1, complete sequence
M00827:12:000000000-AEUNW:1:1101:23350:4251	Thermus thermophilus HB8 chromosome 1, complete sequence
M00827:12:000000000-AEUNW:1:1101:23294:5998	Thermus thermophilus HB8 chromosome 1, complete sequence
M00827:12:000000000-AEUNW:1:1101:12169:8149	Thermus thermophilus HB8 chromosome 1, complete sequence
M00827:12:000000000-AEUNW:1:1101:7922:8647	Thermus thermophilus HB8 chromosome 1, complete sequence

5. Kokių rūšių buvo mėginyje?

Staphylococcus aureus – auksinis stefilokokas (odos ligas sukelianti bakterija)

E.coli – E.coli (žarnyno ligas sukelianti bakterija)

Thermus thermophilus - Thermus thermophilus (karštose versmėse dažnai randama bakterija)