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 padėtas čia:

https://drive.google.com/file/d/0ByNQBUuv7woVNUNqNW83SldmUTQ/view?usp=sharing

Užduotys:

1. Apibūdinkite fastq formatą. (https://en.wikipedia.org/wiki/FASTQ_format). Kokia papildoma informacija pateikiam lyginant su FASTA formatu?
2. Kurią mėnesio dieną Jūs gimėte? Prie dienos pridėkite 33. Koks ASCII simbolis atitinka šį skaičių?

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

4. Parašykite skriptą, kuris:
   a) nustatytu 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?

   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)

   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

5. Kokių rūšių bakterijų buvo mėginyje?