1. Izvršiti kontrolu kvaliteta FASTQ fajlova alatom FastQC. Priložiti izvrštaj i diskutovati rezultate.

   > Paired end 1: Per base sequence content je kriterijum koji je označen kao problematičan i on označava propociju svake baze svih readova. U najboljem slučaju, procenat svih baza bi trebao da bude 25%, ali ovde se taj odnos menja od pozicije 90.

   > Paired end 2: Per base sequence quality je kriterijum koji je označen kao problematičan i on označava kvalitet baza na svim pozicijama svih readova. To je zato što bazni parovi od pozicije 160 imaju veću devijaciju po pitanju kvaliteta.

2. Mapirati sekvencirane readove na referentni genom hg38 upotrebom alata BWA.

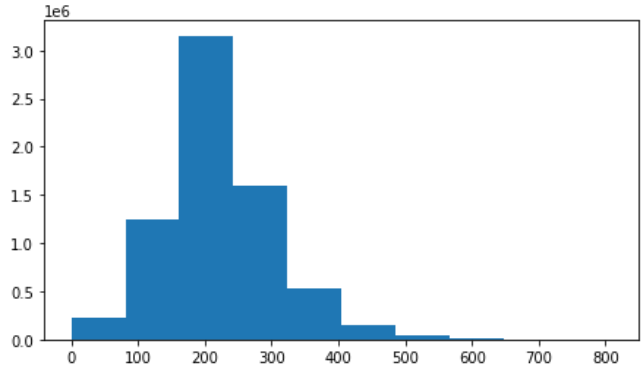
   2a. Koliko je readova uspešno mapirano?

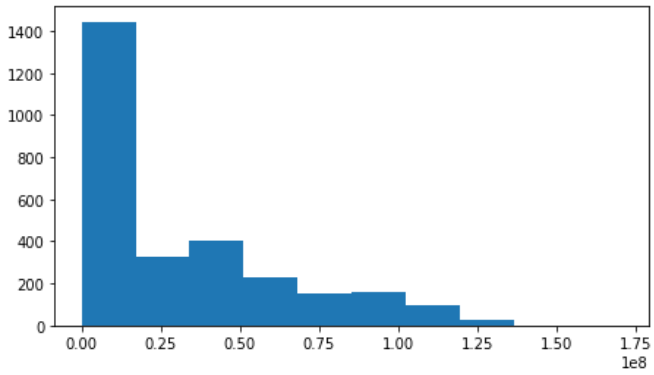
   > Mapiranih readova: 6755505

   2b. Koliko je parova readova mapirano tako da su oba para mapirana?

   > Mapiranih parova: 6735233

   2c. Nacrtati histogram dužina sekvenciranih fragmenata.





3. Izvršiti obradu dobijenog BAM fajla prema GATK protokolu (markiranje duplikata, rekalibracija kvaliteta baza). Koliki su procenati PCR i optičkih duplikata?

   > Ukupno duplikata: 480157

   > PCR duplikati: 100 * 480157 / 6775767 = 7.09%

   > Optički duplikati: 100 * 0 / 6775767 = 0%

4. Identifikovati mutacije upotrebom alata Haplotype Caller i filtirtati mutacije predefinisanim filterima prema Broad preporukama.

   4a. Koliko je ukupno mutacija identifikovano, koliko od njih su SNP, a koliko INDEL?

   > SNP: 14577

   > INDEL: 1796

   4b. Koliko mutacija prolazi, a koliko ne prolazi kriterijume filtriranja.

   > Prolazi: 16368

   > Ne prolazi: 5

   4c. Izračunati TiTv odnos pre i posle filtriranja.

   > Pre: 1.9288728149487644

   > Posle: 1.9294472361809045

5. Anotirati mutacije alatom Funcotator. Izbrojati različite vrednosti ClinVar značajnosti.

   > Benign likely benign: 45

   > Benign: 247

   > Not provided: 11

   > Likely benign: 28

   > Benign other: 1

   > Likely pathogenic: 1

   > Pathogenic: 1

   > Association: 1

   > Uncertain significance: 3

   > Conflicting interpretations of pathogenicity affects: 1

   > Conflicting interpretations of pathogenicity: 2

   > Risk factor: 2

   > Benign likely benign association: 1

6. Svi uzorici sadrže određenu količinu kontaminacije DNK materijalom bakterijskog ili virusnog porekla. Većina ovakvih readova se neće mapirati na ljudski genom. Izvući readove koji nisu mapirani u procesu mapiranja, asemblovati ih alatom abyss, i identifikovati organizam od kojeg potiče najduži skafold upotrebom alata Blast.

   > Nemapirana sekvenca potiče od bakterije Bradyrhizobium Japonicum