Onboarding Q&A 4/5

1. How many DNA-seq secondary analysis pipelines do we support?

    a. 1

    b. 2

    c. 3

2. These pipelines are _____, _____ and _____.

    a. Whole Genome Seq

    b. Intron Seq

    c. Whole Exome Seq

    d. Targeted Seq

3. Coverage is a term used to describe.....

    a. The average number of reads that cover a locus

    b. The number of reads that cover the whole genome

    c. The average number of reads aligned to a chromosome

4. Whole genome analysis is used to discover novel variants.

    a. True

    b. False

5. Whole exome analysis is used to detect coding variants.

    a. True

    b. False

6. Targeted seq analysis is used to detect wild type non-coding variants.

    a. True

    b. False

7. We use targeted seq to analyze variants in a small set of _____. Typically _____.

    a. gene(s), 1

    b. introns, 100

    c. exons, 1000

8. Targeted seq library prep facilitates ~_____X coverage of the targeted region.

a. 10

b. 100

c. 1000

9. Whole genome library prep entails the lowest coverage of the three DNA-seq analyses.

a. True

b. False

10. What are the relative coverages of WGS, WES and TRS?

a. 10, 100, 1000

b. ~30, ~100, ~1000

c. 10, 10, 10

11. We use the GATK to perform alignment post-processing, variant calling and post-variant processing.

a. True

b. False

12. The BWA- aligner produces BAM files.

a. True

b. False

13. The BWA-MEM wrapped version (version on SBPLA) of the aligner produces BAM files.

a. True

b. False

14. What three pre-variant calling/post-alignment steps must we perform.

a. Variant Quality Score Recalibration

b. Indel Realignment

c. De-duplication of duplicate reads

d. Base quality score recalibration

15. INDEL realignment is performed first.

a. True (on presentation slides it is first but de-duplication of duplicate reads might be performed first)

b. False

16. Realignment around INDELs is performed because....

a. INDEL may result in a series of mismatches which affect the variant calling process

b. INDEL are easy to realign

c. We can recalibrate bases if INDELs are not properly aligned

17. Base Recalibration is performed because....

a. Sequencers can produce overconfident base quality scores

b. Base quality scores are bad for business

c. The Phred scale is not as precise as it should be during sequencing

18. De-duplication of reads is performed because.....

a. duplicate reads mess up coverage of loci and thus variant calling

b. duplicate reads take up space

c. duplicate reads are ambiguous and must be eradicated

19. If we want to view FASTQ read statistics we must pass our FASTQ file through which of the following tools?

a. FASTQmcf

b. FASTQC

c. FASTQH

20. FASTQ read quality scores are expected to decrease base by base.

a. True

b. False

21. What is the difference between empirical quality scores and reported quality scores?

a. Empirical scores are based on the read match/mismatch scores, reported scores are based on the base quality scores of individual bases of a read

b. Empirical quality scores are mapping scores while reported scores are base call scores.

c. There is no difference

22. How many possible covariates are there for base quality score recalibration?

a. 3

b. 4

c. 5

23. We recalibrate our base quality scores in order to reduces the qualities of overconfident base calls.

a. True

b. False

24. We use databases such as dbSNP to detect and recalibrate base quality score errors.

a. True

b. False

25. Why is it important to do INDEL realignment first before base quality score recalibration?

a. base quality scoring takes mismatches into context, realignment may resolve previously detected mismatches

b. context of mismatching bases covering a locus may affect the base quality empirical scoring.

c. both a and b make sense

26. Systematic errors during iLLUMINA sequencing where non-fluorescent nucleotides are incorporated in the growing complementary DNA chain may be reported as SNPs or DELETIONS.

a. True

b. False

27. Systematic errors during iLLUMINA sequencing where several non-fluorescent nucleotides followed by a fluorescent nucleotide is a repetition of question 26.

a. True

b. False

28. Systematic errors during iLLUMINA sequencing where terminating groups seize to detach themselves from labeled nucleotides result in DELETIONS.

a. True

b. False

29. Two categories of variant callers are _____ and_____.

a. General

b. Somatic

c. Specialized

30. General callers are typically SNV/INDEL callers.

a. True

b. False

31. Two callers that are commonly used on the SBPLA are GATK _____ and GATK _____.

a. FreeBayes

b. HaplotypeCaller

c. UnifiedGenotyper

32. Humans are diploid organisms. This means that we have two homologous sets of whole genomic material in each of our cells.

a. True (except gametes cells, which are haploid)

b. False

33. A human genotype has two _____.

a. variants

b. types

c. alleles

34. Joint callers call variants simultaneously for many samples.

a. True

b. False

35. GATK _____ can perform joint calling.

a. FreeBayes

b. HaplotypeCaller

c. UnifiedGenotyper

36. Variant callers perform variant calling and _____ .

a. allele calling

b. genotyping

c. phasing

37. GATK HaplotypeCaller is license free as opposed to UnifiedGenotyper.

a. True

b. False

38. HaplotypeCaller and UnifiedGenotyper use Bayes Theorem to infer plausible variants.

a. True

b. False

39. The known set of variants used in Bayes Theorem can be found in database file such as _____.

a. dbSNP

b. SEQ

c. Mills

40. Variant callers produce _____ files.

a. SAM

b. FASTQ

c. VCF

41. VCF files have 7 important columns. What columns are these?

a. chromosome, position, id-name in database (if variant is known), reference value, alternate value(s), quality score and genotype

b. alt, ref, flags, phase, mapping quality, base/variant quality, ti/tv ratio

c. chromosome, position, id-name, alternate value, pass filter, quality score and genotype

42. VCF stands for _____.

a. Variant Coding Format

b. Variant Codon Format

c. Variant Call Format

43. What is a haplotype?

a. A group of two chromosomes exchanging information, and the exchanged information forms a haplotype

b. A group of proximally close loci are physically inherited together (on the same copy of a chromosome) frequently in a population

c. A group of distant loci

44. If two variants are in phase, this means that they are located on the same _____.

a. Sequence

b. Marker

c. Flanking sequence

45. What does the following genotype in a VCF record mean: 0/2 ?

a. Heterozygous reference genotype for the second alternative allele

b. Homozygous reference genotype for the second alternative allele

c. Heterozygous alternative genotype for the zeroth allele