Exam 2 (100 pts)

Name:_____

1. Which of following is the best estimate of the size of the E. coli genome?
    a. 4 base pairs
    b. ~4 Kbp (4,000 base pairs)
    c. ~4 Mbp (4,000,000 base pairs)
    d. ~4 Gbp (4,000,000,000 base pairs)

2. Which of the following is the best estimate of the size of the mouse genome (not discussed in class) given what you know about mammalian genomes?
    a. ~3 Kbp (3,000 base pairs)
    b. ~3 Mbp (3,000,000 base pairs)
    c. ~3 Gbp (3,000,000,000 base pairs)
    d. ~30 Gbp (30,000,000,000 base pairs)

3. DNA sequences from a high-throughput sequencing experiment can be assembled into longer sequences representing continuous regions of the genome called?
    a. Contigs
    b. Reads
    c. Paired-end sequences
    d. Scaffolds

4. Analysis of high-throughput sequencing data is a common application of bioinformatics and computational biology.
    a. True
    b. False

5. Despite tremendous advances in high-throughput sequencing, most genome sequencing projects still use traditional sequencing methods, similar to what we used at the beginning of class, rather than Next-gen methods because of cost.
    a. True
    b. False

6. Illumina sequencing is done on which of the following surfaces?
    a. 96 well plate
    b. Beads
    c. Microarray
    d. Flow cell

7. Which of the following techniques is a component of Illumina sequencing:
    a. Sequencing by synthesis
    b. Bridge amplification and cluster generation
    c. Single or paired end sequencing
    d. All of the above

8. What is a common drawback of Illumina sequencing when compared to other methods discussed in class?
    a. Cost
    b. Time
    c. Throughput
    d. Read length
    e. All of the above

9. Which of the following is a typical size range for Illumna sequencing reads (assume single end reads)?
    a. <50 nt
    b. 50-300 nt
    c. 300-500 nt
    d. >1,000 nt

10. Which of the following are two additional sequencing platforms discussed in class?
    a. MilliQ and NEB
    b. Gnu and LibTech
    c. Nanopore and PacBio
    d. McDonalds and Burger King

11. Short answer: what is the difference between single end and paired end sequencing?

    In single end sequencing, sequencing is done from one end of the DNA fragments.
    In paired end sequencing, sequencing is done from both ends of the DNA fragments

12. Which of the following Phred scores would be used as a cutoff to discard reads from a Next-gen sequencing run that have a probability of >1 error per 100 bases?
    a. 10
    b. 20
    c. 30
    d. 40
    e. 50

13. You're interested in sequencing a mutant strain of C. elegans to identify the causal mutation for a particular phenotype. You determine that you need 30X coverage of the genome and the genome is 100 million base pairs. If your read length is 100 bases, how many reads would you need to sequence the genome at 30X coverage?

    Summary
    Coverage: 30X
    Read length: 100 bases
    Genome: 100,000,000 basepairs

    reads = (30 x 100,000,000)/100 => 30,000,000

14. Suppose you are guaranteed at least 200 million reads from each lane of a flowcell and you need 20 million reads for each of your samples. How many samples can you multiplex into one lane?
    a. 1
    b. 2
    c. 5
    d. 10
    e. 20

15. Which of the following was the first multicellular eukaryote to have its genome sequenced?
    a. E. coli
    b. Drosophila (fruit fly)
    c. C. elegans (worm)
    d. M. musculus (mouse)
    e. H. sapiens (human)

16. In what year was a draft genome of the human genome first published?
    a. 1901
    b. 1961
    c. 1991
    d. 2001
    e. 2011

17. Short answer: What common file format for high-throughput sequencing data is shown below:

    fastq

```
@D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
+
CCCFFFFFHHHHHJIJGHJJJJIJJJJJGGGFFFFEABDHHHFHFF@@DD>
@D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJGFHIJJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
@D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJHIJJJJJGIIJFGIJJJ
```

18. What information is contained in line 2 of the file shown in question 17?
    a. The read ID.
    b. The sequence of the read.
    c. The quality score for each base in the read.
    d. The average quality score across the entire read.

19. How many lines of the file shown in question 17, correspond to a single read?
    a. 1
    b. 2
    c. 3
    d. 4

20. Short answer: what software was used in class to de novo assemble the E. coli genome?

    Velvet

21. Short answer: what software was used in class to map reads to a reference genome sequence?

    Bowtie2 (or Bowtie)

22. Short answer: what GUI software did was used in class to assess library quality?

    FastQC

23. Short answer: what GUI software was used in class to visualize our read mapping data?

    IGV

24. Short answer: what is one critical piece of information contained in a SAM file that is not contained in a raw high-throughput sequencing data file?

    Positional information

25. What is an N50 value in reference to genome assembly?
    a. The contig length at which 50% of the assembled genome is in contigs of N50 or greater.
    b. The percentage of the genome covered by contigs.
    c. The average contig size.
    d. The maximum contig size.

26. Which of the following files are compressed, assuming they are named correctly?
    a. genome.fa
    b. genome.txt.gz
    c. genome.txt.gz.tar
    d. All of the above
    e. B and C

27. Short answer: what is the name of a nuclease commonly used for CRISPR-mediated genome editing discussed in class?

    Cas9

28. Short answer: In addition to a nuclease, what other component is essential for CRISPR-mediated genome editing, assuming short semi-random mutations are acceptable?

    gRNA

29. Short answer: What endogenous mechanism is exploited to introduce a specific mutation or additional sequence into the genome, such as GFP, during CRISPR-mediated genome editing?
    Homologous recombination (or homology-directed repair)

30. Which of the following is an endogenous role of CRISPR?
    a. Adaptive immunity in some archaea and bacteria.
    b. Regulation of endogenous genes in some archaea and bacteria.
    c. Adaptive immunity in some plants and animals.
    d. Regulation of endogenous genes in some plants and animals.

31. Comparative genomics typically involves the comparison of homologous DNA sequences across different species.
    a. True
    b. False

32. Short answer: what is functional genomics?

    The application of genomic data to study gene and protein expression and function.

33. Short answer: what are genome-wide association studies?

    Examination of sequence variants across individuals in an effort to identify genetic causes of a particular trait or disease.