Problem set 2

In this problem set, we will explore applications of DNA sequencing techniques and analyzed some of the data they produced.

From here on out, there can be multiple correct answers because different techniques can achieve the same result and the tradeoff depend on your viewpoint and circumstance. Please provide clear rationale supporting your answers.

DNA sequencing and applications

Among various DNA sequencing technologies that we learned in class (454 Pyrosequencing, Ion Torrent, Illumina, PacBio SMRT-seq, and Nanopore), which platform would you choose to investigate the following research goal, and why?

Hint: I encourage you to approach these problems in two ways: first, come up with an answer using knowledge from lectures and second, try searching for research articles that investigate these questions and identify the platforms used.

Q1: Exome sequencing to identify rare germline mutations in a child.

Q2: 16S rRNA sequencing the V3-V4 hypervariable region to determine the community structure of bacteria in soil.

Q3: Whole genome sequencing of all pathogens in urine sample from a patient with unknown infection.

Q4: Whole genome sequencing of new Thai bat species.

Q5: RNA sequencing to quantify gene expression levels of a cell culture.

Q6: RNA sequencing to discover novel isoforms of coding and non-coding transcripts in a cancer tissue.

Q7: Exome sequencing of a selected 200-gene panel that are relevant in breast cancer to characterize mutation profiles in depth.

Q8: Fast identification of pathogens in patient's blood to enable quick medical decision making.

Processing of DNA sequencing data

Explain the data contained within each file format below. Provide an example of the formatting together with your explanation.

File Format	Content
Q9: FASTA	
Q10: FASTQ	
Q11: SAM	
Q12: BAM	
Q13: VCF	
Q14: MAF	*This one was not mentioned in class, but it is highly related*.

In class, we learned that the "deduplication step" was performed when processing DNA sequencing data.

Q15: Explain what this step does and why it should be performed.

Q16: Do we need to perform deduplication on sequencing data from 3rd generation technique, such as SMRT-seq and Nanopore? Why?

Variant calling

Q17: What are the differences between germline and somatic mutations?

Q18: How would you design an experiment to determine germline mutations of an individual?

Q19: How would you design an experiment to determine somatic mutations in a tumor tissue of an individual?

Let's say you identified two mutations in BRCA1 gene: the first one alters the nucleotide position 5449 from G to T, and the second one alters the nucleotide position 2311 from T to C.

Q20: What are possible functional and clinical impacts of these mutations? *Hint: You need to look up information from online databases. Provide the sources for your answer.*

Applications of sequence alignment

Q21: Compare the similarities and differences between MEGABLAST and BLASTN algorithms. Explain how they yield alignment hits with different level of similarities.

```
Query 213 GCTACCTCAGTTttcttcttcttctcttccttcagattcttccagttctctGCGCTGTCA  272
          |||||  ||  ||||||||  |||||  ||||||||  ||  ||||||||||   |||||  ||  |||
Sbjct 650 GCTACTTCGGTTTTCTTGTTCTTGTCTTCCTTTAGTTTCTTCCAGTCGTCTGCACTATCA  591
```

Q22: In the above alignment result, there are some nucleotides in **lower case letters**. What are they? How did they impact on the scoring of the alignment?

Q23: Perform dynamic programming to determine the best **global alignment** for ATAGC and ATGCAGC under the following scoring scheme: Match +1, Mismatch –1, Gap –2.

Q24: What the best **local alignment** for the sequences in **Q23** would be? You may perform dynamic programming or simply provide some reasonings to support your answer.

You discovered the following transcript from an RNA-sequencing of an indigenous Thai plant:

>CL2742.contig2
GGCAGATACAGTCATAAAAGAAGTAACCAGAATCAGAAAGAATAAAGGTTGATTTG
GATTTAAGCCGACAGAACGATATAGCTACAATACAAAATTAACTTACAGTACAACA
GTATGATTTTAAATTTGCTGATAAACGAAAGGCCATTAACTTGAGAAACTACTAACT
CTTACATTGTTTTTTATCCTTTTAAGAAGCAGCAGTAGTAGTGGCTACCTCAGTTTTC
TTCTTCTTCTCTTCCTTCAGATTCTTCCAGTTCTCTGCGCTGTCAATAATTTCGCCAGT
TTTTATGAGGTAACGAACTCGTTTCCCATTGTCGAGGACTTTGTGACCCACCCGGCT
AACTACATCCTTTTCTTTAGAGTAGAGCATCACATTTGAACTATGAATAGGCGCTTC
GATCTTGATAATCTGGCCAGGTTCTCCTTCTTCTCTGCTCTTCATATGCTTTGTCTTCA
AATTTATTTCACCGACTACCACAGTGCTGTTGTGCTTAAAGATTTTTGTAACTTCACC
GACTTTCCCTTTATCATCTCCAGCTATCACTTTAACTGTGTCTCCTACTTTAACATGC
ATTTTATGTAAAACTGGAAGGCTGTTTGGTTTACATTCCTTCCGCTCCCACCGCTTAA
TCTGCAAGGGGGGAAGAATTGAGTAAATGGAATATAGGTTCATCAGACCCAAAATG
ACTTCAAA

Q25: Which BLAST algorithms would be appropriate for annotating the function of this transcript? Why? *Try to come up with more than one answer!*

Perform a BLAST search using one of the algorithms you picked in **Q25** and answer the following questions.

Q26: How similar is this transcript with the top two hits from your BLAST result? Provide multiple lines of evidence.

Q27: What do you think is the likely function of this transcript?

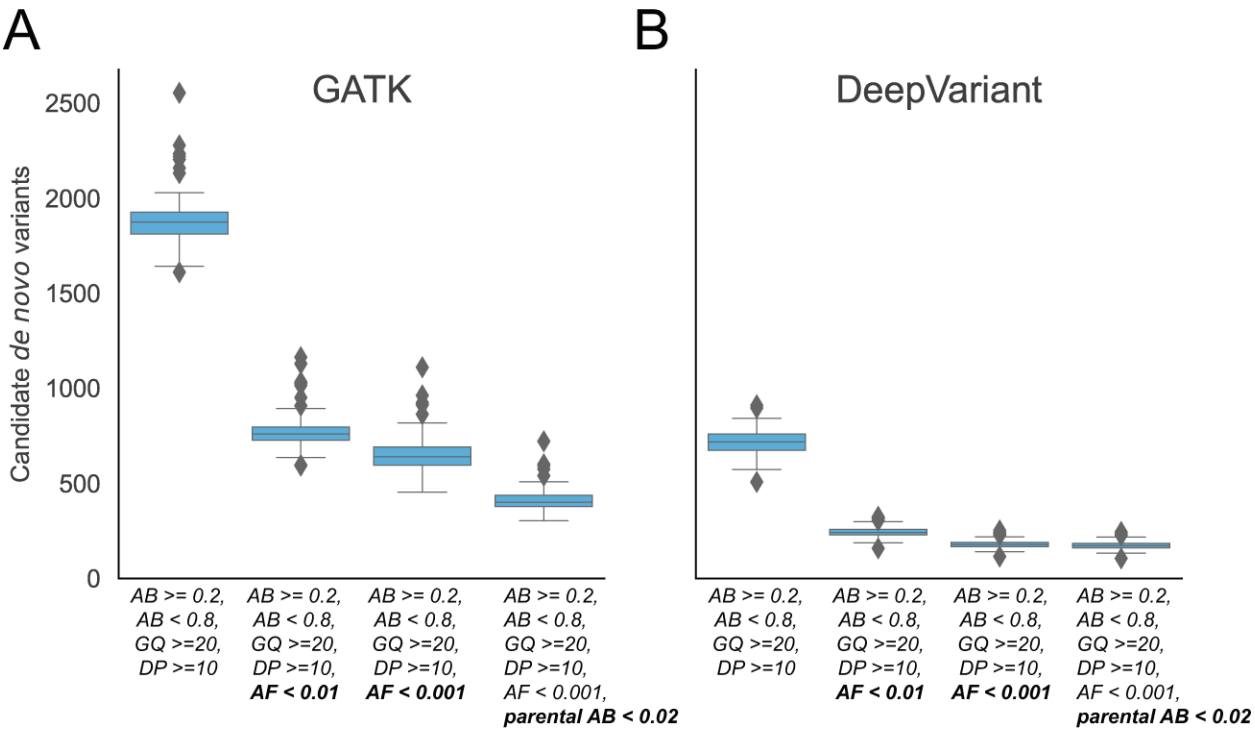
Q28: What is the likely taxonomic group for this plant based on the top BLAST hits?

Q29: What is the likely **open reading frame (ORF)** for this transcript? Provide the translated amino acid sequence from this **open reading frame**. *Hint: You can try translating every frame of this transcript. Alternatively, think about how the right choice of BLAST algorithm can help you.*

Variant filtering (Bonus problem)

Calling mutations is a challenging problem because not all differences in DNA sequence are biological. There are many research papers on the topic of variant filtering.

In Figure 4 from Pedersen, B.S. *et al*. Genomic Medicine 6:60, 2021 (https://www.nature.com/articles/s41525-021-00227-3) shown below, the authors examined the impact of some variant filter parameters, including **AB**, **GQ**, **DP**, and **AF**.



Q1*: Explain what these filter parameters are.

Q2*: Explain the benefits of applying each of these parameter filters. In other words, explain why you would want to apply them on your data?