Problem set 1

This problem set covers the content from week 1: DNA sequencing platform, DNA sequencing applications, and basic processing of DNA sequencing data.

Tips and rules:

- You can answer in English or in Thai.
- There can be more than one correct answer. What I am looking from you is not just the correct answer but the rationale for your answer.
- Please provide evidence of how you think and what sources of information you used.
- AI such as ChatGPT may be used. You can also work together with friends. But you must write the answer in your own words.
- Any incidence of plagiarism and copying of another student's work will be reported to the Graduate Affairs.

DNA sequencing and applications

Among the three major DNA sequencing platforms that we learned in class (Illumina, PacBio SMRT-seq, and Nanopore), which one(s) would you choose for each of the following research goal, and why?

If you think multiple platforms can be used, please designate the best one in your opinion.

If you think a combination of platforms should be used (such as the combination of short-read and long-read data), please indicate so.

Hint: You are encouraged to first came up with an answer using knowledge from the class and then search for papers on these topics to confirm your ideas.

Scenario	Illumina	SMRT-seq	Nanopore
Q1: Exome sequencing to identify rare germline mutations in a child	Yes/No? Why?		
Q2: Sequencing of the V3-V4 hypervariable region of 16S rRNA from patient feces sample			
Q3: Whole genome sequencing of a *K. pneumoniae* isolate from a patient			
Q4: Whole genome sequencing of an unknown Thai bat			

Q5: Transcriptomics profiling of a human cell culture using RNA-seq			
Q6: Searching for novel RNA isoforms in cancer with RNA-seq			
Q7: Targeted sequencing of a cancer gene panel (100 genes) to characterize mutations in depth			
Q8: Fast identification of pathogens in patient's blood to enable medical decision making within 2 hours			

File formats related to DNA sequencing data

For each file format, explain what information are contained in them. Also, for files that have a human-readable format, please provide an example.

Some may not be mentioned in class, but I would like you to try figuring out the answer by searching online.

File	Content / Format
Q9: FASTA	
Q10: FASTQ	
Q11: SAM	
Q12: BAM	
Q13: VCF	
Q14: MAF (Mutation Annotation Format)	
Q15: BED (Browser Extensible Data)	

Basic processing of DNA sequencing data

The following questions concern the “deduplication” of sequencing reads.

Q16: What factors caused duplicated reads in DNA sequencing data?

Q17: What are the benefits of performing deduplication? What could happen if you don’t?

Q18: Is deduplication essential for sequencing data from 3rd generation platforms, SMRT-seq and Nanopore? Why or why not?

The following questions concern “variant calling”.

Q19: What are the key differences between germline mutations and somatic mutations? How do the differences affect the way that we use DNA sequencing to identify them?

Q20: Propose an experimental design to identify germline mutations of a child with a rare genetic disease.

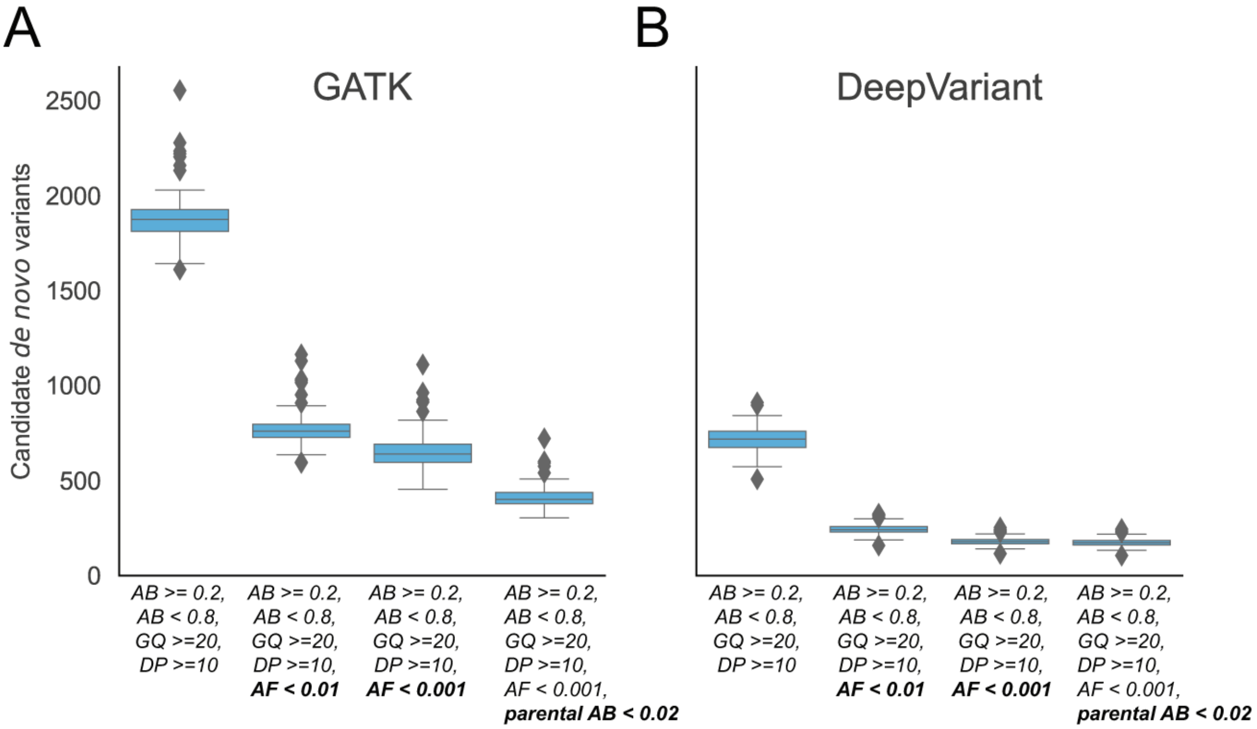
Q21: Propose an experimental design to identify somatic mutations in a liver tumor tissue.

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.

Q22: 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.*

Filtering of variants (Bonus problem)

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 filtering 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.