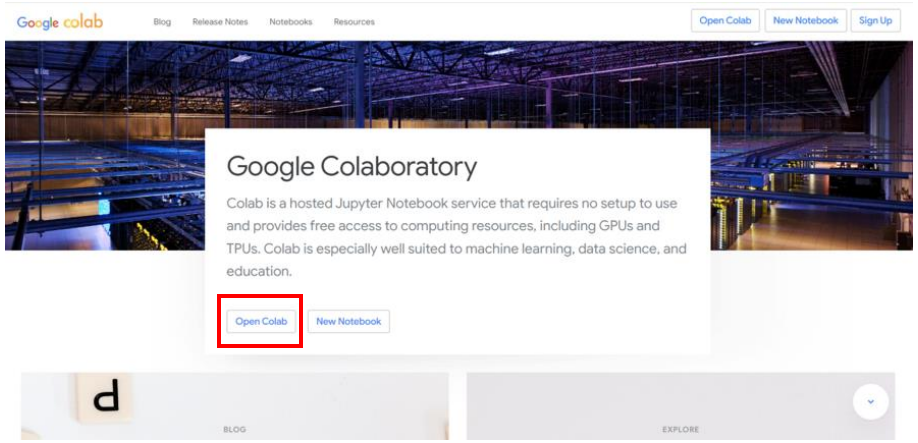
BEHI 5003
Tutorial
DNA sequencing data processing and analysis

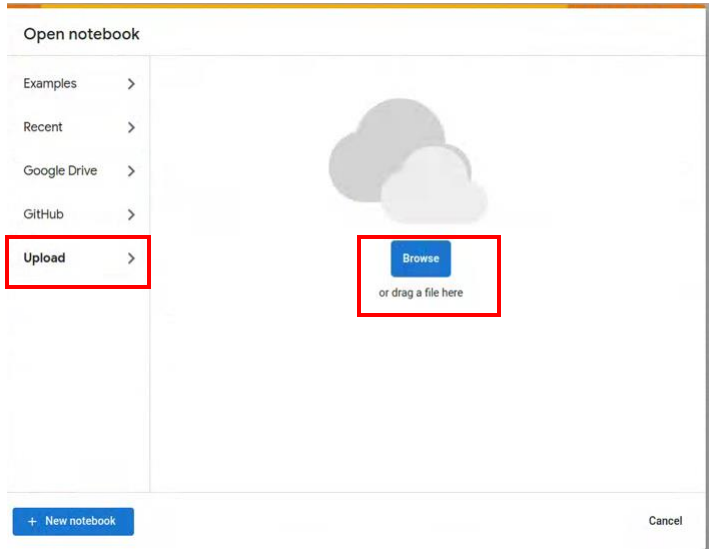
Won Joon Kim

wjkimab@connect.ust.hk

Let's set up Google Colab!

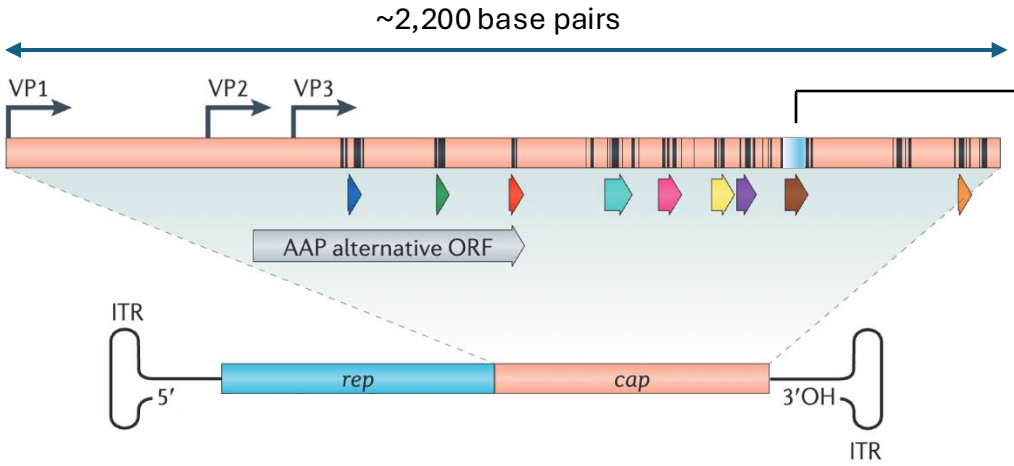
1. Log in to google.com with your (newly created) account.
2. Visit https://colab.google.com/, and click **“Open Colab”**.
3. Click **‘Upload’** and upload the sequencing data analysis demo on Canvas (*aav_illumina_seq_demo.ipynb*).

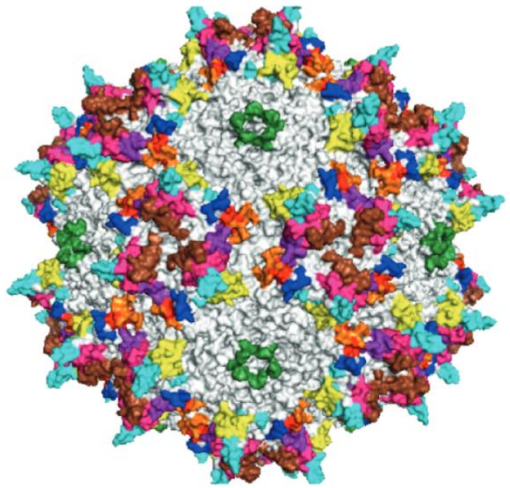




You are NOT running any code on your own computer, so rest assured!

Background to AAV 7-mer insertion libraries





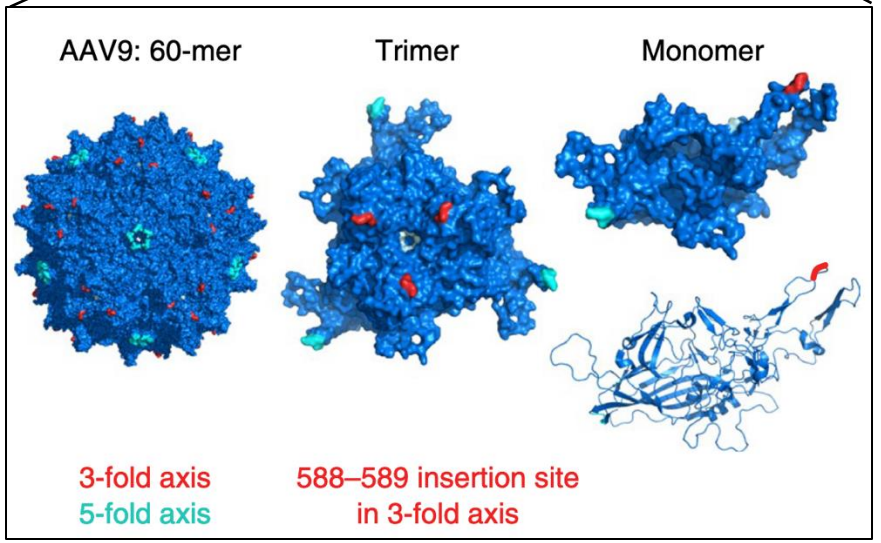
21-base pair insertion

e.g.,
CAGGCCGTGAGGACCAGCCTG

Translate

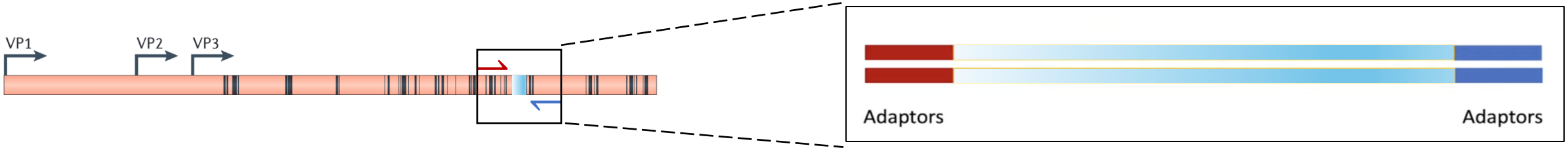
7-mer insertion site

Wild-type SAQ A
SAQQAVRTSLA



Kotterman, M. A. & Schaffer, D. V. *Nat Rev Genet* **15**, 445–451 (2014).; Chan, K. Y. *et al. Nat Neurosci* **20**, 1172–1179 (2017).; Kumar, S. R. *et al. Nat Methods* **17**, 541–550 (2020).

AAV 7-mer library NGS sample preparation



ADAPTOR LENGTHS
NOT TO SCALE

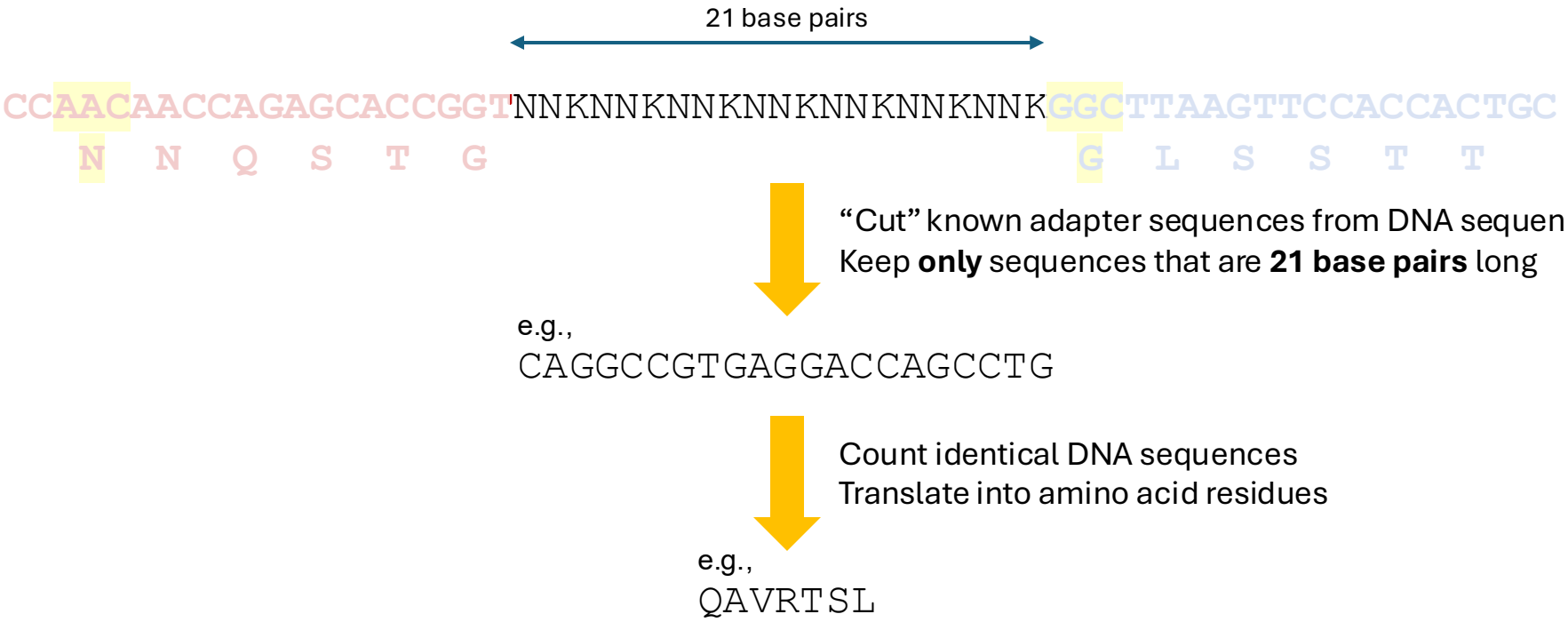
Assumption for today's task:

→ sequence: `CCAACAACCAGAGCACCGGT`

← sequence: `GGCTTAAGTTCCACCACTGC`

Slide adapted from Prof. Zhu's Module 1: Nucleic acid, sequencing, and gene editing

Identifying amino acid-level variants



Example taken from Colab output:

	sequence	count	length	amino_acid
0	AATTATTCTTGGAATTATAAG	1	21	NYSWNYK

Removing “low-quality” sequences

But, what if there are STOP codons (i.e., TAA, TAG, TGA) and/or undetermined bases (N)?

e.g.,
CAGGNCGTGAGGACCTAACTG

Translate

e.g.,
QXVRT*L

Set a “condition” to remove amino acid sequences with X and/or *

Keep the "high-quality" amino acid sequences for downstream analysis

Example taken from Colab output:

278 filtered amino acid variants saved.

	amino_acid	count	occurrence
16	AANFSFL	1	1

Questions (to be answered in Colab)

1. What can you say about the library length?
2. What can you say about the sequencing quality?
3. What are the adapter sequences and library length?
4. What is the desired output FASTQ file, and how do the first 4 lines look?
5. How many reads are left after keeping only the desired lengths?
6. How many valid amino-acid level variants are there in this library?
7. What can you say about this library's amino acid composition?

Thank you for your efforts!

Any questions or feedback would be greatly appreciated!

wjkimab@connect.ust.hk