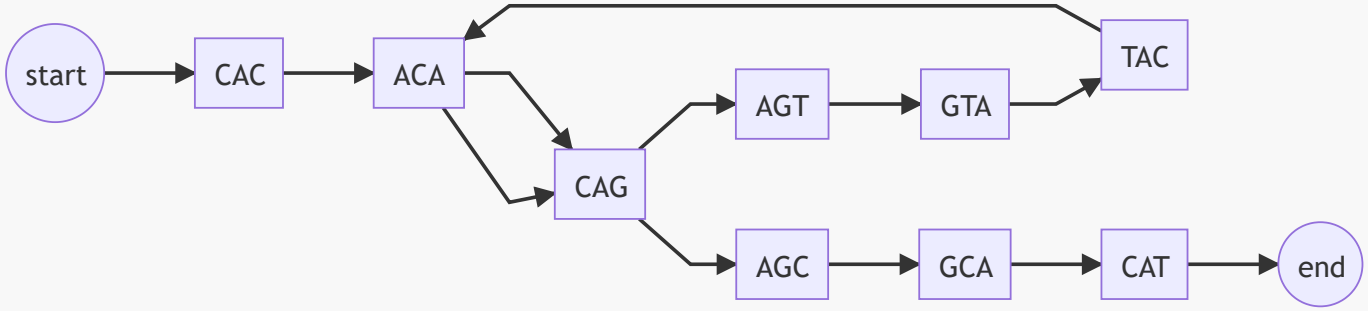
BIOL8706: Dividing and conquering sequence alignment using De Bruijn Graphs



- Student: Richard Morris
- Huttley lab, Australian National University
- Supervisors: Gavin Huttley, Vijini Mallawaarachchi

Australian National University

Introduction to sequence alignment

Given we can sequence genomes of different organisms.

Sequence A: `ATGCATAC` Sequence B: `ATGTAC`

We can compare sequences. But first we have to align these sequences to identify common regions

A	T	G	C	A	T	A	C
↕	↕	↕	X	X	↕	↕	↕
A	T	G	—	—	T	A	C

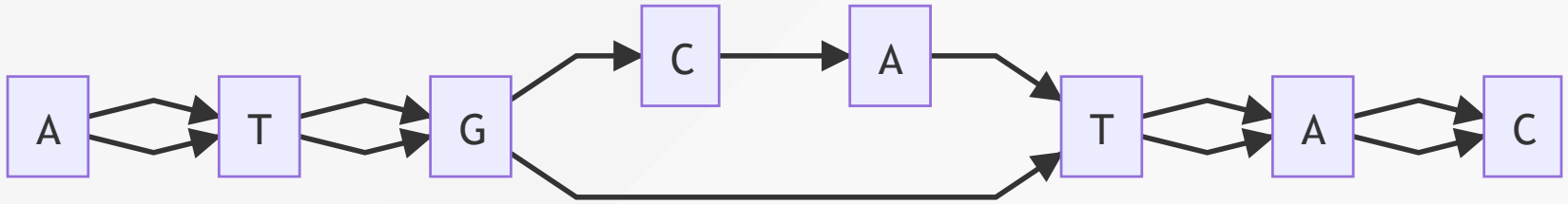
To investigate this difference, we need to identify regions that are different, and regions that are similar. To do that we will put these two sequences in a data structure called a partial order graph

Sequence as a partial order graph

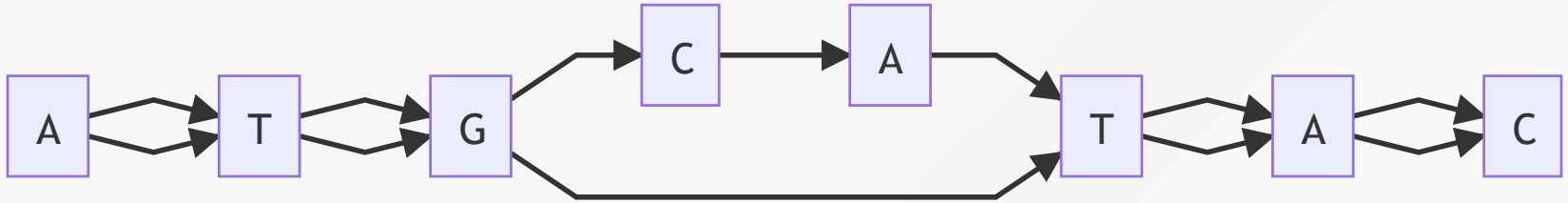
Our alignment

A	T	G	C	A	T	A	C
↕	↕	↕	X	X	↕	↕	↕
A	T	G	—	—	T	A	C

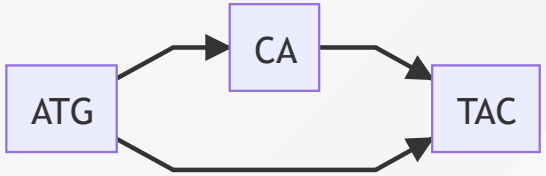
can be represented as the following partial order graph, showing each node and the direction of the alignment.



Extracting regions from the partial order graph



By collecting together adjacent nodes with the same number of edges we can simplify that to



Now we can make some claims about which regions are present in both sequences

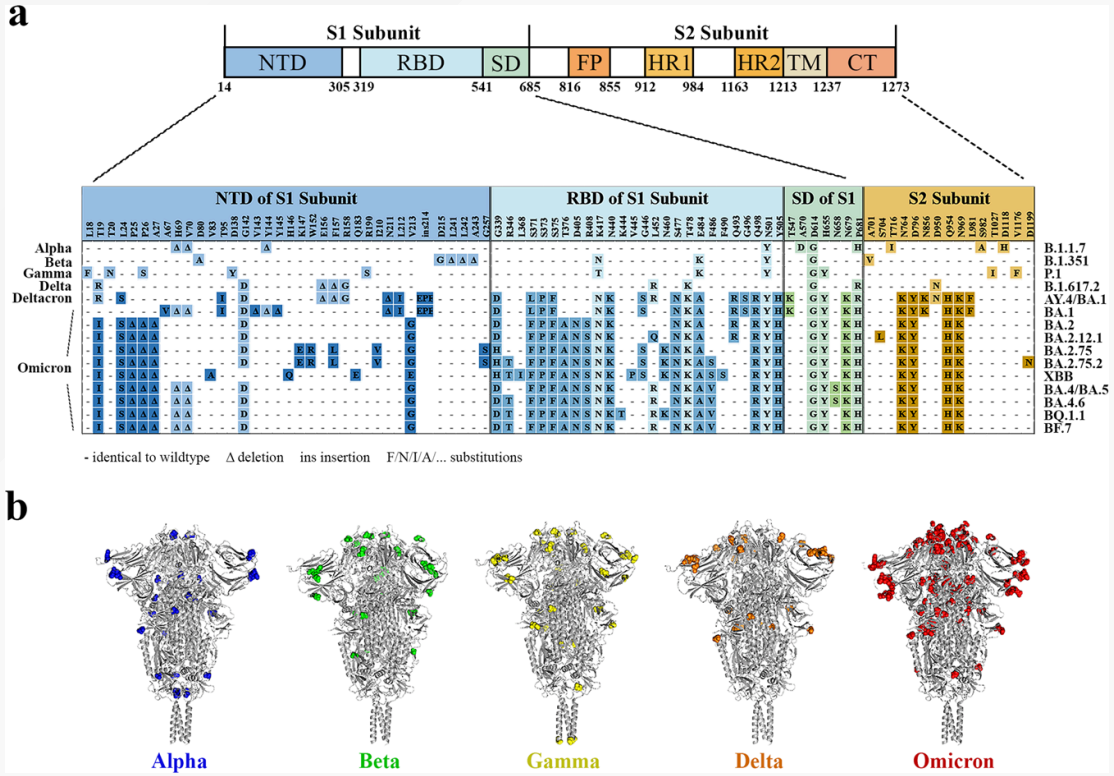
eg: If those regions encoded for genes, then we can make some claims about organism genotype.

	ATG	CA	TAC
Sequence A	✓	✓	✓
Sequence B	✓	⊘	✓

Why is multiple sequence alignment (MSA) important?

Alignment of eg: a viral genome allows us to:

- Identify conserved regions for vaccine/drug development
- Identify changes in function to make predictions about the virus' behaviour
- Identify and prepare for emerging variants



Alignment of S mutation points of SARS-CoV-2 variants

Why is MSA so computationally expensive?

- A complete solution has an order complexity of $O(L^n)$
  - **L** is the length of the sequence
  - **n** is the number of sequences

MSA for SARS-CoV-2 genomes?

SARS-CoV-2

- length: **~29,903** bp
- number: **over 5 million** (as of March 2022) [1]
- $O(29,903^{\text{over 5 million}})$ is a very large number

Required: a method to align large numbers of small sequences

Fig 1: Artists rendition of SARS-CoV-2

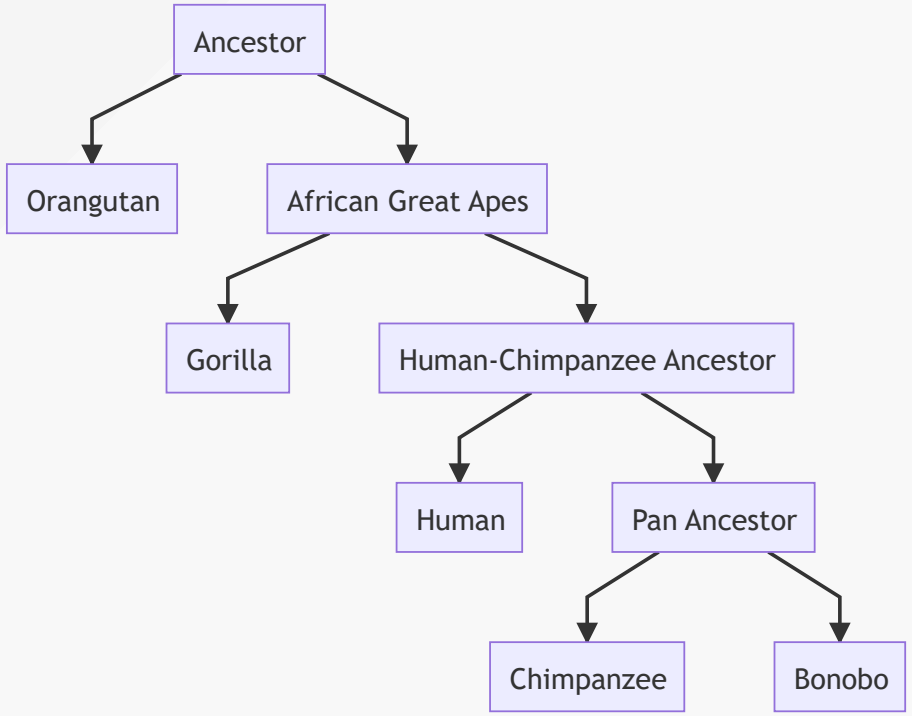
[1] doi.org/10.1038/s41588-022-01033-y | Fig 1 doi.org/10.7875/togopic.2020.199

MSA for great apes genomes?

The great apes

- length: **~3 billion** bp
- number: 5
- $O(3Billion^5)$ is also a very large number.
- However great ape genomes are 97+% identical[1]

Required: a method to identify the few different regions in very long similar sequences



The family tree of great apes

[1] citation needed

Project aims

1. Develop a more efficient method to align

- large numbers of small sequences
- small numbers of very similar long sequences

1. Quantify performance against previous methods
2. Quantify accuracy against previous method

Sequence alignment order complexity

Pairwise sequence alignment

- Compare every letter in one sequence to every letter in the other
- order complexity of $O(mn)$
  - where **m** and **n** are lengths of the sequences

Multiple sequence alignment (MSA)

- Perform a pairwise alignment of every sequence to every other sequence
- order complexity of $O(L^n)$
  - where **L** is the length of the sequences
  - **n** is the number of sequences

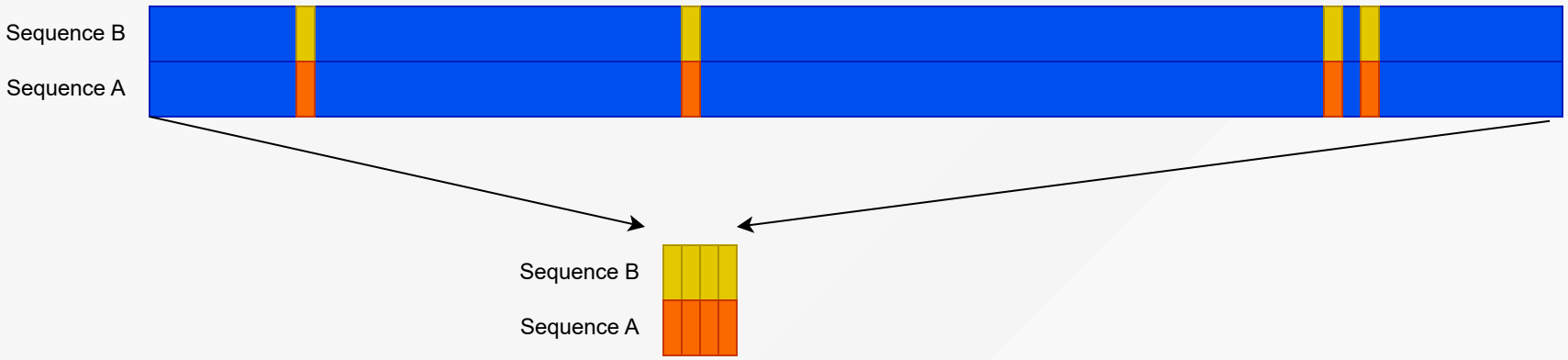
Pairwise sequence alignment methods

- Needlemann-Wunsch algorithm: global alignment for highly similar sequences
- Smith-Waterman algorithm: better for local alignment to find conserved domains

Multiple sequence alignment (MAS) methods

- ClustalW: Constuct a phylogenic tree and align pairs most closely related
- MAAFT: faster but less accurate
- MUSCLE: balances speed and accuracy
- T-Coffee: slower but more accurate

What if we could quickly remove regions that are similar?



We'd be able to focus our computational resources on just the regions that are different.

Sequence alignment using De Bruijn Graphs

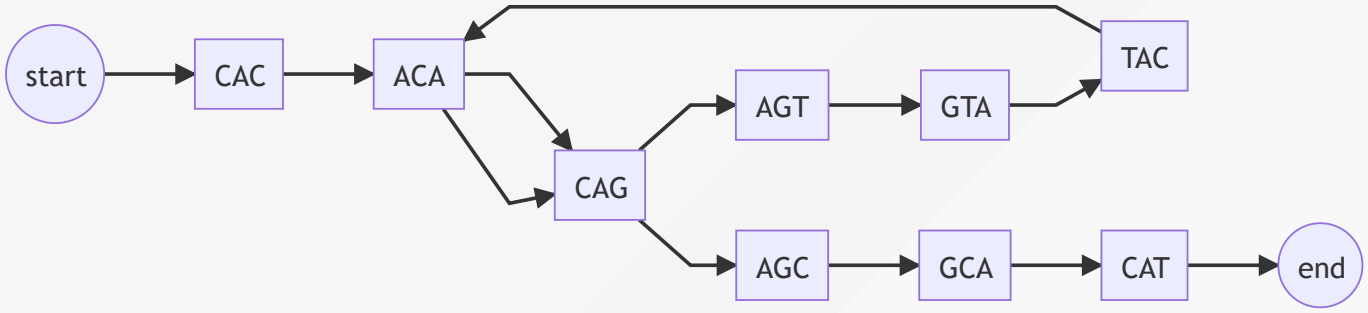
This work builds on the work by Xingjian Leng in a 12 month undergraduate research project in 2022[1] under the supervision of Dr. Yu Lin and Prof. Gavin Huttley.

That project focused on the alignment of closely related viral genomes, with a particular emphasis on SARS-CoV-2. The method is based on the construction and utilization of de Bruijn graphs for both pairwise and multiple sequence alignment tasks.

De Bruijn graphs

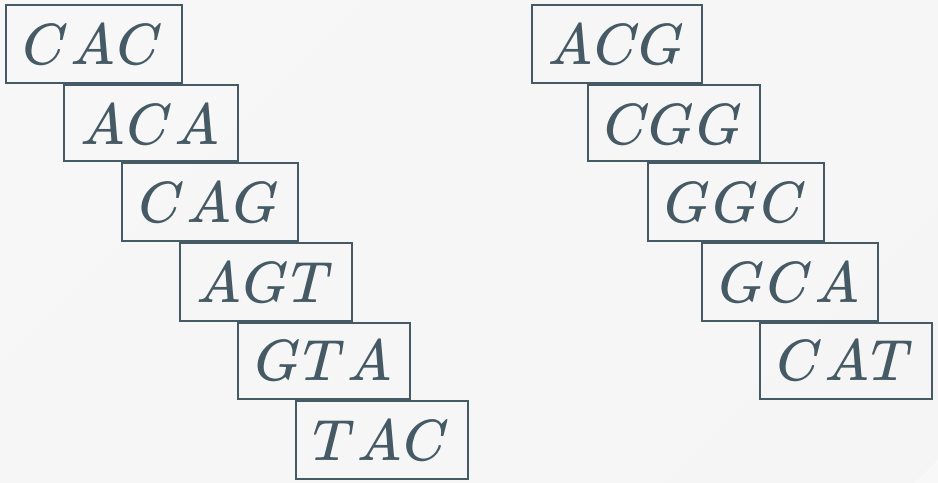
A De Bruijn graph is a directed graph that represents unique overlapping subsequences (or k-mers) at the nodes. This structure is an efficient way to identify sequence overlaps, and common regions.

Building a De Bruijn graph has an order complexity of $O(L)$ where L is the length of the sequence.



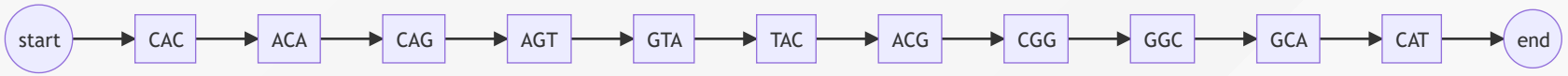
Overlapping k-mers

Consider the DNA sequence $CACAGTACGGCAT$ when broken into 3 character overlapping subsequences (or 3-mers) looks like this:



De Bruijn graphs

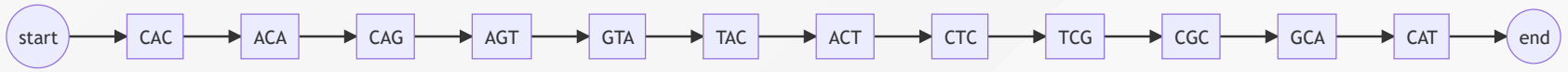
When we represent that as a de Bruijn graph it looks like this:



A second sequence

Consider we want to align that sequence $CACAGTAC\boxed{G}GCAT$ to the very similar sequence $CACAGTAC\boxed{T}CGCAT$

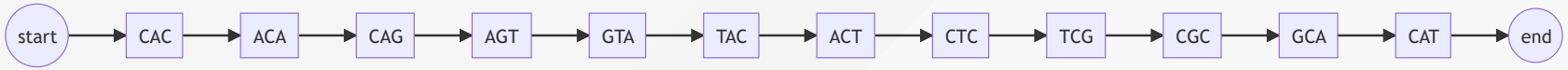
Which as a De Bruijn graph looks like this:



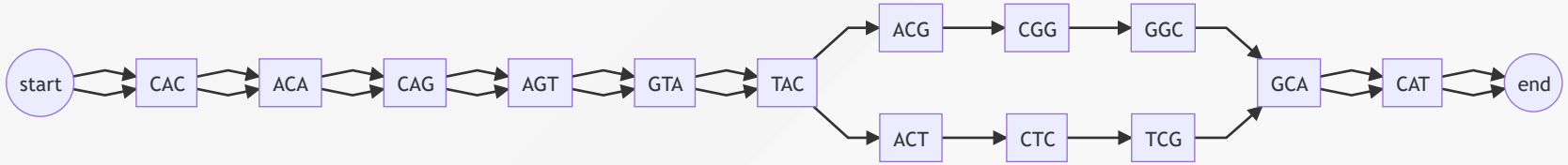
De Bruijn pairwise alignment

Sequence A:

Sequence B:

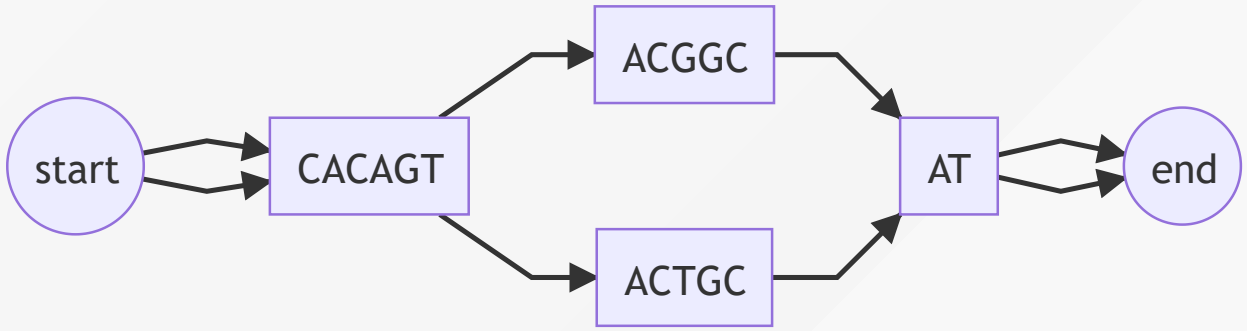


If we combine both sequences into a single de Bruijn graph, we can easily identify the regions that are similar and the regions that are different.



Resolving the graph

We can collect nodes with 2 edge, or 1 edge into single nodes, and we can see the regions that are similar and the regions that are different.

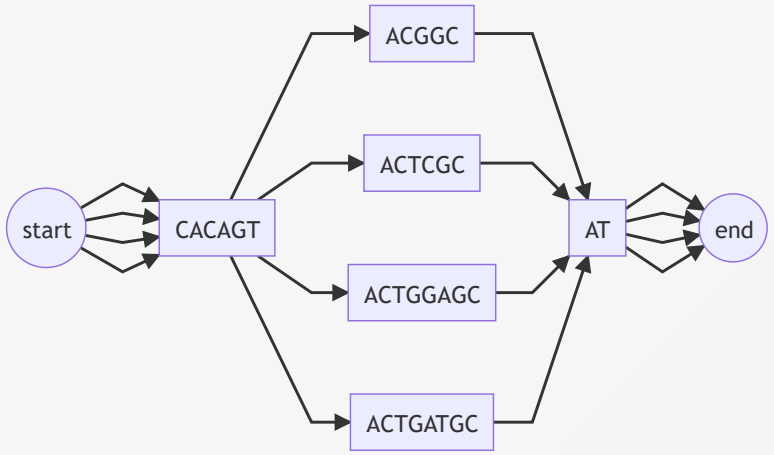


Now we can use a traditional algorithm to align the regions $\boxed{AC\boxed{G}GC}$ and $\boxed{AC\boxed{T}GC}$, and we've reduced $O(14^2)$ down to $O(5^2)$ = **7.8x** less work.

De Bruijn multiple sequence alignment

And we can extend this to multiple sequences. Consider aligning the following sequences

`CACAGTACGGCAT` `CACAGTACTGCAT` `CACAGTACTGGAGCAT` & `CACAGTACTGATGCAT`



Now we've reduced O(13x13x16x16) down to O(6x6x8x8) = **18.8x** less work

Project aims

- Investigate the use of De Bruijn graphs to identify regions of dissimilarity for traditional alignment algorithms
- Build a python library for implementing De Bruijn Graphs
- Quantify the performance of De Bruijn Graph sequence alignment against traditional methods
- Quantify the accuracy of De Bruijn Graph sequence alignment against traditional methods

Results

TBD...

Discussion

TBD...

Future directions

Investigate the potential of using De Bruijn Graphs to;

- identify repeats in sequences
- identify compliment regions in sequences
- identify strategies for choosing alignment methods to align regions of dissimilarity

Thanks

- Gavin Huttley
- Vijini Mallawaarachchi
- Xinjian Leng
- Yu Lin
- Huttley lab

Questions