Lecture 5

<From sequence to gene expression matrix: Assembly and mapping>

Kiwoong Nam **1155113612**

1. More About DP (Dynamic Programming)

1.1. Why we have to do DP and find the optimal sequence alignment (Brief review of what has been covered in Lecture 4)

➢ Pairwise sequence alignment

- In order to look further into biomolecular function and property prediction of different species and to carry out a more sophisticated studies about biological histories, finding out about sequence similarity and identifying regions of similarity are important.

- In **pairwise sequence alignment**, two sequences are arranged to maximize the similarity in between them (insertion of gaps are also allowed)

- However, there are too many possible alignments . . . !

→ Solution : **DYNAMIC PROGRAMMING (DP)**

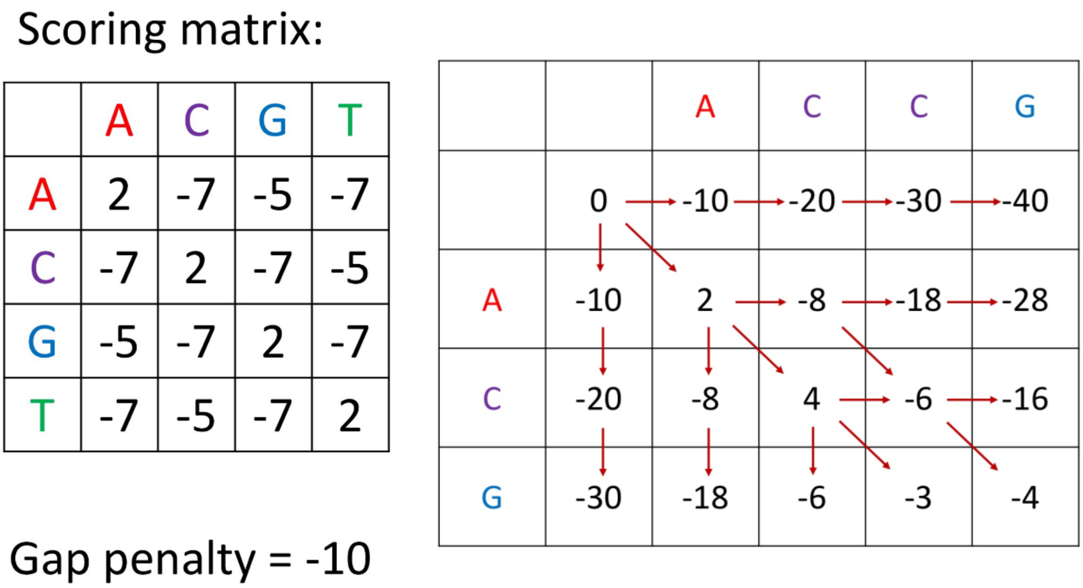
→ Obtain the optimal alignments of sequences that have the highest score (based on the 'scoring matrix') (the score indicates the similarity between the two alignments)

Lecture 5

<From sequence to gene expression matrix: Assembly and mapping>

Kiwoong Nam 1155113612

1.2. Scoring matrix and DP table



Scoring matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10

		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6	-16
G	-30	-18	-6	-3	-4

Figure 1: An image of a scoring matrix (left) and DP table (right) to obtain the best alignments from ACCG and ACG sequences.

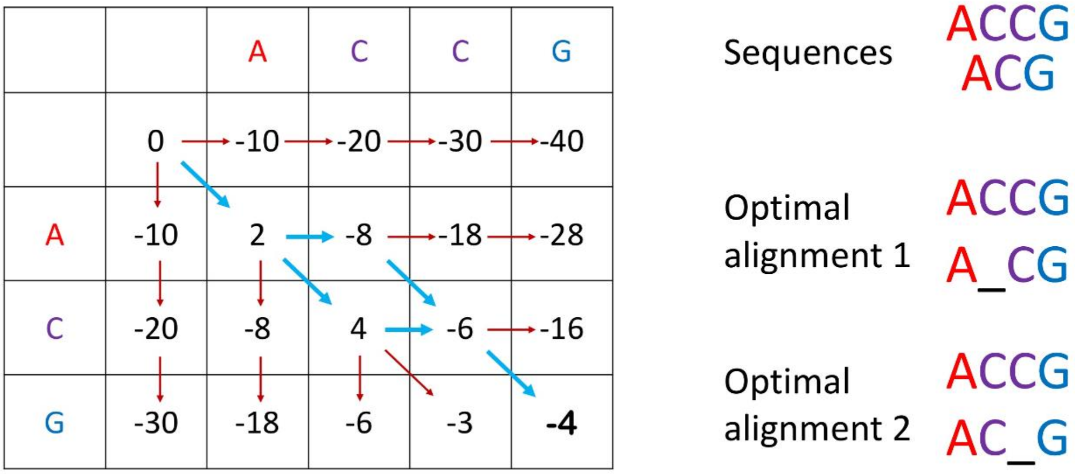
** Scoring matrices can differ based on different databases used for the dynamic programming.

How to do DP (another brief review of Lecture 4)

1. Fill in all the cells in the table (The detailed method is explained in Lecture 4 notes.)

2. Preserve the arrows which show in which direction the alignments of corresponding cell in the table have come with the highest score

3. The value in the last cell (i.e. cell in the bottom right corner of the DP table shown in Figure 1) is the best alignment score (the similarity between the two sequences)



		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6	-16
G	-30	-18	-6	-3	**-4**

Sequences ACCG
ACG

Optimal alignment 1 ACCG
A_CG

Optimal alignment 2 ACCG
AC_G

Figure 2

4. Trace back the arrows to get the optimal alignment with the highest score (follow the blue arrows until you get the optimal alignments (in the above case in Figure 2, there are two sets of alignments that have the same score (indication of sequence similarity))

Scoring matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10

Depending on **<u>how the similarity between two sequences is defined</u>**, the scoring matrix can be different.

- **Mismatch** – due to mutations
- **Gaps** – due to insertion/deletion and gene duplications

Lecture 5

<From sequence to gene expression matrix: Assembly and mapping>

Kiwoong Nam **1155113612**

2. Why and How do we get gene expression matrix?

2.1. Why obtain a sequence data

According to the Central Dogma, processes called transcription and translation involving DNAs and RNAs result in a protein strand.

★The DNA sequences are responsible for genetic information.★

Phenotype =

Genotype (determined by the DNA sequences) + Environment

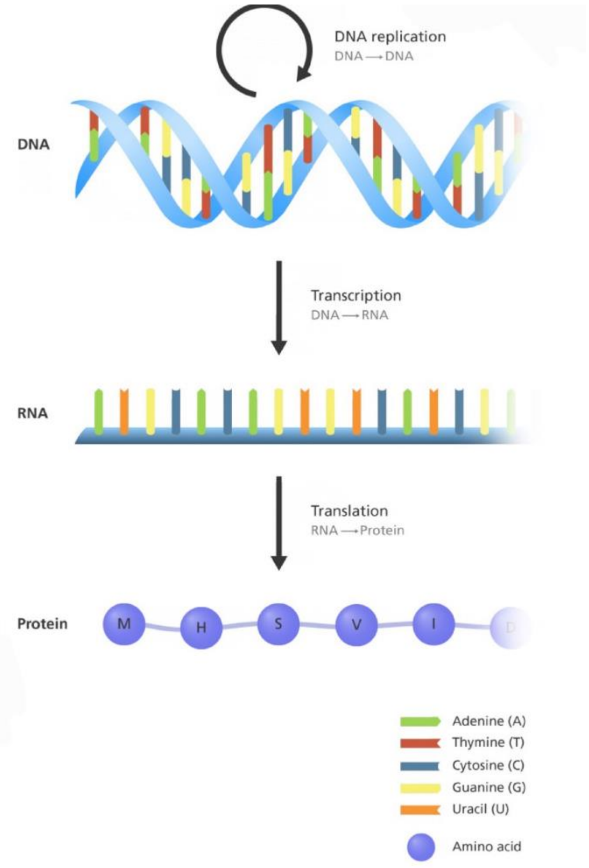


Figure 3: Diagram of the Central Dogma

Lecture 5

<From sequence to gene expression matrix: Assembly and mapping>

Kiwoong Nam **1155113612**

2.2. Gene Expression Matrix

2.2.1. Why do we need gene expression matrix

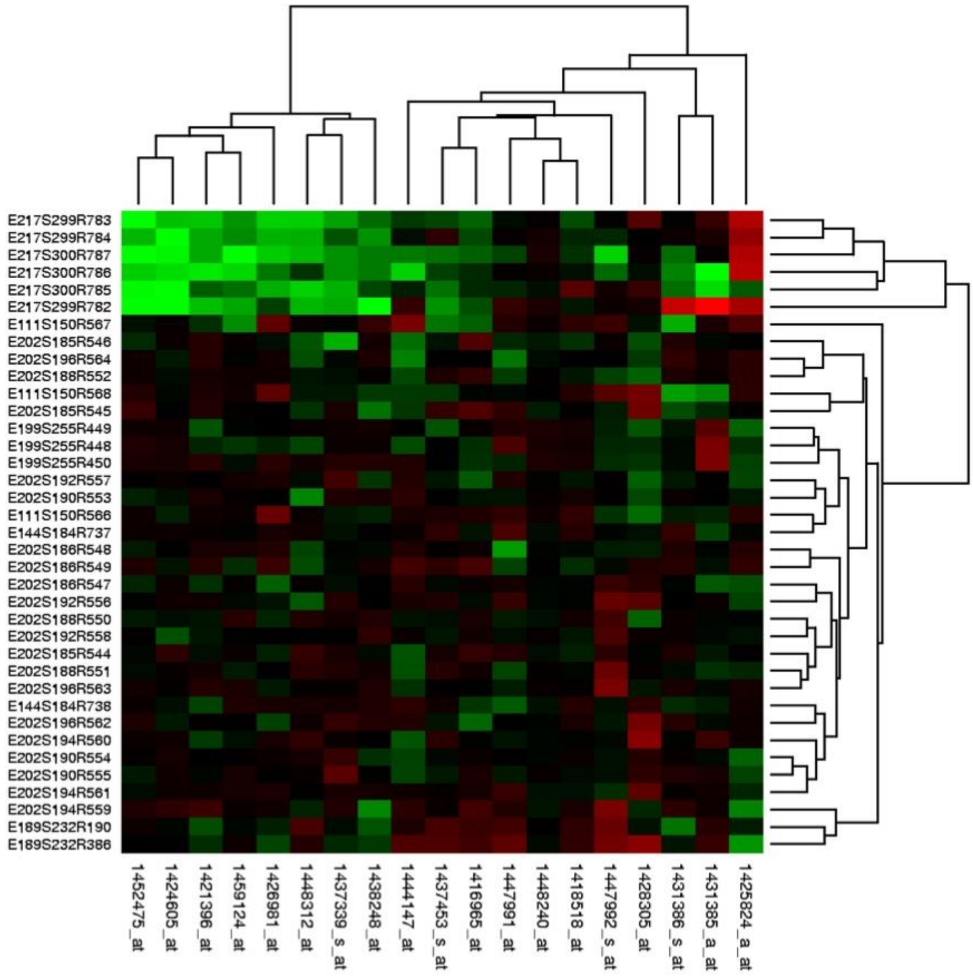


Figure 4: An image of a gene expression matrix (rows indicate the genes, and the columns indicate the value of gene expression counts.

1. Of human genomes, only 0.001% accounts for the genetic difference between different humans.
2. Only 1% of the entire human genome encodes protein.

Therefore, although the DNA sequence alone contains the genetic information of a human, in order to further examine the phenotype difference among different humans, a gene expression matrix (Figure 4) has to be taken into account!

2.2.2. How do we get a gene expression matrix

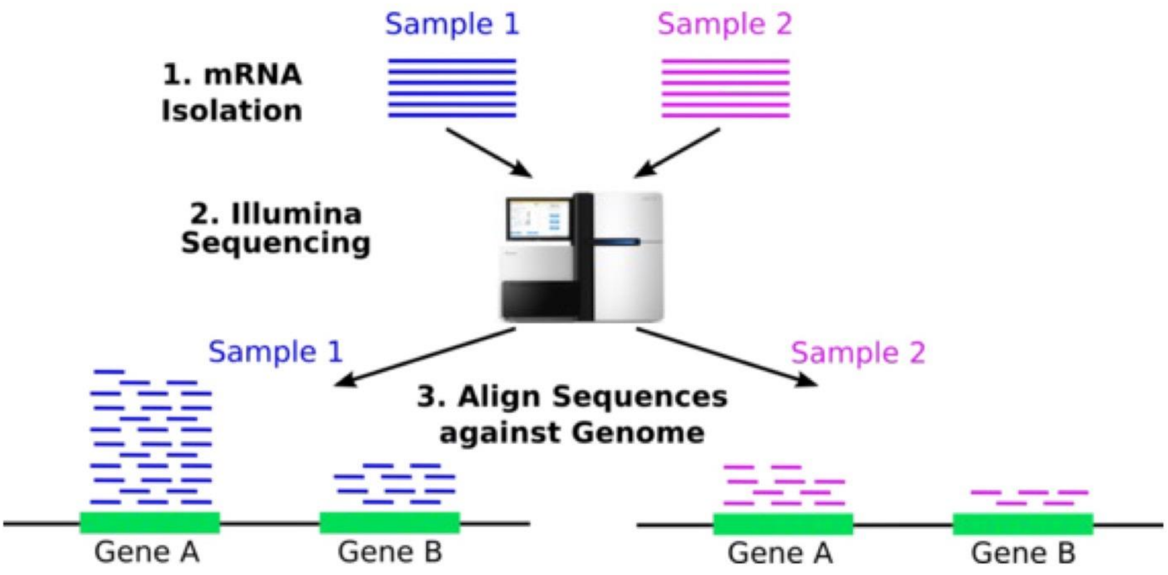


Figure 5

★ The main objective is to count how many mRNAs were produced and mapped into a specific gene in a genome.

Lecture 5

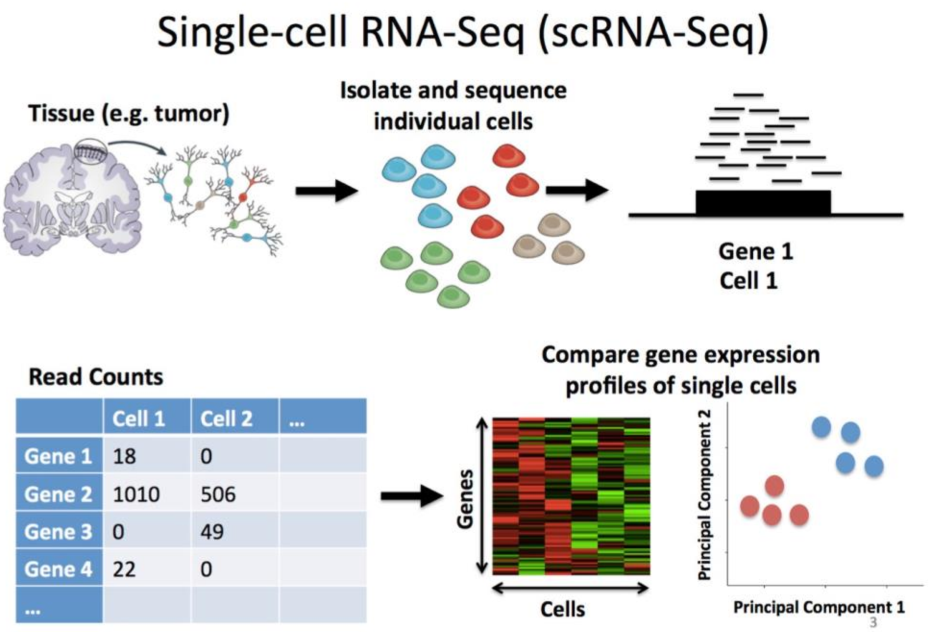
<From sequence to gene expression matrix: Assembly and mapping>

Kiwoong Nam **1155113612**

1. Isolate mRNAs from samples
2. Run the mRNA samples through Illumina Sequencing
3. Align the sequences against a genome
4. In the genome, there are micro-genes, Gene A and Gene B (as shown in Figure 5 above.)
5. Some of the reads (mRNAs) are mapped into Gene A
6. Count the number of the reads mapped into Gene A
7. Put the value of counts into the gene expression matrix
8. Repeat Step 4 – 7 for all the samples and for Gene B as well

(Red color (on the gene expression matrix) → large number of reads

Green color → small number of reads)


Single-cell RNA-Seq (scRNA-Seq)
Tissue (e.g. tumor)
Isolate and sequence individual cells
Gene 1
Cell 1
Read Counts
Compare gene expression profiles of single cells
Genes
Cells
Principal Component 2
Principal Component 1
3

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

Figure 6: An example showing steps of RNA sequencing and transitioning the data to gene expression matrix.

Lecture 5

<From sequence to gene expression matrix: Assembly and mapping>

Kiwoong Nam **1155113612**

3. Introduction to sequence assembly and sequence mapping

3.1. How do we get the genome sequence?

Genome Assembly

- Human genome length : around 3 billion bp
- The genome is assembled based on the short reads and overlap regions.

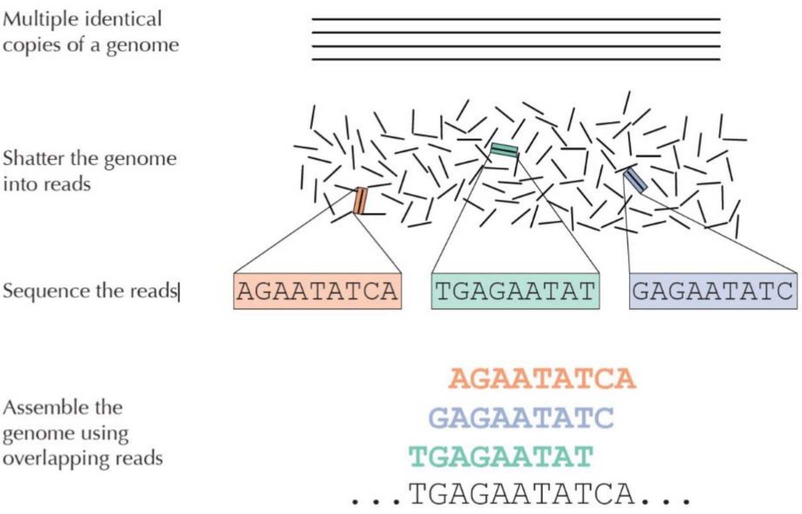
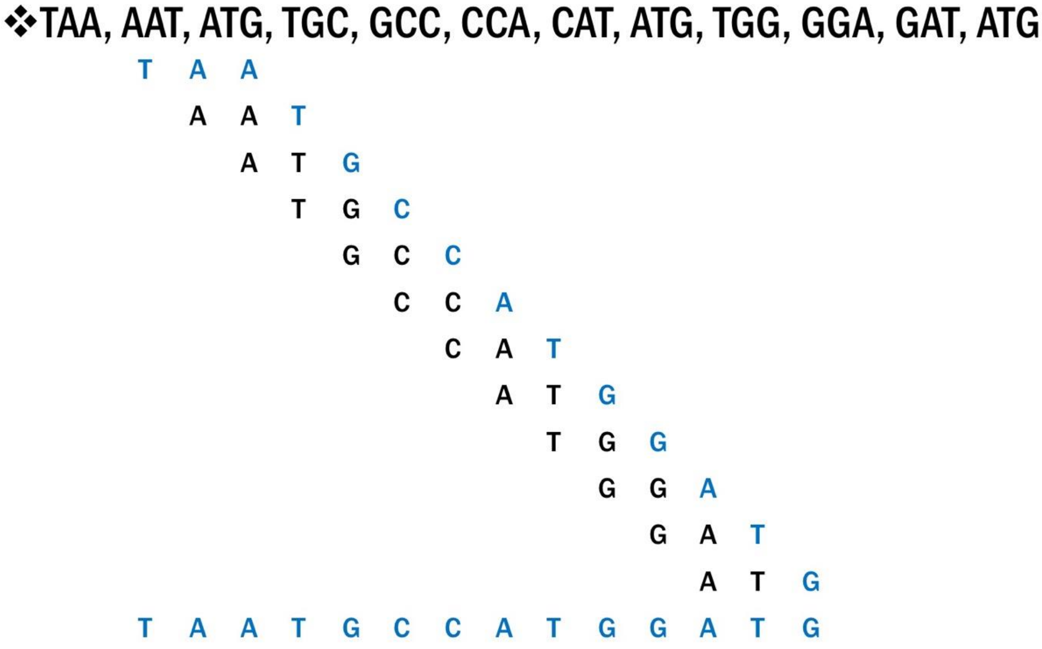

Multiple identical copies of a genome
Shatter the genome into reads
Sequence the reads
AGAATATCA
TGAGAATAT
GAGAATATC
Assemble the genome using overlapping reads
AGAATATCA
GAGAATATC
TGAGAATAT
...TGAGAATATCA...

Figure 7: An example of genome assembly (1)



❖TAA, AAT, ATG, TGC, GCC, CCA, CAT, ATG, TGG, GGA, GAT, ATG

```
T A A
  A A T
    A T G
      T G C
        G C C
          C C A
            C A T
              A T G
                T G G
                  G G A
                    G A T
                      A T G
T A A T G C C A T G G A T G
```

Figure 8: An example of genome assembly (2)

As shown in Figure 7 and 8, the overlapping parts of different short reads will result in a longer read of genome.

3.2. How do we map the short reads to the genome?

Let's set an example where there is a genome : TAATGCCATGGATG

and there are short reads from mRNA samples

: TAA, CCA, GAT, GCC, CCA, ATG

☞ Slide each short read of mRNA along the genome and calculate the number of mismatches within the given section (shown in Figure 9).

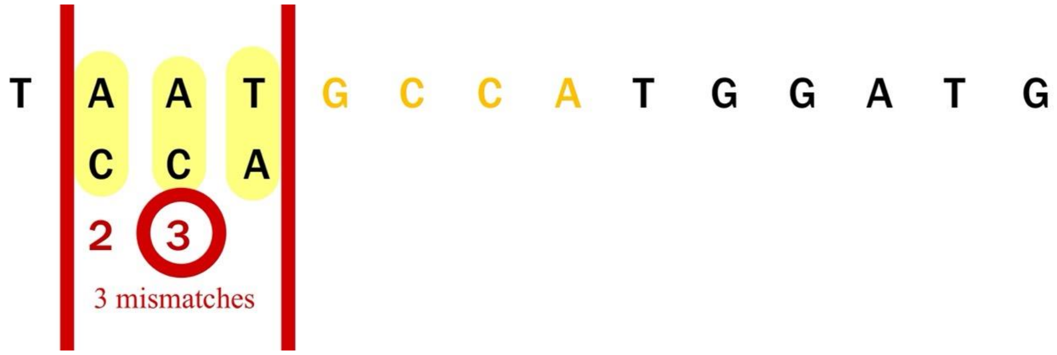


Figure 9

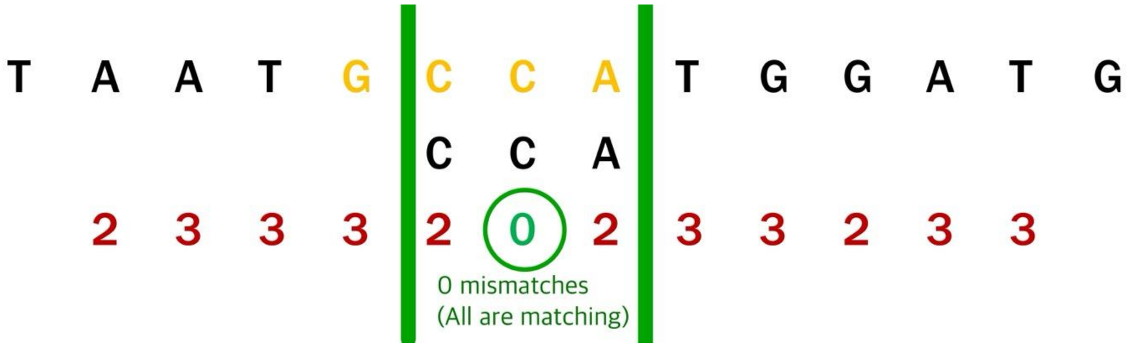


Figure 10

As shown in Figure 10, there is a particular section where the short read of mRNA slides against a genome, and there are no mismatches.

→ CCA is mapped to Gene GCCA in the genome.

Repeat the process with the remaining reads, and the result is shown below (Figure 11) :

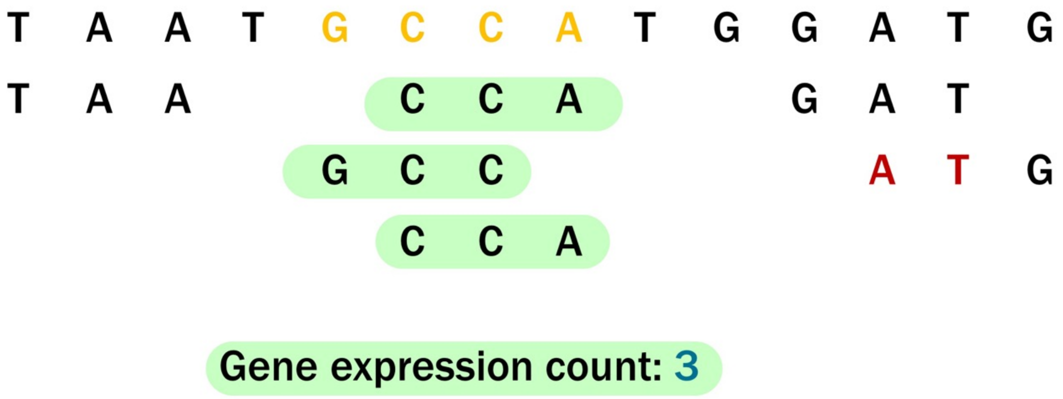


Figure 11

※ If there is a <u>partial match</u> in between a short read and a specific gene, for example:

GCCA

CAT

This **<u>could be</u>** counted as a gene expression count depending on what algorithm is being utilized.