Pendant le cours, vous avez étudié comment, dans le séquençage à haut débit, les reads doivent être mappés sur le génome humain. On va faire un calcul simple pour estimer le nombre d'opérations qui sont nécessaires. Ici, nous ne considérons que les exons [https://en.wikipedia.org/wiki/Exome_sequencing] si nous incluions également les introns, le nombre d'opérations serait beaucoup plus grand, pourquoi ?

Le génome humain contient environ 20 000 gènes codants. Supposons que la longueur moyenne des exons soit de 1 000 nucléotides et que longueur d'un read soit de 100 paires de bases [100bp].

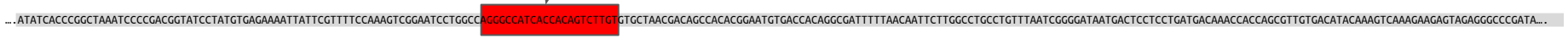
Question 1) Quel est le nombre approximatif d'opérations pour mapper un read ? Et 10^6 reads (le output typique d'une machine à haut débit) ?

Pour votre connaissance, la longueur d'un read dépend de la technologie de séquençage. La longueur typique pour le séquençage Illumina (l'une des plus utilisées dans le monde) est de ~80 paires de bases [80bp].

Our read (100 nt)

```
@WINDU:356:H5GKLBCX3:2:1101:3526:2088 1:N:0:AGGCAGAA+CGTCTAAT
TTTTCCTCTCAGCCCTCAGAACCCAGGGACTCAGCTGTGTACTTCTGTGCCAGCAGTTTTTTCCCCGGGGAGCTGTTTTTTGGAGAAGGCTCTAGGCTGACCGTACTGGAGGACCTGAAAAACGTGTTCCCACCCGAGGTCGCTGTGTTTG
+
DDDDDHHIIIIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIIHHIIIIIIIIIIIHIHIIIIIIIHIIGIIIIIIIIIIIIIIIIIIIHHFHHIIIHIHIHIDEHIIIGIIIIIIIIIIIHIIIIHIIIDHIFHGHHHHHIEFHIIII/
```

Notre objectif : calculer le nombre d'opérations pour mapper le read sur le génome de référence (uniquement les exons)



1) Quelle est la longueur du génome de référence (exons uniquement) ?

20 000 gènes codants. Supposons que la longueur moyenne des exons soit de 1000 nucléotides

Exon 1 (1000 nt)

….ATATCACCCGGCTAAATCCCCGACGGTATCCTATGTGAGAAAATTATTCGTTTTCCAAAGTCGGAATCCTGGCCAGGGCCATCACCACAGTCTTGTGTGCTAACGACAGCCACACGGAATGTGACCACAGGCGATTTTTAACAATTCTTGGCCTGCCTGTTTAATCGGGGATAATGACTCCTCCTGATGACAAACCACCAGCGTTGTGACATACAAAGTCAAAGAAGAGTAGAGGGCCCGATA….

Exon 2 (1000 nt)

….GTAATACCCAGAACTCGCTTTGATTTCGTAGACAGAATTGACACCGCCGTCGCAGTTAGAAGTGCAGGGAGACATATCCCCCTCCCCCATACATGTCCCCGCGCTAGAGTTCAGTCTTGCACGCCGTGTAAATTAGCAAAGCGAATTTTGGCATTTAGCACCATGTAAGTAGATCCCCATACATGTCCCCGCGCTAGAGTTCAGTCTTGCACGCCGTGTAAATTAGCAAAGCGAATTTTGGCA..

Exon 3 (1000 nt)

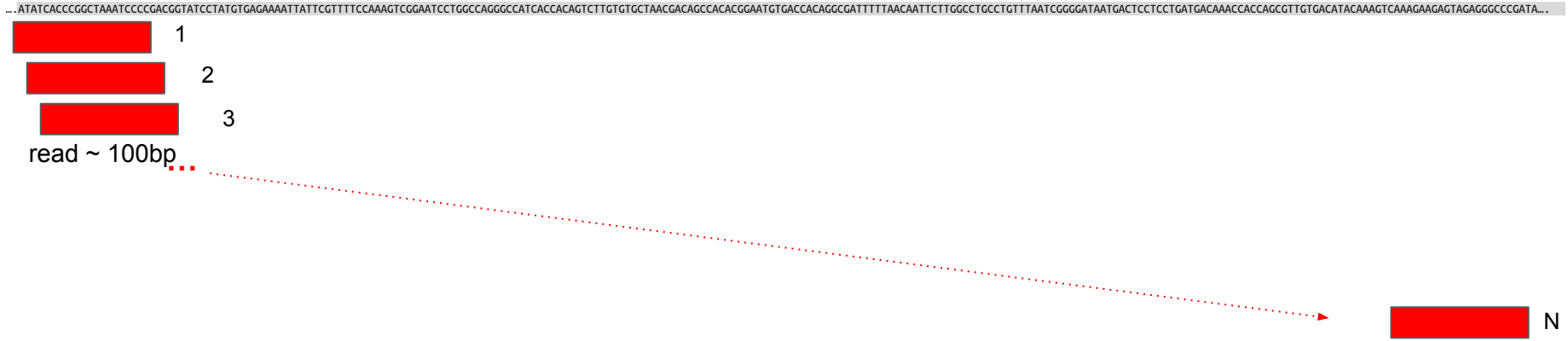
….AGAAAATTATTCGTTTTCCAAAGTCGGAATCCTGGCCAGGGCCATCACCACAGTCTTGTGTGCTCGCTAGAGTTCAGTCTTGCACGCCGTGTAAATTAGCAAAGCGAATTTTGGCATTTAGCACCATGTAAGTAGATCCCCATACATGTCCCCGCGCTGTGTAAATTAGCAAAGCGAATTTTGGCTTGTGTGCTCGCTAGAGTTCAGTCTTGCACGCCGGCACGCCGTGTAAATTAGCAAAGCG..

Etc. etc.

longueur totale du génome de référence: 20000 (gènes codants) * 1000 (taille exon) = 2*10^7 nucléotides

2) **Combien de manières possibles avons-nous pour mapper un read?** Comptez les!

Génome de référence: 20000 (gènes codants) * 1000 (taille exon) = 2*10^7 nucléotides



Calculer N

2) Combien de manières possibles avons-nous pour mapper un read? Comptez les!

Toy example:

```
ACGGTATCCT = reference genome of len 10
ATCC = read of len 4
```

```
ACGGTATCCT      ACGGTATCCT      ACGGTATCCT      ACGGTATCCT
ATCC             ATCC            ATCC            ATCC

ACGGTATCCT      ACGGTATCCT      ACGGTATCCT
    ATCC            ATCC            ATCC
```

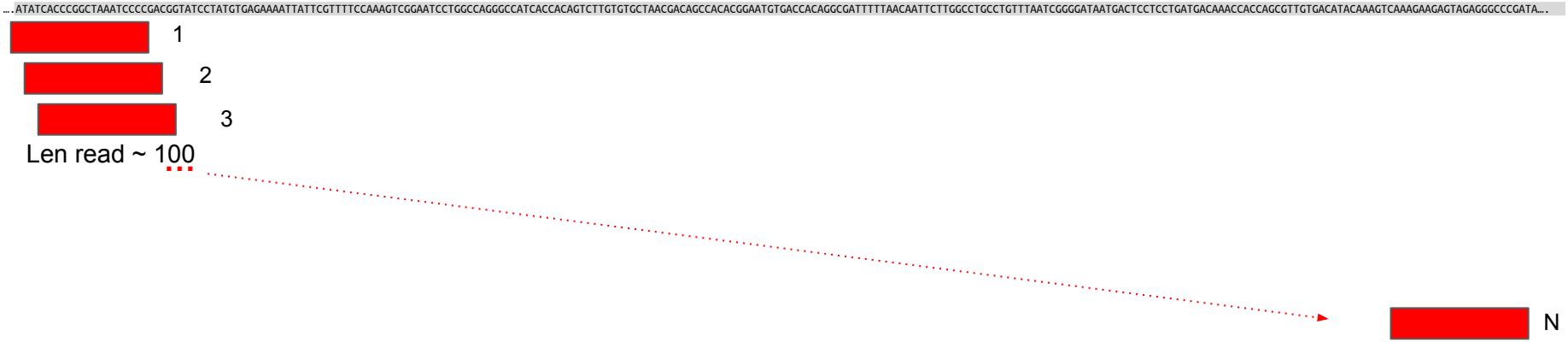
N= 10 - 4 + 1

Num = Len_reference_genome - Len_read + 1

In our case Len_reference_genome >>> Len_read
Num ~= Len_reference_genome

2) **Combien de manières possibles avons-nous pour mapper un read?** Comptez les!

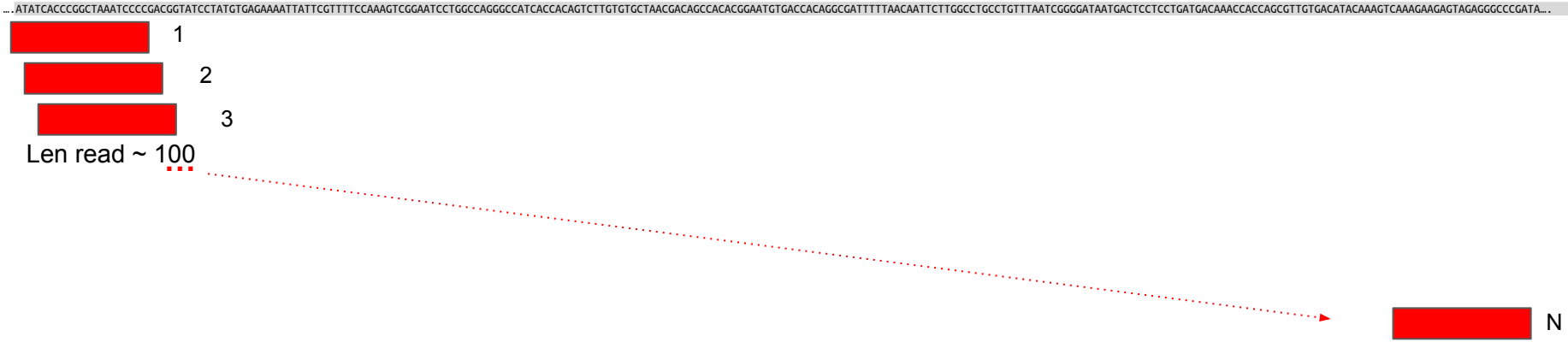
Génome de référence: 20000 (num gènes codants) * 1000 (taille exon) = 2*10^7 nucléotides



N =~ longueur totale du génome de référence = 2*10^7 mapping possible

2) **Combien de manières possibles avons-nous pour mapper un read?** Comptez les!

Génome de référence: 20000 (num gènes codants) * 1000 (taille exon) = 2*10^7 nucléotides



N =~ longueur totale du génome de référence = 2*10^7 mapping possible

En plus,pour chaque mapping, il faut vérifier si le read est égal à cette portion du génome.
Il faut faire une comparaison de nucléotides un à un, donc "longueur du read" opérations ~ 100.

Au total, pour faire le mapping d'un *seul read* on doit faire N_op = 2*10^7 * 100 = 2*10^9 opérations

3) Et pour mapper 10^6 reads (le output typique d'une machine à haut débit)?

Génome de référence: 20000 (num gènes codants) * 1000 (taille exon) = 2*10^7 nucléotides

pour 10^6 reads, vous devez répéter cette recherche pour chaque read donc
N_op = 10^6 * 2*10^9 ~= 2*10^15 opérations . A huge number!

How do they do in practice: https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform