Pendant le cours, vous avez appris comment, dans le séquençage à haut débit, les lectures ("reads") sont mappées sur le génome humain. Nous allons maintenant effectuer un calcul simple pour estimer le nombre d'opérations nécessaires pour cet alignement. Ici, nous nous concentrons uniquement sur les exons ([https://en.wikipedia.org/wiki/Exome_sequencing]). Si nous incluons également les introns, le nombre d'opérations serait bien plus élevé. Pourquoi ?

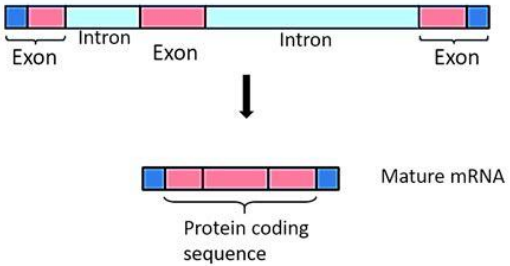
Le génome humain contient environ 20 000 gènes codants, chacun composé de plusieurs exons. Supposons que la longueur totale des exons par gène soit de 1 000 nucléotides. Quelle est la longueur totale du génome de référence en ne considérant que les exons ?

Question 1) Combien d'opérations approximatives sont nécessaires pour mapper un seul "read" ? Et combien d'opérations pour mapper 10^6 "reads" (le résultat typique d'une machine à haut débit) ?

Pour information, 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 d'environ 80 paires de bases [80bp]. Sur un serveur moderne à haute performance avec plusieurs cœurs, l'ensemble du pipeline de exome-sequencing peut prendre entre 6 et 24 heures.

Pendant le cours, vous avez appris comment, dans le séquençage à haut débit, les lectures ("reads") sont mappées sur le génome humain. Nous allons maintenant effectuer un calcul simple pour estimer le nombre d'opérations nécessaires pour cet alignement. Ici, nous nous concentrons uniquement sur les exons ([https://en.wikipedia.org/wiki/Exome_sequencing]). Si nous incluons également les introns, le nombre d'opérations serait bien plus élevé. Pourquoi ?

Le génome humain contient environ 20 000 gènes codants, chacun composé de plusieurs exons. Supposons que la longueur totale des exons par gène soit de 1 000 nucléotides. Quelle est la longueur totale du génome de référence en ne considérant que les exons



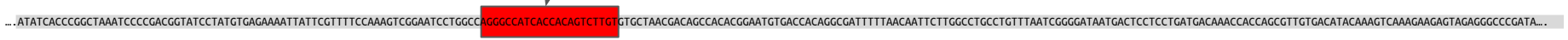
Answer:

Length reference genome = 20.000 (num genes) x 1.000 (total length exons per gene) = 2 x 10^7 base pairs

Our read (80 nt, for Illumina sequencing)

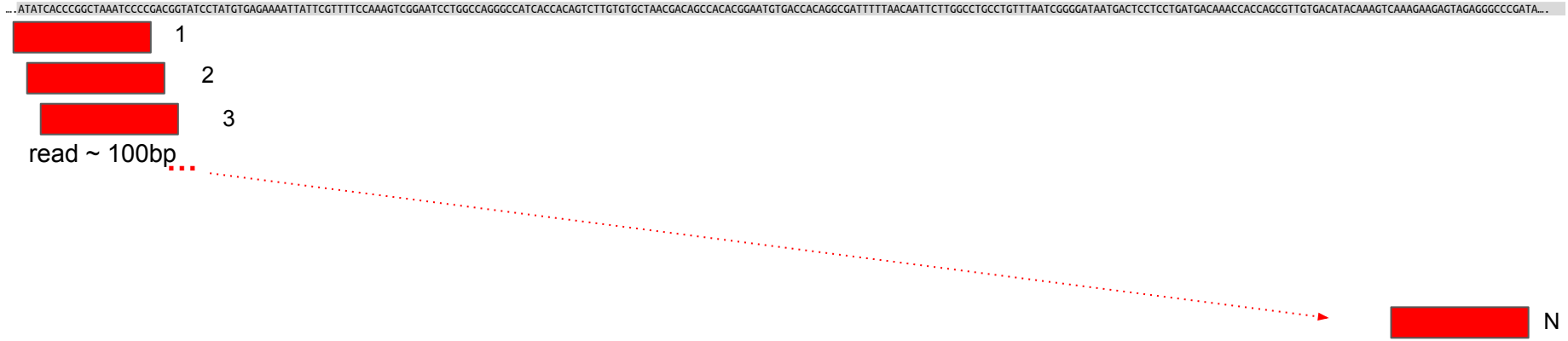
```
@WINDU:356:H5GKLBCX3:2:1101:3526:2088 1:N:0:AGGCAGAA+CGTCTAAT
TTTTCCTCTCAGCCCTCAGAACCCAGGGACTCAGCTGTGTACTTCTGTGCCAGCAGTTTTTTCCCCGGGGAGCTGTTTTTTGGAGAAGGCTCTAGGCTGACCGTACTGGAGGACCTGAAAAACGTGTTCCCACCCGAGGTCGCTGTGTTTG
+
DDDDDHHIIIIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIHHIIIIIIIIIIHIHIIIIIIIHIIGIIIIIIIIIIIIIIIIIIHHFHHIIIHIHIHIDEHIIIGIIIIIIIIIIIHIIIIHIIIDHIFHGHHHHHIEFHIIII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

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



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

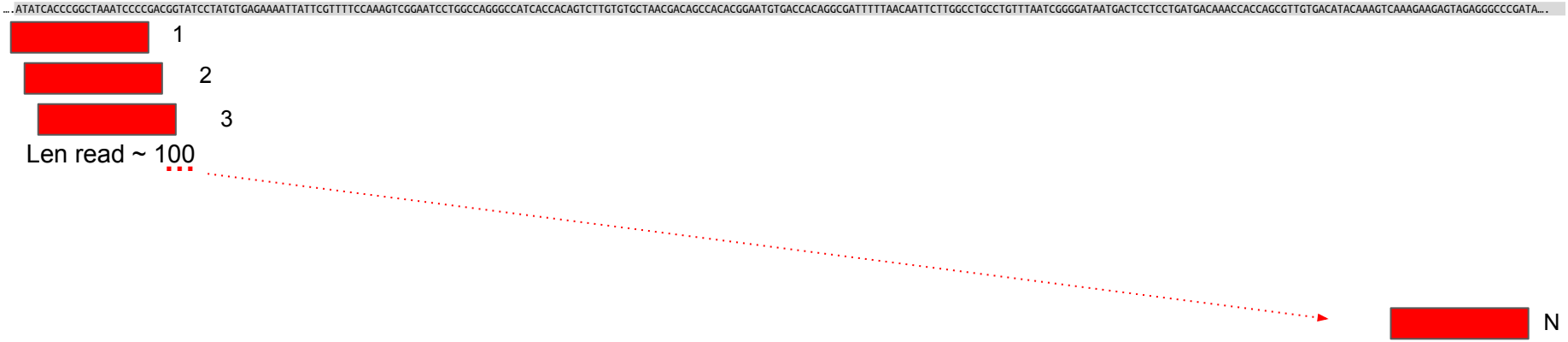
Num = Len_reference_genome - Len_read + 1

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

On pourrait s'arrêter une fois la correspondance trouvée, mais pour un génome réel, la situation est plus complexe (certaines parties du génome peuvent être dupliquées). Donc, on va jusqu'à la fin du génome de référence.

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

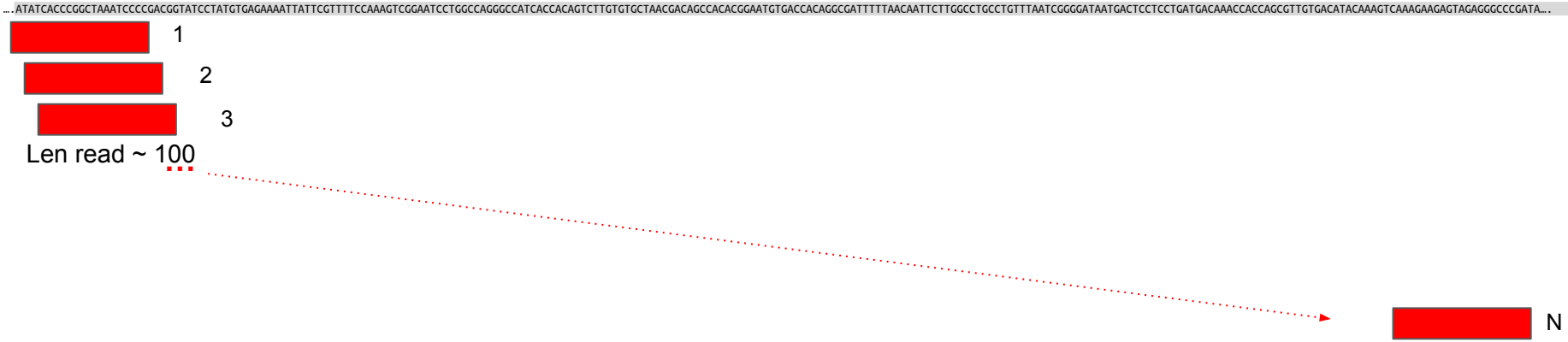
Génome de référence: 20000 (num gènes codants) * 1000 (taille exon) = 2*10^7 nucléotides



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 ~ 80.

Au total, pour faire le mapping d'un *seul read* on doit faire N_op = 2*10^7 * 80 ~ 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