TP 7 - Méthodes des données de séquençage

Données issues de séquenceurs :

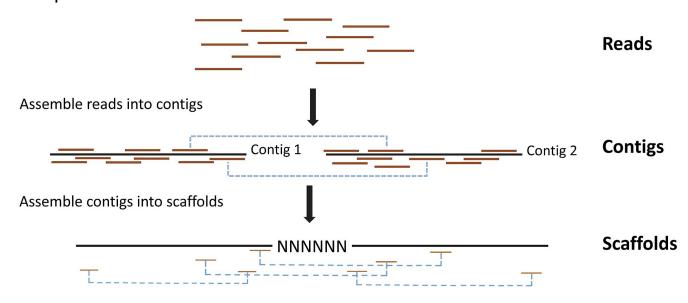
Séquençage 454 : 400 nucléotides.

Illumina NGS : séquences de 100 à 300 nucléotides.

Limitations:

Séquençage 454 impossible de déterminer avec certitude la longueur de ces répétitions si elles sont plus longues que 6 à 10 nucléotides. **Illumina NGS** pour chaque lecture on ne connaît pas sa position ni son orientation dans le génome dont elle est issue.

L'utilisation des lectures permet la formation de scaffolds qui peuvent être utilisés comme squelettes pour la finition de l'assemblage et à la résolution de répétitions.



Différences entre reads et scaffolds © 1

Enjeu : Alignement des lectures sur le génome efficacement.

^{1 -} T. J. Johnson, Marc; Carpenter, Eric J.; Tian, Zhijian; Bruskiewich, Richard; N. Burris, Jason; T. Carrigan, Charlotte; et al. (2012): Schematic representation of the method used to assemble Illumina reads into contigs, and contigs into scaffolds.. PLOS ONE. Figure.

```
TATATTTATGCTATTCAGTTCTAAATATAGAAATTGAAACAGCTGTGTTTAGTGCCTTTGTTCA----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAATTTGI
TATATT ATGCTATTCAGTTCTAAATATAGAAATTGAAACAG GTGTTTAGTGCCTTTGTTCA----ACCCCCTTGCAACAAC
tatatttatgctattcagttctaaatatagaaatt
                                       acagetqtqtttaqtqcctttqttca----acccccttq aacaacettqaqaaccccaqqqaatttqt
TATAT TATGCTATTCAGTTCTAAATATAGAAATTGAAACA etgtgtttagtgeetttgttea----accecettgeaac ACCTTGAGAACCCCAGGGAATTTGT
TATATTTA getatteagttetaaatatagaaattgaaacaget GTTTAGTGCCTTTGTTCACATAGACCCCCTTGCAA aacettgagaaceceagggaatttgt
                              GABATTGABACAGCTGTGTTTAGTGCCTTTGTTCA
TATATTTATGCTATTCAGT
                                                                          ccccttacaacaaccttgagaaccccagggaattt
                                                       GCCTTTGTTCACATAGACCCCCTTGCAACAACCTT
tatatttatgetatteagt
                                                                AG----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGA
tatatttatgetatteagtteta
TATATTTATGCTATTCAGTTCTAA
TATATTTATGCTATTCAGTTCTAAA
                                                                 A----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAA
                                                                 A----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAA
TATATTTATGCTATTCAGTTCTAAA
                                                                               TGCAACAACCTTGAGAACCCCAGGGAATTTGT
TATATTTATGCTATTCAGTTCTAAAT
                                                                               TGCAACAACCTTGAGAACCCCAGGGAATTTGT
TATATTTATGCTATTCAGTTCTAAAT
                                                                               TGCAACAACCTTGAGAACCCCAGGGAATTTGT
                                                                               tgcaacaaccttgagaaccccagggaatttgt
CAACCTTGAGAACCCCAGGGAATTTG1
tatatttatgctattcagttctaaatatagaaatt
tatatttatgctattcagttctaaatatagaaatt
  TATTTATGCTATTCAGTTATAAATATAGAAATTGAAACAG
                                                                                       CCTTGAGAACCCCAGGGAATTTGT
                                                                                         CTTGAGAACCCCAGGGAATTTGT
   atttatgctattcagttctaaatatagaaattgaa
    tttacgctattcagtactaaatatagaaattgaaa
                                                                                        CTTGAGAACCCCAGGGAATTTGT
     ttatgctattcagttctaaatatagaaattgaaac
                                                                                                      gggaatttgt
```

Mapping lectures sur le génome de référence © 2

Outils bio-informatique: BWA, BOWTIE2, HISAT

Fichiers:

```
@SEQ_ID——identifieur
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGFTCAACTCACAGTTT
+—délimiteur
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>CCCCCCC65
```

Fichier des lectures (FASTQ) © 3

```
format sam (fichier d'alignement)_séquences de référence
                                                                             qualité de
                                                                                            code
                         LN: 155270560
                                        -identifieur
@50
          SN:chrX
                                                                                 mapping
                                                                                              CIGAR
@SQ
                         LN:59373566
          SN:chrY
                                                      séq.de réf. position
@SQ
          SN:chrM
                         LN:16571
                     SM:ind1
                                    PL:illumina
HWI-ST1064:130:D289KACXX:1:2102:18718:11896
                                                                                               101M
        55504709
                        233
AATCTTTGCAAATTGAATCTTCTGGAAAGCTGA<del>GCT</del>IGTGCCTACCATAGAATTCTGAATGTACCTATATGACATCTTTGCAAACTTAAAACCTGAATCTT
BBBFFFFBBBBFFIIIIII111IIIIFFIIFFFIIIFBBFFF<del>FFFFBEB</del><FFFFFFFFFF<7BFFF<8BFBFFBBFFBFFBFFBFFBFFBFBBBBBF
                                                                          -séquence
                                                                                            qualité
             AS:i:96
                           XS:i:19
                                          RG:Z:1
HWI-ST1064:130:D289KACXX:1:2102:18718:11896
                                                 147
                                                                       55504709
                                                                                                101M
 score de qualité Phred position de la séquence appariée
                                                                            ~longueur perçue
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Fichier d'alignement (SAM) © 3

²⁻ https://biology.stackexchange.com/questions/1859/what-exactly-are-computers-used-for-in-dna-sequencing

³⁻ Robin Milosz, résumé TP8 BIN1002