MASTER SCIENCES
ET NUMERIQUE POUR LA SANTÉ

Développement d'un script de transfert d'annotation : TransPo-RG

Encadrants : Manuel Ruiz et Gaëtan Droc

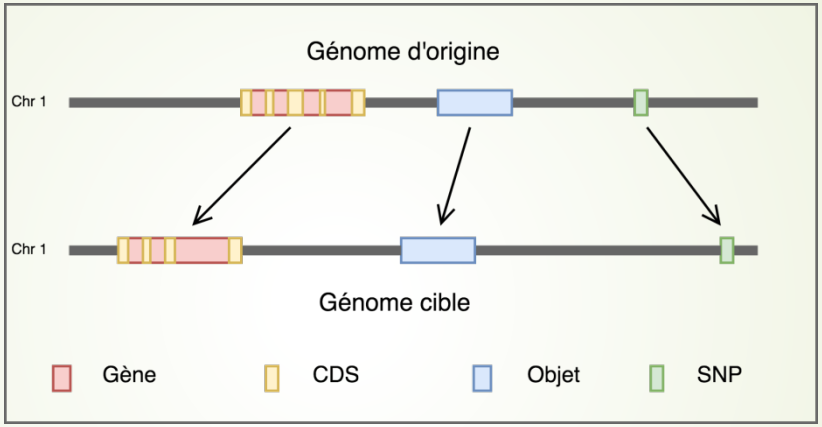
Tuteurs pédagogique : Alban Mancheron

23 avril 2018 - 23 août 2018

Clément Bellot

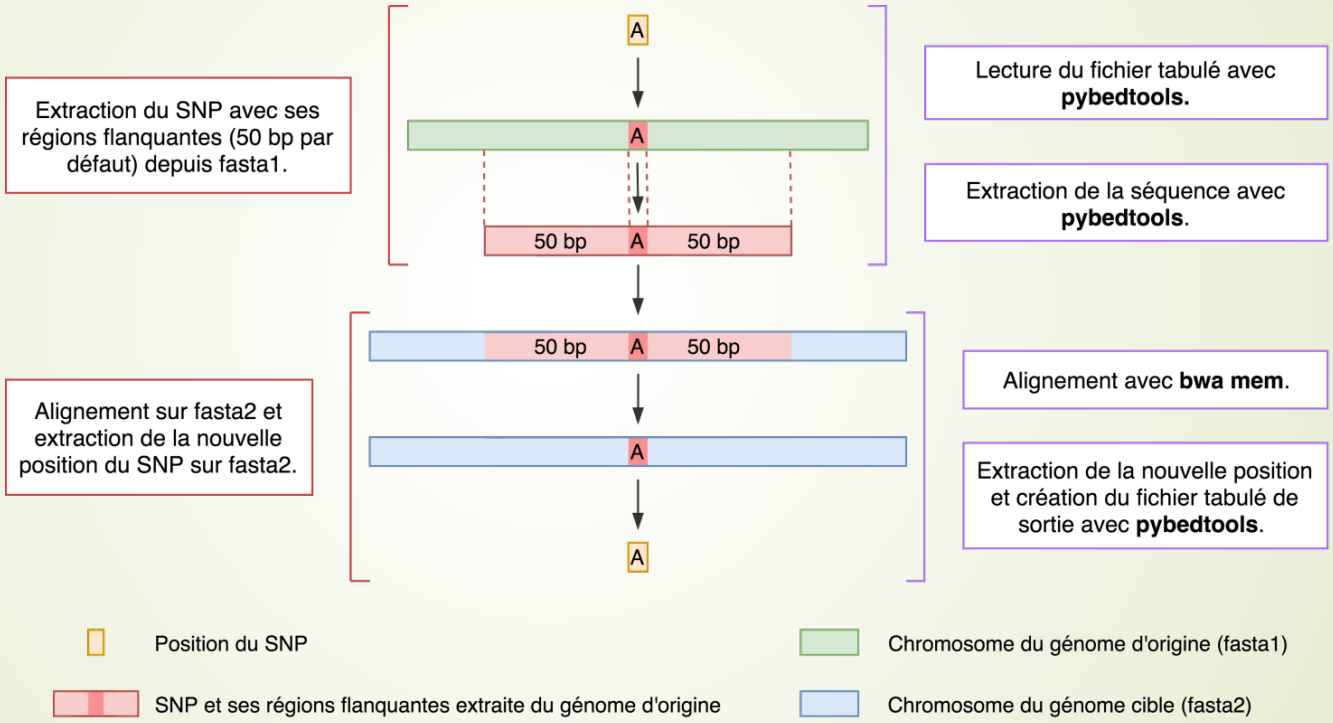
Master 1 – Sciences et Numérique pour la santé – Bioinformatique, Connaissances, Données

Méthodes



- Alignement des séquences des différents objets sur le génome cible.
  - Sélection des séquences avec leurs régions flanquantes

TransPo-RG


Extraction du SNP avec ses régions flanquantes (50 bp par défaut) depuis fasta1.
A
A
50 bp
A
50 bp
Lecture du fichier tabulé avec **pybedtools.**
Extraction de la séquence avec **pybedtools**.
50 bp
A
50 bp
A
A
Alignement sur fasta2 et extraction de la nouvelle position du SNP sur fasta2.
Alignement avec **bwa mem**.
Extraction de la nouvelle position et création du fichier tabulé de sortie avec **pybedtools**.
Position du SNP
Chromosome du génome d'origine (fasta1)
SNP et ses régions flanquantes extraite du génome d'origine
Chromosome du génome cible (fasta2)

TransPo-RG

```
(venv) cbellot@cc2-admin:~/work/TransPo-RG$ python transpo-rg.py -f1 data/example/Sbicolor_79.assembly.fna -f2 data
/example/Sbicolor_313_v3.1.assembly.fna -ti data/example/s10.snp.vcf -n -l -d vcfSNP -v 2
```

Fichier VCF en entrée →

```
chromosome_1    37307   .       T       A       1179.77 .       AC=2;AF=1.00;AN=2;DP=28;FS=0.
000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=28.18      GT:AD:DP:GQ:PL  1/1:0,28:28:84:1208,84,0
chromosome_1    39457   .       G       C       477.77  .       AC=2;AF=1.00;AN=2;DP=15;FS=0.
000;MLEAC=2;MLEAF=1.00;MQ=29.00;MQ0=0;QD=31.85      GT:AD:DP:GQ:PL  1/1:0,15:15:45:506,45,0
```

Cordonnées modifiées pour inclure les régions flanquantes →

```
chromosome_1    37256   37357
chromosome_1    39406   39507
```

Séquences extraites de fasta 1 →

```
>chromosome_1:37256-37357
ACCGGGTATACACAAAAAAGCATATCTATATATTCCTTTTATAAAGAAGATGAATGAGAAGATGAGAACTTGAGAAGGATACATGACATTTTA
GTCTATCA
>chromosome_1:39406-39507
TGCAATGGATAGCAAATATGAGCCTCATGGAAGATCCCATCGTTTTTATCGTGTACAAACAATTGATGATCATTATAGAAGAGAAGTGTATAT
TGGTATTA
```

Mapping des séquences extraites sur fasta 2 →

```
chromosome_1:37256-37357        0       Chr01   37049   60      101M    *       0       0
ACCGGGTATACACAAAAAAGCATATCTATATATTCCTTTTATAAAGAAGATGAATGAGAAGATGAGAACTTGAGAAGGATACATGACATTTTAGTC
TATCA   *       NM:i:0  MD:Z:101        AS:i:101        XS:i:0
chromosome_1:39406-39507        0       Chr01   39199   60      101M    *       0       0
TGCAATGGATAGCAAATATGAGCCTCATGGAAGATCCCATCGTTTTTATCGTGTACAAACAATTGATGATCATTATAGAAGAGAAGTGTATATTGG
TATTA   *       NM:i:0  MD:Z:101        AS:i:101        XS:i:0
```

Fichier VCF de sortie →

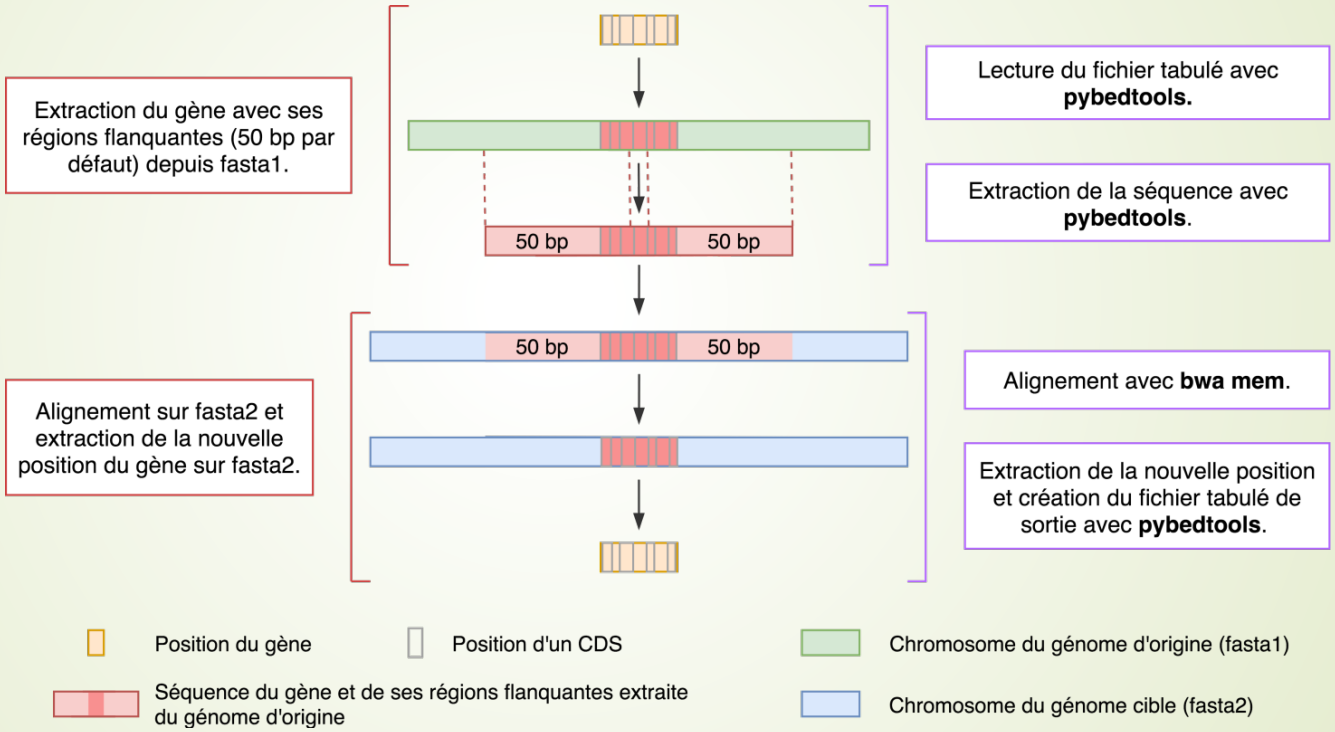
```
chromosome_1    37099   .       T       A       1179.77 .       AC=2;AF=1.00;AN=2;DP=28;FS=0.000;MLEAC=2;
MLEAF=1.00;MQ=60.00;MQ0=0;QD=28.18      GT:AD:DP:GQ:PL  1/1:0,28:28:84:1208,84,0
chromosome_1    39249   .       G       C       477.77  .       AC=2;AF=1.00;AN=2;DP=15;FS=0.000;MLEAC=2;
MLEAF=1.00;MQ=29.00;MQ0=0;QD=31.85      GT:AD:DP:GQ:PL  1/1:0,15:15:45:506,45,0
```

TransPo-RG

Fichier de statistique sur le pourcentage de perte.

#ID	REF	TARGET	LOSS	%
chromosome_1	31199	31154	45	0.14%
chromosome_2	22669	22638	31	0.14%
chromosome_3	25280	25130	150	0.59%
chromosome_4	20441	20360	81	0.40%
chromosome_5	9059	9036	23	0.25%
chromosome_6	15474	15450	24	0.16%
chromosome_7	11212	11171	41	0.37%
chromosome_8	9234	9197	37	0.40%
chromosome_9	13749	13722	27	0.20%
chromosome_10	14475	14461	14	0.10%

TransPo-RG


Extraction du gène avec ses régions flanquantes (50 bp par défaut) depuis fasta1.
50 bp
50 bp
Lecture du fichier tabulé avec **pybedtools.**
Extraction de la séquence avec **pybedtools**.
Alignement sur fasta2 et extraction de la nouvelle position du gène sur fasta2.
50 bp
50 bp
Alignement avec **bwa mem**.
Extraction de la nouvelle position et création du fichier tabulé de sortie avec **pybedtools**.
Position du gène
Position d'un CDS
Chromosome du génome d'origine (fasta1)
Séquence du gène et de ses régions flanquantes extraite du génome d'origine
Chromosome du génome cible (fasta2)

TransPo-RG

```
(venv) cbellot@cc2-admin:~/work/TransPo-RG$ python transpo-rg.py -f1 data/example/Sbicolor_79.assembly.fna -f2 data
/example/Sbicolor_313_v3.1.assembly.fna -ti data/example/Sbicolor_79.gff3 -n -l -d genecutGFF -v 2 -c -t "cds,gene"
```

Fichier GFF3 en entrée →

```
chromosome_1    phytozome       gene    63649   69514   .       -       .       ID=Sb01g000240;Name=
Sb01g000240;Note=similar to Expressed protein
chromosome_1    phytozome       mRNA    63649   69514   .       -       .       ID=Sb01g000240.1;Par
ent=Sb01g000240;Dbxref=Phytozome:1949297,TAIR:AT5G23890.1,MSU:LOC_Os03g64400.1;Name=Sb01g000240.1;No
te=similar to Expressed protein
chromosome_1    phytozome       polypeptide     63906   69290   .       -       1       ID=Sb01g0002
40.1.p;Derives_from=Sb01g000240.1;Name=Sb01g000240.1.p;Note=similar to Expressed protein
chromosome_1    phytozome       exon    63649   64404   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    65203   65734   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    65805   65919   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    66045   66120   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    66194   66293   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    66388   66491   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    66703   66791   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    67615   68591   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    68726   68818   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       exon    69038   69514   .       -       .       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     63906   64404   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65203   65734   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65805   65919   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66045   66120   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66194   66293   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66388   66491   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66703   66791   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     67615   68591   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     68726   68818   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     69038   69290   .       -       0       Parent=Sb01g000240.1
```

Fichier GFF3 filtré pour contenir uniquement les types d'objets demandés →

```
chromosome_1    phytozome       gene    63649   69514   .       -       .       ID=Sb01g000240;Name=
Sb01g000240;Note=similar to Expressed protein
chromosome_1    phytozome       CDS     63906   64404   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65203   65734   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65805   65919   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66045   66120   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66194   66293   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66388   66491   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66703   66791   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     67615   68591   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     68726   68818   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     69038   69290   .       -       0       Parent=Sb01g000240.1
```

TransPo-RG

Fichier GFF3 avec coordonnées modifiés pour inclure les régions flanquantes

```
chromosome_1    phytozome       gene    63599   69564   .       -       .       ID=Sb01g000240;Name=
Sb01g000240;Note=similar to Expressed protein
chromosome_1    phytozome       CDS     63856   64454   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65153   65784   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65755   65969   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65995   66170   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66144   66343   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66338   66541   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66653   66841   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     67565   68641   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     68676   68868   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     68988   69340   .       -       0       Parent=Sb01g000240.1
```

Mapping des séquences extraites sur le génome cible

gène

```
chromosome_1:63598-69564        0       Chr01   63391   60      5966M   *       0       0       CATTTAGGTTTAATAGTAATTATTATTATATC
TATTTGGATTTGGAATTTGGACCTATAATGCTTCAGTGCCTTTGAAGCCTTCACCTCAAGTCATCATCGTAACAATACTAATAACAAGAAGCGGCAGCTCATCCTGCCTGTATGAGGATTCATCAATC
CTGTCTGTACACAGCACGGACTTATTCCACGGGAACTACTGTTCGTTCCAGCATGACTTCAGTAAACTCTGTAACACCCACCACCCCAAAAATTTATTCCAAGGGAATTTACTATTGTTCTTGGTACA
GAGGCGACACAGCCTCCCTCTAGTCTGATGTCTTGAATCTGTGTGAAAACTTTTCCAGGCCGTCCTTGCACTCTTCCATTACCCTCTTCGACTTATCTCCAAGGGTTGCACCAAACGCAGATACACTG
```

CDS

```
chromosome_1:63855-64454        0       Chr01   63648   60      599M    *       0       0       TCCAAGGGAATTTACTATTGTTCTTGGTACAG
AGGCGACACAGCCTCCCTCTAGTCTGATGTCTTGAATCTGTGTGAAAACTTTTCCAGGCCGTCCTTGCACTCTTCCATTACCCTCTTCGACTTATCTCCAAGGGTTGCACCAAACGCAGATACACTGC
CCTGCACTTCTGCTGATACCATGAGTGCCTTGGAAGCAGCAGAAGCCCCAAACTCGGAACACCACTGCCTTGCATCTGCAGCTTGCTGCTTTAGGCTTGCAATGAAGGACCTCACGTGTTGCATCACC
```

Fichier GFF3 de sortie

```
chromosome_1    phytozome       gene    63441   69306   .       -       .       ID=Sb01g000240;Name=
Sb01g000240;Note=similar to Expressed protein
chromosome_1    phytozome       CDS     63698   64196   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     64995   65526   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65597   65711   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65837   65912   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     65986   66085   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66180   66283   .       -       1       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     66495   66583   .       -       0       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     67407   68383   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     68518   68610   .       -       2       Parent=Sb01g000240.1
chromosome_1    phytozome       CDS     68830   69082   .       -       0       Parent=Sb01g000240.1
```

TransPo-RG

Fichier de statistique sur le pourcentage de perte.

#ID	REF	TARGET	LOSS	%
chromosome_1	31199	31154	45	0.14%
chromosome_2	22669	22638	31	0.14%
chromosome_3	25280	25130	150	0.59%
chromosome_4	20441	20360	81	0.40%
chromosome_5	9059	9036	23	0.25%
chromosome_6	15474	15450	24	0.16%
chromosome_7	11212	11171	41	0.37%
chromosome_8	9234	9197	37	0.40%
chromosome_9	13749	13722	27	0.20%
chromosome_10	14475	14461	14	0.10%