HPV EPIMOL

Le projet tente de mettre en relation métadonnées et taxonomie pour HPV42 et les alpha-papillomavirus 11, 6, 13, 44 (avec 55), 74 et des virus de chimpanzés et bonobos.

I. Récupération des items au format GenBank sur NCBI

Les données sont dans le dossier "metadata" au format ".gb"

Pour avoir en possession les métadonnées, il suffit d'aller sur NCBI et de taper dans la barre de recherche :

```
"Human papillomavirus type 42"[ORGN] 100:15000[SLEN]
```

avec ORGN pour caractériser l'organisme d'intérêt et SLEN pour donner un intervalle de longueur des séquences.

Nous avons choisi comme intervalle de longueur [100,15000] pour avoir les portions de séquences et les génomes complets. Pour rappel, une taille de génome d'HPV attendue est aux alentours de 7900pb environ.

Du résultats de recherche a été récupéré le Genbank (.gb) en cliquant sur *Send to --> Complete Record --> File --> GenBank --> Create File*

Il est important de noter que les séquences provenant de brevets n'ont pas été prise en compte, c'est-à-dire celles avec un numéro de brevet ou qui comportent Patent dans les informations.

Attention, pour HPV6 c'est très particulier. En effet, lorsque l'on recherche uniquement "type 6" certaines séquences passent au travers. Du coup, il faut lancer d'autres recherches en désignant :

- type 6a
- type 6b
- type 6vc
- type 6c
- type 6e

Parmis l'ensemble des séquences 6b seules certaines ont été sélectionnées. 2 d'une longueur proche de 100 et 2 Patents / Brevet non sélectionnées. Pour HPV13, 2 Patents ont été omis.

Concernant HPV13, les références X62844, MK303593, OP934206 et AF020905 issues de bonobos et chimpanzés ont été ajoutées aux items.

HPV	Items
42	217
11	735
6	1466
13/Chimp&Bono	14+4

HPV	Items
44/55	72+105
74	17

II. Transformation du fichier GenBank

Les données sont dans le dossier "metadata" au format ".csv"

Le fichier GenBank contenant les métadonnées a été "transformé" en tableau facile de lecture contenant les informations d'intérêts au format CSV à l'aide d'un script python `gbk2table.py`.

```
usage: gbk2table.py [-h] [--output_file OUTPUT_FILE] genbank_file

Script to extract some information from a genbank format file, produced by the
NCBI site, and return a table with all the information retrieved that is :
ACCESSION, PUBMED, DEFINITION, /country, /isolation_source or
/note=*isolate_source, /collection_date and length in LOCUS line.

positional arguments:
  genbank_file          Genbank file with one or more items ;

optional arguments:
  -h, --help            show this help message and exit
  --output_file OUTPUT_FILE, -o OUTPUT_FILE
                        Output file where to write the table. Default : stdout
                        ;
```

Dans le cas de HPV42, un nouveau fichier a été créé (hpv42_with_carcinomes.csv). En effet, Nacho a récupéré des génomes complets. Ainsi, les "Accessions" de ces génomes ont été répertoriés.

III. Récupération des séquences des génomes complets

Les données sont dans le dossier "sequences/hpv/raw_seqs/genomes" au format ".fasta"

Afin de récupérer les séquences des génomes complets, il faut aller sur NCBI et taper dans la barre de recherche :

```
"Human papillomavirus type 42"[ORGN] 7500:15000[SLEN]
```

Pour SLEN, cet intervalle de longueur a été choisi afin de ne récupérer que les génomes complets.

Du résultats de recherche a été récupéré les génomes au format FASTA en cliquant sur *Send to --> Complete Record --> File --> FASTA --> Create File*

Pour HPV11, il est important de noter que le génome JN644142 a été exclut des génomes complet car sa taille beaucoup trop longue (plus de 10 000 bases dû à une grande duplication) pouvait fausser les alignements. Ce qui a été vérifié par la suite et confirmé.

Pour ce qui est des génomes HPV42 issus de carcinomes, ils n'ont pas été récupérés sur NCBI car non publié mais donnés par Nacho.

Pour HPV6, tout comme les métadonnées, des recherches supplémentaires avec les différents types ont été lancées et les résultats concaténés.

Dans le cas de HPV13, les 4 génomes issus de chimpanzés et bonobos ont été ajoutés.

HPV	Nombre génomes
42	32 + 14 + 16 + 3
11	101
6	284
13/Chimp&Bono	3+4
44/55	7+2
74	7

IV. Uniformisation des génomes complets

Les données sont dans le dossier "sequences/hpv/treated_seqs/rotate" au format ".fasta"

Afin de faciliter l'alignement, les génomes complets ont tout d'abord été uniformisés. En effet, avec le pipeline ViRotator, ils ont tous été rotate de telle sorte à ce qu'ils soient tous sur le brin sens et commencent par E6.

Donc, le génomes complets et le gène de E6 de la référence répertorié dans la base de donnée PaVE a été utilisée :

- M73236 pour HPV42
- M14119 pour HPV11
- X00203 pour HPV6
- X62843 pour HPV13
- U31788 pour HPV44 et 55
- AF436130 pour HPV74

Les références ont été mises dans le dossier "data" de ViRotator. Attention, l'outil utilise Blast.

```
mkdir /media/sarahb/Transcend2/HPV_EPIMOL/sequences/hpv42/treated_seqs

./virotator.sh -f
/media/sarahb/Transcend2/HPV_EPIMOL/sequences/hpv42/raw_seqs/genomes -o
/media/sarahb/Transcend2/HPV_EPIMOL/sequences/hpv42/treated_seqs/rotate -d
HPV42
```

Concernant HPV42, une fois l'uniformisation réalisée, les génomes uniformisés issus de carcinomes et ceux issus de NCBI ont été concaténés ensemble.

HPV	Nombre génomes uniformisés
42	32 + 14 + 16 + 3
11	101
6	284
13/Chimp&Bono	3+4
44/55	7+2
74	7

V. Alignement des génomes complets

Les données sont dans le dossier "sequences/hpv/treated_seqs/aligned" au format ".fasta"

Les génomes complets uniformisés ont été alignés à l'aide de MAFFT 7.505 avec comme ligne de commande :

```
mafft --thread 8 --auto
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_rotated.fasta >
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_aligned.fasta
```

Concernant l'alignement des génomes complets HPV42, la fin de l'alignement a été clippée, c'est-à-dire qu'une portion a été enlevée, car il n'y avait que des N pour une séquence et des gaps pour les autres. Donc, pour la suite, seul l'alignement clippé a été utilisé.

VI. Récupération des portions de séquences

Les données sont dans le dossier "sequences/hpv/raw_seqs/portions" au format ".fasta"

Sur NCBI, il a été demandé :

```
"Human papillomavirus type 42"[ORGN] 100:7500[SLEN]
```

Du résultats de recherche a été récupéré les portions de séquences au format FASTA.

Pour HPV11, la séquence L36108 a été splitté en deux, à savoir d'un côté L1-LCR et de l'autre E6-E7-E1. De la même façon pour éviter les "conflits" d'alignement, 12 séquences HPV6 dont l'accession commence par JN5731 ont été splittées en deux : LCR d'un côté, E6 de l'autre. Ainsi, il est important de noter que les fichiers de métadonnées ont été modifiés de telle sorte à incorporer ces changements.

Pour HPV6, tout comme les métadonnées, des recherches supplémentaires avec type 6a, 6b et 6vc ont été lancées et seules certaines séquences ont été choisies par Nacho.

HPV	Nombre portions de séquences
42	185
11	635 (une accession avec 2 séquences)

HPV	Nombre portions de séquences
6	1194 (12 accessions avec 2 séquences)
13	11
44/55	65+103
74	10

VII. Uniformisation des portions de séquences

Les données sont dans le dossier "sequences/hpv/treated_seqs/reverse-complemented" au format ".fasta"

Afin de faciliter l'alignement des portions de séquences contre les génomes complets alignés et par la suite construite l'arbre phylogénétique, pour chaque portion de séquence le brin a été regardé avec un Blastn contre la référence du type d'HPV.

```
blastn -query
/media/sarahb/Transcend1/HPV_EPIMOL/sequences/hpv42/raw_seqs/portions/hpv42_por
tions_sequences.fasta -subject "$ref_genome_for_strand_hpv42" -outfmt "6 qseqid
sstrand" >
/media/sarahb/Transcend1/HPV_EPIMOL/sequences/hpv42/treated_seqs/reverse-
complemented/hpv42_portions_sequences_strand.txt
```

Puis, les séquences qui étaient identifiées sur le brin minus ont été reverse-complement avec le script python `rev-comp.py` de ViRotator.

```
python3 "$rev_comp_script" --input_file
/media/sarahb/Transcend1/HPV_EPIMOL/sequences/hpv42/raw_seqs/portions/hpv42_por
tions_sequences.fasta --file_type fasta --blast_file
/media/sarahb/Transcend1/HPV_EPIMOL/sequences/hpv42/treated_seqs/reverse-
complemented/hpv42_portions_sequences_strand.txt --output_file
/media/sarahb/Transcend1/HPV_EPIMOL/sequences/hpv42/treated_seqs/reverse-
complemented/hpv42_portions_sequences_rc.fasta --output_log
/media/sarahb/Transcend1/HPV_EPIMOL/sequences/hpv42/treated_seqs/reverse-
complemented/hpv42_portions_sequences_rc.log
```

Concernant les séquence où le brin n'a pas pu être déterminé, elles n'ont pas été conservées. Les voici :

```
# HPV42
*---------- Sequences reverse complemented (2) :
MW546447.1
MW546446.1

*---------- Sequences where strand was not find (1) :
MH253555.1

# HPV11
*---------- Sequences reverse complemented (0) :
```

```
*---------- Sequences where strand was not find (6) :
MH253524.1
MH253517.1
MH253513.1
AF548814.1
JC286052.1
EF140813.1

# HPV6
*---------- Sequences reverse complemented (7) :
MT793790.1
MW546456.1
MW546455.1
MW546454.1
MW546453.1
KC706454.1
KC706449.1

*---------- Sequences where strand was not find (6) :
S72322.1
S72317.1
L36842.1
L36840.1
L36837.1
L36839.1

# HPV13
*---------- Sequences reverse complemented (0) :


*---------- Sequences where strand was not find (0) :

# HPV44-55
*---------- Sequences reverse complemented (0) :


*---------- Sequences where strand was not find (0) :

# HPV74
*---------- Sequences reverse complemented (0) :


*---------- Sequences where strand was not find (0) :
```

HPV	Nombre portions brin identifié
42	184
11	628
6	1188
13	11

HPV	Nombre portions brin identifié
44/55	65+103
74	10

IIX. Alignement des portions de séquences contre les génomes uniformisés et alignés

Les données sont dans le dossier "sequences/hpv/treated_seqs/aligned" au format ".fasta"

Pour réaliser cet alignement, tout comme l'alignement des génomes complets, MAFFT a été utilisé avec comme ligne de commande :

```
mafft --thread 8 --inputorder --maxiterate 1000 --localpair --addfragments
$path_to_tmp/hpv42_portions_sequences_rc.fasta
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_aligned_clipped.fasta >
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_clipped_and_portions_se
quences_aligned.fasta
```

Dans le cas d'HPV6 après visualisation de l'alignement, il a été constaté pour l'accession AF126428, qui est caractérisée comme étant seulement le CDS de E6, qu'il y avait la partie LCR. Ainsi, au vu de la caractérisation de la séquence, seule la partie E6 a été conservée dans l'alignement (nouveau fichier "clipped"). La partie LCR, si c'est bien elle, a été croppée.

IX. Construction des arbres phylogénétiques pour les génomes complets

Les données sont dans le dossier "trees/hpv/complete_genomes_tree" et "trees/hpv/complete_genomes_light_trees" (pour HPV6)

Afin de construire les arbres, la version 1.1.0 de raxml-ng a été utilisé. Les arbres ont été réalisés à partir des génomes complets alignés. Voici la ligne de commande :

```
raxml-ng --all --msa
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_aligned_clipped.fasta -
-model GTR+G+I --bs-trees autoMRE{1000} --bs-cutoff 0.01 --prefix
hpv42_complete_genomes_NCBI_and_carcinomes_aligned_clipped --seed 17
```

Dans le cas de HPV6, des génomes ont été supprimés du fichier d'alignement avant construction de l'arbre : MK463906, MK463907, MK463908. Ces génomes ont été enlevés car très différents des autres en termes d'acides aminés.

Aussi, après construction de l'arbre, au vu de la densité du nombre de génomes complets dans HPV6, deux arbres plus light ont été fait à partir d'une sélection réalisée par Nacho après visualisation de l'arbre complet. L'un des deux comportent le génome OL854081 et l'autre non. Ceci a été réalisé car cette séquence était basale et l'on voulait voir la différence topologique avec et sans celle-ci.

Pour la suite des analyses, l'arbre allégé est gardé car cela créé moins d'ambiguité. Aussi, sans le génome OL854081 car il donnait un mauvais racinement du groupe B.

Pour HPV11, un arbre supplémentaire a été réalisé sans les génomes MK463919 et MK463922 afin que le "midpoint" d'enracinement soit "correct". Ce dernier se trouve dans le dossier des analyses supplémentaires.

X. Génération des noms de noeuds des arbres

Les données sont dans le dossier "trees/hpv/nodes_names"

Pour avoir des noms de noeuds, la version 8.2.9 de raxmlHPC-PTHREADS a servi. Les génomes complets + les portions alignées ainsi que les arbres des génomes complets ont été utilisés avec l'option epa de l'outil.

Attention à la personne qui voudra refaire les analyses, souvenez-vous que RAxML ne prends que les paths absolus et non relatifs. Aussi, il faut bien mettre l'extension .nwk à l'arbre donné en entrée. Donc, préalablement l'arbre avec les génomes complets a été renommé de telle sorte à ce qu'il se termine par la bonne extension.

Les fichiers d'intérêts de sortie avec les noms des noeuds et des feuilles sont ceux avec le mot "labelled".

```
mv
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_aligned_clipped.raxml.b
estTree
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_aligned_clipped.nwk

raxmlHPC-PTHREADS -T 8 -f v -G 0.1 -m GTRCAT --epa-keep-placements=100 -t
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_aligned_clipped.nwk -s
$path_to_tmp/hpv42_complete_genomes_NCBI_and_carcinomes_clipped_and_portions_se
quences_aligned.fasta -n hpv42 -w $path_to_tmp/nodes_names/hpv42
```

XI. Délimitation des groupes phylogénétiques

Les données sont dans le dossier "trees/hpv/groups"

Pour définir les groupes phylogénétiques, cela a été fait par Nacho. Les groupes définis sont caractérisés par un code couleur. Ceci se retrouve dans le fichier se terminant par "groups_defined.pdf".

Une fois les groupes définis, pour récupérer les noms des noeuds et feuilles et ainsi créer un fichier de correspondance groupes - noeuds/feuilles, le fichier "originalLabelledTree" a été utilisé.

Dans VScode :

- tout ce qui est entre `:` et `[` (regex : `:(.*?)\[`) à été remplacé par `[`. Donc au lieu d'avoir Accession:Vraisemblance[INuméro], ca devient Accession[INuméro]

Dans FigTree :

- l'arbre avec Accession[INuméro] a été importé
- sélectionner *Branch Labels* --> *Display: label*. Tous les I sont maintenant visibles
- reroot pour avoir le même arbre que celui avec les groupes définis
- sélectionner *Clade* puis le sous arbre d'intérêt --> Ctrl-C / Ctrl-V dans un éditeur de texte

A partir de chacun des sous arbres au format Nexus, le fichier texte de correspondance groupes - noeuds/feuilles a été créé en prenant en compte les I. Attention, dans le format Nexus certains I n'apparaissent pas, ceux à la base de deux sous arbres. Donc, bien vérifier.

XII. Caractérisation de l'appartenance phylogénétique des portions de séquences

Les données sont dans le dossier "genotypes/hpv"

Afin d'obtenir l'appartenance phylogénétique des portions de séquences, le script `genotype.py` de Virophylo a été utilisé. Celui prends en entrée le fichier de sortie de RAxML avec les vraisemblances.

Le script a été modifié de telle sorte qu'en sortie de fichiers il y en ait un avec une colonne contenant le meilleur groupe d'appartenance et non les valeurs de vraisemblance.

```
mkdir -p /media/sarahb/Transcend2/HPV_EPIMOL/genotypes
mkdir /media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42

./genotype.py --raxml_file
/media/sarahb/Transcend2/HPV_EPIMOL/trees/hpv42/nodes_names/RAxML_classificatio
nLikelihoodWeights.hpv42 --groups_file
/media/sarahb/Transcend2/HPV_EPIMOL/trees/hpv42/groups/hpv42_groups.txt --
output_file
/media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42/hpv42_portions_genotype.txt
--output_file2
/media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42/hpv42_portions_bestgroup.tx
t
```

XIII. Récupération de l'appartenance phylogénétique des génomes

Les données sont dans le dossier "genotypes/hpv"

Pour ce qui est des génomes complets, un fichier artefactuel similaire à "RAxML_classificationLikelihoodWeights" été créé à partir de la sortie RAxML "RAxML_originalLabelledTree" en utilisant comme lignes de commande python :

```
>python3

# Open the input file with the tree which contains the labelled nodes and leafs
with
open("/media/sarahb/Transcend2/HPV_EPIMOL/trees/hpv42/nodes_names/RAxML_origina
lLabelledTree.hpv42", "r") as f:
    for line in f: # NB : normally there only one line in the input file
        # Replace "]",")","(",";","\n" and create a list where the line is
separeted by ","
        l = line.replace("]","").replace(")","").replace("
(","").replace(";","").replace("\n","").split(",")


# Open the output file will created
```

```
with
open("/media/sarahb/Transcend2/HPV_EPIMOL/trees/hpv42/nodes_names/RAxML_classif
icationLikelihoodWeights.artefact_complete_genomes.hpv42", "w") as f:
    # For each element of the list created before
    for item in l:
        # Write into the output file : seqId, first labelled name (i.e. I243),
1 and 1
        print(item.split(":")[0], item.split("[")[1].partition(":")[0], 1, 1,
file=f)
```

Puis, le script `genotype.py` a été lancé sur le fichier de sortie nouvellement créé.

```
./genotype.py --raxml_file
/media/sarahb/Transcend2/HPV_EPIMOL/trees/hpv42/nodes_names/RAxML_classificatio
nLikelihoodWeights.artefact_complete_genomes.hpv42 --groups_file
/media/sarahb/Transcend2/HPV_EPIMOL/trees/hpv42/groups/hpv42_groups.txt --
output_file
/media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42/hpv42_complete_genomes_geno
type.txt --output_file2
/media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42/hpv42_complete_genomes_best
group.txt
```

XIV. Concaténation génotypage et métadonnées

Une fois l'appartenance phylogénétique établie, métadonnées et appartenances phylogénétiques ont été concaténées.

Pour cela, tout d'abord les fichiers de résultats de génotypage "bestgroup" des génomes complets et des portions de séquences ont été concaténés en un seul fichier. Il en est de même pour les résultats de génotypage. Il est très important de noter que dans ces nouveaux fichiers, les ".1" des accessions ont été enlevés pour pouvoir matcher par la suite avec les accessions des métadonnées.

Puis la concaténation a été réalisée sous R :

```
library("stringr")

# Import files with best group and  genotype for complete genomes and portions
# NB : in this files pattern ".1" was removed
hpv42_bestgroup <-
read.delim("/media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42/hpv42_complete_
genomes_and_portions_bestgroup.txt", header=FALSE)
hpv42_genotype <-
read.delim("/media/sarahb/Transcend2/HPV_EPIMOL/genotypes/hpv42/hpv42_complete_
genomes_and_portions_genotype.txt", header=TRUE)
# Rename column
colnames(hpv42_bestgroup) <- c("ACCESSION", "BEST GROUP")
colnames(hpv42_genotype) <- c("ACCESSION", "LIKELIHOOD I", "LIKELIHOOD II",
"LIKELIHOOD IIa", "LIKELIHOOD IIb",  "LIKELIHOOD IIb1",  "LIKELIHOOD IIb2")
# Remove word group into bestgroup table
hpv42_bestgroup$`BEST GROUP` <- str_remove(hpv42_bestgroup$`BEST GROUP`,
"group")
# Remove whitespaces column ACCESSION
hpv42_bestgroup$ACCESSION <- str_remove(hpv42_bestgroup$ACCESSION, " ")
```

```
hpv42_genotype$ACCESSION <- str_remove(hpv42_genotype$ACCESSION, " ")

# Import file with metadata for complete genomes and portions
# NB : data from carcinomes was add manually
hpv42_metadata <-
read.delim("/media/sarahb/Transcend2/HPV_EPIMOL/metadata/hpv42_with_carcinomes.
csv")
hpv42_metadata$ACCESSION <- str_remove(hpv42_metadata$ACCESSION, " ")

# Merge metadata and group
hpv42_metadata_and_bestgroup <- merge(hpv42_metadata, hpv42_bestgroup, by=
"ACCESSION", all = FALSE)
hpv42_metadata_and_genotype <- merge(hpv42_metadata_and_bestgroup,
hpv42_genotype, by= "ACCESSION", all = FALSE)

# Write datas merged into an output file
write.table(hpv42_metadata_and_genotype,
"/media/sarahb/Transcend2/HPV_EPIMOL/hpv42_metadata_and_genotype.csv", sep =
"\t", row.names=FALSE, fileEncoding='UTF-8')
```

Pour HPV6, le tableau dans les données supplémentaires de l'article Jelen 2014 (Global Genomic Diversity of Human Papillomavirus 6 Based on 724 Isolates and 190 Complete Genome Sequences) a été ajouté au fichier avec métadonnées et génotypes.

Pour HPV11, deux génomes issus de HPV-Capture ont été ajouté (MPL-9 et MPL-11). L'uniformisation et le génotypage ont été réalisé dans un autre projet contre. L'arbre qui a été utilisé pour génotyper ces séquences est celui de ce projet. Les résultats sont dans le dossier d'analyses supp' sous le nom taxonomy_with_virophylo_hpv11_capture.

Une fois tout réalisé, les tableaux finaux ont été mis en forme à la main, c'est-à-dire que les espaces supplémentaires ont été supprimés, la police caractérisée, etc.

XV. Construction d'un arbre pour les alpha-papillomavirus

Les données sont dans le dossier "sequences/alpha" pour les séquences sélectionnées et dans "tree/alpha" pour l'arbre construit

Pour construire l'arbre des alpha-papillomavirus, pour chacun des types d'HPV deux représentants de chaque groupe ont été choisi. De là, pour chaque HPV, un nouveau fichier avec les représentants alignés ont été créés.

En ligne, le feature merge de MAFFT a été utilisé. Puis, un arbre a été créé sur les représentants alignés avec RAxML

```
raxml-ng --all --msa alpha_genomes_aligned_selected_mafft_merge.fasta --model
GTR+G+I --bs-trees autoMRE{1000} --bs-cutoff 0.01 --prefix
alpha_genomes_aligned_selected_mafft_merge --seed 17
```

XVI. Consolidation des pays par zone géographique

Les tableaux de résultats avec métadonnées et génotypes contiennent pour chacune des ACCESSION, quand cela est indiqué, un pays dans la colonne Country. Afin de renforcer les analyses statistiques, il a été décidé de consolider les pays par région géographique caractérisée selon l'OMS (https://en.wikipedia.org/wiki/List_of_WHO_regions). Attention, USA et Hong Kong ont été rajouté à la liste. Aussi :

Algeria >> Eastern Mediterranean
Eritrea >> Eastern Mediterranean
Israel >> Eastern Mediterranean
North Corea >> Western Pacific Region

Ainsi, un script python `get_region.py` a été développé pour que la consolidation soit fait de façon automatisée. Celui-ci à partir d'un fichier résultat renvoie un fichier de sortie avec ACCESSION et CONSOLIDATED COUNTRY. Il faut noter que si le pays n'est pas renseigné dans le dictionnaire des régions du script, il est renvoyé "NI" pour "Not Identified". Aussi, NA est conservé si le pays est NA.

```
./get_region.py -m ../hpv42_metadata_and_genotype.csv -o ../hpv42_region.csv
```

Par la suite, fichier de sortie et fichier résultats métadonnées - génotypes ont été mergés sous R avec comme ligne de commande ci-dessous. Attention, il faut bien que le premier espace de la colonne COUNTRY ait bien été supprimé dans le fichier metadata_and_genotype !

```
hpv42_metadata_and_genotype <-
read.delim("/media/sarahb/Transcend2/HPV_EPIMOL/hpv42_metadata_and_genotype.csv
")
hpv42_region <-
read.delim("/media/sarahb/Transcend2/HPV_EPIMOL/hpv42_region.csv")

hpv42_metadata_and_genotype_region <- merge(hpv42_metadata_and_genotype,
hpv42_region, by= "ACCESSION", all = TRUE, sort = FALSE)

write.table(hpv42_metadata_and_genotype_region,
"/media/sarahb/Transcend2/HPV_EPIMOL/hpv42_metadata_geno_geo.csv", sep = "\t",
row.names=FALSE, fileEncoding='UTF-8')
```

Dans un soucis de ne pas cumuler des fichiers avec des informations redondantes, les fichier *hpv_region.csv* et *hpv42_metadata_and_genotype.csv* ont été supprimés.