Projet VCF – HMIN113M

Anna-Sophie Fiston-Lavier

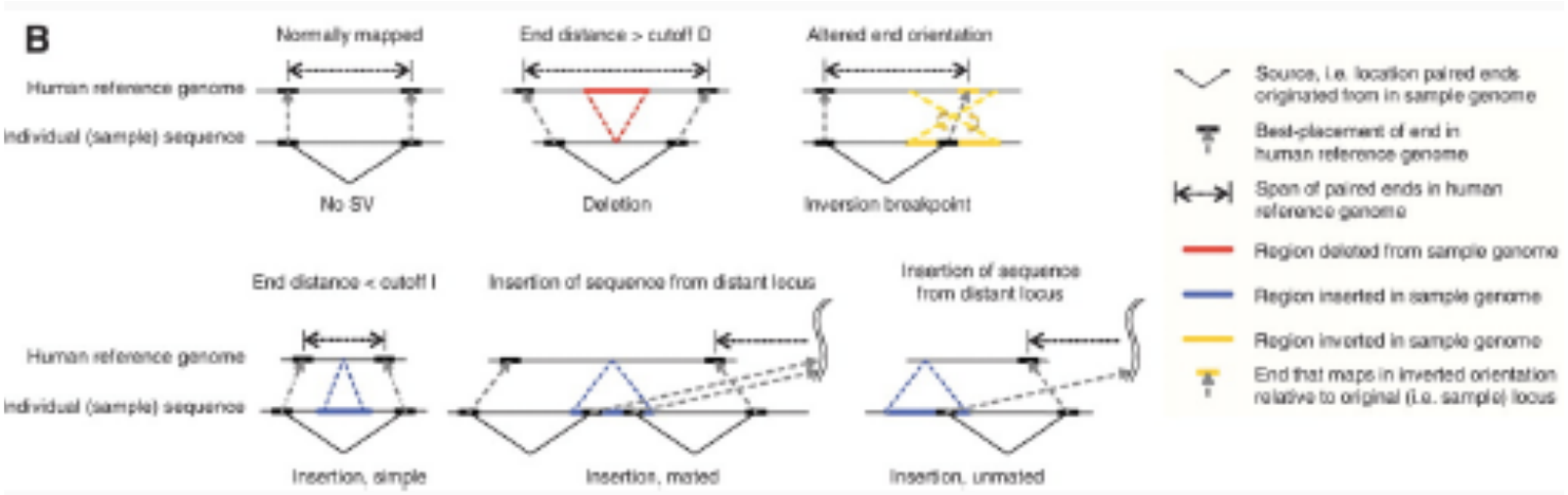
Projet VCF

- Dans le cadre de l'UE HMIN113M, nous vous proposons de réaliser un projet d'analyse et visualisation de données VCF.

- Les données VCF sont des données issues de l'analyse de données de séquençage à haut-débit (NGS).

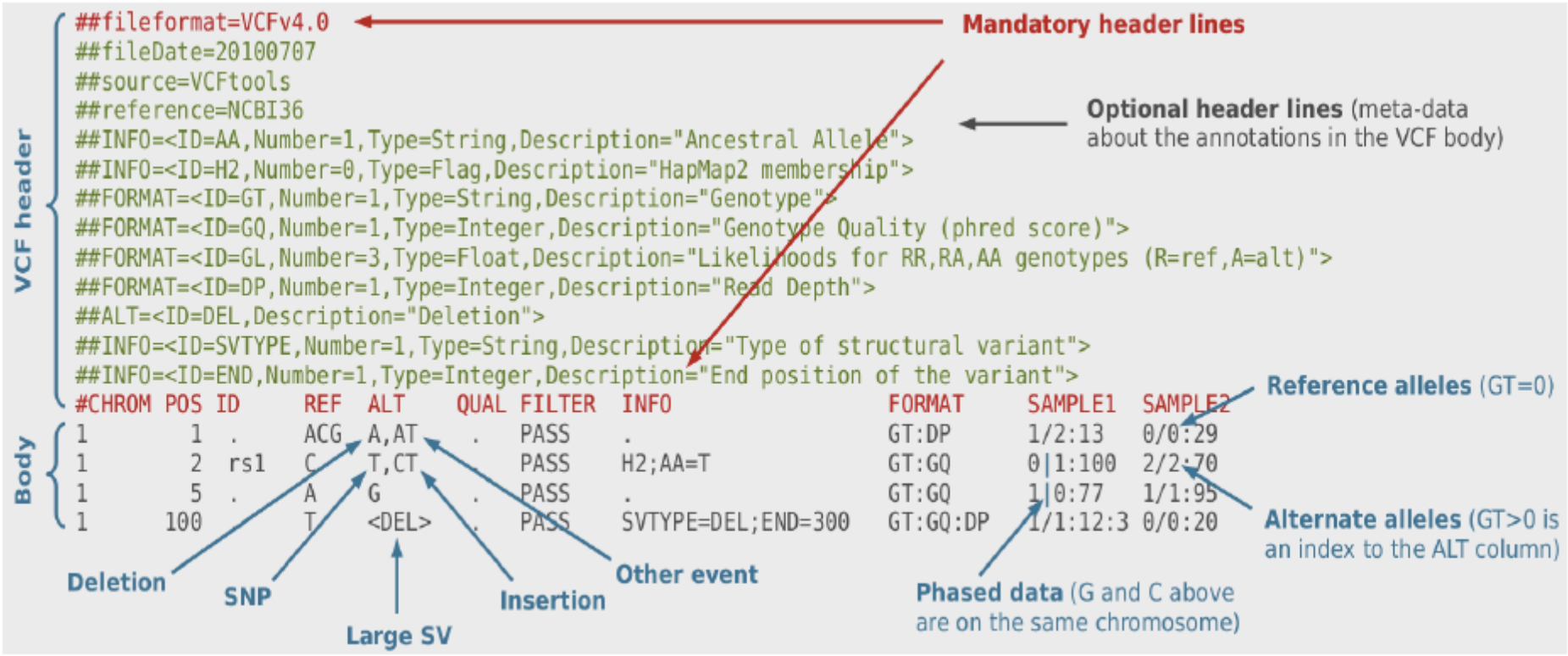
- Un VCF répertorie les variants structuraux (SV ou structural variants) détectés après mapping (alignement) de données NGS (reads ou lectures) sur une séquence de référence. Par exemple, le mapping de reads d'un individu malade sur la séquence d'un individu sain va servir à identifier les variations génomiques responsables de la maladie.

Les variants structuraux (SV)



Korbel et al 2007

Le fichier VCF



Comment le comprendre

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2  2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

- **Variation 1 :** a good SNP
- **Variation 2 :** a possible SNP that has been filtered out because its quality is below 10
- **Variation 3 :** a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error)
- **Variation 4 :** a site that is called monomorphic reference (i.e. with no alternate alleles)
- **Variation 5 :** a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T)

Les attentes

- Votre script python devra impérativement contenir les fonctionnalités « compulsives ». Les fonctionnalités optionnelles donneront lieu à des points de bonus.

- Les « compulsives »: dans votre script, il vous est demandé de:
  1. Vérification des données d'entrée (toujours)
  2. Ouvrir le fichier type vcf
  3. Stockage des informations sur les différents variants (type, position …) dans des dictionnaires
  4. Création d'une interface en Python CGI afin de permettre à l'utilisateur de choisir un type d'analyse (analyse de tous les variants ou d'un type de variant)
  5. Affichage graphique (tableau ou plot) des analyses

Les bonus

- Vous serez notés sur:
  1. La qualité du code
  2. L'analyse biologique des données
  3. La qualité graphique de l'interface
  4. La portabilité
  5. Les commentaires du code

- Des points de bonus seront attribués pour des fonctionnalités supplémentaires aussi bien au niveau de:
  1. L'analyse originale des données (analyse plus poussée comme la combinaison de plusieurs VCF, filtre sur la qualité, représentation des SV le long des chromosomes….)
  2. Les étapes de vérification et control
  3. Interactivité du programme (l'utilisateur suit les différentes étapes)