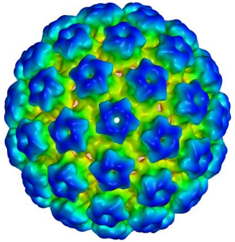
Bioinformatic workflow to detect viral infections in tumors

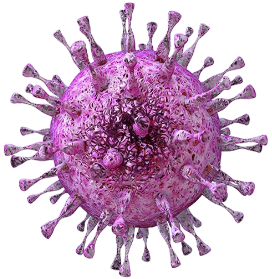
Marion DUFEU, Ombeline TRANCART,
Albane FLOCON, Léa LE LARGE

Context

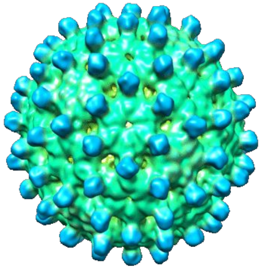
Approximately **10 %** of worldwide cancers are attributable to viral infection



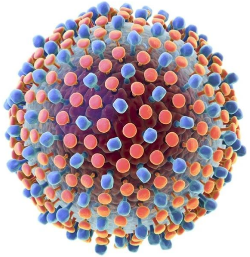
Human Papillomavirus (**HPV**)
dsDNA



Epstein Barr Virus (**EBV**)
dsDNA



Hepatitis B Virus (**HBV**)
dsDNA



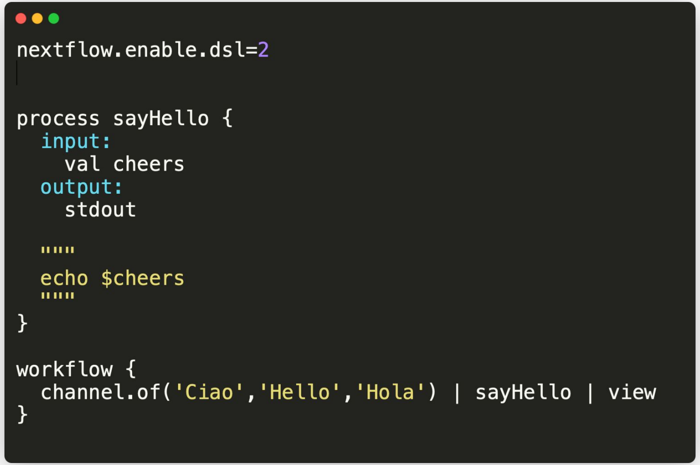
Hepatitis C Virus (**HCV**)
ssRNA

Cervical cancer

Hepatocellular cancer

=> Kaposi's sarcoma-associated herpes virus (**KSHV**), human T-cell leukemia virus (**HTLV-I**), and Merkel cell polyomavirus (**MCPyV**)...

Why Nextflow?



```
nextflow.enable.dsl=2


process sayHello {
  input:
    val cheers
  output:
    stdout

  """
  echo $cheers
  """
}

workflow {
  channel.of('Ciao','Hello','Hola') | sayHello | view
}
```

Automatization & Parallelization

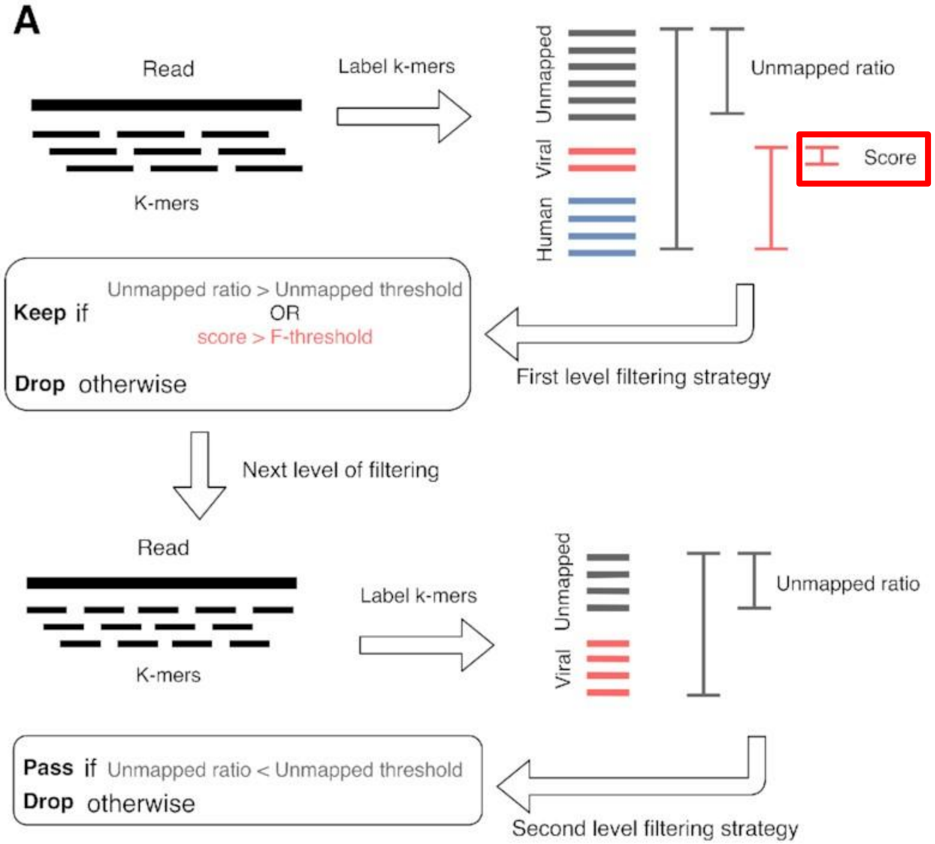
FastViFi

Detection of (Hybrid) Viral DNA and RNA, relies on **ViFi** and **Kraken** tools.

Speed
Sensitivity

Reference paper: Javadzadeh, Sara, Utkrisht Rajkumar, Nam Nguyen, Shahab Sarmashghi, Jens Luebeck, Jingbo Shang, et Vineet Bafna. « FastViFi: Fast and accurate detection of (Hybrid) Viral DNA and RNA ». *NAR Genomics and Bioinformatics* 4, n° 2 (1 juin 2022): lqac032. https://doi.org/10.1093/nargab/lqac032.

FastViFi


A
Read
K-mers
Label k-mers
Unmapped
Viral
Human
Unmapped ratio
Score
Keep if
Unmapped ratio > Unmapped threshold
OR
score > F-threshold
Drop otherwise
First level filtering strategy
Next level of filtering
Read
K-mers
Label k-mers
Unmapped
Viral
Unmapped ratio
Pass if Unmapped ratio < Unmapped threshold
Drop otherwise
Second level filtering strategy

FastViFi

Number of kmers labeled as viral

Read

$$\text{v-score}(r, k_1) = \frac{n_v(r, k_1)}{n_v(r, k_1) + n_h(r, k_1)} \geq t_1$$

kmer length

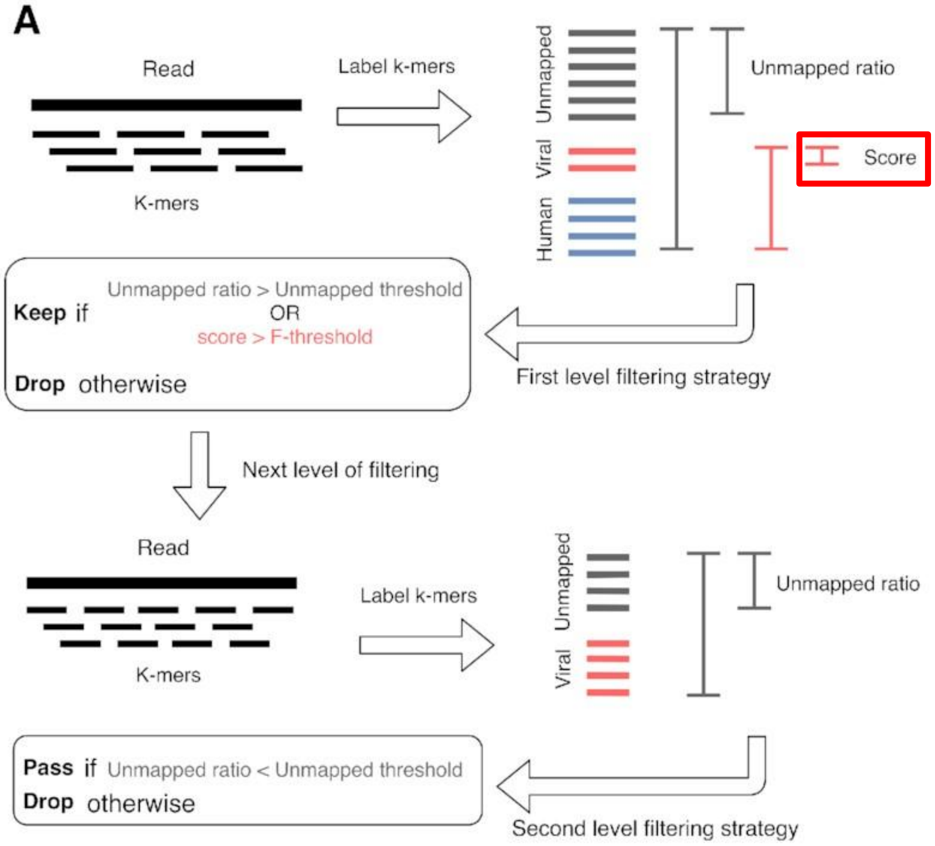
Number of kmers labeled as human

AND

Number of unmapped kmers

$$\frac{n_u(r, k_1)}{n_u(r, k_1) + n_v(r, k_1) + n_h(r, k_1)} \geq u_1$$

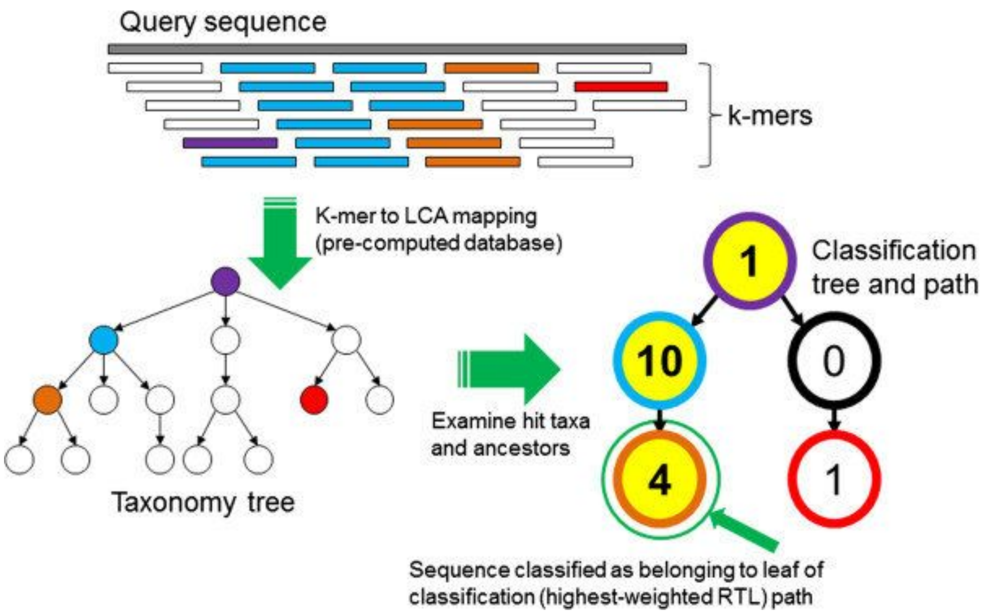
FastViFi


A
Read
K-mers
Label k-mers
Unmapped
Viral
Human
Unmapped ratio
Score
Keep if Unmapped ratio > Unmapped threshold
OR
score > F-threshold
Drop otherwise
First level filtering strategy
Next level of filtering
Read
K-mers
Label k-mers
Unmapped
Viral
Unmapped ratio
Pass if Unmapped ratio < Unmapped threshold
Drop otherwise
Second level filtering strategy

Kraken dataset

Tool to estimate the k-mer composition of the viral reads.

- builds an index
- reports detected taxonomy label for each k-mer
- reports lowest common ancestor for k-mers that mapped to multiple nodes in the taxonomy tree
- reports k-mers that did not map

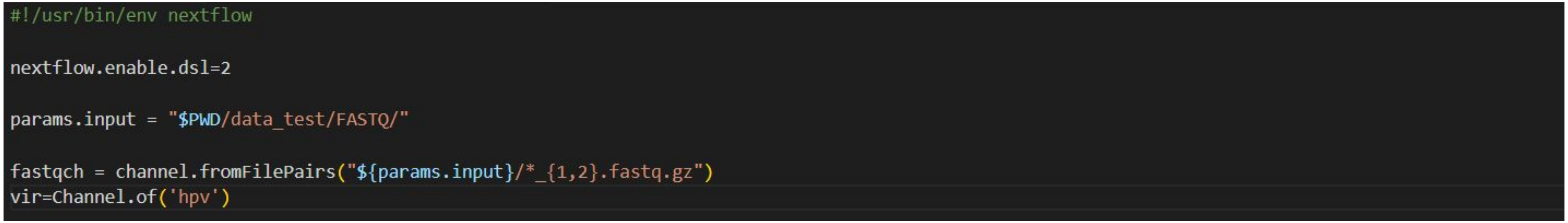

Query sequence
k-mers
K-mer to LCA mapping
(pre-computed database)
Taxonomy tree
Examine hit taxa and ancestors
1
10
0
4
1
Classification tree and path
Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

4 viral references in ViFi : HPV, HBV, HCV and EBV

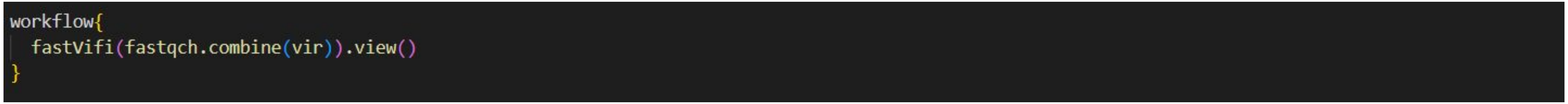
Kraken, a classification algorithm

LCA = Lowest Common Ancestor
RTL = Root To Leaf

Nextflow pipeline



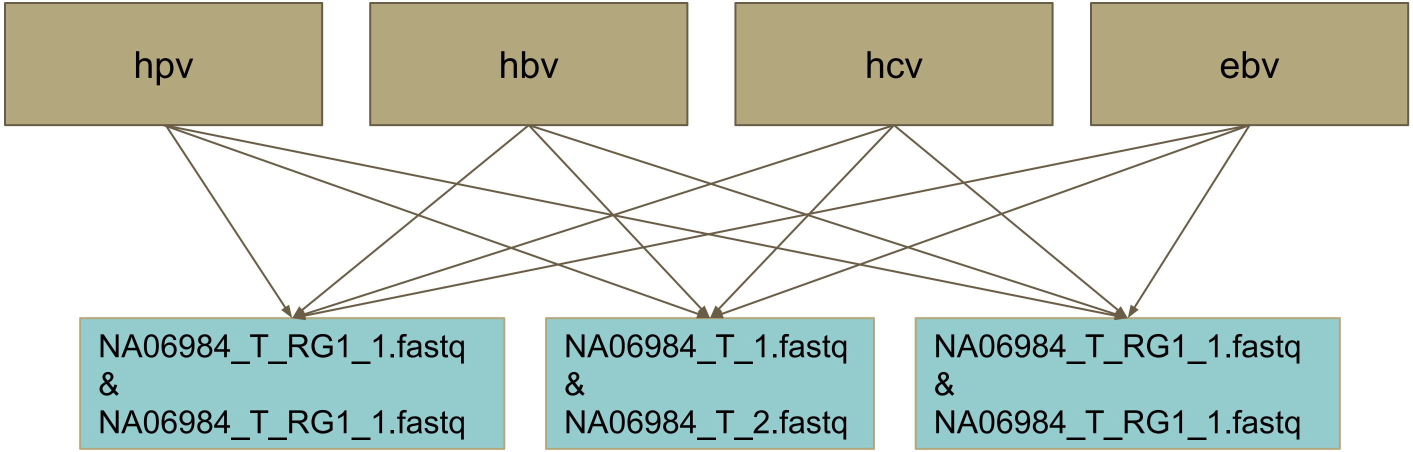
```
#!/usr/bin/env nextflow

nextflow.enable.dsl=2

params.input = "$PWD/data_test/FASTQ/"

fastqch = channel.fromFilePairs("${params.input}/*_{1,2}.fastq.gz")
vir=Channel.of('hpv')
```



```
workflow{
  fastVifi(fastqch.combine(vir)).view()
}
```

Nextflow pipeline

Combinations between viruses and Fastq files



hpv

hbv

hcv

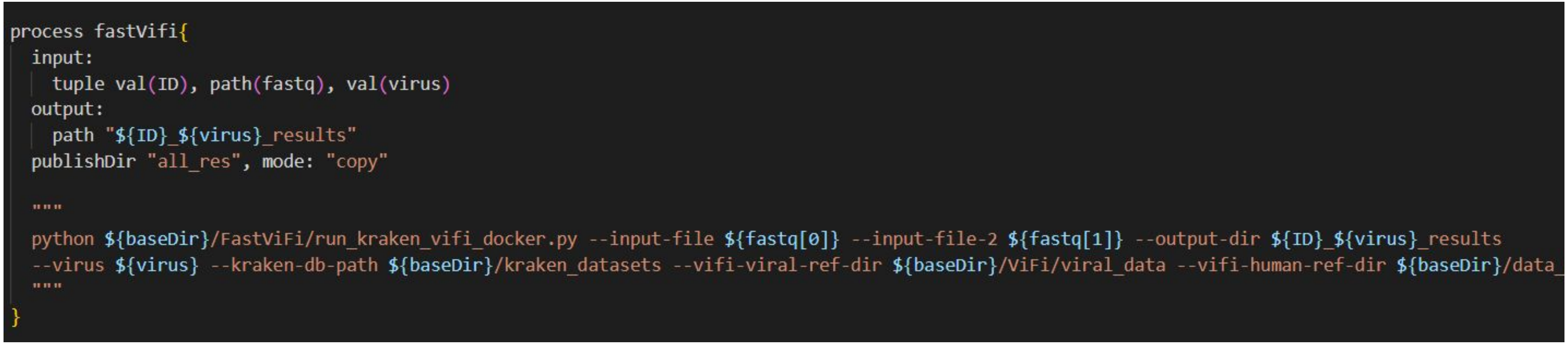
ebv

NA06984_T_RG1_1.fastq
&
NA06984_T_RG1_1.fastq

NA06984_T_1.fastq
&
NA06984_T_2.fastq

NA06984_T_RG1_1.fastq
&
NA06984_T_RG1_1.fastq

Nextflow pipeline



```
process fastVifi{
  input:
    tuple val(ID), path(fastq), val(virus)
  output:
    path "${ID}_${virus}_results"
  publishDir "all_res", mode: "copy"

  """
  python ${baseDir}/FastViFi/run_kraken_vifi_docker.py --input-file ${fastq[0]} --input-file-2 ${fastq[1]} --output-dir ${ID}_${virus}_results
  --virus ${virus} --kraken-db-path ${baseDir}/kraken_datasets --vifi-viral-ref-dir ${baseDir}/ViFi/viral_data --vifi-human-ref-dir ${baseDir}/data_
  """
}
```

Conclusion

- filtering strategy
- trade off between high sensitivity and speed
- utility of NextFlow shows greatly here