Internship

Automated HIV-1 Genotyping

Vera Rykalina

August 28, 2022

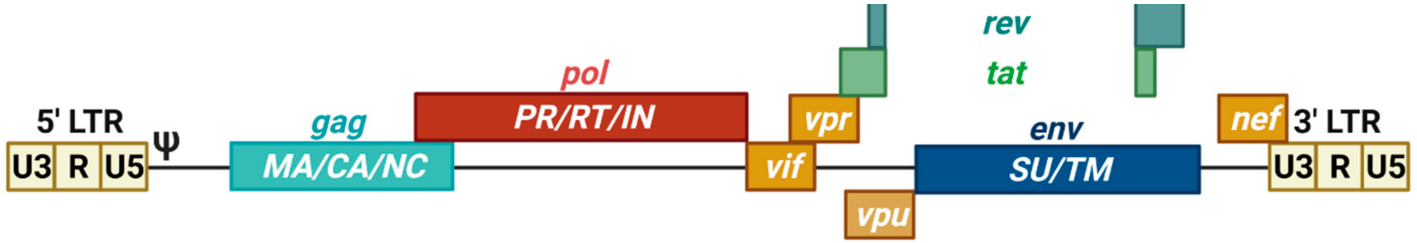
CQ Beratung&Bildung/Robert Koch Institute

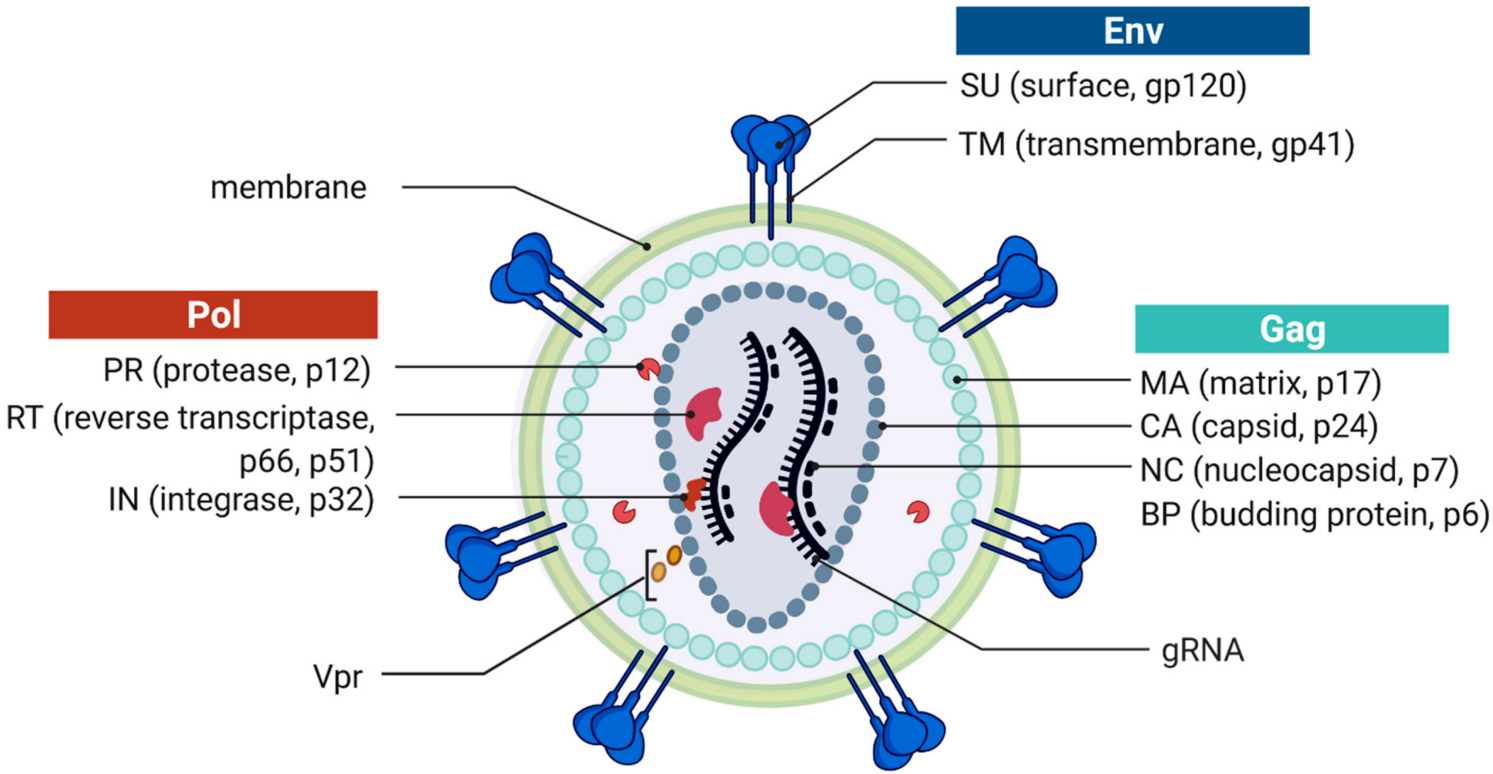
Outline

Outline

Introduction

INTRO: HIV-1 [1]


5' LTR
U3 R U5
Ψ
gag
MA/CA/NC
pol
PR/RT/IN
vif
vpr
vpu
rev
tat
env
SU/TM
nef
3' LTR
U3 R U5


Env
SU (surface, gp120)
TM (transmembrane, gp41)
membrane
Pol
PR (protease, p12)
RT (reverse transcriptase, p66, p51)
IN (integrase, p32)
Gag
MA (matrix, p17)
CA (capsid, p24)
NC (nucleocapsid, p7)
BP (budding protein, p6)
Vpr
gRNA

Introduction

Subtyping is performed by using a combination of the following tools.

Online Subtyping Tools

- STANFORD (sierrapy)
  https://hivdb.stanford.edu/hivdb/by-sequences/
- COMET (https requests)
  https://comet.lih.lu/
- REGA (manual)
  https://www.genomedetective.com/app/typingtool/hiv

Introduction

HIV-1 Groups

Human immunodeficiency virus (HIV) displays an extraordinary genetic diversity with four distinct groups:

- M (major)
- O (outlier)
- N (non M/O)
- P (putative)

Introduction

HIV-1 Subtypes

Within the major M Group are 9 different subtypes, at least 98 circulating recombinant forms (CRFs), and multiple unique recombinant forms (URF).

- A (A1, A2, A3, A4, A5, A6)
- B
- C
- D
- F (F1, F2)
- G
- H
- J
- K

http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html

Analysis

Analysis

SUBTYPING (Toolwise)

ID	PR, RT
1	A
2	B
...	...
n	F

ID	IN
1	A
2	B
...	...
n	F

ID	ENV
1	A
2	B
...	...
n	F

Analysis

SUBTYPING (Fragmentwise)

ID	REGA	STANFORD	COMET
1	B	B	B
2	G	CRF02_AG	G (check for 02_AG)
...	...	...	...
n	CRF11_cpx	A + J	CRF11_cpx

Customer's prefereces

```
$ nextflow ../Scripts/subtyping_pipeline.nf --outdir ../Results --run MS95
N E X T F L O W  ~  version 22.04.4
Launching `../Scripts/subtyping_pipeline.nf` [furious_hodgkin] DSL2 - revision: 333a2ff10d
executor >  local (23)
[cf/8c3339] process > mark_fasta (3)      [100%] 3 of 3 ✔
[81/ccdddc] process > stanford (1)        [100%] 3 of 3 ✔
[28/907f6f] process > json_to_csv (3)     [100%] 3 of 3 ✔
[84/234416] process > rega_to_csv (1)     [100%] 3 of 3 ✔
[6c/14640d] process > comet (2)           [100%] 3 of 3 ✔
[14/7da948] process > prrt_joint (1)      [100%] 1 of 1 ✔
[cf/83225a] process > env_joint (1)       [100%] 1 of 1 ✔
[2f/847e9f] process > int_joint (1)       [100%] 1 of 1 ✔
[15/761046] process > tags_to_csv (3)     [100%] 3 of 3 ✔
[66/8de6bb] process > decision_to_csv (1) [100%] 1 of 1 ✔
[b5/161910] process > full_joint          [100%] 1 of 1 ✔
[-        ] process > report              -
[-        ] process > phylo_fasta         -
```

Predicting sales volume

- Inability to specify product but product type.

Impact of Service and Customer Reviews

- Number of reviews is of great benefit to sales volume

TOOLS

- sierrapy, miller, mafft, iqtree

- DSL2

- sys, re, collections, requests pandas, time, wrap, json, bio

- GitHub

REPO

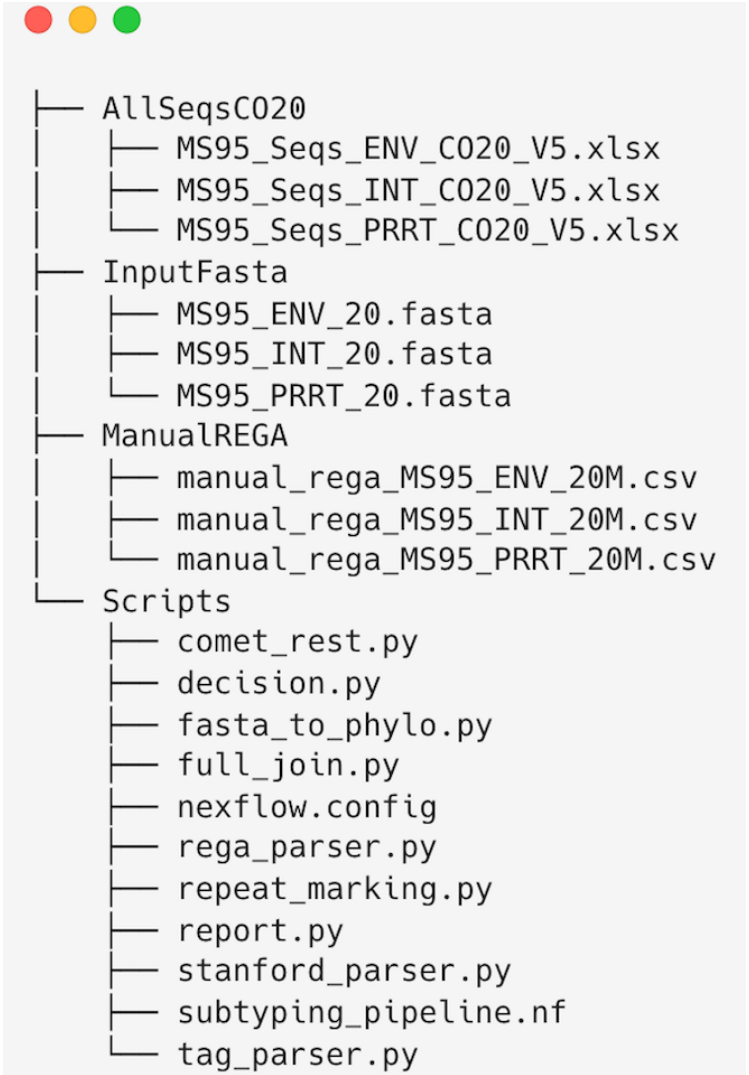
https://github.com/vera-rykalina/rki

Transactional Data

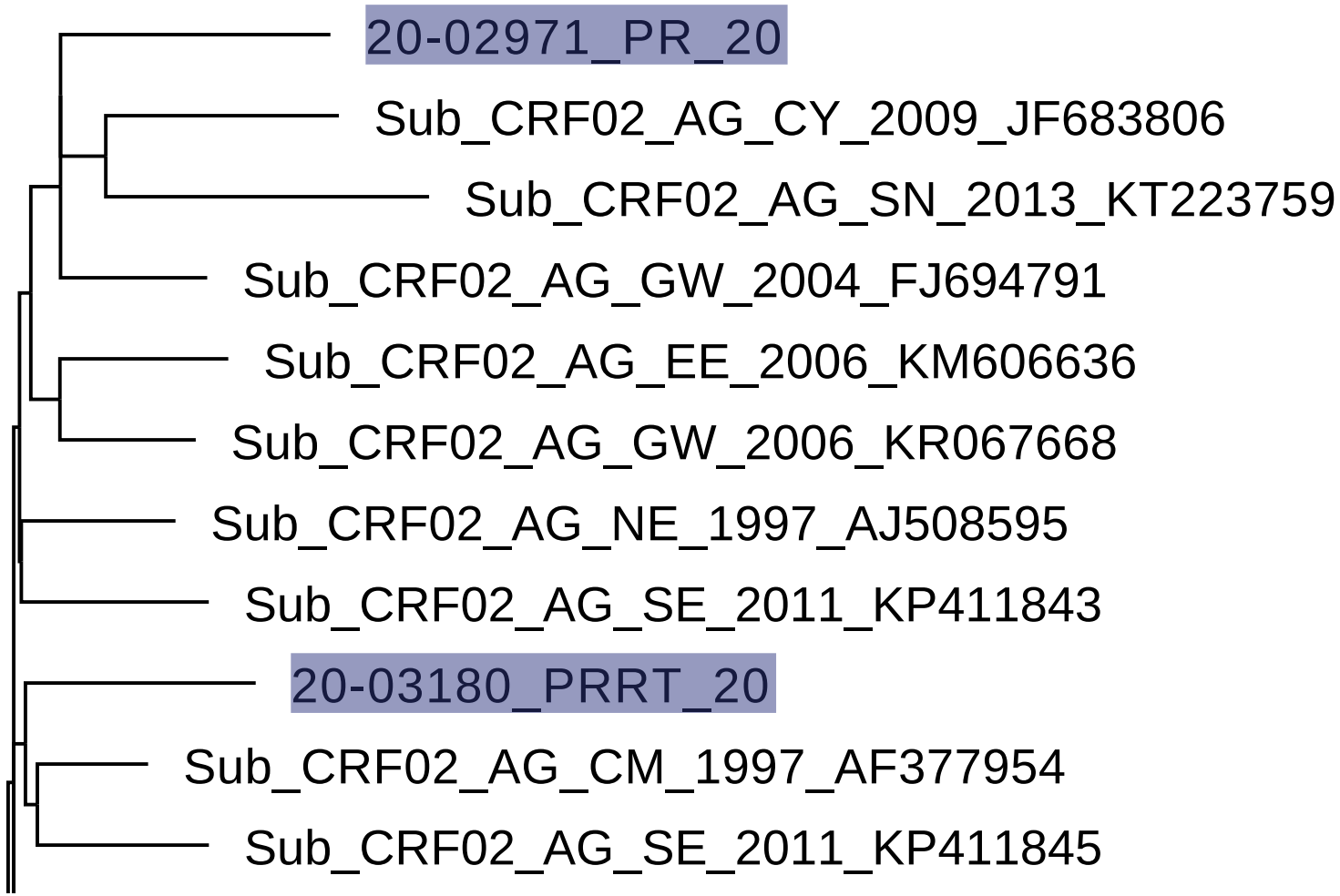
Acquisition

- Complementary profile of both companies (Electronidex is more focused on high end products and Blackwell on lower end).
- Blackwell could expand its customer database (9800 new clients)
- Low frequent products of Electronidex could be easily sold by Blackwell
- Due to financial difficulties Electronidex is a bargain
- Inability to directly collate Electronidex's and Blackwell's (5 product match only)

Carbon


├── AllSeqsCO20
│   ├── MS95_Seqs_ENV_CO20_V5.xlsx
│   ├── MS95_Seqs_INT_CO20_V5.xlsx
│   └── MS95_Seqs_PRRT_CO20_V5.xlsx
├── InputFasta
│   ├── MS95_ENV_20.fasta
│   ├── MS95_INT_20.fasta
│   └── MS95_PRRT_20.fasta
├── ManualREGA
│   ├── manual_rega_MS95_ENV_20M.csv
│   ├── manual_rega_MS95_INT_20M.csv
│   └── manual_rega_MS95_PRRT_20M.csv
└── Scripts
    ├── comet_rest.py
    ├── decision.py
    ├── fasta_to_phylo.py
    ├── full_join.py
    ├── nexflow.config
    ├── rega_parser.py
    ├── repeat_marking.py
    ├── report.py
    ├── stanford_parser.py
    ├── subtyping_pipeline.nf
    └── tag_parser.py

IQTREE


20-02971_PR_20
Sub_CRF02_AG_CY_2009_JF683806
Sub_CRF02_AG_SN_2013_KT223759
Sub_CRF02_AG_GW_2004_FJ694791
Sub_CRF02_AG_EE_2006_KM606636
Sub_CRF02_AG_GW_2006_KR067668
Sub_CRF02_AG_NE_1997_AJ508595
Sub_CRF02_AG_SE_2011_KP411843
20-03180_PRRT_20
Sub_CRF02_AG_CM_1997_AF377954
Sub_CRF02_AG_SE_2011_KP411845

Conclusions

Conclusions

- Need to increase amount of data
- Data quality investment (Demographic & Historical)
- New marketing stratigies based on clients preferences
- Campaigns to induce customers to write reviews
- Development of Recommender System
- More transaction data to make a decision concerning acquisition

Thank you!

Phone: +49 176 412 55 071

Email: verarykalina@gmail.com

LinkedIn: https://www.linkedin.com/in/vera-rykalina-6b005728

Questions?

References

Yasemin van Heuvel et al. “Infectious RNA: Human Immunodeficiency Virus (HIV) Biology, Therapeutic Intervention, and the Quest for a Vaccine”. In: *Toxins* 14.2 (2022). ISSN: 2072-6651. DOI: 10.3390/toxins14020138. URL: https://www.mdpi.com/2072-6651/14/2/138.