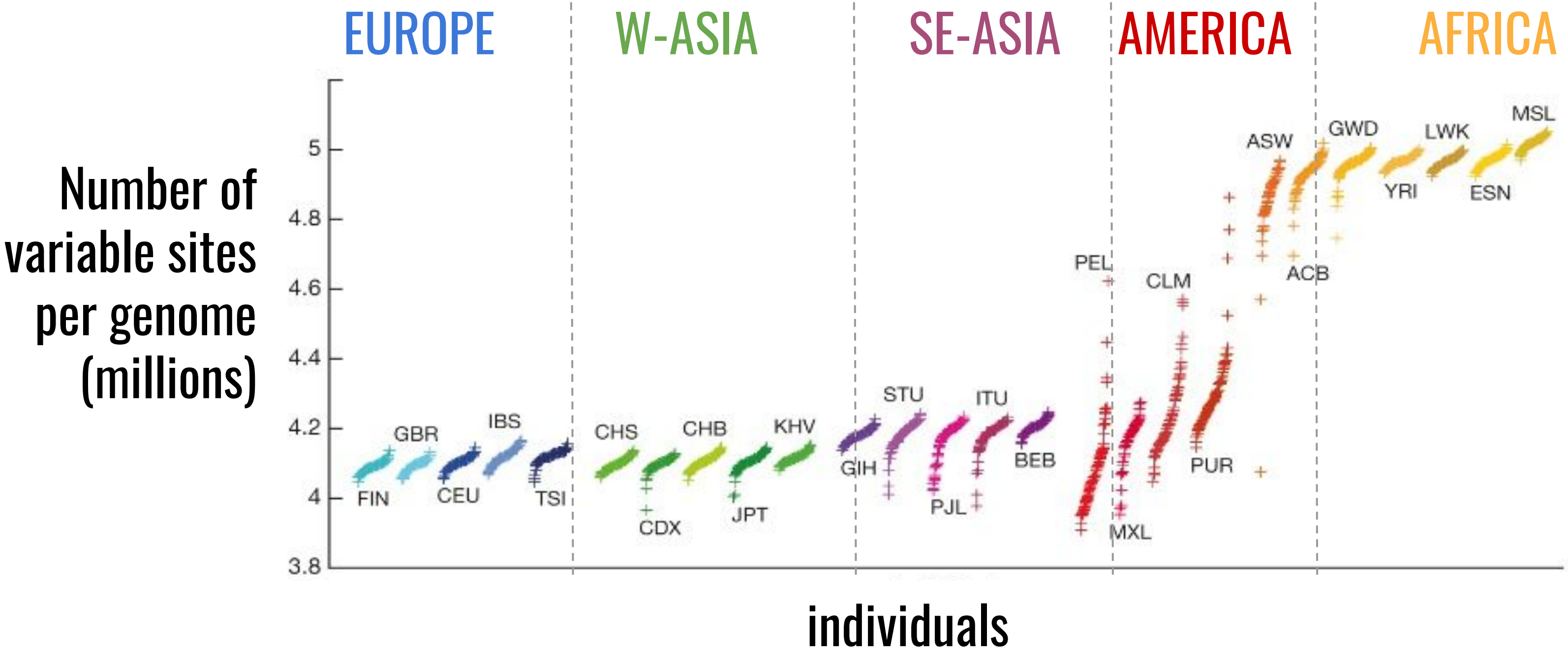
Population Genomics Analyses on pangenome graphs

Flavia Villani

Consiglio Nazionale delle Ricerche | Istituto di Genetica e Biofisica "Adriano Buzzati-Traverso" | Napoli

NETTAB / BBCC 2020 Meeting
November 16-18, 2020
Department of Agricultural Sciences, University of Naples “Federico II”, Portici, Naples, Italy.

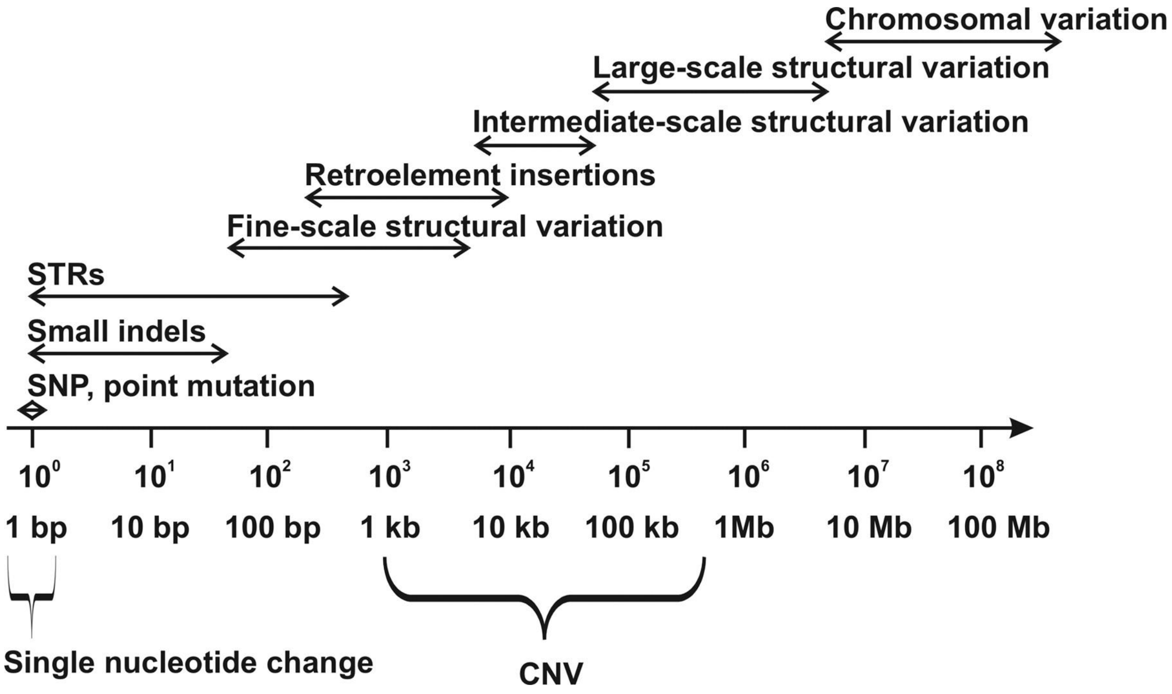
Population Genetics



A Auton et al. Nature 526, 68-74 (2015) doi:10.1038/nature15393

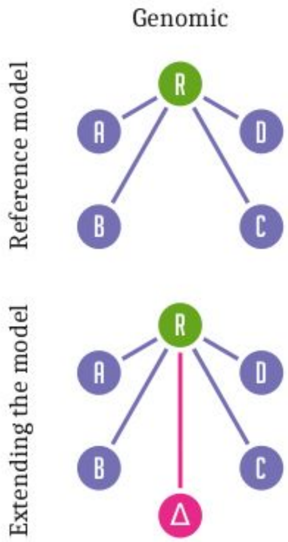
4.3 M differences on average between two individuals

Pangenomics approach for identification structural variants



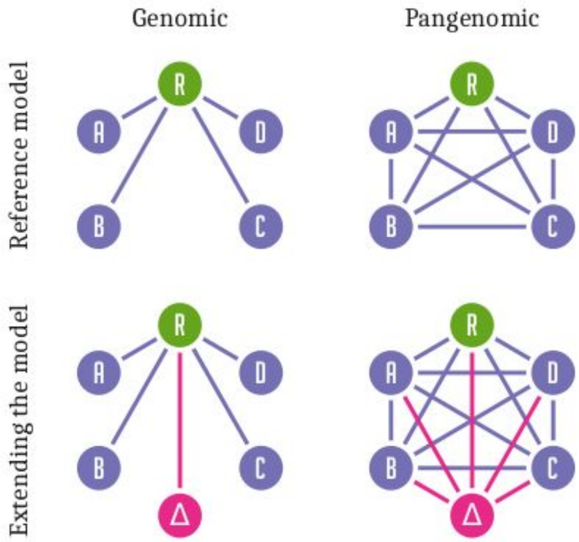
Rebecca L Pollex and Robert A Hegele. “Copy number variation in the human genome and its implications for cardiovascular disease”. In: Circulation 115.24 (2007),pp. 3130–3138

Genomic *versus* pangenomic



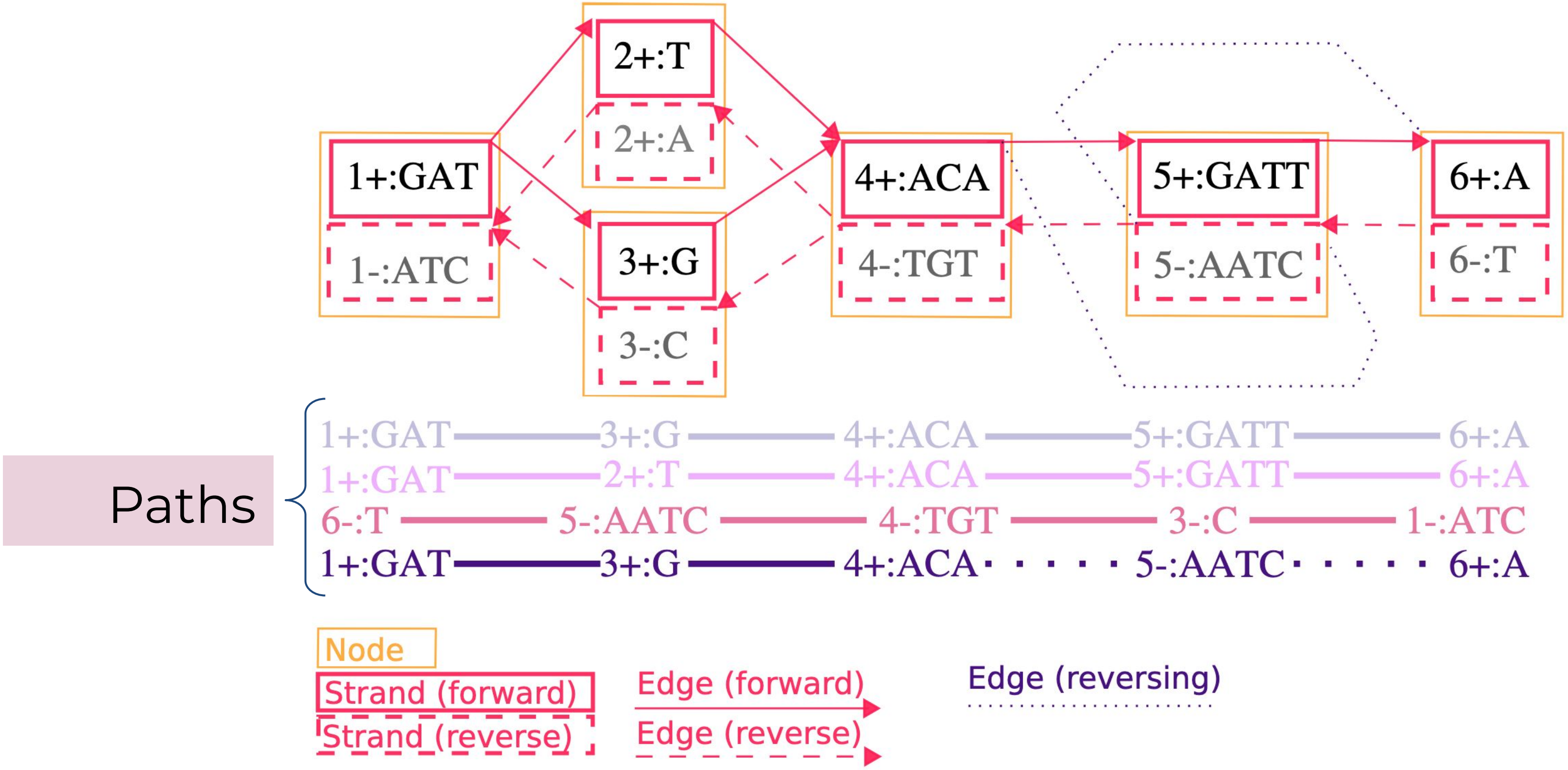
Eizenga, Jordan & Novak, Adam & Sibbesen, Jonas & Heumos, Simon & Ghaffaari, Ali & Hickey, Glenn & Chang, Xian & Seaman, Josiah & Rounthwaite, Robin & Ebler, Jana & Rautiainen, Mikko & Garg, Shilpa & Paten, Benedict & Marschall, Tobias & Sirén, Jouni & Garrison, Erik. (2020). Pangenome Graphs. Annual Review of Genomics and Human Genetics.

Genomic *versus* pangenomic



Eizenga, Jordan & Novak, Adam & Sibbesen, Jonas & Heumos, Simon & Ghaffaari, Ali & Hickey, Glenn & Chang, Xian & Seaman, Josiah & Rounthwaite, Robin & Ebler, Jana & Rautiainen, Mikko & Garg, Shilpa & Paten, Benedict & Marschall, Tobias & Sirén, Jouni & Garrison, Erik. (2020). Pangenome Graphs. Annual Review of Genomics and Human Genetics.

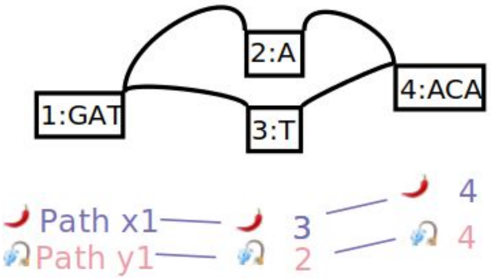
Graphic representation of a pangenome



Eizenga J. M*., Novak A. M.*, Kobayashi E., Villani F., Cisar C., Heumos S., Hickey G., Colonna V., Paten B. & Garrison E. (2020). Efficient dynamic variation graphs. *Bioinformatics Oxford*.

Genetic variants in the linear and graphical model

Pangenome (GFA)



Genomics (VCF)

#CHROM	POS	REF	ALT
x	4	A	T

Genomics standard analyses are based on linear representation of genomes

Goal

To develop a library of functions (vgpop) for population genetic analysis on pangenomic models

Library vgpop

Parsing pangenome

bubblepop

Population genetics

num_sequences

num_segregatingsites

allele_frequencies

fst

Format conversion

gfa2vcf

seqgen2gfa+vcf

Application

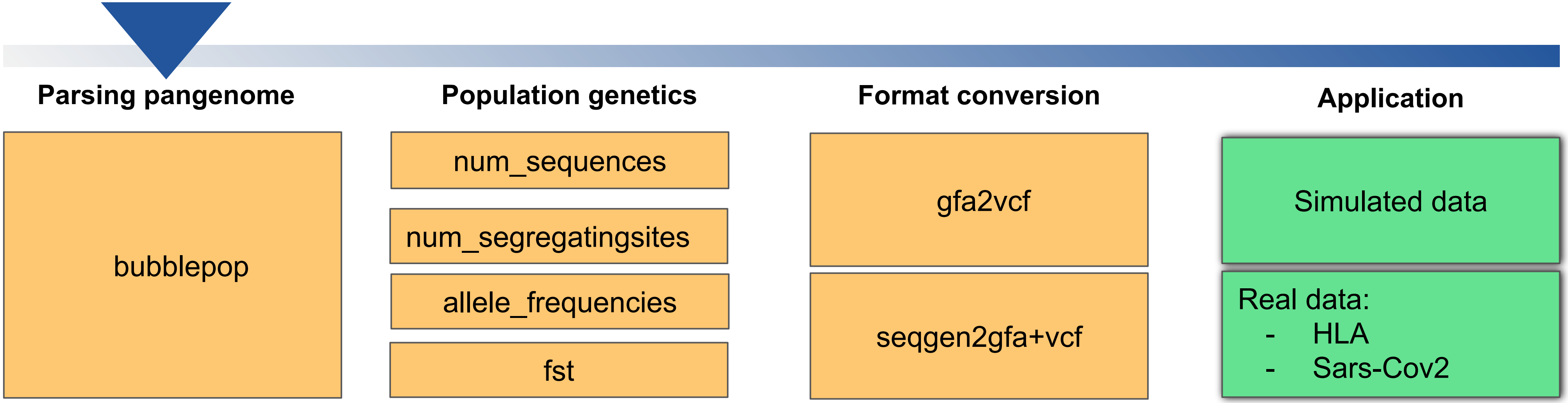
Simulated data

Real data:

- HLA
- Sars-Cov2

https://github.com/Flavia95/VGpop

Library vgpop



Parsing pangenome

bubblepop

Population genetics

num_sequences

num_segregatingsites

allele_frequencies

fst

Format conversion

gfa2vcf

seqgen2gfa+vcf

Application

Simulated data

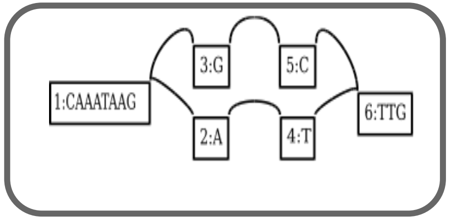
Real data:

- HLA
- Sars-Cov2

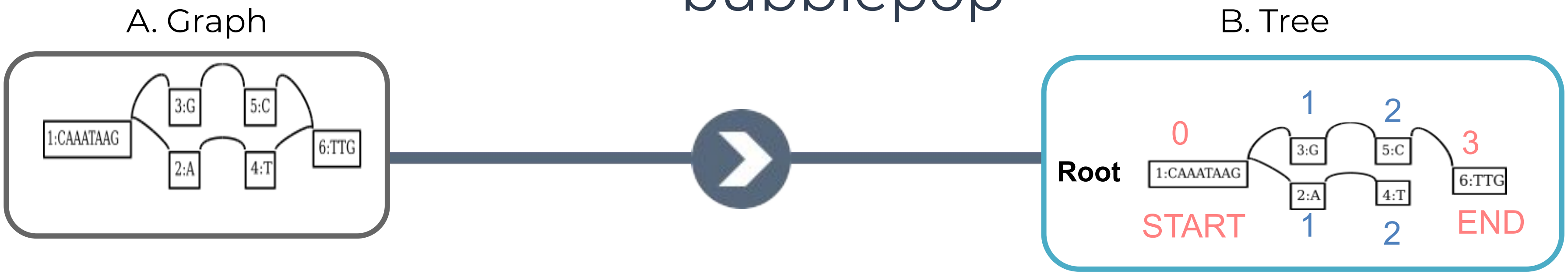
https://github.com/Flavia95/VGpop

bubblepop

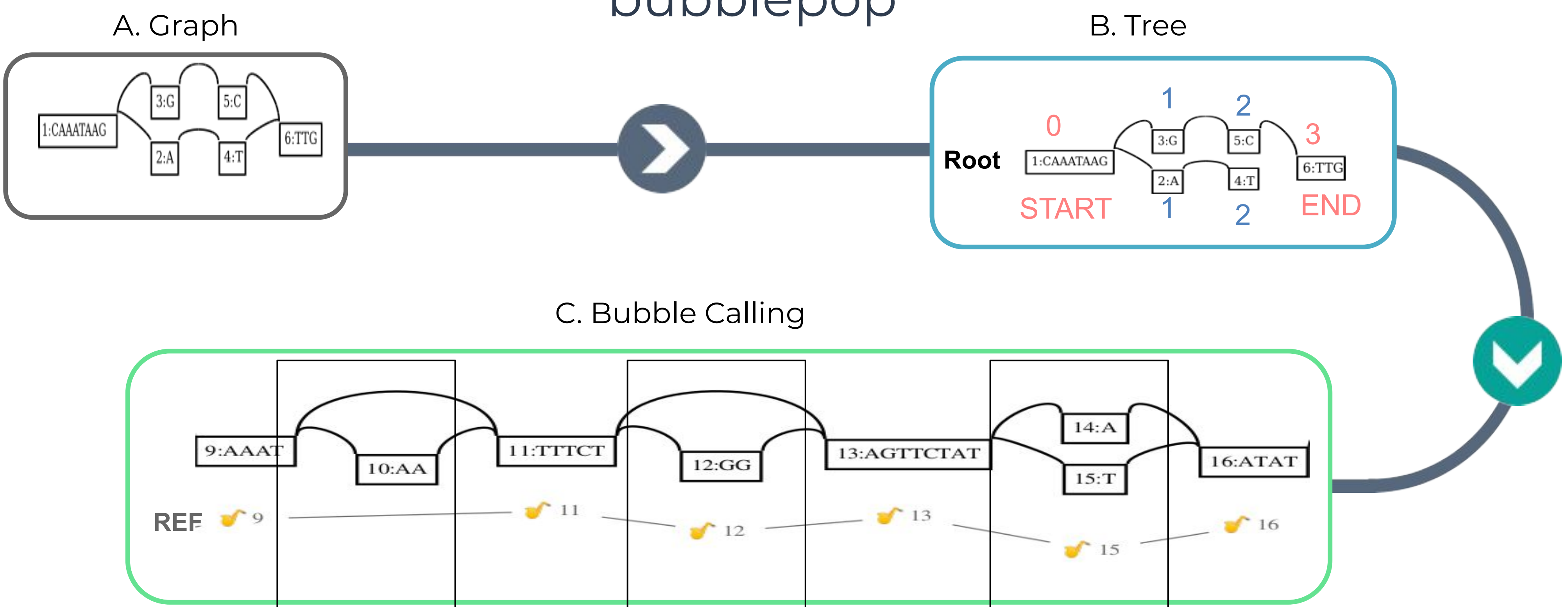
A. Graph



bubblepop


A. Graph
1:CAAATAAG
3:G
5:C
2:A
4:T
6:TTG
B. Tree
Root
0
1:CAAATAAG
START
1
3:G
2:A
1
2
5:C
4:T
2
3
6:TTG
END

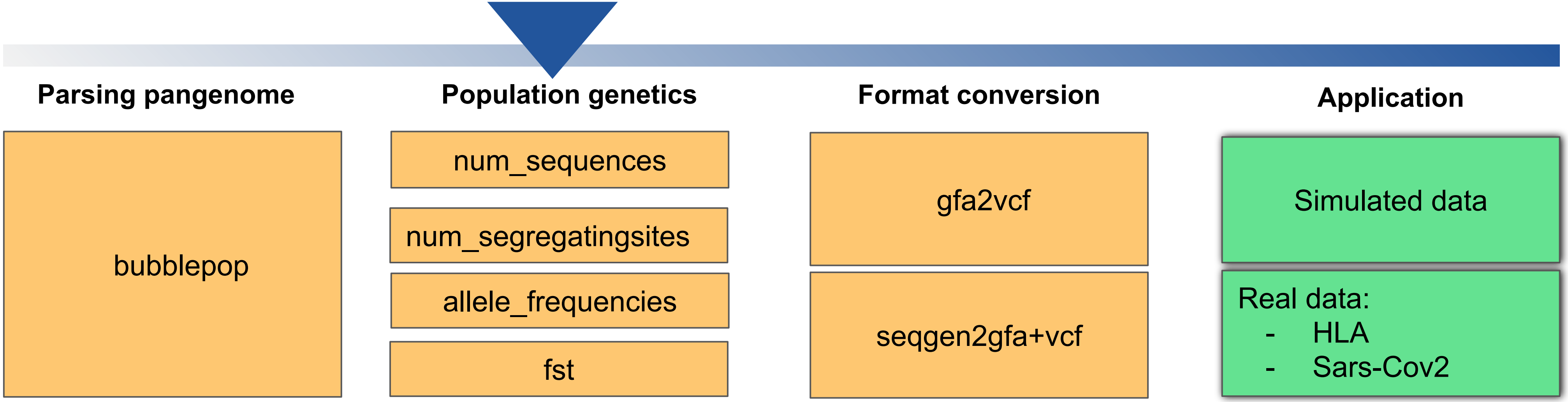
bubblepop



		Insertion		Deletion		SNV	
REF	*AAAT*	*---*	*TTTCT*	*GG*	*AGTTCTAT*	*T*	*ATAT*
ALT	*AAAT*	*AA*	*TTTCT*	*---*	*AGTTCTAT*	*A*	*ATAT*

POSITION	pos1	pos2	pos3	pos4
PATH1	T	T	T	T
PATH2	A	G	A	T
PATH3	T	A	T	A
PATH4	A	T	A	A

Library vgpop



Parsing pangenome

bubblepop

Population genetics

num_sequences

num_segregatingsites

allele_frequencies

fst

Format conversion

gfa2vcf

seqgen2gfa+vcf

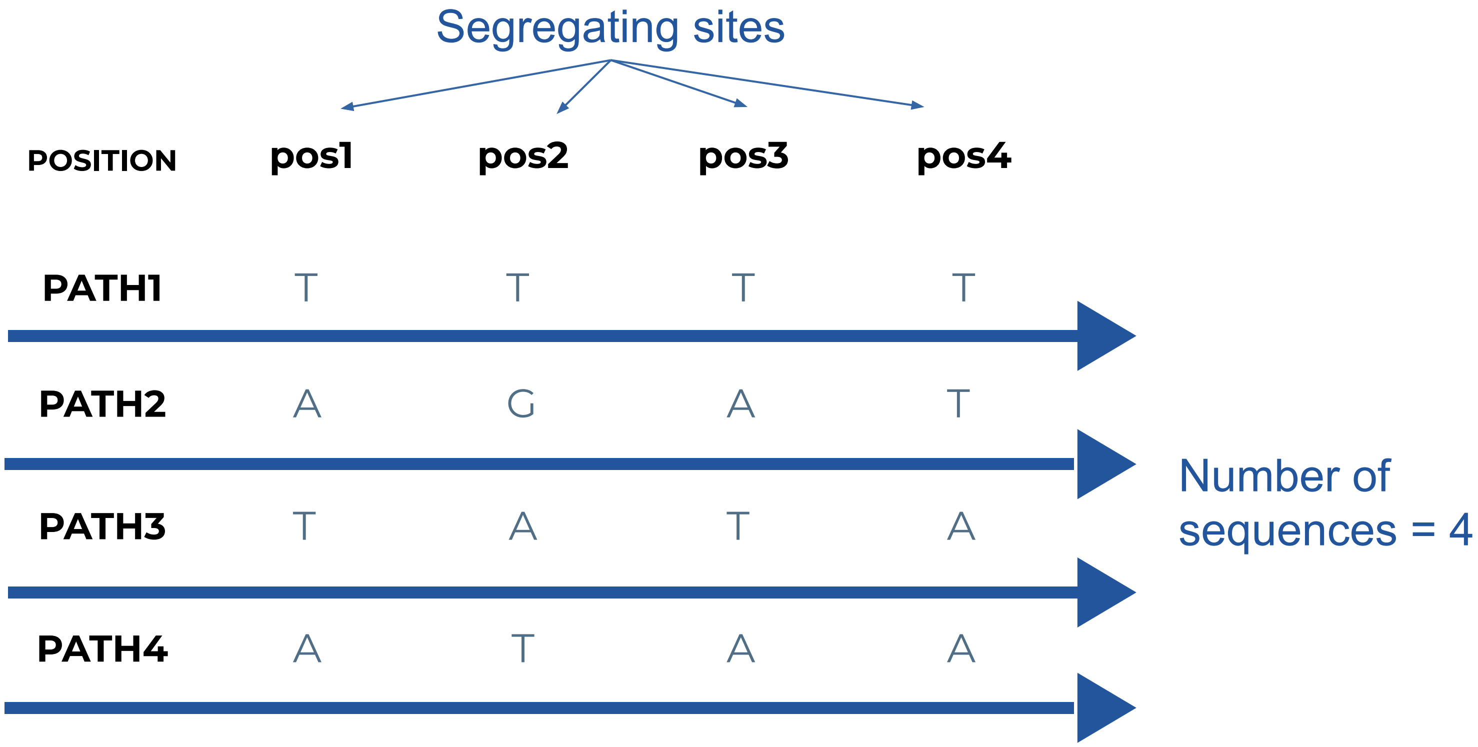
Application

Simulated data

Real data:
- HLA
- Sars-Cov2

https://github.com/Flavia95/VGpop

Segregation sites and sequences



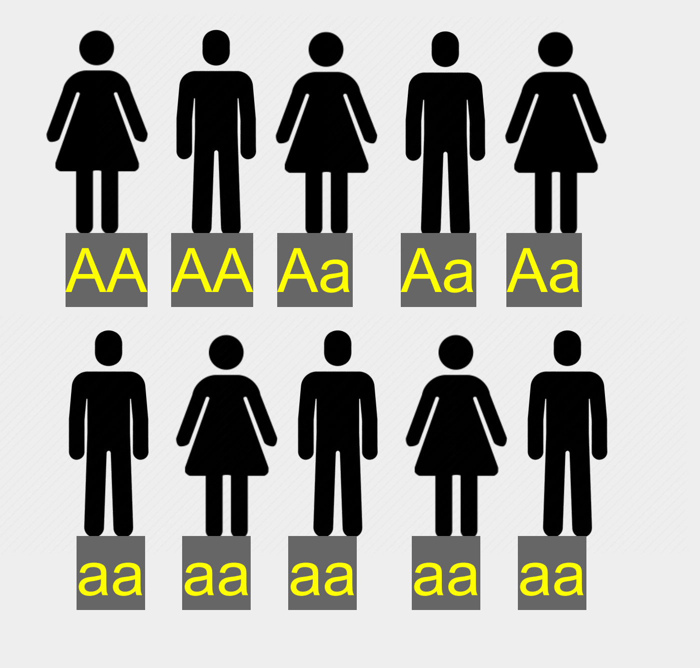
Segregating sites

POSITION	pos1	pos2	pos3	pos4
PATH1	T	T	T	T
PATH2	A	G	A	T
PATH3	T	A	T	A
PATH4	A	T	A	A

Number of sequences = 4

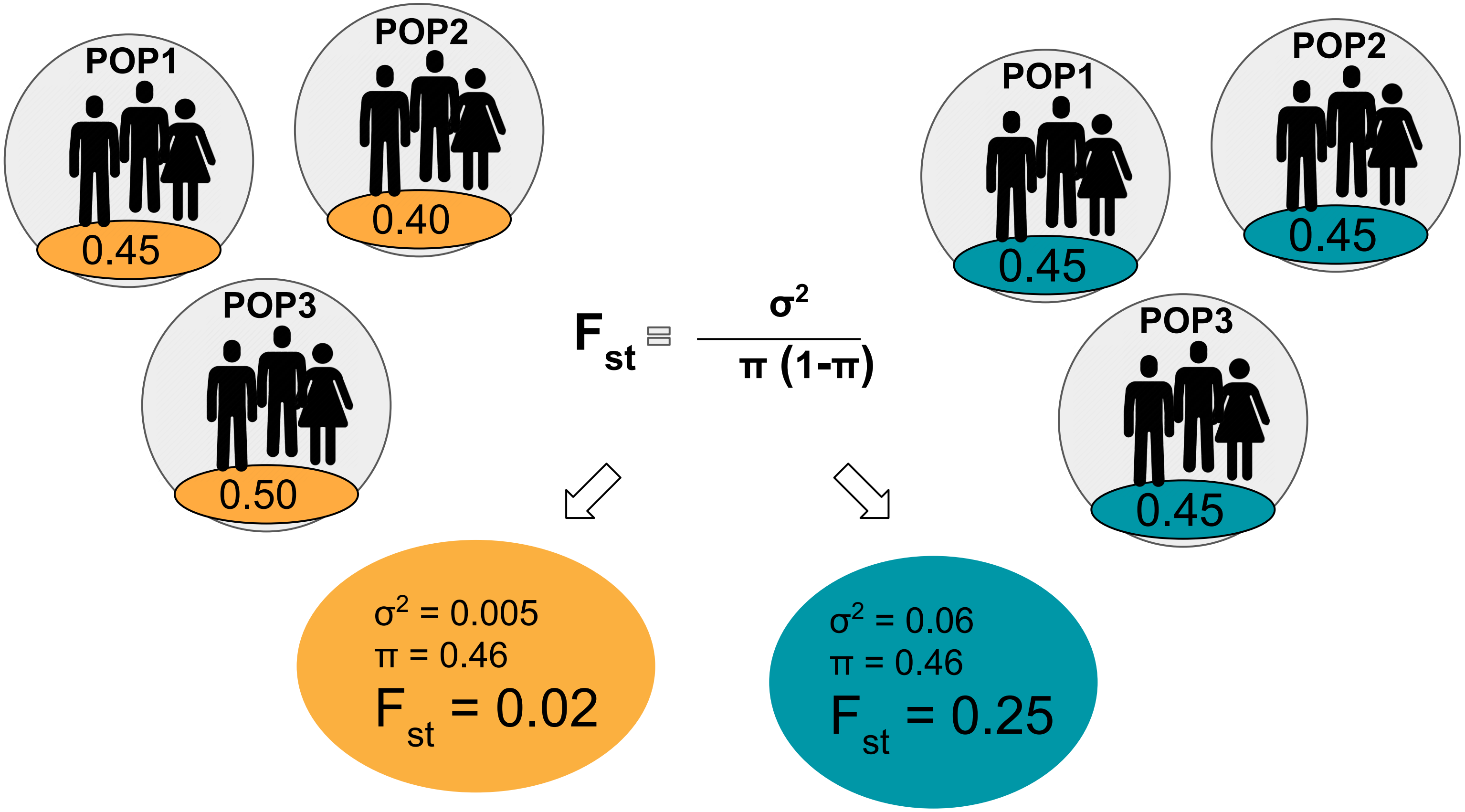
Allele frequencies

2N = 20 chromosomes
(APLOID)



ALLELE	A	a
ALLELE COUNTS	$n_A = 7$	$n_a = 13$
ALLELE FREQUENCIES	$f_A = \frac{n_A}{2N} = 0.35$	$f_a = \frac{n_a}{2N} = 0.65$

Wright's fixation index (F_{st})



POP1 0.45

POP2 0.40

POP3 0.50

$$F_{st} = \frac{\sigma^2}{\pi\,(1-\pi)}$$

$\sigma^2 = 0.005$
$\pi = 0.46$
$F_{st} = 0.02$

POP1 0.45

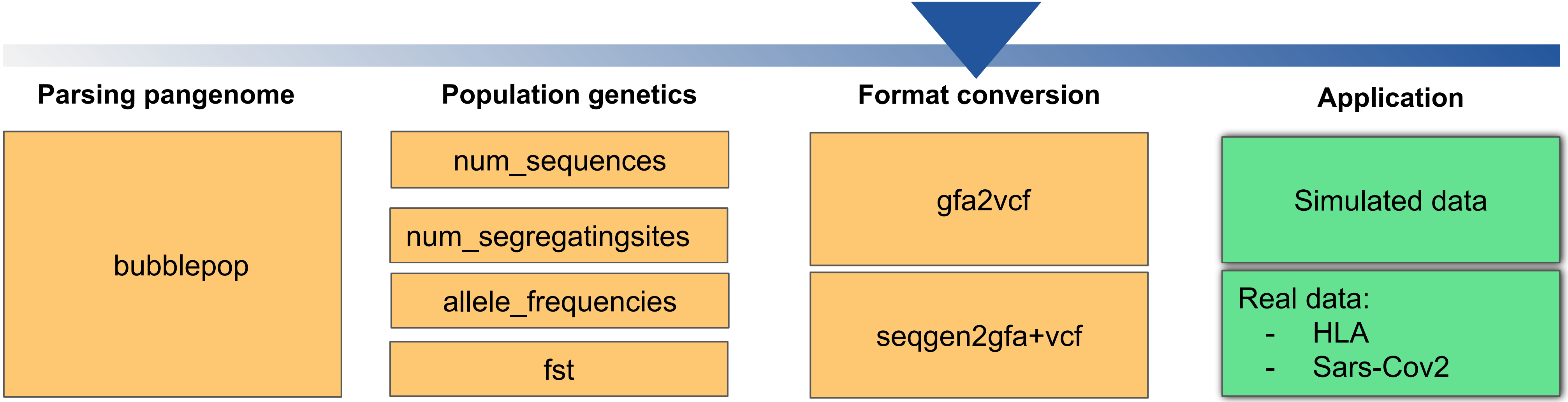
POP2 0.45

POP3 0.45

$\sigma^2 = 0.06$
$\pi = 0.46$
$F_{st} = 0.25$

Barbujani, G., & Colonna, V. (2010). Human genome diversity: frequently asked questions. *Trends in Genetics*, *26*(7), 285-295.

Library vgpop



Parsing pangenome

bubblepop

Population genetics

num_sequences

num_segregatingsites

allele_frequencies

fst

Format conversion

gfa2vcf

seqgen2gfa+vcf

Application

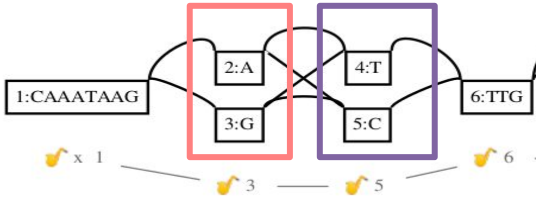
Simulated data

Real data:
- HLA
- Sars-Cov2

https://github.com/Flavia95/VGpop

Format conversion

Pangenomic model (GFA)



gfa2vcf

Linear model (VCF)

#CHROM	POS	ID	REF	ALT	INFO
x	9	.	G	A	TYPE=snv
x	10	.	C	T	TYPE=snv

Simulation sequences (Seq-Gen)

```
2 10
Taxon1  ATCTTTGTAG
Taxon2  ATCCTAGTAG
```

seqgen2gfa+vcf

Pangenomic model (GFA)

```
H    VN:Z:1
S    1     CACTA
S    2     ATTA
L    1     +      2      + 0M
P    x     1+,2+  0M
```

Linear model (VCF)

#CHROM	POS	ID	REF	ALT	INFO
x	2	.	G	A	TYPE=snv
x	3	.	C	T	TYPE=snv

Implementation of vgpop in Rust

Rust is a programming language focused on performance and safety.

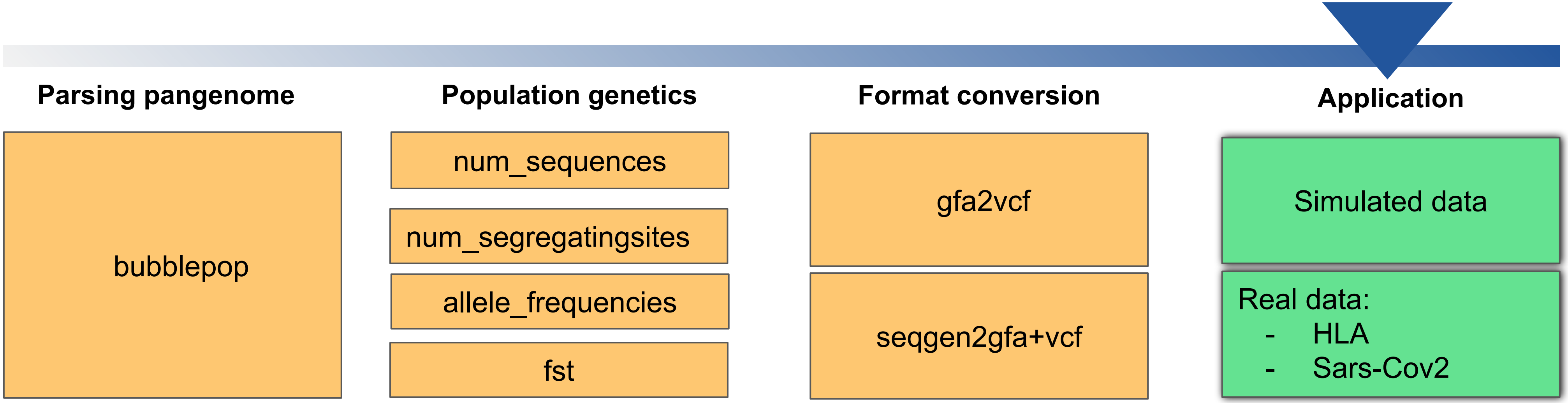
- Great ***ecosystem*** (Cargo, crates.io, docs.rs).
- Much ***safer*** than C++ while having a similar ***speed.***
- Friendly and helpful ***community.***
- Used in many open source projects, such as **Firefox.**

https://www.rust-lang.org/

Francesco Porto
Gianluca Della Vedova

https://github.com/HopedWall/rs-gfatovcf

Library vgpop



Parsing pangenome

- bubblepop

Population genetics

- num_sequences
- num_segregatingsites
- allele_frequencies
- fst

Format conversion

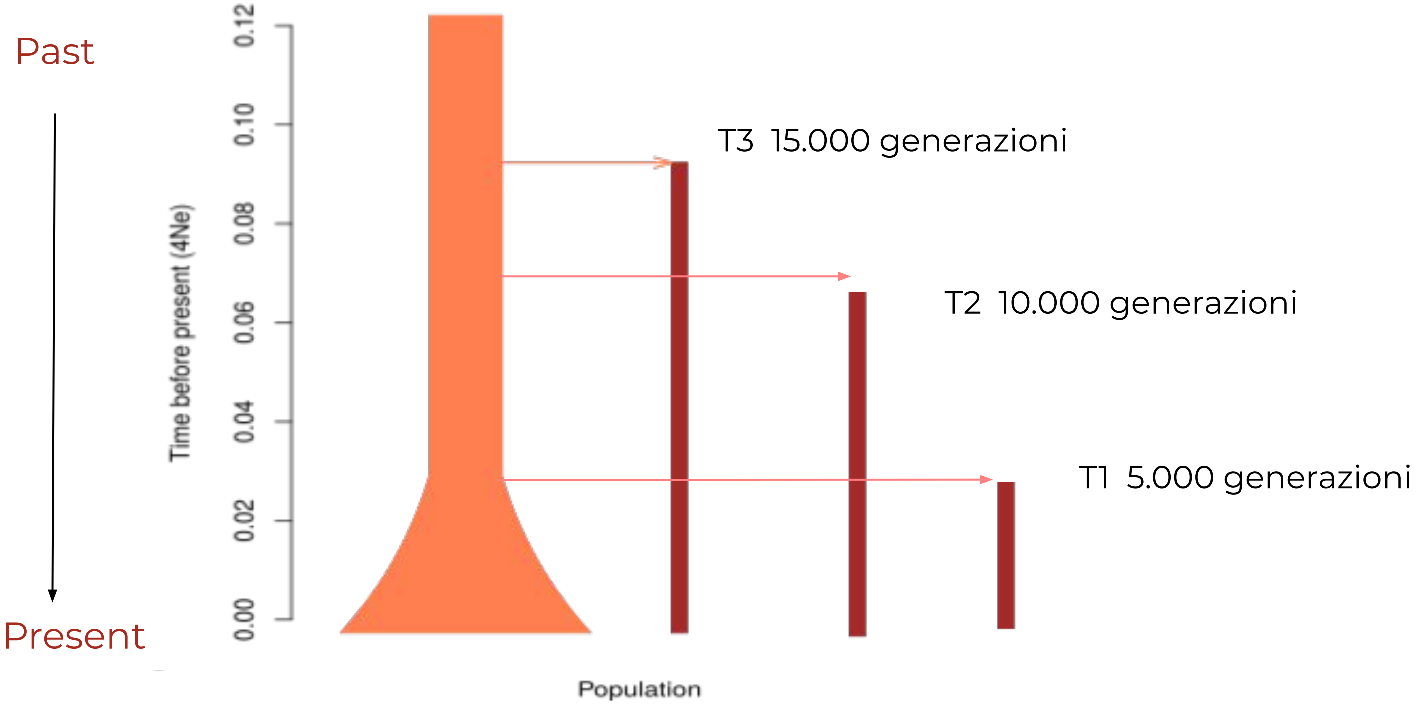
- gfa2vcf
- seqgen2gfa+vcf

Application

- Simulated data
- Real data:
  - HLA
  - Sars-Cov2

https://github.com/Flavia95/VGpop

F_{st} on simulated data



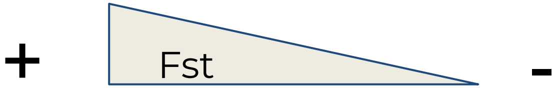
Past

Present

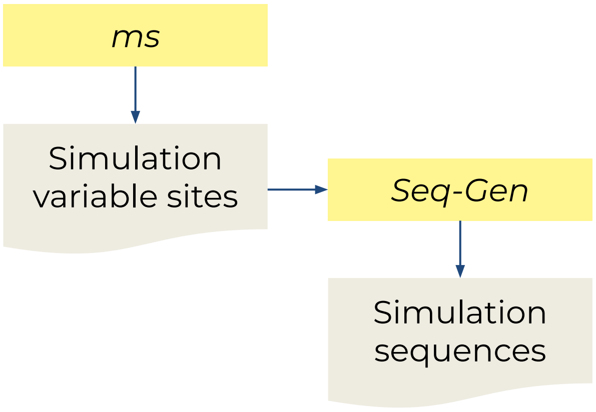
T3 15.000 generazioni

T2 10.000 generazioni

T1 5.000 generazioni



Workflow



ms

Simulation variable sites

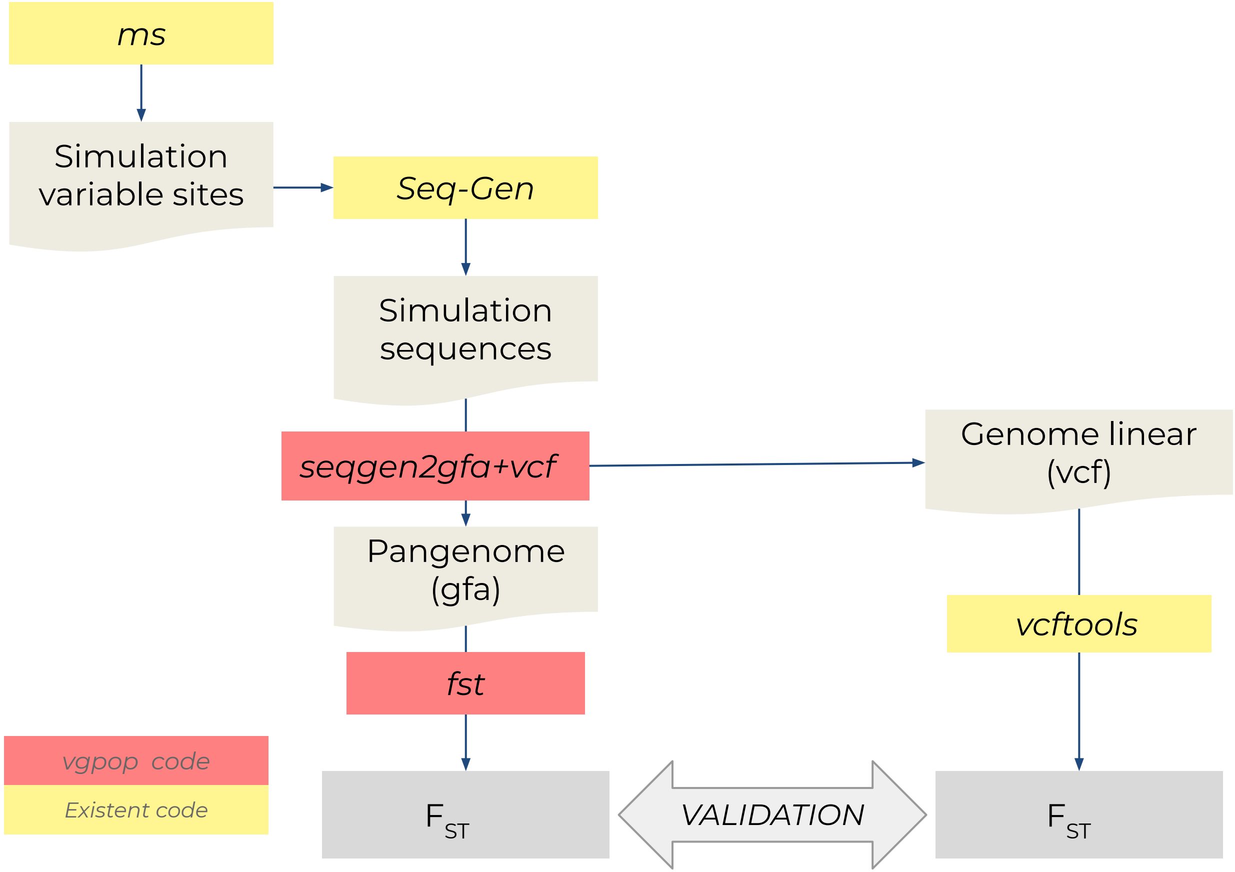
Seq-Gen

Simulation sequences

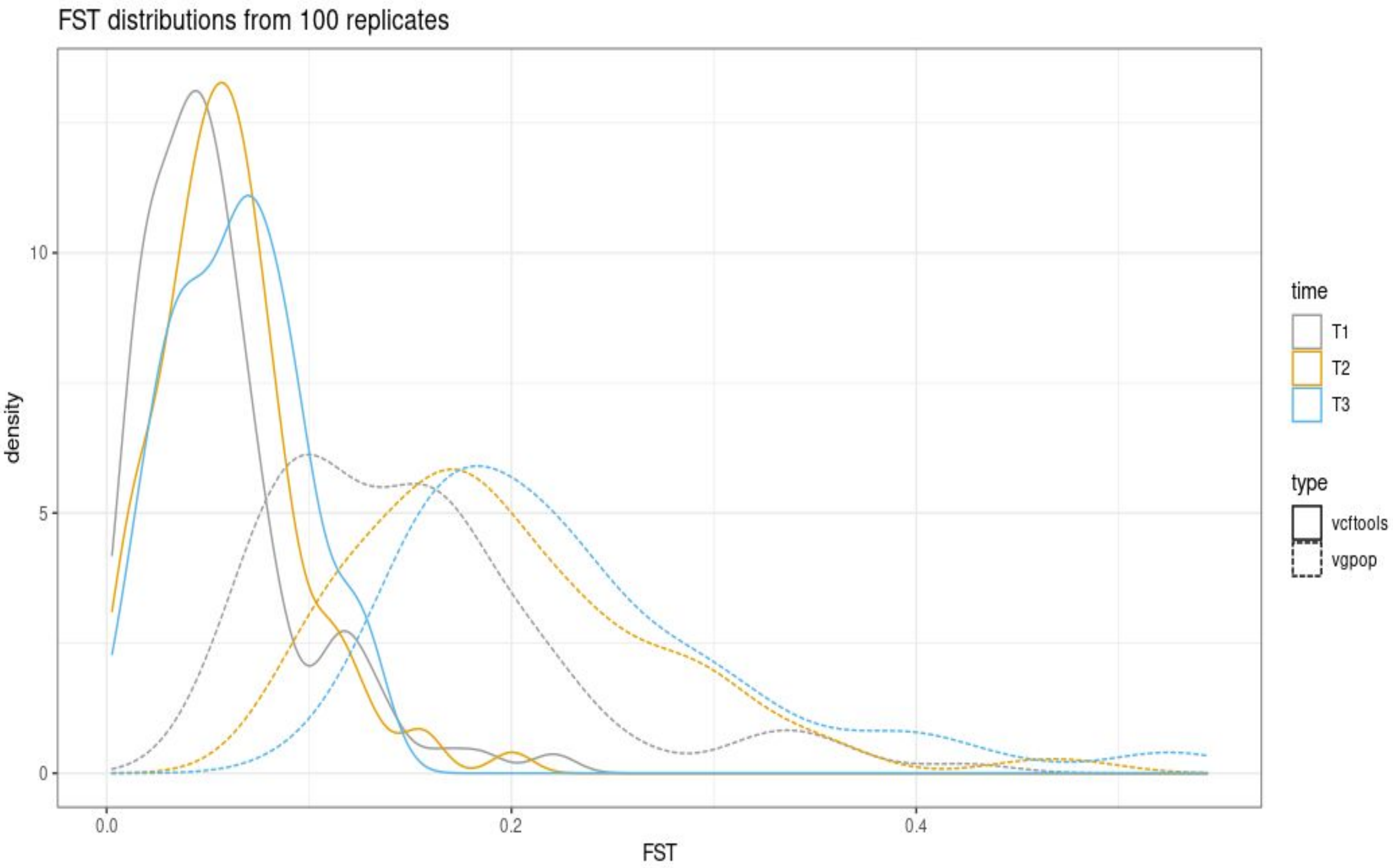
🟥 *vgpop code*

🟨 *Existent code*

Workflow

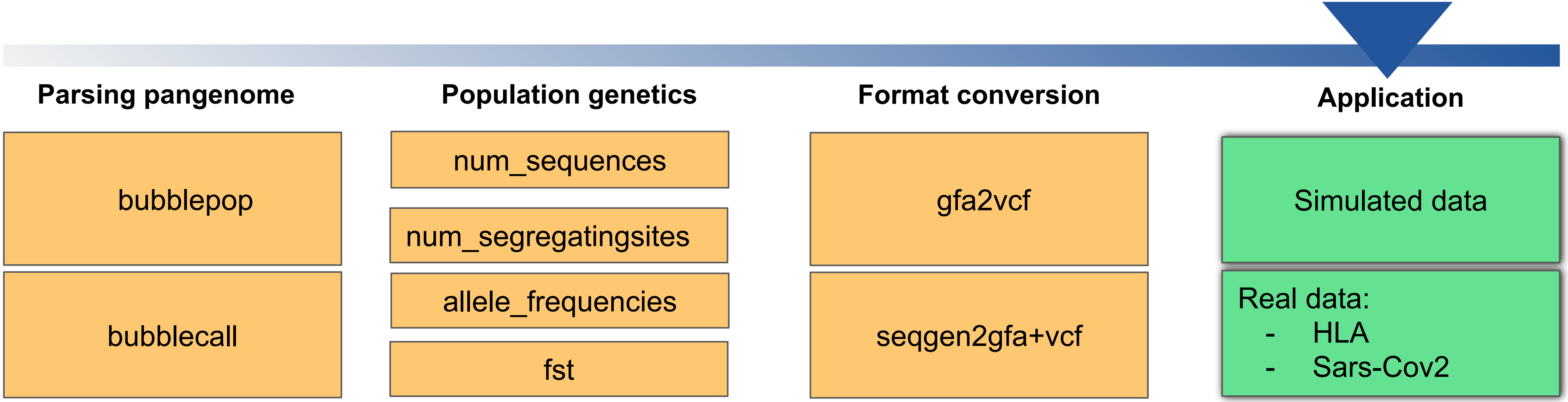

ms
Simulation variable sites
Seq-Gen
Simulation sequences
seqgen2gfa+vcf
Genome linear (vcf)
Pangenome (gfa)
vcftools
fst
vgpop code
Existent code
F_{ST}
VALIDATION
F_{ST}

F_{ST} on 100 replicate use *vgpop e vcftools*



allele_freq
fst

Library vgpop



Parsing pangenome

- bubblepop
- bubblecall

Population genetics

- num_sequences
- num_segregatingsites
- allele_frequencies
- fst

Format conversion

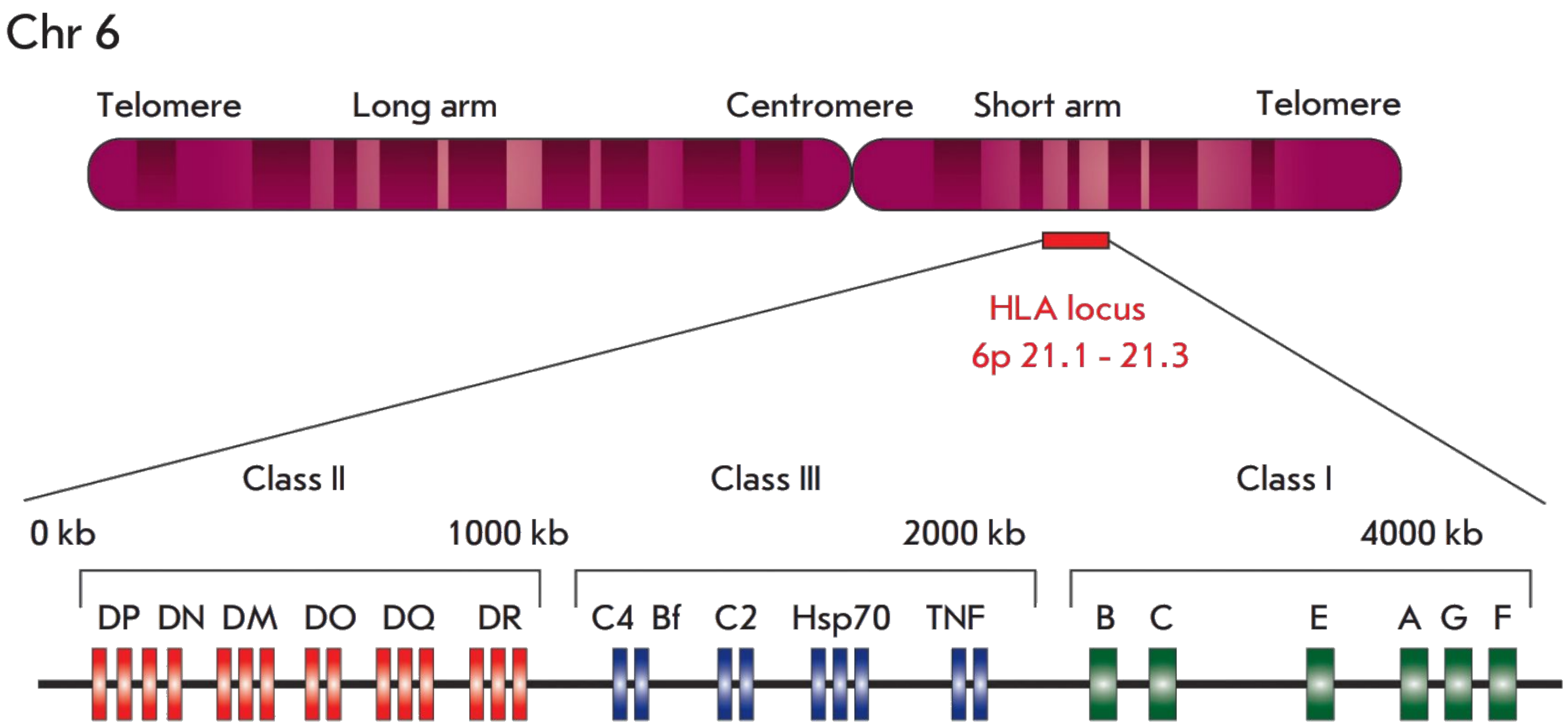
- gfa2vcf
- seqgen2gfa+vcf

Application

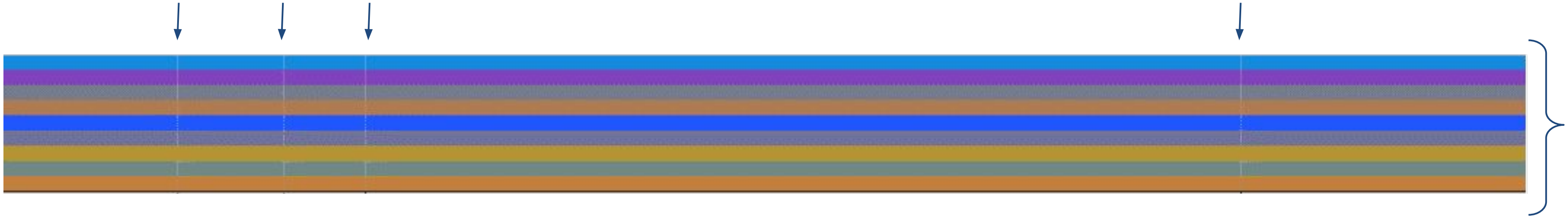
- Simulated data
- Real data:
  - HLA
  - Sars-Cov2

https://github.com/Flavia95/VGpop

Allele frequencies on HLA

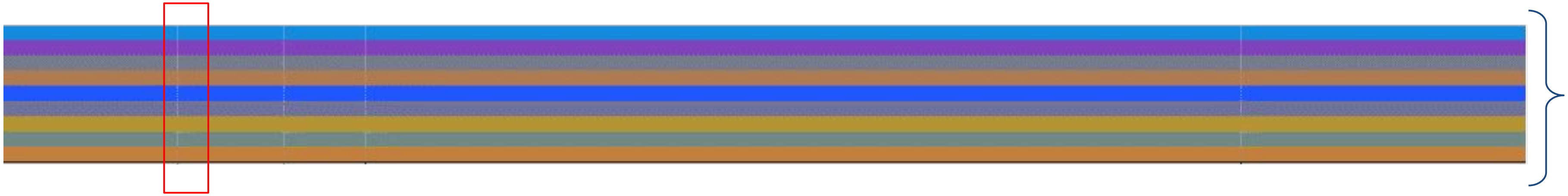


Gene HLA-E

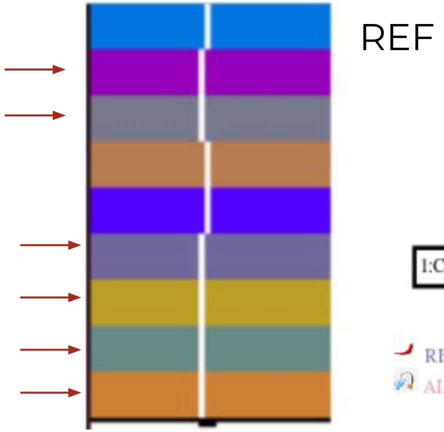


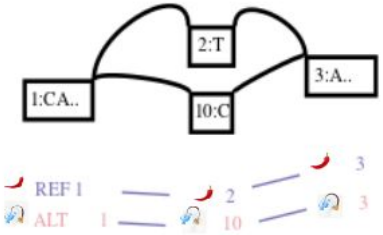
Pangenome
9 sequences

Gene HLA-E



Pangenome
9 sequences





$$\text{Freq} = \frac{6}{9}$$

GENE	PANGENOME	POSITION	REF	ALT	FREQ
HLA-E	HLAE-3133	551	T	C	0.67

Variant discovery in HLA with rust implementation

- ❖ From 12 sequences
- ❖ Size: 163416 nucleotides
- ❖ Run time: ~0.1s
- ❖ Variants found: 7505

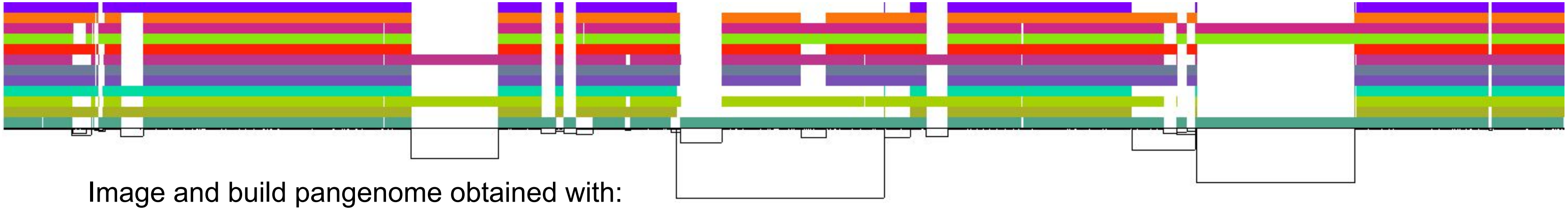
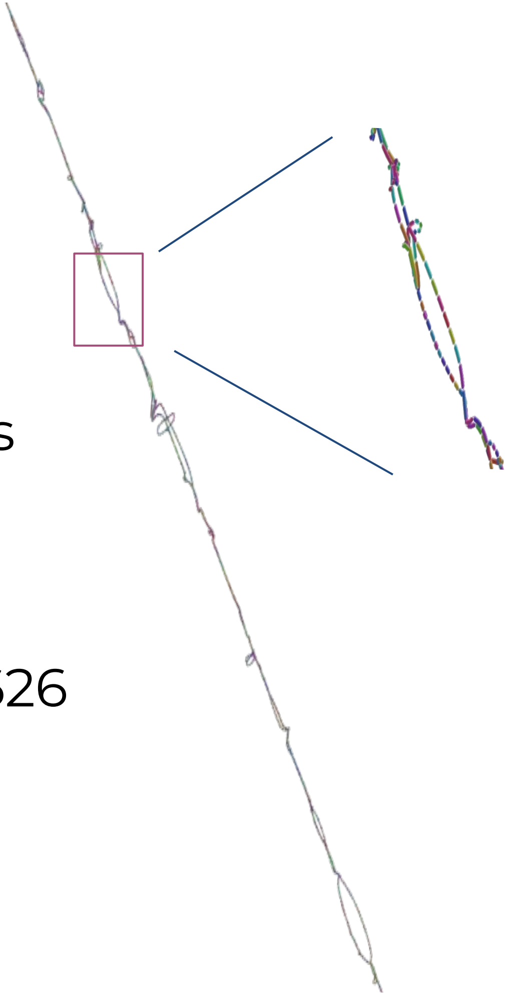


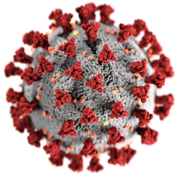
Image and build pangenome obtained with:
https://github.com/pangenome/pggb

Code available at:
https://github.com/HopedWall/rs-gfatovcf

Variant discovery in Sars-Cov2 with rust implementation



- From 15127 genomes
- 1.2 Gbytes
- 78571 fragments
- Run time: ~16m
- Variants found: 294626



COVID-19
PubSeq

Data available at
http://covid19.genenetwork.org/

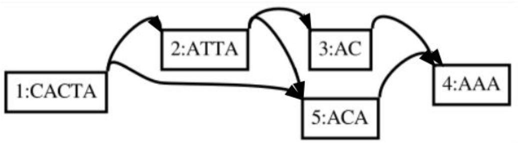
Andrea Guarracino
Pjotr Prins

Conclusion and next steps

vgpop

Software for population genetics analyses on pangenomes

Rust

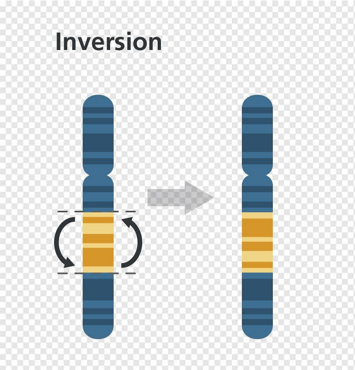


Adding parallel computing to increase performances

https://crates.io/crates/gfautil

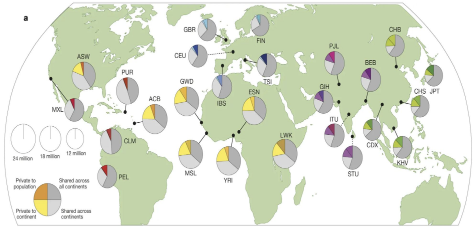
Structural variation

Little considered in the standard population genetics analysis



Population genomics analyses

Based on haplotype and on the differentiation of frequencies between populations



IGB-CNR (US)
Vincenza Colonna
Silvia Buonaiuto
Gianluca Damaggio
Giuliana D’Angelo

University of Milano Bicocca (Italy)
Francesco Porto
Gianluca Della Vedova

University of Rome Tor Vergata (Italy)
Andrea Guarracino

Department of Genetics, Genomics and Informatics (UTHSC)
Pjotr Prins
Robert W. Williams
Christian Fischer

UCSC (US)
Erik Garrison

IGB CNR Institute of Genetics and Biophysics Adriano Buzzati-Traverso

Consiglio Nazionale delle Ricerche

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Università di Roma Tor Vergata

THANKS FOR YOUR ATTENTION!