VCF_creator: Mapping and VCF Creation features in DiscoSnp++

Contact : chloe.riou@inria.fr

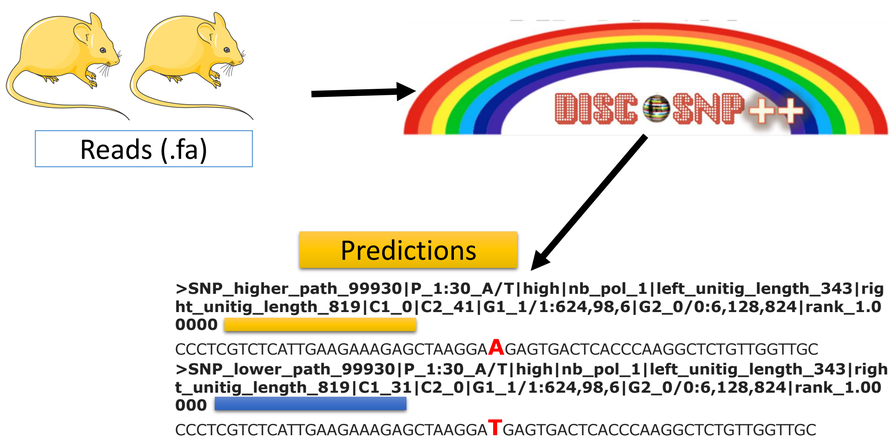
Chloé RIOU[1], Claire LEMAITRE[1] and Pierre PETERLONGO[1]

1 INRIA Rennes-Bretagne Atlantique/IRISA, EPI GenScale, Campus Beaulieu, 263 Avenue Général Leclerc, 35042, Rennes, France

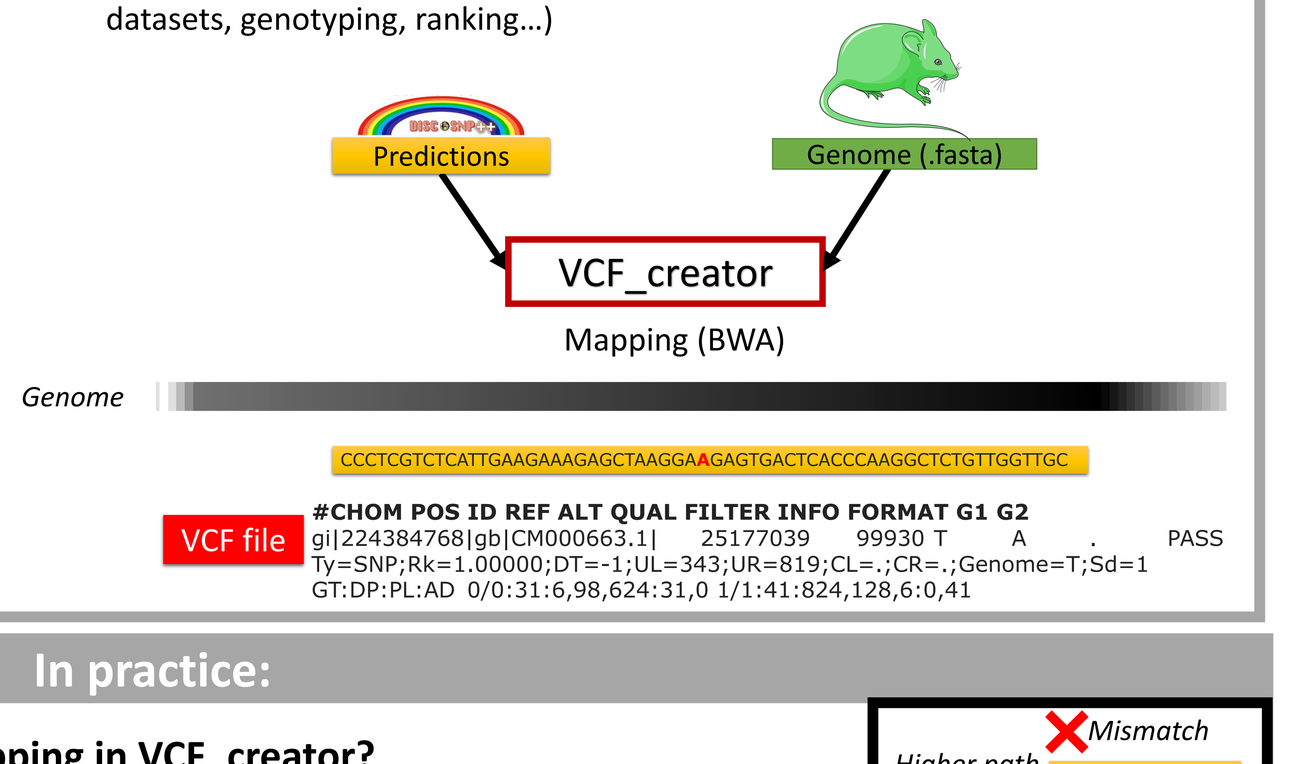
How to locate de novo predicted variants on close reference?

What we propose: DiscoSnp++ + VCF_creator

1. **DiscoSnp++** : Detects genomic variants such as Single Nucleotide Polymorphism (SNPs) or insertion/deletion (INDELs) from raw read set(s) without any reference genome (de novo)



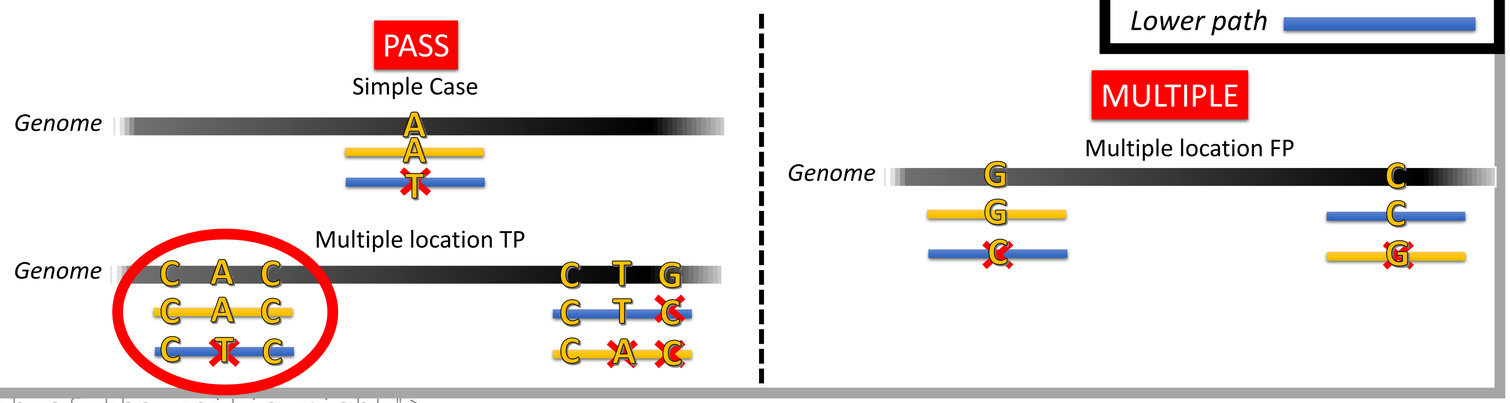
2. **VCF_creator** : Locates and after improves predictions on the given genome. The output is a VCF file with the genomic position, the reference and alternative allele, and the DiscoSnp++ informations (coverage for each datasets, genotyping, ranking…)



In practice:

How alignment is done? How do we validate the mapping in VCF_creator?

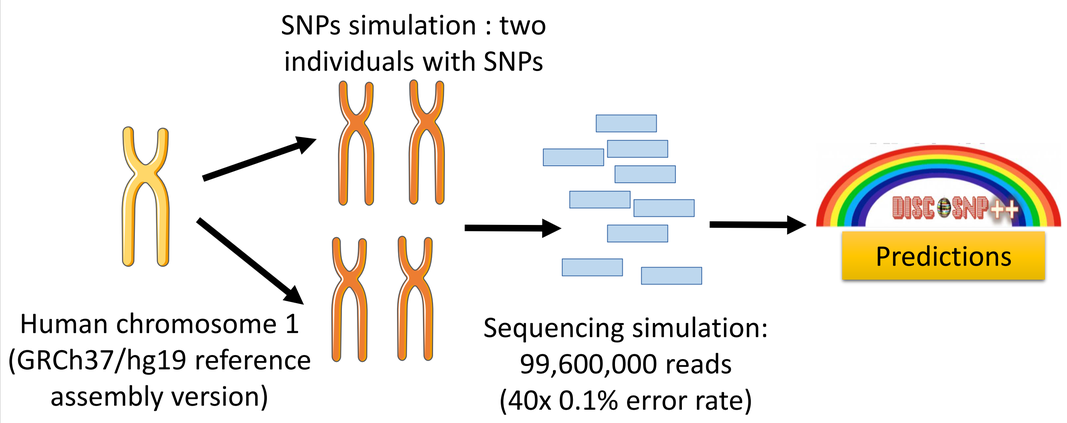
- Maps the two paths of a prediction
- For each path:
  - Record all the mapping positions with minimal mapping distance in a set *S(higher|lower positions)*
- Fills the filter field of the VCF file:
  - PASS : if *S* = 1
  - MULTIPLE : if *S* > 1
  - « . » : if *S* = 0



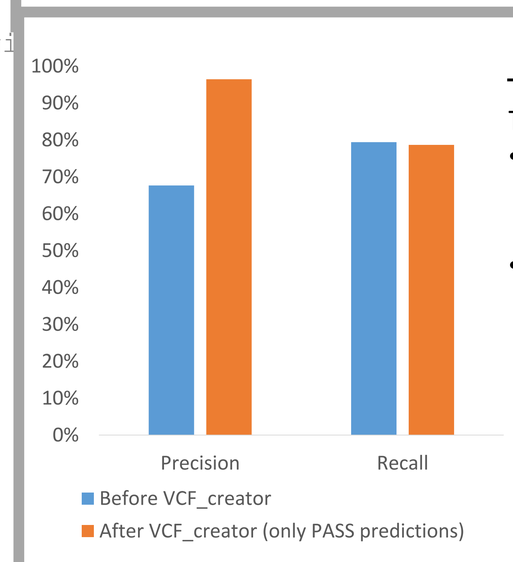
Why should you use VCF_creator?

Validation on simulated datasets

- Simulation of SNPs and INDELs on the human chromosome 1 and after simulation of sequencing with 40x coverage and 0.1% error rate
- Predictions of INDEL and SNPs with DiscoSnp++ on simulated datasets (option b1)
- Mapping on reference genome with VCF_creator



Results



To calculate the recall and the precision:

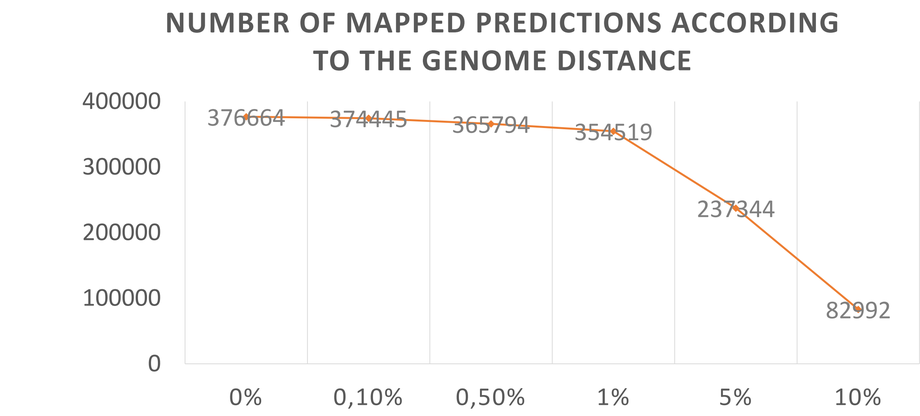
- TP : True positive → predictions of DiscoSnp++ which are true and mapped at the right position on the genome
- FP : False positive → predictions of DiscoSnp++ which are false but still mapped on the genome

Improves precision from 67% to 96%.
Why ? Because the majority of FP are due to repeated regions which are filtered by VCF_creator.

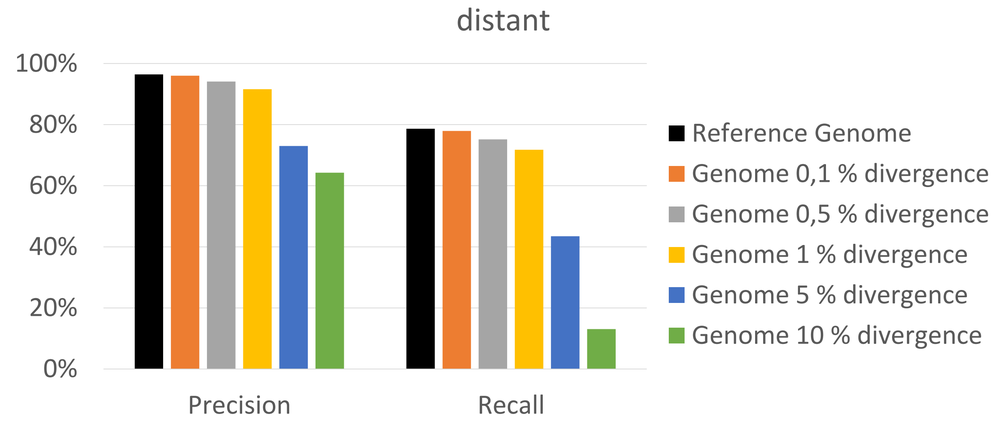
Mapping on genome increasingly distant

Until which divergence could we use VCF_creator?

Finding variants de novo often means that there is no available reference genome (or poor quality one), however it is possible to locate the predictions on close genome. To test the precision and recall of VCF creator on genomes more and more distant (from 0.1 to 10% difference), we simulated several genomes with different number of mutations.



Precision and recall for mapping on genome increasingly distant



It is possible to map effectively with VCF_creator until 1% of divergence.

Conclusion

VCF_creator allows:

- **To locate DiscpSnp++ predictions on a reference genome**
- **To improve the precision of DiscoSnp++ predictions by filtering the multi mapped predictions**
- **To produce a standard output (VCF file) which is used by many tools**

https://colibread.inria.fr/software/discosnp/