# VCF\_creator: Mapping and VCF Creation features in DiscoSnp++

Chloé Riou<sup>1</sup>, Claire Lemaitre<sup>1</sup> and Pierre Peterlongo<sup>1</sup> 1 INRIA Rennes-Bretagne Atlantique/IRISA, EPI GenScale, Campus Beaulieu, 263 Avenue Général Leclerc, 35042, Rennes, France



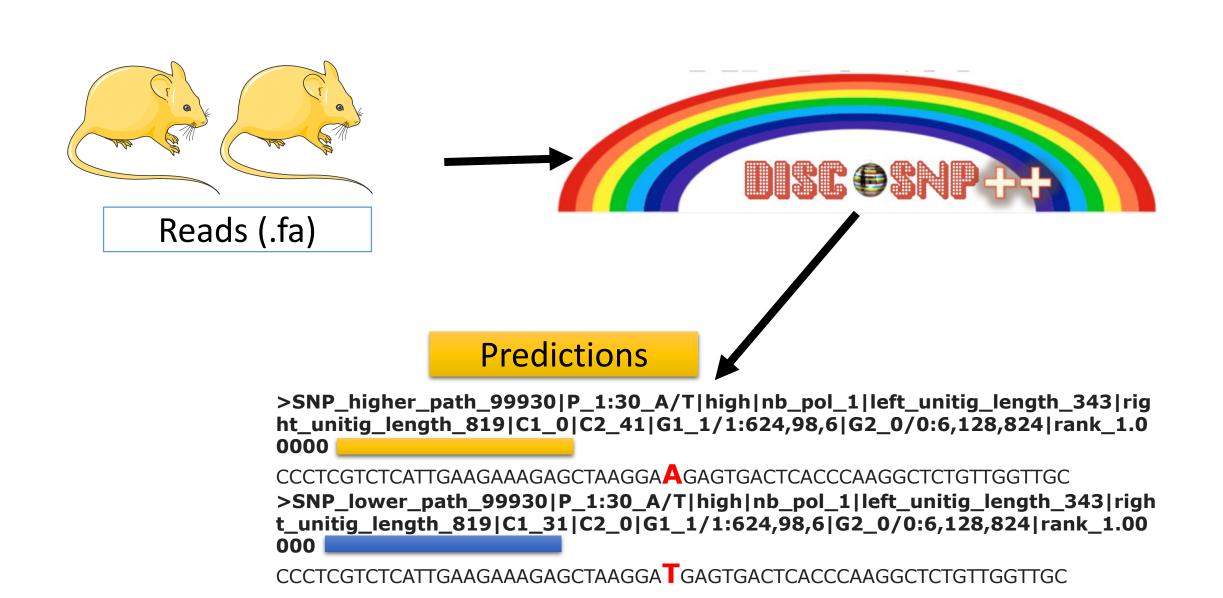




# How to locate de novo predicted variants on close reference?

## What we propose: DiscoSnp++ + VCF creator

**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)



> «.»: if S = 0

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...)

Genome (.fasta) VCF\_creator Mapping (BWA)

CGTCTCATTGAAGAAAGAGCTAAGGA<mark>A</mark>GAGTGACTCACCCAAGGCTCTGTTGGTTG( **#CHOM POS ID REF ALT QUAL FILTER INFO FORMAT G1 G2** 

VCF file gi|224384768|gb|CM000663.1| 25177039 PASS 99930 T Ty=SNP;Rk=1.00000;DT=-1;UL=343;UR=819;CL=.;CR=.;Genome=T;Sd=1 GT:DP:PL:AD 0/0:31:6,98,624:31,0 1/1:41:824,128,6:0,41

#### In practice: ##fileformat **Mismatch** How alignment is done? How do we validate the mapping in VCF\_creator? Higher path ##source=VCF Lower path Maps the two paths of a prediction ##BWA Options • **PASS** ##SAMPLE=file• For each path: Simple Case MULTIPLE ##REF=<ID=REI Record all the mapping positions with Genome Multiple location FP ##FILTER=<ID= minimal mapping distance in a set Genome ##INFO=<ID=T S(higher | lower positions) ##INFO=<ID=R Multiple location TP Fills the filter field of the VCF file: ##INFO=<ID=MU $\triangleright$ PASS : if S = 1 $\rightarrow$ MULTIPLE: if S > 1##INFO=<ID=D]

Genome

#### Why should you use VCF\_creator? ₄tig le ig r Validation on simulated datasets refere orward Simulation of SNPs and INDELs on the human chromosome 1 and after simulation of sequencing with 40x coverage and 0.1% error rate accro Predictions of INDEL and SNPs with DiscoSnp++ on simulated datasets (option b1) notype Mapping on reference genome with VCF\_creator lele SNPs simulation: two individuals with SNPs =T; Sd: e=T;S**Predictions** Sd=1Human chromosome 1 Sequencing simulation: (GRCh37/hg19 reference 99,600,000 reads assembly version) (40x 0.1% error rate) Sd=1Results 100% =.;Sd= 90% 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 50% FP : False positive → predictions of DiscoSnp++ which are false but still mapped on the genome 30% 20% 10% Improves precision from 67% to 96%. Why? Because the majority of FP are due to Recall Precision repeated regions which are filtered by ■ Before VCF creator After VCF\_creator (only PASS predictions) VCF\_creator.

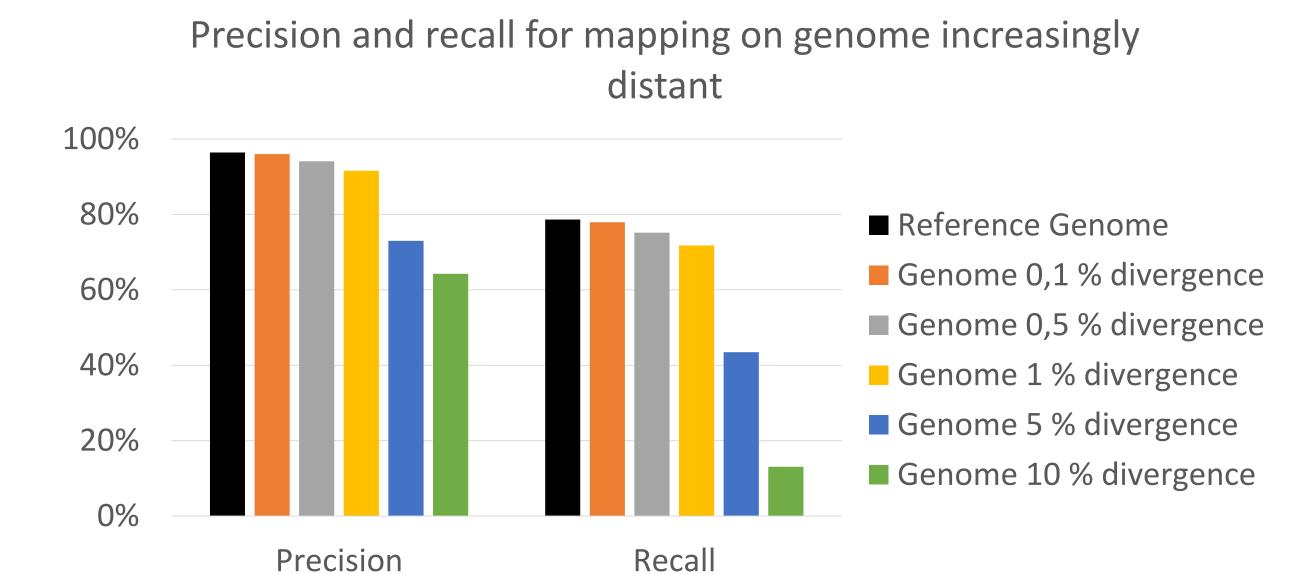
### Mapping on genome increasingly distant

### 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.

NUMBER OF MAPPED PREDICTIONS ACCORDING

### TO THE GENOME DISTANCE 400000 300000 200000 100000 82992 10% 0,10% 0,50% 1%



It is possible to map effectively with VCF\_creator until 1% of divergence.

# Conclusion

False">

##INFO=<ID=UI

### **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