VCF_creator : module of DiscoSnp++

Mapping and VCF Creation features

Table of contents

User part

VCF_creator : one module in DiscoSnp++ Mapping and VCF Creation

VCF_creator is designed to map on a genome the output of DiscoSnp++ the Single Nucleotide Polymorphism (SNP) and small indels. This module create a Variant Calling Format (VCF) from the output of DiscoSnp++ or from an alignment obtained with the software BWA [*]**.**

Quick starting

- Download and uncompress the DiscoSnp++
- Install BWA (you can add it to your PATH)

Running VCF_creator

- The main script ***run_VCF_creator.sh*** has three modes according to your needs :
  - **MODE 1** : You don't have a reference genome but you want to create a vcf (it will summarize the DiscoSnp++ informations and will give the position of the variant on the upper path of DiscoSnp++ variant). The module will create a vcf from the output of DiscoSnp++ :
    - *./run_VCF_creator.sh -p <disco_file> -o <output>*
  - **MODE 2** : You have a reference genome and you want to align your variants against a reference genome. The module will run BWA to make an alignment between your reference genome and the output of DiscoSnp++ :
    - *./run_VCF_creator.sh -**G <ref>** -p <disco_file> -o <output> [-B <path_bwa>] [-w]*
  - **MODE 3** : You already have an alignment (.sam file) and you want to create a vcf file :
    - *./run_VCF_creator.sh -f <sam_file> -o <output>*

Options

General options :

- -p : DiscoSnp++ output file (<file>.fasta) (**Mandatory** unless MODE 3)
- -o : ouput (<file>.vcf) (**Mandatory)**
- -G : reference genome (<file.fasta>) (Only in MODE 2)
- -B : bwa path (*/home/me/my_programs/bwa-0.7.12/)* (note that bwa must be pre-compiled) (Only in MODE 2) (not necessary if BWA is in the path)

---

* : *(Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. [PMID: 20080505]).*

- -f : alignment already done (<file>.sam) (Only in MODE 3)
- -h : show help
- -w : remove waste tmp files (bwa index files)
- -I : create of an output (VCF format) specific to IGV (Integrative Genomics Viewer) : sorting VCF by mapping positions and removing unmapped variants

Output

- **Final results are in your <file>.vcf.** It's a Variant Call Format that will summarize all the mapping information and the header of DiscoSnp++ informations. Example :

- **MODE 1** :

```
#CHROM POS     ID      REF     ALT     QUAL    FILTER INFO    FORMAT G1      G2      G3      G4      G5
SNP_higher_path_13736       30      13736   A       G       .       .
        Ty=SNP;Rk=0.99909;UL=.;UR=.;CL=.;CR=.   GT:DP:PL:AD   0/0:53447:1947,160066,1067676:53425,22
        0/1:3:40,9,20:1,2    1/1:94862:1895477,284398,3575:30,94832  0/1:5:34,9,54:3,2
        1/1:65389:1306661,196101,2493:19,65370
SNP_higher_path_5442        30      5442_1 A       G       .       .
        Ty=SNP;Rk=0.42995;UL=.;UR=.;CL=.;CR=.   GT:DP:PL:AD   0|1:19:98,14,198:12,7      0|
1:18:138,12,138:9,9 0|1:8:66,10,66:4,4  0|1:130:257,124,1813:104,26      0|0:224:14,598,4324:220,4
SNP_higher_path_5442        31      5442_2 G       A       .       .
        Ty=SNP;Rk=0.42995;UL=.;UR=.;CL=.;CR=.   GT:DP:PL:AD   0|1:19:98,14,198:12,7      0|
1:18:138,12,138:9,9 0|1:8:66,10,66:4,4  0|1:130:257,124,1813:104,26      0|0:224:14,598,4324:220,4
INDEL_higher_path_586       30      586     GGAC    G       .       .
        Ty=INS;Rk=0.40534;UL=.;UR=.;CL=.;CR=.   GT:DP:PL:AD   0/1:127:837,17,977:67,60
        1/1:387:6056,518,408:52,335      0/1:70:372,21,651:42,28    0/1:126:1395,53,477:40,86
        0/1:102:671,16,791:54,48
```

- **MODE 2** and **MODE 3 :**

```
#CHROM POS     ID      REF     ALT     QUAL    FILTER INFO    FORMAT G1      G2      G3      G4      G5
Pseudomonas   109860 15103  C       G       .       PASS
        Ty=SNP;Rk=0.65101;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=C;Sd=-1   GT:DP:PL:AD   1/1:4:84,16,4:0,4
        1/1:1:24,7,4:0,1     1/1:1:24,7,4:0,1     0/1:6:17,15,97:5,1  0/1:10:124,14,44:3,7
Pseudomonas   4416118       1416_1 A       G       .       PASS
        Ty=SNP;Rk=0.41638;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=A;Sd=1    GT:DP:PL:AD   1|1:35:606,71,28:3,32
        0|1:26:403,41,43:4,22      1|1:34:617,79,18:2,32      0|1:46:316,14,356:24,22    0|
1:43:547,36,128:11,32
Pseudomonas   4416119       1416_2 A       C       .       PASS
        Ty=SNP;Rk=0.41638;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=A;Sd=1    GT:DP:PL:AD   1|1:35:606,71,28:3,32
        0|1:26:403,41,43:4,22      1|1:34:617,79,18:2,32      0|1:46:316,14,356:24,22    0|
1:43:547,36,128:11,32
Pseudomonas   1312837       14361  C       CGTAGAAGATCTC.       PASS
        Ty=DEL;Rk=0.26896;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=.;Sd=-1   GT:DP:PL:AD
        0/0:792:47,1978,15015:771,21      0/0:745:34,1905,14223:728,17      0/0:697:37,1765,13268:680,17
        0/0:783:32,2016,14979:766,17      0/1:849:1303,868,12339:701,148
Pseudomonas   1753272       6164   TGTGCAGGTCGAAGACTTGA       T       .       PASS
        Ty=INS;Rk=0.13677;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=.;Sd=1    GT:DP:PL:AD
        0/0:1086:34,2817,20829:1064,22    0/0:1158:34,3011,22226:1135,23
        0/0:1003:32,2608,19250:983,20     0/0:1486:124,3555,27823:1437,49
        0/0:1132:585,1967,19224:1033,99
```

- In this example, the three types of DiscoSnp++ variants : in red : simple SNP ; in green close SNPs ; in black INDEL.

- The VCF file has the following fields :

- **CHROM** : chromosome id where the prediction is mapped, or allele id of the upper path if the variant is unmapped or if no reference genome is provided.
- **POS** (1-based leftmost position):
  - If a reference genome is provided and if the variant is mapped on a unique position : the mapping position of the variant
  - If a reference genome is provided and if the variant is not uniquely mapped : one of the positions of the variant (1-based lefmost position)
  - Else (no reference genome provided or unmapped variant) : position of the variant on the upper path of the discoSnp++ prediction (including the left extension)
- **ID** : identification of the variant (used by DiscoSnp++). For the close SNPs, the SNP number is added to the ID. Example : *10388_2*
- **REF** :
  - If one of the two predicted allele maps this position : the corresponding variant
  - Else, or if no reference genome provided : the lexicographically smallest of the two variants
  - In case of close SNPs : the first is defined as previously described. The following SNPs are those located on the same path
- **ALT** : The variant non reported as the “REF” variant
- **QUAL** : “.” (unused)
- **FILTER** :
  - PASS if the variant is mapped at unique position
  - MULTIPLE if the variant is mapped on multiple positions
  - “.” : if the variant is unmapped or if no reference genome is provided
- **INFO** :
  - **Ty** : Type of variant
    - SNP : If the variant is a simple SNP or close SNPs
    - INS : If the variant mapped corresponds to the longuest path ; the alt carries the deletion
    - DEL : If the variant mapped corresponds to the shortest path ; the alt carries the deletion
  - **Rk** : Rank of the prediction computed by DiscoSnp++ (if several datasets are used in DiscoSnp++, ranks the predictions according to their read coverage in each condition favoring SNPs that are discriminant between conditions value between 0 and 1)
  - **UL** : Length of the left unitig (“.” if not computed)

- **UR** : Length of the right unitig (“.” if not computed)
- **CL** : Length of the left contig (“.” if not computed)
- **CR** : Length of the right contig (“.” if not computed)
- **Genome** : Applies only for SNPs when a reference genome is provided (“.” for INDELs and when no reference genome provide or if the variant is unmapped). Reference nucleotide (!!nucleotide in the reference genome !! In general it is correspond to the REF field ; could be different for close snps). **Important Remark** : If one of the two predictions matches the reference : equal to the “REF” field, else equal to the nucleotide of the reference genome.
- **Sd** : Applies only when a reference genome is provided (“.” if no reference genome provided or if the variant is unmapped). Strand of the prediction mapping. “1” : Forward ; “-1” : Reverse. **Important Remark** : Fields “REF” , “ALT” and “Genome” are based on the mapped predictions. If Sd is 1 then these fields correspond to the DiscoSnp++ prediction, else if the Sd is “-1”, then they correspond to the reverse complement of the DiscoSnp++ predictions.

◦ **FORMAT** : Description of the genotype fields (G1, G2, G3 …)

- **GT** : genotype, encodes as allele values with 0 corresponding to the reference and 1 to the alternative. About genotypes :
  - If the separator is a “/” the genotypes are unphased (INDEL, Simple SNP)
  - If the separator is a “|” the genotypes are phased (Close SNPs with the same ID)
- **DP** : Cumulated depth across samples (sum)
- **PL** : Phred-scaled Genotype Likelihood (given by DiscoSnp++)
- **AD** : Depth of each allele by sample
- **HQ** : Haplotype Quality (Q in DiscoSnp++ header)

◦ **Genotype :** G1, G2 , G3 and so on. Each fields contains all the informations of the format fields for each samples.

Examples

◦ **FILTER Fields**:

- PASS if the variant is mapped at unique position
- MULTIPLE if the variant is mapped on multiple positions
- “.” : if the variant is unmapped or if no reference genome is provided

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	G1	G2	G3	G4	G5
Pseudomonas	1945315		8816	C	G	.	PASS						

```
      Ty=SNP;Rk=0.63574;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=C;Sd=1   GT:DP:PL:AD  1/1:8:135,19,16:1,7
      0/0:2:4,10,44:2,0   0/0:1:4,7,24:1,0    0/1:8:30,14,110:6,2 0/1:5:54,9,34:2,3
Pseudomonas  3190754      768    A      G      .      MULTIPLE
      Ty=SNP;Rk=0.06837;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=A;Sd=-1  GT:DP:PL:AD
      0/1:1254:8182,28,9459:659,595    0/1:1235:9809,42,7594:562,673    0/1:903:7217,37,5521:409,494
      0/1:473:4217,52,2520:194,279     0/1:1416:10513,26,9395:680,736
```

- **VCF for IGV :** If you want to use IGV (Integrative Genomics Viewer) to visualize your data, VCF_creator can create a vcf specific to IGV (0-based, sorted by positions, and without unmapped variants)

```
#CHROM POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT G1     G2     G3     G4     G5
Pseudomonas  1206  7667_1 G      T      .      PASS
      Ty=SNP;Rk=0.10501;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=G;Sd=-1  GT:DP:PL:AD  0|
1:115:1172,36,513:41,74    0|1:142:1154,19,875:64,78 0|1:90:707,16,587:42,48    0|
1:111:1190,43,452:37,74    0|1:102:960,26,521:40,62
Pseudomonas  1210  7667_2 T      G      .      PASS
      Ty=SNP;Rk=0.10501;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=T;Sd=-1  GT:DP:PL:AD  0|
1:115:1172,36,513:41,74    0|1:142:1154,19,875:64,78 0|1:90:707,16,587:42,48    0|
1:111:1190,43,452:37,74    0|1:102:960,26,521:40,62
Pseudomonas  1214  8588   T      G      .      PASS
      Ty=SNP;Rk=0.02395;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=T;Sd=1   GT:DP:PL:AD
      0/0:1059:32,2755,20328:1038,21   0/0:1288:16,3518,25081:1272,16   0/0:830:21,2204,16026:816,14
      0/0:907:14,2486,17676:896,11     0/0:993:19,2666,19237:978,15
Pseudomonas  1217  12549_2      A      G      .      PASS
      Ty=SNP;Rk=0.05280;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=A;Sd=1   GT:DP:PL:AD  0|
1:284:1659,27,2378:160,1240|1:379:2586,19,2766:194,185     0|1:231:1509,19,1768:122,109     0|
1:321:1852,30,2710:182,1390|1:260:1820,17,1860:131,129
Pseudomonas  1237  10749  T      G      .      PASS
      Ty=SNP;Rk=0.02921;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=T;Sd=1   GT:DP:PL:AD
      0/0:995:14,2806,19551:987,8      0/0:1252:15,3550,24642:1243,9    0/0:784:14,2240,15461:779,5
      0/0:812:31,2420,16195:811,1      0/0:944:15,2696,18615:938,6
Pseudomonas  1243  860    T      G      .      PASS
      Ty=SNP;Rk=0.02113;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=T;Sd=-1  GT:DP:PL:AD
      0/0:1008:14,2845,19811:1000,8    0/0:1247:15,3535,24542:1238,9    0/0:775:19,2258,15366:772,3
      0/0:782:13,2213,15380:776,6      0/0:921:20,2672,18240:917,4
Pseudomonas  5133  14601  C      A      .      PASS
      Ty=SNP;Rk=0.04497;DT=0;UL=.;UR=.;CL=.;CR=.;Genome=C;Sd=1   GT:DP:PL:AD
      0/0:1458:41,4306,29017:1455,3    0/0:1613:75,4860,32264:1613,0
      0/0:1044:20,3017,20654:1039,5    0/0:1013:48,3054,20264:1013,0
      0/0:1215:46,3631,24253:1214,1
Pseudomonas  6702  8511_1 T      G      .      PASS
      Ty=SNP;Rk=0.05088;DT=1;UL=.;UR=.;CL=.;CR=.;Genome=T;Sd=1   GT:DP:PL:AD  0|
1:200:1360,17,1479:103,97  0|1:196:1177,22,1616:109,87      0|1:118:843,16,843:59,59   0|
1:182:1041,25,1560:104,78  0|1:160:1040,18,1239:85,75
```