

SNP -> DISEASE 软件说明书

一，文件说明

cleaning_scripts/

此文件夹下的文件执行所有的数据清洗工作，最终得到 SNP.db 数据库

cvf_files/

此文件夹下保存需要处理的 vcf 文件

result/

此文件夹下保存所有 vcf 文件处理的结果

SNP.db

1. 此数据库中包含两个表, snp_disease_19 snp_disease_38, 分别对应不同测序的版本号。

```
DORA-MAC:snp ZRC$ sqlite3 SNP.db
SQLite version 3.16.0 2016-11-04 19:09:39
Enter ".help" for usage hints.
sqlite> .tables
snp_disease_19  snp_disease_38
```

2. 对于每个数据库表，其 Columns 的值分别是：

- "index": 索引编号
- "#CHROM": 染色体号
- "POS_START": 变异的起始位置
- "POS_END": 变异的结束位置
- "REF": reference
- "ALT": alteration
- "Variant_class": 关联级别
- "HITag": HITag
- "Group": "Group"

```
sqlite> .schema snp_disease_19
CREATE TABLE "snp_disease_19" (
  "index" INTEGER,
  "disease" TEXT,
  "#CHROM" TEXT,
  "POS_START" REAL,
  "POS_END" REAL,
  "REF" TEXT,
  "ALT" TEXT,
  "Variant_class" INTEGER,
  "HITag" TEXT,
  "Group" TEXT
);
CREATE INDEX "ix_snp_disease_19_index" ON "snp_disease_19" ("index");
```

snp_disease_38.csv / snp_disease_19.csv

数据清洗过后的数据库的 csv 格式

search.py

执行搜寻工作的主文件

二，软件需要安装的包

- numpy
- pandas
- sqlite3

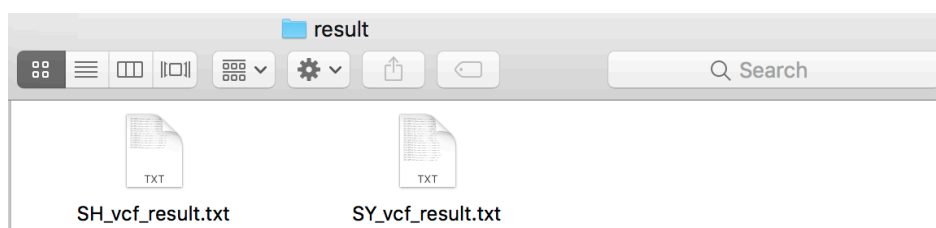
三，程序执行实例

1. 命令行参数接口：

参数一： 数据库名称
参数二： vcf文件所在的文件夹
参数三： 选择测序版本号

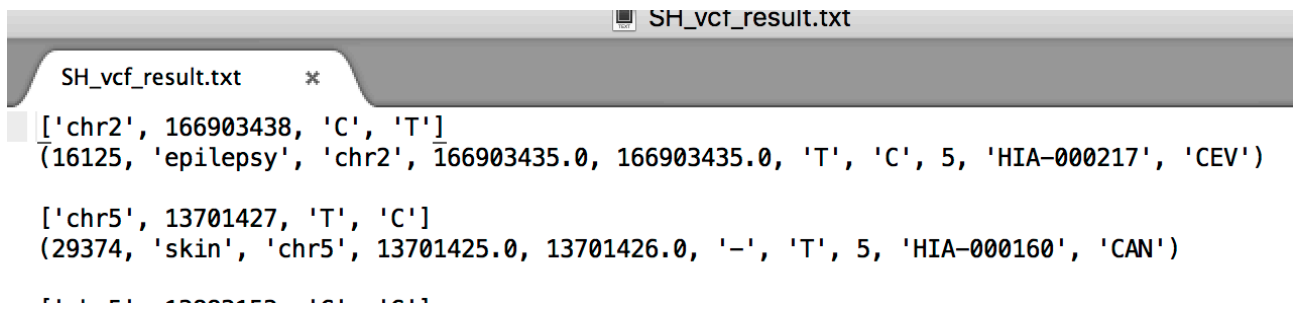
2. 例如，直接执行 `./search.py SNP.db vcf_files 19`，结果保存在 **result** 文件夹中。

```
DORA-MAC:snp ZRC$ ./search.py SNP.db vcf_files 19
precessing the SH.vcf which have 46559 entries..
precessing the SY.vcf which have 46782 entries..
finished.....
```



四，结果描述

连续两行为找到的一条vcf 文件与数据库文件可能存在交叠的记录，第一行为 vcf 文件中的记录，第二行是数据库文件中的记录。



```
SH_vcf_result.txt
SH_vcf_result.txt
['chr2', 166903438, 'C', 'T']
(16125, 'epilepsy', 'chr2', 166903435.0, 166903435.0, 'T', 'C', 5, 'HIA-000217', 'CEV')

['chr5', 13701427, 'T', 'C']
(29374, 'skin', 'chr5', 13701425.0, 13701426.0, '-', 'T', 5, 'HIA-000160', 'CAN')
```