

# RS (Reference SNP) 格式 v2.0 白皮书

2018 年 7 月 30 日

## 1. 什么是 RS 格式?

RS 格式主要是为了定义 SNP Microarray 数据, 是一种文本文件。该格式具有头文件、包含 SNP 芯片产生的数据、基因组坐标等数据信息, 变异类型只包含 SNP 以及 INDEL 等。一个文件中同一个基因组物理位置可包含多个样本。

## 2. RS 格式具体介绍

### 2.1 举例

```
##FileFormat=RSv2.0
##FileDate=20180712
##Platform=ASA-24v1-0-Consort_20022506_A2.bpm;GSGT_1.9.4
##Generated=bioguoke
##VariantsNumber=698500
##Reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##INFO=<ID=CHROM,Number=1,Type=String,Description="chromosome name",Source="1000G",Version="hs37d5">
##INFO=<ID=POS,Number=1,Type=Integer,Description="SNP position",Source="1000G",Version="hs37d5">
##INFO=<ID=ID,Number=1,Type=String,Description="dbSNP rs id",Source="dbSNP",Version="150">
##INFO=<ID=REF,Number=1,Type=String,Description="ref base",Source="dbSNP",Version="150">
##INFO=<ID=NAME,Number=1,Type=String,Description="illumina name/seq name",Source="illumina/defined",Version="ASA-24v1">
##INFO=<ID=FWD,Number=0,Type=Bool,Description="direction of genotype">
#CHROM POS ID REF INFO S1
1 156084877 . . NAME=1:156084877;FWD CC
1 156105743 rs1060502211 . NAME=1:156105743;FWD GG
1 156105833 rs1064796677 . NAME=1:156105833;FWD CC
1 17349122 rs1064794269 . NAME=1:17349122;FWD CC
1 17350508 rs1060503759 . NAME=1:17350508;FWD CC
1 237991690 rs1064796516 . NAME=1:237991690;FWD TT
1 45797134 rs1060501325 . NAME=1:45797134;FWD CC
```

例子中包含 7 个 SNP 的位置、基因型等信息。##后续信息为头文件内容, 其他为 SNP 芯片下机数据信息。

### 2.2 格式说明

1) 头文件以##开头, key 与 value 一一对应, 若有多项, 用英文分号 “;” 隔开

字段	含义
FileFormat	文件格式以及版本号
FileDate	生成数据时间
Platform	检测芯片版本号
Program	生成数据软件、版本以及参数
Generated	生成数据公司/机构/组织等
VariantsNumber	共包含的变异个数

GenomeVersion	基因组信息
Reference	基因组数据链接

2) 头文件##包含上述字段后，包含但不限于以下字段：INFO 等。

INFO 字段信息如下：

##INFO=<ID=ID,Number=number,Type=type,Description=description,Source=source,Version=version>

字段详细信息如下：

字段	含义	包含
ID	描述数据信息字段	包含字段如下： ID、CHROM、POS、INFO、NA0001
Number	字符串长度	暂定方案：1 表示有描述字段，但字段长度不定；0 表示该字段为 flag，只有 True 和 False
Type	数据类型	包含字段如下： Integer, Float, Flag, Character, and String
Description	描述	描述字段含义
Source	描述字段的数据来源	
Version	版本号	

3) 数据字段信息

字段	介绍
CHROM	染色体名称；人常用染色体为 1...22, X, Y, MT
POS	SNP 在基因组的物理位置（1 base）
ID	该字段主要描述该 SNP 位点在 dbSNP 数据库的对应 rs id 信息，若无则用英文 . 表示

REF	reference
INFO	信息描述字段
INFO(NAME)	NAME 字段表示 rs id 名称，可能每个公司对该 SNP 有专属名称
INFO(FWD)	Tag, 正义链
S1	为该样本对应的基因型；若该 SNP 未能检出，用--表示

### 3. 其他待补充