RS（Reference SNP）格式 v1.0 白皮书

2018 年 5 月 18 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=RSv1.0
##fileDate=20180518
##content=GSAMD-24v1-0_20011747_A1.bpm
##source=GSGT_1.9.4
##generated=bioguoke
##totalSNP=700078
##reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##INFO=<ID=ID,Number=1,Type=String,Description="dbSNP rs id",Source="dbsnp",Version="150">
##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=NAME,Number=1,Type=String,Description="illumina name",Source="illunima",Version="GSAMD-24v1">
#ID»    CHROM»  POS»    INFO»   NA0001
rs568121721»    1»      100292476»      NAME=1:100292476;OTHER=xxx»     AA
rs768830171»    X»      9882120»NAME=X:9882120-C-T»     TT
rs371814332»    Y»      2750408»NAME=200610-173»--
.»      MT»     10397»  NAME=200610-22» AA
```

例子中包含 4 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
content	检测芯片版本号
source	生成数据软件以及版本
generated	生成数据公司/机构/组织等
totalSNP	共包含的 SNP 位点数
reference	基因组信息

2）头文件##包含上述字段后，包含但不仅限于以下字段：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
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
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）
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
NA0001	为该样本对应的基因型；若该 SNP 未能检出，用--表示