

基于染色体号和坐标位置转换为RSID

整理: Yongqiang Kong

日期: 2024.5.29

GitHub: <https://github.com/Lonelycube>

基于染色体号和坐标位置转换为RSID


- 1 参考数据准备
- 2 使用R语言进行ID查询转换
 - 2.1 整理input文件
 - 2.2 使用R语言进行rsID转换

参考: <https://zhuanlan.zhihu.com/p/439678589>
<https://zhuanlan.zhihu.com/p/410164485>

1 参考数据准备

下载 hg19 基因组版本的 dbSNP 参考数据: snp150_hg19.txt


- 直链下载: <https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp150.txt.gz>
- 上述连接如果失效, 请重新进入UCSC官网搜索并下载, UCSC官网:
<https://hgdownload.soe.ucsc.edu/downloads.html>

 Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Sequence, Annotation, and Other Downloads

This page contains links to sequence and annotation downloads for the genome assemblies featured in the UCSC Genome Browser. Downloads are also available via our [JSON API](#), [MySQL server](#), or [FTP server](#). Data filtering is available in the [Table Browser](#) or via the command-line [utilities](#).

For access to the most recent assembly of each genome, see the [current genomes](#) directory. Previous versions of certain data are available from our [track archive](#). Data hosted in [Public Hubs](#) exists on external sites. [GenArk](#) (Genome Archive) species data can be found [here](#). All data in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. These data were contributed by many researchers, as listed on the Genome Browser [credits](#) page. Please acknowledge the contributor(s) of the data you use.


[Human](#) 

[SARS-CoV-2 \(COVID\)](#)

[Fruit fly](#)


[Mouse](#)

[Zebrafish](#)

Mammals 

- [Alpaca](#)
- [Armadillo](#)
- [Baboon](#)

下滑，找到hg19版本

- Feb. 2009 (GRCh37/hg19)
- [Genome sequence files and select annotations \(2bit, GTF, GC-content, etc\)](#)
 - [Sequence data by chromosome](#)
 - Annotations
 - [SQL table dump annotations](#) 
 - [Fileserver \(bigBed, maf, fa, etc\) annotations](#) Also see [Data Access](#)
 - [GC percent data](#)
 - [Protein database for hg19](#)
 - [SNP-masked fasta files](#)
 - [SNP151-masked FASTA files](#)
 - [SNP150-masked FASTA files](#)
 - [SNP147-masked FASTA files](#)
 - [SNP146-masked FASTA files](#)
 - [SNP144-masked FASTA files](#)
 - [SNP142-masked FASTA files](#)

下滑，找到snp150.txt.gz, 下载即可

- [snp147OrthoPt4Pa2Rm3.sql](#)
- [snp147OrthoPt4Pa2Rm3.txt.gz](#)
- [snp147Seq.sql](#)
- [snp147Seq.txt.gz](#)
- [snp150.sql](#)
- [snp150.txt.gz](#) 
- [snp150CodingDbSnp.sql](#)
- [snp150CodingDbSnp.txt.gz](#)
- [snp150Common.sql](#)
- [snp150Common.txt.gz](#)

2016-08-18	14:19	2.2K
2016-08-18	14:26	4.0G
2016-08-18	14:08	1.3K
2016-08-18	14:09	934M
2017-08-29	18:44	3.3K
2017-08-29	18:57	5.4G
2017-08-29	19:25	1.7K
2017-08-29	19:25	215M
2017-08-29	19:26	3.3K
2017-08-29	19:27	719M

解压并改名为 snp150_hg19.txt

2 使用R语言进行ID查询转换

2.1 整理input文件

```
chromosome:start
17:69395383
17:69395444
17:69395755
17:69398512
17:69398589
17:69399273
```

2.2 使用R语言进行rsID转换

内存需求较大，建议使用服务器，R语言脚本为：



```
# 确保以下两个R包已安装
```

```

# install.packages(dplyr)
library(dplyr)
library(data.table)

# 读取输入文件: snp_input.txt
tes = read.table("snp_input.txt", header=T, check.names=F, sep="\t")
print("DONE: SNP_input")

# 读取参考文件: snp150_hg19.txt
match = data.table::fread("snp150_hg19.txt", header=T, check.names=F, sep="\t")
print("DONE: SNP_150_hg19")

# 基于参考文件提取rsID, 并保存
# 如果snp150_hg19.txt文件中有对应的RS号, 则比对到test.txt文件中, 如果没有的话, 就变为NA
need = dplyr::left_join(tes, match, by="chromosome:start")
write.table(need, file = "clean.txt", sep = "\t", row.names = FALSE, col.names
=TRUE, quote =FALSE) #保存文件

```

结果示例:

chromosome:start	name
17:69395383	rs2886962
17:69395444	rs917345
17:69395755	rs3047338
17:69398512	rs8073320
17:69398589	rs8077946
17:69399273	rs917344