# schicluster介绍：

schicluster作用在down-sampled Hi-C上，得到imputed Hi-C，以供TOKI和TopDom两种TAD caller识别TAD。

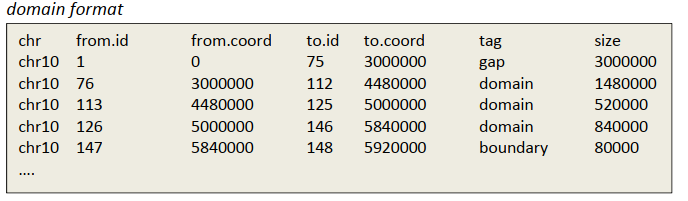
schicluster包括rp，pad，std，sqrtvc等多个参数。其中rp参数控制impute程度，rp取值范围为（0， 1），rp越小，impute程度越高。

# TopDom介绍：

具体参见Manual文件

TopDom包括三步， Step 1 : Generating binSignals by computing bin-level contact frequencies，Step 2 : Detect TD boundaries based on binSignals，Step 3 : Statistical Filtering of false positive TD boundaries。

.domain文件是TAD识别结果，from.id和to.id是一段结构的bin编号，从1开始。tag中的domain是TAD结构，TAD结构中to.id的bin是TAD boundary（根据.binSingal可以看出），gap不是TAD结构（contact稀疏），boundary是第三步中部分domain转换成的结构，应该不是TAD结构。（存疑，<https://github.com/HenrikBengtsson/TopDom/issues/9>有人讨论这个问题）



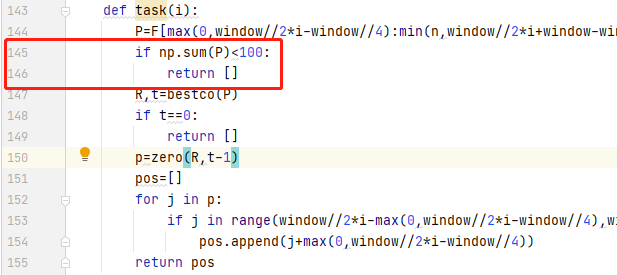
TopDom的第三步设置成了参数statFilter，在默认statFilter设置成True的情况下，对部分染色体的normalized Hi-C进行TAD识别时报错了。



于是我又做了statFilter为False的结果。

# TOKI结果：

TOKI在识别imputed Hi-C时得到的TAD boundary往往不完整（不从0开始，不以最后一个bin结尾）。原因可能是schicluster在impute过程中首先会进行VCnorm（为了RWR可以收敛），最后还会进行sqrtVCnorm。这使Hi-C的每行之和都接近1。而在TOKI.py的第145行对contacts数量小于100的window不予处理，经过VCnorm后的Hi-C在尺度上与原来不一样，很多window都没有100个contacts，可能因为这样导致了TAD boundary不完整。学长是否需要修改一下TOKI来适应schicluster呢？



# 文件结构：

TOKI ------------------------------------------------ TOKI识别TAD的结果

imputed ---------------------------------------- TOKI识别结果schicluster impute之后的down-sampled Hi-C得到的TAD结果

rp0.x ----------------------------------------- schicluster impute过程中取参数rp=0.x的结果

chrx ------------------------------------- 第x条染色体结果

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.TAD----800\_0\_chr1来自于下采样Hi-C文件命名，pad1\_std1\_rp0.1\_sqrtvc来自于schicluster参数

raw --------------------------------------------- TOKI识别未经过schicluster impute的Hi-C得到的TAD结果

chrx ---------------------------------------- 第x条染色体结果

chr1\_800\_0.matrix.TAD--------------------- 下采样Hi-C得到的TAD

chr1\_normed.matrix.TAD------------------- normalized Hi-C得到的TAD

TopDom --------------------------------------------TopDom识别TAD的结果

statFilter=T------------------------------------TopDom设置参数statFilter=True

imputed ------------------------------------------TopDom识别经过schicluster impute之后的down-sampled Hi-C得到的TAD结果

rp0.x ----------------------------------------- schicluster impute过程中取参数rp=0.x的结果

chrx ------------------------------------- 第x条染色体结果

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.topdom.domain

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.topdom.binSignal

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.topdom.bed

.domain和.bed都可以表示TAD结果，.binSingal记录了其他的信息，主要包括记录每个bin是否local minima等。

raw --------------------------------------------- TopDom识别未经过schicluster impute的Hi-C得到的TAD结果

chrx ---------------------------------------- 第x条染色体结果

chr1\_800\_0.topdom.domain等--------------TopDom识别down-sampled Hi-C得到的TAD结果

chr18\_normed.topdom.domain等------------TopDom识别normalized Hi-C得到的TAD结果。有几条染色体的normalized Hi-C在TopDom识别是出现了报错。可能与normalization 方法有关系。（也跟TopDom实现方式有关）

statFilter=F------------------------------------TopDom设置参数statFilter=False

imputed -------------------------------------------TopDom识别经过schicluster impute之后的down-sampled Hi-C得到的TAD结果

rp0.x ----------------------------------------- schicluster impute过程中取参数rp=0.x的结果

chrx ------------------------------------- 第x条染色体结果

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.topdom.domain

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.topdom.binSignal

800\_0\_chr1\_pad1\_std1\_rp0.1\_sqrtvc.topdom.bed

.domain和.bed都可以表示TAD结果，.binSingal记录了其他的信息，主要包括记录每个bin是否local minima等。

raw --------------------------------------------- TopDom识别未经过schicluster impute的Hi-C得到的TAD结果

chrx ---------------------------------------- 第x条染色体结果

chr1\_800\_0.topdom.domain等--------------TopDom识别down-sampled Hi-C得到的TAD结果

chr1\_normed.topdom.domain等------------TopDom识别normalized Hi-C得到的TAD结果