Su

sudo gpasswd -a lzq docker #将登陆用户加入到docker用户组中

newgrp docker #更新用户组

docker ps #测试docker命令是否可以使用sudo正常使用

Ctral +D退出root

重启一下就有权限了

Docker定制镜像

Cd /usr\_storage/lzq/software/paragraph-2.4a/

Docker build . 其中.表示Dockerfile就在当前路径下

/w/00/g/g00/lzq/software/cmake-3.20.0-rc2-linux-x86\_64/bin/cmake ../paragraph-2.4a -DCMAKE\_CXX\_COMPILER=`which g++` -DCMAKE\_C\_COMPILER=`which gcc` -DBOOST\_ROOT=$BOOST\_ROOT

export CXX= /data/apps/gcc/8.3.0/bin/g++

export BOOST\_ROOT= /w/00/g/g00/lzq/software/ boost\_1\_65\_0\_install

/data/apps/cmake/cmake-3.8.2/bin/cmake ../paragraph-2.4a

安装了两三天失败了，悲痛万分。最后选择下载二进制文件

Paragraph仅支持单个VCF输入去genotype二代数据（或者提取多个vcf的info那一列合并）

有人提出用survivor合并个体然后再用paragraph，但是survivor合并之后的断点十分不精确。会影响genotype的结果

While deviations of a few base

pairs can generally be tolerated without issue, deviations

of 20 bp or more reduce recall to around 0.44. For insertions

with differences in breakpoints and/or insertion sequences,

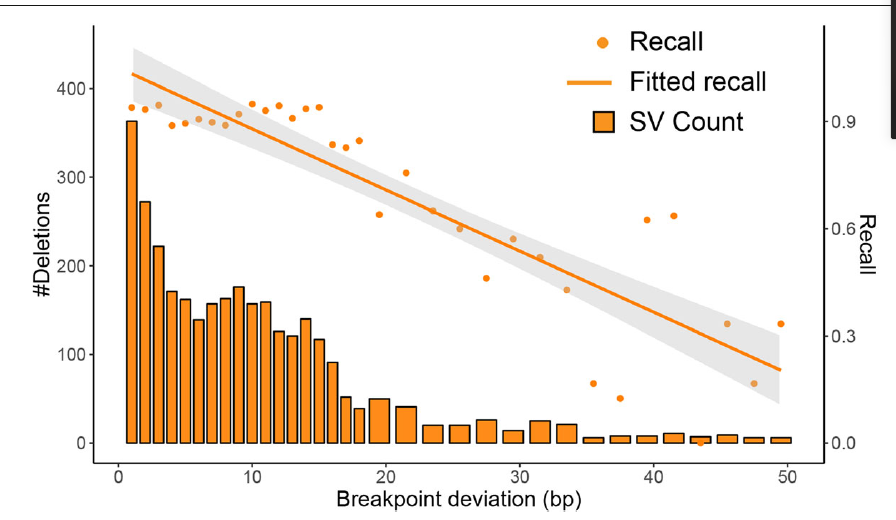
Paragraph recall decreased from 0.88 to 0.66.While deviations of a few base

pairs can generally be tolerated without issue, deviations

of 20 bp or more reduce recall to around 0.44. For insertions

with differences in breakpoints and/or insertion sequences,

Paragraph recall decreased from 0.88 to 0.66



1. So Sniffles measures the agreement of the breakpoints among the individual reads. If the standard deviation of these reads are larger than 5 it reflects this as IMPRECISE. You can see in the info field (STD\_quant\_start=0.707107;STD\_quant\_stop=19.205545. ) that this was triggered mainly from your stop breakpoint. It does not necassarily related to false positives, but rather with the likely imprecision of the breakpoint.
2. Cutesv:the positions of the breakpoints in the reads are quite far from the ground truth breakpoint. Under such circumstances, all the benchmarked SV callers made SV calls, but the sizes and/or positions are incorrect since the read alignments are misleading.

三代vcf

二代比对的Bam bai

Ref.fa 和fai

Sample.txt

less zsp196a-L.minimap2.cutesv.vcf | grep -v 'SVTYPE=TRA'>test.vcf

less -S test.vcf| awk '{if ($0 ~ "#") print $0 ; else if ($2>=150) print $0 }'>test1.vcf

less -S test1.vcf | awk '{split($8,a,";");print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[4]}'>test1.vcf

less test1.vcf | grep -E "#|LG01">test2.vcf

python3 bin/multigrmpy.py -i share/test-data/round-trip-genotyping/candidates.vcf -m share/test-data/round-trip-genotyping/samples.txt -r share/test-data/round-trip-genotyping/dummy.fa -o test

大服务器在线跑是可以的但是提交报错没有pysam模块,其他节点没有安装这个模块

其他版本的pysam

[E::idx\_find\_and\_load] Could not retrieve index file for 'test1/variants.vcf.gz'

conda activate py36

他会先卸载原来安装的版本所以我们没有权限在conda里面安可以

pip install pysam==0.15.3

换了版本以后这个报错就没有啦

在不改变脚本的基础上start和end全部减1

REF=END-START+1

SVLEN= END-START

SNIFFLES 这个软件位置是从零开始的

WARNING Contig1:142667 Padding base in genome is different from VCF. Use the one from genome.

发现报错的位置全是DUP

python3 /usr\_storage/lzq/software/paragraph-2.4a/bin/multigrmpy.py -i test4.vcf -m samples.txt -r /usr\_storage/lzq/work/lrs\_sv/YS.LG.fasta -o test1 --threads 30

DUP就不用减一

POS代表变异位点开始

可是还有很多问题啊

比如群体建立graph 用survivor的合并个体测试了没问题

上面全部用的cuteSV但是cuteSV断点不准确

设置-M 跳过depth\*20的区域For efficiency, it is recommended to manually set the "-M" option (maximum allowed read count for a variant) to skip these high-depth regions. We recommend "-M" as 20 times of your mean sample depth.

群体的genotype尽量单个单个个体来分型

Sniffles没有ref怎么办呢

50 1 1 1 -1 -1

50 5 1 1 -1 -1

less -S sniffles\_merged\_gt\_SURVIVOR\_50bpdist\_typesave.vcf |awk '{split($8,a,";");split(a[3],b,"=");if ($0 ~ "#") print $0 ;else if( b[2] <= 1000000 && b[2] >= -1000000) print $0}'>sniffles\_merged\_gt\_1Mb.vcf

less -S test.vcf| awk '{split($8,a,";");if ($0 ~ "#") print $0 ; else if ($2>150) print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[4]";"a[7]"\t"$9"\t"$10"\t"$11"\t"$12"\t"$13"\t"$14"\t"$15}'>test1.vcf

awk '{split($8,a,";");split(a[2],b,"=");c=b[2]-1;d=$2-1;if ($0 ~ "##") print $0 ;else if ($0 ~ "#") print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9"\t""ZL2020-SXHZ""\t""zsp191a-L""\t""zsp192a-L""\t""zsp194c-L""\t""zsp196a-L";else if ($0 ~ "INV"||$0 ~"DUP") print $0 ;else print $1"\t"d"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[1]";""END="c"\t"$9"\t"$10"\t"$11"\t"$12"\t"$13"\t"$14"\t"$15}' test1.vcf >cutesv.para.vcf

1089319

/usr\_storage/lzq/software/paragraph-2.4a/lib/python3/grm/vcfgraph/vcfgraph.py

大片段的插入

WARNING LG10:17452163 Genome REF is different from VCF. Use genome REF.

>6kb

less -S test.vcf| awk '{split($8,a,";");if ($0 ~ "#") print $0 ; else if ($2>150) print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[4]";"a[7]"\t"$9}'>test1.vcf

awk '{split($8,a,";");split(a[2],b,"=");c=b[2]-1;d=$2-1;if ($0 ~ "##") print $0 ;else if ($0 ~ "#") print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9;else if ($0 ~ "INV"||$0 ~"DUP") print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9"\t"$10 ;else print $1"\t"d"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[1]";""END="c"\t"$9}' test1.vcf >cutesv.para.vcf

less cutesv.ok.gt.recode.vcf | grep -v 'SVTYPE=TRA'|awk '{split($8,a,";");split(a[3],b,"="); if ($0~"SVTYPE=INS" || $0~"SVTYPE=DEL") print $0;else if (b[2]<5000 && b[2]>-5000) print $0}' >test1.vcf

less test1.vcf | grep -v 'Contig' >test2.vcf

awk '{split($8,a,";");if ($0 ~ "#") print $0 ; else if ($2>150) print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[4]";"a[7]"\t"$9}' cutesv.ok.gt.recode.vcf >test1.vcf

awk '{split($8,a,";");split(a[2],b,"=");c=b[2]-1;d=$2-1;if ($0 ~ "##") print $0 ;else if ($0 ~ "#") print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9;else if ($0 ~ "INV"||$0 ~"DUP") print $0 ;else print $1"\t"d"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[1]";""END="c"\t"$9}' test1.vcf >test2.vcf

awk '{if ($4 != "N" && $5 != "N") print $0}' test2.vcf >cutesv.para.vcf

先测试

/usr\_storage/lzq/software/conda\_evns/LRS\_SV/bin/vcftools --vcf /usr\_storage/lzq/work/lrs\_sv/1000\_1\_merge/cutesv.para.vcf --exclude-positions exclud.pos --recode --recode-INFO-all --out cutesv.ok.gt.vcf

正常合并survivor

按照position提取原始位点信息

先测试vcf去除warning的位点 不去除也可以

以及去除N的位点（必须）

过滤掉>5kb的inv和dup

计算前二十个主成分的支持率就好了，画一个累加图，如果Pca主成分前两个累加达到99%，就用这部分数据分析便可以了，也就是只需要画PCA1和PCA2的图就可以了

/usr\_storage/lzq/software/SURVIVOR-master/Debug/SURVIVOR merge sniffles\_vcf\_files\_raw\_calls.txt 1000 1 1 -1 -1 -1 sniffles\_merged\_SURVIVOR\_1kbpdist\_typesave.vcf

/usr\_storage/lzq/software/SURVIVOR-master/Debug/SURVIVOR merge sniffles.gt.txt 1000 1 1 1 -1 -1 sniffles\_merged\_gt\_SURVIVOR\_1kbpdist\_typesave.vcf

less -S ../sniffles\_merged\_SURVIVOR\_1bpdist\_typesave.vcf |awk '{split($8,a,";");split(a[3],b,"=");if ($0 ~ "#") print $0 ;else if( b[2] <= 50000 && b[2] >= -50000) print $0}'>sniffles\_10kb\_1bpdis.vcf

less sniffles\_10kb\_1bpdis.vcf|grep -v 'SVTYPE=TRA'>sniffles\_10kb\_1bpdis\_nonbnd.vcf

/usr\_storage/software/bedtools2/bin/bedtools getfasta -fi /usr\_storage/lzq/data/YS.LG.fasta -bed sniffles\_10kb\_1bpdis\_nonbnd.vcf -fo sniffles\_1bp\_seq

python change\_N\_refseq.py

less -S sniffles\_10kb\_1bpdis\_nonbnd.vcf | grep '#' > refseq\_sniffles\_10kb\_1bpdis\_nonbnd.vcf

删除样品注释行

cat ref\_sniffles\_10kb\_1bpdis\_nonbnd.vcf >> refseq\_sniffles\_10kb\_1bpdis\_nonbnd.vcf

awk '{if ($0 !~ "#") print $1"\t"$2}' ../sniffles\_merged\_gt\_SURVIVOR\_1kbpdist\_typesave.vcf > sniffles.pos

/usr\_storage/lzq/software/conda\_evns/LRS\_SV/bin/vcftools --vcf refseq\_sniffles\_10kb\_1bpdis\_nonbnd.vcf --positions sniffles.pos --recode --recode-INFO-all --out sniffles.para

合并为bnd

awk '{split($8,a,";");split(a[2],b,"=");LEN=b[2]-$2 ;if ($0~"SVTYPE=INS" || $0~"SVTYPE=DEL") print $0;else if (LEN<5000 && LEN>-5000) print $0}' sniffles.para.recode.vcf >test.vcf

less test.vcf | grep -v 'Contig' >test1.vcf

awk '{ if ($0 ~ "#") print $0 ; else if ($2>150) print $0}' test1.vcf >test2.vcf

awk '{split($8,a,";");split(a[2],b,"=");c=int(b[2])+1;d=$2+1;if ($0 ~ "#") print $0 ;else if ($0 ~ "INV"||$0 ~"DUP") print $0 ; else print $1"\t"d"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"a[1]";""END="c"\t"$9}' test2.vcf >test3.vcf

awk '{if ($4 != "N" && $5 != "N") print $0}' test3.vcf > sniffles.ok.vcf

usr\_storage/lzq/software/SURVIVOR-master/Debug/SURVIVOR merge sniffles\_vcf\_files\_raw\_calls.txt 0 1 1 1 -1 -1 sniffles\_merged\_SURVIVOR\_1kbpdist\_typesave.vcf

Exception: Different padding base for REF and ALT at LG01:39833

awk '{split($8,a,";");split(a[2],b,"="); if ($0 ~ "#") print $0 ;else if ($0 ~ "INV"||$0 ~"DUP") print $0 ;else if ($0 ~ "INS") print $1"\t"b[2]"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9 ;else print $0 }' test2.vcf > test3.vcf

nohup bash pda07-a.com 2>pda07-a.com.log &

nohup bash pda07-b.com 2>pda07-b.com.log &

nohup bash pda08-a.com 2>pda08-a.com.log &

nohup bash pda09-a.com 2>pda09-a.com.log &

nohup bash pda10-a.com 2>pda10-a.com.log &

nohup bash pda11-a.com 2>pda11-a.com.log &

nohup bash pda11-b.com 2>pda11-b.com.log &

nohup bash pda12-a.com 2>pda12-a.com.log &

nohup bash pda12-b.com 2>pda12-b.com.log &

nohup bash pda14-a.com 2>pda14-a.com.log &

nohup bash TN1410D3187.com 2>TN1410D3187.com.log &

1、cutesv

Conitg 19 pos:71931 ref:T

tail -n 1 test.fa|while read id;do echo "${id:71930:1}" ;done ::::T

说明cutesv软件是从1开始的

1. Sniffles

同理检查发现sniffles 的pos也是从1开始的

大服务器在线跑是可以的但是提交报错没有pysam模块,其他节点没有安装这个模块

其他版本的pysam

[E::idx\_find\_and\_load] Could not retrieve index file for 'test1/variants.vcf.gz'

conda activate py36

他会先卸载原来安装的版本所以我们没有权限在conda里面安可以

pip install pysam==0.15.3

换了版本以后这个报错就没有啦