SV 检测流程使用说明

V2.1.1 版本

简介

本流程适用于 stLFR 技术以及类似的 linked read 建库的数据。目前在 stLFR 数据上测试运行，理论上也适用于其他 linked read 数据，可通过将 barcode 转换为 read id 上的"read_id#XXX_XXX_XXX" 格式来进行分析测试。
本流程利用 barcode 信息，检出结构变异（SV）的断点信号，例如：平衡易位，INV，部分的缺失重复，以及更复杂的结构断点，可与 cnv 结果，phase 结果一起，联立推导染色体的真实结构变异，其检测精度受限于 linked read 在 DNA 分子上的分布密度和分子长度。例如，以 1.5ng 起始量建库的 stLFR 试剂盒产出 30x 有效深度数据，约可以保证 20K 以上 SV 的检测精度。

目录结构:

```
bin  data  example  LFR-sv  lib  Readme.docx  run.sh  tools
```

主程序：LFR-sv
bin：运行流程所需的相关程序
lib：可能缺少的依赖的支持库
data：预制的一些数据库文件，非必需，可指定其他数据库文件
example：示例说明相关文件，非必需
tools：小工具目录，非必需
Readme.docx：本说明文件
run.sh：运行脚本示例

参数说明:

```
Name:
        LFR-sv
        version 2.1.1
Function:
        Detect the SVs from stLFR WGS data
Usage:
        LFR-sv -bam prefix.sorted.markdup.bam -o out.result.dir
Options:
        -bam <string>   original sorted and markduped bam file,if the index dose no
t exist, will be created.[necessary]
        -out <string>   output SV dir.[necessary](warning:if exists, the output dir
 will be cleaned first!!!!!)
        -ncpu <int>     thread number for running pipeline.[default 1]
        -bar_th <int> at least N read pairs in one barcode.[default 8]
        -seg_th <int> at least N read pairs in one segment.[default 4]
        -gap <int> define the gap size which should be considered as different segm
ent.[default 20000]
        -size <int> output SV length.[default 20000]
        -is <int> proper IS size for read pair library, read pairs with too large I
S will be abandoned.[default 300]
        -bin <int> bin size for cluster the segments.[default 2000]
        -merge1 <int> N continue outline bins could be considered as the same break
 point and will be merged into one evidence.[default 5]
        -merge2 <int> SVs nearby under N binsize will be considered as one event.[d
efault 5]
        -mmax <int> the max SVs allowed in one event.[default 4]
        -low1 <float/int> single end barcode counts threshold, 0-1 float: higher th
an X percentage counts; 1> int: higher than X counts.[default 0.95]
        -low2 <float/int> end to end barcode counts threshold, 0-1 float: higher th
an X percentage counts; 1> int: higher than X counts.[default 0.9995]
        -ex1 <float> when low1 is a float of 0-1, exclude the bins which depth unde
r ex1.[default 0.2]
        -ex2 <float> when low2 is a float of 0-1, exclude the bins which depth unde
r ex2.[default 0.2]
        -phase <string> formatted phase result directory including phased barcode a
nd region by chromosome.[default NULL]
        -bl <string> black list file(BED format).[default NULL]
        -cl <string> sorted control list file(BEDPE format).[default NULL](Be sure
the chromosome and position are sorted in one line!!!)
        -sc <int> allow max sv counts for the same position in one direction.[defau
lt 4]
        -human <Y/N> for Homo sapiens,keep only [1234567890XYM] chromosome.[default
 N]
        -qc1 <float> valid read pair ratio for SV detection.[default 0.60]
        -qc2 <float> average read pair count for one barcode.[default 30]
        -qc3 <float> average segment end count for one bin.[default 8]
        -rlen <int> read length of one read.[default 100]
        -mlen <int> physical limit for the long DNA segment.[default 400000]
        -help Show this message.
```

-bam <string> readname 中带 barcode 信息的 rmdup，sort 后的 bam 文件，

-out <string> 输出目录，请注意每次运行程序时该目录会被清空

-ncpu <int> 多线程参数

-bar_th <int> 设定一个 barcode 下拥有的最少 readpair 数目

-seg_th <int> 设定一个 barcode 中每个 segment 上最少的 readpair 数目

-gap <int> LFR 数据中，一个 barcode 下 readpair 间间距大于此值时会被拆分为 2 个 segment，最好为 bin 值整倍数

-size <int> 设定检出 SV 精度，流程将仅输出大于此大小的 SV 结果，应大于等于 gap 值。

-is <int> read 对插入片段长度估计，用于筛选 proper mapped readpair

-bin <int> 检测 SV 的基本窗口大小

-merge1 <int> 支持同一个 SV 的信号可能落在附近 N 个窗口内，这些窗口会被合并考察

-merge2 <int> 如果检出的 SV 条目之间某端的位置距离小于 N 个窗口大小，会被视为一

个 event。例如，一个倒位会产生 2 个不同类型 SV 断点，但位置靠近，会被归类为同一 event

-mmax <int> 一个 event 允许最大存在的 SV 条目数，当一个 event 内包含过多的条目数时该 event 会被视为假阳不被输出

-low1 <float/int> 单端聚类时 barcode 支持数过滤，超过此阈值被视为高可信断点，如果为大于 1 的整数，则高于该支持数，如果为 0-1 的浮点数，则为统计拟合后高于此分位数的支持数

-low2 <float/int> 端到端共享 barcode 支持数过滤，超过此阈值被视为高可信断点，如果为大于 1 的整数，则高于该支持数，如果为 0-1 的浮点数，则为统计拟合后高于此分位数的支持数

-ex1 <float> 配合 low1 参数使用，当 low1 为浮点数时，在拟合分布的时候，去掉一定比例的低端极值部分，使得拟合结果精确

-ex2 <float> 配合 low2 参数使用，当 low2 为浮点数时，在拟合分布的时候，去掉一定比例的低端极值部分，使得拟合结果精确

-phase <string> 处理后的 hapcut pahse 文件，详情可见 tools/gen_phase/phase.readme 以及下文关于 phase 部分的说明

-bl <string> bed 格式的不良区间，SV 有一端落入此区间会被标记为 BAD_REGION

-cl <string> bedpe 格式的 control 文件，SV 落入此区域会被标记为 COMMON
注意对于每一行应该是有序的,排序的标准为：染色体名如果相同，则对位置大小排序，小的在前，大的在后；染色体名若不同，则按 chrA 与 chrB 去掉 chr 字符后如均是纯数字，则按数值大小排序，如非均纯数字，则按字符串大小排序，小者在前，大的在后。例如
chr1 100000 300000 chr1 500000 800000
chr4 100000 300000 chr5 500000 800000
chr12 2200000 3000000 chr15 500000 800000
12 100000 300000 19 500000 800000
abc 100000 300000 cde 500000 800000
m 1050000 3080000 n 500000 800000

-sc <int> 设定一个断点所允许对应的最大 SV 个数，通常情况下 SV 断点对应的 SV 个数不超过数个，例如，对双倍体而言，极限情况会出现 4-5 个

-human <Y/N> 被检测物种为人类时，设定为 Y，将仅输出 1-22，XY 染色体结果

-qc1 <float> 质控指标 1，segment 上的有效 readpair 数占总有效 readpair 数目比例

-qc2 <float> 质控指标 2，有效 barcode 上平均 readpair 数目

-qc3 <float> 质控指标 3，每个窗口内理论上的 segment 断点覆盖

-rlen <int> 测序数据的 read 长度

-mlen <int> linked read 的 DNA 分子长度的物理极限，例如目前的 stLFR 技术约为 400k

名词解释

一个片段（segment）：一段被认为没有发生变异的 DNA 片段，read 在其上按一定统计规律分布

一个标签（barcode）：一个 barcode 下，可能存在一个或多个 segment，其上的所有 read，为同一个 barcode 所标记。

单端断点：表征多个片段（segment）的某一端断裂在同一个位置，例如：chr1 1000000 R（Right）

SV 断点（端到端）：可以表征一个 SV 的独立断点，由二处共有许多相同的 barcode 的单端断点组成，例如：chr1 1000000 R - chr1 2000000 R

一个事件（Event）：一个真实的 SV 变异，可能由一个（DEL）或者多个（INV，TRA 等）SV 断点组成。例如一个倒位：（chr1 1000000 R - chr1 2000000 R）&& （chr1 1000000 L - chr1 2000000 L）

结果文件

按照生成时间顺序排列：

sin 文件：单端聚类的结果。

Gap 文件：DNA 片段上 readpair 之间的间距，可用于对样本数据的的统计分析。

Stat 文件：对有效 readpiar 和 barcode 的统计，用于 qc 质控计算。

lnd.all 文件：最初生成的 SV 候选文件，包括低质量的条目，可供排查遗漏，以及 control 集合的建立，其内 id 标识对应 sin 文件中的 id。

judge 文件：格式化以及 split 后，经过多条件过滤后的 SV 候选结果。

filter 文件：judge 文件进一步通过物理关系过滤后的结果。

Region 文件：filter 文件经过 bl 和 cl 文件的进一步标记结果。

final 文件：最终的结果文件，分为二个，一个是 final 文件，一个是 final.NoRegionFilter。区别是后者在输出的时候未考虑 region 的标记（BAD_REGION，COMMON）。

Final 文件 header 说明：

EventID 事件 ID，可能包含同一相关事件的多个 SV，例如倒位，易位等，同一事件包含多个 SV 断点

SvID SV 断点 ID,每个 SV 断点拥有唯一 ID

BreakID1 子断点 ID1，可在 sin 文件中根据此 ID 核查相关详细信息

BreakID2 子断点 ID2，可在 sin 文件中根据此 ID 核查相关详细信息

ChrA 染色体 1

PosA 位置 1

ChrB 染色体 2

PosB 位置 2

ShareBarcode 共享 barcode 支持数

RealType 断点真实链接类型的组合方式，分别对应 A 位置和 B 位置，L 表示在 segment 左边断开，R 表示在 segment 右边断开。（LL，RL，RR，LR）

SimpleType 断点对应的可能通俗简单类型。

ComprehensiveFilter 综合各个过滤条件的综合可信度判定

Heatmap 通过共享 barcode 热图的质控，给出对应左右双端三个变化值

Phase 根据 phase 结果得出的可信度判断，单体型别结果，以及该 SV 所处的单体型 block

MapQ 根据 read 比对质量的质控

BlackList 根据给定的 blask list 的质控

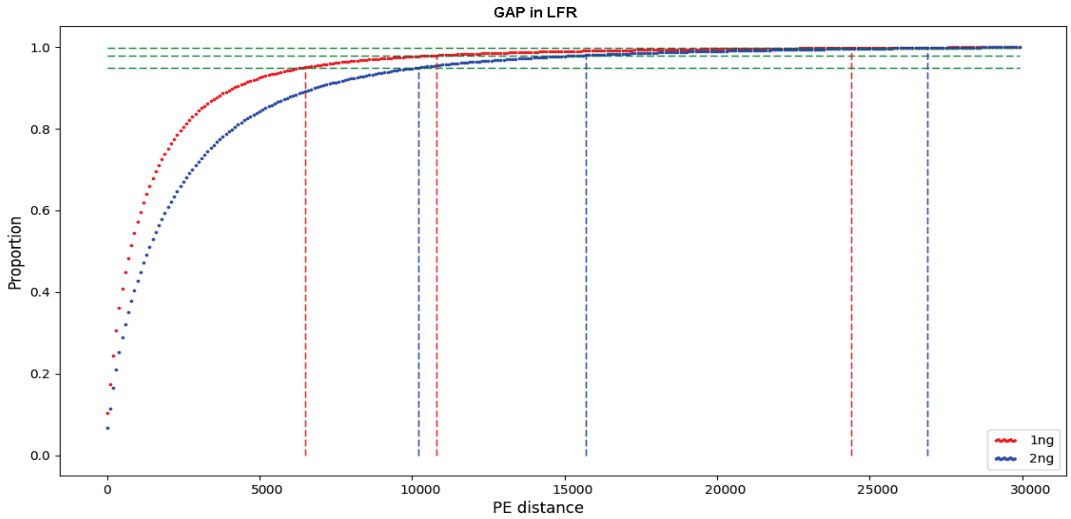
ControlList 根据给定的 control list 的质控
SegmentCheck1 根据不同 LFR segment 的质控
SegmentCheck2 根据相同 LFR segment 的质控
SVchain 根据 LFR segment 将 SV 串联成链，由于目前 LFR 片段长度不长目前效果不显著

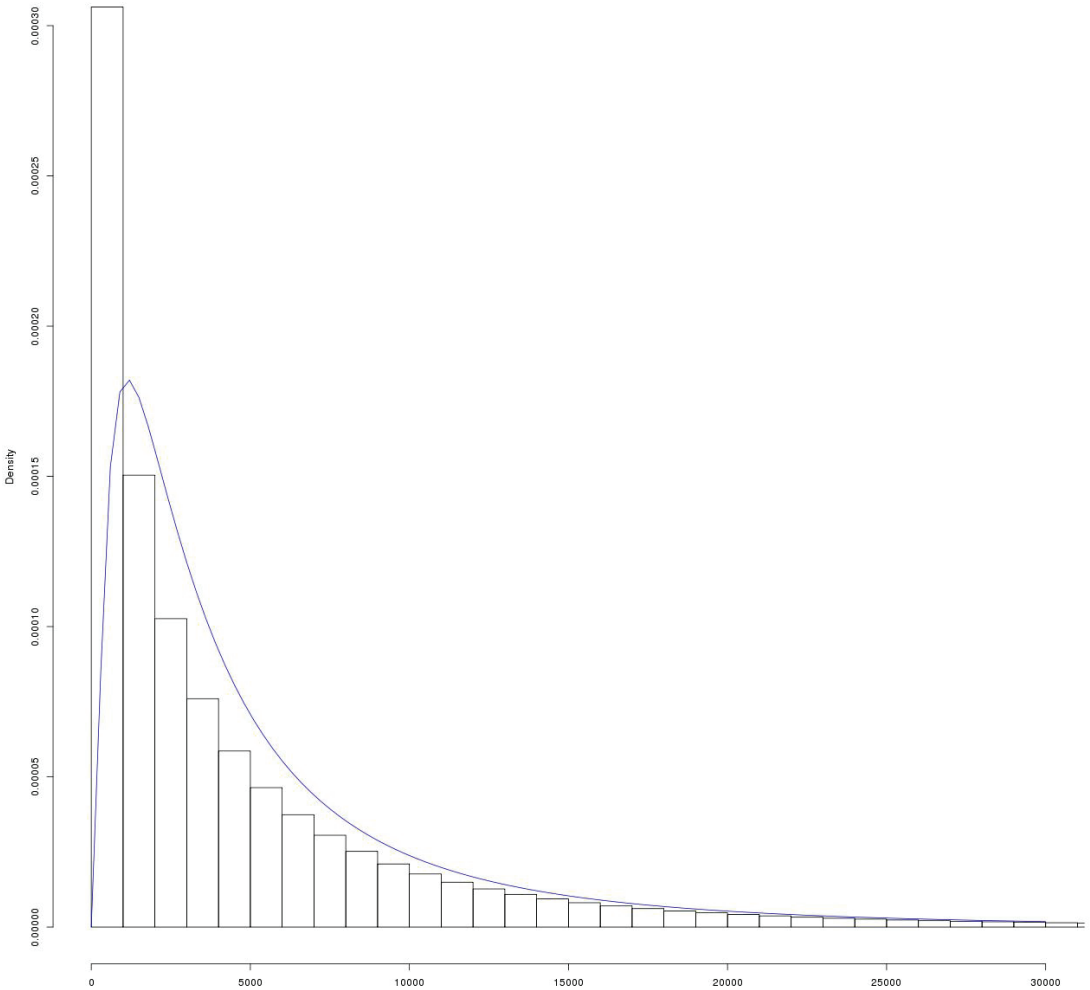
参数选取的说明（生信分析人员重点阅读）

注意: 由于流程没有限定物种或者数据，样本条件，参考序列，测序仪以及建库 protocol 等等都会影响。分析人员应该参考此节内容，针对一些标准样本（验证样本）进行分析，得到符合需求的参数配置，否则可能和理想结果差别较大。

bar_th 和 seg_th 参数：这二个参数分别控制一个 barcode 下和一个 segment 下的 read 对数，可以根据实际的数据情况酌情修改。如果数据量比较足够或者 DNA 片段较长，这个值可以设置得大一些，结果特异性会更好。
通常不建议 bar_th 参数低于 seg_th 的 2 倍，因为对一个 SV 有贡献的断点，需要至少 2 个片段构成。针对 stLFR 数据，流程默认值已经相当低，除非特殊情况，否则不建议继续降低。

bin，merge 和 gap 参数：
gap 参数本质上等同于可能的检测精度参数，即，小于这个长度的 SV，不会被检测出来，根据不同建库的 LFR 片段上 read 的覆盖密度，可以增加或者减小。
（请注意 gap 与 size 参数的区别，gap 源于文库的统计结果表示 LFR 建库的基本特性，由 LFR 文库决定，size 则是检测者期望、关心大小的 SV，例如对于平衡易位，则可以将 size 值设置为相当大的一个值）
bin 参数同 gap 参数，可以由片段上 read 的间距统计来确定，太大则检测的位置精度降低，太小则会导致落入一个窗口的 read 太少，导致 SV 无法检出，但是平衡精确度和检出度的一个参数
merge 参数：merge 控制 segment 断点的合并，由于比对以及 read 在 LFR 片段上分布的缘故，有时支持同一个 SV 的检出结果（设置了一定 bin 值的情况），会落入窗口或者几个 SV 条目中，因此需要 merge。即将支持同一个 SV 的多个截止位置不一片段末端统一或者相关的 SV 断点纳入到一个 Event，方便结果查看。





对于全新数据的测试，分析人员可通过流程生成的 gap 文件进行统计分析以调教参数（如图），比较推荐的配置为：

Bin：60%-65%分位数值

Merge：90%-95%分为数值

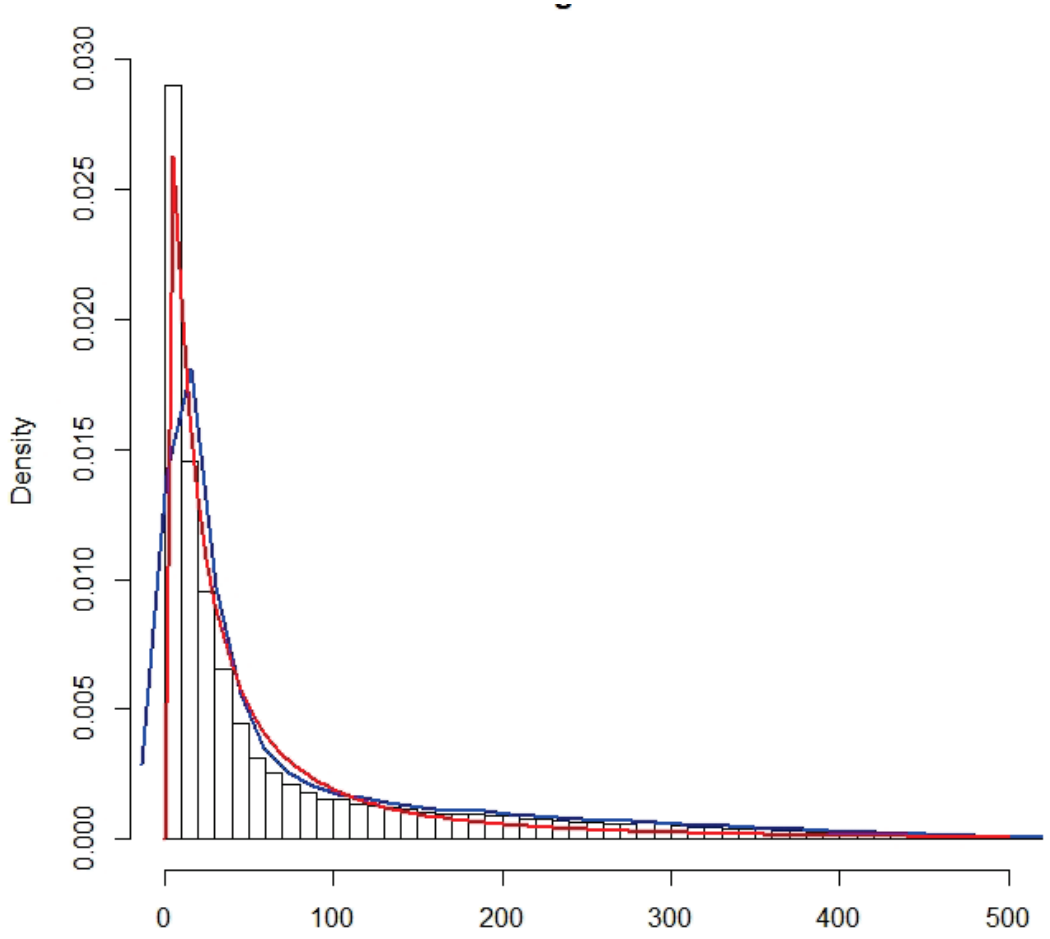
Gap：98%-99%分位数值

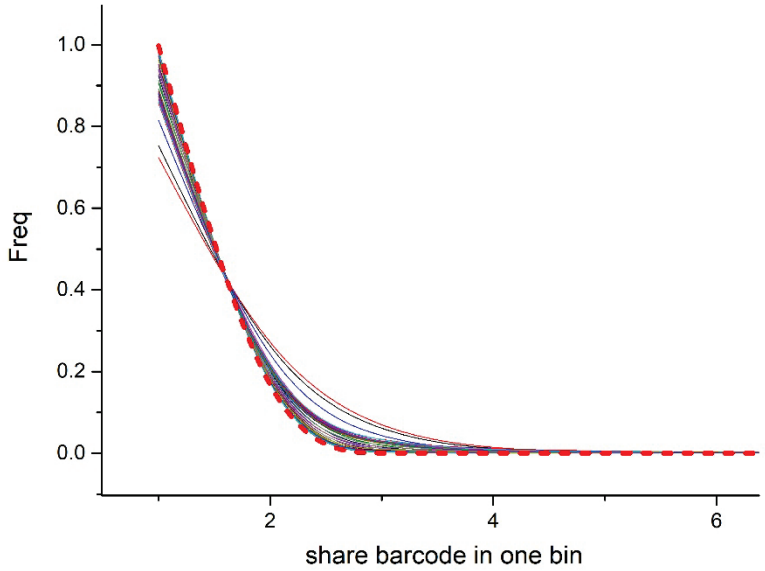
流程在运行时会给出参考建议，分别为 65%，93%，98%值。

而在实际样本分析时，如推荐数值与标准样本推荐值偏差较大，则考虑样本数据异常，对上游数据进行问题排查。

low 参数：

控制灵敏度和特异性的最关键性参数，low1 和 low2 均可以使用 2 种值：0~1 的浮点和正整数，分别意味置信度和硬支持条数。默认情况下，为了适应不同的测序深度，选择为置信度。low1 控制单端断点生成时的深度控制，low2 则控制 SV 断点生成时的深度控制。





在 SV 初筛时，分别对单端和双端的支持进行分布统计和拟合，得到比较合适的阈值（特殊情况也可手动指定 hard cutoff）

其中通过初步 low1 和 low2 质控的才会被输出，然后进行 merge，对 merge 后的 low1 和 low2 再次进行质控，高质量的会被送入下游分析，而全部结果则会保留在 lnd.all 文件中待查。生信分析人员可根据标准品分析确定比较适合的 low1 和 low2

（ex 参数轻易不建议改动，该参数控制拟合精度，如上图 1 中红蓝二个拟合区间的区别，有经验的分析人员，可通过对 share.dat 和 single.dat 二个中间结果文件进行拟合调试该参数）

建议：

（1） 在数据质量较好时，但没有建立好的 control list 的时候，可以选择（low1，low2）为（0.97，0.9999）的组合，可以保证检出结果的特异性。或者适当降低其中一个值，例如（0.95,0.9999）（流程默认）等。

（2） 在建立好 control list 的情况，可以适当降低阈值（0.95,0.9995），实现比较好的灵敏度。

（3） 特殊情况，例如数据不好，需要有目的的寻找某个大致位置的变异，或者根据核型只关注不同染色体的平衡易位，对总特异性要求不高时。可采用（0.95,0.999）甚至极限（0.90,0.999）的参数。此参数可能会产生较多的候选结果，运行时间较长，特异性较差，但是可以最大限度保证灵敏度，例如在 QC 指标提示异常的样本，需要有针对性的分析的时候使用。

bl 和 cl 参数：

即 black list 以及 control list，属于这些 list 的区间，或者 SV 会被 mark 或者过滤。

black list 通常源于参考基因组的一些 gap，segdup，repeat 区等等，亦可添加一些总结的自定义区域。

而 control list 则是 common SV list，通常对于不同数据，不同建库条件，不同测序平台，建议新建立，可以通过对比同批次样本，筛选出高频率出现的 SV 进行过滤，一个优秀的 control 库对于结果的特异性有非常好的效果（如果条件允许，不同性别最好分开建立效果更好）。可以通过对一批稳定样本的 lnd.all 文件进行对比分析，来建立 control list。

qc 参数：
基于稳定条件的样本的质控，流程默认参数为 stLFR 试剂盒 1.5ng 版本目前的统计。
对于尚未确定 qc 的样本类型，可根据标准品的统计结果进行实际情况进行调整。
从 stat 文件中可以得到三个值:
A、有效的 readpair 总数
B、对 segment 有贡献的 readpair 总数
C、有效的 barcode 总数

qc1 B/A
qc2 B/C
qc3 C*2/(ref_total_len/bin_size)

对于已经确定 qc 参数的某类型样本，如果提示 qc 异常，请针对上游数据进行检查。

小工具

Phase 文件生成工具：
将 hapcut2 生成的文件格式化为流程可用的 -phase 参数输入文件，详情见 tools/gen_phase/phase.readme（仅适用 2 倍体）

热图生成工具：
根据给定的坐标范围，绘制共享 barcode 热图，可以通过热图人工核查一个 SV 的具体结构或者是否假阳性 SV
使用方法见 tools/plot_script/readme.txt
如果观察同染色体的较小范围请用 Stat_share.pl
如果观察不同染色体，或者距离较远二个范围请用 Stat_share_dif.pl
针对流程输出结果文件，可使用 bat_plot.pl 脚本实现批量画图。

单体型生成工具：
主要针对不同染色体，例如平衡易位，但理论上同染色体也能通，不排除有 bug。
利用共享 barcode 信息，原始 hapcut2 结果，以及 gen_phase 工具生成的结果，将杂合 SV 进行单体型判断，并将该 phase block 中的最大长度的单体型输出，一个 SV 生成 4 条单体型。（仅适用 2 倍体）
详细使用说明见 tools/make_Haplotype/readme.txt

Control 集合生成工具：
利用流程结果中的 all 文件，集合多个样本，生成由于比对，参考序列等因素造成的过滤集合。
详细使用说明见 tools/make_control/readme.txt

得到的结果格式为 bedpe 格式，示例：
chr4 49094000 49106000 chr10 39078000 39080000 0 0 11 0

其中前 6 列为标准的 bedpe 格式，后 4 列分别为 RL，LR，LL，RR 连接关系出现的次数。可通过对次数的过滤或者占总样本频数的过滤，得到最终的 control 文件。
请注意，得到的 control 文件行与行之间是相互独立可乱序的，但每一行中是符合**参数说明**中关于 cl 参数说明的排序规则的。

说明：data/human_hg19_2000_20000_20000_0.9995_0.95_withchr.conlist
即是用 hg19 的参考序列，2000 为 bin 大小，20000 的 gap 大小，20000 的 size 大小，分别以 0.95 和 0.9995 为 low1 和 low2 得到的 contorl 集合

适用范围：理论上可以适用于同样参考序列下，同等或者更严格的运行参数，例如 size 大于 20000，low1 or low2 大于 0.95 和 0.9995 等。

FAQ

Q：遇见类似错误：
Can't load '/tmp/par-7368616f6c6962696e/cache-5682119e4222baeda8c33df8d64aeff658922f95/82735321.so' for module Bio::DB::HTS: libhts.so.2: 无法打开共享对象文件：没有那个文件或目录 at /home/guojunfu/perl5/lib/5.28.0/x86_64-linux-thread-multi-ld/DynaLoader.pm line 193.

A: 动态库的打开错误，请在运行脚本前配置临时环境变量或者添加如下语句到 sh 脚本
export LD_LIBRARY_PATH=流程目录/lib:$LD_LIBRARY_PATH

Q: final 文件里，为什么会有质控为 Failed 的条目

A: 流程采用聚类和回收机制，即因为有一些真阳性结果会由于各种条件的因素，会被误判造成假阴性，因此在最终结果的 PASS 条目中，会去搜寻与该 PASS 相关条目也纳入一个 event 中，有助于降低假阴性和辅助判断。同理，如果一个 PASS 的结果，拥有很多相关条目，则有可能是一个假阳性结果，-mmax 参数控制此过滤的的阈值，可以酌情设置。
同时，也考虑调整 sc 参数，大多数情况的真实 SV 的一个位置只存在断点（意思是断点某一段的同一位置，例如双倍体情况，没有发生 dup 的话，一个位置一个方向最多可能存在二种突变），某些情况下可以考虑增大此值。

Q: 结果中，phase 信息比较复杂，如何看
A：
例如“PASS:0|1:0.1869:0.1714:0.0385:11:15:51092968-51593456:11:15:51092968-51593456”
该条目为冒号分割,含义为:

phaseQC	Haplotype	ratio1	ratio2	ratio3	chr1	chr1_phase_block_region_index (gen_phase生成region文件行数)	chr1_phase_region	chr2	chr2_phase_block_region_index (gen_phase生成region文件行数)	chr2_phase_region
PASS	0|1	0.1869	0.1714	0.0385	11	15	51092968-51593456	11	15	51092968-51593456

质控规则为:

ratio 1,2,3 为 **phased barcode 比例**。

其中任何一个失败，即未能 phase 的 barcode 比例超过 75%，则质控结果为 NULL，单体型显示为 UNPHASED；

ratio1:支持染色体 1 断点的 barcode 中在 0 单体型上占总 phased 的 barcode 数目的比例,如该位置未能 phase 的 barcode 比例超过 75%,则显示未能 phase 的 barcode 比例;

ratio2:支持染色体 2 断点的 barcode 中在 0 单体型上占总 phased 的 barcode 数目的比例,如该位置未能 phase 的 barcode 比例超过 75%,则显示未能 phase 的 barcode 比例;

ratio3:共享 barcode 在 0 单体型上占总共享 phased 的 barcode 数目的比例,如该位置未能 phase 的 barcode 比例超过 75%,则显示未能 phase 的 barcode 比例.如果染色体 1 与染色体 2 相同,则计算一个值,若为跨染色体,则分别计算在二个染色体上的值。

phased_block_region 上的具体 Haplotype 信息，则在对应的 hapcut 结果中。