疾病基因组事业部自动化流程文档

研究与合作中心-疾病研究部

版本：v4.7

2019 年 2 月 26 日

目录

一、文档编写说明

介绍疾病事业部自动化流程可实现的功能以及实现的方法，以及其的使用方法，同时列举流程运行中可能遇到的问题以及问题的解决方法。

二、流程简介

疾病事业部自动化流程（Human_reseq_pipeline.py）是一个以人和小鼠全基因组、外显子组、目标区域重测序产品的分析特点为基础，实现基本分析的一键式运行流程，最大化地缩短人工参与的时间。流程可从原始下机数据开始，一键式完成全基因组、外显子组、目标区域重测序的基本分析和部分高级分析内容（小鼠只有基本分析内容）。

主流程路径（Version4.7）：

天津集群：/ifs/TJPROJ3/DISEASE/share/Disease_pipeline/Human_reseq/Version_4.7/main/pipeline.py

南京集群：/NJPROJ2/DISEASE/share/Disease_pipeline/Human_reseq/Version_4.7/main/pipeline.py

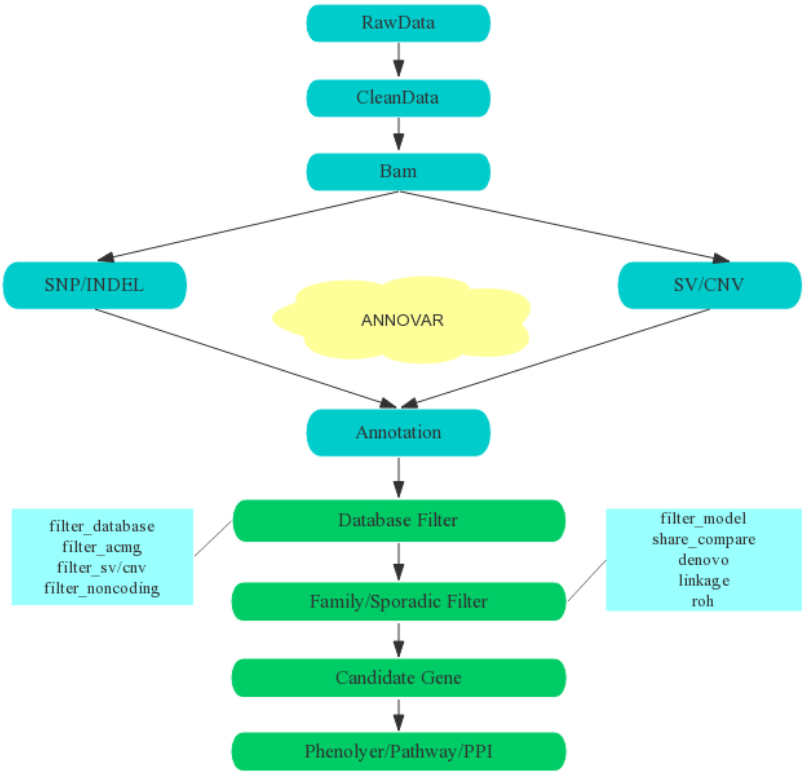


图 1：分析流程图

三、流程详细说明

3.1 流程可以完成的任务：

1. **软链数据**：链接原始下机数据（raw data）到分析路径；

2. **数据质控**：对 raw data 做 QC（质控，Quality Control），质控项目不产生 clean data，基本分析和高级根据项目是否保留 clean data； 质控软件可选择 raw2clean_QC 或 fastp.

3. **数据比对**：将 clean data 跟参考基因组比对，并将同一个样本的 sort.bam 文件合并在一起（多条 lane 时 merge，一条 lane 时直接 move），然后标记重复（一对 reads 的比对位置和方向等完全一致），最终生成*.final.bam 文件用于后续分析；

4. **SNP/INDEL 检测**：可选择使用 Samtools, GATK 或 Sentieon 来检测，并使用 Annovar 对结果注释；默认使用 samtools 检测;

5. **SV 检测**：可选 lumpy, breakdancer 或者 crest 检测；默认使用 lumpy 软件;（注：TR 项目使用 breakmer 检测）

6. **CNV 检测**：可选 freec, cnvnator, conifer，全基因组默认使用 freec 软件；全外显子默认 CoNIFER；(注：小鼠项目只能进行全基因组 CNV 检测)

7. **SNP/INDEL 突变位点过滤**：将所有样本的突变位点 merge(合并)后，默认利用 1000g, Func, ExonicFunc, SIFT, Polyphen, MutationTaster, CADD 进行筛选，其中有害性预测一步对 INDEL 无效。WGS 可以利用突变频率，保守性等进行非编码区筛选；

8. **SNP/INDEL 突变位点有害性分类**：将所有样本突变位点合并（merge）后，进行 7 的突变位点过滤，同时也可以利用家系，频率，基因与疾病相关性，位点保守性，软件预测分值等对位点进行有害性分类，最后取过滤和分类显示有害的位点的并集作为最终的候选位点，并对交集的位点进行标记，可优先考虑交集的位点；

9. **SV/CNV 过滤**：全基因组 SV/CNV，外显子 CNV， 对其进行数据库注释，优先级标注和有害结构变异筛选；

10. **家系显隐性筛选**：在有家系样本存在的情况下，进行显隐性模式筛选（非家系样本也可以根据老师要求筛选纯合或者杂合突变，但要根据实际情况调整）；

11. **家系 de novo 筛选：**患者+患者父母存在情况下可进行 de novo 筛选；SNP/INDEL 默认使用 denovo_samtools 和 denovo_triodenovo 模块；CNV/SV 利用家系样本结构变异检出结果筛选新生结构变异；

12. **家系连锁与单体型分析：**利用 Merlin 检测计算家系的连锁区间，并对连锁区间进一步进行单体型分析；

13. **家系纯合子区域（ROH）分析：**利用 H3M2 进行纯合子区域检测；

14. **Case-Control 筛选：**Case（或者含 control）共有基因筛选（默认 case 10%共有），主要针对散发样本分析；

15. **候选基因排序：**利用 Phenolyzer 分析候选基因与疾病的相关性，并对候选基因排序；

16. **通路富集分析：**利用 R 语言对候选基因进行 GO,KEGG 通路富集并绘图；

17. **蛋白互作分析：**利用 Cytoscape 软件中的 GeneMANIA 模块进行分析；

18. **非编码区分析：**对非编码区位点进行数据库过滤；

19. **关联分析：**在 case 样本检出的 snp 变异位点基础上，以 NovoZhonghua、ExAC 或自定义的 control 群体，可进行位点关联和基因关联分析；

20. **HLA 分型：**HLA 基因分型;

21. 自动化的分析指标检查，并邮件通知未通过检查的样本及信息；

22. 流程整体和分布时间消耗统计；

23. 分析报告生成及邮件通知。

3.2 流程实现：

i. 模块规划

赋予每个分析模块一个编号，简化参数列表，并定义每个分析模块之间的依赖关系，当依赖关系不满足时给出提示信息并退出程序：

表一 分析模块列表

代号	模块	说明	依赖于

1	**quality_control**	**RawData 软链并质控**	**无**
1.1	quality_control_no_clean	RawData 软链并质控，不生成 CleanData, 只适用于质控项目	无
1.2	quality_control_rm_clean	Rawdata 软链并质控，在 bwa 比对成功后删除 CleanData	无
1.3	quality_control_keep_clean(默认)	RawData 软链并质控，保留 CleanData	无
2	**mapping**	**CleanData 比对到参考基因组（默认 b37）**	**1**
2.1	mapping_with_default(默认)	CleanData 比对到参考基因组（默认 b37），使用 bwa 比对，samtools 排序，sambamba 合并和标记重复	1
2.2	mapping_with_sentieon	Clean data 比对到参考基因组（默认 b37），使用 sentieon 比对、排序、合并和标记重复	1
3	**snpindel_call**	**SNP/INDEL 检测并注释**	**1,2**
3.1	snpindel_call_samtools(默认)	使用 Samtools 进行 SNP/INDEL 检测	1,2
3.2	snpindel_call_gatk	使用 GATK 进行 SNP/INDEL 检测	1,2
3.3	snpindel_call_sentieon	使用 Sentieon 进行 SNP/INDEL 检测	1,2
3.4	snpindel_call_mtoolbox	使用 MToolBox 进行 SNP/INDEL 检测（仅适用于线粒体捕获项目）	1,2
4	**sv_call**	**SV 检测并注释 （WGS 项目）**	**1,2**
4.1	sv_call_crest	使用 CREST 进行 SV 检测	1,2
4.2	sv_call_ breakdancer	使用 breakdancer 进行 SV 检测	1,2

4.3	sv_call_breakmer	使用 breakmer 进行 SV 检测（仅适用于 TR 项目）	1.2
4.4	sv_call_lumpy(默认)	使用 Lumpy 进行 SV 检测	
5	**cnv_call**	**CNV 检测并注释(在 WGS 项目中存在 CNV 分析时会在 CNV 分析路径下画 Circos 图)**	**1,2**
5.1	cnv_call_freec(WGS)	使用 freec 进行 CNV 检测(WGS 项目)	1,2
5.2	cnv_call_cnvnator	使用 cnvnator 进行 CNV 检测	1,2
5.3	cnv_call_conifer(WES)	使用 conifer 进行 CNV 检测(WES 项目)	1,2
6	**filter**	**对结果进行过滤**	**1,2,3**
6.1	filter_db	Merge 样本的 SNP/INDEL 进行数据库过滤	1,2,3
6.2	filter_acmg	Merge 样本的 SNP/INDEL 进行数据库过滤和位点的有害性分类，然后找两种方法得到的有害性位点的并集作为最终的有害性结果。	1,2,3
6.3	filter_sv	对 SV 结果进行数据库过滤	1,2,4
6.4	filter_cnv	对 CNV 结果进行数据库过滤	1,2,5
6.5	filter_noncoding	对非编码区位点进行数据库过滤: 1）需要提供疾病名称（必要条件） 2) 如果提供疾病或表型所对应的组织，可以筛选相应组织中活性增强子和活性启动子(可选条件)	1,2,3
7	**filter_model**	**家系筛选**	**1,2,3,6**
7.1	model_dominant	显性模式筛选	1,2,3,6

7.2	model_recessive	隐性模式筛选，包括复合杂合筛选	1,2,3,6
7.3	share_compare	患者共有基因筛选，一般适用于散发样本	1,2,3,6
8	**denovo**	**新生突变检测，一般选 8.1,8.3，即使用 samtools 和 triodenovo 检测新生突变并取交集**	**1,2,3**
8.1	denovo_samtools	使用 Samtools 进行 SNP/INDEL 新生突变检测	1,2
8.2	denovo_denovogear	使用 denovogear 进行 SNP/INDEL 新生突变检测	1,2
8.3	denovo_triodenovo	使用 triodenovo 进行 SNP/INDEL 新生突变检测	1,2,3
8.4	denovo_sv	使用家系内每个样本 SV 检测结果检测 de novo SV	1,2,4
8.5	denovo_cnv	使用家系内每个样本 CNV 检测结果检测 de novo CNV	1,2,5
9	**linkage**	**家系连锁分析**	**1,2**
9.1	merlinkage	利用 Merlin 进行家系连锁分析	1,2
10	**other**	**其他高级分析**	
10.1	roh	利用 H3M2 进行纯合子区间分析	1,2,3
10.2	phenolyzer	利用 Phenolyzer 进行候选基因排序	1,2,3,6
10.3	pathway	利用 R 语言对候选基因进行 GO,KEGG 通路富集并绘图	1,2,3,6
10.4	ppi	利用 Cytoscape 软件中的 GeneMANIA 模块进行分析	1,2,3,6

10.5	site_association	位点关联分析	1,2,3
10.6	gene_association	基因关联分析	1,2,3
10.7	hla	HLA 基因分型	1,2

有了上面两步的铺垫，就可以使用一个参数告诉主流程我们要做什么分析（比如--analy_array 1.2,2.1,3.1,4.1,5.1）。

ii. 规划并建立路径

通过选定分析模块对应的编号进行分析，流程对自动建立对应模块的分析路径，比如使用 QC 模块，则会建立 QC 文件夹，并在其下建立以每个样本名命名的文件夹，存放每个样本 QC 结果。

表二 分析路径下目录说明

目录	说明
RawData	存放每个样本 Rawdata 的软链接，子目录以 sampleID 命名
QC	存放每个样本的质控结果（cleandata），子目录以 sampleID 命名
Mapping	存放每个样本的比对结果（bam&bam.bai），子目录以 patientID.sampleID 命名
Alnstat	存放每个样本 depth&covrage 等信息的统计结果，子目录以 sampleID 命名
Other	存放自动化检查结果和相关处理脚本，数据删除脚本等
Mutation	存放每个样本 SNP/InDel calling 的结果，并含有注释信息，子目录以 sampleID.software 命名，如 xxx.samtools。
SV	存放每个样本的 CNV（如果是外显子则存放这个项目全部外显子 CNV 分析结果），SV 检测结果，并含注释信息，子目录以 sampleID 命名；全基因组 CNV 检测时，在分析路径下存在 Circos 目录，画该样本的 Circos 图
log	以对应 job 名存放所有 jobs 的的标准输出和标准错误输出，目录先按照 job 名区分，默认提交的 job 名是以日期+'job'命名，比如 2015.06.08.job

job	每次执行主流程脚本生成的 job 文件，用于 sjm 投递，投递后还会生成 status 文件和 log 文件，status 文件可用于重投失败的任务
Report	存放各级 report 结果（qc、mapping、primary、advance, Advance_brief），目录下按照 job 名区分
Result	存放项目的数据释放结果（raw data，cleandata，vcf 等），目录下按照 job 名区分。包含释放脚本和打包脚本
Advance	存放基本分析统计结果及高级分析结果，目录下按照 job 名区分。全部可能的子目录： |-- Summary：样本突变检出相关统计数据 |--IntegrateResult：家系 SNP/INDEL 候选突变位点整合结果（例如将隐性筛选与 de novo 筛选的结果整合到一个表格中），以及全部样本候选基因整合结果 |-- Merged_vcf：样本 merge 及过库的结果 | |-- Filter

	| |-- VCF |-- ACMG：样本有害性分类结果 |-- FilterSV：SV 有害性筛选结果 |-- FilterCNV：CNV 有害性筛选结果 |--Noncoding：非编码区位点筛选结果 |-- ModelF：按照遗传模式（显隐性）筛选的结果 |-- Share：共有突变基因筛选和统计结果 |-- Denovo：de novo 分析的结果 | |-- DenovoSam | |-- DenovoGear | |-- DenovoTrio | |-- DenovoRate | |-- DenovoSV | |-- DenovoCNV |-- Linkage：连锁分析的结果 |-- ROH：纯合子区间分析结果 |--Pathway：通路富集结果 |--Network：Phenolyzer 分析排序结果 |--PPI：蛋白互作结果 |--SiteAS：位点关联分析结果 |--GeneAS：基因关联分析结果 |--HLA：HLA 分型结果

iii. 流程参数说明

主流程通过识别输入的每个参数，来确定分析样本，内容，路径等信息。每个参数的解释如下表。

表三 流程参数说明

主要参数	值类型	参数说明[default]
--pwd	[dir]	分析路径[./]
--ref	[string]	使用的参考基因组，可选 b37,hg19,hg38[b37]
--samp_list	[file]	需要分析的样本信息
--samp_info	[file]	样本附加信息文件
--pn	[file]	子项目编号<Tab>项目名称
--seqstrag	[string]	测序方法 WES_ag(WES), WES_illu, WGS
--TR	[file]	目标区域[b37.chr25Region.bed]
--rmdup	[logic]	是否标记 duplicate reads [Y]
--analy_array	[string]	需要进行的分析编号列表[1,2.2,3.2]
--rawdata	[float]	用于自动化指标检查，样本需求的数据量[0]
--depth	[float]	用于自动化指标检查，样本需求的测序深度[0]
--PE	[int]	用于自动化指标检查，样本的测序 PE 长度[150]
--Q20	[float]	用于自动化指标检查，数据需求的最低 Q20 百分值，默认 90，表示 90%[90]
--Q30	[float]	用于自动化指标检查，数据需求的最低 Q20 值，默认 85，表示 85%[85]
--Error	[float]	用于自动化指标检查，数据容忍最高的 error，默认 0.1，表示 0.1%[0.1]

--dup	[float]	用于自动化指标检查，数据容忍最高的 duplication 比例，默认 30，表示 30%[30]
--qcsuffix	[string]	qc_list 的后缀（required）
--startpoint	[string]	指定开始分析的的位置[NULL]
--queues	[string]	指定 qsub 所用的队列
--newjob	[string]	生成的 jos 文件名[year.month.day.job]
--callTR	[logic]	是否只对目标区域进行 snp/indel 检测[N]
--moduledir	[dir]	模块路径
--mail	[string]	报告生成后发送邮件的联系人
--yymail	[string]	释放数据打包信息发送邮件给对应项目的运营
--WES_xten	[logic]	该项目使用 Xten 测外显子数据[N]
--MT	[string]	是否分析线粒体上突变[N]
--disease_type	[string]	研究疾病的类型（复杂还是单基因疾病）[N]
--datastat	[logic]	是否记录样本信息，默认记录；非正常项目、测试项目需要设置为 N
--pdf	[logic]	generate pdf report or not[default=Y]
--confidence	[logic]	mark confidence for integrate results or not[default=N]
--hla-gene	[list]	the gene to do HLA typing for ATHLATES, default will do all genes
--hla-software	[string]	the software for hla typing[default="athlates,hlahd"]
-sps, --show-startpoints		Show the available startpoints
--software	[string]	speify software, like "roh=plink;", "aligment=sentieon;merge=picard"

iv. startpoint 参数可选值及顺序说明

分析过程会遇到各种情况，比如先分析到 QC（Mapping），再进行变异检测，或者由于数据库升级等引起的某步之后的分析需要重新提交，此时为了避免重复分析不需要再次分析的内容，会使用到--startpoint 参数，表示从某个指定分析内容开始往后分析。

查看可用的 startpoint 命令：pipeline4.7 –sps

```
$pipeline4.7 -sps
use configration: /ifs/TJPROJ3/DISEASE/share/Disea
1.0.0 -- qc
1.0.0 -- md5_raw
1.0.1 -- qc_check
1.0.1 -- qc_report
2.0.0 -- bwa_mem
2.0.0 -- sentieon_bwa_mem
2.0.0 -- mapping
2.0.1 -- gzip_md5_clean
2.0.2 -- stat_flag
2.0.2 -- stat_depth
2.0.2 -- sambamba_merge
2.0.3 -- sentieon_markdup
2.0.3 -- sambamba_markdup
2.0.3 -- combine_stat
2.0.3 -- stat_uncover
2.0.4 -- mapping_report
2.0.4 -- mapping_check
2.0.6 -- finalbam
2.0.6 -- final_bam
2.3.0 -- mtoolbox_call
3.0.0 -- mutation
3.0.0 -- sentieon_realign
3.0.0 -- samtools_call
3.0.0 -- gatk_hc_call
3.0.1 -- gatk_concat
3.0.1 -- bcftools_concat
3.0.1 -- sentieon_recal
3.0.2 -- bcftools_filter
```

说明：--后面为可选的 startpoint，前面为优先级；优先级由.分隔的三个数字组成，比较规则如下：

- ✧ 第一个数字不同时，数字大的分析在后；
- ✧ 前两个数字相同时，表示同一分支，第三个数大的分析在后；
- ✧ 第一个数字相同，第二个数字不同时，表示不同分支，没有先后关系；

例如做完了 QC，想从比对开始分析，可以指定 startpoint 为 mapping 或 bwa_mem;

做完了 Mapping 想从 finalbam 开始接着进行后续变异检测则可指定 startpoint 为 finalbam 或 final_bam

v. 任务调度系统说明

分析会涉及到大量的分析脚本，为了有序且自动化的运行这些脚本，流程基于 sjm 任务调度系统，构建了 sjm 能够识别的 job 文件。job 文件中包含每个分析脚本的名字，使用的内存，投递的参数，脚本的路径，脚本之间的先后分析顺序等信息。sjm 能识别这个文件，让各个分析脚本按照给定的顺序和内存依次投上计算节点运行。Job 文件包含两部分 job 命令部分和 order 定义部分：

Job 命令部分：

```
job_begin
  name samtoolsMpileup_chr_11_liujia
  memory 300M
  status waiting
  sched_options -V -cwd -S /bin/bash -q disease.q -q novo.q -q all.q -P joyce -q joyce.q
  cmd_begin
    sh /ifs/TJPROJ3/DISEASE/ProcessPipeline/WES.Pipeline.4.6/Mutation/liujia.samtools/samtoolsMpileup_chr_11_liujia.sh
  cmd_end
job_end
```

Order 部分：

```
order qc_liujia_ZTD17040251_HCT25ALXX_L6 after ln_liujia_ZTD17040251_HCT25ALXX_L6
order md5_liujia_ZTD17040251_HCT25ALXX_L6 after ln_liujia_ZTD17040251_HCT25ALXX_L6
order gzip_fq_liujia_ZTD17040251_HCT25ALXX_L6 after bwa_mem_liujia_ZTD17040251_HCT25ALXX_L6
order QC_check_liujia after qc_liujia_ZTD17040251_HCT25ALXX_L6
order bwa_mem_liujia_ZTD17040251_HCT25ALXX_L6 after QC_check_liujia
order samtools_sort_liujia_ZTD17040251_HCT25ALXX_L6 after bwa_mem_liujia_ZTD17040251_HCT25ALXX_L6
order picard_mergebam_liujia after samtools_sort_liujia_ZTD17040251_HCT25ALXX_L6
order picard_rmdupBam_liujia after picard_mergebam_liujia
order finalbam_liujia after Map_check_liujia
order Map_check_liujia after combine_liujia
order Map_check_liujia after picard_rmdupBam_liujia
order remove_liujia after finalbam_liujia
```

Job 命令部分指出如何 qsub 任务，order 定义部分指出各个任务之间的优先级。这样我们的流程就只需要做 5 件事：

① 建立各个分析模块结果路径；

② 生成 job 命令部分所需脚本；

③ 定义各个分析模块对应的各个分析任务的 job 命令字符块并输出到 job 文件；

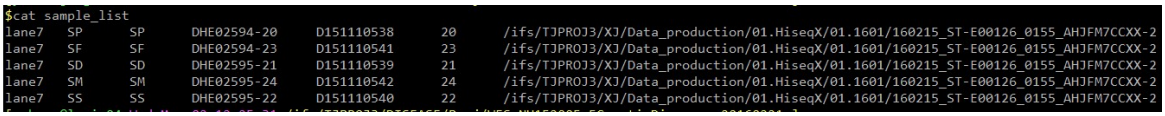
④ 定义各个分析任务的优先级并输出到 job 文件；

⑤ 最后由我们自己运行 sjm，执行 job 文件完成全部任务的投递，执行 job 文件后会生成 job.status 文件记录此次运行的各个 job（shell）的最终状态，failed 则可重跑。

vi. 流程配置文件说明

流程要分析的样本，需要的家系信息，项目信息等均来自与不同的配置文件。只有使用规范准确的配置文件，才能得到正确的结果，因此配置文件的格式和内容都非常重要。下面详细介绍不同的配置文件的格式与内容。

1) **sample_list: <tab> 分隔**



```
$cat sample_list
lane7   SP      SP      DHE02594-20     D151110538      20      /ifs/TJPROJ3/XJ/Data_production/01.HiseqX/01.1601/160215_ST-E00126_0155_AHJFM7CCXX-2
lane7   SF      SF      DHE02594-23     D151110541      23      /ifs/TJPROJ3/XJ/Data_production/01.HiseqX/01.1601/160215_ST-E00126_0155_AHJFM7CCXX-2
lane7   SD      SD      DHE02595-21     D151110539      21      /ifs/TJPROJ3/XJ/Data_production/01.HiseqX/01.1601/160215_ST-E00126_0155_AHJFM7CCXX-2
lane7   SM      SM      DHE02595-24     D151110542      24      /ifs/TJPROJ3/XJ/Data_production/01.HiseqX/01.1601/160215_ST-E00126_0155_AHJFM7CCXX-2
lane7   SS      SS      DHE02595-22     D151110540      22      /ifs/TJPROJ3/XJ/Data_production/01.HiseqX/01.1601/160215_ST-E00126_0155_AHJFM7CCXX-2
```

注：如果第一行是 title,则需要以#号开头，每列的意思如下：

第一列：样本在测序仪原始的 Lane 号(Ori_Lane)

第二列：被采集样本人的个体编号(PatientID)

第三列：被采集样本人的样本名称（同一个人可能有不同样本，SampleID）

第四列：样本文库编号(LibID)

第五列：样本诺禾编号，来自于下机单，一个样本有唯一诺禾编号(Novo ID)

第六列：构建文库所使用的 Index(Index)

第七列：样本分析需要使用的数据存储路径，一般是原始下机数据路径(Path)

2) **sample_info: <tab> 分隔**

若家系信息未知，写成'.'(若需要进行高级分析，家系名不能为'.')，其他信息未知，写成"U"。第一行需要以#FamilyID 号开头，且至少包含 SampleID，SEX，Normal/Patient，PN （忽略大小写），如果有以**'#disease:'** 开头的行，表示这个项目对应的疾病信息，如果有以**'#tissue:'** 开头的行，表示这个项目对应疾病的组织信息，如果有**'#gene:'** 开头的行表示这个项目老师提供的 gene list，其他以'#'号开头的行会被认为是注释而跳过。每列解释如下：

#FamilyID：样本家系信息

SampleID：样本名称

SEX：样本性别信息（M 男，F 女，U:unknown）

Normal/Patient：样本患病情况（N 正常，P 患者，U：unknown）

PN:此合同的项目编号

Data:只有在项目中有家系需要连锁分析时需要填写。对应的样本是否有测序数据，对应数字 0,1,2,3。0 表示样本无测序数据（在连锁分析时，有些样本没有测序但是要用于连锁分析的信息构建中，0 的样本只会出现在需要连锁分析时），1,2,3 均表示样本有测序数据。**多个家系需求不同分析时**，其中家系中样本数据为 1 表示只用于连锁分析，家系中患者数据为 2 表示只用于新生突变，家系中患者数据为 3 表示既用于连锁分析也用于新生突变分析。**如果分析的家系中没有需要连锁分析的家系，这列可以不要。**

Example1:不需要做 de novo 或者连锁分析，不需要 data,pa,ma 列

#Familyid	Sampleid	sex	Normal/Patient	PN
S	SP	F	P	P2015110473
S	SD	F	P	P2015110473
S	SS	F	N	P2015110473
S	SF	M	N	P2015110473
S	SM	F	N	P2015110473

Example2:高级分析含 de novo 分析，但是不含连锁分析，需要 pa,ma 列，但是不需要 data 列

#FamilyID	SampleID	SEX	Normal/Patient	PN	Pa	Ma
F1	AVM0000001	M	P	P2016010286	AVM0000002	AVM0000003
F1	AVM0000002	M	N	P2016010286		
F1	AVM0000003	F	N	P2016010286		
F2	AVM0000010	F	P	P2016010286	AVM0000011	AVM0000012
F2	AVM0000011	M	N	P2016010286		
F2	AVM0000012	F	N	P2016010286		

Example3：即有需要做 de novo 分析，又有需要做连锁分析的家系，且提供了疾病信息，需要‘#disease:’行和 data,pa,ma 列。如下 S 家系，需要做连锁分析，且家系中 SII-10 成员没有测序数据，所以 SII-10 的 data 编号是 0；同时 S 家系的 SP 样本，也需要做 de novo 分析，所以他的 data 编号是 3；其他的样本只用于连锁分析，编号为 1（SF,SM 虽然用于 de novo 分析，但不是患者，只是作为父母，所以编号为 1）。另一个 R 家系，只需要做 de novo 分析，所以患者 R1 的 data 编号是 2。

#Familyid	Sampleid	sex	Normal/Patient	PN	Data	Pa	Ma
#disease:Polydactylism							
S	SP	F	P	P2015110473	3	SF	SM
S	SD	F	P	P2015110473	1	SII-10	SP
S	SS	F	N	P2015110473	1	SF	SM
S	SF	M	N	P2015110473	1	0	0
S	SM	F	N	P2015110473	1	0	0
S	SII-10	M	N	P2015110473	0	0	0
R	R1	M	P	P2015110473	2	R2	R3
R	R2	M	N	P2015110473			
R	R3	F	N	P2015110473			

3) pn.txt：包含子项目编号和项目名称(空格，tab 分割均可)

```
X101SC19010260-Z01      10例DNA样本全外显子测序分析技术服务（委托）合同
```

vii. 模块与配置文件对应关系说明

流程识别的分析样本信息来源于 sample_list，流程在运行时会将 sample_list 中的信息转化写入 qc_list_suffix(suffix 配置文件中的参数)。同一个项目，在不人为改变 qc_list_suffix 的情况下，后续刷脚本或者新分析使用的 sample_list 中的信息会被全部追加到 qc_list_suffix 。

四、项目执行

4.1 准备输入文件

```
※ sample_list
※ sample_info
※ TR 文件（WGS 项目无需指定；WES 项目可指定 V5 或 V6；TS 项目需指定区间文件路径）
※ pn.txt
```

4.2 项目执行流程

1) 创建项目路径，在疾病项目路径下建立新的文件夹(注：创建之前需要先检查路径的剩余的存储空间是否足够支撑项目的运行，如果存储资源不够，则联系相关负责人进行协调，存储足够支撑分析需求时才可投递任务)，文件夹命名方式：测序策略.合同号.疾病名.第一批数据下机时间，4 个字符串以“.”连接，疾病名使用我们给出的疾病中英文对照文件中的英文名，查不到的以汉语拼音代替，未提供疾病名的可以写单位名称，如：WES.NH160126.Polydactylism.20160304

2) 准备 sample_list，sample_info 文件，目前通常是抓取下机信息单中信息来生成。

3) 准备 pn.txt 文件，包含项目编号和项目全称(来自 OMS 系统)

4) 书写执行流程的脚本，之后在命令行直接执行脚本即可

5) sjm job/*.job 以此方式在登陆节点执行第四步生成的 job 文件。

4.3 分析示例

1. WES 基本分析

```
pipeline4.7 \
 --samp_info sample_info_B1 \
 --samp_list sample_list_B1 \
 --qcsuffix B1 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B1.primary.v4.7.20190221.job \
 --analy_array 1,2.1,3.1,5.3
```

2. WES 高级分析

```
pipeline4.7 \
 --samp_info sample_info_B1 \
 --samp_list sample_list_B1 \
 --qcsuffix B1 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B1.advance.v4.7.20190221.job \
 --analy_array 1,2.1,3.1,5.3,6.2,6.4,7.1,8.1,8.3,10.2,10.3,10.4,10.5
```

3. WGS 基本分析

```
pipeline4.7 \
 --samp_info sample_info_B1 \
 --samp_list sample_list_B1 \
 --qcsuffix B1 \
 --seqstrag WGS \
 --newjob B1.primary.v4.7.20190221.job \
 --analy_array 1,2.1,3.1,4.4,5.3
```

4. WGS 高级分析

```
pipeline4.7 \
 --samp_info sample_info_B1 \
 --samp_list sample_list_B1 \
 --qcsuffix B1 \
 --seqstrag WGS \
 --newjob B1.advance.v4.7.20190221.job \
 --analy_array 1,2.1,3.1,4.4,5.3,6.2,6.3,6.4,6.5,7.1,8.1,8.3,10.2,10.3,10.4,10.5
```

4. TS 基本分析

```
pipeline4.7 \
 --samp_info sample_info_B1 \
 --samp_list sample_list_B1 \
 --qcsuffix B1 \
 --seqstrag TS \
 --TR xxx.bed \
 --newjob primary.v4.7.20190221.job \
 --analy_array 1,2.1,3.1
```

5. 样本加测

需要加测的样本，**需要重新构建只含加测数据的 sample_list**,，流程会识别工作目录下的 qc_list_suffix 以判断样本是否为加测，所以加测时，注意保留上批样本的 qc_list_suffix【如果 qc_list_suffix 与 sample_list 中某个样本的 FC_lane 信息不一致，则判定 sample_list 中的为加测】。

生成报告和数据释放时，流程会自动识别 qc_list_suffix 中的样本，和 sample_info 中的样本信息进行报告生成和数据释放，如果遇到分批需要分批释放或者出报告的情况，则需要更改 qc_list_suffix 中的样本信息，再生成报告和生成数据释放结果。

```
pipeline4.7 \
 --samp_info sample_info_B1 \
 --samp_list sample_list_B1_jiace \
 --qcsuffix B1 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B1_jiace.mapping.v4.7.20190221.job \
 --analy_array 1,2.1
```

6. Sentieon 分析

```
pipeline4.7 \
 --samp_info sample_info_B2 \
 --samp_list sample_list_B2 \
 --qcsuffix B2 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B2.primary.v4.7.20190221.job \
 --analy_array 1,2.2,3.4
```

7. 项目分期

对于分期项目，在原路径下，使用新的 sample_list 和 sample_info 文件，刷新脚本，再执行新生成的 job 文件即可。

```
pipeline4.7 \
 --samp_info sample_info_B2 \
 --samp_list sample_list_B2 \
 --qcsuffix B2 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B2.mapping.v4.7.20190221.job \
 --analy_array 1,2.1
```

8. 指定 startpoint

可以通过--startpoint 参数来指定分析开始的节点，生成的 job 中 startpoint 之前的分析会设为 done，查看可用的 startpoint 命令：pipeline4.7 –sps

例如做完 Mapping 没有问题后想接着做后续的基本分析，可指定 startpoint 为 finalbam

```
pipeline4.7 \
 --samp_info sample_info_B2 \
 --samp_list sample_list_B2 \
 --qcsuffix B2 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B2.primary.v4.7.20190221.job \
 --analy_array 1,2.1,3.1,5.3 \
 --startpoint finalbam
```

9. 意外跑断

意外跑断的话，找到最新的 job.status 文件（需要确定之前的 sjm 进程已经结束，可使用 ps xf 看进程是否存在，需注意要在提交 sjm 任务的节点看）。

运行命令：sjm *.job.status

10. 指定软件

可以通过--software 参数来指定一些分析软件，例如使用 fastp 来进行质控

```
pipeline4.7 \
 --samp_info sample_info_B2 \
 --samp_list sample_list_B2 \
 --qcsuffix B2 \
 --seqstrag WES_ag \
 --TR V6 \
 --newjob B2.primary.v4.7.20190221.job \
 --analy_array 1,2.1,3.1,5.3 \
 --software 'qc=fastp'
```

11. 线下项目分析

对于线下项目，如果给的是原始数据，就按照普通项目分析，利用原始数据构建 sample_list 进行分析。如果给的是 bam 文件，则利用已有测序数据构建一个非真实的 sample_list，将 bam 文件在 Mapping 文件夹下对应的样本名的文件夹中，命名为“样本名.nodup.bam“，然后 startpoint 指定 finalbam， 刷脚本和进行投递。此时 Rawdata 和 QC 部分都不是线下项目样本的真实数据，但是分析是从 bam 文件开始，所以对分析没有影响。

4.4 自动化流程

查看帮助：prepare4.7 -h 主要参数说明：

```
-pid, --projectid          项目编号
-n, --fenqi_number         分期编号，如 B1, B1S2
-info, --info-file         sample_info 文件，可以使原始的 excel 信息搜集表，也可以是配置好的 sample_info 文本
-disease, --disease-name   疾病名称，如果没提供但是提供了信息搜集表，则从表中自动提取疾病名(仅限英文)
-seq, --seqstrag           测序策略，默认根据分析路径自动判断
-TR, --target-region       目标区域文件，可以写 V5,V6，也可以指定文件路径
-array, --analy-array      分析编号，默认 1,2.1
-job, --job-name           Job 名称，默认为{FENQI}.{ANALY_ARRAY}.{DATE}.job
-shell, --shell-name       Shell 脚本名称，默认为{PROJECTID}.{FENQI}.{ANALY_ARRAY}.{DATE}.sh
-other, --other-args       其他流程可用参数，如 "--dup 35 --software 'qc=fastp'"
```

使用示例：

```
prepare4.7 -pid X101SC19010260-Z01 -n B1S2 -info sample_info_B1 -array 1.2,2.1,3.1,5.3
```

运行结果：

```
$prepare4.7 -pid X101SC19010260-Z01 -n B1S2 -info sample_info_B1 -array 1.2,2.1,3.1,5.3
>>> preparing sample_info ...
use text sample_info: sample_info_B1
>>> preparing sample_list ...
> checking: X101SC19010260-Z01 - BKDN190004673-1A
> checking: X101SC19010260-Z01 - BKDN190004672-1A
write sample_list: sample_list_B1S2
write shell: X101SC19010260-Z01.B1S2.1.2,2.1,3.1,5.3.20190221.sh
2个样均下机，开始执行项目...
sh X101SC19010260-Z01.B1S2.1.2,2.1,3.1,5.3.20190221.sh && sjm2 job/B1S2.1.2_2.1_3.1_5.3.20190221.job
use configration: /NJPROJ2/DISEASE/share/Disease_pipeline/Human_reseq/Version_4.7/config/config_nanjing.ini
hello, suqingdong
qc status: waiting
mapping status: waiting
check queues...
  used queues: ['disease.q', 'disease1.q', 'sentieon1.q']
check analy_array...
Check files...
check target region...
  [warn] no TR was supplied for reference b37
  default V6 TR was used: /NJPROJ2/DISEASE/share/Disease/Agilent/SureSelectXT.Human.All.Exon.V6/S07604514_Regions_extract.bed
analysis items:
 1.2  quality_control_rm_clean
 2.1  mapping_with_default
 3.1  snpindel_call_samtools
 5.3  cnv_call_conifer
extract sample informations...
update qc_list ...
  updated qc_list ...
  report number: B1
  samples (2): ['S2377_001', 'S2377_003']
set analysis memory...
>>> pipeline start...
  mutation_soft:samtools, sv_soft:, cnv_soft:conifer, denovo_soft:[]
> QC
>  qc start...
>  clean data will be removed after mapping_check
> Mapping
>  mapping start...
['S2377_001_BKDN190004672-1A_HHCNJDSXX_L3.sort.bam']
```

自动化项目状态查看：projstatus

```
$projstatus -h
usage: projstatus [-h] [-u USER] [-k {waiting,stop,delete}] [id]

positional arguments:
  id                    the id to search

optional arguments:
  -h, --help            show this help message and exit
  -u USER, --user USER  the username to search[default=suqingdong]
  -k {waiting,stop,delete}, --operation {waiting,stop,delete}
                        the operation to do with sepcific id, choose from
                        waiting, stop, delete
```

```
$projstatus
_id      status  projectid       projpath        shell   job
------------------------------------------------------------
5c6e674e8ea1660382927bac        running X101SC19010260-Z01      /NJPROJ2/DISEASE/Proj/PipelineTest/suqingdong/temp/WES.H101SC19010260.101iDNAyangben.20190202    X101SC19010260-Z01.B1S2
.1.2,2.1,3.1,5.3.20190221.sh    B1S2.1.2_2.1_3.1_5.3.20190221.job
```

修改状态或删除：projstatus -k stop/delete <id>

```
$projstatus
_id      status  projectid       projpath        shell   job
------------------------------------------------------------
5c6e674e8ea1660382927bac        running X101SC19010260-Z01      /NJPROJ2/
.1.2,2.1,3.1,5.3.20190221.sh    B1S2.1.2_2.1_3.1_5.3.20190221.job
[suqingdong@njlogin04 17:11:02 /NJPROJ2/DISEASE/Proj/PipelineTest/suqing
$
[suqingdong@njlogin04 17:11:02 /NJPROJ2/DISEASE/Proj/PipelineTest/suqing
$projstatus -k delete 5c6e674e8ea1660382927bac
1 objects was deleted...
```

五、注意事项

1) 注意 sample_list, qc_list 与分析模块之间的关系，sample_info 中的性别信息用于变异结果的按性别过滤，**如果性别信息不明则 X, Y 染色体结果都保留。**

2) 测序策略参数--seqstrag 支持{WES_ag，WES_illu，WGS, TS }，注意拼写。

3) 由于外显子捕获区域外也能检测到很多 SNP/INDEL，所以在 WES 时，最好使用--callTR N 参数（默认使用）。

4) 保留项目中产生的每个执行 shell 脚本，有分期，加测，模块更改等情况时新建 shell 脚本，**不要在原来的脚本上做改动**。

5) 项目因为 sample_list 书写错误报错时，重新执行必须删除生成的错误的 qc_list_suffix，否则流程会识别错误的 qc_list_suffix 中的信息进行 merge 或者生成报告，释放数据等。

6) 家系的显隐性模式筛选识别 sample_info 中的家系信息，sample_info 中有的样本都会做此部分分析，所以如果筛选模式和分析家系情况较多，可能需要更改脚本输入的 sample_info 进行投递（如果忘记修改，导致 sample_info 中的家系 or 样本全部做了此分析，数据释放时，需要把不需要的从 Advance 文件夹下的 ModelF 文件夹中删掉，然后重新投递数据释放的脚本）

7) 对于 WGS 的 SV 检测，默认使用 lumpy（分析编号 4.4） 代替 CREST

8) 主流程下 Other 文件夹中的 extract_info.sh 用于项目信息统计，非常重要，要保证执行成功

9) 主流程 Other 文件夹中的 Deep_remove.sh 脚本**谨慎执行**，此脚本会删除分析产生的各种中间文件，一般是数据释放完成至少 3 个月以后才能执行！

10) **非编码区分析，需提供疾病名称，**没有疾病名称不能进行分析，如果还提供疾病或表型所对应的组织，可以筛选相应组织中活性增强子和活性启动子。

六、出错处理

✧ 执行生成 job 文件的 shell 时报错，注意检查 sample_list 中的内容和格式是否正确，检查 qc_list_suffix 是否包含错误信息。

✧ 若流程运行完毕，而结果未产生，则先查看 xx.job.status.log 中哪些任务 failed，再查看 log 中这个任务的报错信息，寻找原因。当脚本修改完毕后，可用 startpoint 参数定点重新提交。

七、环境要求

- ✧ 64 位中央处理器，Linux 操作系统（内核版本号不低于 2.6），GNU Compiler Collection（版本不低于 3.4）。
- ✧ 针对全基因组和外显子组重测序，内存至少需要 12G。

八、更新

1. QC 部分增加了质控软件 fastp

   对于没有 adapter 的数据可使用 fastp 进行质控（参见分析示例：10. 指定软件）

2. Mapping 部分增加了 Sentieon 方法，分析时可指定为分析代码 2.2,3.4

   Mapping 结果均为 nodup.bam 作为 final.bam，Sentieon 和 GATK 的 recall 或 realign 均在 Mutaion 目录下进行

3. Mutation 部分为每个样本 call gvcf，然后 merge 和注释，再拆分每个样本的 vcf 和注释结果；

   变异检测软件可选 Samtools，Sentieon 或 GATK4

   WES 项目和 TR 项目以后默认 Call Bed 区间内全部位点，这样 merged 后能够区分"."和"0/0"

4. 内部 数据库更新（WES_2827， WGS_542）
5. HGMD 数据库更新（2018.4）
6. WGS 项目 SV 检测默认替换为 Lumpy（分析代号 4.4）
7. 优化自动化检查，包括 QC，Maping 和 Primary 的检查，结果会自动记录到数据库；
8. 新生突变检测去掉了 DenovoF 模块，新增了 triodenovo 检测方法，默认和 Samtools 方法取交集（分析代码 8.1,8.3）
9. 关联分析调整为 10.5 位点关联，10.6 基因关联
10. 增加 HLA 分析（10.7）
11. 增加 PDF 报告的生成
12. 增加材料与方法模块，根据分析内容生成 word 版的英文材料与方法

13. 增加文章图表模块，根据分析内容生成 excel 版的文章图表

14. 整合脚本优化

15. 增加了几种特定疾病的病种分析和报告

九、联系

如有问题请联系：suqingdong@novogene.com