基于扩增片段靶向测序变异分析

项目文件夹：

/opt/data/db/sra/jx\_161118\_changhai\_19samples

/opt/data/db/sra/jx\_140219/SIPPR\_exon

/opt/data/db/sra/jx\_170918\_blood\_germline\_2smp

/opt/data/db/sra/jx\_240518\_mix2

/opt/data/db/sra/jx\_291218\_changhai/changhai

/opt/data/db/sra/jx\_amde\_pharmVar/adme

/opt/data/db/sra/jx\_amde\_pharmVar/pharmVar

靶向测序变异分析先后经历很多相似项目，但是都有所不同，从实验方法到分析方法，所以代码一直在调整。整体分析流程如下：

1. 用fastp工具对原始数据进行预处理，此处要注意两头是否有两碱基GC，若有则加上--trim\_front1 2 --trim\_front2 2两个参数。此工具会自动检测adapter序列，亦可提供adapter序列。另外一个参数是最小长度过滤，主要是将非特异扩增及二聚体的数据进行过滤，如果需要统计这部分情况，则最小长度过滤应设定为15。
2. 在清理原始数据后，可以用cutPrimer工具（<https://github.com/ray1919/cutPrimers>）对引物进行统计和切除。引物切除主要目的是将测序结果中的引物序列清除，避免与真实序列同时进行variant calling。
3. 将上一步的结果与基因组进行比对，使用基因组比对工具bwa（<http://bio-bwa.sourceforge.net/>）这里用的都是默认参数。但是值得注意的是，由于部分基因在基因组scaffold序列上有相似基因，所以比对仅用染色体1-22，X，Y序列进行比对。
4. 用bamstats05（<https://lindenb.github.io/jvarkit/>）这个工具对指定区域的coverage进行统计。这主要是出于质控要求，明确得到各个靶向区域的覆盖情况，以免出现没有覆盖或低覆盖的情况出现而不知道。
5. 对比对的结果用Pisces（<https://github.com/Illumina/Pisces>）工具做variant calling。这个工具是Illumina原测序仪上自带的分析工具进行了开源开发和公布，主要是针对amplicon data做germline和somatic的变异分析。虽然我们多数用来做Germline的分析，但是Germline的默认设定variant frequency只能精确到1%，所以统一用somatic的分析模式进行分析，得到精确度0.1%的variant结果，然后再用R代码进行基因型的矫正。这里主要分为靶向位点SNP分型和基因外显子突变分析。一种是指定位置甚至基因型，一种是在某一区域内找出variant。这个变化主要体现在最后的R代码中的一些变化。
6. 用R代码1统计fastp预处理的结果，整理成Excel表格
7. 用R代码2处理最后variant calling得到的额vcf文件，得到变异分析的Excel结果。

步骤a-e用到代码：target\_region\_variant\_calling.sh，参数较多，但是在代码中对每个参数都做出了说明。

步骤f用到代码：fastp\_report.R

步骤g用到代码：pisces\_xlsx.R，pisces\_exon+rs.R，pisces\_vcf.R区别在于从excel、文件文件、vcf文件中读取目标位点信息。