**{****PGD单基因病检测}**

**信息分析流程操作手册**

# 目的

根据父、母、先症者和若干个胚胎组成的家系样本，基因捕获测序的方法进行PGD-单基因病的信息分析检测流程。首先对每个样本进行QC、比对、去重、call SNP&Indel，再根据父、母、先症者的遗传关系分析出父母的正常和治病单体型，进而确定胚胎是否遗传治病单体型，从而确定胚胎是否正常。

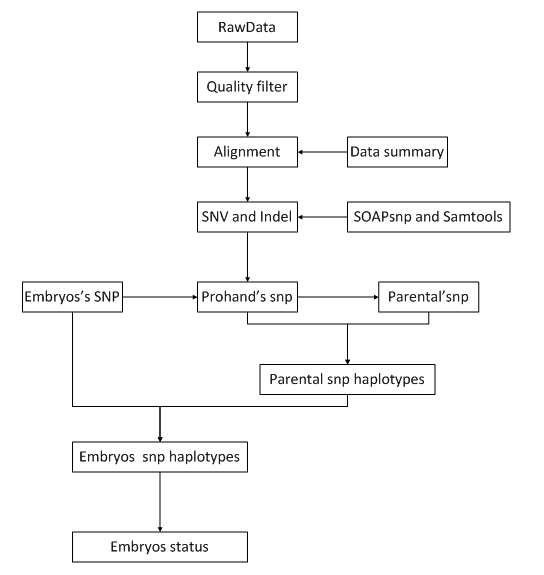
# 适用范围

该流程能够对样本测序数据进行QC、比对、去重、call SNP&Indel，以及目标基因单体型分析、胚胎单体型确定。适用于PGD检测产品升级子项目-基因病检测项目所产测序数据的分析，且样本类型（包括家系类型）、数据情况符合项目要求。

# 运行环境

操作系统：Linux，CPU：10，内存：100G

# 信息分析流程图



# 流程详细步骤

* 1. 步骤一 : 原始序列过滤

5.1.1 操作原理

过滤低质量、带接头、N%>5%的序列。

5.1.2 输入文件

原始fastq序列文件，adapter文件

5.1.3 运行程序

程序名：Filter\_raw\_data.pl，运行时间：~10min

-Q\_system：为1时，质量系统为64；为2时，质量系统为32；

5.1.4输出结果分析

Clean\_fastq文件

* 1. 步骤二：BWA比对

5.2.1 操作原理

参考BWA软件介绍及使用说明[1]。

5.2.2输入文件

Clean\_fastq文件，hg19参考序列

5.2.3 运行程序

软件：BWA，参数：-I -i 15 -L -k 2 -l 31 -t 4，运行时间：~4hours

5.2.4 输出结果

Sam文件

* 1. 步骤三：Samtools处理

5.3.1 操作原理[2]

SAM文件转为BAM文件，去duplication，取uniq比对reads，排序，加index；

5.3.2输入文件

比对完的SAM文件，hg19参考序列

5.3.3 运行程序

软件：Samtools，参数：无，运行时间：~0.5hours

5.3.4 输出结果

BAM文件

* 1. 步骤四：call SNP

5.4.1 操作原理

根据处理完的BAM文件进行SNP、indel calling

5.4.2输入文件

过滤、去重后的BAM文件，hg19参考序列

5.4.3 运行程序

软件：GATK，运行时间：~10hours

5.4.4 输出结果

VCF文件

* 1. 步骤五：Haplotyping

5.4.1 操作原理

根据VCF文件进行家系单体型分析，确定治病单体型，进而判断胚胎是否正常。

5.4.2输入文件

VCF文件

5.4.3 运行程序

程序：Haplotype\_based\_merged\_GATK\_V2.pl，运行时间：~10min

5.4.4 输出结果

Hap文件：

文件格式：Gene、Chr、Pos、Hap[F1，F2，M1，M2]，Embryo-Support-Hap [E1~En]，Genotype[gDNAs，Embryos]，Depth[gDNAs，Embryos]

Score文件：基因上中下游informative SNP统计

# 参数说明及注意事项

文件名说明：

1、Fastq文件：原始下机序列文件；

2、Adapter文件：原始下机序列中包含adapter接头的序列；

3、Clean\_fastq文件：经QC过滤后的fastq文件；

4、Reference文件：hg19参考序列文件；

5、SAM文件：Clean\_fastq 文件经BWA比对后产生的文件，包含比对结果；

6、BAM文件：SAM文件经Samtools处理后产生的文件，包含比对结果；

7、VCF文件：GATK 通过对BAM文件call SNP&Indel后产生的文件，包含突变结果；

8、Hap文件：VCF文件经单体型分析后产生的单体型文件；

9、score文件：单体型分析后每个基因的informative SNP数统计；

10.report文件：项目检查报告文件，包括测序数据、捕获覆盖度、深度、目标基因单体型、治病位点等信息。

# 软件列表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | | 软件名 | 软件版本 | 大型机路径（或软件本库路径） |
| (1) | Main | PGD-MD-automatic-pipeline-  V2.pl | V2 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/New\_pipeline/ |
| (2) | QC | Filter\_raw\_data.pl | V1 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/bin/New\_Pipeline/ |
| Aln | Bwa | V0.5.9-r16 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/soft/bwa/bw |
| Stat | Sam\_sta.V1.03.pl | V1.03 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/bin/New\_Pipeline/ |
| Rmdup | Samtools | V0.1.17 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/soft/ |
| Calling | GATK | v2.3-4 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/soft |
| Anno | convert2annovar.pl  summarize\_annovar.pl |  | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/soft/ |
| Hap | Haplotype\_based\_merged\_GATK\_V3.pl | V2 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/bin/New\_Pipeline |
| Report | PGD\_MD\_summary\_report.pl | V1 | /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/bin/New\_Pipeline/ |

例子操作目录：/hwfssz1/ST\_MCHRI/REHEAL/USER/heyuxuan/Project/PGD/

主程序使用范例：

perl /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/New\_pipeline/Back.Ceshi\_seq500\_pgd.pl -Lib /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/04.work/26\_Run20170726\_GroupTest/Lib -Target /hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/00.rawdata/ifs1/bin/PGD/bed/PGD-chipA -queue P17Z18000N0155 -Gene HBA -Auto Y -User chendayang

几个重要的参数：

Lib: sample Library 样本名与文库名对应，样本名命名需严格遵循家系成员关系。

Target：芯片捕获目标区域，包含Target\_region.bed、target.gene等文件，文件名与格式均固定。

Gene：治病基因名，可不给定，不给定的话不会做最后report一步。

Quasym：测序质量体系，1为64，2为33。

Rmdup：是否去重。

Memory：投递任务申请内存大小。

OutDir：输出路径，如不给定，即以Lib所在文件夹为输出目录。

Auto：是否自动投递任务。

Queue：设置投递任务队列。

User: 流程完成后的邮件通知地址

**8．如何操作？**

此次测试为一个含有父源样本的家系，共4个样本，为节约时间，每个样本只有10M PE reads.

最终结果展示目录：/hwfssz1/ST\_MCHRI/REHEAL/USER/heyuxuan/Project/PGD/

跑完可在上面的目录中核对相关文件。

操作步骤：

第一步： 拷贝主程序：

/hwfssz1/ST\_MCHRI/REHEAL/USER/chendayang/04.work/26\_Run20170726\_GroupTest/work.sh

第二步： 修改work.sh中的项目编号-queue F15HQSQS1987为自己的项目编号，-User 参数为自己的邮箱前缀。即可完成文件配置。

第三步：sh work.sh 运行时间~1min

运行完成后会在当前目录下生成目录：anno chr Embryo1 Father Mother Paternal sh

第四步：qstat查看任务运行状态，投递任务完成，待所有任务跑完后会有邮件通知。运行时间~3 hours (10M)

最终结果有文件 /hwfssz1/ST\_MCHRI/REHEAL/USER/heyuxuan/Project/PGD/PGD-MD-Project-case-report.xls

且一致，即表示流程操作成功。

# 9．常见问题及建议解决方案

1、当胚胎目标检测基因单体型支持SNP数过少。首先加入indel信息，并对SNP准确性进行人工确认，重新进行单体型SNP/indel统计，确定胚胎单体型；

2、当加入indel分析后目标基因单体型SNP/indel支持数依旧不足。对数据进行人工分析，找出原因，根据不同原因，确定解决方案。（一般有三方面原因：目标基因ADO非常严重，一条单体型丢失较多；目标基因测序数据太低，无法获得准确SNP信息；目标基因某单体型原始支持SNP数较少。）

3、当确定胚胎遗传父方或母方单体型时，父方或母方的两个单体型均有SNP支持。如两单体型SNP支持数具有明显差距，可忽略SNP数少的单体型；如两单体型SNP支持数差距不明显，需人工对各单体型进行确认，如对错误SNP进行校正，根据SNP位置信息判断是否有基因重组或者三体现象等。

# 参考文献

[1]. Li H, Durbin R. Fast and accurate short read alignment with Burrows-wheeler transform. Bioinformatics. 2010; 26(5): 1754–1760.

[2]. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078-2079.