肿瘤事业部 Control 集合建立及更新方案

撰写人		日　期	
/Draft：	刘博	/Date：	2020.09
审核人		日　期	
/Review：		/Date：	
批准人		日　期	
/Approve：		/Date：	
文件密级/	□普通/Unclassified　□秘密/Secret		
Classification：	■机密/Highly Secret　□绝密/Top Secret		

目录

1 目的（Objectives）

经过生信部门的内部讨论，确定了肿瘤事业部 A0 版本的 Control 集合构建标准，制订本方案是为保证明确记录肿瘤事业部内部信息分析流程的 Control 集合构建流程，并在事业部内部形成共识，规范化事业部内部各个产品的 Control 集合的构建标准，并建立统一的规范，形成标准化管理。同时也为后期可能的数据追溯，Control 集合数据集更新建立参考文档，指导后期相关方案的落实。

2 适用范围（Scopes）

本方案适用于肿瘤事业部内部所有需要进行 Control 集合构建的生信分析流程，如无特殊说明，肿瘤事业部内部所有 Control 集合均基于本方案进行构建，后期如果在使用过程中发现了本方案的缺陷或不足，则应该对本方案方法进行优化及迭代说明，并对所有基于本方案生成的 Control 集合于指定时间内进行同步更新。

3 职责（Responsibilities）

3.1 生信部负责本方案的整体撰写及第一版脚本的梳理准备工作。

3.2 生信部负责人负责本方案的审核工作。

3.3 生信部各个产品生信分析流程负责人员负责本方案的执行，并生成相应的 Control 集合应用于各个产品对应的生产流程。生信所有相关人员负责使用及了解本方案的使用方法及原理，如果因发现缺陷不足需要对本方案进行升级，则应通知生信部门全员（bgi_tumor@genomics.cn）知晓相关调整，从而确保各个产品的 Control 集合可以得到及时更新并保持构建逻辑的一致性。

4 术语和定义（Terms and definitions）

Control 集合：基于华大基因（简称“华大”，下同）接收的历史临床样本的白细胞数据构建临床收样患者在人群层面表现出来的多态性位点（人群高发，大概率不致病）和检测体系的假阳性位点（普发低频变异，体系偏好性导致），用于后期对患者变异结果进行过滤从而剔除一些和患者的肿瘤发生发展无关的变异集合。

批注 [刘博(Bo1]: 和 ErrorBaseLine 似乎是一样的功能，是否有必要保留。

ErrorBaseLine：基于历史检出样本，构建位点假阳性错误波动范围文件，各个产品由于上游在实验、测序平台的差异，因此会存在不同的错误偏好性。不同的产品应该独立的进行本集合的构建。

Vcf：变异检测结果文件格式，用于记录样本的变异检出结果。

热点变异： 基于解读收录整理的一些已知致病变异，这类变异由于已知和肿瘤的发生发展存在相关性，因此需要从 Control 集合中进行过滤，从而避免造成临床解读层面的假阴性。

5 数据来源和数量（Data Sources）

为了更好的反馈检测系统的偏差因此构建 Control 集合所用的**变异结果（原始 VCF 文件）应和产品信息分析流程保持一致**，从而避免分析流程中存在的差异对 Control 集合带来偏好性的影响。

构建 Control 集合时，应该优先使用真实的临床样本，同时避免癌种的过渡集中（癌种的过渡集中会导致该癌种中的热点变异被富集，被错误的识别为人群多态性位点，尤其是一些暂未被广泛研究的隐藏热点变异被过滤）。

样本量应该尽可能的多，从而更准确的反应人群的分布情况，减少偶然性带来的影响。建议样本量在 1000+以上。如果转产初期样本量少，建议在样本量分别达到 100 例、500 例、1000 例时，对 Control 集分别进行定期更新。同时由于样本量较少的情况下，为确保变异不会因为偶然性出现非多态性位点的富集，因此在样本量较少（500 例以下时），建议人群频率使用 10%进行过滤。

6 集合构建（Data Analysis）

6.1. 构建流程获取

为了便于整个生信分析流程的管理，所有 Control 集合构建流程均保存在华大 Gitlab，生信工具集仓库中（bioinfotoolkit：https://gitlab.genomics.cn/liubo4/bioinfotoolkit/-/tree/master/），如有权限问题，联系（liubo4@genomics.cn）

首先下载 git 仓库 git clone ssh://git@gitlab.genomics.cn:2200/liubo4/bioinfotoolkit.git

下载完成后目录 bioinfotoolkit/02.toolkit/04.Create_ControlSite 中对应的即为 Control 集合构建的流程。使用方式参考下文及示例 shell：02.toolkit/04.Create_ControlSite/work.demo.sh

下载完成后目录 02.toolkit/04.Create_ErrorBaseLine 中对应的即为 ErrorBaseLine 集合构建流程，使用方式参考下文及示例 shell：02.toolkit/04.Create_ErrorBaseLine/demo.sh

6.2. Control 集合构建 – 基于血细胞检测结果

Control 集合构建参考如下命令：

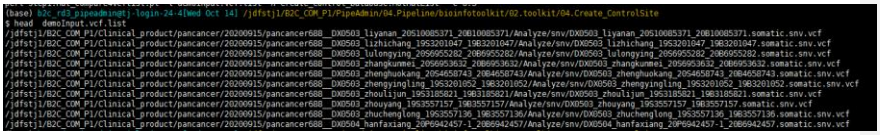
perl Create_Control_Database.pl -l demoInput.vcf.list -h Create_Control_Database.HotMutList -c 0.05

其中：

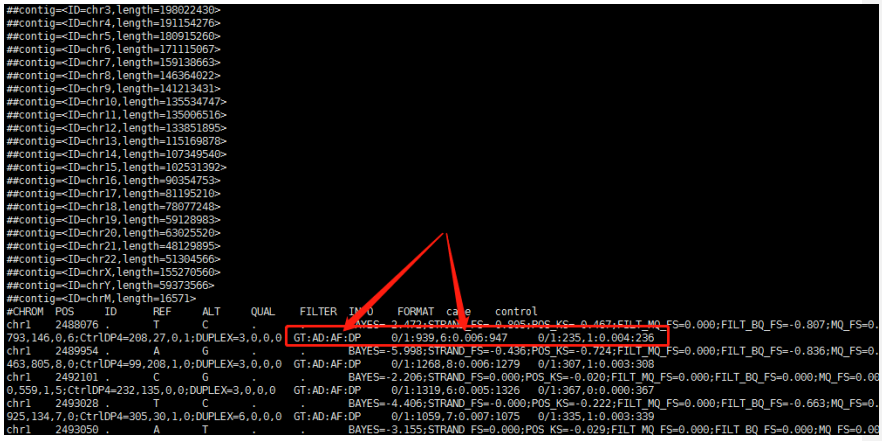
Create_Control_Database.pl 为 Control 集合构建主流程

demoInput.vcf.list 为所有用于构建 Control 集合的 vcf 文件 list，**注意由于 vcf 格式不一致，需要确定 vcf 文件中 AF 列是否存在，若不存在或 AF 所在列不一致，需要对流程进行调整）**

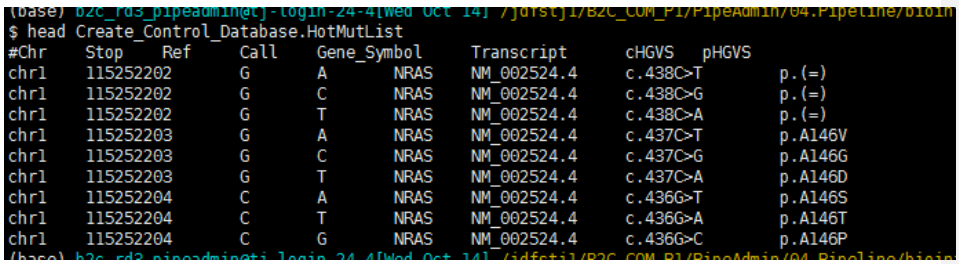
输入的 list 文件示例如下：



本流程构建所用 vcf 文件格式示意如下：



Create_Control_Database.HotMutList 为解读提供的热点变异文件列表，如无更新可参考流程自带的 list 文件，文件格式如下：



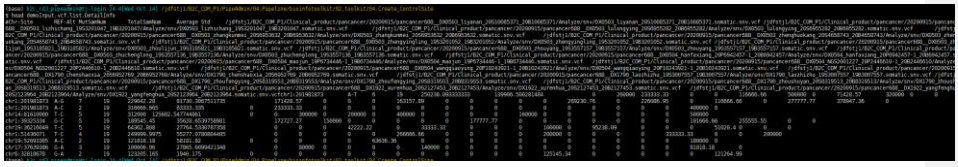
-c 0.05 为指定人群频率过滤阈值，人群发生频率低于该阈值的变异则会在 Control 集合中进行剔除。目前默认人群频率阈值为 5% 以上的会在临床检测中进行过滤。

集合构建结果

Control 集合构建结束后会生成如下两个文件：

```
$ ^C
(base) b2c_rd3_pipeadmin@tj-login-24-4[Wed Oct 14] /jdfstj1/B2C_COM_P1/PipeAdmin/04.Pipeline/bioinfotoolkit/02.toolkit/04.Create_ControlSite
$ ll
total 6806
-rw-r--r-- 1 b2c_rd3_pipeadmin b2c_rd3   11393 Oct 14 00:14 Create_Control_Database.HotMutList
-rwxr-xr-x 1 b2c_rd3_pipeadmin b2c_rd3    7610 Oct 14 00:14 Create_Control_Database.pl
-rw-r--r-- 1 b2c_rd3_pipeadmin b2c_rd3    3408 Oct 14 00:14 demoInput.vcf.list
-rw-r--r-- 1 b2c_rd3_pipeadmin b2c_rd3 4292344 Oct 14 00:14 demoInput.vcf.list.DetailInfo
-rw-r--r-- 1 b2c_rd3_pipeadmin b2c_rd3 2652680 Oct 14 00:14 demoInput.vcf.list.PopCutoff_0.05.countInfo
-rw-r--r-- 1 b2c_rd3_pipeadmin b2c_rd3      102 Oct 14 00:15 work.demo.sh
(base) b2c_rd3_pipeadmin@tj-login-24-4[Wed Oct 14] /jdfstj1/B2C_COM_P1/PipeAdmin/04.Pipeline/bioinfotoolkit/02.toolkit/04.Create_ControlSite
```

其中后缀 * DetailInfo 的文件记录了这些变异在各个样本中的变异检出频率等相关细节，用于可能需要的结果核查，数据回溯等。示例如下：



其中后缀 *PopCutoff_0.05.countInfo 的文件，为最终生成的 Control 集合文件，结果示例如下：

```
(base) b2c_rd3_pipeadmin@tj-login-24-4[Wed Oct 14] /jdfstj1/B2C_COM_P1/PipeAdmin/04.Pipeline/bioinfotoolkit/02.toolkit/04.Create_ControlSite
$ head demoInput.vcf.list.PopCutoff_0.05.countInfo
#Chr    Site    REF     Alt     >25%SamPopFreq  5%-25%SamPopFreq        <5%SamPopFreq   LocalPopFreq    >25%SamNum      5%-25%SamNum    <5%SamNum       altSampleNums   totalSampleNum
chr1    201981873       A       T       0       0       0.315789473684211       0.315789473684211       0       0       6       6       19
chr1    201981873       A       G       0       0       0.368421052631579       0.368421052631579       0       0       7       7       19
chr1    201981873       A       C       0       0       0.105263157894737       0.105263157894737       0       0       2       2       19
chr14   81610000        T       C       0       0       0.263157894736842       0.263157894736842       0       0       5       5       19
chr1    39325334        G       C       0       0       0.263157894736842       0.263157894736842       0       0       5       5       19
chr19   36216049        T       C       0       0       0.263157894736842       0.263157894736842       0       0       5       5       19
chr1    51436071        T       C       0       0       0.210526315789474       0.210526315789474       0       0       4       4       19
chr19   52693305        A       C       0       0       0.105263157894737       0.105263157894737       0       0       2       2       19
chr17   37639386        G       A       0       0       0.157894736842105       0.157894736842105       0       0       3       3       19
```

其中前 4 列为变异信息分别为染色体、位置、Ref、Alt；

其中 5-8 列为不同变异频率阈值下的人群频率，（>25%SamPopFreq，表示该变异检出频率大于 25%的患者在所有患者中的占比，LocalPopFreq 为不进行频率过滤的人群占比，）

其中 9-12 列为不同变异频率阈值下的变异患者数目。

其中第 13 列为样本总数。

注：流程中的人群频率过滤仅针对“>25%SamPopFreq” 和“<5%SamPopFreq”两个变异频率筛选后的人群频率进行过滤。

其中“>25%SamPopFreq”大于 5%，则认为该位点为人群多态性位点；

其中“<5%SamPopFreq”大于 5%，则认为该位点存在系统性偏好错误。

后续需要基于流程使用位置确定是否需要**对变异信息进行左对齐或 3’对齐。**

6.3. ErrorBaseLine 集合构建 – 基于 case 样本

构建数据集合原理：

a) 使用天津本地分析流程对所有临床样本进行变异检测分析。通过分析获得变异检出结果文件原始（vcf），该文件获取方式，应和临床实际分析模式保持一致，避免分析过程等其他差异导致错误集合和临床真实错误错误集合出现偏差。

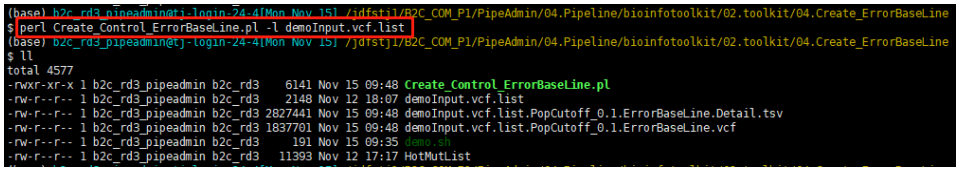
b) 对文件进行整理和汇总，获得组织样本人群检出频率大于 10%的所有变异位点及相关检出频率信息；

c) 针对保留下来的所有变异，统计检出的突变频率，并基于所有检出频率结果计算出位点的检测阈值，cutoff=均值+3*标准差。

批注 [刘博(Bo2]: 目前未检出的样本未纳入方差及均值的计算；
是否有必要考虑

ErrorBaseLine 集合构建方法：
使用： 仓库脚本
bioinfotoolkit/02.toolkit/04.Create_ErrorBaseLine/Create_Control_ErrorBaseLine.pl

整理历史临床样本，对应组织结构检测得到的 vcf 文件对应的文件路径 list 作为 ErrorBaseLine 集合构建脚本的输入文件。流程运行后会生成对应的 ErrorBaseLine.vcf，为原始的过滤集合，后续需要基于流程使用位置确定是否需要**对变异信息进行左对齐或 3’对齐**。



6.4. 集合使用方法

完成*PopCutoff_0.05.countInfo 文件的获取后，需要在原有的生信流程中添加 Control 集合过滤步骤，将过滤后人群频率大于 5%的变异位点进行剔除，仅保留剩余的变异位点用于后期的遗传解读。基于染色体、位置、Ref、Alt 作为变异信息的唯一识别码进行识别并过滤。

完成* ErrorBaseLine.vcf 文件的获取后，需要在原有的生信流程中添加 ErrorBaseLine 过滤步骤，将样本中检出同时频率低于 ErrorBaseLine 确定频率阈值的相关变异进行剔除，仅保

留不位于 ErrorBaseLine 集合，或检出频率大于 ErrorBaseLine 集合频率阈值的变异用于后期的遗传解读。

7 相关记录（Related Records）

无

8 参考文献资料（References）

无

9 附录（Appendix）

附表 1 主要标准作业指导书一览表

附表 1 主要标准作业指导书一览表

操作	文件名称	文件编号	版本号

终止符