SHINE方法[1]的数据集确认：

1. 训练集的确认

1.1、下载SHINE的训练集的时候遇到一个问题:github提供的数据数量大于论文中使用的数据量。（下载地址https://github.com/xf-omics/SHINE）

我联系了文章作者，作者手上没有原始数据，但是给出了建议：提供的数据比论文中的数据多可能是因为①提供的样本没有去重，②应该使用VEP注释，并选择the canonical transcripts 为inframe indels的样本。

根据作者的回复，我的处理结果为①没有重复样本，②使用VEP官网选择canonical转录本，并且根据canonical转录本上的结果筛选出注释结果为inframe的突变，结果在下表第三行。

训练集量数据表格

|  |  |  |
| --- | --- | --- |
|  | del\_1 | ins\_1 |
| github提供的数量 | 2238 | 680 |
| 论文中的数量 | 2151 | 679 |
| 注释并筛选出inframe后 | 2197 | 667 |

注：Del\_1表示缺失一个氨基酸的inframe突变，ins\_1表示插入一个氨基酸的inframe突变，SHINE使用这两部分样本作为训练集

结果表明处理后数据量也没有对上。可能原因是使用VEP版本的不同可能导致注释结果的差异，并且SHINE作者回复训练集的少量样本差异并不影响结果。

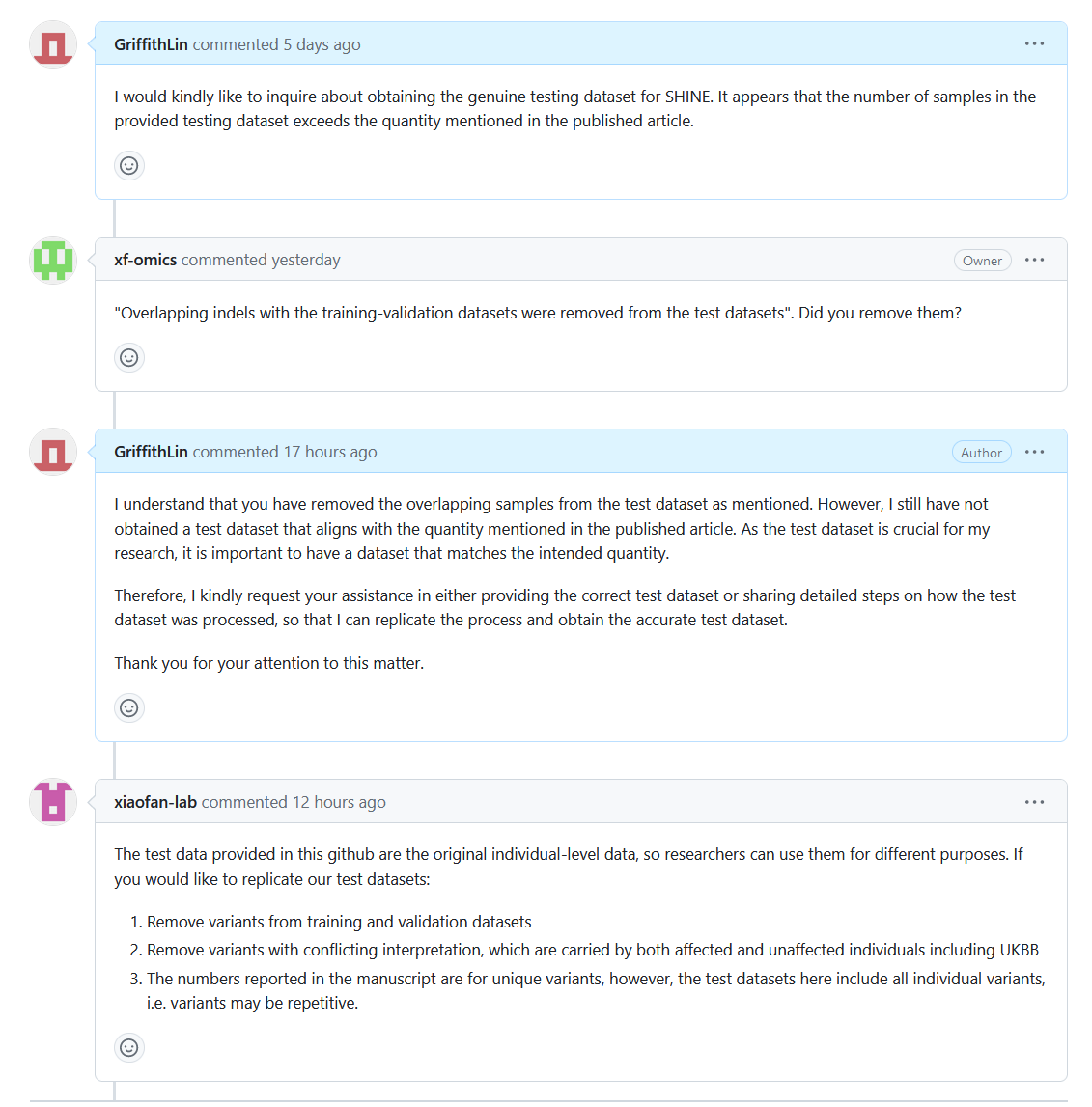
2、测试集的确认

github提供的测试集数据量大于论文中使用的数据量。

作者回复概括：①在测试集中，移除训练集和验证集中的数据。

②移除affect和unaffected两者矛盾的样本

③数据中心可能存在重复的样本，需要去除



具体处理步骤：

说明：训练集中有"ID", "Pathogenicity"两列。ID字段就包含了vcf文件中的染色体号、突变位点、REF。应该是作者拿来去重的标志。Pathogenicity字段包含了上文所提的affect和unaffected、还有根据来源不同Benign\_UKBB或者Pathogenic\_hotspot ，是来标记致病性的。

步骤：（四个测试集文件分开进行）。

先以"ID", "Pathogenicity"关键字去重，重复样本仅保留一条。

再以"ID"关键字去除矛盾样本，全部去除。

最后以"ID"关键字，去除和训练集重复的样本。

结果描述：只有hotspots数据集的插入突变数量是对得上的。

论文数量：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 来源 | NDD | | hotspots | |
| 种类 | del | ins | del | ins |
| （致病+中性） | 146+2808 | 35+1504 | 307+132 | 119+54 |

处理结果:

ndd\_ukb\_del.txt：

Benign\_UKBB 2886

Affected 155

Unaffected 2

. 1

ndd\_ukb\_ins.txt:

Benign\_UKBB 1501

Affected 30

Unaffected 1

hotspots\_ukb\_del.txt:

Pathogenic\_hotspot 367

Benign\_UKBB 136

hotspots\_ukb\_ins.txt:

Pathogenic\_hotspot 119

Benign\_UKBB 54

附：

训练集的确认步骤

1. 下载数据 ，并且转为vcf

下载地址为：<https://github.com/xf-omics/SHINE>

Vcf的转化程序：shine2vcf.py——将下载的数据集txt文件转为同名的vcf格式文件

其中训练集：clinvar\_gnomAD\_del\_1aa.vcf，clinvar\_gnomAD\_ins\_1aa.vcf

验证集：clinvar\_gnomAD\_ins\_2aa.vcf、clinvar\_gnomAD\_del\_2aa.vcf

1. vep官网注释

<https://asia.ensembl.org/Tools/VEP>

额外添加的注释选项为：“Identify canonical transcripts”，“Return results for variants in coding regions only”。作用是识别canonical transcripts，并且只显示编码区的结果。

注释结果选择下载txt文件。clinvar\_gnomAD\_del\_1aa\_canonical.txt和clinvar\_gnomAD\_ins\_1aa\_canonical.txt。

1. 处理注释后的文件筛选出经典转录本上，注释结果带有inframe，和仅为inframe的突变的条数。

程序：confirm\_traindata.py

输入：clinvar\_gnomAD\_del\_1aa\_canonical.txt、clinvar\_gnomAD\_ins\_1aa\_canonical.txt

输出：处理后的条数only\_inframe:2197、find\_inframe:2234、only\_inframe:667、find\_inframe:673

测试集的确认步骤

confirm\_testdata.py

输入：

ndd\_del\_file = "ndd\_ukb\_del.txt"

ndd\_ins\_file = "ndd\_ukb\_ins.txt"

hotspots\_del\_file = "hotspots\_ukb\_del.txt"

hotspots\_ins\_file = "hotspots\_ukb\_ins.txt"

del\_train\_file = "clinvar\_gnomAD\_del\_1aa.txt"

del\_val\_file = "clinvar\_gnomAD\_del\_2aa.txt"

ins\_train\_file = "clinvar\_gnomAD\_ins\_1aa.txt"

ins\_val\_file = "clinvar\_gnomAD\_ins\_2aa.txt"

输出：处理前和处理后的四个样本文件的数量情况。

如下所示：

ndd\_ukb\_del.txt:

Benign\_UKBB 3143

Affected 186

Unaffected 5

. 1

Name: Pathogenicity, dtype: int64

after\_drop!!!!!:

Benign\_UKBB 2886

Affected 155

Unaffected 2

. 1

Name: Pathogenicity, dtype: int64