单细胞可变剪切分析研发复盘

By：张春媛

1. **项目背景**

单细胞产品市场竞争越来越激烈，需要有区别于市场的衍生类产品为单细胞产品增加竞争力，单细胞可变剪切分析得到越来越多的关注，是单细胞衍生产品的不二之选。目前10X单细胞分析是公司最重要的单细胞产品之一，未来M20单细胞分析也会成为重要的单细胞产品。10X和M20均属于droplet-based单细胞数据，但是目前市场上尚缺乏针对droplet-based单细胞数据检测可变剪切的产品（Table 1），本项目研发同时适用于10X和M20的单细胞可变剪切分析流程。

Table 1．单细胞转录组分析市场调研

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 同行公司 | 常规单细胞分析 | 常规单细胞分析 | 可变剪切分析 | 可变剪切分析 |
| droplet-based | plate-based | droplet-based | plate-based |
| 诺禾致源 | √ | × | × | × |
| 美吉生物 | √ | × | × | × |
| 华大科技 | √ | √ | × | √ |
| 乐备实 | √ | × | × | × |
| 烈冰 | √ | × | × | × |
| 中科新生命 | × | × | × | × |
| 博奥晶典 | √ | × | × | × |
| 基迪奥 | √ | √ | × | √ |
| 晶能生物 | √ | √ | × | √ |
| 生工 | √ | √ | × | √ |
| 欧易 | √ | x | √ | x |

1. **项目目标**

开发同时适用于10X和M20的单细胞可变剪切分析流程。

1. **项目成员**

|  |  |  |
| --- | --- | --- |
| **项目经理** | 阳秀凤 | |
| **项目统筹人** | 张春媛 | |
| **项目主要负责人** | 张春媛 | |
| **项目成员** | 研发 | 林博、阳秀凤、公光业、张春媛、孙坤 |
| 技术支持 | 章璠、叶晓欣、李凯晴、王子衡 |
| 运营 | 陆瑶 |
| 产品 | 巴永兵 |

1. **项目过程**
   1. **项目里程碑**

本次研发从7月中旬开始做相关调研，并与8月正式立项并开始搭建流程，流程搭建同时考虑了实用性和降本增效，于9月20日出具第一版demo报告，和技术支持、产品、运营讨论后进一步升级，10月19日完成转产培训后，正式提交测试。



Figure 1．项目进展甘特图

* 1. **项目实施过程**
     1. **背景调研**

在项目立项之前针对单细胞水平的可变剪切、基因融合、SNP calling、lncRNA分析进行过一次可行性调研，并撰写调研报告。在正式立项之后系统的调研近期发表的相关工具，综合考量后，最终选择唯一专门为10X数据设计pipeline的工具（MARVEL）。

Table 2．检测单细胞可变剪切工具。



* + 1. **立项**

2023年8月7日正式正式提交立项，同时为10X数据和M20数据搭建可变剪切分析流程。



* + 1. **研发过程**

研发过程中遇到的各种问题、解决过程和最终方案都已整理归档，详情请参考《研发记录》。最终分析流程展示如下：

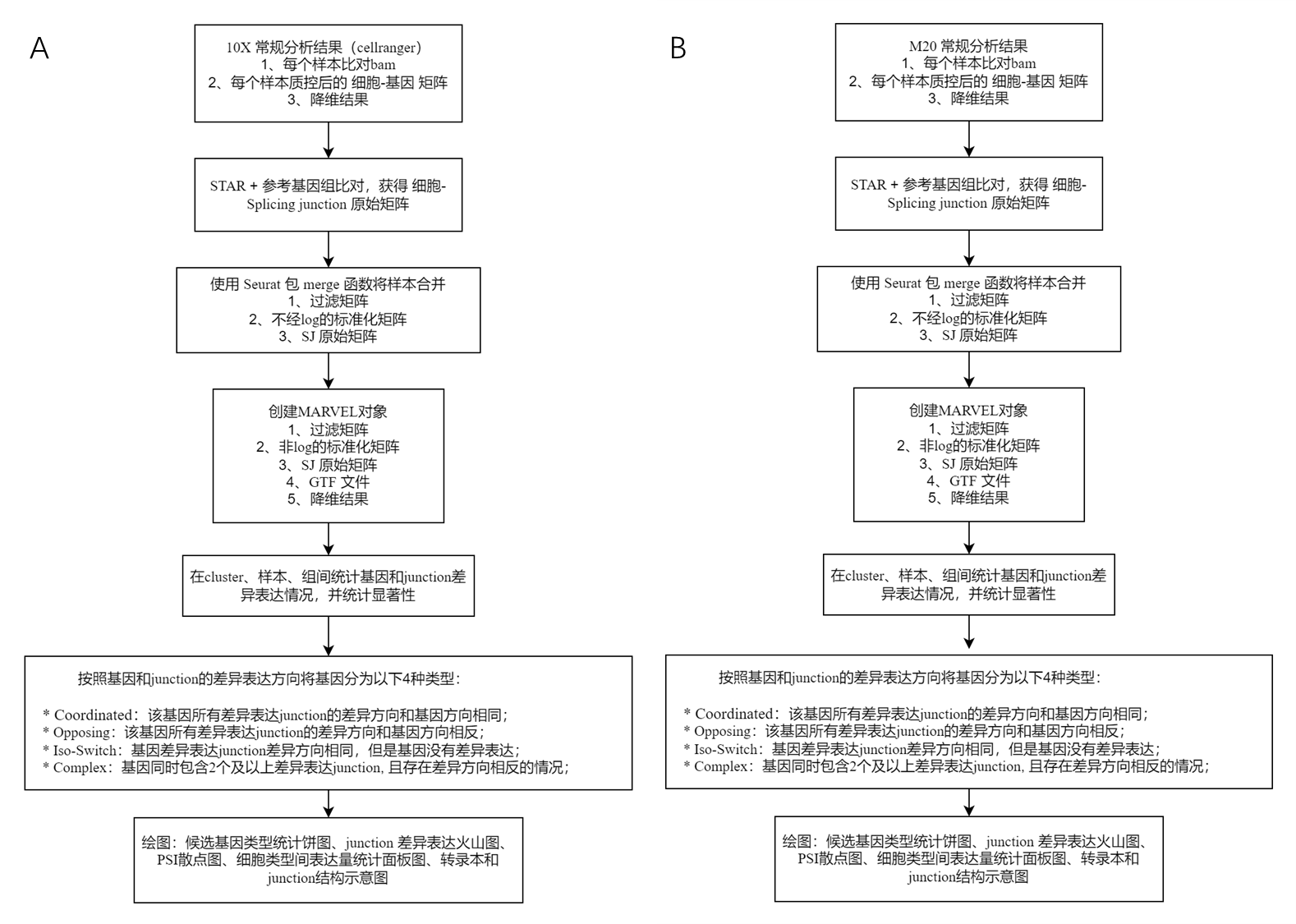


Figure 2．可变剪切流程图。（A）10X数据分析流程；（B）M20数据分析流程。10X和M20分析流程的主要不同在于获取上游数据和重新比对，在构建MARVEL对象后流程完全相同。

* + 1. **第一版demo报告**

报告主要由项目概况、MARVEL检测可变剪切基本原理、生信分析、项目分析结果、申明和附录组成。



* 1. **项目难点**
     1. **选择合适的工具**

单细胞可变剪切分析目前分析热度较高，处理针对单细胞数据做可变剪切分析的工具发布数量要大于实际应用的文章数量，尚没有人对当前已经发表的工具进行系统性的评估，且很少有针对10X数据的应用型文章，这就需要我们自己根据原理结合经验自行判断并作出选择。在比对过25款工具（Table 2）后，选择MARVEL工具作为可变剪切分析的基本工具。

针对本项目，选择工具的基本原则是：1、选择专门为droplet-based数据设计的工具，同时适用于两种数据的工具不选择；2、由于M20项目使用核RNA，不成熟RNA占50%以上，考虑多种可变剪切形式的工具不选择，选择仅考虑完整外显子剪切的工具；3、同2原因，使用cDNA作为参考基因组的工具不选择；

* + 1. **设计符合公司产品的方案**

MARVLE工具提供了现成的pipeline，但其并不符合公司目前的产品，本项目所定方案（Figure 2）在原有pipeline上做了部分升级和优化。详细内容见Table 3。

Table 3．在原有pipeline基础上升级优化。

|  |  |  |
| --- | --- | --- |
| MARVEL pipeline 处理方式 | 自行优化升级部分 | 优化理由 |
| 使用STAR比对获得基因-细胞矩阵和junction-细胞矩阵，并用这两个矩阵做后续分析。 | 仅使用STAR比对获取junction-cell矩阵，使用上游分析输出的gene-cell矩阵。10X数据使用cellranger输出矩阵，M20数据使用上游STAR输出矩阵。 | STAR比对的基因-细胞矩阵和cellranger输出的矩阵存在一定的区别，区别大小会随着项目的不同而发生改变，若使用STAR比对得到的基因-细胞矩阵，会导致可变剪切分析结果和原报告结果基因列表差距较大。 |
| 在比对时使用10X提供的原始白名单。 | 不再使用原始白名单，而是使用上游比对输出的barcode.tsv.gz文件作为可变剪切分析比对白名单。 | 因版本等原因，10X官方提供多个barcode白名单，cellranger会根据数据的特征和内置算法自动选择合适的白名单。但是我们使用STAR重新比对时需要手动选择正确的白名单，为了保证白名单的正确使用，我们不再直接使用库中提供白名单，而是使用上游常规分析输出的barcode.tsv.gz文件作为白名单，既保证了白名单使用的正确性，又降低了比对计算量。 |
| 统计基因在细胞中表达比例，以最高细胞表达密度值作为筛选基因的标准，过滤掉在较少细胞中表达的基因。 | 升级为两种方式：1、自动统计基因在细胞中表达的密度分布，获取最大值并以此筛选基因；2、使用固定值，默认为1%。 | 对7个项目进行统计，最高细胞表达密度值介于6~11%之间，在M20常规分析中该值设置为1%，一方面本项目和M20常规分析保持一致，另一方面较低的标准可以提供更多的结果。 |
| 使用delta和p-value作为候选基因的筛选标准。 | 丰富候选基因的筛选标准，基因差异显著性增加q-value，junction差异增加log2FoldChang | q-value和log2FoldChang是RNA差异表达分析中常用的筛选参数，但是MARVEL pipeline中并没有这两项，在分析比较7个项目后发现，delta和log2FoldChang对候选基因的筛选有不同的倾向，前者更倾向检测总体表达水平较高的基因，后者更倾向于检测总体表达水平较低的基因。 |
| 按照基因和junction差异方向将基因分为4类。 | 扩大Complex类别候选基因。 | 本项目调整了4中类型基因的分类标准，原pipeline的分类标准中仅考虑基因和junction同时差异表达的基因，但是前人的研究表明有不少的可变剪切在基因水平上并没有差异表达，但是相互表达不同的转录本，因此我们将只要含有两个及以上差异junction，且junction表达方向存在相反的基因都作为Complex类型的基因，并绘制该类型基因全部图片。 |

* + 1. **技术难点或流程亮点**

本项目执行过程中主要难点总结在Table 4中，更加详细内容请参考《研发记录》。

Table 4．技术难点或流程亮点。

|  |  |
| --- | --- |
| 难点或亮点 | 解决办法 |
| 改写MARVEL工具函数 | 虽然现有MARVEL pipeline中提供了各函数的使用方法，但是由于部分依赖R包升级等原因，部分函数并不能正确使用，加之为了方便传递参数，美化和重新排版图片等原因，本项目修改了几乎所有MARVEL自带函数。 |
| STARsolo比对结果和cellranger比对结果差异 | 在本项目中，可变剪切通过比较junction的差异表达，并结合基因差异表达筛选可能含有可变剪切的候选基因，为了获取junction的表达信息需要重新比对，为了让重新比对后的矩阵和cellranger比对的矩阵尽可能的相似，我们测试了STARsolo与比对相关的8个参数，并结合cellranger官方网站的建议，确定了当前的比对方案，详细信息请参考《研发记录》。 |
| Junction-cell矩阵过大导致运行失败 | 根据gene-cell矩阵提取高质量的细胞，注释junction，仅保留注释到完整外显子的junction，这样分别从cell和feature两个维度大大降低了junction-cell矩阵。 |
| 同时适配10X和M20数据以及外部数据 | 10X和M20数据常规处理方式不同，可变剪切流程也做了相应的适配。可变剪切流程对于两种数据在使用过程和结果呈现上没有任何不同。 |
| 默认参数设置 | 整个分析过程中，针对数据筛选和结果筛选的参数共有26个，结合5个真实项目的分析情况，设置好了默认参数，并给出参数使用建议。 |
| 一个基因对应多个ID处置方案 | 一般情况下一个gene\_name对应一个gene\_id，但是少数基因存在对应多个gene\_id的情况，这会导致在使用gene\_name做基因注释时损失信息或出现错误，在本项目中，重复出现的gene\_name添加“.1”、“.2”后缀以便区分（与常规分析保持一致），同时采用gene\_id进行注释，再按照对应关系转回gene\_name。既保证了信息的准确性，又避免手动修改注释文件。 |
| 任意比较组设置 | 可变剪切的比较需要在不同的比较组（细胞类型）间进行比较，流程目前采用完全传参的方式，任何常规分析中已经完成的分类结果均可以作为可变剪切分析的比较组。建议使用的比较组有：clusters、celltype、raw\_main\_celltype、sampleid、groups |

1. **项目的价值和结果**
   1. **项目价值**
      1. **推出全新产品**

此前，市场上尚缺乏基于droplet-based数据的单细胞水平的可变剪切分析产品，本次推出全新产品，能够利用10X和M20数据，在样本间、cluster间、细胞类型间等多种分组水平上检测可能含有可变剪切的候选基因。以示例数据为例（Figure 3），我们成功检测到*ATP5F1C* 基因的三个junction（chr10:7802855:7806973、chr10:7807011:7807658 以及chr10:7802855:7807658）在第2天和第4天的IPS细胞中差异表达，结合基因表达的差异方向和junction差异表达的方向以及转录本结构图，我们推断*ATP5F1C*基因在两种细胞间存在可变剪切，主要体现在是否包含第9个外显子上。详细信息请参考推广软文《新品上新 | 谁说10X单细胞的可变剪切做不了？来找欧易生物吧》

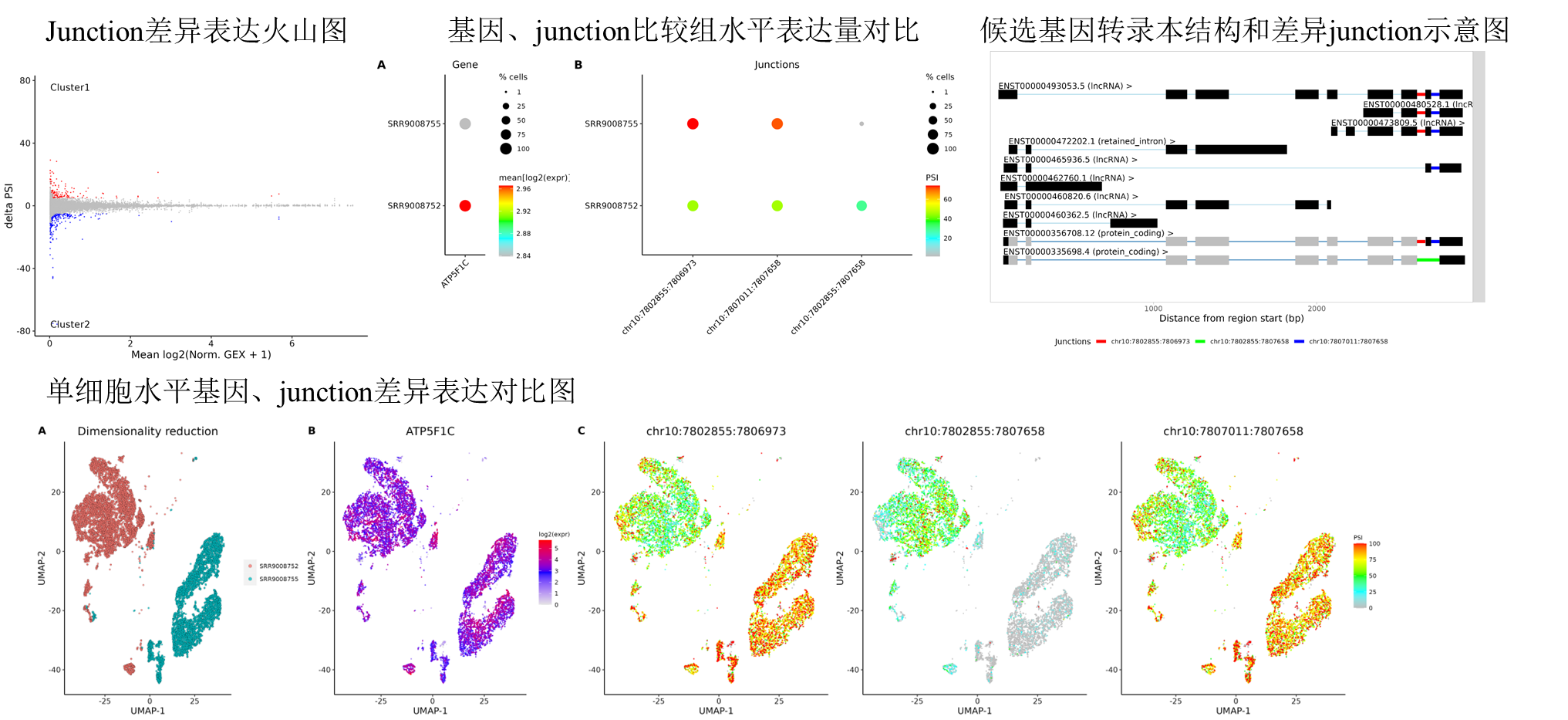
****

Figure 3．分析报告主要图片示例。

* + 1. **同时适配10X和M20项目**

公司目前有两款droplet-based产品，分别是10X和M20，新推出的可变剪切产品可以同时适配10X数据和M20数据，在操作上也不需要进行任何设置，流程会自动识别当前项目并进行后续分析，最终报告完全相同。

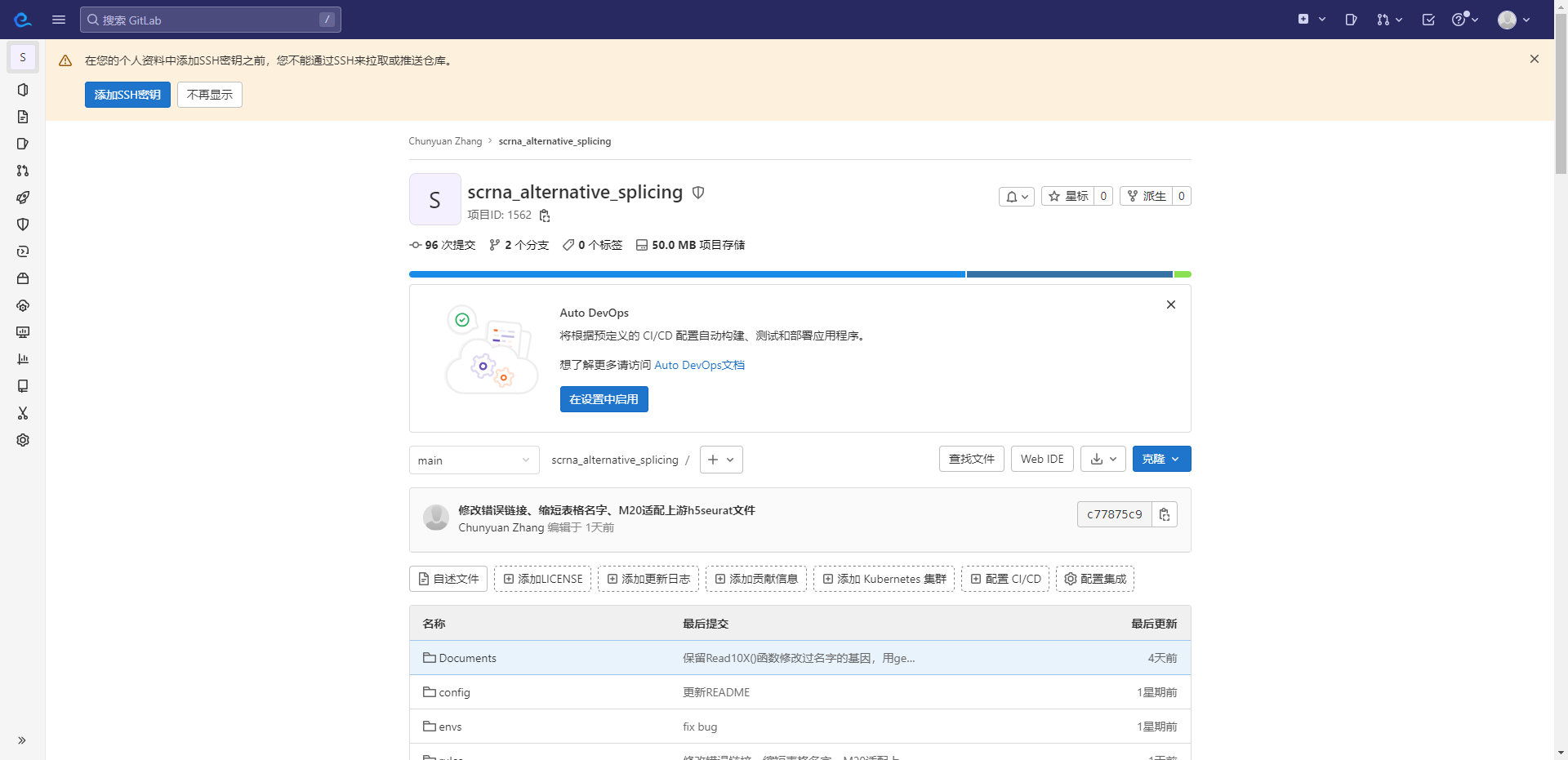
* + 1. **同时兼顾降本增效**

流程实现一键化操作，仅需要简单填写一下配置文件，即可提交运行，在配置好线上订单后可实现完全自动化。采用预先处理的方式降低矩阵占用内存。

【需添加一个统计项目消耗资源的表格】

1. **项目产出**

流程代码



[一篇推广软文](https://mp.weixin.qq.com/s?__biz=MjM5MjEyODY0OQ==&mid=2650657076&idx=1&sn=8f5aeaf750480b1d284757664d3d923e&chksm=bea27ad289d5f3c4da8aad663ed9bad3bd5f1366b6a1ceef93693e5d8d0709311a822ee689a0&mpshare=1&scene=1&srcid=10272l2b9vitOX1JMXFU38Zo&sharer_shareinfo=e3f1b45d95d41940486961ba124204c3&sharer_shareinfo_first=e3f1b45d95d41940486961ba124204c3&version=4.1.10.6015&platform=win#rd)



1. **经验教训**
2. **未来展望**

配套单细胞可变剪切后续验证方案。

1. **致谢**

感谢林博、阳秀凤、公光业、孙坤在项目执行过程中提供的支持。