可变剪切分析结果说明

我们使用MARVEL[1]工具为10X数据设计的[pipeline](https://wenweixiong.github.io/MARVEL_Droplet.html)检测不同细胞类型之间的可变剪切，首先使用STARsolo[2]工具获得splicing junction的数据矩阵，再利用MARVEL工具统计细胞类型之间的PSI值，经过100次重排检验（permutation）获得差异显著性，再根据显著性标准筛选splicing junction和基因（Figure 1）。

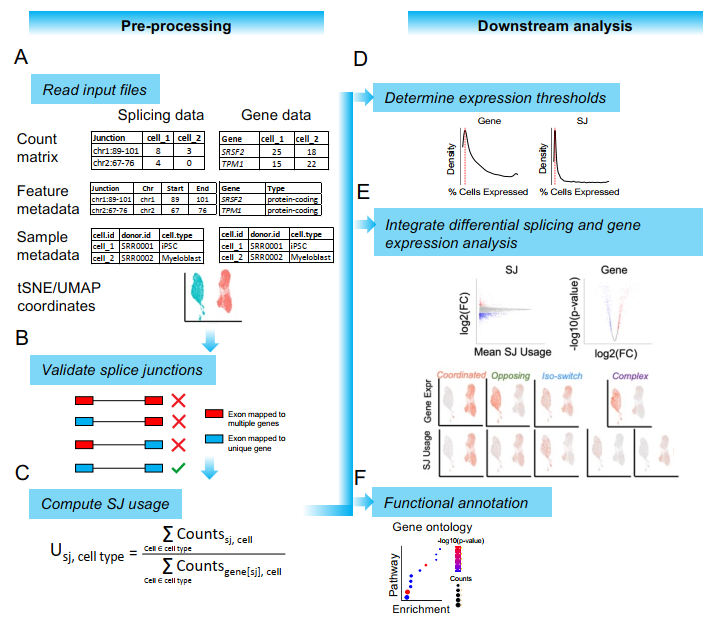


Figure ．MARVEL工具在不同细胞类型间检测可变剪切原理示意图。

**数据流动过程**：

1. 每个样本单独比对，获得SJ矩阵，从原项目中获得Gene矩阵
2. SJ矩阵未经过滤，非常巨大，服务器完全算不动，从Gene矩阵中提取细胞信息，按照细胞对SJ矩阵进行第一轮数据缩减，再对junction做注释，仅保留注释到外显子的junction，进行第二轮数据缩减，此时SJ矩阵的大小能够处理。
3. diff\_group.csv文件指定了比较组，在正式比较之前先做一下数据的筛选。比较组双方各自分别根据至少在百分之{filter\_gene\_in\_percent\_cell}的细胞中表达筛选基因和至少在百分之{filter\_sj\_in\_percent\_cell}的细胞中表达筛选junction。两个步骤分别进行，然后取基因和junction的交集。【实际效果】junction是基因的子集，如果基因不满足，那对应基因的所有junction都会被删除；在基因满足该条件的前提下，过滤掉基因对应的junction。
4. 数据筛选结束后，以junction为单位，计算比较组之间junction的差异水平和基因的差异水平，统计p-value和q-value。
5. 虽然计算时是 以junction为单位进行计算的，但是最终结果应以基因为单位进行统计。
6. 将候选基因分为4种类型，如果一个基因含有两个及以上junction，且含有差异方向相反的junction，这样的基因被定义为Complex，Complex是最有可能含有可变剪切的基因。

**参数设置方面的建议**：

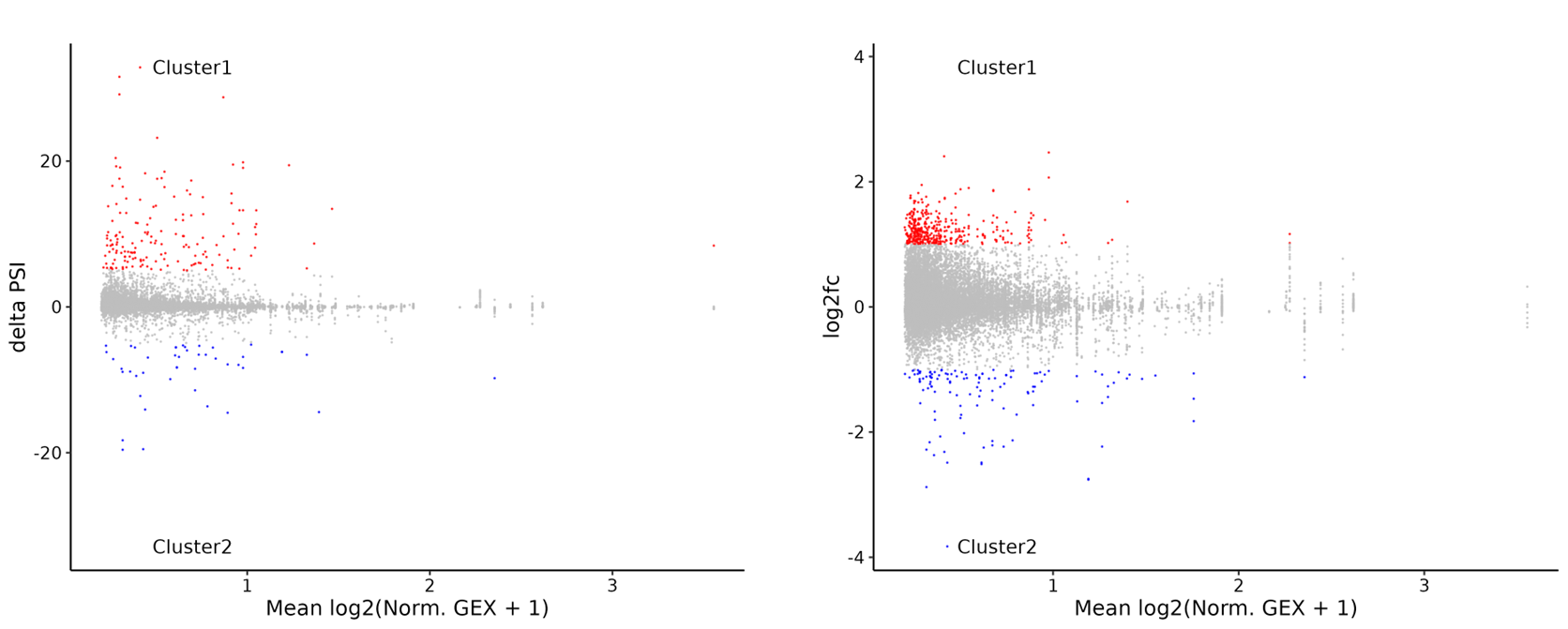


Figure ．差异junction火山图。左侧为以PSI绝对值为标准进行筛选时的结果，右侧为以log2fc(PSI)为标准进行筛选时的结果。

1. Wen, Wei X., A.J. Mead, and S. Thongjuea, *MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data.* Nucleic Acids Research, 2023. **51**(5): p. e29-e29.

2. Benjamin, K., Y. Dinar, and D. Alexander, *STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data.* bioRxiv, 2021: p. 2021.05.05.442755.