**真实基因组组装结果比较**

使用相同的原始数据，分别用新旧版本的过滤程序进行过滤，后续用相同的程序进行组装。

1. **数据过滤结果比较**

**表 1-1 新版本过滤程序数据量统计表**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Libraries** | **Insert Size(bp)** | **Raw**  **reads(M)** | **Raw**  **bases(Mb)** | **adapter**  **(%)** | **usable**  **reads(M)** | **usable**  **bases(Mb)** |
| CERqqxDBHDCAAPEI-12 | 200 | 74.61 | 7461.34 | **0.18** | 67.43 | **6540.79** |
| **FIGpqnDBHDIAPEI-52** | 500 | 189.54 | 18953.77 | **0.57** | 168.52 | **16346.59** |
| CERqqxDBHDIAAPEI-4 | 500 | 89.19 | 8919.18 | **0.03** | 69.79 | **6769.85** |
| CERqqxDBHDMAAPEI-6 | 800 | 80.14 | 8013.71 | **0.01** | 57.83 | **5609.94** |
| FIGruoDABDWAAPEI-9 | 2000 | 123.20 | 6036.70 | **0.23** | 97.74 | **4300.71** |
| FIGruoDAADLAAPEI-12 | 5000 | 101.72 | 4984.07 | **0.09** | 53.14 | **2337.98** |
| FIGruoDACDTAAPEI-7 | 10000 | 89.36 | 4378.65 | **0.18** | 34.51 | **1518.57** |
| FIGruoDACDUAAPEI-2 | 20000 | 110.28 | 5403.69 | **0.03** | 20.99 | **923.42** |
| Total |  | 858.04 | 64151.11 |  | 569.95 | 44347.85 |

**表 1-2 旧版本过滤程序数据量统计表**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Libraries** | **Insert Size(bp)** | **Raw**  **reads(M)** | **Raw**  **bases(Mb)** | **adapter**  **（%）** | **usable**  **reads(M)** | **usable**  **bases(Mb)** |
| CERqqxDBHDCAAPEI-12 | 200 | 74.61 | 7461.34 | **0.61** | 67.11 | **6509.36** |
| **FIGpqnDBHDIAPEI-52** | 500 | 189.54 | 18953.77 | **1.00** | 167.45 | **16744.67** |
| CERqqxDBHDIAAPEI-4 | 500 | 89.19 | 8919.18 | **0.45** | 69.45 | **6736.46** |
| CERqqxDBHDMAAPEI-6 | 800 | 80.14 | 8013.71 | **0.44** | 57.53 | **5580.02** |
| FIGruoDABDWAAPEI-9 | 2000 | 123.20 | 6036.70 | **0.45** | 97.52 | **4291.06** |
| FIGruoDAADLAAPEI-12 | 5000 | 101.72 | 4984.07 | **0.34** | 53.01 | **2332.42** |
| FIGruoDACDTAAPEI-7 | 10000 | 89.36 | 4378.65 | **0.44** | 34.44 | **1515.28** |
| FIGruoDACDUAAPEI-2 | 20000 | 110.28 | 5403.69 | **0.29** | 20.93 | **920.97** |
| Total |  | 858.04 | 64151.11 |  | 567.44 | 44630.24 |

以上两个版本过滤条件除了**FIGpqnDBHDIAPEI-52**外均相同，可以看出在相同过滤条件下，新版本过滤程序过滤掉的含adapter污染的数据量减少，得到的有效数据比旧版本的多。

1. **K-mer分析结果比较**

图1-1 K-mer分析

从上图来看，新旧版本的过滤程序得到的K-mer分析的结果没有明显变化。

1. **组装结果比较**

**表1-3 新旧版本过滤程序的补洞结果统计**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | estimated gap length | extended gap length | gap number | fully filled gap number | fully filled gap ratio |
| old version | 9,213,502 | 6,641,548 | 84,111 | 61,162 | **72.72%** |
| new version | 8,825,466 | 6,434,346 | 80,528 | 63,872 | **79.32%** |

可以看出，新版本过滤程序对补洞率的提升有比较大的作用。

**表 1-4 新版本过滤程序得到的基因组组装长度统计表**

|  | **Contig** | | **Scaffold** | |
| --- | --- | --- | --- | --- |
| **Size(bp)** | **Number** | **Size(bp)** | **Number** |
| N90 | 30,027 | 2058 | 1,730,846 | 36 |
| N80 | 53,734 | 1379 | 3,289,210 | 25 |
| N70 | 81,684 | 962 | 5,282,720 | 19 |
| N60 | 111,254 | 688 | 7,272,407 | 14 |
| N50 | **147,755** | 449 | **9,156,408** | 11 |
| Longest | 1,788,803 |  | 26,023,237 |  |
| Total Size | 279,664,763 |  | 280,553,330 |  |
| Total Number(>=100bp) |  | 10,307 |  | 6,021 |
| Total Number(>=2kb) |  | 4,255 |  | 564 |

＊此表中，包括拼接的scaffold和最后的contig的N50~N90的长度信息，总长度和最长的长度信息，统计信息不包括长度短于100bp的片段。其中，contig为补洞后得到的contig。计算N50时以组装得到的序列总长作为标准。

**表 1-5 旧版本过滤程序得到的基因组组装长度统计表**

|  | **Contig** | | **Scaffold** | |
| --- | --- | --- | --- | --- |
| **Size(bp)** | **Number** | **Size(bp)** | **Number** |
| N90 | 17,413 | 3,956 | 2,085,069 | 30 |
| N80 | 29,751 | 2,764 | 5,315,380 | 21 |
| N70 | 41,901 | 1,983 | 6,357,092 | 16 |
| N60 | 56,273 | 1,414 | 8,037,603 | 13 |
| N50 | **74,395** | 986 | **9,558,897** | 10 |
| Longest | 683,425 |  | 27,400,720 |  |
| Total Size | 276,647,649 |  | 277,939,842 |  |
| Total Number(>=100bp) |  | 15,018 |  | 7,397 |
| Total Number(>=2kb) |  | 6,946 |  | 256 |

＊此表中，包括拼接的scaffold和最后的contig的N50~N90的长度信息，总长度和最长的长度信息，统计信息不包括长度短于100bp的片段。其中，contig为补洞后得到的contig。计算N50时以组装得到的序列总长作为标准。

从表1-4和表1-5可以看出，新版本过滤程序得到的组装结果中，1）contig N50显著提高。2）contig总长和scaffold总长有一些增加。3）scaffold N50略有下降。

1. **结论**

新版本的过滤程序能够保留更多的有效数据，使得contig的N50有显著提升，对scaffold的N50等有比较轻微的负面影响。