Recently, I have been working on two things:

1. The modification method for calculation QUAL value in GAP5\_snp (and GAP5\_indel) to make this value more reasonably. In GATK, the bigger the QUAL, the higher the value of TP is (as Fig.1). But in GAP5, it’s not (as Fig.2). I’ve tried several methods, but results are still not very good.

Fig1：QUAL vs TP of INDELs result in GATK

|  |  |  |  |
| --- | --- | --- | --- |
| QUAL | total | TP | FP |
| 0-500 | 3982 | 441 | 3541 |
| 500-1000 | 4621 | 1073 | 3548 |
| 1000-1500 | 5930 | 2353 | 3577 |
| 1500-2000 | 7556 | 3905 | 3651 |
| 2000-2500 | 6480 | 3771 | 2709 |
| 2500-3000 | 4090 | 2575 | 1515 |
| 3000-3500 | 1984 | 1320 | 664 |
| 3500-4000 | 1016 | 710 | 306 |
| 4000-4500 | 451 | 317 | 134 |
| 4500-5000 | 256 | 181 | 75 |
| 5000- | 488 | 292 | 196 |
| total | 36854 | 16938 | 19916 |

Fig2：QUAL vs TP of INDELs result in GAP5\_indel

|  |  |  |  |
| --- | --- | --- | --- |
| QUAL | total | TP | FP |
| 0-1 | 2049 | 1140 | 909 |
| 1-10 | 4234 | 2312 | 1922 |
| 10-50 | 6298 | 3278 | 3020 |
| 50-100 | 6291 | 2818 | 3473 |
| 100-150 | 5005 | 2450 | 2555 |
| 150-200 | 3262 | 1724 | 1538 |
| 200-250 | 2165 | 1245 | 920 |
| 250-300 | 755 | 465 | 290 |
| 300-350 | 254 | 141 | 113 |
| 350-400 | 136 | 80 | 56 |
| 400-450 | 74 | 47 | 27 |
| 450-500 | 57 | 32 | 25 |
| 500- | 170 | 73 | 97 |
| total | 30750 | 15805 | 14945 |

1. Trying to propose a larger-INDELs calling pipeline. After testing data by some larger INDELs calling tools (Pindel, Delly, BreakDancer, SOAPindel ect.), I find the best tools are Pindel and SOAPindel. I am learning from these two pipelines now.