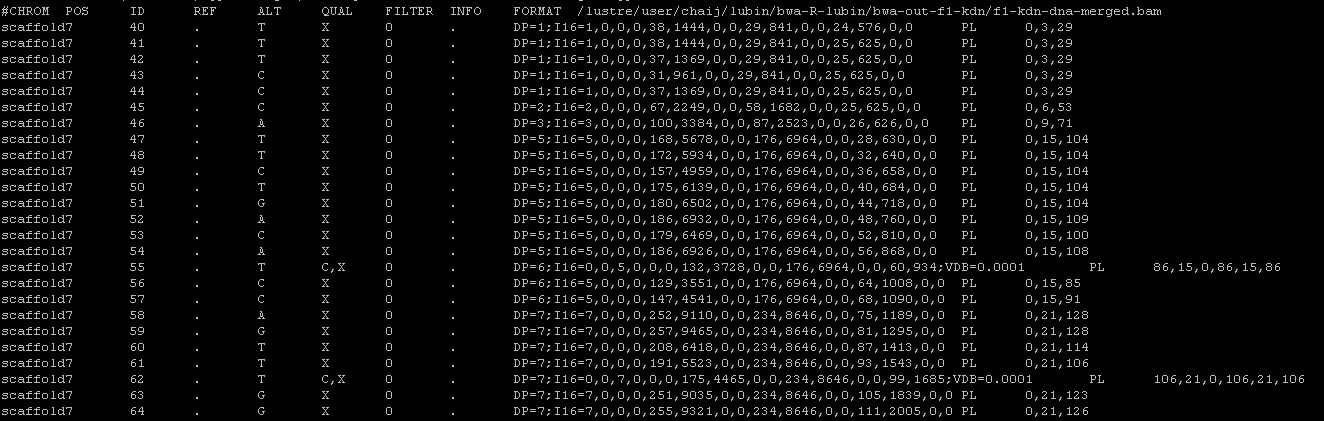
1. **根据bcf文件，在dna和rna的mapping范围中，求它们mapping的交集（mapping重叠区域）**

以f1个体肾脏组织为例（路径/lustre/user/chaij/lubin/bwa-R-lubin/bwa-out-f1-kdn/）

输入F1-kdn-dna-merged.bcf和F1-kdn-dna-merged.bcf

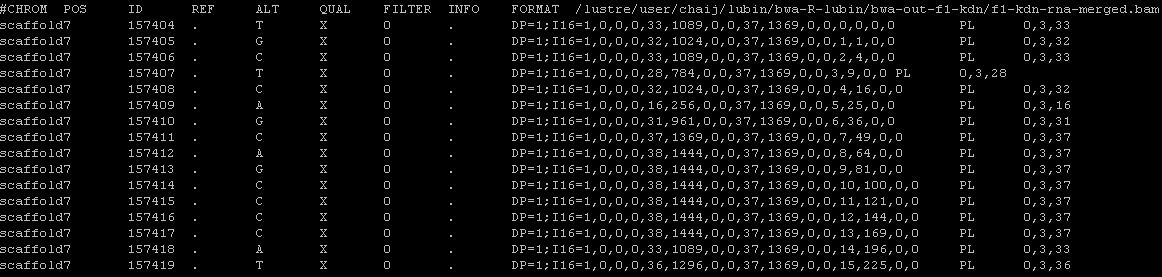
F1-kdn-dna-merged.bcf格式如下：

**列1 列2 列3 列4 列5 列6 列7 列8**



F1-kdn-rna-merged.bcf格式如下：

**列1 列2 列3 列4 列5 列6 列7 列8**



两个输入文件中，在列1相同下，取出两个输入文件列2相同的数值范围存于输出文件列2 中，输入文件列1存入输出文件列 1

输出文件名：f1-kdn-dna-rna shared mapping region，输出文件名输出格式如下：

|  |  |
| --- | --- |
| 列1 （scaffold） | 列2（dna与rna mapping交集范围） |
| CHROM | Region |
| Scaffold 7 | 2900-7500 |
|  |  |

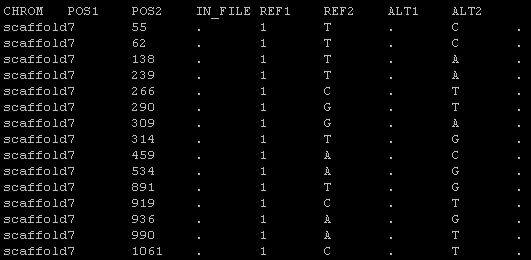
1. **根据交集，处理diff文件，只保留在交集范围的位点**

输入 f1-kdn-dna-rna shared mapping region（上一步生成文件）和f1-kdn-dnavsrna.diff.sites\_in\_files

输入文件1：f1-kdn-dna-rna shared mapping region格式如下：

|  |  |
| --- | --- |
| 列1 | 列2（dna与rna mapping交集范围） |
| CHROM | Region |
| Scaffold 7 | 2900-7500 |
|  |  |

输入文件2：f1-kdn-dnavsrna.diff.sites\_in\_files格式如下：



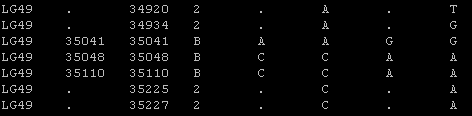
**列1 列2 列3 列4 列5 列6 列7 列8**

1. 在两个输入文件列1 相同的情况下（相同的scaffold），保留输入文件2列2（POS2）处于输入文件1列2 （region）范围内的所有位点，并将各个位点的整行信息存于输出文件中；
2. 在输出文件的列9增加突变类型（type）

突变类型的判定如下：

1. 识别输入文件2列5和列7为碱基，列6和列8为“.” 的位点，列8碱基就与列5一致，突变类型记为列7-列8（537为例，type记为G-A），并存入输出文件的列9中；

**列1 列2 列3 列4 列5 列6 列7 列8**



B. 识别输入文件2列5和列7为“.”，列6和列8为碱基的位点，列7碱基就与列6一致，突变类型记为列7-列8（34934为例，type记为G-A）

C. 识别输入文件2列7和列8碱基一致情况下（如35110处位点），此种情况不作统计。

输出文件名为f1-kdn-dnavsrna.diff-filter.out，输出格式如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| CHROM | POS1 | POS2 | IN\_FILE | REF1 | REF2 | ALT1 | ALT2 | TYPE |
| Scaffold 7 | 534 | . | 1 | A | . | G | . | G-A |
| LG49 | . | 34934 | 2 | . | A | . | G | A-G |

3）得出以上结果后，生成统计报告，内容包括：突变类型（Type），各种类型下的突变数分别为多少（NO.），突变总数为多少（Total），各种类型下突变占总数的百分比是多少(Rate)。

输出的统计报告格式如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 |
| Tpye | NO. | Total | Rate |
| A-T |  |  |  |
| A-G |  |  |  |
| A-C |  |  |  |
| T-A |  |  |  |
| T-G |  |  |  |
| T-C |  |  |  |
| …… |  |  |  |

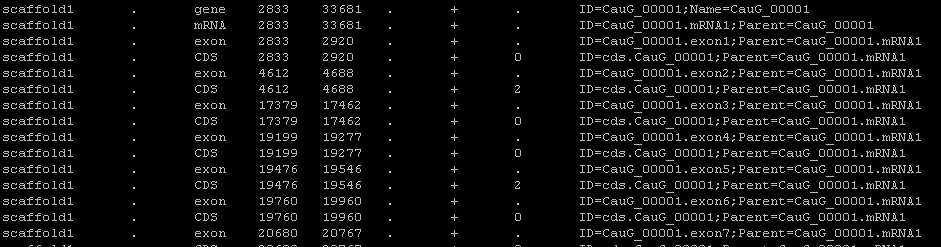
三、**对已注释的父母本基因,进行四个区域的划分 CDs, UTR, intron和intergenic**

输入母本参考基因组 fish.gff3和父本参考基因组V1.0.gff（路径：/lustre/user/chaij/reference-gf/以及/lustre/user/chaij/genome/carp-1507/）

输入文件：母本参考基因组 fish.gff3格式如下：

1.正、反链及正反链交集拆分

**列1 列2 列3 列4 列5 列6 列7 列8 列9**



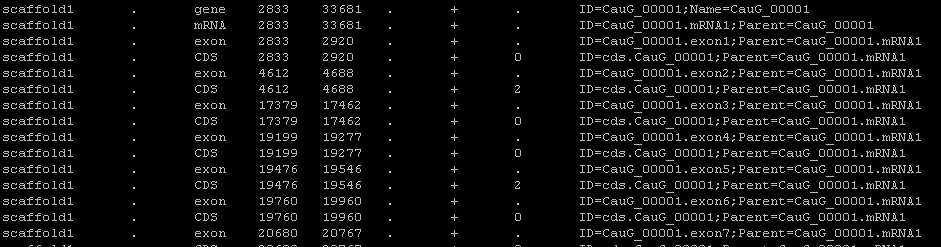
1)识别列7为“+”的位点，将该位点正行都存入输出文件中，输出文件命名为fish.gff3-plus（正链）

2)识别列7 为“-”的位点，将该位点正行都存入输出文件中，输出文件命名为fish.gff3-anti（反链）

3) 正、反链交集的gene，输出文件命名为fish.gff3-shared（正反链共有区域）

输出文件格式如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| Scaffold1 | . | gene | 2833 | 33681 | . | + | 0 | ID=cds.CauG\_00001;Name=CauG\_00001 |
|  |  |  |  |  |  |  |  |  |



**列1 列2 列3 列4 列5 列6 列7 列8 列9**

2.CDs区统计

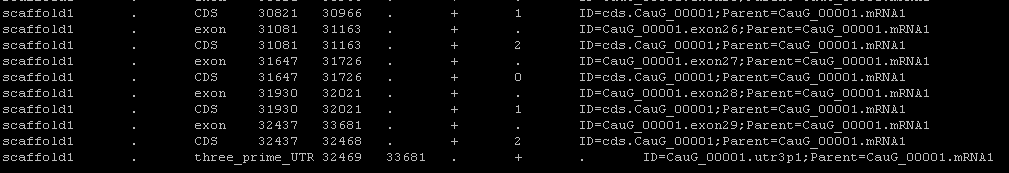
识别列3中为CDs的字符，统计该行的所有数据(列1——列9)到输出文件中

输出文件名为fish-giff3- plus -cds，（**如果在反链中，输出文件名为**fish-giff3- anti-cds）输出格式如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| Scaffold1 | . | CDs | 2833 | 2920 | . | + | 0 | ID=cds.CauG\_00001;Parent=CauG\_00001.mRNA1 |

3.UTR区统计

**列1 列2 列3 列4 列5 列6 列7 列8 列9**



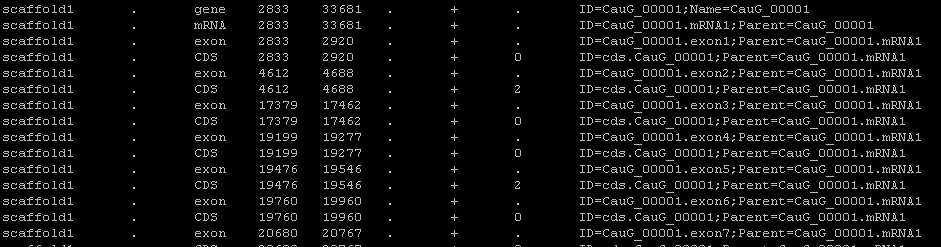
识别列3中带UTR的字符串，统计该行的所有数据(列1——列9)到输出文件中

输出文件名为fish-giff3-plus-UTR，输出格式如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| Scaffold1 | . | Three\_prime\_UTR | 32469 | 33681 | . | + | . | ID= CauG\_00001.utr3p1;Parent=CauG\_00001.mRNA1 |

4.intron区统计

**列1 列2 列3 列4 列5 列6 列7 列8 列9**



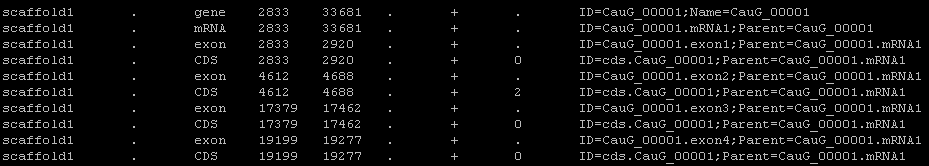
1. 在识别列1相同的情况下（相同scaffold），识别列3为exon的字符，前一个exon字符所在行的列5数值加“1”所得结果存于输出文件的列4中，后一个exon字符所在行的列4数值减“1”所得结果存于输出文件的列5中;
2. 前一个exon的列6，列7，列8字符分别存于输出文件的列6，列7，列8中。输出文件列1中字符存于输入文件列1，在输出文件的列2中存入“intron”，输出文件列9中存入gene行的列9字符串

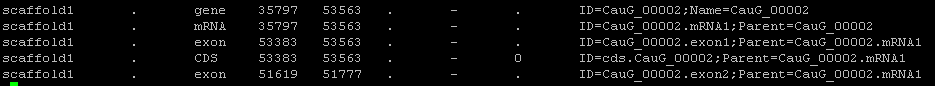
输出文件名为输出文件名为fish-giff3-plus-intron，输出格式如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| Scaffold1 | . | intron | 2921（2920+1） | 4611（4612-1） | . | + | . | ID=CauG\_00001;Name= CauG\_00001 |

5.interginc区统计

**列1 列2 列3 列4 列5 列6 列7 列8 列9**





**列1 列2 列3 列4 列5 列6 列7 列8 列9**

在识别列1字符一致下（相同scaffold），识别输入文件列3中为gene的字符，前一个gene所在行的列5数值加“1”所得结果存入输出文件的列4中，后一个gene所在行的列4数值减“1”所得结果存于输出文件列5中，在输出文件列1中存入对应的scaffold（输出文件1列1），在输出文件的列2中存入“intergenic”，输出文件列6中存入两个gene所在行的列9信息

输出文件名为输出文件名为fish-giff3-plus-intergenic，输出格式如下：

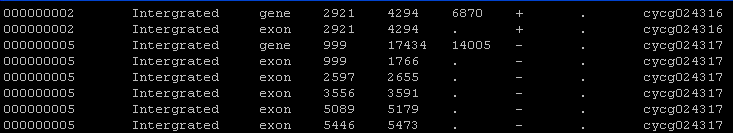
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列 5 | 列6 | 列7 | 列8 | 列9 |
| Scaffold1 | . | Intergenic | 33682（33681+1） | 35796（35797-1） | . | - | . | ID=CauG\_00001;Name= CauG\_00001  ID=CauG\_00002;Name= CauG\_00002 |

**父本参考基因组**V1.0.gff

（路径：/lustre/user/chaij/genome/carp-1507/）

输入文件V1.0.gff格式如下：

**列1 列2 列3 列4 列5 列6 列7 列8 列9**



1.exon区统计

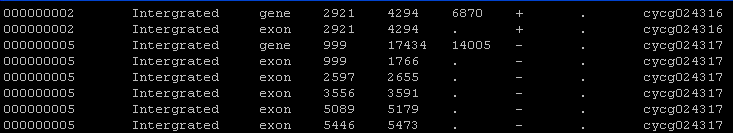
识别列3中为exon的字符串，统计该行的所有数据(列1——列9)到输出文件中。

输出文件名为V1.0.gff-plus-exon，输出格式如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| 000000002 | Intergrated | exon | 2921 | 4294 | . | + | . | cvcg024316 |

2.intron区统计

**列1 列2 列3 列4 列5 列6 列7 列8 列9**



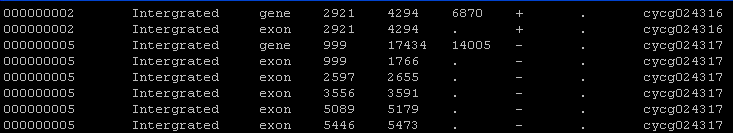
1. 在识别列1相同情况下，识别列3为exon的字符，前一个exon字符所在行的列5数值加“1”所得结果存于输出文件的列4中(1766+1)，后一个exon字符所在行的列4数值减“1”所得结果存于输出文件的列5中(3556-1);
2. 前一个exon的列6，列7，列8字符分别存于输出文件的列6，列7，列8中。输出文件列1中字符存于输入文件列1，在输出文件的列2中存入“intron”，输出文件列9中存入gene行的列9字符串

输出文件名为输出文件名为V1.0.gff-plus-intron，输出格式如下：

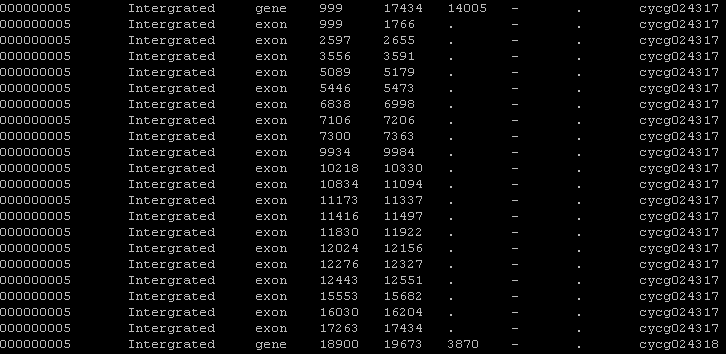
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| 000000005 | Intergrated | intron | 1767(1766+1) | 3555(3556-1) | . | + | . | cvcg024317 |

3.interginc区统计

**列1 列2 列3 列4 列5 列6 列7 列8 列9**



1. 识别输入文件列3中为gene的字符，排除V1.0.gff中第一个gene区域（从2921-4294），从第二个gene开始计数。（因为两个gene间区域有所重叠）



**列1 列2 列3 列4 列5 列6 列7 列8 列9**

1. 前一个gene所在行的列5数值加“1”所得结果存入输出文件的列4中(17434+1)，后一个gene所在行的列4数值减“1”所得结果存于输出文件列5中(18900-1)，在输出文件列1中存入输出文件1列1，在输出文件的列2中存入“intergenic”，输出文件列6中存入两个gene所在行的列9信息

输出文件名为输出文件名为V1.0.gff-anti-intergenic，输出格式如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列 5 | 列6 |
| 000000005 | Intergrated | Intergenic | 17435(17434+1) | 18899(18900-1) | cvcg024317  cvcg024318 |

**四、统计出cds, UTR, intron, intergenic 中突变位点信息，输出的文件名为：**

输入文件f1-kdn-dnavsrna.diff-filter.out（第2步生成）和fish-giff3-plus-cds, fish-giff3-plus-UTR, fish-giff3-plus-intron, fish-giff3-plus-intergenic,

输入文件1格式f1-kdn-dnavsrna.diff-filter.out

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 | 列10 |
| CHROM | POS1 | POS2 | IN\_FILE | REF1 | REF2 | ALT1 | ALT2 | TYPE | Total：？ |
| Scaffold 7 | 534 | . | 1 | A | . | G | . | G-A |  |
| LG49 | . | 34934 | 2 | . | A | . | G | A-G |  |

输入文件2 格式fish-giff3-cds (以CDs为例)

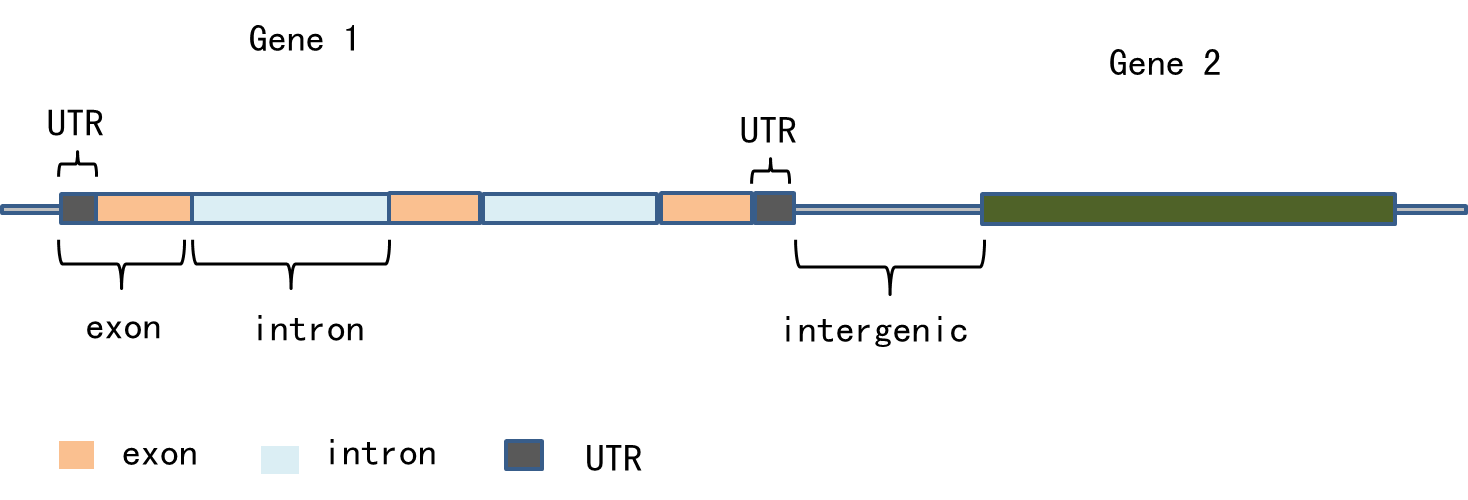
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 |
| Scaffold1 | . | CDs | 2833 | 2920 | . | + | 0 | ID=cds.CauG\_00001;Parent=CauG\_00001.mRNA1 |

1. 在两个输入文件列1相同下，输入文件1列2或列3数值处于输入文件2列4与列5范围内的位点，将该位点的所有突变类型统计到输出文件的列12中。（输出文件是在fish-giff3-plus-cds基础上加了列10——列12）
2. 输出文件列10的数值为列12中所有突变的总数；
3. 输出文件列11的百分比为列10数值除以总的突变数（统计报告中的total值）

输出文件1为：f1-kdn-dnavsrna.diff-filter-plus-cds.out，输出文件格式为：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 | 列10 | 列11 | 列12 |
| Scaffold1 | . | CDs | 2833 | 2920 | . | + | 0 | ID=cds.CauG\_00001;Parent=CauG\_00001.mRNA1 | 3 | ?% | G-A,A-T,C-G |

**以上是CDs区 的统计，其他区域（UTR，intron，intergenic）的统计情况与上相同**

****