Information are extracted from the BAM files, using the following samtools command,

```
samtools view  restseq_data_hg_sorted.bam |perl -lane 'print "$F[1]\t$F[2]\t$F[3]\t$F[4]\t$F[5]\t$F[6]\t$F[7]\t$F[8]"' > restseq_data_hg_sorted.hits

samtools view  bliss_data_hg_sorted.bam |perl -lane 'print "$F[1]\t$F[2]\t$F[3]\t$F[4]\t$F[5]\t$F[6]\t$F[7]\t$F[8]"' > bliss_data_hg_sorted.hits
```

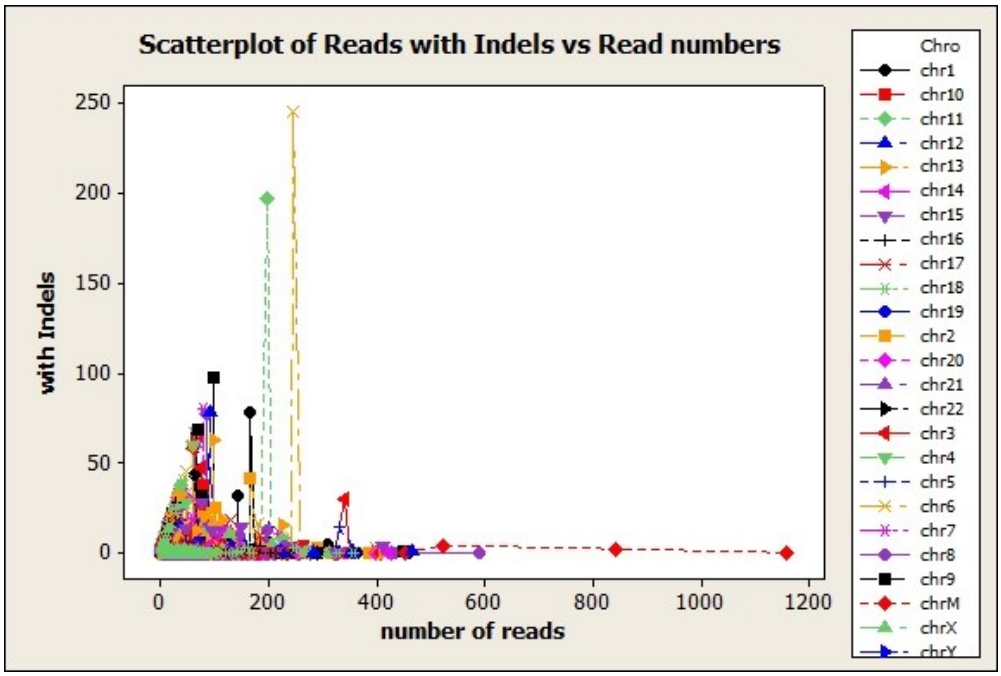
Information in the restseq_data_hg_sorted.hits and bliss_data_hg_sorted.hits files are consequently summarized using codes, written in R.

In restseq data:

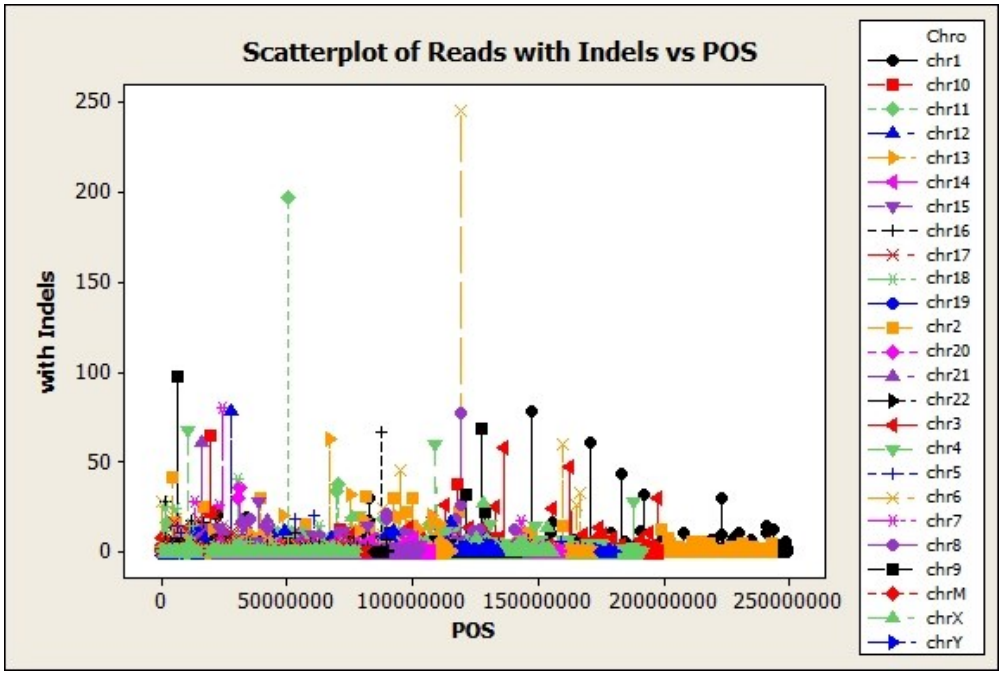
I have written a code in R to extract information from the restseq file. In each genomic position, this code calculates the total read counts and also the number of reads with indels, mapped to that position.

https://github.com/CNVdetection/Quiz/blob/master/restseq.R

Scatter plot for the total number of reads vs. number of reads with Indels, for each genomic position in the restseq data.



Scatter plot for the genomic position vs. number of reads with Indels, in the restseq data.

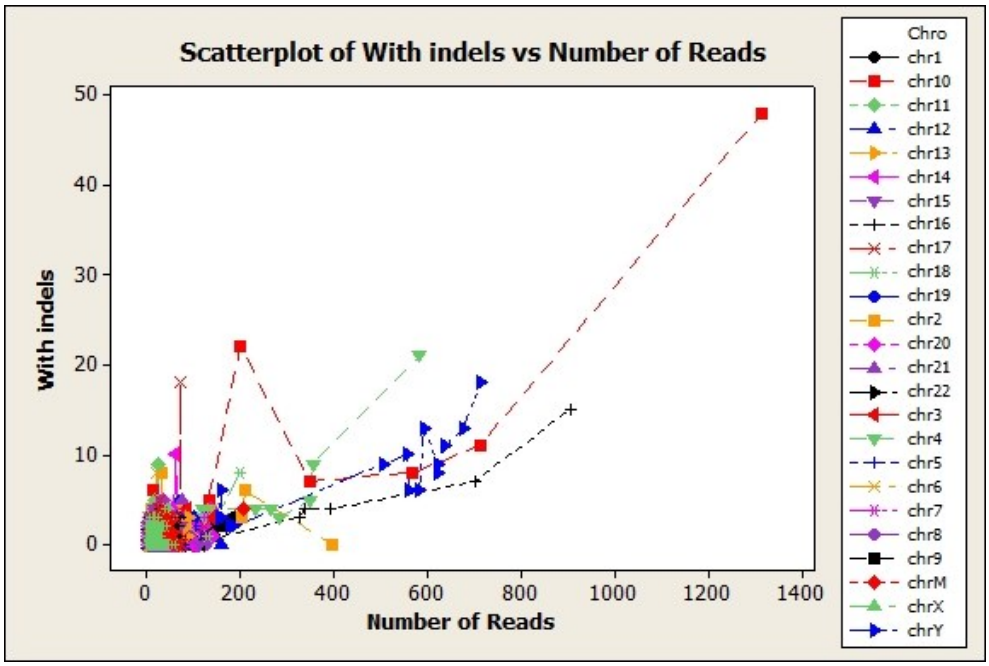


In bliss data:

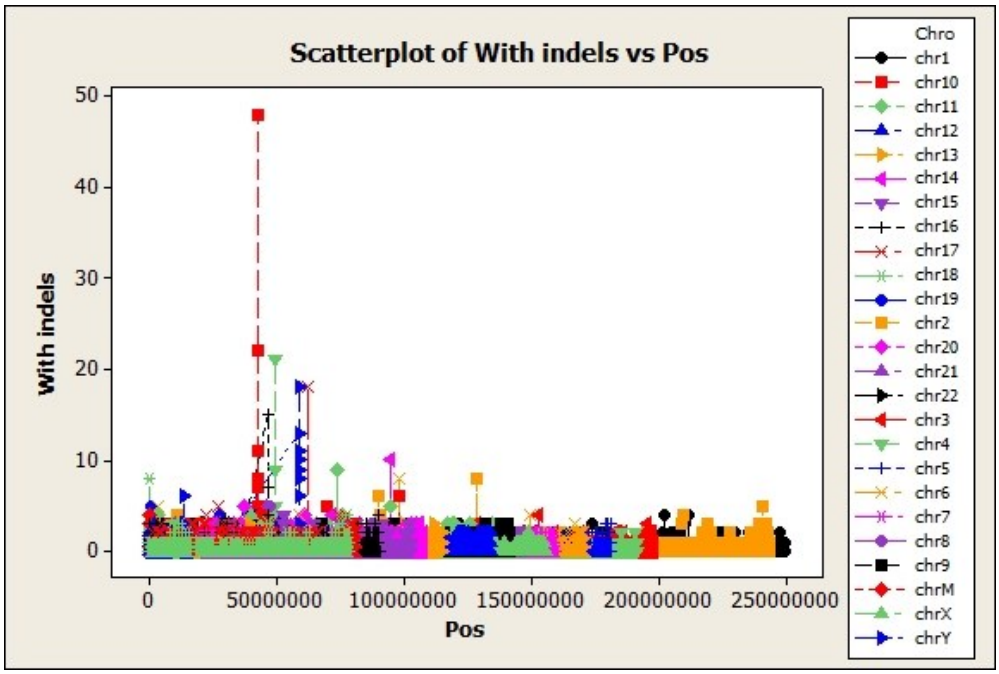
I have written a code in R to extract information from the bliss file. Since in the bliss data we have the whole-genome sequencing data and many reads are mapped to locations which are close to each other, to have a better visualization of data I have summarized information in genomic segments of length 10 kb. This code calculates the total read counts and also the number of reads with indels, mapped to genomic segments with length 10 kb.

https://github.com/CNVdetection/Quiz/blob/master/bliss.R

Scatter plot for the total number of reads vs. number of reads with Indels, for each genomic position in the bliss data.



Scatter plot for the genomic position vs. number of reads with Indels, in the bliss data.



Further Analysis:

In the restseq data, the average number of reads which are mapped to each targeted regions is 6.22, with a standard deviation of 16.42.

Average + 3*standard deviation=6.22+3*16.42=55.48

In the bliss data, the average number of reads which are mapped to a genomic segment of length 10 kb is 6.2, with a standard deviation of 11.

Average + 3*standard deviation=6.2+3*11=39.2

Therefore,

- In restseq data, the critical value for the number of reads which are mapped to those targeted genomic regions is 55.48.
- In bliss data, the critical value for the number of reads, which are mapped to each genomic segment of length 10 kb, is 39.2.

Based on the above critical values,

- **In the restseq data, there are many regions in which the number of abnormal reads i.e. with indels, are statistically significant.**
- **In bliss data, there is only one region in chromosome 10 in which the number of abnormal reads are higher that the critical level of 39.2.**