NGS DATA COMPARISON

Perl Script : comapre.pl

R Script : compare.R

Created By : Ankur Ganveer

Motivation

The development of ‘compare pipeline’ was motivated by the need of having a handy robust tool to compare two different algorithms or files at different steps of a pipeline which processes NGS data. In the field of bioinformatics there are several algorithms and tools which serve the similar purposes and claims to be the best, but they produce different results. It is important to find out which of the tools is better than other, thus guiding the pipeline in right direction further. Performing comparison on the results produce by these different algorithms or tools, is a way to find out which one of them is better. This is the objective of ‘compare pipeline’. This pipeline is robust and produces results in a form of plots which are easily understandable.

Overview

As of now compare pipeline, is built to handle GTF/GFF files only but it will be enhance further to compare all different types of NGS data files like BAM, Sam and VCF. Compare, quantitatively express the differences or similarity of 2 GTF files in a final output PDF file which include venn diagrams, barplots and histograms.

Compare, is ready to run perl script. It takes 2 GTF files as an input and give2 PDF files as an output. One PDF is ‘Plots.pdf’ which have venn diagrams, histograms and some bar plots. Other PDF is ‘DetailedPlots.pdf’ which have more detailed multiple barplots. For data extraction and filtering perl hashes and awk utility is used. To find overlaps intersectBed, one of the tools from BEDTools is used, which is set to find overlap of minimum 50% of base pairs of the shortest feature of 2 files. Here feature is referred to genome features(exon, transcript, gene, CDS,UTR etc). A separate R script is written to generate different type of plots and print them all in one PDF file. This R script is called inside compare perl script, thus no need to run Rscript separately.

According to working of BEDTools, whenever there is comparison of two files, file B is loaded into memory and file A is processed line by line and compared with the features of file B. Therefore, to minimize the memory usage, one should set the smaller of the two files as the B file. Also, intersectBed do comparison of file-A over file-B, thus in a manner to find overlap of file-B over file-A one has to run the script one more time reversing the file order.

Execution compare.pl

Usage: $ perl compare.pl [-g1= <fileA>] [-g2=<fileB>]

Option –g1 indicates name of file A

Option –g2 indicates name of file B

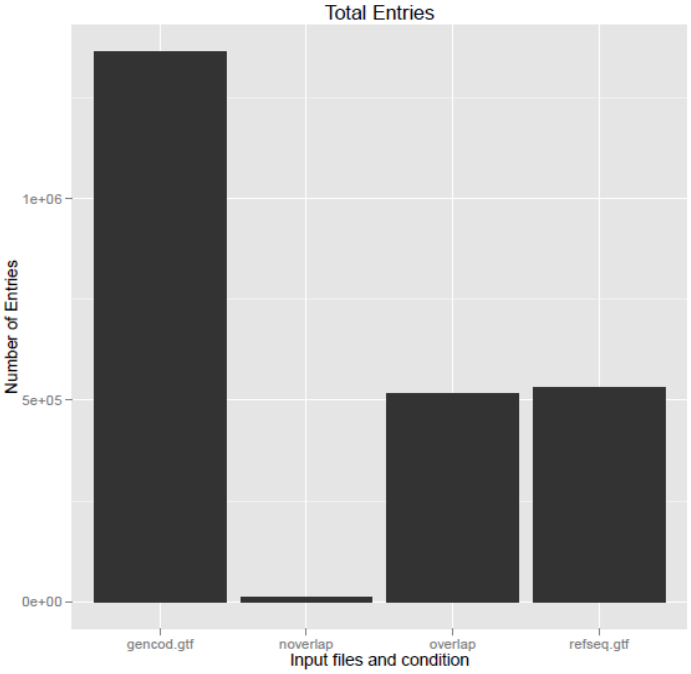
Output: 2 PDF files

Plots.pdf

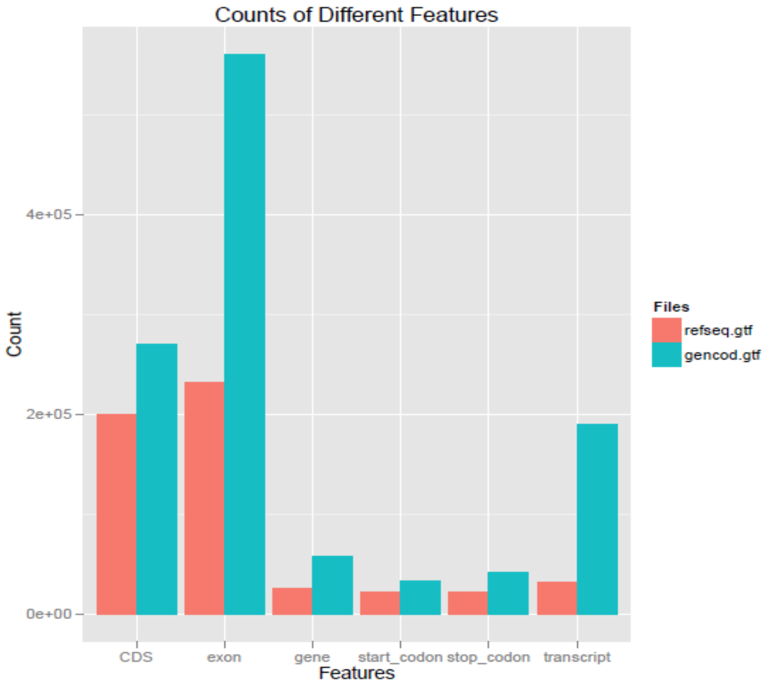
DetailedPlots.pdf

Plots.pdf , will have the following plots:

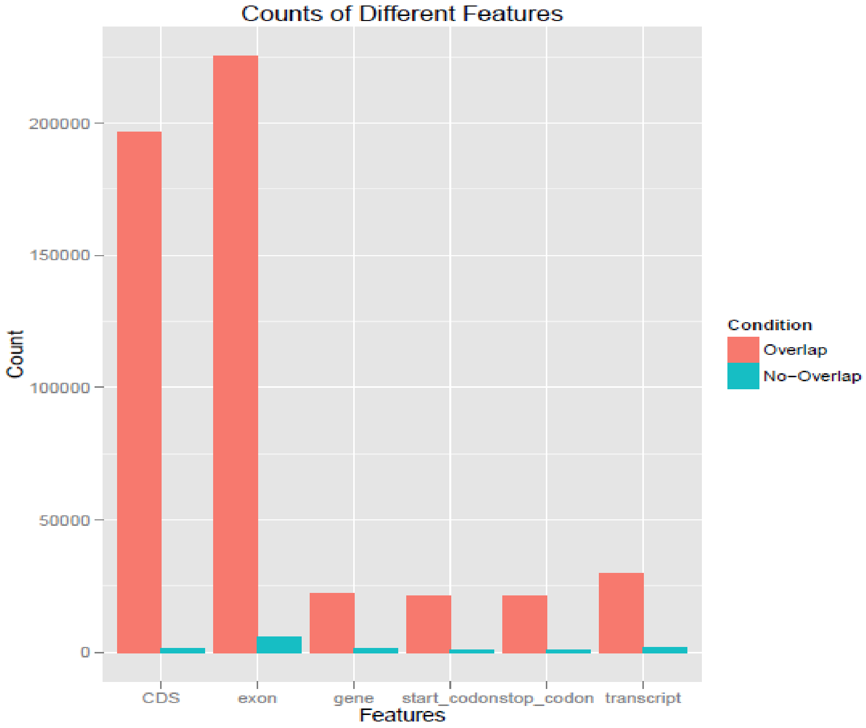
1. Total Entries in Different Files : this bar plot shows the total number of entries in <fileA> <fileB> and number of entries of <fileA> which overlaps <fileB> as “overlap” bar and number of entries of <fileA> which do not overlaps in <fileB> as “non-overlap” bar. Sum of overlap and non-overlap bar is equal to <fileA> bar.



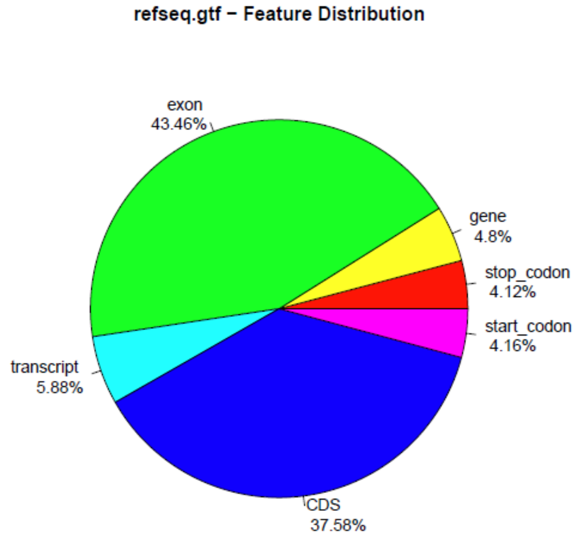
2. Count of Genome Features by <fileA> <fileB>: This bar plot shows count of different genome features by <fileA> <fileB>.



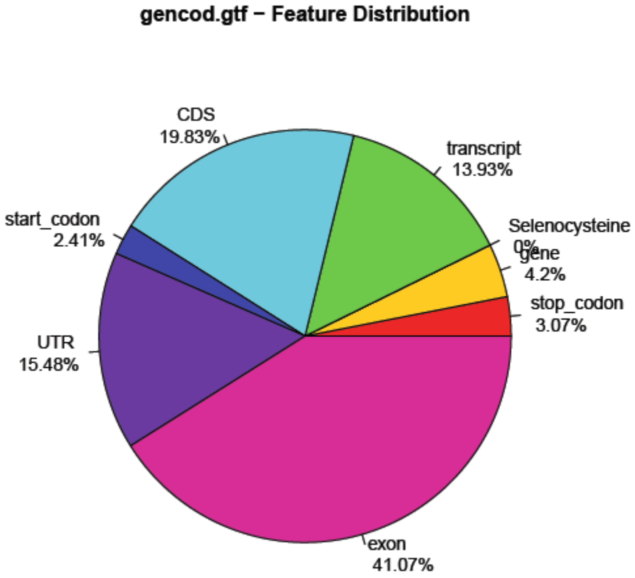
3. Count of Genome Features by Overlap-NonOverlap condition: This bar plot shows count of different genome features by Overlap-NonOverlap condition.



4. <fileA>-feature distribution: this pie chart shows the % distribution of all unique features present in <fileA>. Percentage is rounded to 2 decimal digits.

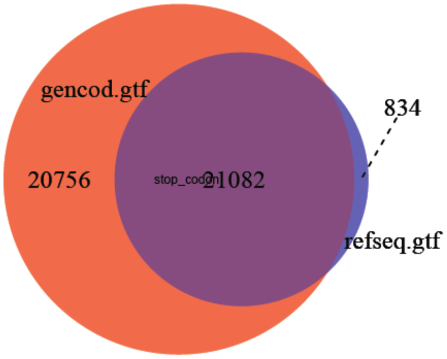


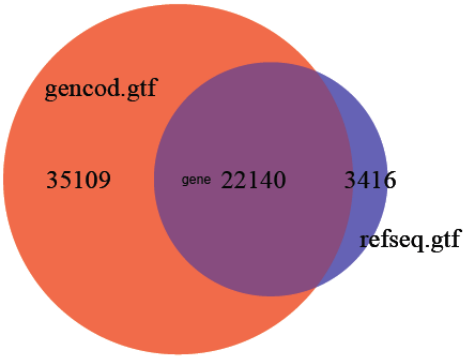
5. <fileB>-feature distribution: this pie chart shows the % distribution of all unique features present in <fileB>. Percentage is rounded to 2 decimal digits.

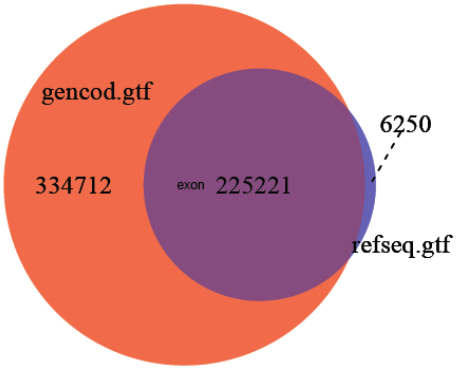


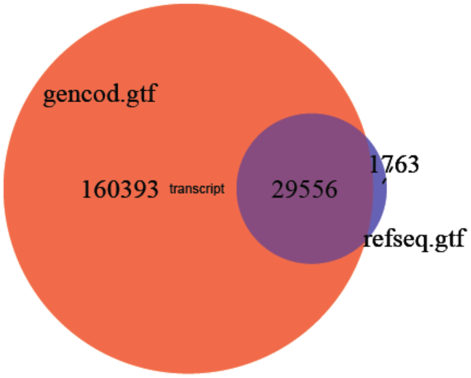
6. Venndiagrams-genome features: Number of venndiagrams will depend on total number of unique genome features that are overlapping. A venndiagram, shows how many entries from

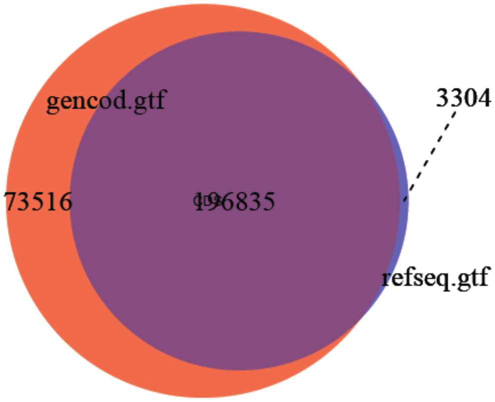
<fileA> overlaps <fileB> for a particular feature. There will be one venndiagram for each genome feature(exon, gene, transcript, stop_codon, start_codon etc).

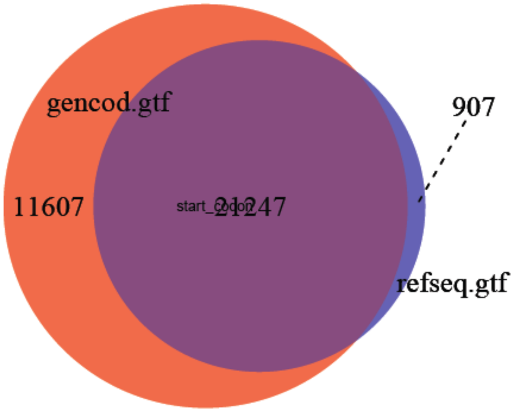




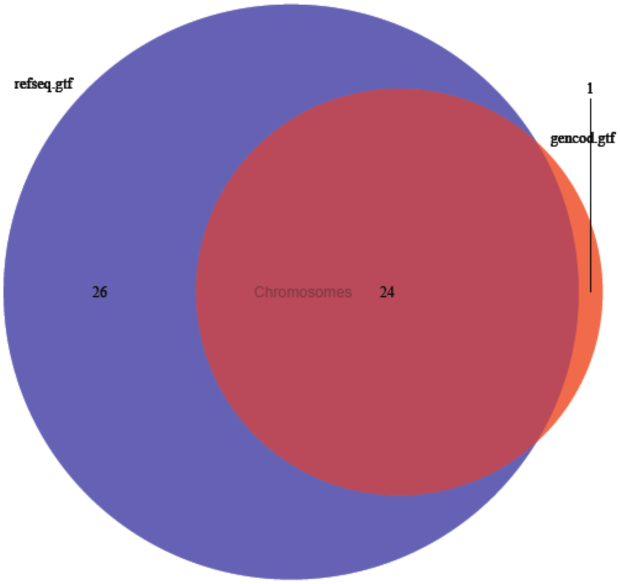




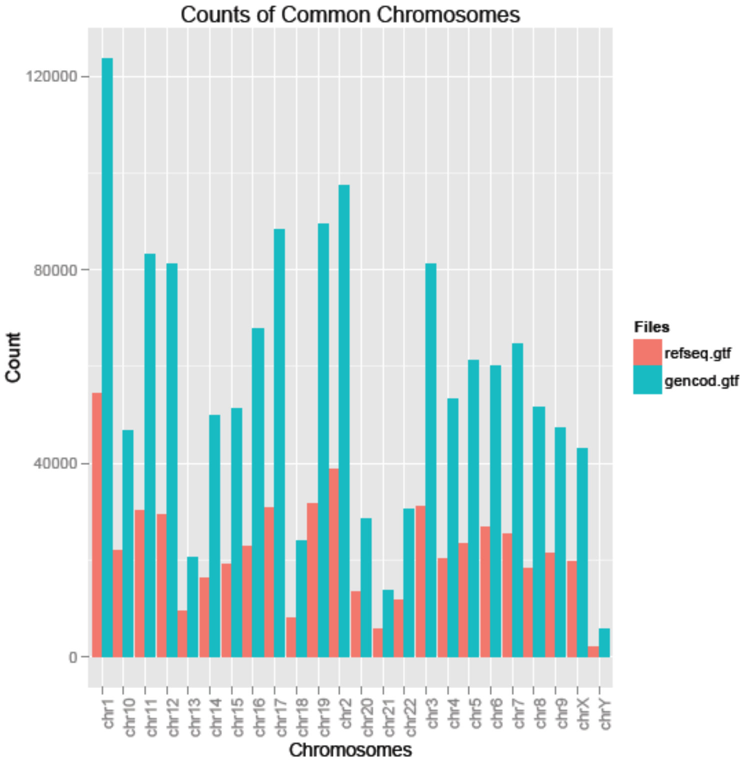




7. Venndiagram-chromosome: This venndiagram shows how many unique chromosomes are present in each <fileA><fileB> and shows how many are common chromosomes.

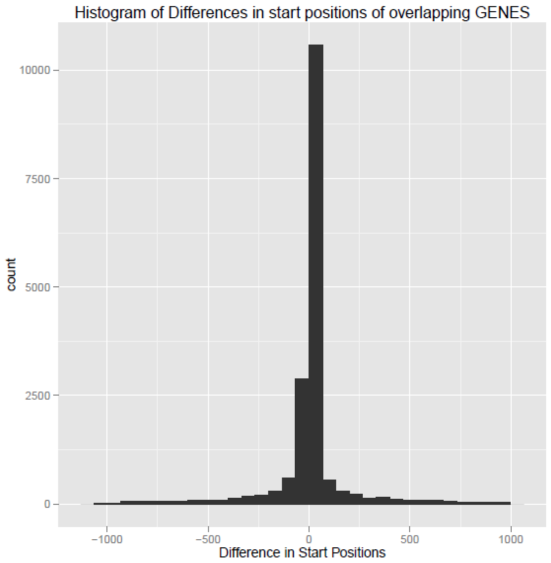


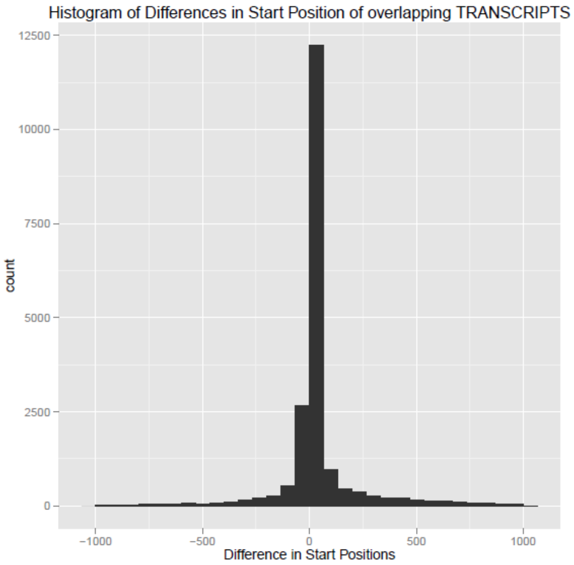
8. Common Chromosomes count: This bar plot shows the count of each common chromosomes found in <fileA> and <fileB>.

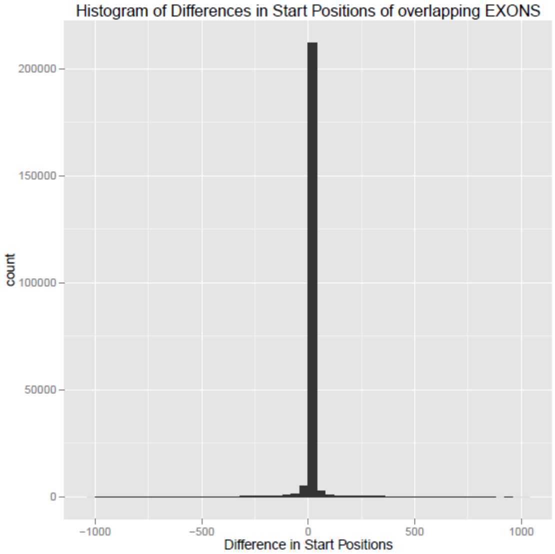


9. Histogram of Differences in bps of Start Positions of Overlapping <Genes/Transcripts/Exons>:

In case of any overlapping genes/transcripts/exons found, this histogram will show difference in bps of start position of a gene/transcript/exon in <fileA> to corresponding start position of overlapped gene/transcript/exon found in <fileB>. Instead of plotting differences in complete range(max to min difference), the range is restricted to -1000 to 1000 base pairs only. Each bar is modified to have upto only 50 bps. Notice difference could be negative as start position of genes/transcripts/exons in <fileB> be greater than start position of overlapped genes/transcripts/exons in <fileB>.



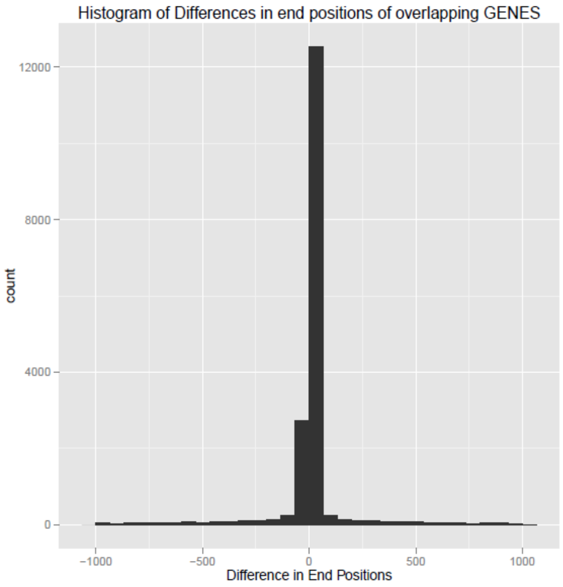


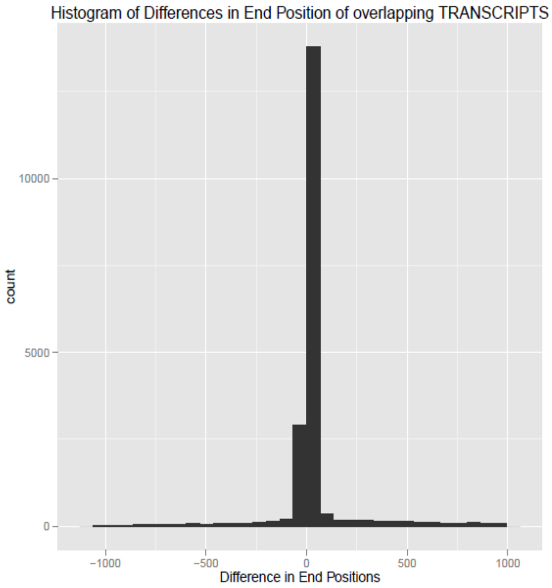


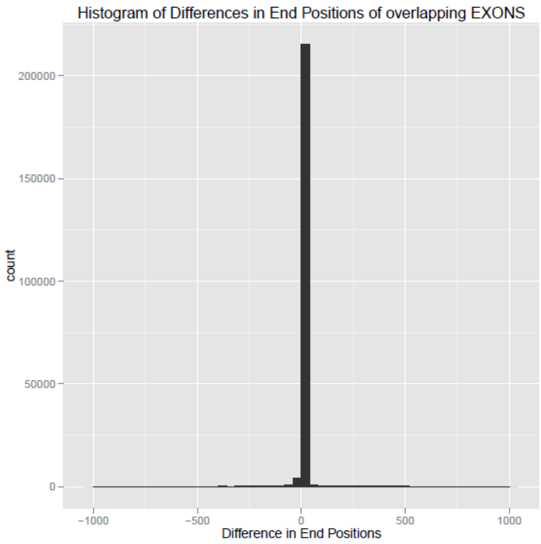
10. Histogram of Differences in bps of End Positions of Overlapping <Genes/Transcripts/Exons>:: In case of any overlapping genes/transcripts/exons found, this histogram will show difference in

bps of end position of a gene/transcript/exon in <fileA> to end position of overlapped gene/transcript/exon found in <fileB>. Instead of plotting differences in complete range(max to min difference), the range is restricted to -1000 to 1000 base pairs only. Each bar is modified to have upto only 50 bps. Notice difference could be negative as start position of gene/transcript/exon in <fileB> be greater than start position of overlapped gene/transcript/exon in <fileB>.

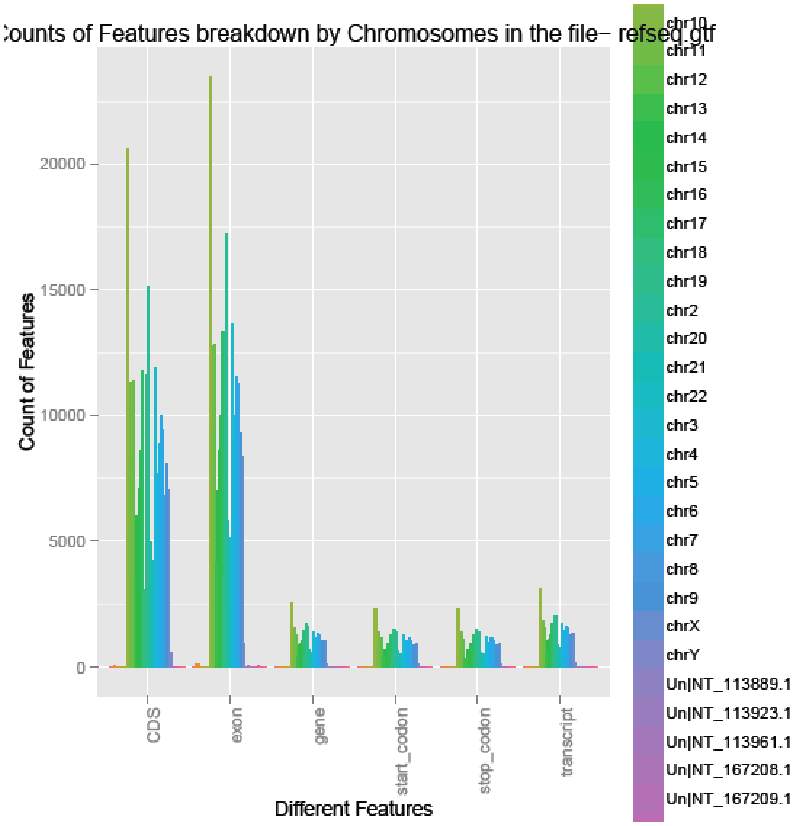
There will a different histogram for different genomic feature(gene/transcript/exon) both start position and end position.





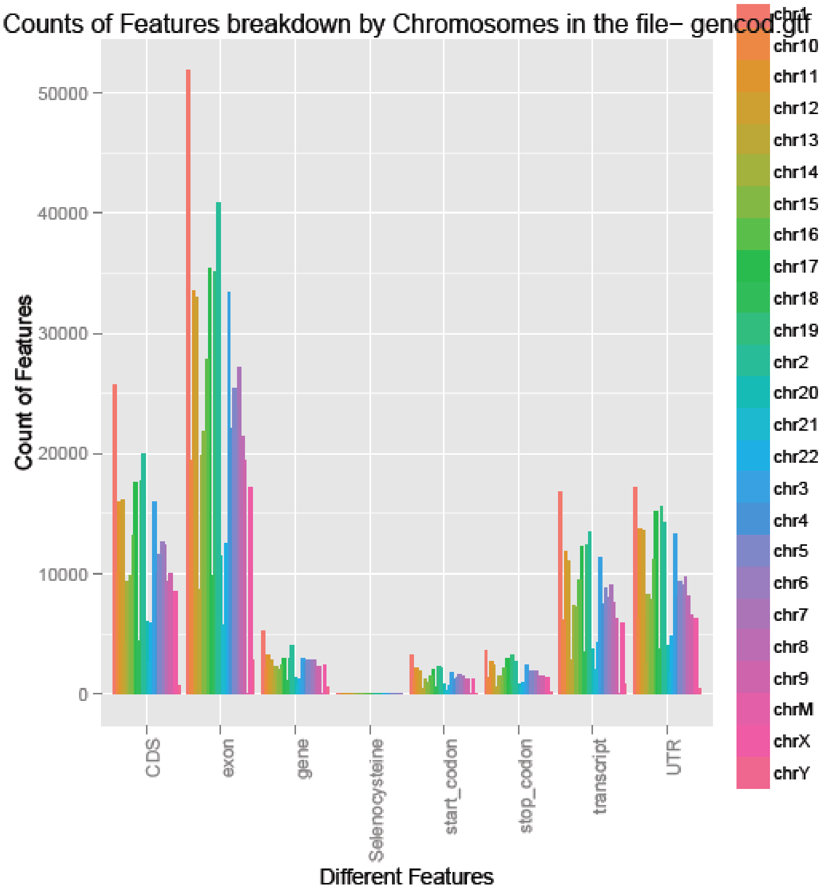


11. Counts of <fileA> genomic features breakdown by Chromosomes: this bar plot shows the counts for all the features breakdown by all the chromosomes found in <fileA>.
    - This plot is breakdown into different number of plots depending on number of genomic feature in <fileA>. And all these plots can be find in “**DetailedPlots.pdf**”.
      For example: If there are 6 unique features in <fileA>, then in “DetailedPlots.pdf” there will 6 different plots each for each feature, showing counts of feature by chromosomes.

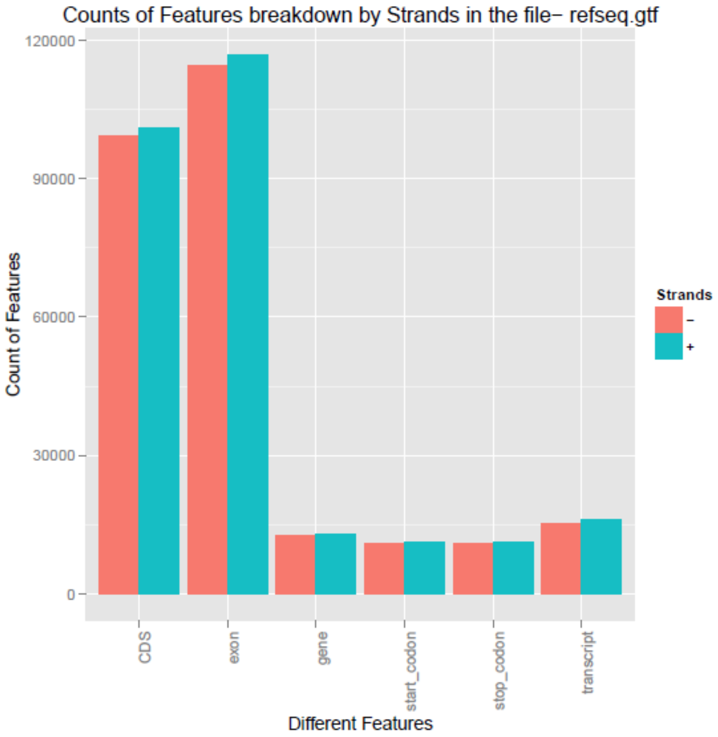


12. Counts of <fileB> genomic features breakdown by Chromosomes: this bar plot shows the counts for all the features breakdown by all the chromosomes found in <fileB>.

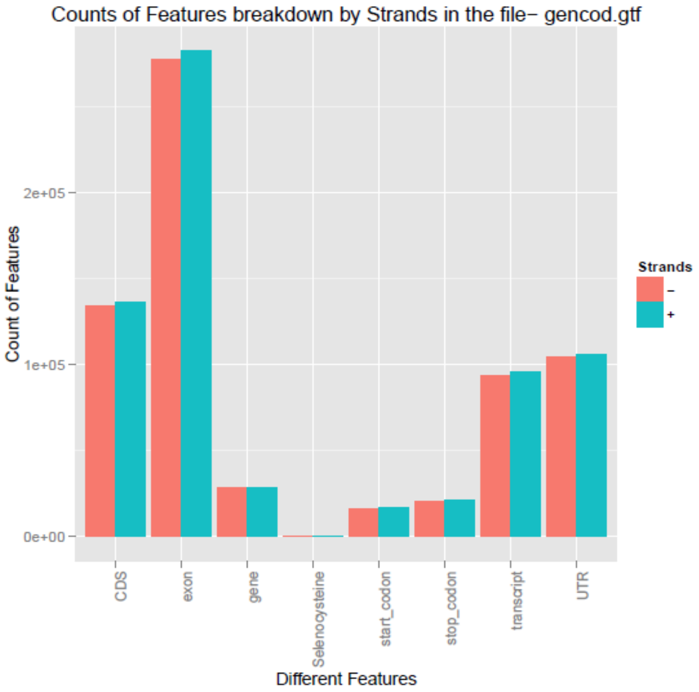
    - This plot is breakdown into different number of plots depending on number of genomic feature in <fileB>. And all these plots can be find in "**DetailedPlots.pdf**".
      For example: If there are 6 unique features in <fileB>, then in "DetailedPlots.pdf" there will 6 different plots each for each feature, showing counts of feature by chromosomes.



13. Counts of <fileA> genomic features by Strands: this bar plot shows the counts for all the features breakdown by all the strands(+,-,.) found in <fileA>.

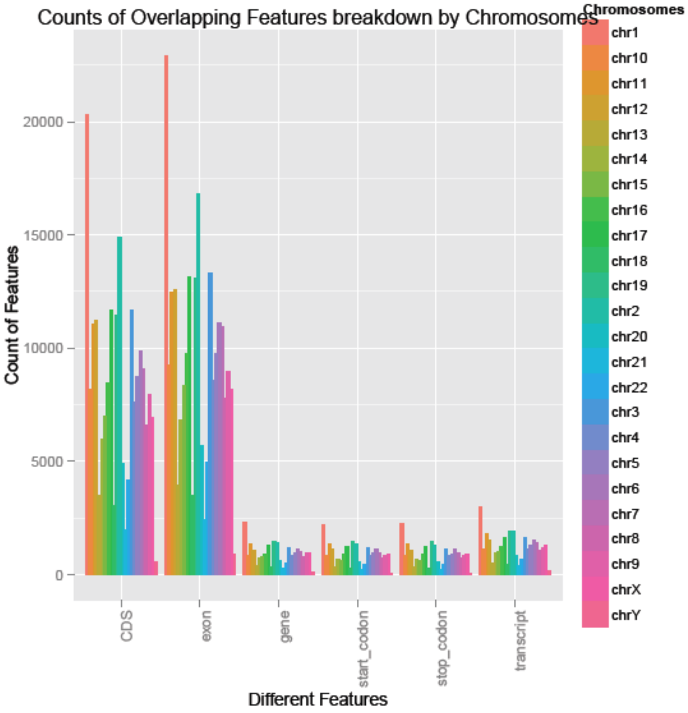


14. Counts of <fileB> genomic features by Strands: this bar plot shows the counts for all the features breakdown by all the strands(+,-,.) found in <fileB>.

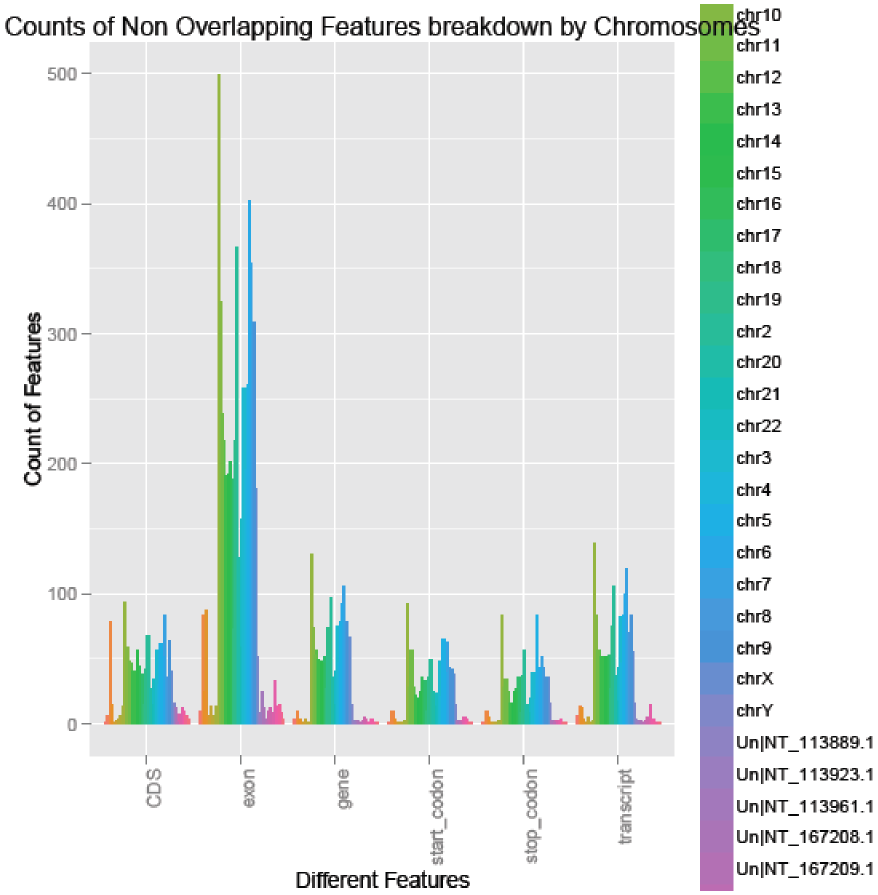


15. Counts of Overlapping genomic features by Chromosomes: this bar plot shows counts of all overlapping features breakdown by all unique common chromosomes.
    - This plot is breakdown into different number of plots depending on number of unique overlapping features found. And all these plots can be find in "**DetailedPlots.pdf**".

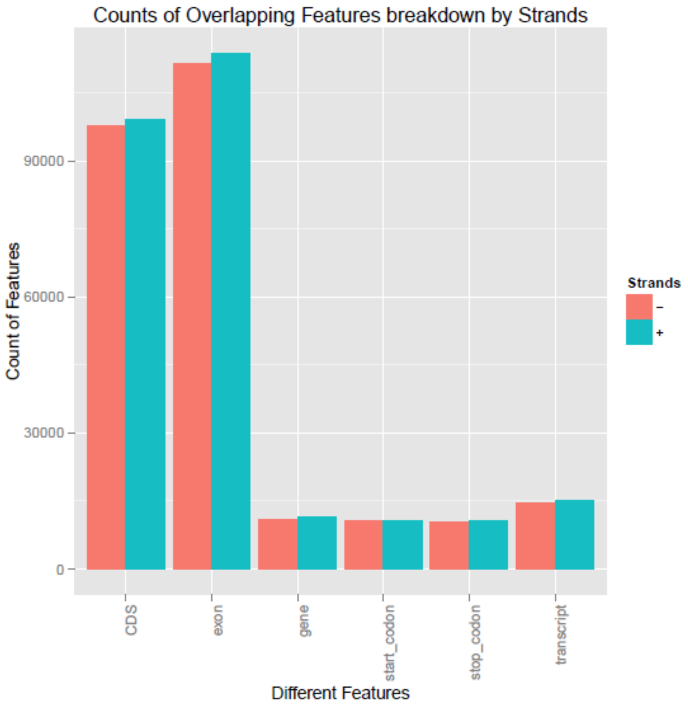
      For example: If there are 6 unique overlapping features, then in "DetailedPlots.pdf" there will 6 different plots each for each feature, showing counts of feature by chromosomes.



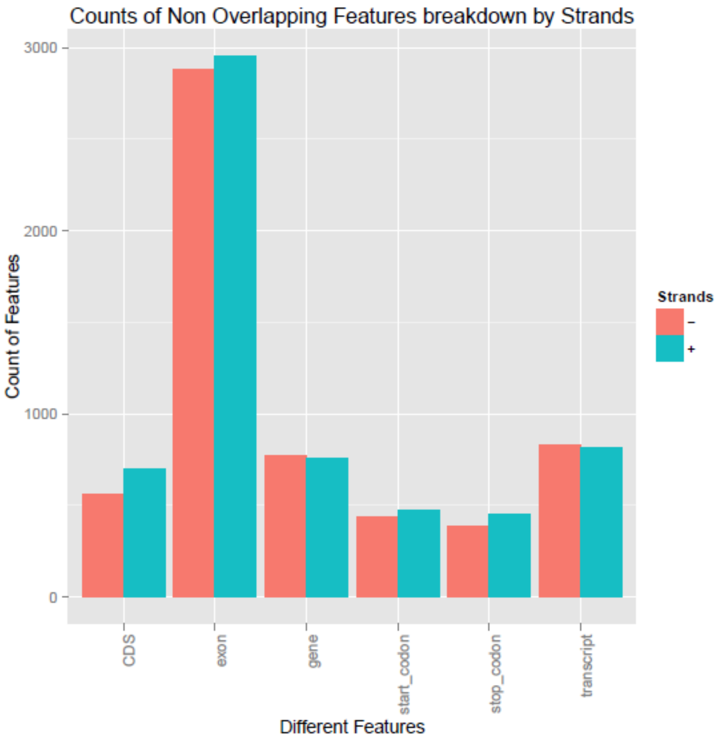
16. Counts of Non-Overlapping genomic features by Chromosomes: this bar plot shows counts for all non-overlapping features breakdown by all unique chromosomes.
    - This plot is breakdown into different number of plots depending on number of unique non-overlapping genomic feature found. And all these plots can be find in "**DetailedPlots.pdf**". For example: If there are 6 unique non-overlapping features, then in "DetailedPlots.pdf" there will 6 different plots each for each feature, showing counts of feature by chromosomes.



17. Counts of Overlapping genomic features by Strands: this bar plot shows counts for all overlapping features breakdown by strands(+,-,.) .



18. Counts of Non-Overlapping genomic features by Strands: this bar plot shows counts for all features breakdown by all the strands(+,-,.).



Supplementary Material:

1. Compare.pl
2. Compare.R
3. Sample- Plots.pdf
4. Sample- DetailePlots.pdf