**Comparison of various tools for RNA sequencing**

There are three approaches applied to obtain transcript quantification: NanoCount

STAR

Kallisto

**NanoCount**:

NanoCount is used for quantification of transcripts based on EM algorithm(Expectation Maximization) of nanopore reads mapped to transcriptome by minimap2.

Input files:

1. This method of RNA sequencing takes into consideration FASTQ file of RNA sequence generated by latest version of guppy-basecaller available at <https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md>

2. Human DNA fasta sequence available in ENSEMBL FTP site

3. GTF files for human annotations in FTP site are used

Tools used:

* Generation of transcriptome fasta using bedparse and bedtools

GTF files from ensembl are converted to BED format using bedparse gtf2bed

Then the BED file is converted to FASTA file using bedtools getfasta and reference human genome sequence, the obtained fasta sequence is transcriptome FASTA

* Mapping transcriptome FASTA sequence to RNA FASTQ from consortium using minimap2

Minimap2 maps the FASTQ file against transcriptome FASTA to obtain the sequences which correspond to coding and non coding transcripts. The output of minimap2 is a SAM file

* Obtaining the transcript expression values using NanoCount

NanoCount is used for estimation of transcript abundance in the SAM file obtained in minimap2 mapping against RNA FASTA sequence

Final Output:

A datasheet which contains transcript names and their corresponding values such as estimated counts,raw and tpm(Transcripts per Million)

**STAR:**

Spliced Transcripts Alignment to a Reference – contains mainly two steps of generating genome indexes and running mapping jobs. This is used for short length fragments where as NanoCount is used for long length fragments.

NOTE: These reads are of different lengths so the reads should be normalised with respect to the library size to make the expression values comparable

Input files:

For Genome indexes:

Human DNA FASTA sequence from ENSEMBL and corresponding GTF file for annotation data

For Mapping jobs:

Paired end FASTQ sequences of polyA-plus RNA with length of minimum 100 and library strand specific (filters to be applied in ENCODE : Homo\_sapiens,GM12878 cell line and GRCh38 assembly,polyA+ RNA)

Output:

Quantmode in mapping job produces ReadsPerGene file which contains tab seperated values of counts corresponding to pair ended FASTQ sequences and unstranded counts.

**KALLISTO:**

There are two steps in this method of RNA sequencing based on pseudoalignment for rapidly determining compatibility of reads with targets without alignment. This is fastest method of quantification of transcripts compared to others.

The two steps are kallisto index and kallisto quant.

Input files:

The kallisto index uses reference transcriptome FASTA(obtained in Nanopore sequencing)

kallisto quant uses index generated and paired end FASTQ sequences from ENCODE which produces output folder

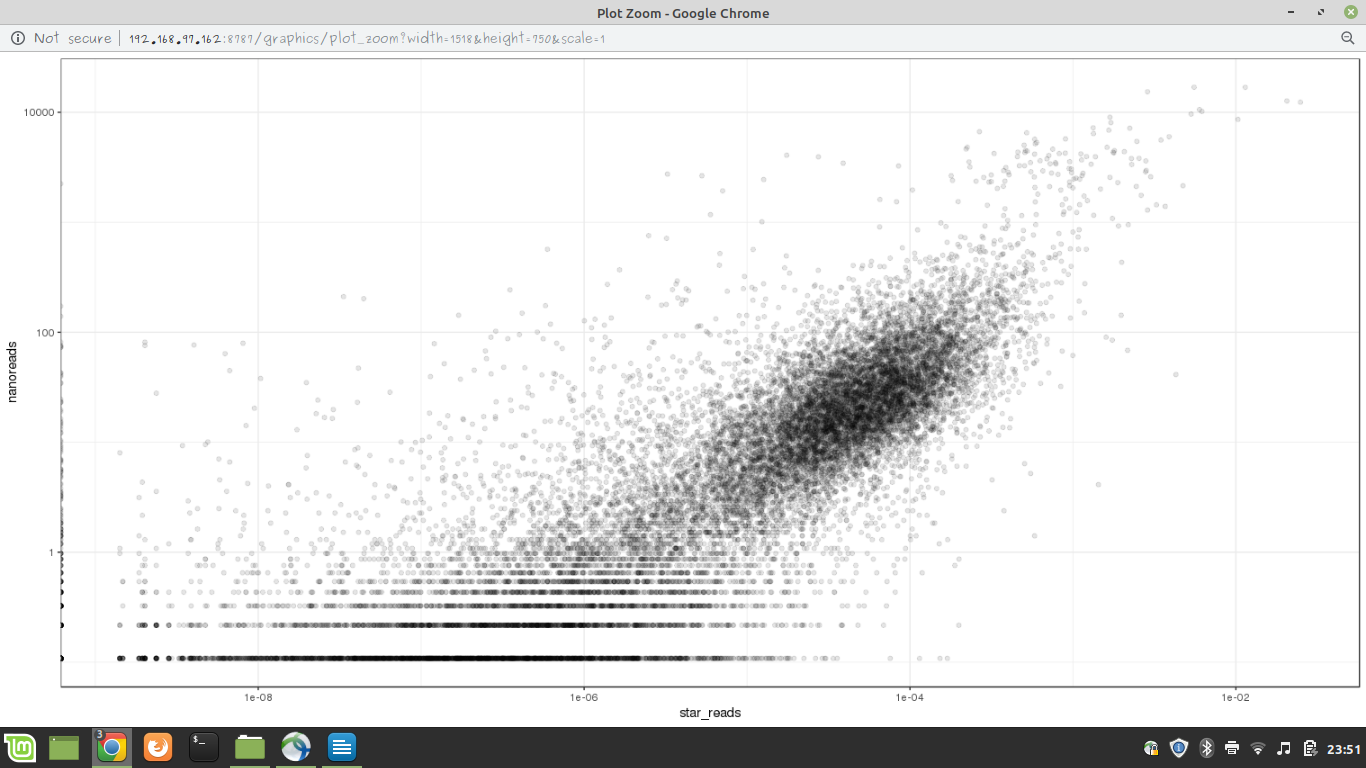
Output:

kallisto quant produces output folder which contains three files abundance.tsv,abundance.h5,run\_info.json

The tsv file contains three columns of est\_counts,tpm,effective length

Pearson correlation coefficient is a measure of linear relationship between two variables where as Spearman correlation coefficient is a measure for monotonic relationship between two variables that is a group of variables tend to change together either increase or decrease but not necessarily at a constant rate. These two coefficients are calculated to identify the relationship between expression values of transcripts obtained using different sequencing tools.

**Plot between nanocount and star:**



**OBSERVATION:**

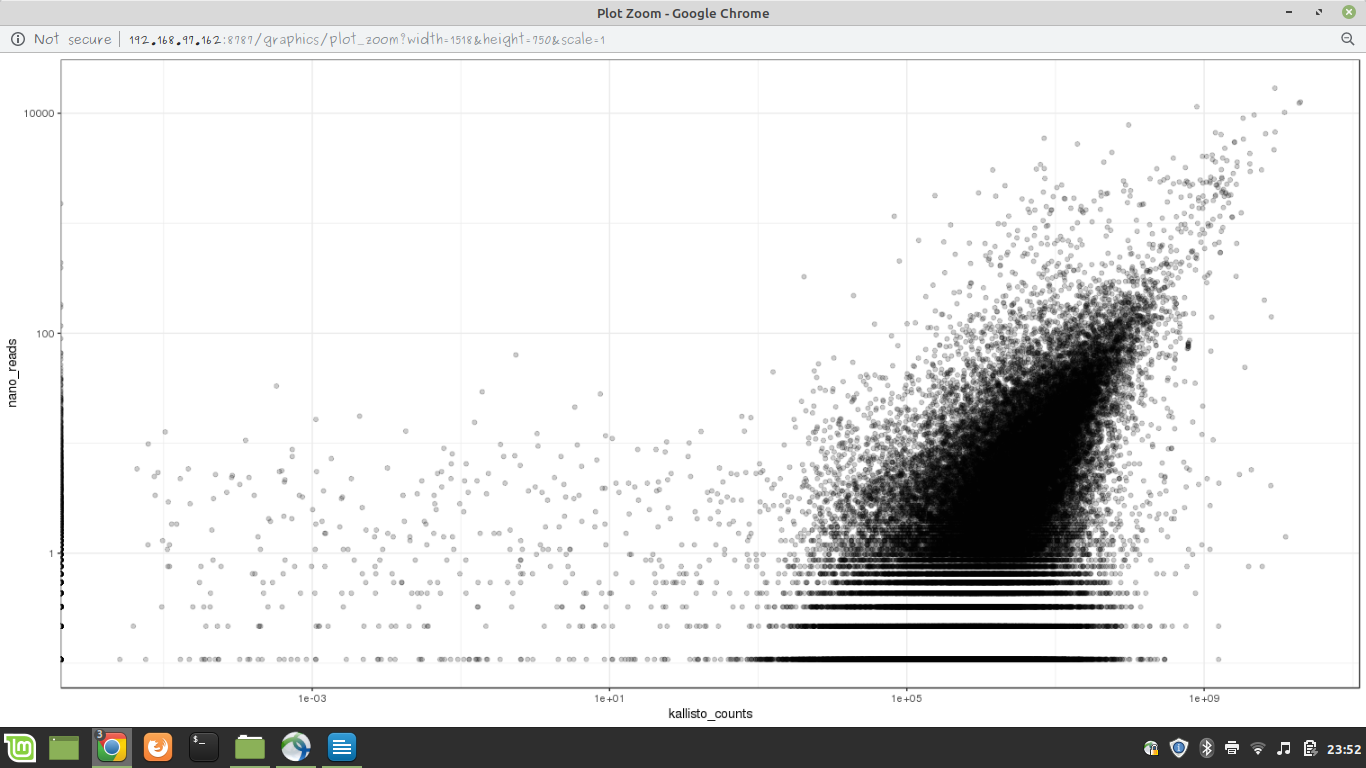
Spearman Correlation coefficient = 0.8439805

Pearson Correlation coefficient = 0.6822731

More than linear relationship the values are monotonically related which can be inferred from higher Spearman compared to Pearson.

R square – pearson coefficient square

**Plot between nanocount and kallisto(transcript\_quantification):**



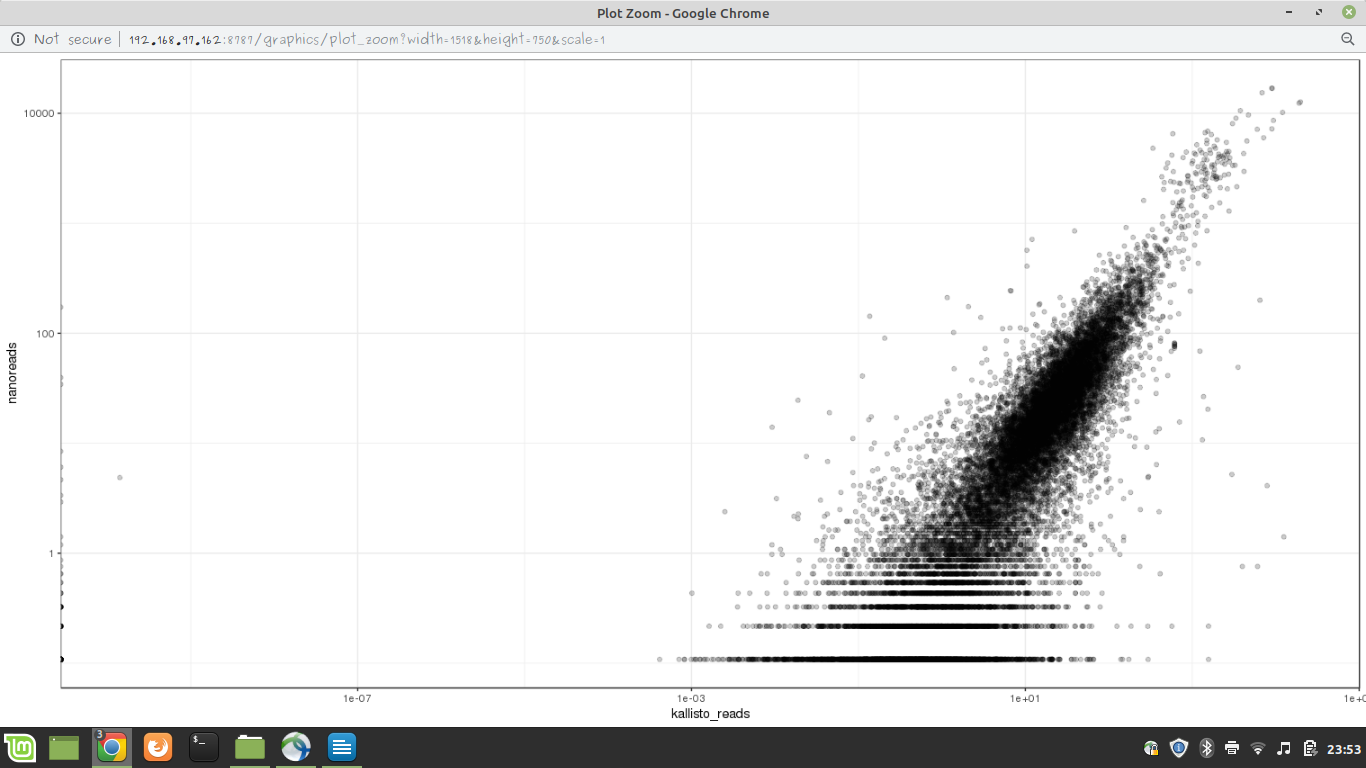
OBSERVATION:

Spearman Correlation coefficient = 0.5104981

Pearson Correlation coefficient = 0.6589651

In this case the linear relationship is more higher compared to monotonic relation among the expression values. The plot contains a dense region which shows values are more closely related from nanocount and KALLISTO

**Plot between nanocount and kallisto(gene quantification):**



OBSERVATION:

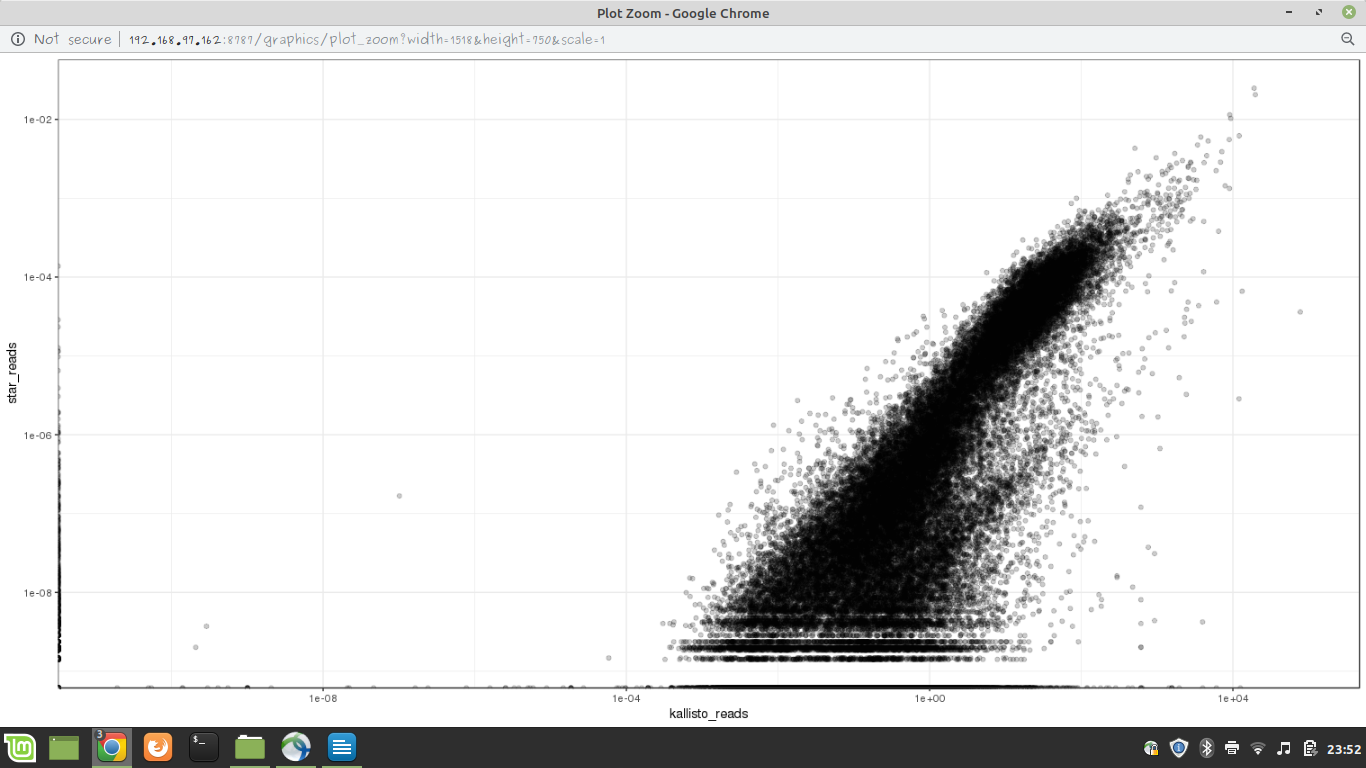
Spearman Correlation coefficient = 0.8680328

Pearson Correlation coefficient = 0.7690679

This plot is comparison of expression values of genes which have transcripts expressed. All the expression values of transcripts of one particular gene are added up to be the expression value of that particular gene.

Plot shows that expression values of genes are more closely related in both the sequencing methods compared to transcript expression. This explains there is a higher variance of expression values at transcript level compared to gene level

**Plot between STAR and kallisto:**



**OBSERVATION:**

Spearman Correlation coefficient = 0.7768924

Pearson Correlation coefficient = 0.4160974

The linear relationship is low but values have higher monotonic relationship that is they changed together. In this case of comparison we are plotting the expression values of genes not transcripts as STAR data contains expression values of genes where as KALLISTO gives transcript quantification. So gene aggregation is done on transcripts and gene expression values are compared with STAR.