Comparison of sequencing performance across different Illumina sequencers

Hemant Kelkar, Bioinformatics and Analytics Research Collaborative

Illumina has introduced a range of sequencer types over the years. There were changes in color chemistry (going from 4- to 2- to 1-color). Yield of the data also increased in leaps as new sequencer types came out.

One of the frequently asked questions HTSF gets is equivalence of sequencing data generated by different sequencers.

We were lucky to have a pool of well characterized mouse RNAseq libraries (courtesy of Dr. Charles Perou, Lineberger Cancer Center at UNC). These libraries originally ran HiSeq 2500. Over the years we have also run them on HiSeq 4000, NovaSeq 6000 and NextSeq 2000 sequencers. As a result, we have a unique source of data that allowed us to directly compare the sequence quality across different sequencer types.

We used a couple of different methods to do these comparisons.

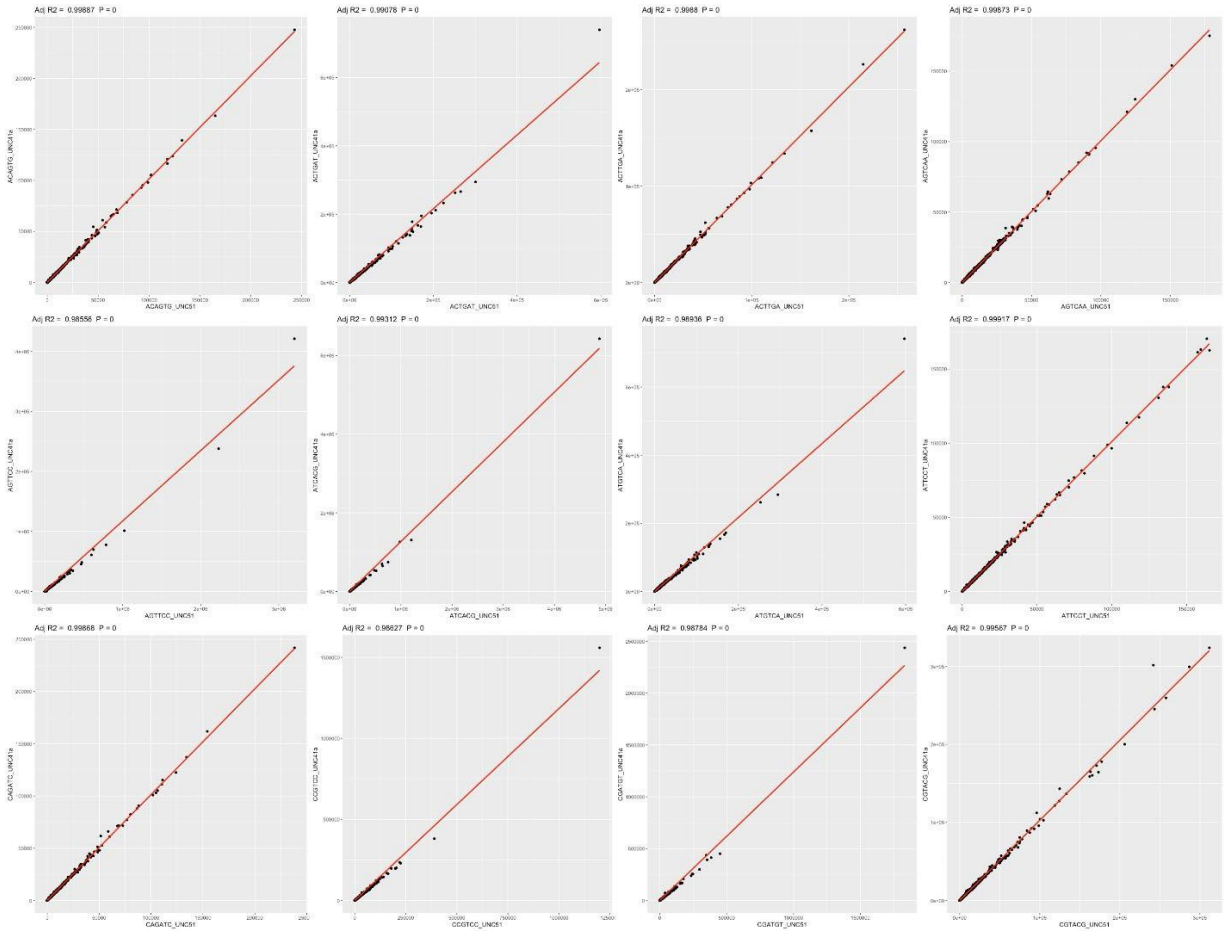
Method 1: Traditional alignments with BBMap (mouse genome)

BBMap aligner is part of the BBTools package (https://sourceforge.net/projects/bbmap/) and is a splice aware aligner. We used v. 38.87. We chose to align reads that multi-mapped to a random location amongst all locations where the read could map. Rest of the parameters were left default. We aligned the sequence data from various sequencers against GRCm38 genome.

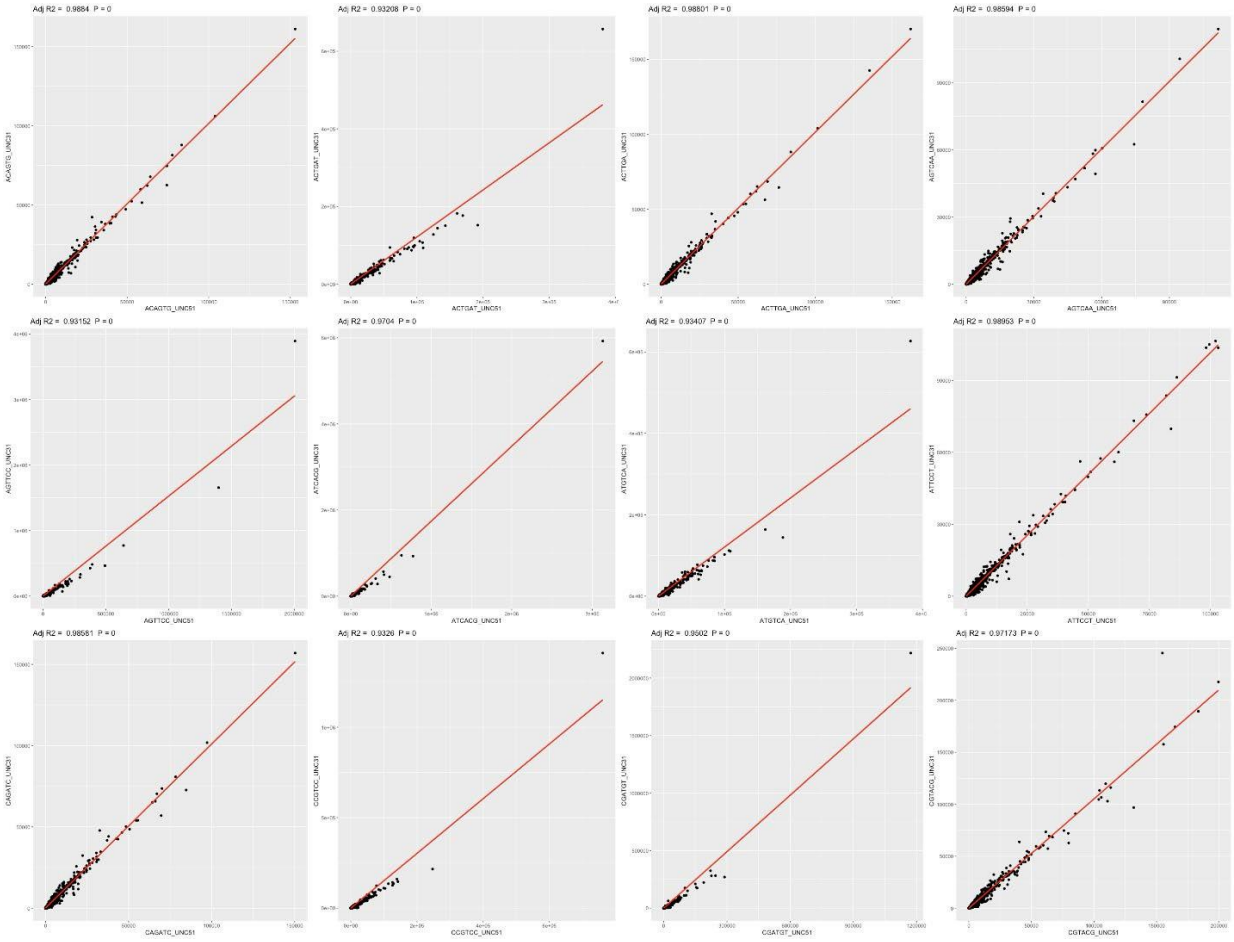
Alignment was then followed by "featureCounts" program to generate raw gene-level counts matrix (genes in rows, samples in columns). featureCounts program is part of the "Subread" package. We used
v.2.0.3 of the subread package.

The count matrix was then imported into DESeq2 and separated into sequencer specific objects for downstream comparisons. Data was normalized using DESeq2's and then plotted using regression function and ggplot package.

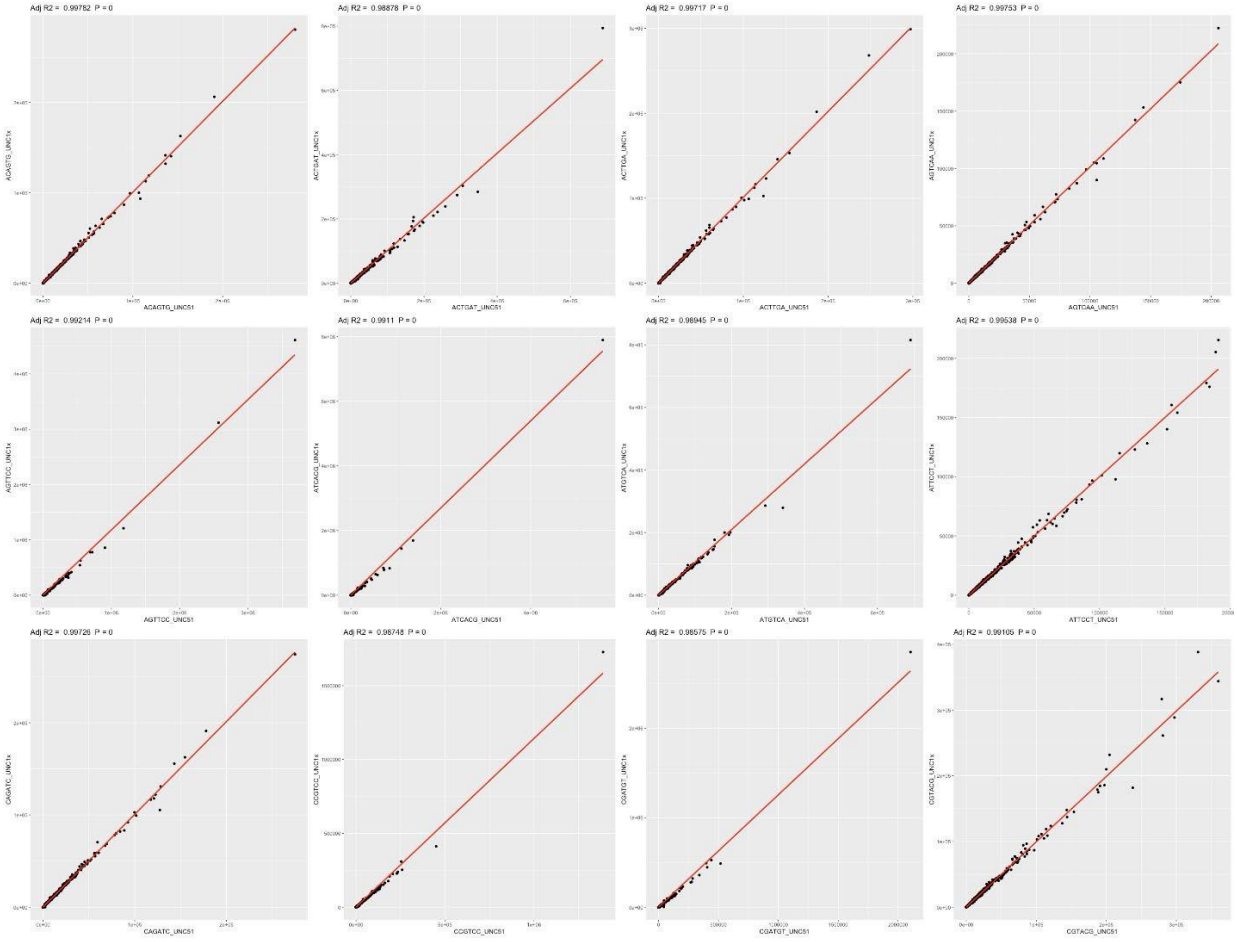
Comparison of NextSeq 2000 (X axis) with NovaSeq 6000 (Y axis)



Comparison of NextSeq 2000 (X-axis) with HiSeq 4000 (Y-axis)



Comparison of NextSeq 2000 (X-axis) with HiSeq 2500 (Y-axis)

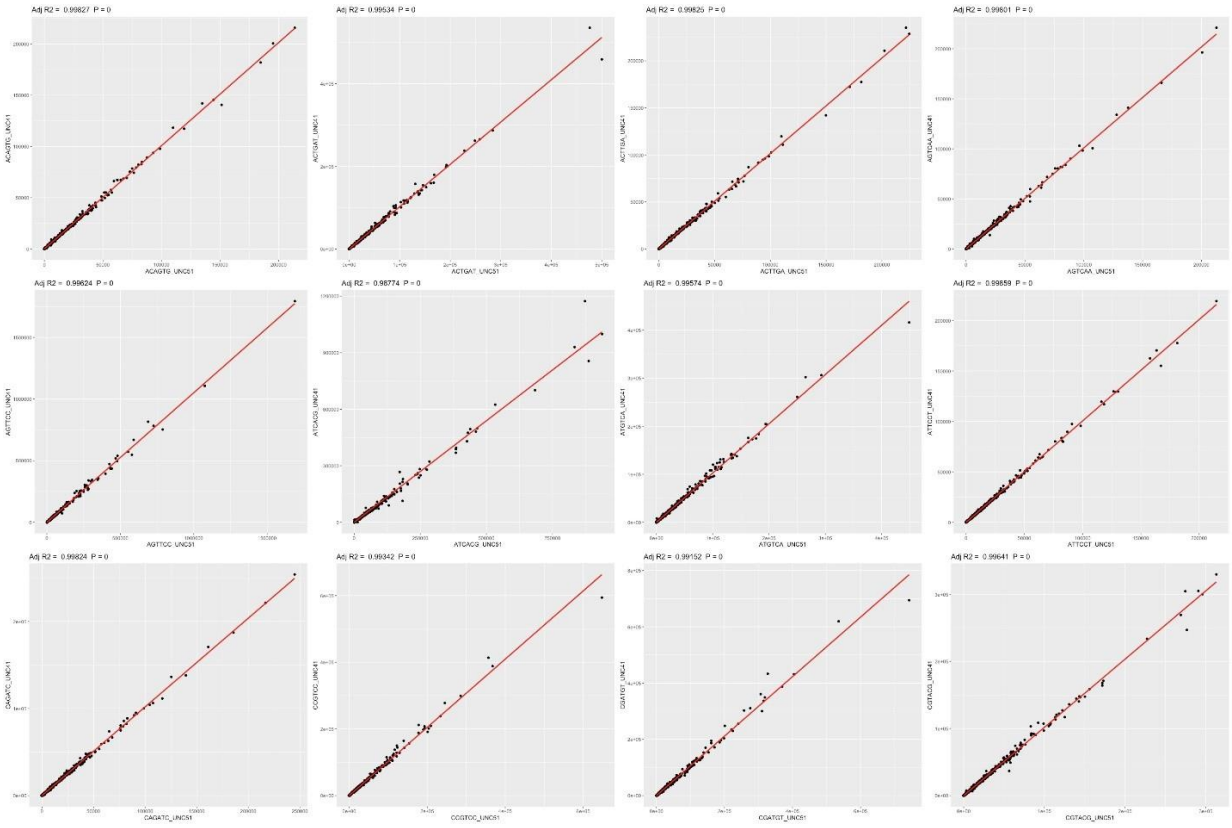


Method 2: Selective alignment using Salmon (mouse transcriptome)

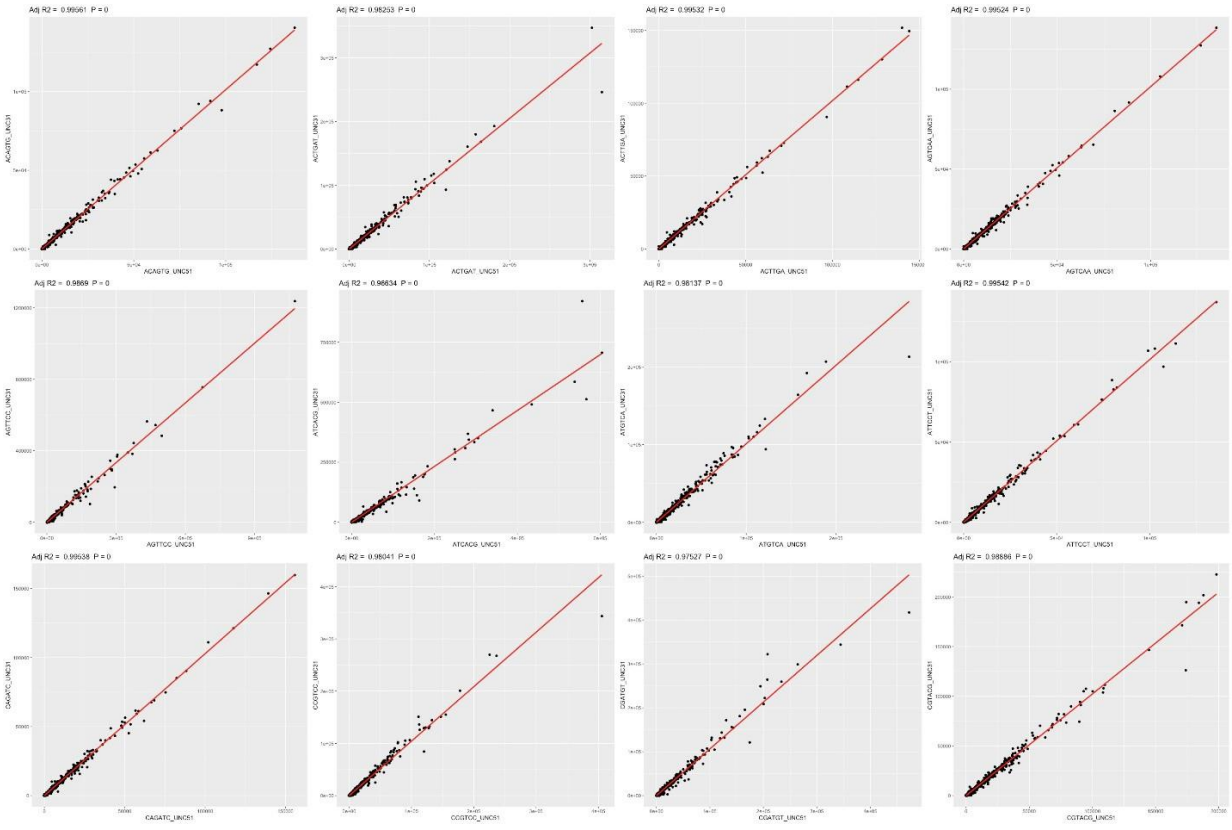
"salmon" is a popular tool that makes use of a transcriptome and a selective alignment algorithm to achieve rapid quantitation of RNAseq data (https://salmon.readthedocs.io/en/latest/).

We used salmon v. 1.6.0 with mouse transcriptome reference (GRCm38/mm10) with decoy genome as recommended by authors of salmon. Pre-built salmon indexes for GRCm38 were downloaded from http://refgenomes.databio.org/. Salmon results are stored in individual quant.sf files for each sample. These were then imported into Rstudio using tximport package from Bioconductor (v. 3.14) and converted into 4 separate objects. These were analyzed with DESeq2. Normalized count data was plotted using a regression function with ggplot.

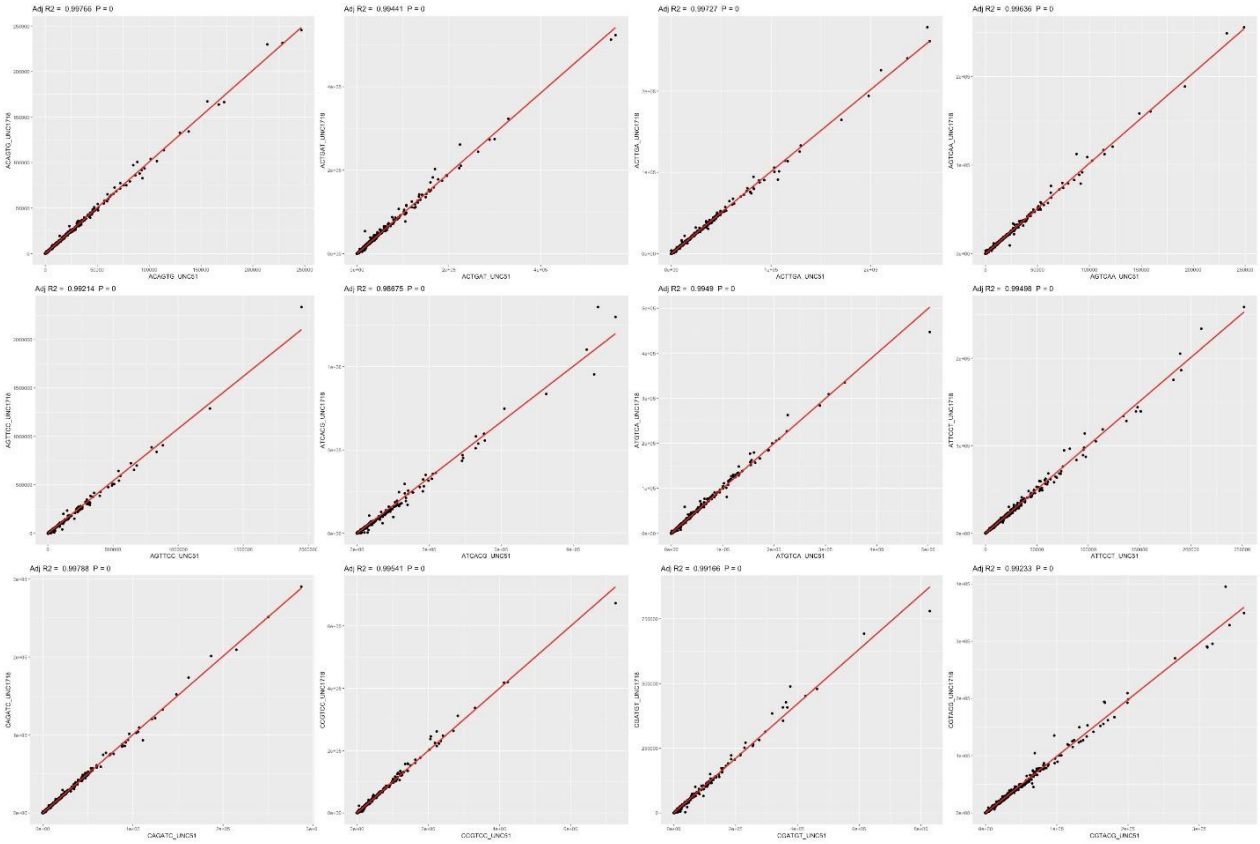
Comparison of NextSeq 2000 (X axis) with NovaSeq 6000 (Y axis)



Comparison of NextSeq 2000 (X-axis) with HiSeq 4000 (Y-axis)



Comparison of NextSeq 2000 (X-axis) to HiSeq 2500 (Y-axis)



Conclusion

We observed excellent correlation between sequence of samples (measured as normalized gene counts) across different types of sequencers using two different RNAseq data analysis methods.

Sequencing samples on a newer generation sequencer should provide the best value for your money while providing quality sequence data.