

OPEN

# QuantSeq. 3' Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis

Susan M. Corley<sup>1\*</sup>, Niamh M. Troy<sup>2</sup>, Anthony Bosco<sup>2</sup> & Marc R. Wilkins<sup>1</sup>

RNA-Seq is increasingly used for the diagnosis of patients, targeting of therapies and for single cell transcriptomics. These applications require cost effective, fast and reliable ways of capturing and analyzing gene expression data. Here we compared Lexogen's QuantSeq which captures only the 3' end of RNA transcripts and Illumina's TruSeq, using both Tophat2 and Salmon for gene quantification. We also compared these results to microarray. This analysis was performed on peripheral blood mononuclear cells stimulated with Poly (I:C), a viral mimic that induces innate antiviral responses. This provides a well-established model to determine if RNA-Seq and QuantSeq identify the same biological signatures. Gene expression levels in QuantSeq and RNA-Seq were strongly correlated (Spearman's rho ~0.8), Salmon and Tophat2 (Spearman's rho > 0.9). There was high consistency in protein coding genes, non-concordant genes had a high proportion of shorter, non-coding features. RNA-Seq identified more differentially expressed genes than QuantSeq, both methods outperformed microarray. The same key biological signals emerged in each of these approaches. We conclude that QuantSeq, coupled with a fast quantification method such as Salmon, should provide a viable alternative to traditional RNA-Seq in many applications and may be of particular value in the study of the 3'UTR region of mRNA.

Changes in gene expression reflect changes in cellular function and behaviour, in development and in disease states. Over the past two decades, methods of measuring gene expression have improved dramatically with a plethora of hybridization arrays available followed by RNA-Seq, the sequencing of short or long RNA reads using massively parallel sequencing technology. There is now a wealth of scientific literature detailing new analysis methods and describing the applications of these approaches<sup>1-4</sup>. Transcriptomics research has matured to the extent that its benefits for clinical diagnosis and treatment of patients can be explored<sup>5</sup>. Already, research is being done to utilize RNA-Seq in a clinical setting to detect early stage head and neck cancer<sup>6</sup>, to inform treatment options in breast cancer<sup>7,8</sup>, in predicting adverse immune responses in organ transplant cases<sup>9</sup> and in cancer immunotherapy<sup>10</sup>. The advent of single cell RNA-Seq, involving the simultaneously sequencing of RNA from thousands of individual cells<sup>11</sup>, brings additional new potential for understanding the mechanisms of disease. Moving from a research environment to a clinical environment has many additional challenges. Results need to be obtained quickly, so that patient diagnosis and care is not unreasonably delayed. Sample volume may be high and cost to the health system needs to be minimised. How best to do this are important questions at this time.

QuantSeq. 3' mRNA library preparation predominantly produces fragments for sequencing close to the 3' end of polyadenylated mRNA, generally from the last exon and the 3' untranslated region (3' UTR)<sup>12</sup>. The QuantSeq method uses total RNA as input, there is no prior poly(A) enrichment or rRNA depletion. The QuantSeq Forward kit has an oligo (dT) primer which contains the Illumina-specific Read 2 linker (P7), which is annealed to the 3' end of the mRNA fragment to synthesise the first cDNA strand via reverse transcriptase. The second strand synthesis is commenced by random priming and DNA polymerase extension. The random primer contains the

<sup>1</sup>Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, UNSW Sydney, New South Wales, Australia. <sup>2</sup>Telethon Kids Institute Australia, The University of Western Australia, Perth, Australia. \*email: [s.corley@unsw.edu.au](mailto:s.corley@unsw.edu.au)

Illumina-specific Read 1 linker sequence (P5). Sequencing commences from the Read 1 sequencing primer and goes toward the poly(A) tail with only one fragment produced per transcript. QuantSeq therefore differs from traditional RNA-Seq where multiple reads are generated from any part of the RNA transcript.

As QuantSeq sequences a smaller part of the transcript and produces only one read per transcript, less sequencing should be required than for standard RNA-Seq. The QuantSeq vendor, Lexogen, recommends that 10 M reads per sample are required for QuantSeq mammalian transcriptomics. In addition to requiring less sequencing, the 3' sequencing method was originally conceived as a way of avoiding gene length bias seen in RNA-Seq<sup>13</sup>. Several studies have investigated the effect of sequencing depth on RNA-Seq and whilst higher power is achieved as sequencing depth increases, beyond 20–30 M reads the gain in power is minimal<sup>14,15</sup>. Nevertheless, ideal read depth will depend on the goals of the experiment, for example the GTEx project used 50 M reads per sample to assess allele-specific expression<sup>16</sup>.

Salmon<sup>17</sup> like kallisto<sup>18</sup>, is a new quantification method which avoids direct base-by-base mapping to the genome, such as occurs with Tophat2<sup>19</sup>, the newer version HISAT2<sup>20</sup> or STAR<sup>21</sup>, and instead quantifies transcripts through quasi-mapping and probabilistic modelling of transcript abundance. To commence quantification the user provides a transcripts file (fasta format) as input and reads are assigned to these transcripts. The output of Salmon is abundance per transcript. The R package Tximport<sup>22</sup> can be used and was used here to summarise transcript abundance to the gene level. The motivation for using QuantSeq and Salmon is similar, with both potentially providing a straightforward means of gene expression analysis. This may be of particular benefit in emerging clinical applications as well as for single cell RNA-Seq which by its very nature involves transcript quantification of a very large number of samples, although at lower number of reads per sample.

Here we compare Illumina's TruSeq RNA-Seq and the Lexogen QuantSeq, and also compare the results from use of read mapping by TopHat2 and Salmon. It is of course most important that any new approach gives reliable informative results. We have therefore used a test case in a well-established model involving peripheral blood mononuclear cells (PBMCs) stimulated with Polyinosinic-polycytidylic acid (Poly (I:C)), a potent inducer of anti-viral responses. Poly(I:C) activates the viral RNA-sensing pattern recognition receptors (PRRs) TLR3, RIG-I/MDA5 and PKR. Activation of these PRRs induces a series of intracellular signalling cascades that trigger nuclear translocation of transcription factors NF- $\kappa$ B and IRF family, leading to cellular expression of pro-inflammatory and antiviral genes<sup>23,24</sup>. Using this model we expect to see changes in gene expression consistent with the innate anti-viral response.

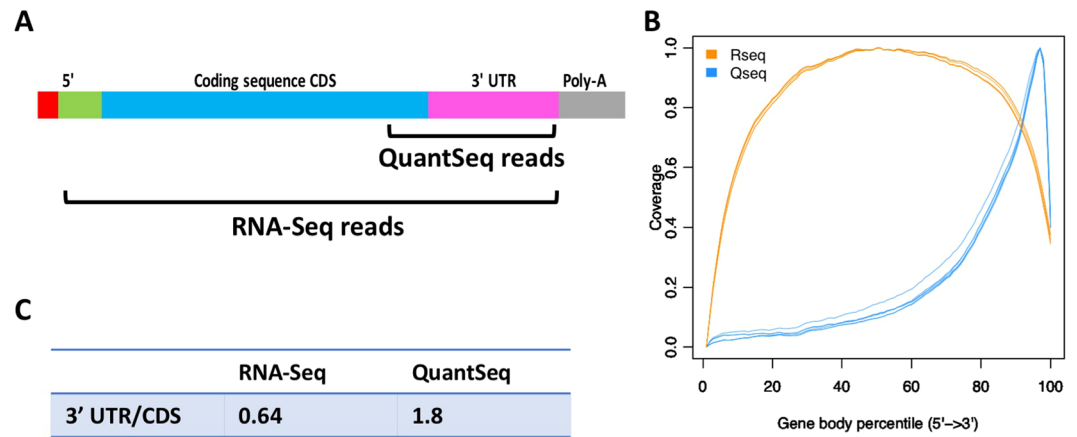
## Results

Our experimental design consisted of six biological replicates of PBMCs, three of which were treated with Poly(I:C) as described further in Methods. RNA libraries were prepared from the six samples using the Lexogen QuantSeq kit or Illumina TruSeq kit (the latter will be referred to as RNA-Seq). In addition, RNA from the six samples were analysed with Clariom S Human microarrays. Following library generation with either QuantSeq or TruSeq kits and 75 bp single-end and paired-end sequencing, respectively, on the Illumina NextSeq500 the sequenced reads were mapped using two different strategies. First, they were mapped to the human genome using Tophat2<sup>19</sup> with reads assigned to features using featureCounts of the Subread package<sup>25</sup>. For brevity, we refer to this pipeline as the Tophat method. As an alternative we performed quantification using Salmon<sup>17</sup>, which does not provide a base-by-base mapping, followed by summarisation to the gene level using the R package, Tximport<sup>22</sup>. We refer to this as the Salmon method. Our next-generation sequencing workflow produced four result sets (i) RNA-Seq mapped with Tophat, (ii) QuantSeq mapped with Tophat (iii) RNA-Seq quantified with Salmon (iv) QuantSeq quantified with Salmon. Our primary data sets consisted of approximately 92 M reads per sample for RNA-Seq and 30 M reads per sample for QuantSeq.

In addition to analysing our full data sets we also conducted analyses on a subset of our data, consisting of the typical number of RNA-Seq reads used in mammalian transcriptomics (30 M reads per sample) and Lexogen's recommended coverage for QuantSeq (10 M reads per sample). Whilst our full data sets might be considered a gold standard, the subsets represent a more typical experimental design. It is appropriate to use less reads for QuantSeq as sequencing is only performed on fragments from the 3' end of the transcript rather than the entire transcript. Herein, lies a major potential advantage of QuantSeq, in that less sequencing reduces the total cost of the experiment. However, our design also enables us to compare RNA-Seq and QuantSeq using the same number of reads in both sets (QuantSeq full data (30 M reads per sample) vs RNA-Seq subset (30 M reads per sample)). We use the following shorthand notation and colour scheme in the Figures throughout this manuscript (RNA-Seq with Tophat ("Rseq", dark orange), QuantSeq with Tophat ("Qseq", dark blue), RNA-Seq with Salmon ("RseqSal", light orange), QuantSeq with Tophat ("QseqSal", light blue), Microarray (green), where we have used a subset of the data we append "SS" to the name).

The Salmon quantification method avoids base-to-base mapping of the entire read and is therefore very fast. The computational time required for the Tophat mapping was around 60 times greater than for Salmon. On our computer cluster Salmon quantification took 10 mins per sample using 6 threads while Tophat took in the order of 10 hours using 6 threads. As Salmon does not perform base-to-base mapping of the entire read it does not produce bam files. The Tophat method produced bam files, each with a size of 6–10 GB. By contrast, Salmon produced an output file of feature abundance (sf file) for each sample which is in the order of 10 MB, that is, approximately 1/100 the size of a bam file.

**QuantSeq reads originate from the 3' end and UTR of RNA transcripts.** We first analysed the location of reads generated by the QuantSeq and RNA-Seq methods. We expected the QuantSeq reads to align to the 3'UTR and the final exon/s of the coding sequence (CDS) whereas RNA-Seq reads cover the entire transcript (Fig. 1A). This was confirmed by calculating the average read coverage across genes using the python script `geneBody_coverage.py` from the RSeQC package<sup>26</sup>. The results showed that the six RNA-Seq libraries (orange lines in



**Figure 1.** QuantSeq reads are predominantly from the 3' end of RNA transcripts. **(A)** Schematic diagram of mRNA indicating the 3' region from which QuantSeq reads are derived and noting that RNA-Seq reads may come from the entire transcript. **(B)** Coverage profile of our RNA-Seq samples (~92 M reads per sample) and QuantSeq samples (~30 M reads per sample) based on the number of reads covering each nucleotide position along the gene body including the 3'UTR and 5'UTR calculated using the `geneBody_coverage.py` of RSeQC. RNA-Seq samples 1–6, orange lines, QuantSeq samples 1–6, blue lines. **(C)** The ratio of reads aligned to 3'UTR versus CDS exons (Tags/Kb) calculated using the `read_distribution.py` script of RSeQC.

Fig. 1B) can be depicted by an inverted u-curve in which coverage plateaus in the middle of the transcript but is lower at both the 3' and 5' ends, whereas the QuantSeq reads peak at the 3' end of the gene (blue lines in Fig. 1B). QuantSeq coverage is higher than RNA-Seq over the last 10% of the gene body including the 3'UTR. There is a great deal of variability in the length of the 3'UTR in human genes (range 60 bp – 4000 bp)<sup>27</sup> and as QuantSeq produces a mean library fragment size between 335–456 bp ([www.lexogen.com](http://www.lexogen.com)), reads will match both the coding sequence (CDS) and the 3'UTR. We expected that the ratio of 3'UTR reads to CDS reads in the QuantSeq data would be higher than in RNA-Seq, where the entire CDS is sampled. Comparing the reads per kilobase derived from the 3'UTR and the reads per kilobase of the CDS we did indeed find a three-fold increase in the 3'UTR/CDS ratio in QuantSeq compared to RNA-Seq (Fig. 1C).

We then compared the allocation of reads to genes in the RNA-Seq and QuantSeq data. We found that a higher percentage of the RNA-Seq reads were uniquely assigned to features compared with the QuantSeq reads (average 78% for RNA-Seq and 64% for Quant-Seq with TopHat; average 88% for RNA-Seq and 79% for QuantSeq with Salmon). The difference between RNA-Seq and QuantSeq was mainly due to the greater number of multimapped reads with QuantSeq (9% average for RNA-Seq versus 18% for QuantSeq). This indicates that a larger proportion of the QuantSeq single end 3' reads can be aligned to multiple genomic regions. A decrease in uniquely mapped reads for QuantSeq has been proposed to arise from sequences found in 3' regions that reduce the number of uniquely assignable reads<sup>12</sup>. Furthermore, our QuantSeq reads were single end reads whereas our RNA-Seq reads were paired end which will also affect the number of multimapped reads<sup>28</sup>. To summarise the mapping results, QuantSeq reads mapped towards the 3' end of CDS and the 3'UTR where coverage was higher than for RNA-Seq but a higher percentage of reads were discarded compared to RNA-Seq.

Most eukaryotic genes harbour multiple polyadenylation sites (PAS) with most being located in the 3'UTR<sup>29</sup>. It is possible that the QuantSeq pipeline may fail to identify alternative polyadenylation isoforms thus affecting the overall gene count and expression level of these genes. To assess this we compared expression distributions for genes with 1, 2, 4, 6, 8 and 10 PAS. This analysis indicates that QuantSeq did not perform differently to RNA-Seq in quantifying genes with multiple PAS. Boxplots of these distributions are Supplementary Fig. S1.

**Protein coding and non-coding genes found by the different quantification methods.** We next investigated the capacity of RNA-Seq and QuantSeq to detect protein coding and non-coding features (such as pseudogenes, antisense, snoRNA, lincRNA), together referred to as “genes”. A gene was “identified” if at least one read from one sample was mapped to that gene. The total number of identified genes are set out in Table 1 and compared by barplot in Fig. 2(A–C). Filtering of lowly-expressed genes was undertaken to arrive at a more robust list of genes we consider to be expressed. We used a typical filtering strategy to remove genes with low counts<sup>30</sup>, excluding genes with less than one count per million (1 CPM) in at least half of the samples. The remaining “expressed genes” were used in downstream differential expression analysis. The total number of expressed genes for each data set can be found in Table 1 and are plotted to provide a comparison between the methods (Fig. 2D–F). We found consistency between RNA-Seq and QuantSeq in the number of protein coding genes before filtering (Fig. 2B) and after filtering (Fig. 2E). There was more variability in the number of non-coding features (Fig. 2C,F). The most striking difference seen from the barplots is that more non-coding genes are included with Salmon quantification (Fig. 2C,F) and that this is more prominent for the QuantSeq data than the RNA-Seq data. Distribution plots (violin plots) of the expression levels of the protein coding and non-coding features showed a broad range of expression values, from very low to very high expression (Fig. 2G–J). The non-coding features

|                                   | Identified Genes <sup>a</sup> | Protein Coding % <sup>b</sup> | Expressed Genes <sup>c</sup> | Protein Coding % <sup>b</sup> | DEGs <sup>d</sup> | Protein Coding % <sup>b</sup> |
|-----------------------------------|-------------------------------|-------------------------------|------------------------------|-------------------------------|-------------------|-------------------------------|
| RNA-Seq (Tophat) <sup>e</sup>     | 23857                         | 0.71                          | 13286                        | 0.92                          | 1597              | 0.92                          |
| QuantSeq (Tophat) <sup>f</sup>    | 21425                         | 0.75                          | 12767                        | 0.92                          | 1129              | 0.92                          |
| RNA-Seq SS (Tophat) <sup>g</sup>  | 22664                         | 0.73                          | 13245                        | 0.92                          | 1247              | 0.94                          |
| QuantSeq SS (Tophat) <sup>h</sup> | 19433                         | 0.78                          | 12706                        | 0.92                          | 958               | 0.93                          |
| RNA-Seq (Salmon) <sup>e</sup>     | 39657                         | 0.46                          | 15396                        | 0.83                          | 1795              | 0.85                          |
| QuantSeq (Salmon) <sup>f</sup>    | 34689                         | 0.50                          | 17026                        | 0.76                          | 1160              | 0.79                          |
| RNA-Seq SS (Salmon) <sup>g</sup>  | 36271                         | 0.49                          | 15354                        | 0.83                          | 1519              | 0.86                          |
| QuantSeq SS (Salmon) <sup>h</sup> | 31676                         | 0.53                          | 16956                        | 0.76                          | 961               | 0.83                          |
| Microarray                        | 16039                         | 0.99                          |                              |                               | 327               |                               |

**Table 1.** Genes expressed in PBMCs as detected by RNA-Seq or QuantSeq and those that are differentially expressed after treatment with polyinosinic-polycytidylic acid. <sup>a</sup>Number of genes with an associated gene symbol identified as having at least one read in at least one sample. <sup>b</sup>The percentage of protein coding genes in the total number from the previous column. <sup>c</sup>After filtering to remove those genes with <1 CPM in at least 3 samples. <sup>d</sup>All differentially expressed genes with FDR < 0.05 and logFC > 1 determined as set out in Methods. <sup>e</sup>All data being an average of 92 M reads per sample for RNA-Seq. <sup>f</sup>All data being an average of 30 M reads per sample for QuantSeq. <sup>g</sup>Subset (SS) of data being an average of 30 M reads per sample for the RNA-Seq dataset. <sup>h</sup>Subset (SS) of data being an average of 10 M reads per sample for the QuantSeq dataset.

were of generally low expression (Fig. 2H,J) and a significant proportion of these features were excluded after filtering (compare Fig. 2G–H,I–J). After filtering to remove lowly expressed genes we found consistent distributions of expression levels within the protein coding and non-coding data sets.

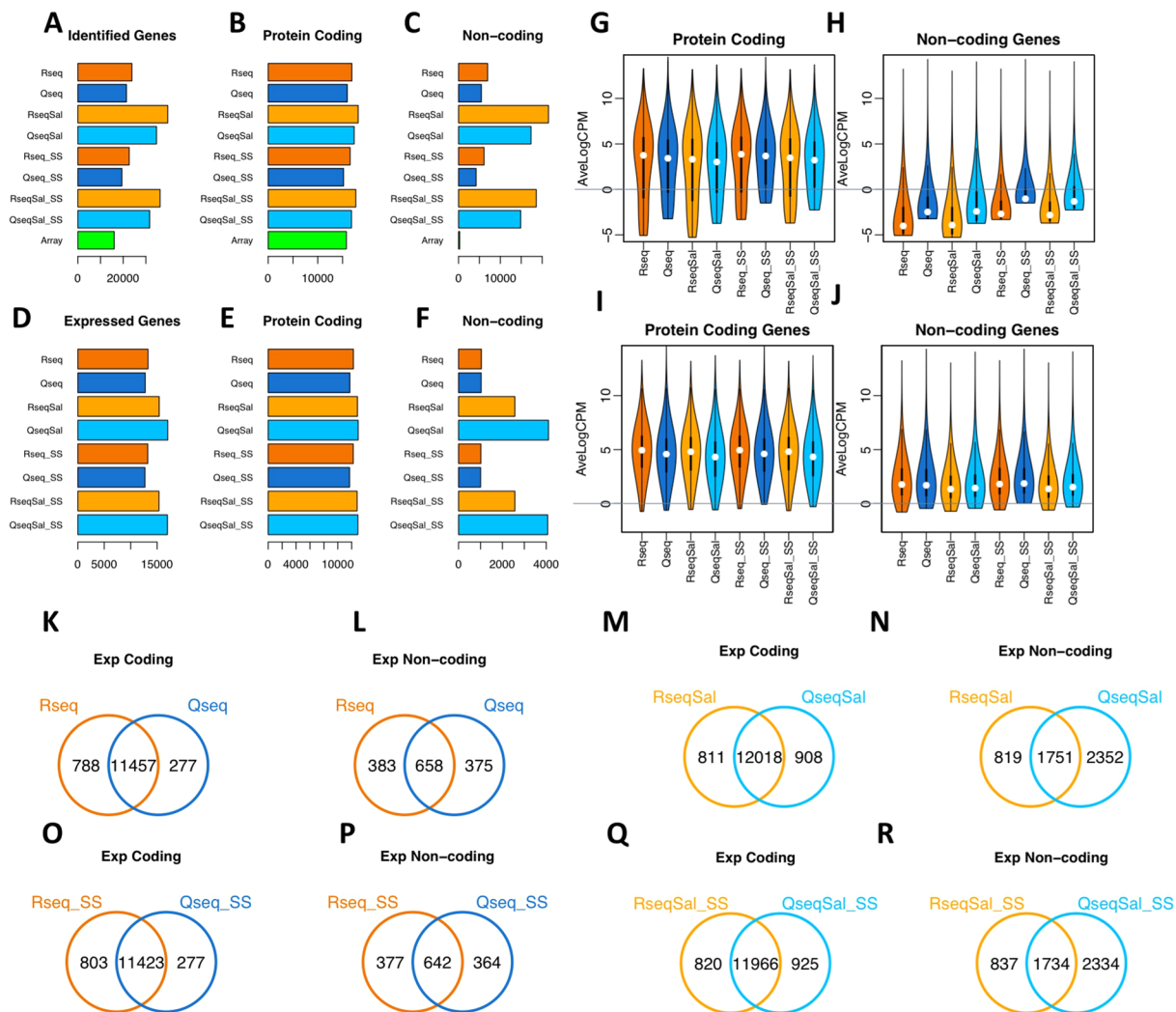
Clariom S arrays used for microarray analysis have probes for well-annotated genes rather than the full transcriptome of coding and non-coding features, with 99% of the detected genes being protein coding. The number of protein coding genes detected by microarray is similar to the number identified in our RNA-Seq and QuantSeq data sets (Fig. 2B). Overall, we see internal consistency in the number of protein coding genes across all the next-generation sequencing methods but variation in the number of non-coding genes with more found when using Salmon.

To compare the expression measures from RNA-Seq, QuantSeq and microarray, we calculated the Spearman's rank correlation co-efficient of the average expression values (AveLogCPM) for all expressed genes (Table 2). The RNA-Seq and QuantSeq data was strongly correlated (Spearman's rho 0.83, Tophat; 0.80, Salmon). We also found very high correlation when comparing the Tophat and Salmon mapping for either the RNA-Seq (Spearman's rho 0.98) or QuantSeq (Spearman's rho 0.93). Correlation between the RNA-Seq and microarray (Spearman's rho 0.79) and QuantSeq and microarray (Spearman's rho 0.7) was reduced but still revealed a high degree of correlation.

To compare the gene sets found by the different approaches, Venn diagrams were made for four pairwise comparisons of the coding and non-coding expressed genes (Fig. 2K–R). We saw a similar high overlap in the protein coding genes when comparing QuantSeq with RNA-Seq irrespective of quantification method, for both the full data set and the subsets (Tophat: Jaccard coef = 0.91, Salmon: Jaccard coef = 0.87). The non-coding genes accounted for a modest proportion of the full complement of expressed genes, 8% of the Tophat expressed genes and between 17–24% of the Salmon expressed genes. These subsets of non-coding genes show greater variability between RNA-Seq and QuantSeq (Tophat: Jaccard coef = 0.46, Salmon: Jaccard coef = 0.36). The most striking difference between the RNA-Seq and QuantSeq expressed genes can be seen in Venn diagrams comparing the non-coding genes found by Salmon (Fig. 2N,R). The QuantSeq expressed genes include 2352 non-coding genes not expressed in the RNA-Seq data. Overall, 72% of the QuantSeq exclusive genes, were non-coding RNA which represents a doubling or tripling in the proportion of non-coding RNA in the set of genes found to be exclusive to one method. Pairwise comparisons of the expressed protein coding and non-coding genes in each of our data sets can be found in Supplementary Table S1.

To summarise, with Tophat mapping the RNA-Seq data sets have a slightly higher number of expressed protein coding genes (~4% higher) than QuantSeq but a very similar number of non-coding genes (~1% higher). Furthermore, QuantSeq recovers around 94% of the protein coding genes and around 60–70% of the non-coding genes found by RNA-Seq. With Salmon quantification we found that RNA-Seq and QuantSeq produce a similar number of expressed protein coding genes (<0.1% difference) and again QuantSeq recovers around 94% of the protein coding genes found by RNA-Seq. However, we see a difference in the non-coding genes, with Salmon reporting a substantial increase in the number of non-coding genes from the QuantSeq data.

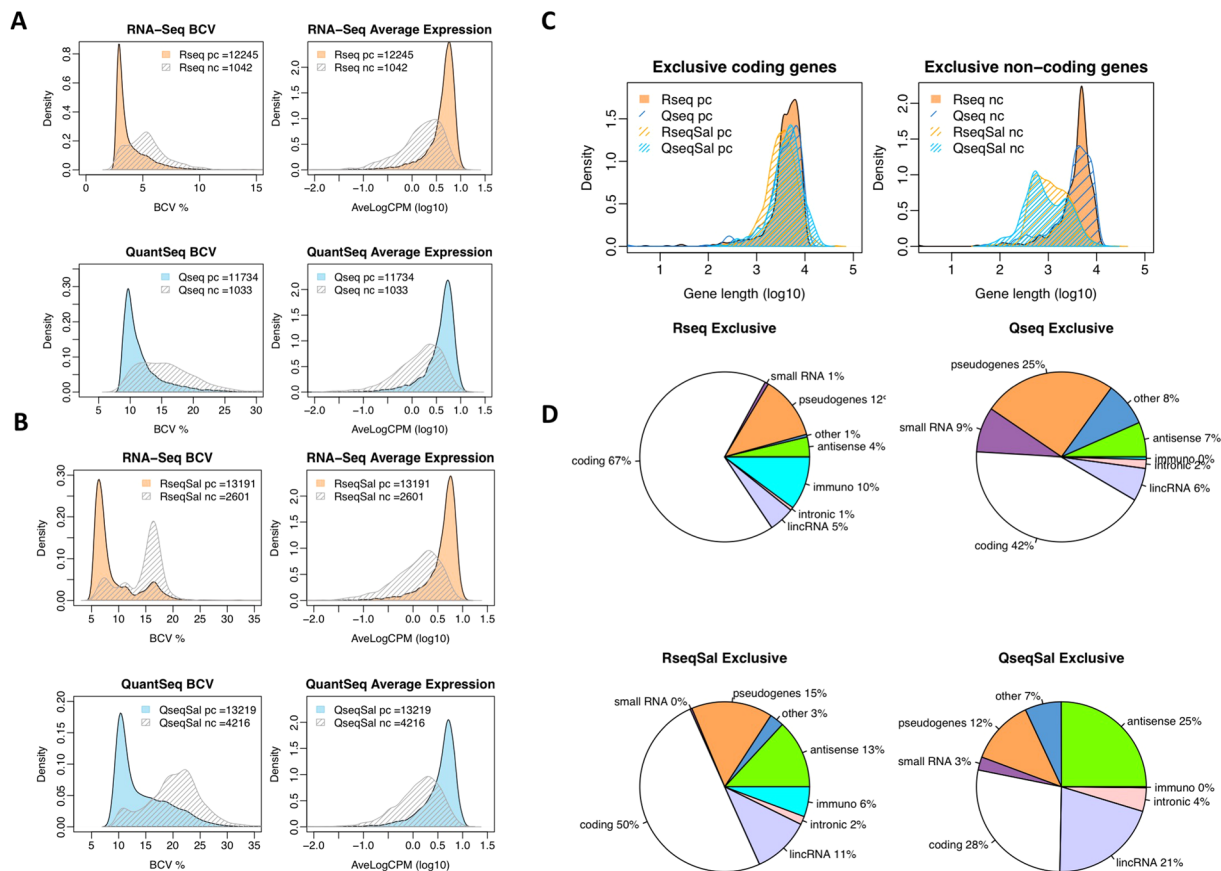
The substantial increase in the non-coding genes reported from Salmon analysis of the QuantSeq data raised the question of whether these observations are true. Characteristics of this data set may help to understand this issue. Firstly, the non-coding genes tend to be 'noisy' in that they have lower expression and higher variability between samples. This was evident when comparing average expression values and the biological co-efficient of variation (BCV) for the coding and non-coding genes (Fig. 3A,B). Secondly, we considered whether gene length could be a factor contributing to the differences seen. Distribution plots of gene length for the protein coding and non-coding genes expressed exclusively in QuantSeq or RNA-Seq were generated (Fig. 3C). We saw little difference in gene length distribution between QuantSeq and RNA-Seq in the protein coding genes. However, for the non-coding features we saw a marked shift in the distributions arising from



**Figure 2.** Comparing gene identification and gene expression for protein coding and non-coding genes. (A–C) Barplots comparing the total number of identified genes found in each of the datasets and splitting these into protein coding genes and non-coding genes. (D–F) Barplots comparing the total number of expressed genes found in each of the datasets and splitting these into protein coding genes and non-coding genes. (G–H) Violin plots showing distributions of average expression values (AveLogCPM) for protein coding genes and non-coding genes identified by the different methods. (I–J) Violin plots for all expressed genes (after filtering to retain only those genes with > 1 CPM in at least 3 samples, and with an associated gene symbol). (K–R) Venn diagrams showing pairwise comparison of expressed (‘Exp’) protein coding genes (K,M,O,Q) and non-coding genes (L,N,P,R). General colour scheme: RNA-Seq (Tophat): dark orange. QuantSeq (Tophat): dark blue, RNA-Seq (Salmon): light orange, QuantSeq (Salmon): light blue. Subsets (30 M reads per sample for RNA-Seq and 10 M reads per sample for QuantSeq are denoted with SS).

|            | Rseq | Qseq | Rseq SS | Qseq SS | Rseq Sal | Qseq Sal | Rseq Sal SS | Qseq Sal SS | Array |
|------------|------|------|---------|---------|----------|----------|-------------|-------------|-------|
| Rseq       | 1    | 0.83 | 1       | 0.83    | 0.98     | 0.78     | 0.98        | 0.78        | 0.79  |
| Qseq       |      | 1    | 0.83    | 1       | 0.82     | 0.93     | 0.82        | 0.93        | 0.7   |
| Rseq SS    |      |      | 1       | 0.83    | 0.98     | 0.78     | 0.98        | 0.78        | 0.79  |
| Qseq SS    |      |      |         | 1       | 0.82     | 0.93     | 0.82        | 0.93        | 0.7   |
| RseqSal    |      |      |         |         | 1        | 0.8      | 1           | 0.8         | 0.79  |
| QseqSal    |      |      |         |         |          | 1        | 0.8         | 1           | 0.68  |
| RseqSal SS |      |      |         |         |          |          | 1           | 0.8         | 0.79  |
| QseqSal SS |      |      |         |         |          |          |             | 1           | 0.68  |
| Array      |      |      |         |         |          |          |             |             | 1     |

**Table 2.** Spearman’s Rank Correlation for Average expression of Expressed Genes.



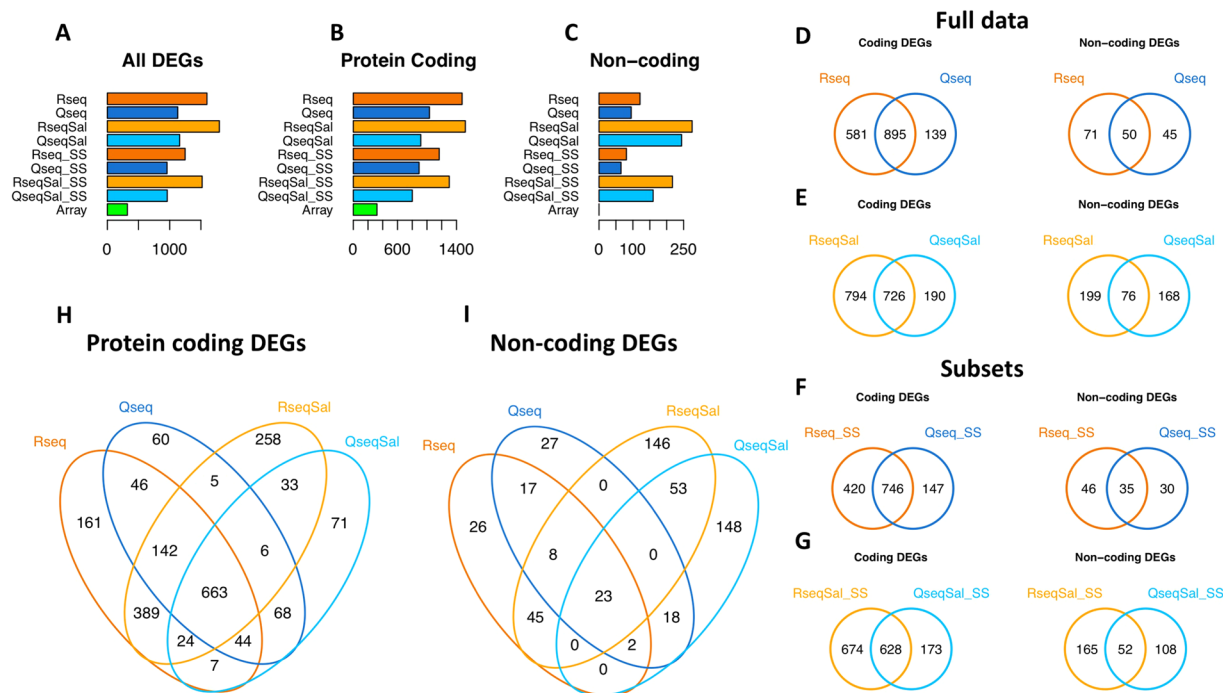
**Figure 3.** Comparing genes expressed in RNA-Seq or QuantSeq. Distribution plots of biological coefficient of variation and average expression comparing the protein coding (pc) genes and non-protein coding (nc) genes for (A) Tophat data (B) Salmon quantification. (C) Distribution plots of gene length for genes expressed exclusively in RNA-Seq or QuantSeq. (D) Pie charts showing proportion of protein coding and non-coding gene types for the sets of genes expressed exclusively in RNA-Seq or QuantSeq, top row Tophat mapping, lower row Salmon quantification.

Salmon quantification, with shorter non-coding features being seen in both the QuantSeq and RNA-Seq exclusive non-coding genes (Fig. 3C). Thirdly, we assessed the types of non-coding features which were being found exclusively in RNA-Seq or QuantSeq. Pie charts of the gene biotype are presented in Fig. 3(D). Interestingly, with RNA-Seq we saw an increase in the immunity related non-coding genes, the TR\_V genes and the IG\_V genes and with QuantSeq an increase in small RNAs, pseudogenes and antisense. In the Salmon-generated counts we saw an increase in lincRNA and antisense transcripts as compared to counts generated by Tophat. In summary the exclusive non-coding genes tended to be lowly expressed, variable across samples and of shorter length.

**Differential expression analysis.** We next sought to understand the capacity of the QuantSeq, RNA-Seq and microarray methods to discover changes in gene expression between normal and Poly(I:C)-treated PBMCs. We performed differential expression analysis using (i) RNA-Seq with approximately 92 M reads/for each of the 6 samples (ii) QuantSeq with approximately 30 M reads/for each of the 6 samples (iii) RNA-Seq with approximately 30 M reads/for each of the 6 samples (iv) QuantSeq with approximately 10 M reads/for each of the 6 samples (v) Clariom S microarrays of the 6 samples. Each of the non-microarray experiments were quantified using the Tophat and also the Salmon methods. This produced 9 result sets in total. We used an FDR cut-off of  $<0.05$  and incorporated a fold change threshold into the statistical testing using the TREAT approach<sup>31</sup> to be more certain of obtaining biologically significant results.

The total number of differentially expressed genes (DEGs) found in each different result set are given in Table 1 and are compared in barplots (Fig. 4A–C). RNA-Seq identified more DEGs than QuantSeq (~30–40% with Tophat and ~50–60% with Salmon, Fig. 4A). Microarray identified far fewer DEGs, around 75–80% less than those found by RNA-Seq or QuantSeq with only 2% not being found by the other methods. If the DEGs were split into protein coding and non-coding genes (Fig. 4B,C respectively), it is apparent that RNA-Seq identifies a greater number of both protein coding and non-coding DEGs than QuantSeq. Figure 4C also demonstrates that Salmon quantification led to an increase in the number of non-coding DEGs compared to Tophat mapping.

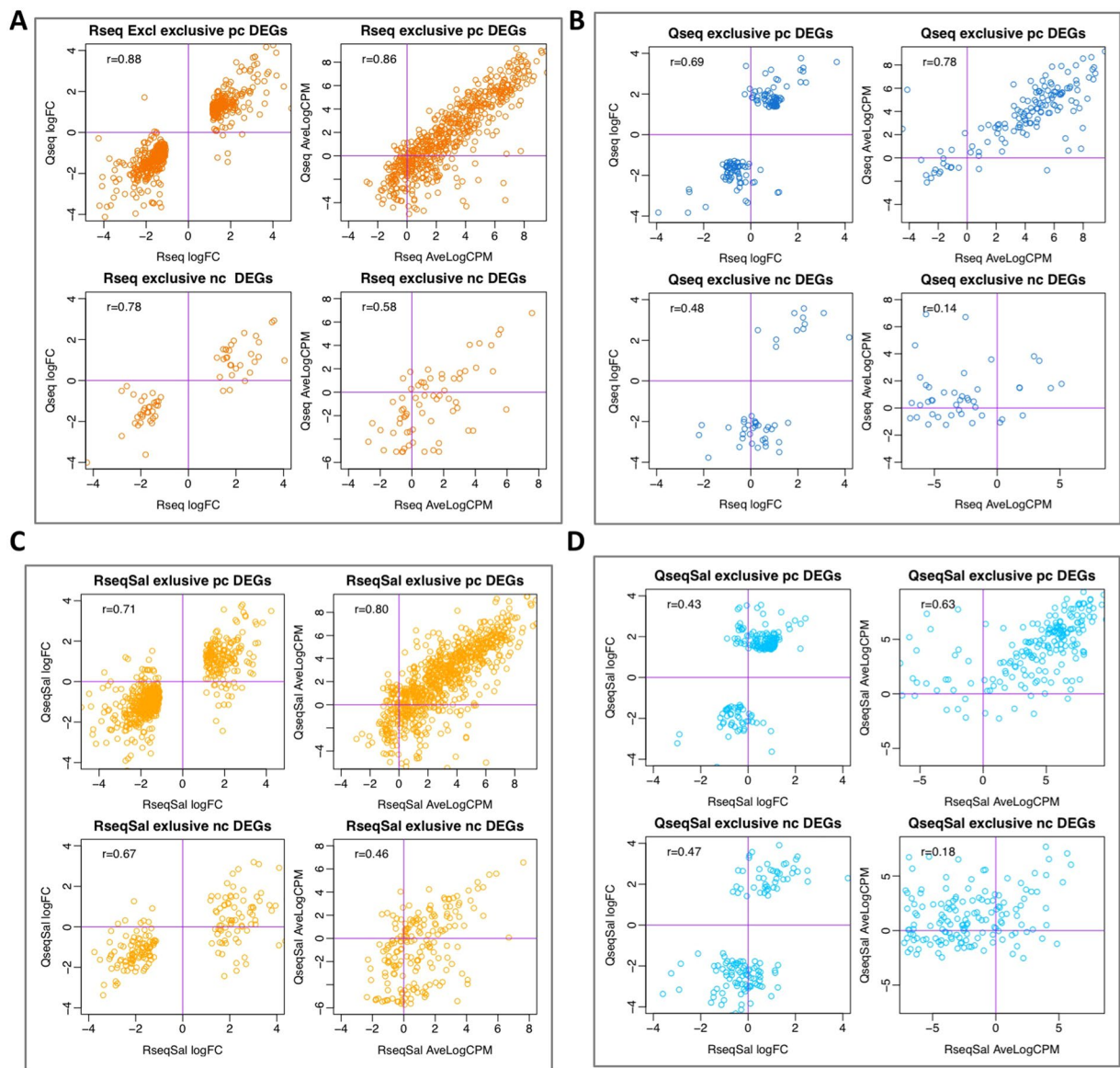
We carried out a pairwise comparison of all RNA-Seq and QuantSeq DEGs and calculated the number of protein coding and non-coding DEGs shared in each pairwise comparison (Supplementary Tables S2). Venn



**Figure 4.** Comparing differentially expressed genes (DEGs). (A–C) Barplots comparing the total number of differentially expressed genes (DEGs) found in each of the result sets (A) and split into the number of protein coding DEGs (B) and the number of non-coding DEGs (C). (D–G) Venn diagrams comparing RNA-Seq and QuantSeq DEGs, split into protein coding genes (left side) and non-coding genes (right side). (H–I) Venn diagrams of DEGs found with RNA-Seq and QuantSeq using Tophat or Salmon quantification. General colour scheme: RNA-Seq (Tophat): dark orange. QuantSeq (Tophat): dark blue, RNA-Seq (Salmon): light orange, QuantSeq (Salmon): light blue, Microarray: green. Subsets (30 M reads per sample for RNA-Seq and 10 M reads per sample for QuantSeq are denoted with SS).

diagrams for four pairwise comparisons of both protein coding and non-coding DEGs are shown in Fig. 4(D–G). These comparisons showed that around 80% of the QuantSeq protein coding DEGs are also identified by RNA-Seq. There was less agreement between the RNA-Seq and QuantSeq methods for the non-coding DEGs. However, these subsets of non-coding genes form a smaller percentage of the entire set, less than 8% of the DEGs produced using Tophat are non-coding whereas up to 21% are non-coding if Salmon quantification is used. We also found that the subsets comprised of 30 M and 10 M reads per sample for RNA-Seq and QuantSeq respectively versus the full data set of 92 M reads per sample for RNA-Seq and 30 M reads per sample for QuantSeq recovered 80–85% of the DEGs found by the full data sets (Table 1, Fig. 4A, Supplementary Table S2). Furthermore, the RNA-Seq subset of 30 M reads per sample identified around 10% more DEGs than the full QuantSeq dataset, also having 30 M reads per sample. Comparisons of the DEGs found in RNA-Seq and QuantSeq with both Tophat and Salmon quantification are presented in Venn diagrams (Fig. 4(H–I)). Interestingly, this shows that a core set of coding genes are found as differentially expressed, irrespective of approach, but that the non-coding genes found are strongly influenced by the sequencing and read mapping approach taken.

**Why are some genes differentially expressed by RNA-Seq or QuantSeq but not by both?** We examined the DEGs found in RNA-Seq but not in QuantSeq and vice versa to identify any characteristics which could help to explain why these genes are identified as being differentially expressed by one method but not the other. We calculated Spearman's correlation coefficient for the logFC values and average expression values of protein coding DEGs found by RNA-Seq but not QuantSeq. This revealed a high correlation of the fold change and average expression (logFC  $r = 0.88$ , AveExpr  $r = 0.86$  (Tophat,  $n = 576$ )). When we did the same for the 780 protein coding DEGs found exclusively in RNA-Seq using Salmon we found reduced but still strong correlation (logFC  $r = 0.71$ , AveExpr  $r = 0.80$  (Salmon,  $n = 780$ )). This demonstrates that the expression profile of DE protein coding genes found uniquely by RNA-Seq is likely due to the failure of the QuantSeq data to meet the thresholds for differential expression. There were far fewer DEGs found exclusively in the QuantSeq data. In this case we found reduced correlation of fold change and average expression (logFC  $r = 0.69$ , AveExpr  $r = 0.78$  (Tophat,  $n = 139$ ) and (logFC  $r = 0.43$ , AveExpr  $r = 0.63$  (Salmon,  $n = 196$ )). Scatterplots of fold change and average expression for genes found to be differentially expressed exclusively in RNA-Seq or QuantSeq are presented in Fig. 5. We found lower correlation for the non-coding DEGs, however there were also fewer of these genes. Looking more closely at the DEGs unique to QuantSeq, we found a slight shift towards shorter non-coding genes (Fig. 6B) but similar length distributions for the protein coding genes (Fig. 6A). We also saw that a greater proportion of the QuantSeq exclusive DEGs were non-coding genes (Fig. 6C), with a notable increase in the small

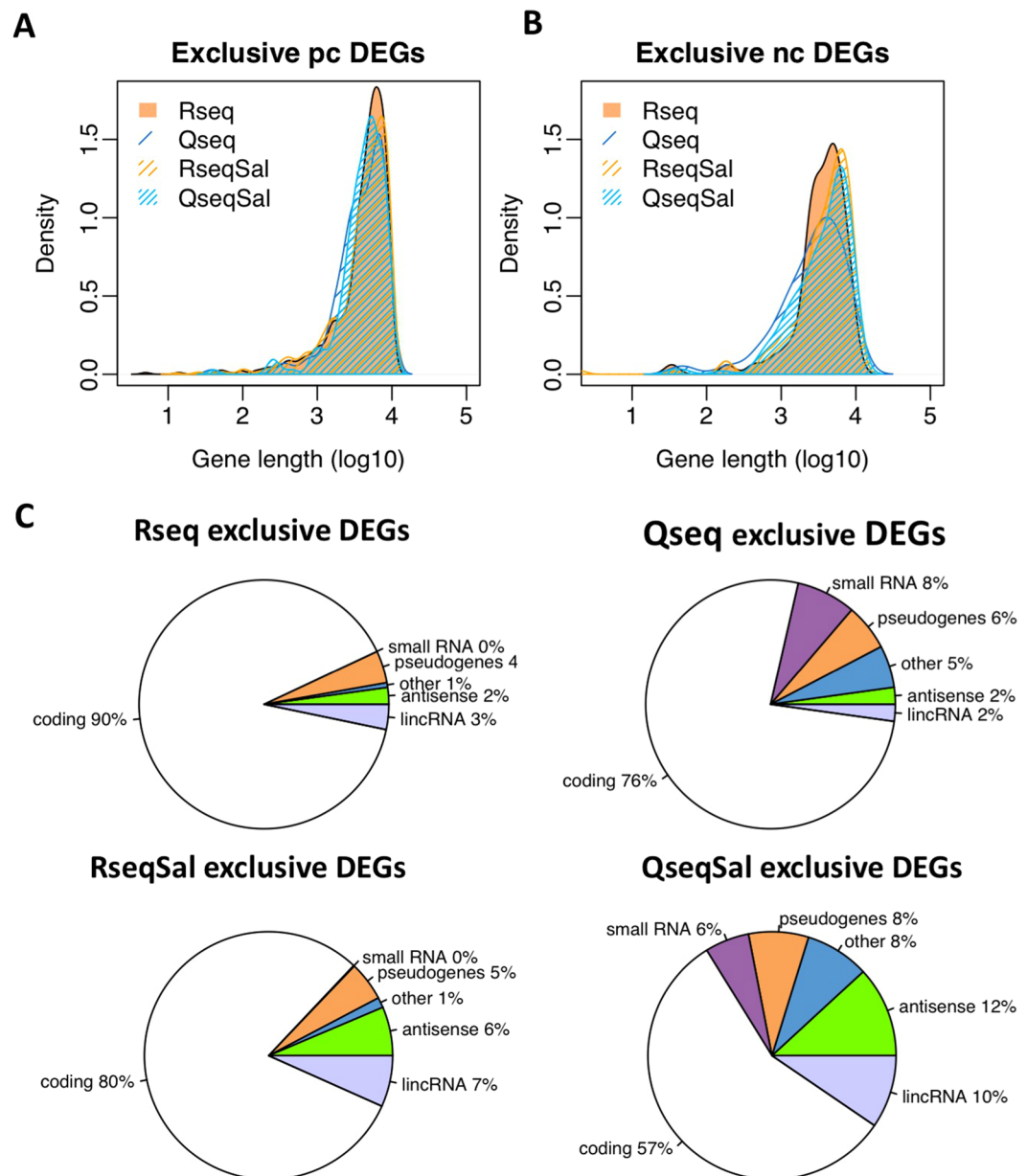


**Figure 5.** Comparing fold change and expression in DEGs found exclusively by RNA-Seq or QuantSeq. Scatterplots of fold change and average expression comparing values derived from RNA-Seq and QuantSeq data for the DEGs found by one method and not the other. Each panel, top row LogFC and AveLogCPM for protein-coding (pc) DEGs, bottom row non-coding (nc) DEGs. **(A)** DEGs found by RNA-Seq not QuantSeq using Tophat. **(B)** DEGs found by QuantSeq not RNA-Seq using Tophat. **(C–D)** using Salmon.

RNAs. The small RNAs (snoRNA, scrRNA) are by definition short sequences and accordingly the two factors, gene length and biotype are related.

**Biological insight and gene set analysis.** The endpoint of these analyses is to obtain insight into the biological or functional differences between the conditions being compared. We performed gene set analysis using the Camera function<sup>32</sup> to obtain sets of functionally related genes associated with the poly(I:C) treatment. Gene Set Analysis is based on the principle of co-regulation; that genes which are functionally related are likely to have similar expression patterns. This analysis method considers the expression values of all expressed genes rather than being restricted to a group of DEGs chosen by an arbitrary cut-off (such as,  $FDR < 0.05$  and  $LogFC > 1$ ). In the top 20 gene sets identified in each of our data sets (Fig. 7A) we saw a strong signature of up-regulation and that the most significant gene sets are those contributing to the interferon induced anti-viral response, including a gene set previously described by us, the Bosco Interferon Induced Antiviral Module<sup>33</sup>. These highly significant gene sets were detected by all versions of the RNA-Seq and QuantSeq methods tested, including the subset data. Furthermore, we saw agreement in the microarray data set in these gene signatures. The microarray gene sets included 5 down-regulated sets, these were also significantly down-regulated in the RNA-Seq and QuantSeq data under both quantification methods but did not fall within the top 20. These findings are consistent with

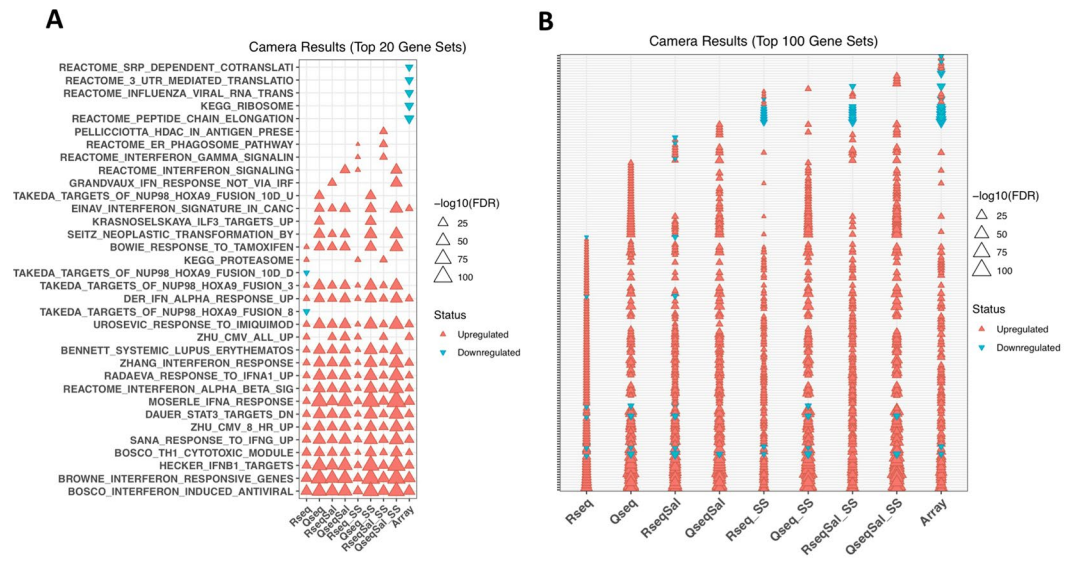




**Figure 6.** Gene length and biotype in the exclusive DEGs. (A,B) Distribution plots of gene length for the non-common differentially expressed protein coding (pc) genes and non-coding (nc) genes (DEGs) show a slight shift towards genes with a shorter sequencing region in the QuantSeq data, most noticeable in the non-coding genes. (C) Classification of these DEGs by biotype reveals enrichment of non-coding RNAs, particularly in the QuantSeq data.

our expectations for this model system and serves as proof of principle that each of the expression quantification methods is detecting differential expression in the groups of genes we expect for this system. When we looked more broadly at the top 100 gene sets we still saw good agreement across the methods (Fig. 7B). 95 of the top 100 gene sets found by QuantSeq were also found to be significantly different in the RNA-Seq data and all of the top 100 gene sets identified by RNA-Seq were found to be significant in the QuantSeq data. Barcode plots which show the fold changes for the genes in the Bosco Interferon Induced Antiviral Module can be found in Supplementary Fig. S3.

Gene set analysis is particularly powerful as it avoids the problem of some genes failing to meet an imposed FDR cut-off in order to be classified as differentially expressed. However, the use of a restricted set of DEGs in functional analysis is also a common analytical approach. If we took only the DEGs and compare the top gene ontology terms and top KEGG pathways enriched in those lists we still found good agreement between RNA-Seq and QuantSeq with approximately 90% of the top 500 gene ontology functional classifications common to RNA-Seq and QuantSeq and 70% of the top 20 KEGG pathways (Supplementary Fig. S4). The four top KEGG



**Figure 7.** Gene Set analysis identifies antiviral signatures. Gene set analysis using Camera identifies enrichment of gene sets involving antiviral and inflammatory responses. **(A)** The top 20 gene sets arising in each of the data sets are plotted, as the top 20 are not exactly the same for each data set the total number of gene sets is greater than 20 in this plot. **(B)** The top 100 gene sets are plotted as the top 100 are not exactly the same for each data set the total number of gene sets is greater than 100 in this plot. Columns are ordered as RNA-Seq (Tophat), QuantSeq(Tophat), RNA-Seq (Salmon), QuantSeq (Salmon), RNA-Seq subset (30 M) (Tophat), QuantSeq subset (10 M) (Tophat), RNA-Seq subset (30 M) (Salmon), QuantSeq subset (10 M) (Salmon).

pathways that appeared in each of our four data sets are Cytokine-cytokine receptor interaction, NOD-like receptor signalling pathway, Herpes simplex infection and Influenza A.

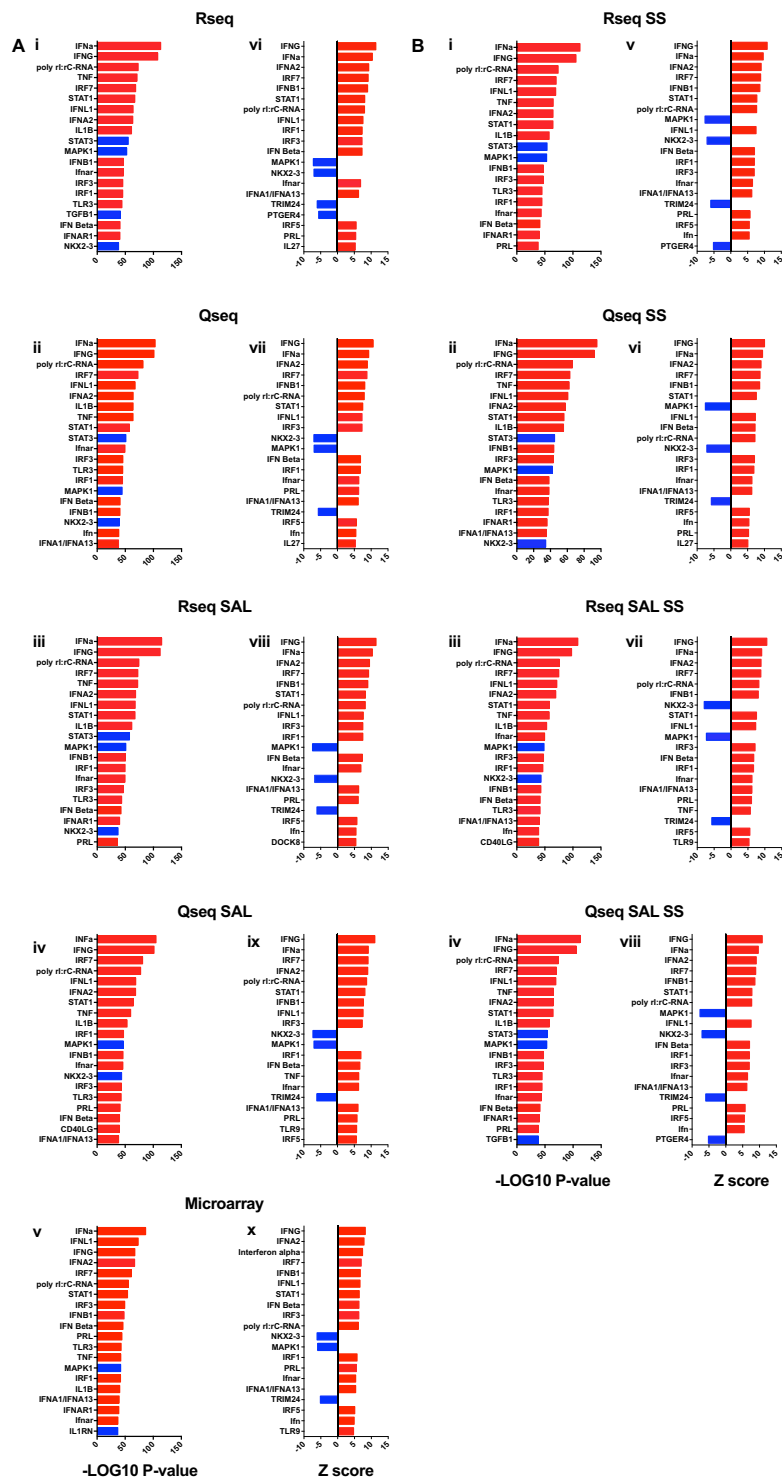
Finally, we performed Upstream Regulator Analysis (Ingenuity Systems) on the DEGs from all four datasets and the microarray dataset (Fig. 8A), and the four subset data sets (Fig. 8B). This analysis uses prior knowledge to identify transcriptional regulators (TR) that drive the observed gene expression changes to gain further biological insight. Two statistical measures are presented; (1) the overlap p-value, which measures the overlap between the DEGs and the target genes that are regulated by a TR (Fig. 8A(i-v), B(i-iv)); and (2) the activation Z-score, which measures the pattern match between expected relationship direction of a TR (activation or inhibition) and observed gene expression changes (up/downregulation) (Fig. 8A(vi-x), B(v-viii)). As illustrated in Fig. 8, PolyI:C is a top driver of gene expression changes in all datasets and the microarray data. Furthermore, the top drivers in all 5 datasets are dominated by well-known antiviral (type I and III interferons, IFNG, IRF3 and IRF7) and proinflammatory genes (TNF and IL1b). When the data sets are subset (Fig. 8B) the same key response drivers emerge as seen for the full data sets.

## Discussion

In this study we compared QuantSeq to RNA-Seq and microarrays and assessed its ability to return similar gene expression profiles and identify differentially expressed genes. This was done in a test case involving PBMCs stimulated with a viral double-stranded RNA mimic. As well as evaluating QuantSeq we also compared a new quantification method, Salmon, with Tophat2. Our design allowed us to compare the effect of library preparation (QuantSeq vs RNA-Seq), quantification method (Tophat2 vs Salmon) and sequencing depth (Full datasets vs subsets), on the identification of transcribed genes and the calling of DEGs.

Overall, we saw consistency in the expressed protein coding genes, with QuantSeq finding 94% of the expressed protein coding genes found by RNA-Seq, irrespective of quantification method. Comparing QuantSeq and RNA-Seq to microarray analysis we also found high correlation between ranked expression levels for the protein coding genes. We found reduced correlation and more variability in the non-coding genes which comprised between 8–24% of the expressed genes. At the outset it should be borne in mind that our library preparation depended on the presence of a polyA tail on the transcripts to be sequenced. As non-coding RNA transcripts fall into three categories, (i) those with a polyA tail, (ii) those without a poly A tail and (iii) those which are bimorphic (appearing both with and without a polyA tail)<sup>34</sup> we would not expect to see the full population of non-coding RNA in our data sets. Using Tophat mapping we saw a similar number of non-coding features in the RNA-Seq and QuantSeq experiments (~1000). However, when we employed Salmon quantification we found an overall increase in non-coding features with a greater number being identified in the QuantSeq experiment, QuantSeq (~4000) versus RNA-Seq (~2500).

We also found a marked shift in the length distribution of the non-coding features with Salmon identifying a population of short non-coding transcripts, including a number of snRNA and snoRNA with a length less than 300bp (around 10% of the Salmon/QuantSeq exclusive non-coding features). The remaining non-coding genes were enriched for antisense and lincRNA. It has previously been reported that most lincRNA are polyadenylated<sup>35</sup> which is consistent with our sequencing approach. Interestingly, it has been reported that lincRNAs have similar sequence and structural features to 3'UTR<sup>36</sup>.



**Figure 8.** Identification of Transcriptional Regulators (TRs) of the polyI:C response in PBMCs. Molecular drivers of gene expression changes were inferred using Upstream Regulator Analysis on the full expression sets (A), and the subset data (B). Drivers are ranked by P-value (A (i-v) and B (i-iv)) and activation Z score (A (vi-x) and B (v-viii)). Red bars indicate pathway activation (activation Z-score > 2.0) and blue bars indicate pathway inhibition (activation Z-score < -2.0).

In summary, we found high concordance in the expressed protein coding genes across the methods tested. The small proportion of discordant genes tended to be non-coding genes with lower expression and shorter length than the concordant genes. Other benchmarking studies contrasting RNA-Seq quantification methods have also reported that non-concordant genes tend to have lower expression and shorter length<sup>28,37,38</sup>.

Performing differential expression analysis to assess the effect of Poly(I:C) on gene expression we found that that library preparation method (QuantSeq vs RNA-Seq) had a greater effect on our differential expression results than quantification method or read depth. RNA-Seq identified a greater number of DEGs than QuantSeq irrespective of quantification method. Nevertheless, a high percentage of the protein coding DEGs found by QuantSeq were also common to RNA-Seq (89% using Tophat, 79% using Salmon). Again, there were fewer differentially expressed non-coding DEGs and more variability in the sets found by QuantSeq and RNA-Seq. When we looked more closely at the DEGs called by RNA-Seq but not by QuantSeq we found that a majority had similar expression characteristics (logFC and average expression) whilst failing to meet our logFC and FDR cut-offs to be classified as DEGs. Both QuantSeq and RNA-Seq clearly outperformed microarray in identifying differentially expressed genes.

We found remarkable consistency in the gene sets emerging as most biologically significant in our differential expression analysis. The most significant gene set in each of our data sets was the Bosco Interferon Induced Anitviral Module. This gene signature meets our expectation for our test case in which we anticipated elevation of gene sets involved in inflammation and anti-viral responses. We see this result clearly emerging from all the methods tested including when we use subsets of data consisting of 10 M QuantSeq reads per sample. In addition, we see similar transcription factors acting as drivers of the changed gene expression in our data sets. We are mindful however that not all biological investigations will be as clear cut as our test case, which involved a strong biological response. It may be that we would see a divergence of RNA-Seq and QuantSeq outcomes in a more nuanced scenario.

As QuantSeq only sequences the 3' end of the RNA fragment less reads should be required than for RNA-Seq where reads are derived from the entire fragment. Lexogen suggests that 10 M reads per sample are adequate for quantification. We tested this and compared results using 10 M reads per sample to those obtained with three times that read depth. We found very little difference in the number of expressed genes (less than 0.5%) using the QuantSeq subset compared to the full data set. When we performed differential expression analysis we found a modest decrease in the number of DEGs of around 15%. Nevertheless, the same biological signals were highlighted in the subset data as found for the full data set. Comparing QuantSeq with 30 M reads per sample to RNA-Seq with 30 M reads per sample we found that RNA-Seq had a small increase in expressed genes and around a 10% increase in DEGs compared to QuantSeq. This draws us to conclude that QuantSeq data consisting of 10 M reads per sample is adequate for straight forward gene expression analysis and that little is to be gained by deeper sequencing in this case.

As well as evaluating QuantSeq and RNA-Seq we have compared the new quantification method, Salmon with Tophat2. Quantification with Salmon took less than 10 minutes per sample (6 CPUs) delivering the speed advantages we expected<sup>17</sup>. The resulting data file was around 1/100 the size of a bam file. The speed of processing and reduced storage requirements are important considerations in applying these techniques to projects involving many samples such as longitudinal cohort studies, single cell RNA-Seq and clinical applications.

## Conclusions

Overall, this study shows that combining QuantSeq with a fast quantification method such as Salmon is capable of providing a clear diagnostic signal in a relatively quick and inexpensive manner and may therefore be suited to high throughput clinical applications or to single cell transcriptomics. An additional benefit of QuantSeq is that the high read coverage over the 3' UTR is likely to be of particular interest to those studying this important gene regulatory region. It should be noted however that the QuantSeq - Salmon workflow is restricted to assessing gene expression changes and does not provide information regarding mutations/base changes or novel transcripts. This pipeline is ideally suited to situations where straight forward gene expression changes would assist in diagnostics or treatment.

## Methods

**Cellular stimulation and RNA preparation.** Whole blood was obtained from a healthy adult volunteer, and peripheral blood mononuclear cells (PBMC) were isolated over Ficoll and cryobanked as previously described<sup>39</sup>. The cryobanked PBMC were thawed and cultured in AIM-V medium (Gibco) with 50 µg/ml polyI:C extract (Invivogen), or medium alone (control) in triplicate. After 24 h, cells were harvested and total RNA was extracted using TRIzol (Invitrogen) followed by purification on an RNeasy column (Qiagen)<sup>40</sup>. Quality of the total RNA was assessed on the Bioanalyser (ThermoFisher) and RIN scores were  $\geq 8$  for all samples. Six samples (3 treated with Poly(I:C) and 3 controls) were used in each of the RNA-Seq, QuantSeq and microarray experiments.

**RNA library preparation.** RNA-Seq libraries were prepared using the Illumina TruSeq Stranded mRNA Prep Kit according to the manufacturer's instructions. The RNA-Seq libraries were sequenced using the Illumina NextSeq 500 to produce 75 bp paired-end reads for each sample. The QuantSeq libraries were prepared using Lexogen's QuantSeq 3' mRNA-Seq Library Prep Kit for Illumina, according to the manufacturer's instructions. The QuantSeq libraries were sequenced using the Illumina NextSeq 500 to produce 75 bp single-end reads for each sample.

Library preparation and sequencing was done at the Ramaciotti Centre for Genomics (University of New South Wales, Australia).

**Microarray.** Clariom S Human Assays (ThermoFisher) were used for microarray analysis. The analysis was undertaken in accordance with the manufacturer's instructions, by the Ramaciotti Centre for Genomics.

**Bioinformatic analysis of RNA-Seq and QuantSeq reads.** The QuantSeq reads were trimmed to remove the first 12 bp from the 3' end in accordance with Lexogen's recommendation. For this we used trim\_galore (v 0.4.4) with the command `three_prime_clip_R1 12`. RNA-Seq reads were not trimmed.

The RNA-Seq and QuantSeq reads (fastq files) were mapped to the Ensembl *Homo sapiens* genome (GRCh38). Mapping was performed with Tophat2 (v 2.0.12)<sup>41</sup> calling Bowtie2 (v 2.2.3)<sup>19</sup>. The featureCounts function of Subread (v 1.4.6-p5)<sup>25</sup> was used to generate counts of reads uniquely mapped to annotated genes using the GRCh38 gtf file. In addition we used the abundance quantification method Salmon<sup>17</sup> to calculate transcript abundance based on the transcript fasta files downloaded from ensemble-release87, (Homo\_sapiens.GRCh38.cdna.all.fa and Homo\_sapiens.GRCh38.ncrna.all.fa). We ran Salmon with the following default settings for the RNA-seq data: quant -i transcripts\_index -l ISR -1 \*R1\*fastq\* -2 \*R2\*fastq\* -o \$output -p 6 -posBias-gcBias-seqBias -writeUnmappedNames. For the QuantSeq data which would only capture the 3' end of the transcript we changed the settings to: quant -i transcripts\_index -l SR -r \*R1\_trimmed.fq\* -o \$output -p 6 -writeUnmappedNames. The Bioconductor package tximport<sup>22</sup> was used to convert transcripts to gene counts which could be used in downstream differential expression analysis.

We used the polyA\_DB v3 database<sup>29</sup> to ascertain polyadenylation sites (PAS). We selected those sites which are found in the 3' UTR and which have an extension. This yielded 32969 PASs found in 9967 genes. We took those genes which have 1, 2, 4, 6, 8 and 10 PASs and compared expression (AveLogCPM) in RNA-Seq and Quant-Seq by plotting boxplots of distribution.

Differential expression analysis was performed using functions from the Bioconductor packages edgeR (v 3.14.0)<sup>42</sup> and voom limma (v 3.32.7)<sup>43,44</sup>. We excluded lowly expressed genes and tested those genes with expression of at least 1 CPM (counts per million) in at least three samples. Counts were normalized using the TMM method and generalized linear models were used for differential expression analysis, using the functions glmQLFit and glmTreat<sup>31</sup> with  $\text{lfc} = \log_2(2)$ . In all cases differentially expressed genes were defined as those genes with a Benjamini-Hochberg corrected p value less than 0.05<sup>45</sup>.

**Bioinformatic analysis of microarrays.** The Bioconductor packages oligo and limma were used to perform normalization and differential expression analysis. The probes were annotated using the database clarimshumantranscriptcluster.db\_8.6.0, available through Bioconductor. Normalization was performed using the Robust Multichip Average (RMA) algorithm. Linear models were constructed for all samples and the contrast of interest in this case the PBMCs with and without Poly(I:C) treatment was tested using empirical Bayes moderated t tests. The Treat function was used with  $\text{lfc} = \log_2(2)$ . In all cases differentially expressed genes were defined as those genes with a Benjamini-Hochberg corrected p value less than 0.05.

**Gene set analysis.** Gene set analysis was performed using the camera function<sup>32</sup> within limma. The genes tested for differential expression were evaluated against the human\_c2\_v5p2 gene set collection downloaded from <http://bioinf.wehi.edu.au>. The camera algorithm analyses enrichment of genes in a set of interest and produces p-values and FDR values for the gene set being up-regulated or down-regulated in the data.

**Gene ontology and KEGG pathway enrichment and upstream regulator analysis.** The most highly significant enriched gene ontology (GO) terms were assessed using the fea\_topGO function of the package FGNet package<sup>46</sup>. The most significant KEGG pathways were assessed using the kegg function of limma. The input for these functions were the lists of DEGs derived in our four data sets. Molecular drivers of the gene expression responses were identified utilizing upstream regulator analysis in the Ingenuity Pathway Analysis package (Qiagen Bioinformatics)<sup>47</sup>.

**Ethics approval and consent to participate.** This study was performed in accordance with the National Statement on Ethical Conduct in Human Research from the National Health and Medical Research Council. The study was approved by the Human Research Ethics Committee at the University of Western Australia. Written informed consent was obtained from study participants.

## Data availability

The RNA-Seq sequencing data has been submitted to the NCBI Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>) and under accession number GSE123523.

Received: 28 May 2019; Accepted: 26 November 2019;

Published online: 11 December 2019

## References

- Engstrom, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10**, 1185–1191, <https://doi.org/10.1038/nmeth.2722> (2013).
- Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115–121, <https://doi.org/10.1038/nmeth.3252> (2015).
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet* **30**, 418–426, <https://doi.org/10.1016/j.tig.2014.07.001> (2014).
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610–620, <https://doi.org/10.1016/j.molcel.2015.04.005> (2015).
- Byron, S. A. *et al.* sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* **17**, 257–271, <https://doi.org/10.1038/nrg.2016.10> (2016).
- Kaddi, C. D., Coulter, W. H. & Wang, M. D. Developing Robust Predictive Models for Head and Neck Cancer across Microarray and RNA-seq. *Data. ACM BCB* **2015**, 393–402, <https://doi.org/10.1145/2808719.2808760> (2015).
- Lesurf, R. *et al.* Genomic characterization of HER2-positive breast cancer and response to neoadjuvant trastuzumab and chemotherapy—results from the ACOSOG Z1041 (Alliance) trial. *Ann Oncol* **28**, 1070–1077, <https://doi.org/10.1093/annonc/mdx048> (2017).

8. Fumagalli, D. *et al.* Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. *BMC Genomics* **15**, 1008, <https://doi.org/10.1186/1471-2164-15-1008> (2014).
9. Deng, M. C. *et al.* Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. *Am J Transplant* **6**, 150–160, <https://doi.org/10.1111/j.1600-6143.2005.01175.x> (2006).
10. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211, <https://doi.org/10.1126/science.aad0095> (2015).
11. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998–1013 e1020, <https://doi.org/10.1016/j.cell.2018.10.038> (2018).
12. Moll, P., Ante, M., Seitz, A. & Reda, T. QuantSeq. 3' mRNA sequencing for RNA quantification. *Nature Methods* **11**, 952 (2014).
13. Tandonnet, S. & Teixeira Torres, T. Traditional versus 3' RNA-seq in a non-model species. *Genomics Data* **11**, 9–16 (2017).
14. Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20**, 1684–1696, <https://doi.org/10.1261/rna.046011.114> (2014).
15. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30**, 301–304, <https://doi.org/10.1093/bioinformatics/btt688> (2014).
16. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585, <https://doi.org/10.1038/ng.2653> (2013).
17. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419, <https://doi.org/10.1038/nmeth.4197> (2017).
18. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525, <https://doi.org/10.1038/nbt.3519> (2016).
19. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, <https://doi.org/10.1186/gb-2013-14-4-r36> (2013).
20. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360, <https://doi.org/10.1038/nmeth.3317> (2015).
21. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21, <https://doi.org/10.1093/bioinformatics/bts635> (2013).
22. Sonesson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, <https://doi.org/10.12688/f1000research.7563.2> (2015).
23. Troy, N. M. B. A. Respiratory viral infections and host responses; insights from genomics. *Respiratory Research* **17**, <https://doi.org/10.1186/s12931-016-0474-9> (2016).
24. Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–263, <https://doi.org/10.1126/science.1179050> (2009).
25. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930, <https://doi.org/10.1093/bioinformatics/btt656> (2014).
26. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185, <https://doi.org/10.1093/bioinformatics/bts356> (2012).
27. Hesketh, J. In *Encyclopedia of Life Sciences* (2005).
28. Corley, S. M., MacKenzie, K. L., Beverdam, A., Roddam, L. F. & Wilkins, M. R. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC Genomics* **18**, 399, <https://doi.org/10.1186/s12864-017-3797-0> (2017).
29. Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA\_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Research* **46**, 315–319, [10.1093/nar/gkx1000](https://doi.org/10.1093/nar/gkx1000) (2018).
30. Chen, Y., Lun, A. T. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438, <https://doi.org/10.12688/f1000research.8987.2> (2016).
31. McCarthy, D. J. & Smyth, G. K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25**, 765–771, <https://doi.org/10.1093/bioinformatics/btp053> (2009).
32. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* **40**, e133, <https://doi.org/10.1093/nar/gks461> (2012).
33. Bosco, A., Ehteshami, S., Stern, D. A. & Martinez, F. D. Decreased activation of inflammatory networks during acute asthma exacerbations is associated with chronic airflow obstruction. *Mucosal Immunol* **3**, 399–409, <https://doi.org/10.1038/mi.2010.13> (2010).
34. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L. L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* **12**, R16, <https://doi.org/10.1186/gb-2011-12-2-r16> (2011).
35. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17**, 47–62, <https://doi.org/10.1038/nrg.2015.10> (2016).
36. Niazi, F. & Valadkhan, S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3'UTRs. *RNA* **18**, 825–843, <https://doi.org/10.1261/rna.029520.111> (2012).
37. Everaert, C. *et al.* Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci Rep* **7**, 1559, <https://doi.org/10.1038/s41598-017-01617-3> (2017).
38. Robert, C. & Watson, M. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol* **16**, 177, <https://doi.org/10.1186/s13059-015-0734-x> (2015).
39. Heaton, T. *et al.* An immunoepidemiological approach to asthma: identification of *in-vitro* T-cell response patterns associated with different wheezing phenotypes in children. *Lancet* **365**, 142–149, [https://doi.org/10.1016/S0140-6736\(05\)17704-6](https://doi.org/10.1016/S0140-6736(05)17704-6) (2005).
40. Troy, N. M., Hollams, E. M., Holt, P. G. & Bosco, A. Differential gene network analysis for the identification of asthma-associated therapeutic targets in allergen-specific T-helper memory responses. *BMC Med Genomics* **9**, 9, <https://doi.org/10.1186/s12920-016-0171-z> (2016).
41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359, <https://doi.org/10.1038/nmeth.1923> (2012).
42. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140, <https://doi.org/10.1093/bioinformatics/btp616> (2010).
43. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, <https://doi.org/10.1093/nar/gkv007> (2015).
44. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, <https://doi.org/10.1186/gb-2014-15-2-r29> (2014).
45. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289–300 (1995).
46. Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31**, 1686–1688, <https://doi.org/10.1093/bioinformatics/btu864> (2015).
47. Kramer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530, <https://doi.org/10.1093/bioinformatics/btt703> (2014).

## Acknowledgements

M.R.W. acknowledges funding from the Australian Government NCRIS scheme, administered by Bioplatforms Australia, and by the New South Wales State Government RAAP scheme. A.B. is supported by a Fellowship from the Simon Lee Foundation. N.T. is supported by Asthma Australia Mickie Hardy PhD Scholarship.

## Author contributions

A.B., M.W., S.C. and N.T. made contributions to study conception, design and analysis. A.B. and N.T. collected and prepared the RNA samples. S.C. designed and performed the bioinformatic analysis in consultation with M.W., S.C. drafted the manuscript. M.W., A.B. and N.T. made contributions to the final manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-019-55434-x>.

**Correspondence** and requests for materials should be addressed to S.M.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019