Chapter 16

Impact of Gene Annotation on RNA-seq Data Analysis

Shanrong Zhao and Baohong Zhang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61197

Abstract

RNA-seq has become increasingly popular in transcriptome profiling. One of the major challenges in RNA-seq data analysis is the accurate mapping of junction reads to their genomic origins. To detect splicing sites in short reads, many RNA-seq aligners use reference transcriptome to inform placement of junction reads. However, no systematic evaluation has been performed to assess or quantify the benefits of incorporating reference transcriptome in mapping RNA-seq reads. Meanwhile, there exist multiple human genome annotation databases, including RefGene (RefSeq Gene), Ensembl, and the UCSC annotation database. The impact of the choice of an annotation on estimating gene expression remains insufficiently investigated.

In this chapter, we systematically characterized the impact of genome annotation choice on read mapping and gene quantification by analyzing a RNA-seq dataset generated by Illumina's Human Body Map 2.0 Project. The impact of a gene model on mapping of non-junction reads is different from junction reads. We demonstrated that the choice of a gene model has a dramatic effect on both gene quantification and differential analysis. Our research will help RNA-seq data analysts to make an informed choice of gene model in practical RNA-seq data analysis.

Keywords: RNA-seq, gene quantification, gene model, RefSeq, UCSC, Ensembl

1. Introduction

In recent years, RNA-seq has become a powerful approach for transcriptome profiling [1–3]. RNA-seq not only has considerable advantages for examining transcriptome fine structure—

for example, in the detection of novel transcripts, allele-specific expression, and alternative splicing—but also provides a far more precise measurement of levels of transcripts than that of other methods such as microarray [4–7]. Previously, we had performed a side by side comparison of RNA-seq and microarray in investigating T cell activation, and demonstrated that RNA-seq is superior in detecting low abundance transcripts, differentiating biologically critical isoforms, and allowing the identification of genetic variants [7]. In addition, RNA-seq has a much broader dynamic range than microarray, which allows for the detection of more differentially expressed genes with higher fold-change. Furthermore, RNA-seq avoids technical issues in microarray related to probe performance such as cross-hybridization, limited detection range of individual probes, and nonspecific hybridization [5–7]. Thus, RNA-seq delivers unbiased and unparalleled information about the transcriptome and gene expression. By RNA-seq technology, the Genotype-Tissue Expression (GTEx) project generated large amount of RNA sequence data to investigate the patterns of transcriptome variation across individuals and tissues [8–9]. An analysis of RNA sequencing data in the GTEx project from 1,641 samples across 43 tissues from 175 individuals revealed the landscape of gene expression across tissues, and catalogued thousands of tissue-specific expressed genes. These findings provide a systematic understanding of the heterogeneity among a diverse set of human tissues.

Current RNA-seq approaches use shotgun sequencing technologies such as Illumina, in which millions or even billions of short reads are generated from a randomly fragmented cDNA library. The first step and a major challenge in RNA-seq data analysis is the accurate mapping of sequencing reads to their genomic origins including the identification of splicing events. Despite of the fact that a large number of mapping algorithms have been developed for read mapping [10–13] and RNA-seq differential analysis [14–15] in recent years, however, accurate alignment of RNA-seq reads is a challenging and yet unsolved problem because of exon-exon spanning junction reads, relatively short read lengths and the ambiguity of multiple-mapping reads. Nowadays, many RNA-seq alignment tools, including GSNAP [16], OSA [17], STAR [18], MapSplice [19], and TopHat [20], use reference transcriptomes to inform the alignments of junction reads. In fact, this has become a common practice in RNA-seq data analysis. However, no systematic evaluation has been performed to assess and/or quantify the benefits of incorporating reference transcriptome in mapping RNA-seq reads.

The second aspect of transcriptome research is to quantify expression levels of genes, transcripts, and exons. Acquiring the transcriptome expression profile requires genomic elements to be defined in the context of the genome. Gene models are hypotheses about the structure of transcripts produced by a gene. Like all models, they may be correct, partly correct, or entirely wrong. In addition to RefGene [21], there are several other public human genome annotations, including UCSC Known Genes [22], Ensembl [23], AceView [24], Vega [25], and GENCODE [26]. Characteristics of these annotations differ because of variations in annotation strategies and information sources. RefSeq human gene models are well supported and broadly used in various studies. The UCSC Known Genes dataset is based on protein data from Swiss-Prot/TrEMBL (UniProt) and the associated mRNA data from GenBank, and serves as a foundation for the UCSC Genome Browser. Vega genes are manually curated transcripts produced by the

HAVANA group at the Welcome Trust Sanger Institute, and are merged into Ensembl. Ensembl genes contain both automated genome annotation and manual curation, while the gene set of GENCODE corresponds to Ensembl annotation since GENCODE version 3c (equivalent to Ensembl 56). AceView provides a comprehensive non-redundant curated representation of all available human cDNA sequences.

Although there are multiple genome annotations available, researchers need to choose a genome annotation (or gene model) while performing RNA-seq data analysis. However, the effect of genome annotation choice on downstream RNA-seq expression estimates is under-appreciated. Wu et al. [27] demonstrated that the selection of human genome annotation results in different gene expression estimates. Chen et al. [28] systematically compared the human annotations present in RefSeq, Ensembl, and AceView on diverse transcriptomic and genetic analyses. They found that the human gene annotations in the three databases are far from complete, although Ensembl and AceView annotate many more genes than RefSeq. In this paper, we performed a more comprehensive evaluation of different annotations on RNA-seq read mapping and gene quantification, including RefGene, UCSC, and Ensembl, and reported the main findings. More comprehensive reports were presented elsewhere [29–30].

2. Method

The Human Body Map 2.0 Project, using Illumina sequencing, generated RNA-seq data for 16 different human tissues (adipose, adrenal, brain, breast, colon, heart, kidney, leukocyte, liver, lung, lymph node, ovary, prostate, skeletal muscle, testis, and thyroid) and is accessible from ArrayExpress (accession number E-MTAB-513). We chose to analyze this public dataset because gene expression is tissue specific [9] and analyzing those 16 high-quality RNA-seq samples as a whole could result in less biased conclusions. The read length is 75 bp in all 16 samples, and there are 70 to 80 million reads for each sample (Supplementary Table 1 in [30]). To demonstrate the impact of read length on analysis results, we created a new dataset in which each original 75-bp long sequence read was trimmed to 50 bp. The same analysis protocol described below was applied to both datasets. In this chapter, we mainly presented the results for the read length of 75 bp, and for 50 bp reads, the detail reports could be found in [29–30]. We used the total number of reads mapped to each individual gene to represent expression level. For a given tissue sample, we analyzed the same RNA-seq dataset using the same aligner but with different gene models. The raw reads mapped to each gene across gene models can be compared directly.

The RefGene, Ensembl, and UCSC annotation files in GTF format were downloaded from the UCSC genome browser. Primary sequencing reads were first mapped to the reference transcriptome and the human reference genome GRCH37.3 using Omicsoft Sequence Aligner (OSA) [17]. Benchmarked with existing methods such as TopHat and others, OSA improves mapping speed 4–10 fold, with better sensitivity and fewer false positives.

As shown in Figure 1A, the mapping result of a sequence read is gene model dependent. For instance, read #2 can be uniquely mapped to gene #b if the gene model #A is chosen in the

mapping step. However, this read became a multiple-mapped read when either gene model #B or #C is used instead, because it can be mapped to genes #b and #e equally well. For a junction read with short overlap with an exon, it can be aligned to genome with the help of a reference transcriptome. Otherwise, it might fail to map to a genomic loci without the usage of a gene model when mapping reads.

Note that none of the gene annotation is 100% complete. As a result, for those RNA-seq reads not covered by a gene annotation, whether to use the gene model in the mapping step has no impact on their mappings. Therefore, to fairly assess the impact of a gene model on RNA-seq read mapping, only those reads covered by a gene model were used. In this study, we devised a two-stage mapping protocol (Figure 1B) for our evaluation. In Stage #1, all RNA-seq reads were mapped to a reference transcriptome only, and then only mapped reads are saved into a new FASTQ file. In Stage #2, all remaining reads were re-mapped to the reference genome with and without the use of a gene model, respectively. The role of a gene model in the mapping step was then quantified and characterized by comparing the mapping results in Stage #2. The two-stage mapping protocol is crucial for a fair evaluation. Otherwise, the impact of a gene annotation on RNA-seq data analysis will be diluted or underestimated.

The effect of a gene model on RNA-seq read mapping could be characterized and quantified by comparing the read mapping results in different mapping modes. We focused on those uniquely mapped reads with a gene annotation and divided them into four categories (Figure 1C) with respect to their mapping results without a gene annotation in the mapping step: (1) "Identical", the same alignment results were obtained regardless of the use of a gene model; (2) "Alternative", the read was still mapped but mapped differently. It turns out that the majority of reads in this category were junction reads. A junction read could be either mapped as a non-junction read, or remain mapped as a junction read but with different start, end, and splicing positions; (3) "Multiple", a uniquely mapped read became a multiple-mapped one. When a read is mapped across the whole reference genome, it is more likely to be mapped to multiple locations; and (4) "Unmapped", i.e., a read could not be mapped to anywhere in the genome without the assistance of a gene model. Nearly all reads in this category were junction reads.

The impact of a reference transcriptome on read mapping is dependent upon whether a sequence is a junction read and how much it overlaps with an exon. Therefore, we split all mapped reads into junction and non-junction ones based upon the CIGAR string in the SAM files. Then we compared the mapping difference with and without a reference transcriptome in the mapping step, and summarized the difference in each category shown in Figure 1C. Additional analysis was performed on "Alternative" and "Unmapped" junction reads to characterize the splicing patterns in terms of their overlaps with exons.

3. Results

3.1. The coverage of different gene annotations

The RNA-seq read mapping summaries for all 16 samples are shown in Figure 2. There are two different mapping modes. In the "transcriptome only" mapping mode, all RNA-seq reads

http://dx.doi.org/10.5772/61197

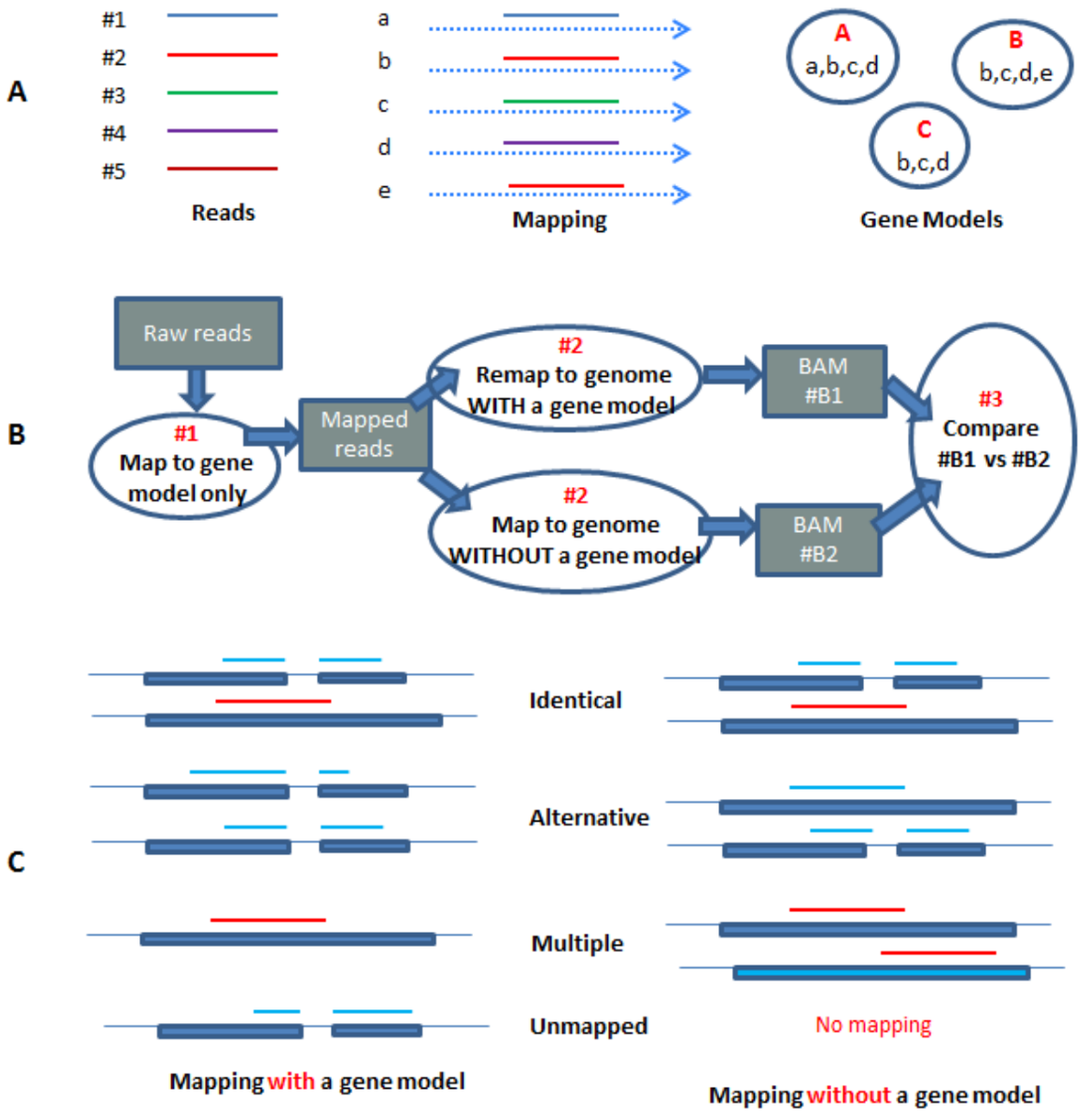

Figure 1. Analysis protocol. (A) The mapping result for a sequence read is gene model dependent; (B) "two-stage" mapping protocol: at Stage #1, all RNA-Seq reads are mapped to a reference transcriptome; at Stage #2, the mapped reads at Stage #1 are re-mapped to the genome with and without the use of a gene model, respectively; (C) the protocol for classifying uniquely mapped sequence reads into four categories, i.e., "Identical", "Alternative", "Multiple" and "Unmapped" (or Fail).

were mapped to a reference transcriptome only. If a read could not be mapped to a known gene region, it became unmapped, even though it could potentially be aligned to a genomic region without annotations. While in the "transcriptome + genome" mapping mode, reads were first mapped to a reference transcriptome, and then the unmapped ones were mapped to the reference genome. The impact of a reference transcriptome on the mapping of RNA-seq reads is attenuated in the "transcriptome + genome" mapping mode because every unmapped read has a second chance to be mapped to a genome.

In the "transcriptome only" mapping mode, more reads were mapped in Ensembl than in RefGene and/or UCSC. For each tissue type, the mapping rate was similar between RefGene and UCSC. The average read mapping rates across all 16 samples were 86%, 69%, and 70% for Ensembl, RefGene, and UCSC annotations, respectively. Short-read mapping is a basic step in RNA-seq data analyses, and to a certain extent, the percentage of reads mapped to a given transcriptome can roughly reflect the completeness or coverage of its annotated genes and transcripts. Thus, Ensembl annotation has much broader gene coverage than RefGene and UCSC. The patterns in "transcriptome + genome" mapping mode was different from those in "transcriptome only" mode (left panel on Figure 2). In the "transcriptome + genome" mapping mode, the average mapping rates for Ensembl, RefGene, and UCSC increased to 96.7%, 94.5%, and 94.6%, respectively, and the mapping rate difference among different gene models decreased. This large difference in the mapping rates between the two modes suggests the incompleteness of gene models: there are many reads that were mapped to the genomic regions without annotations.

Figure 2 shows that the read mapping percentage is also sample dependent, and this holds true for every gene model. For instance, only 52.5% of sequence reads in the heart were mapped to the RefGene model; while in leukocytes, 84.2% of reads could be mapped to RefGene. This mapping difference between heart and leukocyte results from, at least in part, the incompleteness of the RefGene annotation. As more expressed genes are annotated in a gene model, a higher percentage of reads will be mapped in the "transcriptome only" mapping mode.

In the "transcriptome only" mapping mode (the right panel in Figure 2), an average of 6.9%, 1.4%, and 1.8% of reads were multiple-mapped reads in Ensembl, RefGene, and UCSC gene models, respectively. The percentage of multiple-mapped reads in Ensembl is higher than in RefGene or UCSC. Usually, a more comprehensive annotation generally annotates more genes and isoforms, and thus, increases the possibility of ambiguous mappings. These ambiguous mappings directly translate to an increase in the percentage of non-uniquely mapped reads.

Different gene identifiers are used in different annotation databases; therefore, we mapped those database-specific identifiers into the unique HGNC gene symbols from the HUGO Gene Nomenclature Committee when comparing their gene quantification results across the different gene models originating from these databases. Considering that annotations are more or less incomplete in these databases, we only focused on common genes when comparing the results from different annotations. The Venn diagram in Figure 3 showed the overlap and intersection of RefGene, UCSC, and Ensembl annotations. Clearly RefGene has fewest unique genes, while more that 50% of genes in Ensembl are unique. In general, the different annotations have very high overlaps: 21,598 common genes are shared by all three gene annotations.

3.2. The impact of a gene model on RNA-seq read mapping

To evaluate the impact of a gene model on read mapping, the mapping summaries in Figure 2 were not sufficient. For instance, a read could be aligned differently with and without the assistance of a gene model in mapping, and in this scenario, the mapping summary could not tell such a difference. Thus, we compared the mapping details for every read, including start and end positions and splicing sites. For simplicity, in Stage #2, we focused on only uniquely

http://dx.doi.org/10.5772/61197

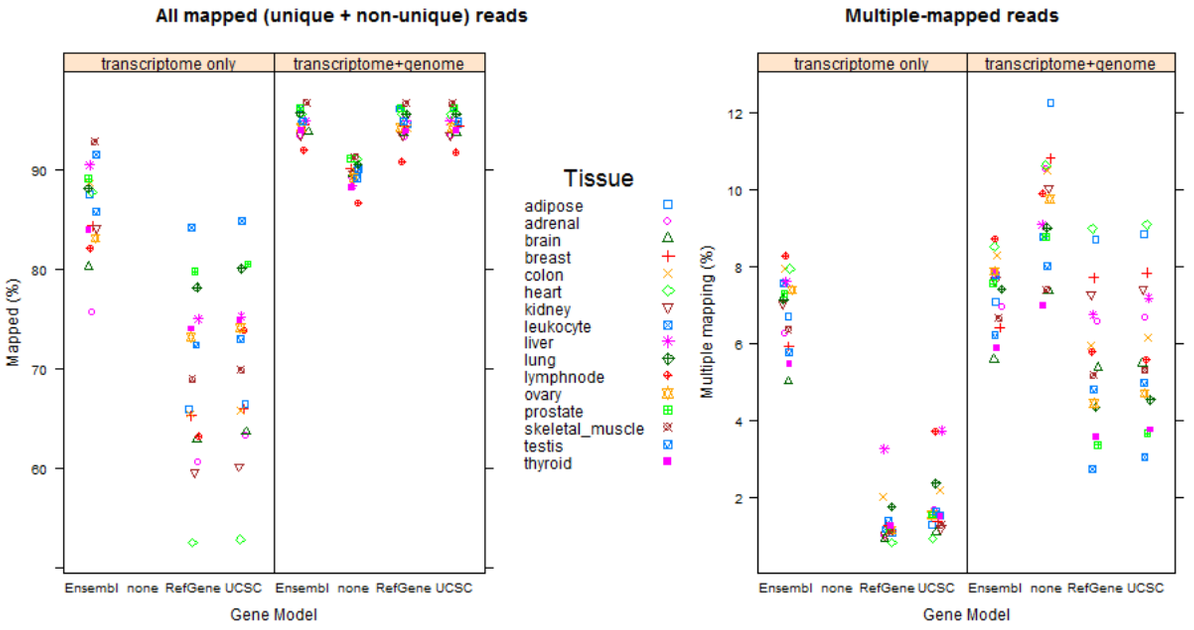

Figure 2. The read mapping summary for 16 tissue samples in the "transcriptome only" and "transcriptome+genome" mapping modes (note: read length = 75 bp). In the "transcriptome only" mode, more reads are mapped in Ensembl than in RefGene and UCSC (left panel), and more reads become multiple-mapped in Ensembl than in RefGene and UCSC (right panel). Note: the gene model "none" means the RNA-Seq reads are mapped to the reference genome directly without the use of a gene model.

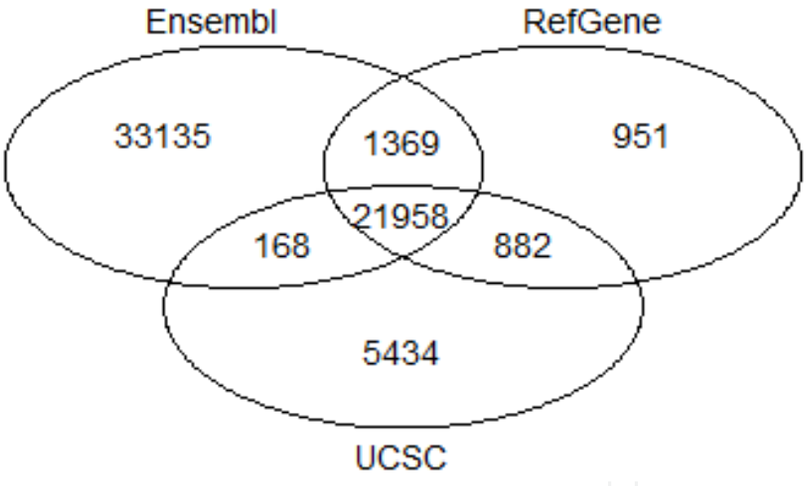

Figure 3. The overlap and intersection among RefGene, UCSC, and Ensembl annotations.

mapped reads in the "transcriptome only" mapping mode. A uniquely mapped read could be classified into four categories (Figure 1C) with respect to its corresponding mapping information without a gene model: (1) "Identical"—remaining mapped to the same genomic region; (2) "Alternative"—still uniquely mapped but differently; (3) "Multiple"—mapped to more locations; and (4) "Unmapped". The detailed evaluation results are summarized in Figure 4 (read length = 75 bp).

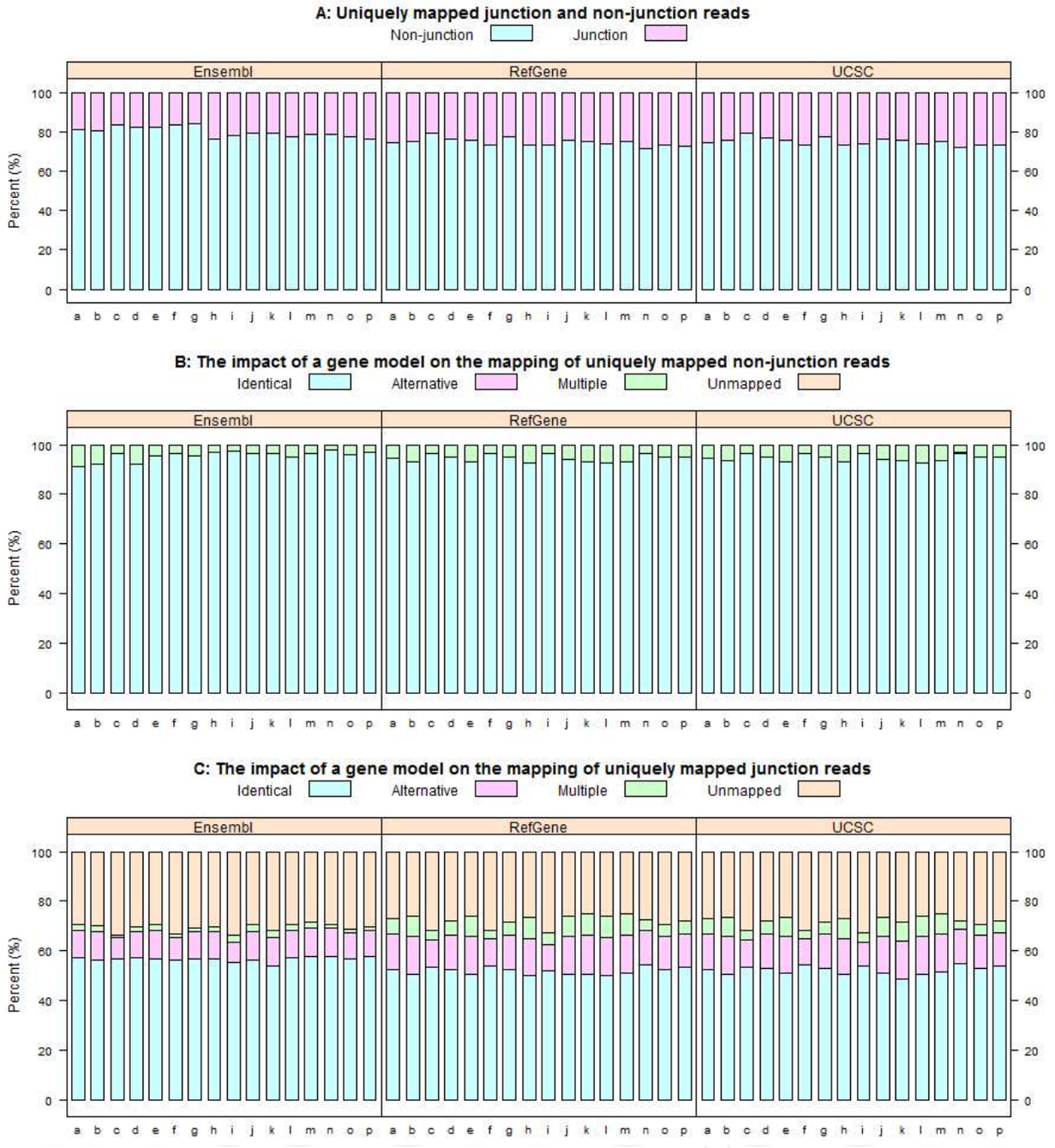

Figure 4. The impact of a gene model on RNA-Seq read mapping (read length = 75 bp). (A) Composition of mapped reads; (B) effect on mapping of non-junctions reads; (C) effect on mapping of junctions reads. (Note: The 16 tissue sample names are denoted as follows: a: adipose; b: adrenal, c: brain; d: breast; e: colon; f: heart; g: kidney; h: leukocyte; i: liver; j: lung; k: lymph node; l: ovary; m: prostate; n: skeletal muscle; o: testis; and p: thyroid.)

In Figure 4A, we divided uniquely mapped reads into two classes, i.e., non-junction reads and junction reads, and investigated the impact of a gene model on their mapping. Accordingly to Figure 4A, approximately 23% of mapped reads were junction reads, and the remaining 77% were non-junction reads. For non-junction reads (see Figure 4B), 95% remained mapped to exactly the same genomic location regardless of the use of a gene model. Without a gene model,

http://dx.doi.org/10.5772/61197

3% to 9% of non-junctions reads became multiple mapped reads. However, it is very rare for a non-junction read to become unmapped or alternatively mapped. In contrast, the mapping of junction reads was strongly impacted by the gene models (see Figure 4C). Without using a gene model, an average of 53% of junction reads remained mapped to the same genomic regions, 30% failed to map to any genomic region, and 10–15% of them mapped alternatively. Such alternative mappings are generally inferior compared to their corresponding mapping results using a gene model [29]. Similar to non-junction reads, an average of 5% of junction reads were mapped to more than one location without using a gene model. As shown in Figure 4C, more uniquely-mapped junction reads became multiple mapped reads in RefGene and/or UCSC than in Ensembl when the sequence reads were aligned to the reference genome without the use of gene models.

As we demonstrated, a gene model mainly affects the alignment of junction reads, but has little impact on non-junction reads. On average, 23% of reads in our samples were junction reads, and usually about one third of them failed to be mapped without the use of a gene model. Therefore, it is expected that when the read length is 75 bp, ~6% (23% * 0.33) of the mapped reads become unmapped without the use of a gene model. The percentage is expected to be higher when the read length is longer since a long read is more likely to span two or more exons.

3.3. The splicing patterns for "Identical", "Alternative", and "Unmapped" reads

As concluded above, a reference transcriptome mainly affects the mapping of junction reads. One interesting question is what kind of junction reads tend to be mapped identically, alternatively, or unmapped. In order to characterize the splicing patterns, we focus on only two-exon junction reads that are uniquely mapped when the RefGene annotation is used. For every junction read, we calculate the number of overlapping nucleotide bases with its left exon (OL) and right exons (OR), respectively. Then the minimum of OL and OR is chosen for histogram analysis (Figure 5). Only the results for lung, liver, kidney, and heart samples are shown in Figure 5, and for the rest of 12 samples, the patterns were very similar to those in Figure 5 (data not shown). Since the full read length is 75 bp long, the MOE (Minimum Overlap with an Exon, MOE = min(OL,OR)) ranges from 1 to 37 for any junction read.

For "Identical" junction reads, the typical MOE ranges from 15 to 37, and the frequency drops to nearly 0 when MOE is less than 10 (left panels in Figure 5). For "Alternative" junction reads, the most dominant MOE is 1 (middle panels in Figure 5), representing an average of one-third of cases. In general, those "Alternative" reads have very small MOE. For those junction reads with MOE of 1, 2, and 3, it is virtually impossible to map them 'correctly' without the prior knowledge on transcripts. The MOE for "Unmapped" reads has a much broader range with peaks from 4 to 12 (right panels in Figure 5). In order to map a junction read without a reference transcriptome, the read should have sufficient overlaps with exons at both ends. The majority of "Identical" reads meet this requirement (left panels in Figure 5). However, if the overlap with one end is too short, let's say 1 or 2 nucleotide bases, this read will be more likely mapped to only a single exon with the remaining couple of bases mapping to the intron region adjacent to that exon (middle panels in Figure 5). Otherwise, such junction reads become either

unmapped or mapped to different genomic regions as non-junction reads if the overlap is something between (right panels in Figure 5).

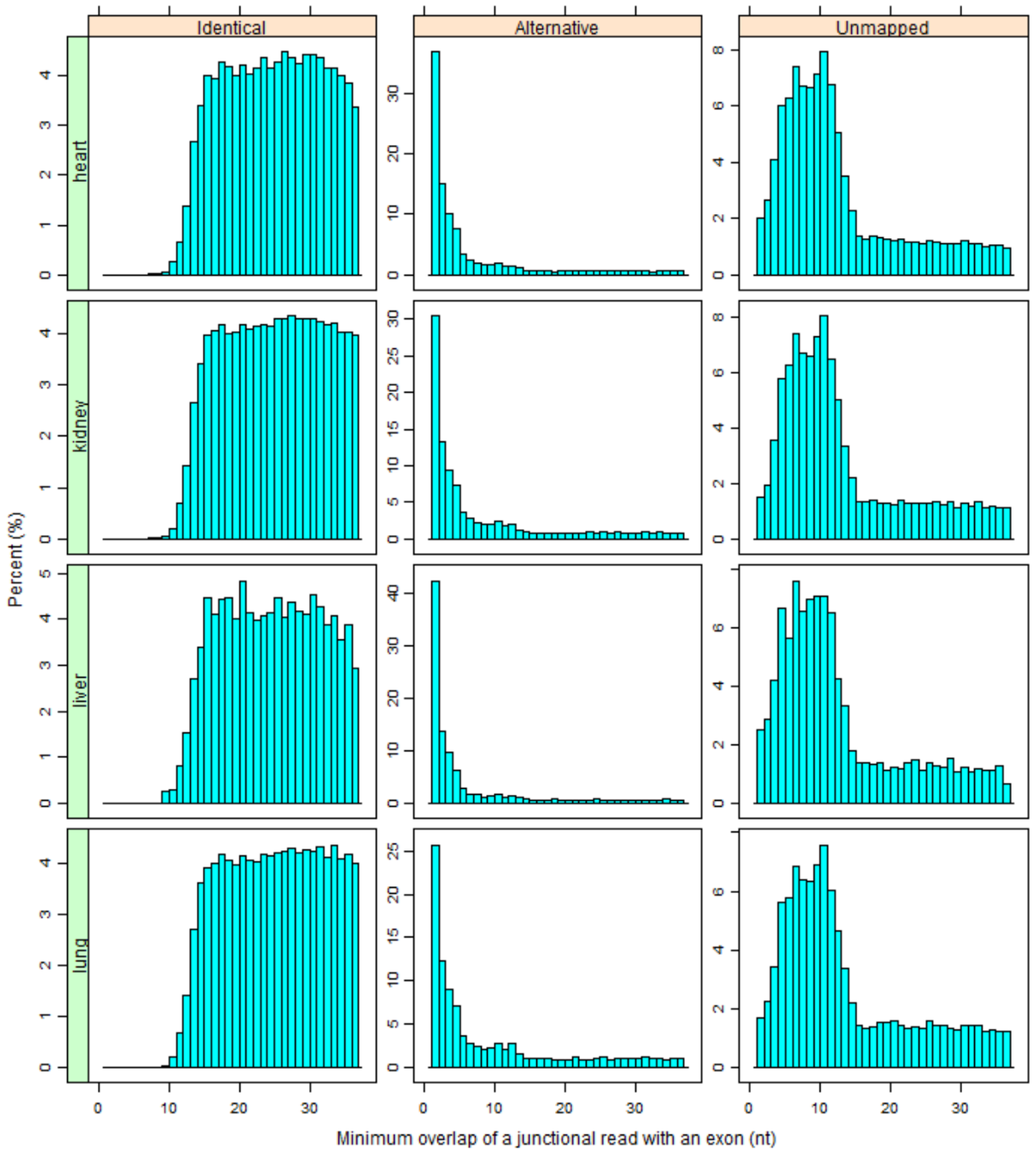

Figure 5. The splicing patterns and distribution of MOE (Minimum Overlap with an Exon) for junction reads. The typical MOE for "Identical" junction reads ranges from 15 to 37. For "Alternative" junction reads, the most dominant MOE is 1, representing an average of one-third of cases. In contrast, the MOE for "Unmapped" reads has a much broader range with peaks from 4 to 12. Note the scale for y-axis is not uniform.

3.4. Comparison of the mappings of "Alternative" reads

Since "Alternative" reads remain mapped but differently, we are more interested in the mapping difference in detail and the main reasons for alternative mapping. A typical example of "Alternative" reads is shown in Figure 6, in which 19 unique junction reads are nearly perfectly mapped to gene HSP90AB1 when RefGene is used in the mapping step. Without a reference transcriptome, four junction reads indicated by the red arrow remain mapped to the same gene HSP90AB1 but as non-junction reads with mismatches at one end. A few bases previously mapped to another exon are now mapped to the intron region. The remaining 15 junction reads are aligned to pseudogene gene HSP90AP3P as non-junction reads instead. The comparison reveals that the original mappings to HSP90AB1 for those 15 reads are nearly perfect, while they all have more mismatches when mapped to HSP90AP3P. Clearly, the alternative mapping for those junction reads in Figure 6 is getting worse without a reference transcriptome. In a sense, those 15 junction reads indicated by the blue arrow in Figure 6A are "forced" to be mapped to a different genomic region without the help of reference transcriptome.

"Alternative" junction reads are also likely to remain mapped to the same start and end positions but spliced differently. Two cases in point are shown in Figure 7. For those junction reads mapped to gene TCEA3 with and without RefGene model, both mappings are equally well in terms of alignment scores and gaps between exons. So there is no way to tell which one is right without the assistance of reference transcriptome. Likewise, the mappings of junction reads in gene FBXL3 are also equally well regardless of the usage of RefGene model. Despite the minor difference in splicing sites, the read mapped with RefGene model is considered as fully compatible to a known gene, and thus is counted in gene quantification. Collectively, the examples in Figure 6 and 7 illustrate the important role of a gene annotation in proper alignment of junction reads.

3.5. The impact of gene model choice on gene quantification

To investigate the impact of different gene models on gene quantification results, we focused on the set of 21,598 common genes (Figure 3). The overall correlation between RefGene and Ensembl was shown in Figure 8. Both x and y-axes represented log2(count+1). For all genes, 1 was added to the counts to avoid a logarithmic error for those genes with zero counts. Ideally, we should get identical counts of mapped reads for all common genes, regardless of the choice of a gene model; however, this was clearly not the case. Although the majority of genes had highly consistent or nearly identical expression levels, there were a significant number of genes whose quantification results were dramatically affected by the choice of a gene model. As shown in Figure 8, there were many genes for which the number of reads mapped to them was 0 in one gene model, but many in others.

To quantify the concordance between RefGene and Ensembl annotations, we first calculated the ratio of mapped read for each gene. For a given gene, we defined the raw read counts in RefGene and Ensembl annotations as #C1 and #C2, respectively. To prevent division by 0, 1 was added to all raw read counts before the ratios were calculated. The adjusted counts were denoted as #C1′ (=#C1+1) and #C2′ (=#C2+1), respectively. The ratio was calculated as

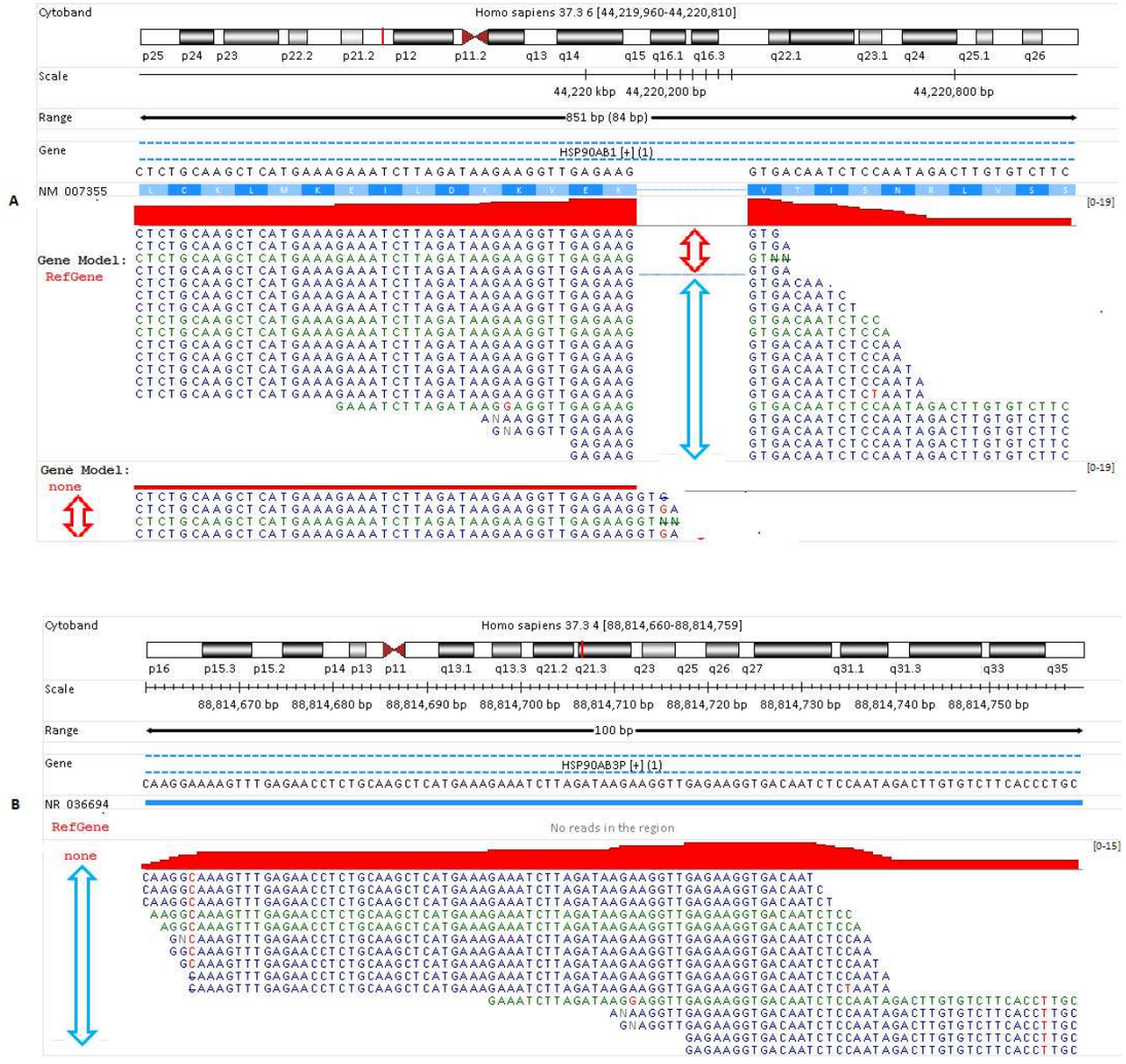

Figure 6. The impact of a reference transcripotome on the mapping of junction reads in gene HSP90AB1. (A) When RefGene is used, 19 unique junction reads are mapped to gene HSP90AB1 nearly perfectly. Four junction reads become non-junction ones with a few bases mapped to the intron region with mismatches without the usage of the RefGene model; (B) The remaining 15 reads (indicated by the blue arrow) are alternatively aligned to gene HSP90AP3P as non-junction reads without the assistance of RefGene annotation. Note the reads colored in blue are mapped to "+" strand, and colored in green when mapped to "-" strand. The mismatched nucleotide bases are colored in red.

Max(#C1',#C2')/Min(#C1',#C2'). Therefore the calculated ratio was always equal or greater than 1. The distribution of ratios was summarized in Table 1 (read length = 75 bp). Among the 21,958 common genes, about 20% of genes had no expression at all in both annotations. Identical counts were obtained for only 16.3% of genes. Approximately 28.1% of genes' expression levels differed by 5% or higher, and among them, 9.3% of genes (equivalent to 2,038) differed by 50% or greater. As shown in Table 1 and Figure 8, the choice of a gene model had a large impact on gene quantification. Compared with Ensembl, UCSC had a much better

http://dx.doi.org/10.5772/61197

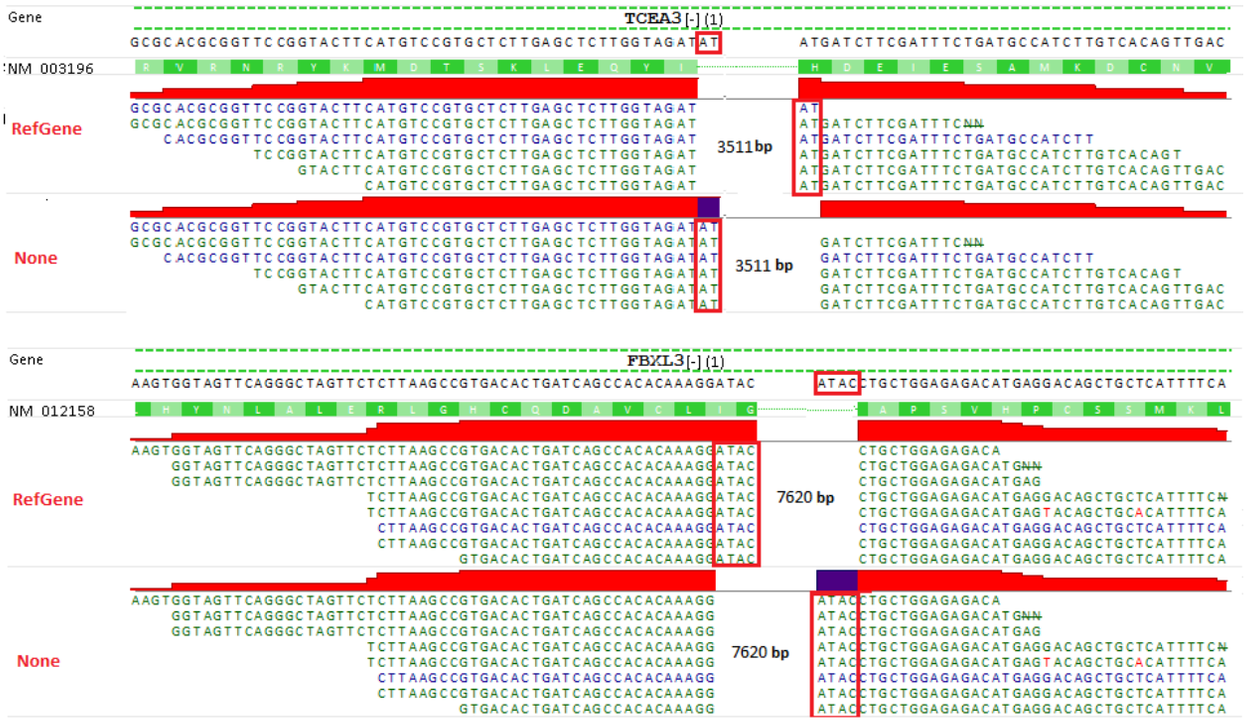

Figure 7. Alternative splicing with and without the use of RefGene annotation. All junction reads are still mapped to the same gene with the same start/end positions and intron size regardless of gene model, but are spliced differently.

concordance with RefGene, in terms of the gene quantification results [30]. 38.3% of genes had identical read counts, much higher than the 16.3% between Ensembl and RefGene. The percentage of genes with expression levels differing by 5% or more was only 11.3%, which was much less than the corresponding 28% between Ensembl and RefGene. Furthermore, only 3.24% of genes differed by 50% or greater, which was lower than the 9.3% between Ensembl and RefGene.

Why does the choice of a gene model have so dramatic an effect on gene quantification? If the gene definition is the same among different annotations, we expect the identical number of reads mapped to a given gene. Unfortunately, the gene definition varies from annotation to annotationm and can differ singnificantly for some genes. PIK3CA is a good example. The PIK3CA gene definition in both Ensembl and RefGene, and the mapping profile of RNA-seq reads were shown in Figure 9. In the liver sample, there were 1,094 reads mapped to PIK3CA in Ensembl annotation, while only 492 reads were mapped in RefGene. Clearly, the big difference in gene definition gives rise to the observed discrepancy in quantification. In Ensembl, there are three isoforms for PIK3CA, and the longest isoform is ENST00000263967. The total length of this transcript is 9,653 bp, comprising 21 exons, with a very long exon #21 (6,000 bp, chr3: 178,951,882-178,957,881). In RefGene, PIK3CA has only one transcript named NM_006218. This transcript is 3,909 bp long with a very short exon #21 (only 616 bp, located at chr 3:178,951,882-178,952,497). The definition of the PIK3CA gene in Ensembl seems more accurate than the one in RefGene, based upon the mapping profile of the sequence reads.

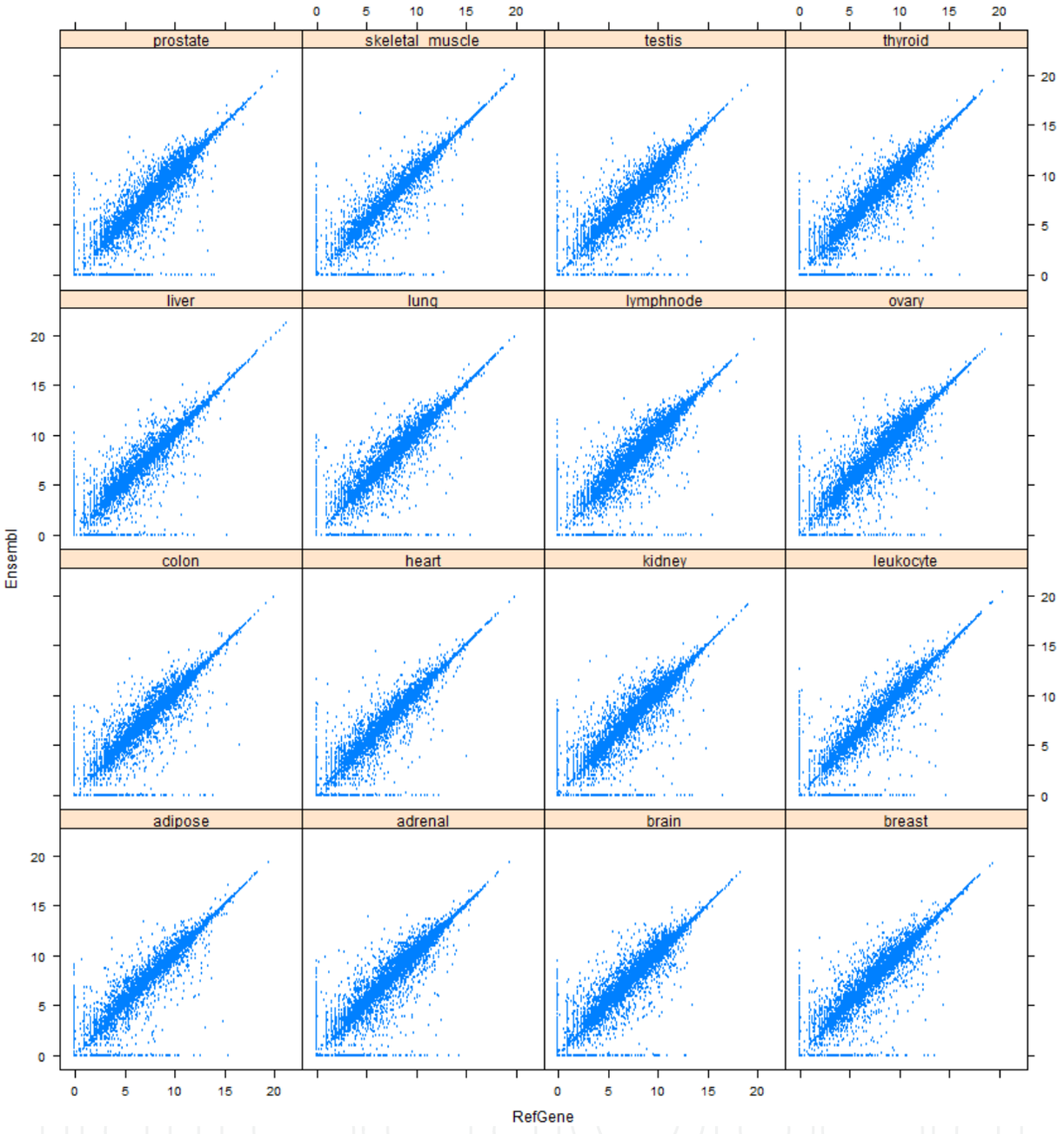

Figure 8. The correlation of gene quantification results between RefGene and Ensembl. Note both x and y-axes represent Log2(count + 1).

3.6. The effect of gene models on differential analysis

Generally, RNA-seq differential analysis requires biological replicates. However, we analyzed 16 different single tissue samples. To demonstrate the effect of gene models on differential analysis, the fold changes between heart and liver samples were calculated using RefGene and Ensembl annotations. The correlation of the calculated Log2Ratio (liver/heart) was depicted in Figure 10. The graph should show a perfect diagonal line if the choice of a gene model has no effect on differential analysis. Although the majority of genes have highly consistent or

Sample	No Expr	Same	1.05	1.10	1.20	1.50	2	5	10	100
adipose	19.97	16.53	26.16	19.64	14.51	8.81	5.65	1.96	0.94	0.16
adrenal	16.92	14.04	36.18	27.09	19.07	11.28	7.14	2.45	1.24	0.24
brain	16.79	15.22	32.94	24.91	17.95	10.78	6.73	2.29	1.08	0.20
breast	18.04	15.22	29.63	22.21	16.06	9.80	6.52	2.38	1.19	0.20
colon	20.50	17.41	25.85	19.43	14.30	8.95	6.10	2.30	1.17	0.19
heart	21.23	16.43	26.39	20.10	14.39	8.88	5.47	1.73	0.82	0.19
kidney	18.86	16.08	28.88	21.50	15.51	9.55	6.40	2.55	1.30	0.26
leukocyte	29.53	17.37	20.03	15.29	11.62	7.58	5.37	2.47	1.33	0.26
liver	24.60	19.16	23.20	17.43	12.84	8.24	5.42	2.00	1.02	0.15
lung	19.65	16.46	29.22	21.35	15.07	9.09	6.15	2.61	1.43	0.24
lymph node	20.94	16.79	31.74	24.16	17.21	10.26	6.65	2.69	1.44	0.24
ovary	16.90	13.42	31.46	23.30	16.72	10.23	6.63	2.31	1.13	0.20
prostate	18.21	16.29	28.33	21.14	15.17	9.43	6.51	2.49	1.27	0.23
skeletal muscle	29.60	23.48	18.65	14.40	10.73	6.88	4.81	2.34	1.39	0.21
testis	10.15	13.35	31.35	22.57	15.84	9.35	5.92	2.08	1.05	0.28
thyroid	17.41	14.25	30.08	22.23	15.88	9.39	5.88	1.97	1.03	0.24
Average	19.96	16.34	28.13	21.05	15.18	9.28	6.09	2.29	1.18	0.22

Note: Column "No Expr" represents the percentage of genes that do not express at all in both annotations. Column "Same" denotes the percentage of genes that have the same number of reads mapped to them in both gene models. The number in each cell after the column "Same" corresponds to the percentage of genes whose ratio is equal or greater than the threshold represented by the number.

Table 1. The distribution of the ratio of read counts between RefGene and Ensembl annotations (read length = 75 bp).

comparable expression changes, there are a number of genes whose ratios are dramatically affected by the choice of a gene model. Interestingly, some genes have a very high fold change in one gene model, but no change at all in another gene model. Evidently, the choice of a gene model has an effect on the downstream differential expression analysis, in addition to gene quantification.

4. Discussions

4.1. The effect of a gene model on read mapping is read length dependent

We performed the same analyses of the dataset with a 50-bp read length, and the results were detailed in [30]. Intuitively, the shorter a read, the more likely it is to map to multiple locations. As a result, the percentage of uniquely mapped reads decreases, and the percentage of

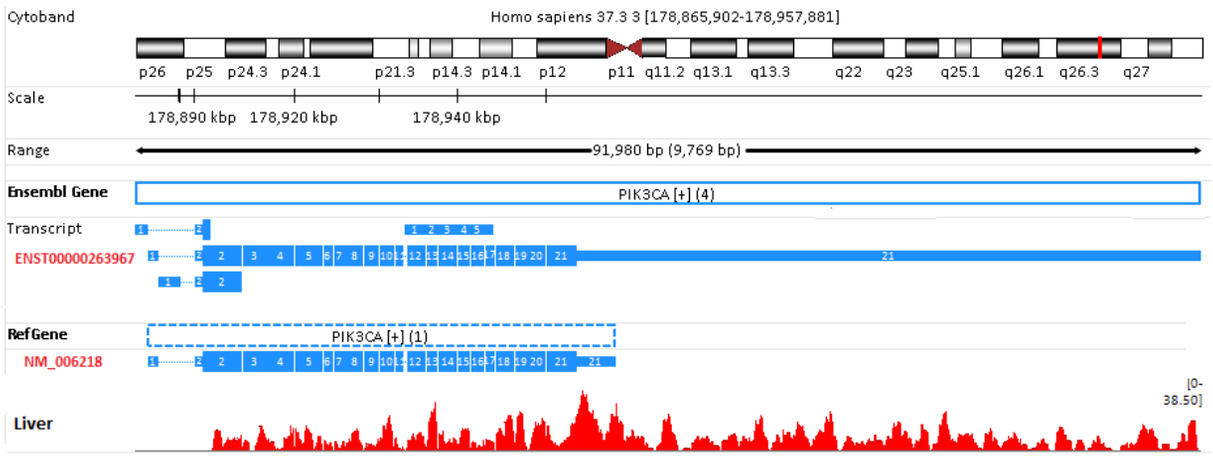

Figure 9. The different gene definitions for PIK3CA give rise to differences in gene quantification. PIK3CA in the Ensembl annotation is much longer than its definition in RefGene, explaining why there are 1,094 reads mapped to PIK3CA in Ensembl, while only 492 reads are mapped in RefGene.

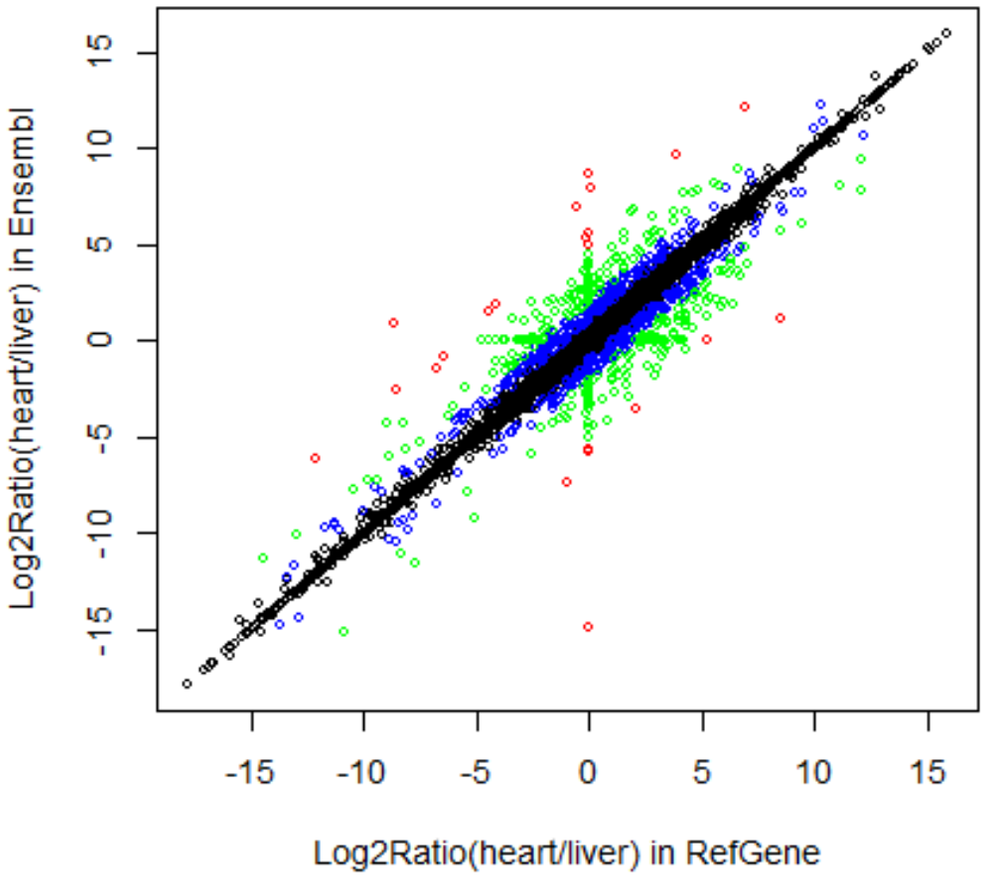

Figure 10. The correlation of the calculated Log2Ratio (heart/liver) between RefGene and Ensembl. The green, blue, and red points indicate corresponding absolute difference between the two Log2Ratios that were greater than 1, 2, or 5, respectively. Although the majority of genes have highly consistent expression changes, there are many genes that are remarkably affected by the choice of different gene models.

multiple-mapping reads increases. No matter which gene model was used in the mapping step, this observation held true. Thus, the mapping fidelity for a sequence read increases with its length, and this is especially true for junction reads. As demonstrated in Figure 4, when the read length was 75 bp, an average of 53% of junction reads remained mapped to the same genomic regions no matter whether a gene annotation was used. However, this percentage dropped to 42% when the read length was 50 bp long [30]. Thus, the effect of a gene model on the mapping of junction reads is significantly influenced by read length.

In the meantime, the relative abundance of junction reads is heavily determined by read length as well. According to Figure 4, on average, roughly 23% of sequence reads were junction reads when the read length was 75 bp. This percentage dropped to 16% when the read length was 50 bp [30]. This is explained by the fact that the longer the read, the more likely that it spans more than one exon. As sequencing technology evolves, the read length will become longer and longer. Consequently, more junction reads will be generated by short-gun sequencing technologies. Therefore, the need to incorporate genome annotation in the read mapping process will greatly increase.

4.2. The incompleteness and inaccuracy in gene annotation

Pyrkosz et al. [31] have explored the issue of "RNA-Seq mapping errors when using incomplete reference transcriptome" in detail. They used simulated reads generated from real transcriptomes to determine the accuracy of read mapping, and measured the error resulting from using an incomplete transcriptome. When 10% increments of the chicken reference transcriptome are missing, the true positive rate decreases by approximately 6–8%, while the false positive rate remains relatively constant until the reference is more than 50% incomplete. The number of false positives grows as the reference becomes increasingly incomplete. For model organisms such as human and mouse, their transcriptome models are relatively more complete compared to non-model organisms. Admittedly, RefGene, UCSC, and Ensembl are all not 100% complete and accurate, though the qualities in their annotations are constantly improving. For transcriptome-guided mapping of RNA-Seq reads, the more complete and accurate the transcriptome, the better. In addition, Seok et al. [32] have demonstrated that incorporating transcript annotations from reference transcriptome significantly improved the de novo reconstruction of novel transcripts from short sequencing reads for transcriptome research. The prior knowledge helped to define exon boundaries and fill in the transcript regions not covered by sequencing data. As a result, the reconstructed transcripts were much longer than those from de novo approaches that assume no prior knowledge.

4.3. The impact of gene annotation on variant effect prediction

The choice of a gene annotation has a big impact not only on RNA-seq data analysis, but also on variant effect prediction [33–34]. Variant annotation is a crucial step in the analysis of genome sequencing data. Functional annotation results can have a strong influence on the ultimate conclusions of disease studies. Incorrect or incomplete annotations can cause researchers both to overlook potentially disease-relevant DNA variants and to dilute interesting variants in a pool of false positives.

McCarthy et al. [33] recently used the software ANNOVAR [35] to quantify the extent of differences in annotation of 80 million variants from a whole-genome sequencing study with the RefSeq and Ensembl transcript sets as the basis for variant annotation. They demonstrated the large differences in prediction of loss-of-function (LoF) variation when RefSeq and Ensembl transcripts are used for annotation, highlighting the importance of the reference transcripts on which variant functional annotation is based. Choice of transcript set can have a large effect on the ultimate variant annotations obtained in a whole-genome sequencing study.

Frankish et al. [34] performed a detailed analysis of the similarities and differences between the gene and transcript annotation in the Gencode (v21) and RefSeq (Release 67) genesets in order to identify the similarities and differences between the transcripts, exons and the CDSs they encode. They demonstrated that the Gencode Comprehensive set is richer in alternative splicing, novel CDSs, and novel exons and has higher genomic coverage than RefSeq, while the Gencode Basic set is very similar to RefSeq. They presented evidence that the reference transcripts selected for variant functional annotation do have a large effect on variant annotation.

4.4. Which genome annotation to choose for gene quantification?

In practice, there is no simple answer to this question, and it depends on the purpose of the analysis. In this chapter, we compared the gene quantification results when RefGene and Ensembl annotations were used. Among 21,958 common genes, the expressions of 2,038 genes (i.e., 9.3%) differed by 50% or more when choosing one annotation over the other. Such a large difference frequently results from the gene definition differences in the annotations. Some genes with the same HUGO symbol in different gene models can be defined as completely different genomic regions. When choosing an annotation database, researchers should keep in mind that no annotation is perfect and some gene annotations might be inaccurate or entirely wrong.

Wu et al. [27] suggested that when conducting research that emphasizes reproducible and robust gene expression estimates, a less complex genome annotation, such as RefGene, might be preferred. When conducting more exploratory research, a more complex genome annotation, such as Ensembl, should be chosen. Based upon our experience of RNA-seq data analysis, we recommend using RefGene annotation if RNA-seq is used as a replacement for a microarray in transcriptome profiling. For human samples, Affymetrix GeneChip HT HG-U133+ PM arrays are one of the most popular microarray platforms for transcriptome profiling, and the genes covered by this chip overlap with RefGene very well, according to Zhao et al. [7]. Despite the fact that Ensembl R74 contains 63,677 annotated gene entries, only 22,810 entries (roughly one-third) correspond to protein coding genes. There are 17,057 entries representing various types of RNAs, including rRNA (566), snoRNA (1,549), snRNA (2,067), miRNA (3,361), misc_RNA (2,174), and lincRNA (7,340). There are 15,583 pseudogenes in Ensembl R74. For most RNA-seq sequencing projects, only mRNAs are presumably enriched and sequenced, and there is no point in mapping sequence reads to RNAs such as miRNAs or lincRNAs. Ensembl R74 contains 819 processed transcripts that were generated by reverse transcription of an mRNA transcript with subsequent reintegration of the cDNA into the genome, and are

usually not actively expressed. In this scenario, a read truly originating from an active mRNA can be mapped to a processed transcript equally well or mapped to the processed transcript only, which is especially true for junction reads. Consequently, the true expression for the corresponding mRNA may be underestimated. Another downside of using a larger annotation database is calculation of adjusted P values, because the adjustment of the raw P value to allow for multiple testing is mainly determined by the number of genes in the model. If genes of interest are defined inconsistently across different annotations, it is recommended that an RNA-seq dataset is analyzed using different gene models.

5. Conclusions

RNA-seq has become increasingly popular in transcriptome profiling. Acquiring transcriptome expression profiles requires researchers to choose a genome annotation for RNA-seq data analysis. In this chapter, we assessed the impact of gene models on the mapping of junction and non-junction reads, characterized the splicing patterns for junction reads, and compared the impact of genome annotation choice on gene quantification and differential analysis. To fairly assess the impact of a gene model on RNA-seq read mapping, we devised a two-stage mapping protocol, in which sequence reads that could not be mapped to a reference transcriptome were filtered out, and the remaining reads were mapped to the reference genome with and without the use of a gene model in the mapping step. Our protocol ensured that only those reads compatible with a gene model were used to evaluate the role of a genome annotation in RNA-seq data analysis.

Ensembl annotates more genes than RefGene and UCSC. On average, 95% of non-junction reads were mapped to exactly the same genomic location without the use of a gene model. However, only an average of 53% junction reads remained mapped to the same genomic regions. About 30% of junction reads failed to be mapped without the assistance of a gene model, while 10–15% mapped alternatively. It is also demonstrated that the effect of a gene model on the mapping of sequence reads is significantly influenced by read length. The mapping fidelity for a sequence read increases with its length. When the read length was reduced from 75 bp to 50 bp, the percentage of junction reads that remained mapped to the same genomic regions dropped from 53% to 42% without the assistance of gene annotation.

There are 21,958 common genes among RefGene, Ensembl, and UCSC annotations. Using the dataset with the read length of 75 bp, we compared the gene quantification results in RefGene and Ensembl annotations, and obtained identical counts for an average of 16.3% (about one-sixth) of genes. Twenty percent of genes are not expressed, and thus have zero counts in both annotations. About 28.1% of genes showed expression levels that differed by 5% or higher; of these, the relative expression levels for 9.3% of genes (equivalent to 2,038) differed by 50% or greater. The case studies revealed that the difference in gene definitions caused the observed inconsistency in gene quantification.

In this chapter, we demonstrate that the choice of a gene model not only has a dramatic effect on both gene quantification and differential analysis, but also has a strong influence on variant

effect prediction and functional annotation. Our research will help RNA-seq data analysts to make an informed choice of gene model in practical RNA-seq data analysis.

Author details

Shanrong Zhao* and Baohong Zhang

*Address all correspondence to: Shanrong.Zhao@pfizer.com

Clinical Genetics and Bioinformatics, Pfizer Worldwide Research & Development, Cambridge, MA, USA

References

[1] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods. 2008;5(7):621-8.

[2] Wang Z, Gerstein M, Snyder M. RNA-seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1):57-63.

[3] Mutz KO, Heilkenbrinker A, Lönne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol. 2013;24(1):22-30.

[4] McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, et al. RNA-seq: technical variability and sampling. BMC Genomics. 2011;12:293.

[5] Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic. 2009;8(3):174-83.

[6] Malone J, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011;9:34.

[7] Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.

[8] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):648-60.

[9] Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660-5.

[10] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8(6): 469-77.

[11] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013;10(12):1185-91.

[12] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91.

[13] Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-seq. PLoS One. 2013;8(10):e76935.

[14] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.

[15] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1): 139-40.

[16] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873-81.

[17] Hu J, Ge H, Newman M, Liu K. OSA: A fast and accurate alignment tool for RNA-seq. Bioinformatics. 2012;28(14):1933-4.

[18] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

[19] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010;38(18):e178.

[20] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.

[21] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(Database):D61-5.

[22] Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics. 2006;22(9):1036-46.

[23] Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res. 2014;42(Database issue):D749-55.

[24] Thierry-Mieg D, Thierry-Mieg J. AceView: A comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 2006;7 Suppl 1:1-14.

[25] Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. The vertebrate genome annotation (Vega) database. Nucleic Acids Res. 2008;36(Database):D753-60.

[26] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;22(9):1760-74.

[27] Wu P-Y, Phan JH, Wang MD. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics. 2013;14 Suppl 11:S8.

[28] Chen G, Wang C, Shi L, Qu X, Chen J, Yang J, et al. Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. RNA. 2013;19(4):479-89.

[29] Zhao S. Assessment of the impact of using a reference transcriptome in mapping short RNA-seq reads. PLoS One. 2014;9(7):e101374.

[30] Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16:97.

[31] Pyrkosz AB, Cheng H, Brown CT. RNA-Seq mapping errors when using incomplete reference transcriptomes of vertebrates. arXiv.org. 2013;arXiv:1303.2411v1.

[32] Seok J, Xu W, Jiang H, Davis RW, Xiao W. Knowledge-based reconstruction of mRNA transcripts with short sequencing reads for transcriptome research. PLoS ONE. 2012;7:e31440.

[33] McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 2014;6(3):26.

[34] Frankish A, Uszczynska B, Ritchie GR, Gonzalez JM, Pervouchine D, Petryszak R, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015;16 (Suppl 8):S2.

[35] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.