RNA-seq Data Employed in The Sequencing Quality Control (SEQC) Project

Yang Liao and Wei Shi

December 3, 2014

1 Introduction

This vignette briefly describes the content in the seqc package.

The seqc package provides a series of data frames derived from the SEQC project between 2011 to 2014. Three types of data frames are included in this package: RNA-seq read count tables, junction tables and an additional gene intensity table generated by using the TaqMan RT-PCR technology. All the data frames are ready to use in the R environment once the library is loaded.

```
> library(seqc)
> options(width=110, digits=2)
> ls(2)
```

```
 [1] "ILM_aceview_gene_AGR" "ILM_aceview_gene_BGI" "ILM_aceview_gene_CNL" "ILM_aceview_gene_COH"
 [5] "ILM_aceview_gene_MAY" "ILM_aceview_gene_NVS" "ILM_junction_AGR_A"   "ILM_junction_AGR_B"
 [9] "ILM_junction_AGR_C"   "ILM_junction_AGR_D"   "ILM_junction_BGI_A"   "ILM_junction_BGI_B"
[13] "ILM_junction_BGI_C"   "ILM_junction_BGI_D"   "ILM_junction_CNL_A"   "ILM_junction_CNL_B"
[17] "ILM_junction_CNL_C"   "ILM_junction_CNL_D"   "ILM_junction_COH_A"   "ILM_junction_COH_B"
[21] "ILM_junction_COH_C"   "ILM_junction_COH_D"   "ILM_junction_MAY_A"   "ILM_junction_MAY_B"
[25] "ILM_junction_MAY_C"   "ILM_junction_MAY_D"   "ILM_junction_NVS_A"   "ILM_junction_NVS_B"
[29] "ILM_junction_NVS_C"   "ILM_junction_NVS_D"   "ILM_refseq_gene_AGR"  "ILM_refseq_gene_BGI"
[33] "ILM_refseq_gene_CNL"  "ILM_refseq_gene_COH"  "ILM_refseq_gene_MAY"  "ILM_refseq_gene_NVS"
[37] "LIF_aceview_gene_LIV" "LIF_aceview_gene_NWU" "LIF_aceview_gene_PSU" "LIF_aceview_gene_SQW"
[41] "LIF_junction_LIV_A"   "LIF_junction_LIV_B"   "LIF_junction_LIV_C"   "LIF_junction_LIV_D"
[45] "LIF_junction_NWU_A"   "LIF_junction_NWU_B"   "LIF_junction_NWU_C"   "LIF_junction_NWU_D"
[49] "LIF_junction_PSU_A"   "LIF_junction_PSU_B"   "LIF_junction_PSU_C"   "LIF_junction_PSU_D"
[53] "LIF_junction_SQW_A"   "LIF_junction_SQW_B"   "LIF_junction_SQW_C"   "LIF_junction_SQW_D"
[57] "LIF_refseq_gene_LIV"  "LIF_refseq_gene_NWU"  "LIF_refseq_gene_PSU"  "LIF_refseq_gene_SQW"
[61] "ROC_aceview_gene_MGP" "ROC_aceview_gene_NYU" "ROC_aceview_gene_SQW" "ROC_junction_MGP_A"
[65] "ROC_junction_MGP_B"   "ROC_junction_NYU_A"   "ROC_junction_NYU_B"   "ROC_junction_SQW_A"
[69] "ROC_junction_SQW_B"   "ROC_refseq_gene_MGP"  "ROC_refseq_gene_NYU"  "ROC_refseq_gene_SQW"
[73] "taqman"
```

Six (6) samples were employed in the SEQC project [1]. Sample A and B were obtained from two human RNA reference libraries, while a small amount of Ambion ERCC RNA Spike-in Mix was added into both samples. Sample A and B were then mixed up to different ratios, creating Sample C and D. The pure ERCC Spike-in Mix 1 and 2 were also used as sample E and F, ending up with totally six samples being studied in this

project. The details about the samples are further described in the next section. Huge amounts of RNA-seq data (short reads) were yielded from the six samples at twelve (12) sequencing site by using three (3) different platforms (sequencers).

Genes that have different expression levels among samples can be detected from the RNA-seq data. The read count tables in this package are useful for the purpose of Differently Expressed (DE) gene analysis. For generating the read count tables, the RNA-seq data was mapped to the human genome by using the Subread aligner [2], and the mapping result was assigned to the annotated genes by using the featureCounts program [3] for estimating their expression levels. Two sets of gene annotations were referred when the mapping result was assigned to the genomes: the AceView annotations and the RefSeq annotations, both from the National Center for Biotechnology Information (NCBI) in the United States.

The RNA-seq technologies give an opportunity to precisely call the exon-exon junctions that are coded in the RNA transcriptome. We used the Subjunc program [2] to process the mapping result from the Subread aligner, therefore detecting the exon-exon junctions by using the seed-and-vote strategy. The junction tables from Subjunc were summarized on the sample level and are included in this package.

In addition to the analyses on the RNA-seq data, samples A, B, C and D were also analyzed by using the TaqMan RT-PCR technology. The intensity values of the selected genes are included in this package as a reference.

2 Samples and Sequencing

The SEQC consortium prepared six RNA samples and distributed the samples to twelve sequencing sites. The first two samples, A and B, were derived from Agilent's Universal Human Reference RNA (UHRR) and Life Technologies' Human Brain Reference RNA (HBRR) cell lines respectively. Sample A and B were then mixed with Ambion ERCC RNA Spike-In Mix 1 and 2 accordingly, which are two mixtures of RNA molecules from 92 contigs, but each contig is known to have different density levels in the two mixtures. Sample C and D were then created by mixing sample A and B to different ratios. In sample C, there is 75% of the volume from sample A and 25% of the volume from sample B; the ratio between sample A and B is 25% and 75% in sample D. As we mentioned above, sample E and F are the pure ERCC RNA Spike-In Mix 1 and 2. All the sequencing sites sequenced Sample A and B, but some sites did not sequence sample C, D, E and/or F.

Each of the samples has a number of replicates. Samples A, B, C and D each has five (5) replicates, and Sample E and F each has two (2) replicates. Some sequencing sites sequenced all of these replicates, but the other sites only sequenced a part of them. The set of replicates sequenced at each site can be found from the column names of the data frames in this package. The naming scheme is described in the next chapter.

Three RNA-seq platforms were examined in the SEQC project: the Illumina HiSeq 2000 devices (ILM), the Roche 454 GS FLX platform (ROC) and Life Technologies'

SOLiD 5500 instruments (LIF). Each sequencing site uses a number of lanes and flowcells to sequence each replicate, but the numbers of lanes and flowcells used in the SEQC project differ between the platforms and event between the sequencing sites using the same platform. This project generated 2758 libraries in total: 1832 of them from the Illumnia HiSeq platform, 12 from the 454 Life Science platform and 914 from the SOLiD systems. Table 1 lists the sequencing sites, the platforms employed, the number of samples/replicates that were sequenced at each site, and the numbers of short read libraries yielded.

The 2758 libraries have different read formats. All the libraries from Illumnia HiSeq are paired-end, and the length of every single read is around 100bp. The 454 Life Science libraries, on the other hand, are all single-end, and the read length is variable from tens of bases to above one thousand bases. Different to the libraries from Illumnia HiSeq and 454 Life Science, which contain only base-space reads, all the SOLiD libraries contain only color-space reads. There are 50 single-end libraries and 864 paired-end libraries among all the 914 SOLiD libraries. The read lengths in the SOLiD library vary from 36 to 76 colors long.

Site	Platform	Samples	Replicates	Libraries	PE/SE
AGR	Illumina (ILM)	A, B, C, D	4 for each sample	256	PE
BGI	Illumina (ILM)	A, B, C, D, E, F	5 for A, B, C & D; 2 for E & F	384	PE
CNL	Illumina (ILM)	A, B, C, D, E, F	5 for A, B, C & D; 2 for E & F	384	PE
COH	Illumina (ILM)	A, B, C, D	4 for each sample	128	PE
MAY	Illumina (ILM)	A, B, C, D, E, F	5 for A, B, C & D; 2 for E & F	384	PE
NVS	Illumina (ILM)	A, B, C, D, E, F	4 for A, B, C & D; 2 for E & F	320	PE
LIV	SOLiD (LIF)	A, B, C, D	2 for each sample	50	SE
NWU	SOLiD (LIF)	A, B, C, D, E, F	5 for A, B, C & D; 2 for E & F	288	PE
PSU	SOLiD (LIF)	A, B, C, D, E, F	5 for A, B, C & D; 2 for E & F	288	PE
SQW	SOLiD (LIF)	A, B, C, D, E, F	5 for A, B, C & D; 2 for E & F	288	PE
SQW	454 (ROC)	A, B	1 for each sample	4	SE
MGP	454 (ROC)	A, B	1 for each sample	4	SE
NYU	454 (ROC)	A, B	1 for each sample	4	SE

Table 1: List of the sequencing sites and processed samples/replicates. SQW used two platforms: SOLiD and 454. The last column indicates if the libraries are paired-end (PE) or single-end (SE)

3 Read Mapping and Read Assignment

All the libraries were mapped to the human genome by using the Subread package.

The reference genome index was first built from the GRCg37/hg19 human genome, plus the 92 ERCC Spike-In contigs[1], the Subread aligner then mapped the RNA-seq libraries to the index using the default settings. The mapping results were then assigned

[1] `http://tools.lifetechnologies.com/downloads/ERCC_Controls_Annotation.txt`

to the annotated genes by featureCounts, running on the single-end mode or the paired-end mode according to the format of the library. Two sets of annotations were used in the study: the AceView annotations and the RefSeq annotations. There are 1,383,372 exons (from 55,950 genes) in the AceView annotations, and there are 225,074 exons (from 25,072 genes) in the RefSeq annotations; only the exon regions of the genes were extracted from the annotations for read assignment. It is noticed that, although there are much more genes in the AceView annotations, the genes in the RefSeq annotations are not a subset of genes in the AceView annotations. Both sets of annotations have some unique genes not in the other set.

The gene-level read assignment results are provided in this package as 22 data frames, which contain the numbers of single-end reads or paired-end fragments overlapping with exons of each annotated gene. The names of the data frames are synthesized following this scheme: `(PLATFORM)_(ANNOTATION)_gene_(SITE)`, for example, data frame `LIF_refseq_gene_NWU` includes the read assignment results on the libraries sequenced at NMU by using the SOLiD platform, and the reads were assigned to the RefSeq annotations. The column names in the data frames describe the sample names, the replicate numbers, the lane numbers and the flowcell numbers (if applicable) associated with the libraries. For example, the read counts in `LIF_refseq_gene_NWU$B_4_L02_FlowCell2` are the fourth replicate of Sample B, sequenced by using the second flowcell in lane 2. Another example, `ROC_aceview_gene_MGP$A_1_R02` include the RNA-seq library generated from the first replicate of Sample A at MGP by using the second region of the 454 platform.

```
> ROC_aceview_gene_MGP[1:15,]
```

	EntrezID	Symbol	GeneLength	IsERCC	A_1_R01	A_1_R02	B_1_R01	B_1_R02
1	<NA>	2-OXOACID_DH	1624	FALSE	0	0	0	0
2	<NA>	A1BGAS	5364	FALSE	1	3	0	2
3	29974	A1CF	2511	FALSE	1	3	0	0
4	<NA>	A2BP1	9426	FALSE	0	0	68	78
5	<NA>	A2LD1	3848	FALSE	5	2	2	0
6	2	A2M	8153	FALSE	213	169	46	61
7	144568	A2ML1	7492	FALSE	0	1	3	0
8	3	A2MP1	1741	FALSE	0	0	0	0
9	53947	A4GALT	3184	FALSE	5	2	2	1
10	51146	A4GNT	1771	FALSE	0	0	0	1
11	100329167	AAA1	7197	FALSE	0	3	0	0
12	8086	AAAS	4614	FALSE	19	14	7	4
13	65985	AACS	13098	FALSE	14	11	27	20
14	<NA>	AACSL	3594	FALSE	2	0	0	0
15	13	AADAC	3106	FALSE	0	0	0	0

```
> colnames(ILM_aceview_gene_BGI)
```

```
 [1] "EntrezID"          "Symbol"            "GeneLength"        "IsERCC"            "A_1_L01_FlowCellA"
 [6] "A_1_L01_FlowCellB" "A_1_L02_FlowCellA" "A_1_L02_FlowCellB" "A_1_L03_FlowCellA" "A_1_L03_FlowCellB"
[11] "A_1_L04_FlowCellA" "A_1_L04_FlowCellB" "A_1_L05_FlowCellA" "A_1_L05_FlowCellB" "A_1_L06_FlowCellA"
[16] "A_1_L06_FlowCellB" "A_1_L07_FlowCellA" "A_1_L07_FlowCellB" "A_1_L08_FlowCellA" "A_1_L08_FlowCellB"
[21] "A_2_L01_FlowCellA" "A_2_L01_FlowCellB" "A_2_L02_FlowCellA" "A_2_L02_FlowCellB" "A_2_L03_FlowCellA"
[26] "A_2_L03_FlowCellB" "A_2_L04_FlowCellA" "A_2_L04_FlowCellB" "A_2_L05_FlowCellA" "A_2_L05_FlowCellB"
[31] "A_2_L06_FlowCellA" "A_2_L06_FlowCellB" "A_2_L07_FlowCellA" "A_2_L07_FlowCellB" "A_2_L08_FlowCellA"
[36] "A_2_L08_FlowCellB" "A_3_L01_FlowCellA" "A_3_L01_FlowCellB" "A_3_L02_FlowCellA" "A_3_L02_FlowCellB"
```

```
 [41] "A_3_L03_FlowCellA" "A_3_L03_FlowCellB" "A_3_L04_FlowCellA" "A_3_L04_FlowCellB" "A_3_L05_FlowCellA"
 [46] "A_3_L05_FlowCellB" "A_3_L06_FlowCellA" "A_3_L06_FlowCellB" "A_3_L07_FlowCellA" "A_3_L07_FlowCellB"
 [51] "A_3_L08_FlowCellA" "A_3_L08_FlowCellB" "A_4_L01_FlowCellA" "A_4_L01_FlowCellB" "A_4_L02_FlowCellA"
 [56] "A_4_L02_FlowCellB" "A_4_L03_FlowCellA" "A_4_L03_FlowCellB" "A_4_L04_FlowCellA" "A_4_L04_FlowCellB"
 [61] "A_4_L05_FlowCellA" "A_4_L05_FlowCellB" "A_4_L06_FlowCellA" "A_4_L06_FlowCellB" "A_4_L07_FlowCellA"
 [66] "A_4_L07_FlowCellB" "A_4_L08_FlowCellA" "A_4_L08_FlowCellB" "A_5_L01_FlowCellA" "A_5_L01_FlowCellB"
 [71] "A_5_L02_FlowCellA" "A_5_L02_FlowCellB" "A_5_L03_FlowCellA" "A_5_L03_FlowCellB" "A_5_L04_FlowCellA"
 [76] "A_5_L04_FlowCellB" "A_5_L05_FlowCellA" "A_5_L05_FlowCellB" "A_5_L06_FlowCellA" "A_5_L06_FlowCellB"
 [81] "A_5_L07_FlowCellA" "A_5_L07_FlowCellB" "A_5_L08_FlowCellA" "A_5_L08_FlowCellB" "B_1_L01_FlowCellA"
 [86] "B_1_L01_FlowCellB" "B_1_L02_FlowCellA" "B_1_L02_FlowCellB" "B_1_L03_FlowCellA" "B_1_L03_FlowCellB"
 [91] "B_1_L04_FlowCellA" "B_1_L04_FlowCellB" "B_1_L05_FlowCellA" "B_1_L05_FlowCellB" "B_1_L06_FlowCellA"
 [96] "B_1_L06_FlowCellB" "B_1_L07_FlowCellA" "B_1_L07_FlowCellB" "B_1_L08_FlowCellA" "B_1_L08_FlowCellB"
[101] "B_2_L01_FlowCellA" "B_2_L01_FlowCellB" "B_2_L02_FlowCellA" "B_2_L02_FlowCellB" "B_2_L03_FlowCellA"
[106] "B_2_L03_FlowCellB" "B_2_L04_FlowCellA" "B_2_L04_FlowCellB" "B_2_L05_FlowCellA" "B_2_L05_FlowCellB"
[111] "B_2_L06_FlowCellA" "B_2_L06_FlowCellB" "B_2_L07_FlowCellA" "B_2_L07_FlowCellB" "B_2_L08_FlowCellA"
[116] "B_2_L08_FlowCellB" "B_3_L01_FlowCellA" "B_3_L01_FlowCellB" "B_3_L02_FlowCellA" "B_3_L02_FlowCellB"
[121] "B_3_L03_FlowCellA" "B_3_L03_FlowCellB" "B_3_L04_FlowCellA" "B_3_L04_FlowCellB" "B_3_L05_FlowCellA"
[126] "B_3_L05_FlowCellB" "B_3_L06_FlowCellA" "B_3_L06_FlowCellB" "B_3_L07_FlowCellA" "B_3_L07_FlowCellB"
[131] "B_3_L08_FlowCellA" "B_3_L08_FlowCellB" "B_4_L01_FlowCellA" "B_4_L01_FlowCellB" "B_4_L02_FlowCellA"
[136] "B_4_L02_FlowCellB" "B_4_L03_FlowCellA" "B_4_L03_FlowCellB" "B_4_L04_FlowCellA" "B_4_L04_FlowCellB"
[141] "B_4_L05_FlowCellA" "B_4_L05_FlowCellB" "B_4_L06_FlowCellA" "B_4_L06_FlowCellB" "B_4_L07_FlowCellA"
[146] "B_4_L07_FlowCellB" "B_4_L08_FlowCellA" "B_4_L08_FlowCellB" "B_5_L01_FlowCellA" "B_5_L01_FlowCellB"
[151] "B_5_L02_FlowCellA" "B_5_L02_FlowCellB" "B_5_L03_FlowCellA" "B_5_L03_FlowCellB" "B_5_L04_FlowCellA"
[156] "B_5_L04_FlowCellB" "B_5_L05_FlowCellA" "B_5_L05_FlowCellB" "B_5_L06_FlowCellA" "B_5_L06_FlowCellB"
[161] "B_5_L07_FlowCellA" "B_5_L07_FlowCellB" "B_5_L08_FlowCellA" "B_5_L08_FlowCellB" "C_1_L01_FlowCellA"
[166] "C_1_L01_FlowCellB" "C_1_L02_FlowCellA" "C_1_L02_FlowCellB" "C_1_L03_FlowCellA" "C_1_L03_FlowCellB"
[171] "C_1_L04_FlowCellA" "C_1_L04_FlowCellB" "C_1_L05_FlowCellA" "C_1_L05_FlowCellB" "C_1_L06_FlowCellA"
[176] "C_1_L06_FlowCellB" "C_1_L07_FlowCellA" "C_1_L07_FlowCellB" "C_1_L08_FlowCellA" "C_1_L08_FlowCellB"
[181] "C_2_L01_FlowCellA" "C_2_L01_FlowCellB" "C_2_L02_FlowCellA" "C_2_L02_FlowCellB" "C_2_L03_FlowCellA"
[186] "C_2_L03_FlowCellB" "C_2_L04_FlowCellA" "C_2_L04_FlowCellB" "C_2_L05_FlowCellA" "C_2_L05_FlowCellB"
[191] "C_2_L06_FlowCellA" "C_2_L06_FlowCellB" "C_2_L07_FlowCellA" "C_2_L07_FlowCellB" "C_2_L08_FlowCellA"
[196] "C_2_L08_FlowCellB" "C_3_L01_FlowCellA" "C_3_L01_FlowCellB" "C_3_L02_FlowCellA" "C_3_L02_FlowCellB"
[201] "C_3_L03_FlowCellA" "C_3_L03_FlowCellB" "C_3_L04_FlowCellA" "C_3_L04_FlowCellB" "C_3_L05_FlowCellA"
[206] "C_3_L05_FlowCellB" "C_3_L06_FlowCellA" "C_3_L06_FlowCellB" "C_3_L07_FlowCellA" "C_3_L07_FlowCellB"
[211] "C_3_L08_FlowCellA" "C_3_L08_FlowCellB" "C_4_L01_FlowCellA" "C_4_L01_FlowCellB" "C_4_L02_FlowCellA"
[216] "C_4_L02_FlowCellB" "C_4_L03_FlowCellA" "C_4_L03_FlowCellB" "C_4_L04_FlowCellA" "C_4_L04_FlowCellB"
[221] "C_4_L05_FlowCellA" "C_4_L05_FlowCellB" "C_4_L06_FlowCellA" "C_4_L06_FlowCellB" "C_4_L07_FlowCellA"
[226] "C_4_L07_FlowCellB" "C_4_L08_FlowCellA" "C_4_L08_FlowCellB" "C_5_L01_FlowCellA" "C_5_L01_FlowCellB"
[231] "C_5_L02_FlowCellA" "C_5_L02_FlowCellB" "C_5_L03_FlowCellA" "C_5_L03_FlowCellB" "C_5_L04_FlowCellA"
[236] "C_5_L04_FlowCellB" "C_5_L05_FlowCellA" "C_5_L05_FlowCellB" "C_5_L06_FlowCellA" "C_5_L06_FlowCellB"
[241] "C_5_L07_FlowCellA" "C_5_L07_FlowCellB" "C_5_L08_FlowCellA" "C_5_L08_FlowCellB" "D_1_L01_FlowCellA"
[246] "D_1_L01_FlowCellB" "D_1_L02_FlowCellA" "D_1_L02_FlowCellB" "D_1_L03_FlowCellA" "D_1_L03_FlowCellB"
[251] "D_1_L04_FlowCellA" "D_1_L04_FlowCellB" "D_1_L05_FlowCellA" "D_1_L05_FlowCellB" "D_1_L06_FlowCellA"
[256] "D_1_L06_FlowCellB" "D_1_L07_FlowCellA" "D_1_L07_FlowCellB" "D_1_L08_FlowCellA" "D_1_L08_FlowCellB"
[261] "D_2_L01_FlowCellA" "D_2_L01_FlowCellB" "D_2_L02_FlowCellA" "D_2_L02_FlowCellB" "D_2_L03_FlowCellA"
[266] "D_2_L03_FlowCellB" "D_2_L04_FlowCellA" "D_2_L04_FlowCellB" "D_2_L05_FlowCellA" "D_2_L05_FlowCellB"
[271] "D_2_L06_FlowCellA" "D_2_L06_FlowCellB" "D_2_L07_FlowCellA" "D_2_L07_FlowCellB" "D_2_L08_FlowCellA"
[276] "D_2_L08_FlowCellB" "D_3_L01_FlowCellA" "D_3_L01_FlowCellB" "D_3_L02_FlowCellA" "D_3_L02_FlowCellB"
[281] "D_3_L03_FlowCellA" "D_3_L03_FlowCellB" "D_3_L04_FlowCellA" "D_3_L04_FlowCellB" "D_3_L05_FlowCellA"
[286] "D_3_L05_FlowCellB" "D_3_L06_FlowCellA" "D_3_L06_FlowCellB" "D_3_L07_FlowCellA" "D_3_L07_FlowCellB"
[291] "D_3_L08_FlowCellA" "D_3_L08_FlowCellB" "D_4_L01_FlowCellA" "D_4_L01_FlowCellB" "D_4_L02_FlowCellA"
[296] "D_4_L02_FlowCellB" "D_4_L03_FlowCellA" "D_4_L03_FlowCellB" "D_4_L04_FlowCellA" "D_4_L04_FlowCellB"
[301] "D_4_L05_FlowCellA" "D_4_L05_FlowCellB" "D_4_L06_FlowCellA" "D_4_L06_FlowCellB" "D_4_L07_FlowCellA"
[306] "D_4_L07_FlowCellB" "D_4_L08_FlowCellA" "D_4_L08_FlowCellB" "D_5_L01_FlowCellA" "D_5_L01_FlowCellB"
[311] "D_5_L02_FlowCellA" "D_5_L02_FlowCellB" "D_5_L03_FlowCellA" "D_5_L03_FlowCellB" "D_5_L04_FlowCellA"
[316] "D_5_L04_FlowCellB" "D_5_L05_FlowCellA" "D_5_L05_FlowCellB" "D_5_L06_FlowCellA" "D_5_L06_FlowCellB"
[321] "D_5_L07_FlowCellA" "D_5_L07_FlowCellB" "D_5_L08_FlowCellA" "D_5_L08_FlowCellB" "E_1_L01_FlowCellA"
[326] "E_1_L01_FlowCellB" "E_1_L02_FlowCellA" "E_1_L02_FlowCellB" "E_1_L03_FlowCellA" "E_1_L03_FlowCellB"
[331] "E_1_L04_FlowCellA" "E_1_L04_FlowCellB" "E_1_L05_FlowCellA" "E_1_L05_FlowCellB" "E_1_L06_FlowCellA"
[336] "E_1_L06_FlowCellB" "E_1_L07_FlowCellA" "E_1_L07_FlowCellB" "E_1_L08_FlowCellA" "E_1_L08_FlowCellB"
[341] "E_2_L01_FlowCellA" "E_2_L01_FlowCellB" "E_2_L02_FlowCellA" "E_2_L02_FlowCellB" "E_2_L03_FlowCellA"
[346] "E_2_L03_FlowCellB" "E_2_L04_FlowCellA" "E_2_L04_FlowCellB" "E_2_L05_FlowCellA" "E_2_L05_FlowCellB"
[351] "E_2_L06_FlowCellA" "E_2_L06_FlowCellB" "E_2_L07_FlowCellA" "E_2_L07_FlowCellB" "E_2_L08_FlowCellA"
[356] "E_2_L08_FlowCellB" "F_1_L01_FlowCellA" "F_1_L01_FlowCellB" "F_1_L02_FlowCellA" "F_1_L02_FlowCellB"
```

```
[361] "F_1_L03_FlowCellA" "F_1_L03_FlowCellB" "F_1_L04_FlowCellA" "F_1_L04_FlowCellB" "F_1_L05_FlowCellA"
[366] "F_1_L05_FlowCellB" "F_1_L06_FlowCellA" "F_1_L06_FlowCellB" "F_1_L07_FlowCellA" "F_1_L07_FlowCellB"
[371] "F_1_L08_FlowCellA" "F_1_L08_FlowCellB" "F_2_L01_FlowCellA" "F_2_L01_FlowCellB" "F_2_L02_FlowCellA"
[376] "F_2_L02_FlowCellB" "F_2_L03_FlowCellA" "F_2_L03_FlowCellB" "F_2_L04_FlowCellA" "F_2_L04_FlowCellB"
[381] "F_2_L05_FlowCellA" "F_2_L05_FlowCellB" "F_2_L06_FlowCellA" "F_2_L06_FlowCellB" "F_2_L07_FlowCellA"
[386] "F_2_L07_FlowCellB" "F_2_L08_FlowCellA" "F_2_L08_FlowCellB"
```

```
> ILM_aceview_gene_BGI[1:15,1:7]
```

	EntrezID	Symbol	GeneLength	IsERCC	A_1_L01_FlowCellA	A_1_L01_FlowCellB	A_1_L02_FlowCellA
1	<NA>	2-OXOACID_DH	1624	FALSE	0	0	0
2	<NA>	A1BGAS	5364	FALSE	17	15	18
3	29974	A1CF	2511	FALSE	21	14	21
4	<NA>	A2BP1	9426	FALSE	7	5	1
5	<NA>	A2LD1	3848	FALSE	9	15	17
6	2	A2M	8153	FALSE	2683	2777	2660
7	144568	A2ML1	7492	FALSE	4	0	6
8	3	A2MP1	1741	FALSE	1	3	0
9	53947	A4GALT	3184	FALSE	44	58	56
10	51146	A4GNT	1771	FALSE	1	0	0
11	100329167	AAA1	7197	FALSE	5	10	8
12	8086	AAAS	4614	FALSE	369	358	327
13	65985	AACS	13098	FALSE	202	207	184
14	<NA>	AACSL	3594	FALSE	15	12	10
15	13	AADAC	3106	FALSE	0	0	0

4 Junction Detection

The subread aligner can very efficiently map reads to the reference genome. From the mapping results, subjunc furthers detected exon-exon junctions that were coded in the RNA-seq reads[2], and generated a junction table for each library. We merged the junction tables from the libraries on the platform-site-sample level, then provide the resulted data frames in this package.

The data frames are named following this scheme: `(PLATFORM)_junction_(SITE)_SAMPLE`. For example, data frame `ILM_junction_AGR_B` includes the junctions that were detected from all the libraries of Sample B sequenced at AGR by using the Illumina HiSeq platform. Each data frame contains four columns: the chromosome name, the first junction side, the second junction side and the number of supporting reads (i.e., the reads that spanning the junction location and include the two exons surrounding the junction). The two junction sides are defined as the chromosomal position (one starting) of the last base in the first exon, and the chromosomal position of the first base in the second exon both surrounding this junction point.

```
> ILM_junction_AGR_B[1:15,]
```

	Chromosome	Location1	Location2	nSupportingReads
1	ERCC-00002	195	228	2
2	ERCC-00002	407	430	2

[2]The libraries were processed by using an old version of the subread package. The subjunc in that package only takes SAM files from the subread aligner as input. The subjunc program in the newer package (after version 1.4.0) is able to directly take FASTQ and FASTA files as input.

```
3   ERCC-00002        75       228               2
4   ERCC-00096       527       563               2
5   ERCC-00113       130       146               2
6   ERCC-00130       757       850               6
7   GL000210.1      5430      5545               2
8  NT_113878.1     29833     30763              12
9  NT_113878.1     30969     31271              19
10 NT_113878.1     37549     41405              11
11 NT_113878.1     37782     41405               1
12 NT_113878.1     41272     41405               3
13 NT_113878.1     41456     41682               5
14 NT_113878.1     41459     41682              20
15 NT_113885.1     27618     31602               2
```

5 TaqMan RT-PCR

The gene expression levels in Samples A, B, C and D were further measured by using the TaqMan RT-PCR technologies. The expression levels of the 1044 selected genes in replicates 1, 2, 3 and 4 of the four samples are stored in data frame `taqman`. The first column in this data frame are the entrez ids and symbols of the 1044 genes, followed by the gene intensity values in 16 columns, which are named as `(SAMPLE).(SAMPLE)(REPLICATE)_values`, for example, `taqman$C.C3_values` are the gene intensity values in the third replicate of sample C.

```
> colnames(taqman)

 [1] "EntrezID"   "Symbol"     "A.A1_value" "A.A2_value" "A.A3_value" "A.A4_value" "B.B1_value" "B.B2_value"
 [9] "B.B3_value" "B.B4_value" "C.C1_value" "C.C2_value" "C.C3_value" "C.C4_value" "D.D1_value" "D.D2_value"
[17] "D.D3_value" "D.D4_value"

> taqman[1:15, 1:9]

   EntrezID Symbol A.A1_value A.A2_value A.A3_value A.A4_value B.B1_value B.B2_value B.B3_value
1      1543 CYP1A1    0.00861    0.00875    0.00827    0.00819    0.00243    0.00174    0.00173
2      1950    EGF    0.02697    0.03002    0.02657    0.03169    0.00895    0.01291    0.00943
3      1153  CIRBP    2.11583    2.11533    2.15163    2.08393    3.91501    3.98488    3.85148
4      1613  DAPK3    0.19717    0.20006    0.16115    0.19256    0.20755    0.20065    0.20932
5      1665  DHX15    0.11647    0.11893    0.10562    0.11335    0.05930    0.06412    0.06543
6      1786  DNMT1    0.77968    0.81197    0.84169    0.76226    0.19767    0.18879    0.19459
7      1982 EIF4G2    7.14649    6.76277    7.65979    7.17128    4.25443    4.47529    4.28752
8      2069   EREG    0.03583    0.03379    0.03350    0.03665    0.00032    0.00052    0.00038
9      2196   FAT2    0.00447    0.00432    0.00412    0.00488    0.88969    0.93208    0.94607
10     3491  CYR61    0.65991    0.66509    0.68998    0.65407    0.09393    0.09249    0.08683
11     1527  TEX28    0.00043    0.00043    0.00043    0.00043    0.00032    0.00032    0.00032
12     1346 COX7A1    0.00291    0.00231    0.00222    0.00177    0.46515    0.47453    0.47124
13     1357   CPA1    0.00043    0.00043    0.00043    0.00043    0.00032    0.00032    0.00032
14     1360   CPB1    0.00043    0.00043    0.00043    0.00043    0.00075    0.00079    0.00062
15     1503  CTPS1    1.10256    1.06406    1.07379    1.12382    0.14391    0.13963    0.14414
```

6 Citation

[1] Su Z, Labaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, Thompson JF, et al. (SEQC/MAQC-III Consortium). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the

Sequencing Quality Control Consortium. Nature Biotechnology, 32(9), 903-914

[2] Yang Liao, Gordon K Smyth and Wei Shi (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108.

[3] Yang Liao, Gordon K Smyth and Wei Shi (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30

7 Authors

Yang Liao and Wei Shi
Bioinformatics Division
The Walter and Eliza Hall Institute of Medical Research
1G Royal Parade, Parkville, Victoria 3052
Australia

8 Contact

Please post to Bioconductor mailing list (`http://bioconductor.org/`) if you find any bugs and have any inquires. Or, you may contact Wei Shi (shi at wehi dot edu dot au) directly.