Cancer Letters 469 (2020) 246–255

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Original Articles

Technical differences between sequencing and microarray platforms impact transcriptomic subtyping of colorectal cancer

Ina A. Eilertsen[a,b,c], Seyed H. Moosavi[a,b,c], Jonas M. Strømme[a,b,d], Arild Nesbakken[b,c,e], Bjarne Johannessen[a,b,c], Ragnhild A. Lothe[a,b,c], Anita Sveen[a,b,c,*]

[a] Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital, P.O. Box 4953, Nydalen, NO-0424, Oslo, Norway
[b] K. G. Jebsen Colorectal Cancer Research Centre, Oslo University Hospital, P.O. Box 4953, Nydalen, NO-0424, Oslo, Norway
[c] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, P.O. Box 1171 Blindern, NO-0318, Oslo, Norway
[d] Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, P.O. Box 1080, Blindern, NO-0316, Oslo, Norway
[e] Department of Gastrointestinal Surgery, Oslo University Hospital, P.O. Box 4950, Nydalen, NO-0424, Oslo, Norway

ARTICLE INFO

Keywords:
Colorectal cancer
Microarray
RNA sequencing
Consensus molecular subtypes
CRC intrinsic subtypes
Classification concordance

ABSTRACT

Gene expression profiling has increasing relevance in the molecular screening of patients with colorectal cancer (CRC). We investigated potential platform-specific effects on transcriptomic subtyping according to established frameworks by comparisons of expression profiles from RNA sequencing and exon-resolution microarrays in 126 primary microsatellite stable CRCs. There was a strong platform correspondence in global gene expression levels, albeit with systematic technical bias likely attributed to few sequencing reads covering short (< 2000 nucleotides) and/or lowly expressed genes (< 1 FPKM), as well as over-saturation of highly expressed genes on microarrays. Classification concordances according to both the consensus molecular subtypes and CRC intrinsic subtypes (CRIS) were also strong, but with disproportionate subtype distributions between platforms caused by frequent disagreements in adherence to sample classification thresholds. Subtypes defined largely by genes expressed at low levels, including the CRIS-D subtype and the estimated level of tumor-infiltrating cytotoxic lymphocytes, had a weaker correspondence in classification metrics between platforms. In conclusion, even subtle differences between platforms suggest that clinical translation of transcriptomic CRC subtyping frameworks is dependent on assay standardization, and systematic technical biases reinforce the need for careful selection of classifier genes.

1. Introduction

Gene expression profiling of colorectal cancers (CRCs) has strong clinical relevance [1]. Incorporation of gene expression changes in the molecular screening of patients in clinical trials may improve treatment predictions compared with genomic screening alone [2]. Nonetheless, the current evidence of clinical relevance is strongest in relation to patient prognosis, and both tumor microenvironment-related expression signals arising from cancer-associated fibroblasts and infiltrating cytotoxic T cells [3–5], as well as transcriptomic classification frameworks [6–11], have strong associations with patient survival independent of cancer stage. Gene expression-based subtyping reached a milestone with the publication of the four consensus molecular subtypes (CMS) in 2015 [12], and CMS has now been accepted as a robust taxonomy of CRCs. However, this framework is vulnerable to tumor sampling bias [13,14], and both the CMS1-immune and CMS4-mesenchymal groups are strongly influenced by nonmalignant cells in the tumor microenvironment. To provide a stroma-independent framework, the five CRC intrinsic subtypes (CRIS) were identified based on cancer cell-intrinsic transcriptional traits [15]. In particular the CMS framework has been scrutinized for a potential predictive value in retrospective analyses of clinical studies investigating standard chemotherapies and targeted agents. However, indications of subtype-associated differences in treatment benefits have so far been inconsistent between studies, and discrepant results might be explained by both biological and technical analysis variation [16,17].

Microarrays have for many years been the technology of choice for large-scale transcriptomic studies, and the availability of large data resources with thousands of CRC samples has been essential for the development of the classification frameworks. However, RNA

[*] Corresponding author. Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital, P.O. Box 4953, Nydalen, NO-0424, Oslo, Norway.

E-mail address: anita.sveen@rr-research.no (A. Sveen).

https://doi.org/10.1016/j.canlet.2019.10.040
Received 26 August 2019; Received in revised form 24 October 2019; Accepted 27 October 2019

sequencing (RNA-seq) addresses several of the limitations of microarrays and has in the last years emerged as the preferred method for whole-transcriptome profiling. RNA-seq enables absolute quantification of gene expression, resolution at the nucleotide level, detection of novel transcript structures and splice junctions [18], as well as quantitative measurement of allele-specific expression [19]. RNA-seq also avoids inherent technical biases observed with microarrays, related to cross-hybridization and a limited dynamic range of expression [18]. On the other hand, the considerably higher cost per sample for RNA-seq compared with microarrays is an important limiting factor, particularly in translational studies, where large sample numbers are often needed. Additionally, the accuracy of RNA-seq is influenced by both the abundance and length of transcripts [20,21], as well as the choice of assay for library construction [22] and tools for bioinformatic data processing [23].

Several studies have compared the performance of sequencing and microarray-based technologies [24–31], but the majority of studies have used older versions of microarrays that only target the 3′ ends of transcripts, and/or have relied on small sample numbers. Furthermore, studies including CRCs are few, and downstream analyses have primarily been focused on detection of differentially expressed genes and evaluation of dynamic ranges. The two technologies have not been compared with respect to clinically relevant CRC subtyping frameworks, for which sufficient sample numbers to adequately represent each subtype is required. Prior to clinical implementation of transcriptomic classifications, the potential impact of technical factors, such as choice of analysis technology, needs to be resolved. Here, we have compared gene expression profiles generated by both RNA-seq and exon-resolution microarrays for 126 primary CRCs. Expression correspondence was analyzed both sample-wise and gene-wise, and we investigated the potential effects of analysis platform on CRC classification according to established frameworks.

2. Materials and methods

2.1. Patient samples

Altogether 126 primary CRCs from a consecutive, population-representative patient series were included. The patients were treated surgically at Oslo University Hospital, Oslo, Norway, between 2005 and 2014. The CRCs were selected to include only microsatellite stable (MSS) and mostly stage II and III tumors (Supplementary Table S1). Details of RNA extraction and microsatellite instability analysis are included as Supplementary Methods. RNA from the same extraction procedure was analyzed by both microarrays and RNA-seq.

2.2. Microarray analyses

All samples have been analyzed for gene expression at exon-resolution using the GeneChip Human Exon Array 1.0 (HuEx; $n = 85$) or GeneChip Human Transcriptome Array 2.0 (HTA; $n = 41$) according to the manufacturer's instructions (Thermo Fisher Scientific, Waltham, MA, USA). The majority of the data ($n = 123$) have previously been published [32–36] and the three remaining samples have been deposited to the Gene Expression Omnibus (GEO). All samples can be accessed from the GEO series GSE139170. Raw intensity data CEL files were pre-processed by the robust multi-array average (RMA) algorithm using the justRMA function in the R package affy with custom CDF files from Brainarray (v22 GENECODEG) [37]. According to Brainarray, the CDF files were built using the human genome GRCh38 and the annotation file gencode.v26.annotation.gft from GENCODE [38]. R commands for pre-processing of microarray data are included as Supplementary Methods.

2.3. RNA sequencing

RNA-seq was done using the Illumina HiSeq 2500 platform with a 2x101 base-pair paired-end mode (details in Supplementary Methods). The quality of raw reads was controlled by the FastQC v0.11.2 tool, confirming good quality control metrics for all samples (Supplementary Table S2). Adapter sequences and low quality bases were removed using TRIMMOMATIC v0.38 [39]. Trimmed reads were aligned to the human genome GRCh38 with STAR v2.5.3a (2-pass mode) [40]. Aligned reads were sorted by read name using SAMtools v1.1 [41], and quantification of reads mapping to each gene was done with the htseq-count tool from the python package HTSeq v0.10.0 [42]. The same annotation file used to build the CDF files for microarray analyses, gencode.v26.annotation.gft, was applied for both STAR and htseq-count. Command line parameters are detailed in the Supplementary Methods.

Three methods for normalization of gene expression levels for protein-coding genes were applied for comparison. Within-sample normalization was done by estimation of the fragments per kilobase of transcripts (non-overlapping exonic length of a gene) per million mapped reads (FPKM) and the transcripts per million (TPM). Both the FPKM and TPM values were log2-transformed after a constant of 1 was added to each value. In addition, across-sample normalization based on variance-stabilizing transformation (VST) of the htseq-count data was done using the vst function in the R package DESeq2 [43]. Formulas and R commands for normalization of RNA-seq data are included as Supplementary Methods. Data can be accessed by contacting the corresponding author.

2.4. Gene-expression-based subtyping

The CRCs were classified according to both the CMS and CRIS frameworks based on both microarray and RNA-seq data. For the microarray dataset, classification was done separately for samples analyzed on HuEx and HTA arrays. To enable direct comparison, the normalized RNA-seq data were similarly divided into the two matching sample sets. Ensembl Gene IDs were translated to Entrez IDs using the R packages AnnotationDbi and org.Hs.eg.db. CMS classification was done using the random forest (RF) classifyCMS.RF function in the R package CMSclassifier [12], and the default posterior probability of 0.5 was used as a threshold for confident sample classification. CRIS classification was done using the nearest template predictor (NTP) algorithm implemented in the R package CMScaller [44] with CRIS template genes [15], and samples with a false discovery rate adjusted *P*-value (P_{adjusted}) $<$ 0.05 were considered confidently classified.

2.5. Gene expression-based abundances of microenvironment-related cell populations

The R package MCPcounter [45] was applied on both log2-transformed microarray and RNA-seq data (FPKM normalized) to estimate the abundances of tumor-infiltrating cytotoxic lymphocytes and fibroblasts in each CRC sample. These abundance scores represent the mean log2 expression level of the signature genes of each cell population in each sample.

2.6. Comparison of gene expression between platforms

Comparisons of the microarray and RNA-seq platforms were performed both sample-wise (across all genes per sample) and gene-wise (across all samples per gene). Thresholds for genes considered “underestimated by RNA-seq” compared with microarrays, and *vise versa*, were determined by visual inspection of sample-wise scatter plots (details in Supplementary Methods).

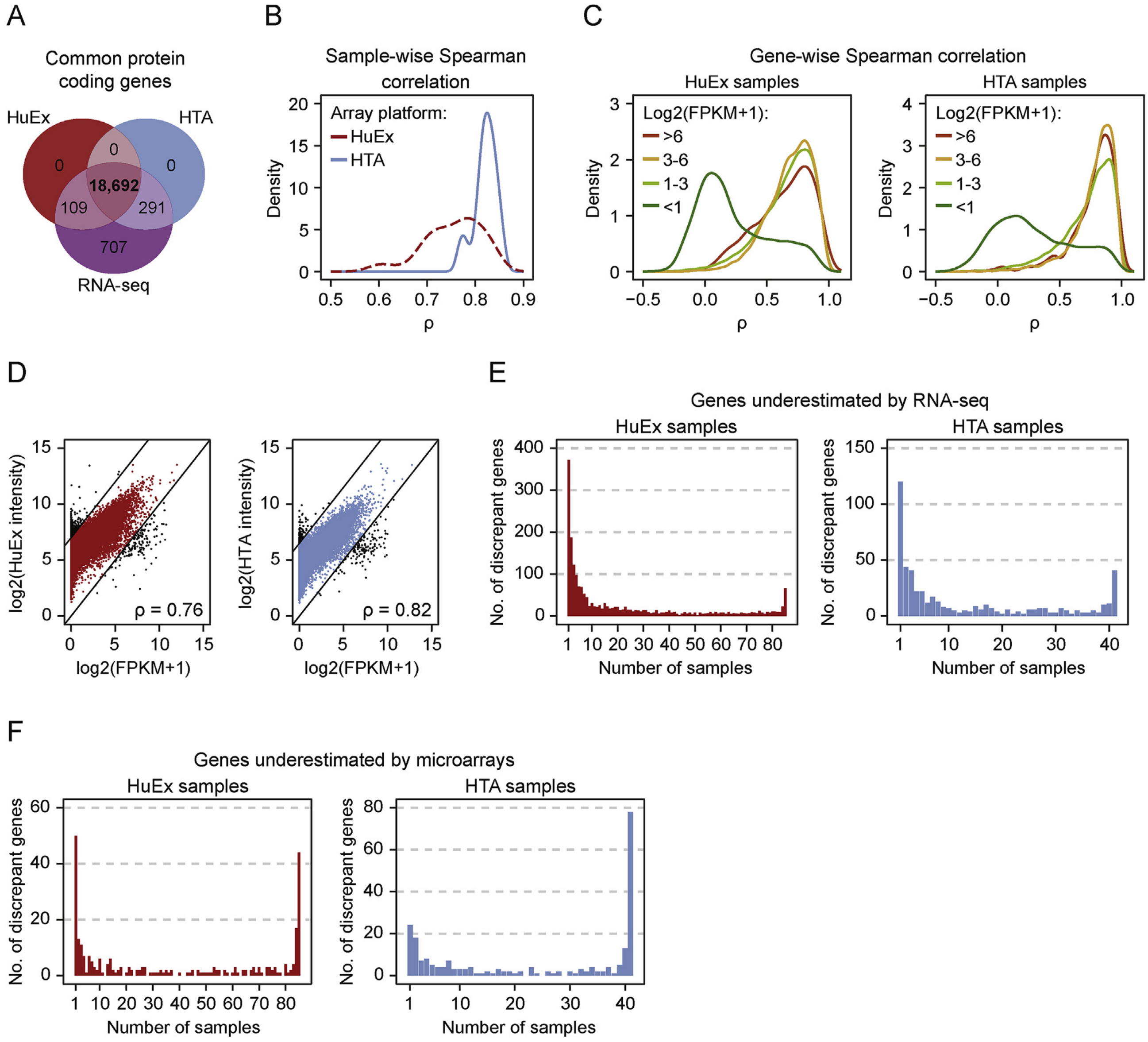

Fig. 1. Comparison of gene-level expression between microarrays and RNA-seq. (A) By matching the datasets by Ensembl Gene IDs, a total of 18,692 protein-coding genes were found in common between the three platforms and retained for downstream analysis. Protein-coding genes were identified from the gencode.v26.annotation.gft file. (B) Density plot of sample-wise correlations in gene expression between matched microarray and RNA-seq profiles, grouped by microarray platform. (C) Density distribution of cross-sample correlations for genes measured by both microarrays and RNA-seq, grouped according to their median expression level across samples as measured by RNA-seq. Gene-wise correlations were done after dividing the RNA-seq data according to samples analyzed on HuEx and HTA arrays, and the median log2-transformed FPKM value of each gene across samples was determined for the two RNA-seq sample sets separately. (D) Scatter plots illustrating the microarray and RNA-seq expression profiles for two samples, with genes underestimated or overestimated by RNA-seq compared to microarrays colored in black. Bar plots illustrating the number of genes with (E) underestimated expression levels based on RNA-seq relative to microarrays and (F) *vice versa*, according to the number of samples with the respective discrepant expression pattern.

2.7. Statistical analyses

All statistical analyses were performed in the R software environment v3.4.2 (details in Supplementary Methods).

3. Results

3.1. Good overall concordance, but with systematic biases, in gene expression between microarrays and RNA-seq

Gene expression profiles of 126 MSS CRCs were generated using both exon-resolution microarrays (HuEx or HTA) and RNA-seq and matched by Ensembl Gene IDs to generate a common list of 18,692 protein-coding genes for cross-platform comparisons (Fig. 1A). Sample-wise comparisons between microarrays and RNA-seq showed good correspondence in gene expression levels for all samples, although the correlation with RNA-seq was stronger and less varying for samples analyzed on HTA arrays (median Spearman's ρ = 0.82, interquartile range, IQR, 0.81–0.84) than for samples analyzed on HuEx arrays (median Spearman's ρ = 0.76, IQR 0.71–0.80; Fig. 1B). Similarly, gene-wise comparisons across all samples analyzed on each microarray type also showed that HTA arrays had a better correlation with RNA-seq (HTA: median Spearman's ρ = 0.71, IQR 0.36–0.86; HuEx: median Spearman's ρ = 0.60, IQR 0.25–0.78), with 52% (9,475 of 18,312) and 37% (6,682 of 18,263) of the genes having correlation coefficients > 0.70 for samples analyzed on HTA and HuEx arrays, respectively.

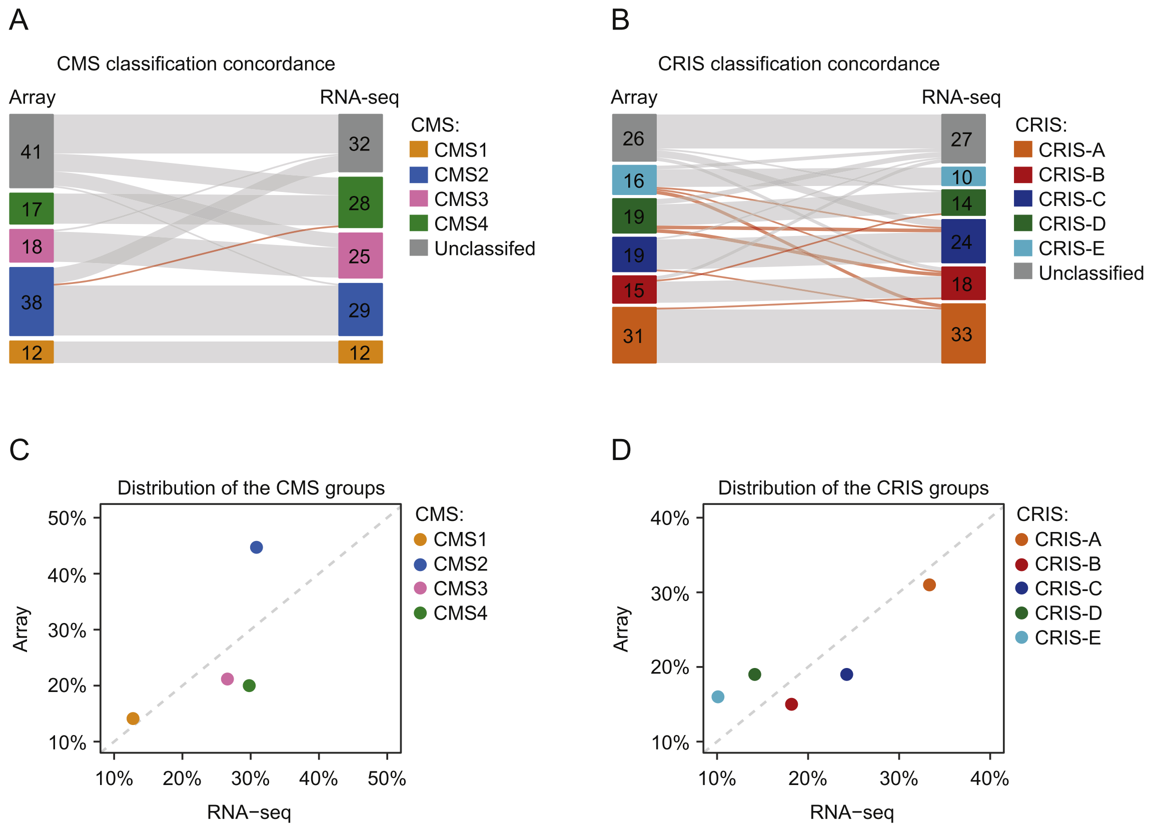

Fig. 2. CRC subtyping by microarrays and RNA-seq. Sample-wise concordance in (A) CMS and (B) CRIS classification between matched microarray and RNA-seq gene expression profiles from 126 CRCs. The number of tumors assigned to each subtype is specified within the blocks. Alluvia colored in red represent tumors with confident but discordant classification by the two technologies. Comparison of the distribution of the (C) CMS and (D) CRIS groups among the confidently classified tumors according to analysis platform.

However, the gene-wise correlations were found to vary according to the expression level (Fig. 1C). Genes expressed at low levels (< 1 FPKM in the RNA-seq data) had poor correlations between platforms (HTA: median Spearman's ρ = 0.24, IQR 0.05–0.53; HuEx: median Spearman's ρ = 0.16, IQR 0.01–0.44), while the median Spearman correlation for genes expressed at higher levels (> 1 FPKM) was 0.81 (IQR 0.68–0.89) and 0.71 (IQR 0.56–0.82) on HTA and HuEx arrays, respectively. Grouping of the genes into expression quantiles showed that this effect was continuous, with a steady increase in gene expression correlation with an increase in expression levels up to approximately 1 FPKM (Supplementary Figure S1). An effect of gene length on correspondence between platforms was also found, and there was a progressive increase in the median correlation in gene expression levels with increasing gene lengths up to approximately 4,000 nucleotides, although the distribution in correlation was wide across all tested quantiles of gene lengths (Supplementary Figure S2). For comparison, normalization of the RNA-seq count data by the TPM approach resulted in similar gene-wise correspondences between platforms (HTA: median Spearman's ρ = 0.68, IQR 0.35–0.84; HuEx: median Spearman's ρ = 0.58, IQR 0.24–0.76), and had no impact on the bias associated with gene length and gene expression level (Supplementary Figure S3). To test a potential impact of the gene-length adjustment inherent to expression level normalization by FPKM and TPM calculation, the RNA-seq count data were also normalized by the VST method [43]. This confirmed the good overall gene-wise correspondence (HTA: median Spearman's ρ = 0.72, IQR 0.37–0.86; HuEx: median Spearman's ρ = 0.60, IQR 0.26–0.78), but also confirmed the poor technical robustness for short genes and genes with low expression levels (Supplementary Figure S4).

To further analyze these systematic effects, we focused on genes with discrepant expression measures in the sample-wise comparisons, that is, genes with "underestimated" expression in RNA-seq compared with microarray data (threshold defined based on visual inspection of scatter plots), and *vice versa* (shown as black dots in two example samples in Fig. 1D). Most genes denoted as underestimated by RNA-seq were underestimated in only one or a few samples each, indicating no systematic platform effects (Fig. 1E; Supplementary Table S3). However, 66 (3% of 1,936) and 41 (8% of 502) genes were consistently underestimated by RNA-seq in all samples compared with HuEx and HTA arrays, respectively. Compared with a thousand random gene sets of equal sizes, these genes were more likely to be short (< 2,000 nucleotides; Supplementary Figure S5), which corroborates the results from the gene-wise analyses and indicates that RNA-seq systematically underestimates the expression levels of short genes. For genes denoted as underestimated by microarrays, 19% (44 of 278) and 34% (78 of 232) were consistently underestimated in all HuEx and HTA samples, respectively (Fig. 1F; Supplementary Table S4). A PANTHER overrepresentation test showed that these genes were strongly enriched for "housekeeping" genes in the class "ribosomal protein" (Supplementary Table S5).

3.2. Variation in confident CRC subtyping between microarrays and RNA-seq

To investigate whether CRC subtyping is consistent across technologies, we classified the samples according to both the CMS and CRIS frameworks using both microarray and RNA-seq expression profiles. The representativeness of the tumor series for transcriptomic classification (selected for MSS and mostly stage II and III cancers) was evaluated by comparing the microarray-derived CMS labels with previously published labels for the same samples analyzed as part of a larger, consecutive tumor series (n = 409 primary CRCs) [35]. This showed a subtype agreement of 99% among the samples that were confidently

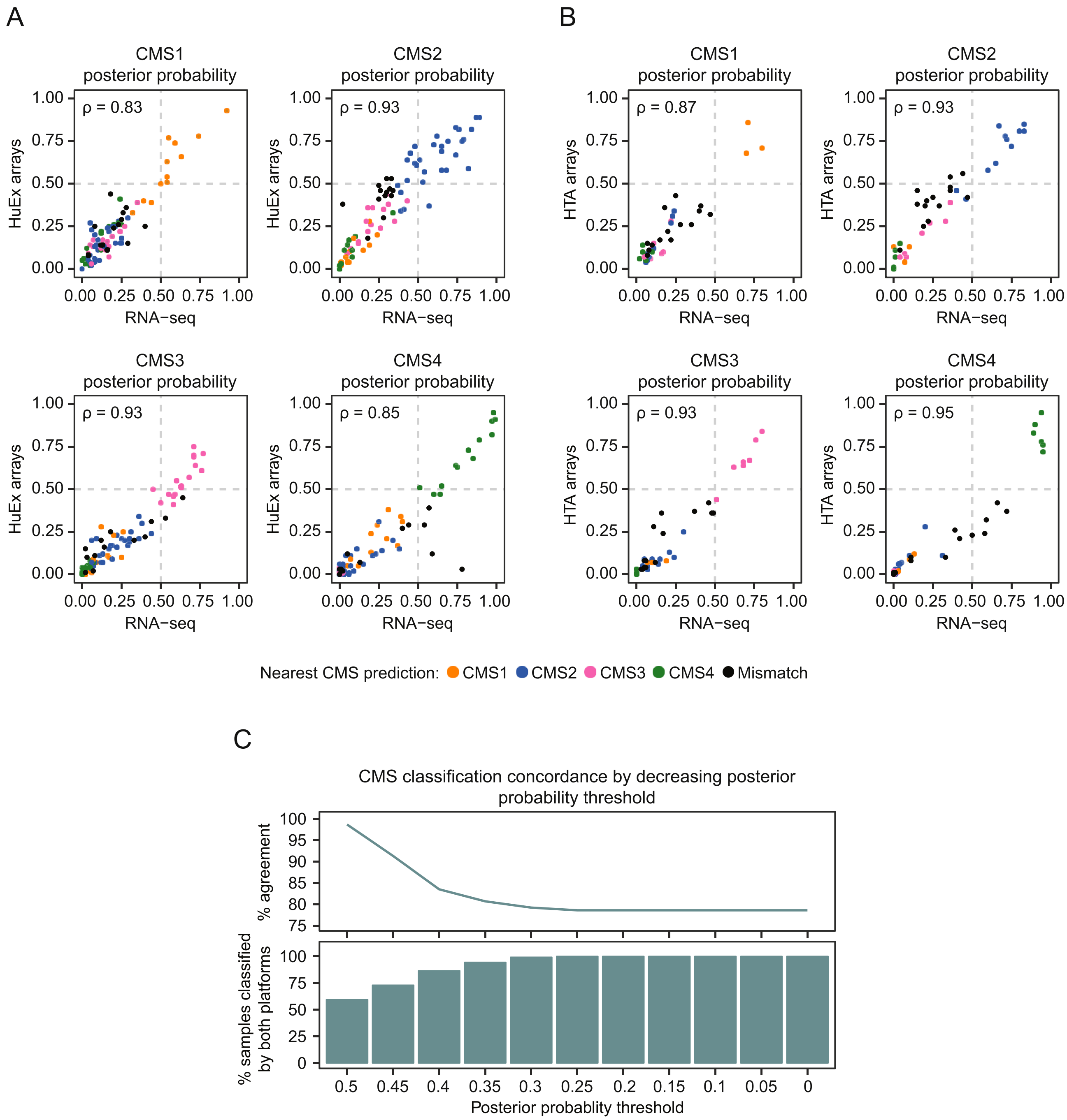

Fig. 3. Sample-wise CMS posterior probability scores by microarrays and RNA-seq. Scatter plots comparing the posterior probabilities assigned to each sample from RNA-seq and either (A) HuEx or (B) HTA data for each of the four CMS groups. The samples are colored according to the nearest predicted CMS subtype as determined by RNA-seq and microarrays, with samples not assigned the same nearest predicted subtype colored in black. The dotted lines at 0.5 represents the default threshold for confident sample classification. (C) Line plot (upper panel) showing the percentage of agreement in CMS classification among samples confidently classified by both platforms according to decreasing posterior probability threshold (bottom panel).

classified in both studies, although 18% of the samples that were assigned a CMS label in the previous study were unclassified in this study. The proportion of the 126 CRCs that were confidently classified (CMS: RF posterior probability ≥0.5; CRIS: $P_{\text{adjusted}} < 0.05$; see Methods for details) by both microarrays and RNA-seq was relatively low in both the CMS and CRIS classification frameworks (CMS: 60%, 75 of 126 samples; CRIS: 73%, 92 of 126 samples). Among these samples, there was a near-perfect subtype agreement of 99% (74 of 75 samples) for CMS (Cohen's κ = 0.98, 95% confidence interval, CI, 0.95–1.02; Fig. 2A; Supplementary Table S6; Supplementary Figure S6A), but only 88% agreement for CRIS (Cohen's κ = 0.84, 95% CI 0.76–0.93; Fig. 2B; Supplementary Table S7; Supplementary Figure S6B). The more frequent discordance in the CRIS framework (11 of 92 samples) was likely related to the larger number of samples designated as confidently classified, although sample classification was based on a more strict threshold than in the original CRIS study [15]. Using the original threshold ($P_{\text{adjusted}} < 0.2$), 80% of the samples were given a CRIS label by both platforms and the classification concordance remained the same (Supplementary Table S8). However, analyses of all samples in the patient series, including the unclassified as a separate group, resulted in reduced cross-platform classification agreements for both CMS and CRIS (CMS: 76%, kappa statistic 0.69 [95% CI 0.60–0.79]; CRIS:

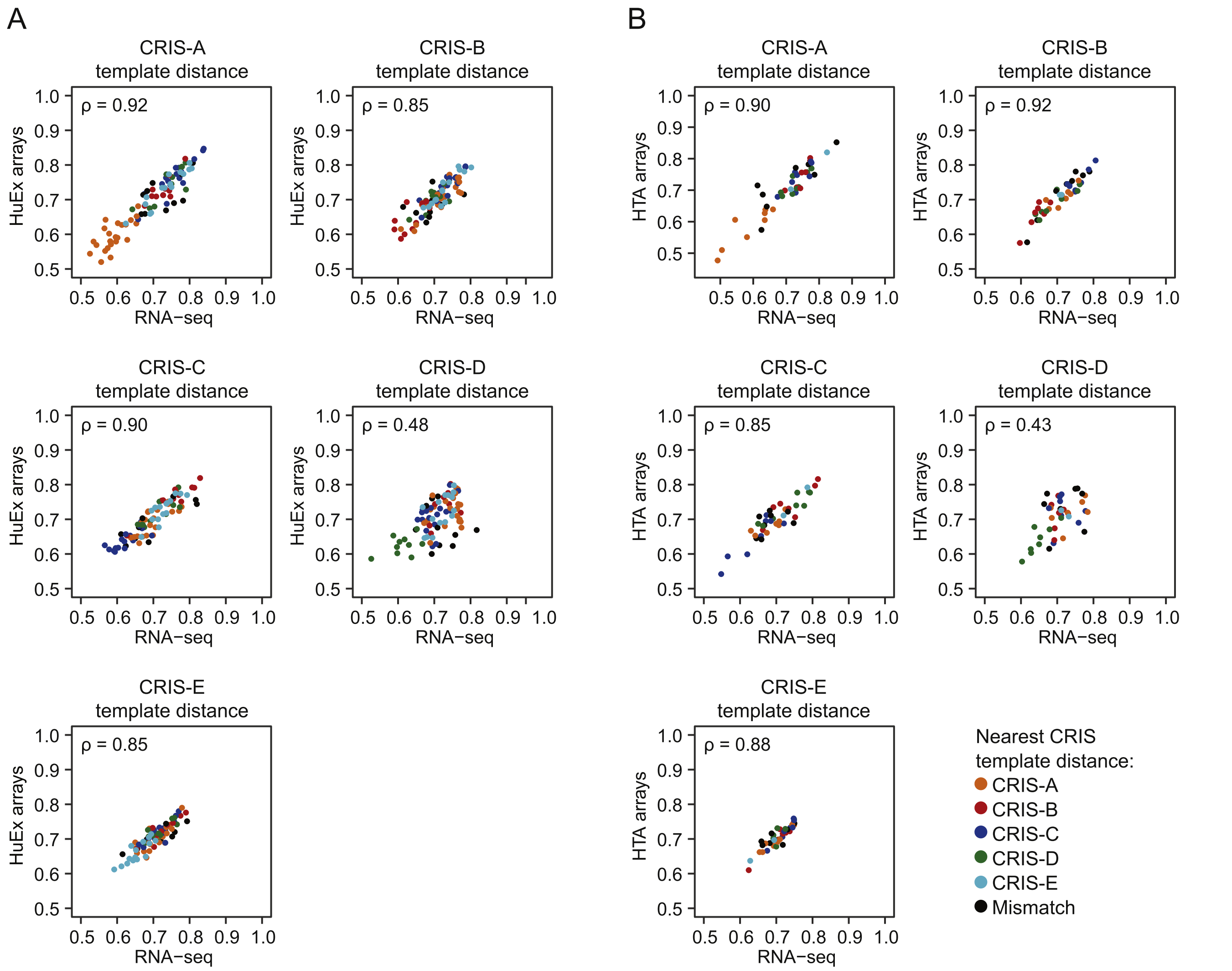

Fig. 4. Sample-wise CRIS template distances by microarray and RNA-seq. Scatter plots comparing the sample-wise distances in the template gene signatures for each of the five CRIS groups between RNA-seq and either (A) HuEx or (B) HTA data. The samples are colored according to the CRIS group with the nearest template distance as determined by both RNA-seq and microarrays, with samples not assigned the same nearest CRIS group colored in black.

79%, kappa statistic 0.75 [95% CI 0.66–0.83]; Supplementary Tables S6-S7). This indicated frequent disagreements between platforms with respect to which samples adhered to the confident sample classification thresholds, and as many as 23% and 12% of the tumors were confidently classified by only one platform for CMS and CRIS, respectively. This resulted in a pronounced difference in the subtype distributions of confidently classified tumors according to analysis platform, in particular in the CMS framework (Fig. 2C and D). In the RNA-seq data, the prevalence of CMS2 and CMS4 was similar (31% and 30%, respectively), while CMS2 was more prevalent than CMS4 in the microarray data (45% and 20%, respectively). Classification of tumor samples using either TPM or VST normalized RNA-seq count data resulted in similar cross-platform classification concordances and subtype distributions compared with FPKM normalization, indicating minimal impact of RNA-seq normalization method on both classification frameworks (Supplementary Tables S9-S12; Supplementary Figures S7-8).

For closer investigation, we compared the posterior probability scores of each CMS group across samples. Irrespective of the subtype call of the samples, the posterior probabilities from the microarray and RNA-seq data had a strong correlation for all four CMS groups (Spearman's ρ range: 0.83–0.95; Fig. 3A and B). This indicates that discordances were caused by only subtle platform differences, except for two samples with a strong discordance in CMS4 scores. Nonetheless, a lowering of the posterior probability threshold to allow subtype calling of more samples resulted in a reduction in the classification concordance (Fig. 3C), terminating in a classification concordance of 79% among all 126 samples (Cohen's κ = 0.70, 95% CI 0.61–0.80), with a corresponding posterior probability ≤0.25. Corresponding analyses were performed for CRIS classification, in which each sample was assigned a correlation distance to each of the five template gene signatures. Correlation distances were proportional between the two expression platforms for four of the five CRIS subtypes (Spearman's ρ range: 0.85–0.92). Surprisingly, CRIS-D had a poor correlation (Spearman's ρ 0.48 on HuEx arrays and 0.43 on HTA arrays; Fig. 4A and B).

3.3. *Low gene expression levels contribute to discrepancies in transcriptomic classification*

We investigated whether the weaker correspondence in CRIS-D template distances were related to the expression levels of the template genes, and found that the CRIS-D genes had significantly lower expression levels than the genes in the other four CRIS template signatures (Kruskal-Wallis test, $P < 0.001$; Fig. 5A). Consistently, the CRIS-D genes also had poorer and more varying cross-platform correlations (HuEx: median Spearman's ρ = 0.71, IQR 0.47–0.88; HTA: median Spearman's ρ = 0.78, IQR 0.59–0.93; Fig. 5B). Exclusion of both

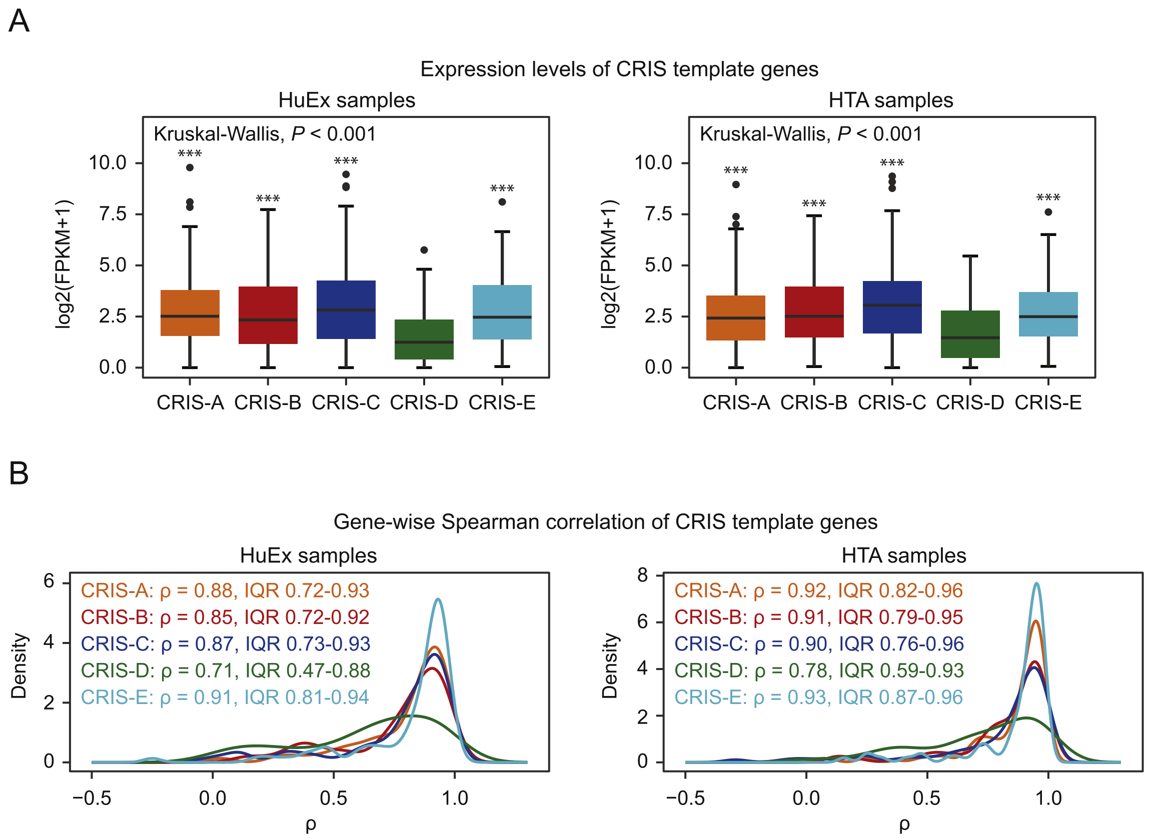

Fig. 5. Expression of CRIS template genes. (A) Box plots showing the expression levels of genes constituting the five CRIS template gene signatures as measured by RNA-seq. The expression level of each gene represents the median log2-transformed FPKM value across samples after dividing the RNA-seq data according to samples analyzed on HuEx and HTA arrays. ****P* < 0.001 compared to the CRIS-D group from Dunn's multiple comparisons test with Benjamini-Hochberg correction. (B) Density distribution of gene expression correlations of the CRIS genes between microarrays and RNA-seq, stratified according to their respective CRIS templates.

lowly expressed (< 1 FPKM) and short (< 2,000 nucleotides) genes from the CRIS template gene sets resulted in stronger correlations in template distances between platforms, especially for the CRIS-D subtype (Spearman's ρ 0.79 on HuEx arrays and 0.80 on HTA arrays), as well as more proportionate subtype distributions (Supplementary Table S13; Supplementary Figure S9). In the CMS framework, however, corresponding analyses had minimal impact on the already strong cross-platform correspondences in posterior probabilities and did not improve the correspondence in subtype distributions, but resulted in a higher proportion of unclassified samples for both technologies (17%, 22 of 126 samples versus 33%, 41 of 126 samples; Supplementary Table S14; Supplementary Figure S10).

To test if the association between technical robustness and gene expression levels also applied to other clinically relevant gene expression characteristics, we further measured the abundances of tumor infiltrating cytotoxic lymphocytes and cancer-associated fibroblasts using the MCPcounter algorithm [45]. Low expression levels of cytotoxic lymphocyte signature genes compared with fibroblast signature genes was confirmed in our set of MSS tumor samples (Mann-Whitney *U* test, *P* = 0.001; Fig. 6A). The cytotoxic lymphocyte signature genes also had weaker cross-platform correlations (HuEx: median Spearman's ρ = 0.49, IQR 0.25–0.65; HTA: median Spearman's ρ = 0.62, IQR 0.48–0.70; Fig. 6B). Comparisons of the estimated abundances showed a strong cross-platform correspondence for the fibroblasts (Spearman's ρ 0.96 on HuEx arrays and 0.98 on HTA arrays), and all samples with CMS4 as their nearest CMS subtype had a high level of fibroblast infiltration (Fig. 6C). The corresponding correlations for the less abundant cytotoxic lymphocytes were only 0.60 and 0.74 between RNA-seq and HuEx and HTA arrays, respectively (Fig. 6D).

4. Discussion

This study confirms a strong correspondence in the global gene expression profiles generated by RNA-seq and exon-resolution microarrays [25,30,31] in a series of 126 MSS CRCs, both in sample-wise comparisons of all genes and in gene-wise comparisons of all samples. However, even modest differences between the gene expression platforms had an impact on tumor classification according to clinically relevant frameworks, primarily related to a variable adherence to sample classification thresholds. This resulted in platform-dependent subtype distributions within the analyzed set of tumors, which is of potential clinical importance due to a likely influence on the analyses of subtype-dependent therapeutic benefits [17]. Both the CMS and CRIS frameworks were originally shown to accommodate several gene expression platforms, including Affymetrix microarrays and RNA-seq [12,15]. The strong correspondence in subtype probability metrics across the CRCs in our study confirms the appropriateness of both expression platforms, and the observed classification discordances therefore primarily highlight the need for assay-standardization prior to clinical translation. However, for development of standardized assays to prospectively stratify patients in clinical trials, it will be important to keep in mind that lowering of the classification thresholds resulted in an inferior classification concordance between platforms, down to a subtype concordance of 79% when forcing CMS classification on all tumors. Clinical applicability is also likely to increase with accommodation of formalin-fixed paraffin-embedded samples, for which assays based on the NanoString platform have shown promise [46]. Adaptation of the template gene sets, both to the sample type (formalin-fixed paraffin-embedded or fresh frozen) and analysis platform, may be needed.

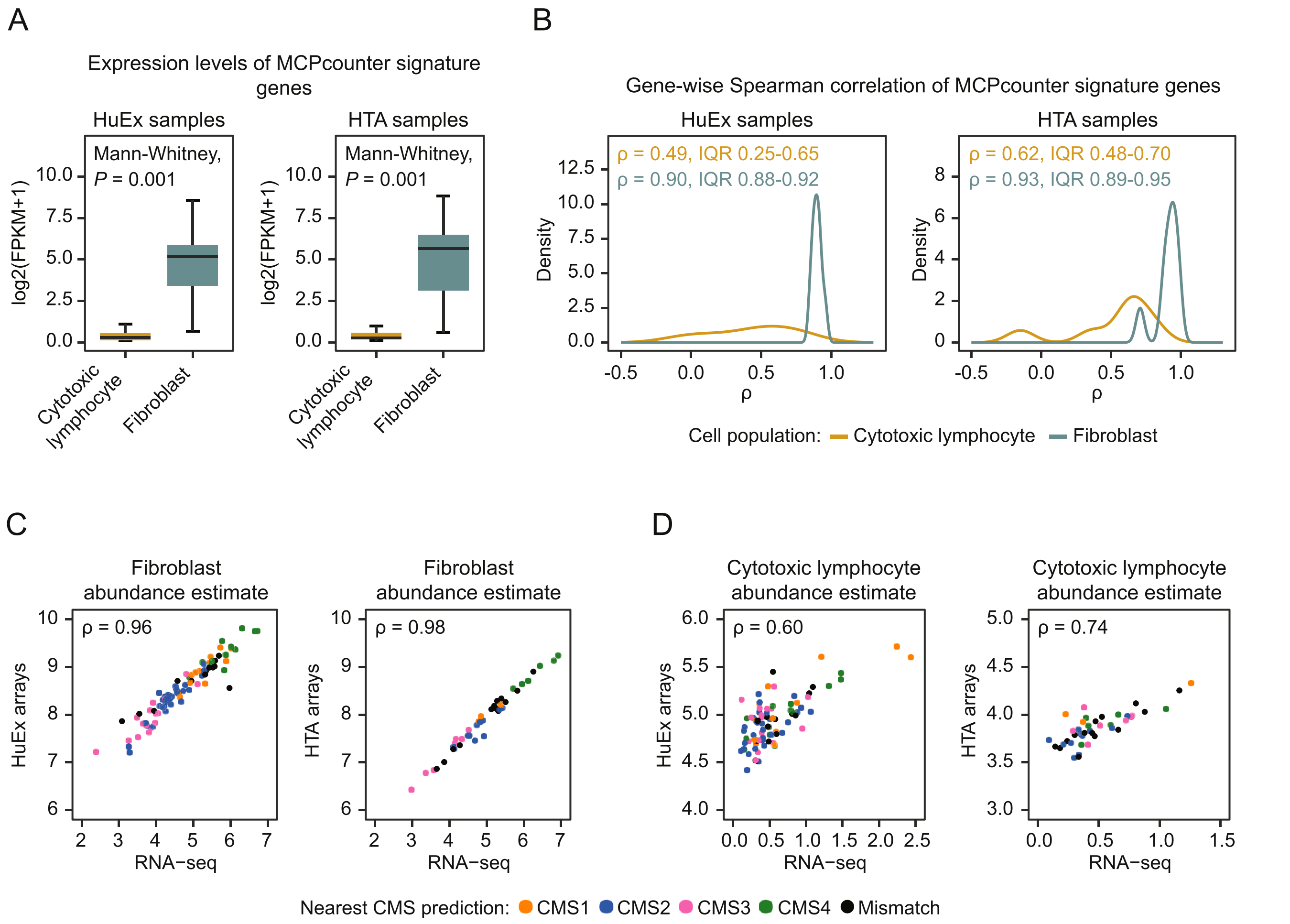

Fig. 6. Sample-wise fibroblast and cytotoxic lymphocyte abundance estimates by microarrays and RNA-seq. (A) Box plots showing the expression levels of genes constituting the cytotoxic lymphocyte and fibroblast signatures as measured by RNA-seq. The expression level of each gene represents the median log2-transformed FPKM value across samples after dividing the RNA-seq data according to samples analyzed on HuEx and HTA arrays. (B) Density distribution of gene expression correlations of the cytotoxic lymphocyte and fibroblast genes, stratified according to their respective cell population. Scatter plots comparing the estimated abundances of (C) cancer-associated fibroblasts and (D) tumor infiltrating cytotoxic lymphocytes in each sample between RNA-seq and microarrays. The samples are colored according to the nearest predicted CMS subtype as determined by RNA-seq and microarrays, with samples not assigned the same nearest predicted subtype colored in black.

This study confirmed the previously reported poor technical robustness between microarray and RNA-seq platforms for the quantification of genes with low expression levels (< 1 FPKM) [29–31]. We further demonstrated that this had a strong effect on the classification metrics of the CRIS-D subtype, as well as on the measured abundances of tumor infiltrating cytotoxic lymphocytes, which was expected to be low in this series of MSS CRCs [47]. Furthermore, we also observed a systematic bias likely resulting from underestimation of short genes by RNA-seq. Notably, this consistently affected a small set of genes across all the 126 samples. Low expression levels by RNA-seq relative to microarrays could also be the result of cross-hybridization of microarray probes to multiple targets. However, our analyses strongly suggest an association with gene length (< 2,000 nucleotides), which is consistent with previous studies [30,48]. Both short genes and genes expressed at low levels will produce less sequencing reads than longer genes or genes expressed at higher levels, and studies of both technical and biological replicates have shown that the variability in quantification of low-level gene expression is higher with RNA-seq compared with microarrays [25,28,30]. Although normalization of read counts by transcript lengths (as with FPKM and TPM estimates) might increase the technical variability of short transcripts compared to longer transcripts [20], testing of an alternative expression normalization method not dependent on scaling read counts by gene lengths (VST) confirmed the poor technical robustness of short genes. Consequently, the discordant expression measures of short genes are also likely a result of technical bias in the sequencing platform. The improved correspondences in both the CRIS-D classification metrics and the CRIS subtype distributions observed after excluding short and lowly expressed genes from the template gene sets suggest that such genes should be avoided in further development of classification algorithms. However, exclusion of short and lowly expressed template genes should not be done at the expense of markers with a strong subtype discriminatory power, as illustrated by a reduction in the proportion of confidently classified samples in the CMS framework. This may argue for development of novel classifier template gene sets rather than filtering of existing gene sets.

Systematic bias was also observed as underestimation of gene expression levels by microarrays relative to RNA-seq, consistently affecting a small proportion of genes encoding ribosomal proteins. These genes may constitute a significant proportion of highly expressed "housekeeping" genes, thus producing a high number of sequencing reads with RNA-seq and simultaneously saturating the probes on the microarray. Although over-saturation may be expected to affect all highly expressed genes, saturation levels have been shown to vary considerably amongst probes on microarrays [49]. Furthermore, all comparisons showed that HTA arrays had better correspondence with RNA-seq than HuEx arrays, suggesting that the probe densities on the microarrays also had an impact on the accuracy of the gene expression measurements. The HTA arrays have an average of 109 probes targeting each gene [50], compared with 40 probes on the HuEx arrays [51]. GC-content correction in combination with standard RMA pre-processing of HTA data may further increase the correspondence in the expression profiles of protein-coding genes with RNA-seq (Spearman's ρ = 0.782) [30], however, we obtained the same high level of correlation using standard RMA (Spearman's ρ = 0.82).

This is to our knowledge the largest study to analyze classification concordances of CRC subtypes between microarray and RNA sequencing platforms, although custom assays based on the NanoString platform have been compared with either microarray- or RNA sequencing-derived classification of smaller sets of matched samples [46,52]. CRC subtype distributions are known to vary according to both cancer stage and genomic variables [17]. In this respect, we cannot dismiss the possibility that the presence of only MSS tumors and enrichment with stage II and III cancers in our patient series may have contributed to prediction inaccuracies. However, representativeness of the tumor series for transcriptomic classification was shown by correspondence with the microarray-derived CMS labels of the same tumors previously analyzed as part of a larger, unselected patient series.

In conclusion, we showed that despite good overall concordance in gene expression levels between exon-resolution microarrays and RNA-seq, the two technologies differed in their sensitivity for genes represented by few RNA-seq reads (short genes and/or genes with low expression levels) and for highly expressed genes likely prone to over-saturation of microarray features. We further demonstrated that such platform differences had an impact on subtype distributions according to clinically relevant classification frameworks, and we suggest that the technical robustness of tumor subtyping may be improved by avoiding genes subjected to systematic technical biases.

Ethical approval and consent to participate

The study was approved by the Regional Committee for Medical and Health Research Ethics, South Eastern Norway (REC number 1.2005.1629). All patients provided written informed content, and the study was conducted in accordance with the Declaration of Helsinki.

Funding

This work was supported by the Norwegian Cancer Society, Norway [grant numbers 6824048-2016, 182759-2016]; the Research Council of Norway, Norway [grant number 250993]; and the South-Eastern Norway Regional Health Authority, Norway [grant number 2016003].

Declaration of competing interest

The authors declare no potential conflict of interest.

Acknowledgements

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.canlet.2019.10.040.

References

[1] R. Dienstmann, L. Vermeulen, J. Guinney, S. Kopetz, S. Tejpar, J. Tabernero, Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer, Nat. Rev. Cancer 17 (2) (2017) 79–92.

[2] J. Rodon, J.C. Soria, R. Berger, W.H. Miller, E. Rubin, A. Kugel, A. Tsimberidou, P. Saintigny, A. Ackerstein, I. Brana, Y. Loriot, M. Afshar, V. Miller, F. Wunder, C. Bresson, J.F. Martini, J. Raynaud, J. Mendelsohn, G. Batist, A. Onn, J. Tabernero, R.L. Schilsky, V. Lazar, J.J. Lee, R. Kurzrock, Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial, Nat. Med. 25 (5) (2019) 751–758.

[3] A. Calon, E. Lonardo, A. Berenguer-Llergo, E. Espinet, X. Hernando-Momblona, M. Iglesias, M. Sevillano, S. Palomo-Ponce, D.V. Tauriello, D. Byrom, C. Cortina, C. Morral, C. Barcelo, S. Tosi, A. Riera, C.S. Attolini, D. Rossell, E. Sancho, E. Batlle, Stromal gene expression defines poor-prognosis subtypes in colorectal cancer, Nat. Genet. 47 (4) (2015) 320–329.

[4] C. Isella, A. Terrasi, S.E. Bellomo, C. Petti, G. Galatola, A. Muratore, A. Mellano, R. Senetta, A. Cassenti, C. Sonetto, G. Inghirami, L. Trusolino, Z. Fekete, M. De Ridder, P. Cassoni, G. Storme, A. Bertotti, E. Medico, Stromal contribution to the colorectal cancer transcriptome, Nat. Genet. 47 (4) (2015) 312–319.

[5] R. Dienstmann, G. Villacampa, A. Sveen, M.J. Mason, D. Niedzwiecki, A. Nesbakken, V. Moreno, R.S. Warren, R.A. Lothe, J. Guinney, Relative contribution of clinicopathological variables, genomic markers, transcriptomic subtyping and microenvironment features for outcome prediction in stage II/III colorectal cancer, Ann. Oncol. (2019), https://doi.org/10.1093/annonc/mdz287.

[6] E. Budinska, V. Popovici, S. Tejpar, G. D'Ario, N. Lapique, K.O. Sikora, A.F. Di Narzo, P. Yan, J.G. Hodgson, S. Weinrich, F. Bosman, A. Roth, M. Delorenzi, Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer, J. Pathol. 231 (1) (2013) 63–76.

[7] L. Marisa, A. de Reynies, A. Duval, J. Selves, M.P. Gaub, L. Vescovo, M.C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J.F. Flejou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig, V. Boige, Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value, PLoS Med. 10 (5) (2013) e1001453.

[8] P. Roepman, A. Schlicker, J. Tabernero, I. Majewski, S. Tian, V. Moreno, M.H. Snel, C.M. Chresta, R. Rosenberg, U. Nitsche, T. Macarulla, G. Capella, R. Salazar, G. Orphanides, L.F. Wessels, R. Bernards, I.M. Simon, Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition, Int. J. Cancer 134 (3) (2014) 552–562.

[9] E.M.F. De Sousa, X. Wang, M. Jansen, E. Fessler, A. Trinh, L.P. de Rooij, J.H. de Jong, O.J. de Boer, R. van Leersum, M.F. Bijlsma, H. Rodermond, M. van der Heijden, C.J. van Noesel, J.B. Tuynman, E. Dekker, F. Markowetz, J.P. Medema, L. Vermeulen, Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions, Nat. Med. 19 (5) (2013) 614–618.

[10] A. Sadanandam, C.A. Lyssiotis, K. Homicsko, E.A. Collisson, W.J. Gibb, S. Wullschleger, L.C. Ostos, W.A. Lannon, C. Grotzinger, M. Del Rio, B. Lhermitte, A.B. Olshen, B. Wiedenmann, L.C. Cantley, J.W. Gray, D. Hanahan, A colorectal cancer classification system that associates cellular phenotype and responses to therapy, Nat. Med. 19 (5) (2013) 619–625.

[11] A. Schlicker, G. Beran, C.M. Chresta, G. McWalter, A. Pritchard, S. Weston, S. Runswick, S. Davenport, K. Heathcote, D.A. Castro, G. Orphanides, T. French, L.F. Wessels, Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines, BMC Med. Genomics 5 (2012) 66.

[12] J. Guinney, R. Dienstmann, X. Wang, A. de Reynies, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B.M. Bot, J.S. Morris, I.M. Simon, S. Gerster, E. Fessler, E.M.F. De Sousa, E. Missiaglia, H. Ramay, D. Barras, K. Homicsko, D. Maru, G.C. Manyam, B. Broom, V. Boige, B. Perez-Villamil, T. Laderas, R. Salazar, J.W. Gray, D. Hanahan, J. Tabernero, R. Bernards, S.H. Friend, P. Laurent-Puig, J.P. Medema, A. Sadanandam, L. Wessels, M. Delorenzi, S. Kopetz, L. Vermeulen, S. Tejpar, The consensus molecular subtypes of colorectal cancer, Nat. Med. 21 (11) (2015) 1350–1356.

[13] P.D. Dunne, D.G. McArt, C.A. Bradley, P.G. O'Reilly, H.L. Barrett, R. Cummins, T. O'Grady, K. Arthur, M.B. Loughrey, W.L. Allen, S.S. McDade, D.J. Waugh, P.W. Hamilton, D.B. Longley, E.W. Kay, P.G. Johnston, M. Lawler, M. Salto-Tellez, S. Van Schaeybroeck, Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer, Clin. Cancer Res. 22 (16) (2016) 4095–4104.

[14] P.D. Dunne, M. Alderdice, P.G. O'Reilly, A.C. Roddy, A.M.B. McCorry, S. Richman, T. Maughan, S.S. McDade, P.G. Johnston, D.B. Longley, E. Kay, D.G. McArt, M. Lawler, Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification, Nat. Commun. 8 (2017) 15657.

[15] C. Isella, F. Brundu, S.E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, A. Bertotti, Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer, Nat. Commun. 8 (2017) 15107.

[16] D. Aderka, S. Stintzing, V. Heinemann, Explaining the unexplainable: discrepancies in results from the CALGB/SWOG 80405 and FIRE-3 studies, Lancet Oncol. 20 (5) (2019) e274–e283.

[17] E. Fontana, K. Eason, A. Cervantes, R. Salazar, A. Sadanandam, Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials, Ann. Oncol. 30 (4) (2019) 520–527.

[18] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (1) (2009) 57–63.

[19] J.K. Rhee, S. Lee, W.Y. Park, Y.H. Kim, T.M. Kim, Allelic imbalance of somatic mutations in cancer genomes and transcriptomes, Sci. Rep. 7 (1) (2017) 1653.

[20] A. Oshlack, M.J. Wakefield, Transcript length bias in RNA-seq data confounds systems biology, Biol. Direct 4 (2009) 14.

[21] L. Jiang, F. Schlesinger, C.A. Davis, Y. Zhang, R. Li, M. Salit, T.R. Gingeras, B. Oliver, Synthetic spike-in standards for RNA-seq experiments, Genome Res. 21 (9) (2011) 1543–1551.

[22] P. Cui, Q. Lin, F. Ding, C. Xin, W. Gong, L. Zhang, J. Geng, B. Zhang, X. Yu, J. Yang, S. Hu, J. Yu, A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing, Genomics 96 (5) (2010) 259–265.

[23] Y. Han, S. Gao, K. Muegge, W. Zhang, B. Zhou, Advanced applications of RNA sequencing and challenges, Bioinf. Biol. Insights 9 (Suppl 1) (2015) 29–46.

[24] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, Y. Gilad, RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, Genome Res. 18 (9) (2008) 1509–1517.

[25] N. Raghavachari, J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, C.J. O'Donnell, P.J. Munson, G.J. Kato, A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease, BMC Med. Genomics 5 (2012) 28.

[26] Y. Guo, Q. Sheng, J. Li, F. Ye, D.C. Samuels, Y. Shyr, Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data, PLoS One 8 (8) (2013) e71462.

[27] X. Xu, Y. Zhang, J. Williams, E. Antoniou, W.R. McCombie, S. Wu, W. Zhu, N.O. Davidson, P. Denoya, E. Li, Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets, BMC Bioinf. 14 (Suppl 9) (2013) S1.

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

[29] D. Fumagalli, A. Blanchet-Cohen, D. Brown, C. Desmedt, D. Gacquer, S. Michiels, F. Rothe, S. Majjaj, R. Salgado, D. Larsimont, M. Ignatiadis, M. Maetens, M. Piccart, V. Detours, C. Sotiriou, B. Haibe-Kains, Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology, BMC Genomics 15 (2014) 1008.

[30] P.V. Nazarov, A. Muller, T. Kaoma, N. Nicot, C. Maximo, P. Birembaut, N.L. Tran, G. Dittmar, L. Vallar, RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples, BMC Genomics 18 (1) (2017) 443.

[31] J.P. Romero, M. Ortiz-Estevez, A. Muniategui, S. Carrancio, F.J. de Miguel, F. Carazo, L.M. Montuenga, R. Loos, R. Pio, M.W.B. Trotter, A. Rubio, Comparison of RNA-seq and microarray platforms for splice event detection using a cross-platform algorithm, BMC Genomics 19 (1) (2018) 703.

[32] A. Sveen, T.H. Agesen, A. Nesbakken, G.I. Meling, T.O. Rognum, K. Liestol, R.I. Skotheim, R.A. Lothe, ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients, Clin. Cancer Res. 18 (21) (2012) 6001–6010.

[33] T.H. Agesen, A. Sveen, M.A. Merok, G.E. Lind, A. Nesbakken, R.I. Skotheim, R.A. Lothe, ColoGuideEx: a robust gene classifier specific for stage II colorectal cancer prognosis, Gut 61 (11) (2012) 1560–1567.

[34] A.M. Hoff, B. Johannessen, S. Alagaratnam, S. Zhao, T. Nome, M. Lovf, A.C. Bakken, M. Hektoen, A. Sveen, R.A. Lothe, R.I. Skotheim, Novel RNA variants in colorectal cancers, Oncotarget 6 (34) (2015) 36587–36602.

[35] A. Sveen, J. Bruun, P.W. Eide, I.A. Eilertsen, L. Ramirez, A. Murumagi, M. Arjama, S.A. Danielsen, K. Kryeziu, E. Elez, J. Tabernero, J. Guinney, H.G. Palmer, A. Nesbakken, O. Kallioniemi, R. Dienstmann, R.A. Lothe, Colorectal cancer consensus molecular subtypes translated to preclinical models uncover potentially targetable cancer cell dependencies, Clin. Cancer Res. 24 (4) (2018) 794–806.

[36] A. Sveen, T.H. Agesen, A. Nesbakken, T.O. Rognum, R.A. Lothe, R.I. Skotheim, Transcriptome instability in colorectal cancer identified by exon microarray analyses: associations with splicing factor expression levels and patient survival, Genome Med. 3 (5) (2011) 32.

[37] M. Dai, P. Wang, A.D. Boyd, G. Kostov, B. Athey, E.G. Jones, W.E. Bunney, R.M. Myers, T.P. Speed, H. Akil, S.J. Watson, F. Meng, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, Nucleic Acids Res. 33 (20) (2005) e175.

[38] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, T.J. Hubbard, GENCODE: the reference human genome annotation for the ENCODE Project, Genome Res. 22 (9) (2012) 1760–1774.

[39] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.

[40] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (1) (2013) 15–21.

[41] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.

[42] S. Anders, P.T. Pyl, W. Huber, HTSeq–a Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2) (2015) 166–169.

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

[44] P.W. Eide, J. Bruun, R.A. Lothe, A. Sveen, CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models, Sci. Rep. 7 (1) (2017) 16618.

[45] E. Becht, N.A. Giraldo, L. Lacroix, B. Buttard, N. Elarouci, F. Petitprez, J. Selves, P. Laurent-Puig, C. Sautes-Fridman, W.H. Fridman, A. de Reynies, Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression, Genome Biol. 17 (1) (2016) 218.

[46] C. Ragulan, K. Eason, E. Fontana, G. Nyamundanda, N. Tarazona, Y. Patil, P. Poudel, R.T. Lawlor, M. Del Rio, S.L. Koo, W.S. Tan, F. Sclafani, R. Begum, L.S. Teixeira Mendes, P. Martineau, A. Scarpa, A. Cervantes, I.B. Tan, D. Cunningham, A. Sadanandam, Analytical validation of multiplex biomarker assay to stratify colorectal cancer into molecular subtypes, Sci. Rep. 9 (1) (2019) 7665.

[47] R. Dolcetti, A. Viel, C. Doglioni, A. Russo, M. Guidoboni, E. Capozzi, N. Vecchiato, E. Macri, M. Fornasarig, M. Boiocchi, High prevalence of activated intraepithelial cytotoxic T lymphocytes and increased neoplastic cell apoptosis in colorectal carcinomas with microsatellite instability, Am. J. Pathol. 154 (6) (1999) 1805–1813.

[48] H. Rehrauer, L. Opitz, G. Tan, L. Sieverling, R. Schlapbach, Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching, BMC Bioinf. 14 (2013) 370.

[49] D. Skvortsov, D. Abdueva, C. Curtis, B. Schaub, S. Tavare, Explaining differences in saturation levels for Affymetrix GeneChip arrays, Nucleic Acids Res. 35 (12) (2007) 4154–4163.

[50] Affymetrix, DataSheet: GeneChip human transcriptome array 2.0, http://tools.thermofisher.com/content/sfs/brochures/hta_array_2_0_datasheet.pdf , Accessed date: 3 April 2019.

[51] Affymetrix, Application focus: whole-transcript expression analysis, Gene Expr. (2019), https://assets.thermofisher.com/TFS-Assets/LSG/brochures/wt_appnote.pdf , Accessed date: 3 April 2019.

[52] R. Piskol, L. Huw, I. Sergin, C. Kljin, Z. Modrusan, D. Kim, N. Kljavin, R. Tam, R. Patel, J. Burton, E. Penuel, X. Qu, H. Koeppen, T. Sumiyoshi, F. de Sauvage, M.R. Lackner, E.M.F. de Sousa, O. Kabbarah, A clinically applicable gene-expression classifier reveals intrinsic and extrinsic contributions to consensus molecular subtypes in primary and metastatic colon cancer, Clin. Cancer Res. 25 (14) (2019) 4431–4442.