



# Quartet RNA reference materials improve the quality of transcriptomic data through ratio-based profiling

Received: 28 October 2022

Accepted: 15 June 2023

Published online: 07 September 2023

Check for updates

Ying Yu <sup>1,17</sup>, Wanwan Hou <sup>1,17</sup>, Yaqing Liu <sup>1,17</sup>, Haiyan Wang <sup>1,17</sup>, Lianhua Dong <sup>2,17</sup>, Yuanbang Mai<sup>1</sup>, Qingwang Chen <sup>1</sup>, Zhihui Li<sup>1</sup>, Shanyue Sun<sup>1</sup>, Jingcheng Yang <sup>1,3</sup>, Zehui Cao<sup>1</sup>, Peipei Zhang<sup>1</sup>, Yi Zi<sup>1</sup>, Ruimei Liu<sup>1</sup>, Jian Gao<sup>1</sup>, Naixin Zhang<sup>1</sup>, Jingjing Li <sup>1,4</sup>, Luyao Ren<sup>1</sup>, He Jiang <sup>1</sup>, Jun Shang <sup>1</sup>, Sibo Zhu<sup>1</sup>, Xiaolin Wang<sup>1</sup>, Tao Qing<sup>1</sup>, Ding Bao<sup>1</sup>, Bingying Li<sup>1</sup>, Bin Li<sup>1</sup>, Chen Suo<sup>1</sup>, Yan Pi<sup>1</sup>, Xia Wang<sup>2</sup>, Fangping Dai<sup>5</sup>, Andreas Scherer <sup>6,7</sup>, Pirkko Mattila <sup>6,7</sup>, Jinxiong Han<sup>8</sup>, Lijun Zhang<sup>8</sup>, Hui Jiang<sup>9</sup>, Danielle Thierry-Mieg <sup>10</sup>, Jean Thierry-Mieg <sup>10</sup>, Wenming Xiao <sup>11</sup>, Huixiao Hong <sup>12</sup>, Weida Tong <sup>12</sup>, Jing Wang<sup>2</sup>, Jinming Li <sup>13,14</sup>, Xiang Fang<sup>2</sup>, Li Jin <sup>1</sup>, Joshua Xu <sup>12</sup>✉, Feng Qian <sup>1,15</sup>✉, Rui Zhang <sup>13,14</sup>✉, Leming Shi <sup>1,16</sup>✉ & Yuanting Zheng <sup>1</sup>✉

Certified RNA reference materials are indispensable for assessing the reliability of RNA sequencing to detect intrinsically small biological differences in clinical settings, such as molecular subtyping of diseases. As part of the Quartet Project for quality control and data integration of multi-omics profiling, we established four RNA reference materials derived from immortalized B-lymphoblastoid cell lines from four members of a monozygotic twin family. Additionally, we constructed ratio-based transcriptome-wide reference datasets between two samples, providing cross-platform and cross-laboratory ‘ground truth’. Investigation of the intrinsically subtle biological differences among the Quartet samples enables sensitive assessment of cross-batch integration of transcriptomic measurements at the ratio level. The Quartet RNA reference materials, combined with the ratio-based reference datasets, can serve as unique resources for assessing and improving the quality of transcriptomic data in clinical and biological settings.

RNA sequencing (RNA-seq) is an indispensable tool for transcriptome-wide analysis of differential gene expression and is widely used in biomedical research to discover biomarkers for clinical diagnosis, prognosis and therapeutic action<sup>1–5</sup>. As transcriptome-based biomarker discovery continues to advance, RNA-seq-based assays will routinely be used within the clinic<sup>3,6,7</sup>. For example, clinical tests complemented by measuring the differential expression of clinically relevant genes will facilitate the prediction of clinical outcomes and treatment

decisions<sup>8–10</sup>. It should be noticed that clinically relevant differences in gene expression among study groups are often small<sup>11–13</sup>. Hence, there is a consistent need for making RNA-seq more reliable to enhance its power of detecting subtle differential expression, especially for clinical applications such as companion diagnostics and prognostics. The reliability of RNA-seq technology comprises two aspects. It must be ensured that data from a certain laboratory or batch are acquired with the best proficiency obtainable with the technology (intra-batch)<sup>14</sup>, and

A full list of affiliations appears at the end of the paper. ✉ e-mail: [joshua.xu@fda.hhs.gov](mailto:joshua.xu@fda.hhs.gov); [fengqian@fudan.edu.cn](mailto:fengqian@fudan.edu.cn); [ruizhang@nccl.org.cn](mailto:ruizhang@nccl.org.cn); [lemingshi@fudan.edu.cn](mailto:lemingshi@fudan.edu.cn); [zhengyuanting@fudan.edu.cn](mailto:zhengyuanting@fudan.edu.cn)

similar differential expression results from replicate samples processed with different platforms, laboratories, protocols or batches should be required (cross-batch)<sup>15</sup>. Cross-batch reproducibility also refers to multi-batch integrability, which is the ability to provide similar results between within-batch analysis and cross-batch integrative analysis in the existence of widespread batch effects<sup>16,17</sup>.

Reference materials are valuable tools for evaluating the reliability of omic data<sup>18,19</sup>. Based on RNA-seq data generated with reference materials from different platforms, laboratories or batches, reliability can be objectively evaluated according to the two aforementioned aspects of intra-batch (or laboratory) proficiency and cross-batch reproducibility. The MicroArray/Sequencing Quality Control (MAQC/SEQC) consortia previously established two publicly available transcriptome-wide RNA reference materials that are derived from 10 cancer cell lines and brain tissues of 23 donors<sup>15</sup>. Based on these RNA reference materials, the MAQC/SEQC consortia systematically evaluated the performances of different platforms and laboratories in using the microarray<sup>15</sup> and RNA-seq<sup>20,21</sup> technologies, which have served as resources for the research community to develop and validate new RNA quantification technologies<sup>22</sup>.

However, the ability to successfully distinguish the two MAQC RNA reference materials does not guarantee that the underlying transcriptomic profiling system can be used to detect subtle differential expression for clinical diagnosis purposes. First, the considerable biological differences between the two MAQC reference materials<sup>23</sup> are substantially greater than groupwise differences commonly seen in most clinically relevant scenarios. Second, the ability of distinguishing two MAQC RNA sample groups does not translate to the ability of reliably distinguishing more than two sample groups as commonly seen in clinical applications. Third, the current stock of the MAQC B sample is almost exhausted<sup>24</sup>, and it is difficult to be regenerated. Therefore, there is an urgent need for a multiple-group RNA reference materials suite with subtle inter-sample differences, high stability, long-term availability and easy manufacturability.

Furthermore, reference datasets can be used as 'ground truth' in performance assessment. Previous studies have shown that genome-wide reference datasets of genetic variants enable improvement of the reproducibility and accuracy of clinical applications of cancer<sup>25–27</sup> and genetic diseases<sup>28–30</sup>. However, there is a paucity of transcriptome-wide reference datasets<sup>3,18</sup>. Therefore, transcriptome-wide reference datasets associated with publicly available RNA reference materials are urgently needed but are lacking<sup>18</sup>.

As a part of the Quartet Project for the quality control and data integration of multi-omics profiling (<http://chinese-quartet.org/>), we established four RNA reference materials derived from immortalized B-lymphoblastoid cell lines (LCLs) from the four members of a monozygotic twin family quartet, which exhibited subtle inter-sample differences, high stability, long-term availability and easy manufacturability. Furthermore, matched multi-omics reference materials, including DNAs<sup>31</sup>, proteins<sup>32</sup> and metabolites<sup>33</sup>, were established along with RNAs from the same culturing of the LCLs to enable integrative omics analyses. In this study, we performed a multi-laboratory RNA-seq study based on 21 batches of multi-laboratory RNA-seq datasets generated with different protocols, established ratio-based reference datasets of gene expression and developed quality metrics for assessing reliability of RNA-seq technology in terms of intra-batch proficiency and cross-batch reproducibility.

## Results

### Overview of study design

The Quartet RNA reference materials were derived from the Epstein–Barr virus (EBV) LCLs from four members of a Chinese family quartet, including monozygotic twin daughters (D5 and D6), father (F7) and mother (M8) (Fig. 1a). They have been certified by China's State Administration for Market Regulation as the First Class

of National Reference Materials and are extensively being used for proficiency testing and method validation. The certified reference material numbers are GBW09904 (D5), GBW09905 (D6), GBW09906 (F7) and GBW09907 (M8).

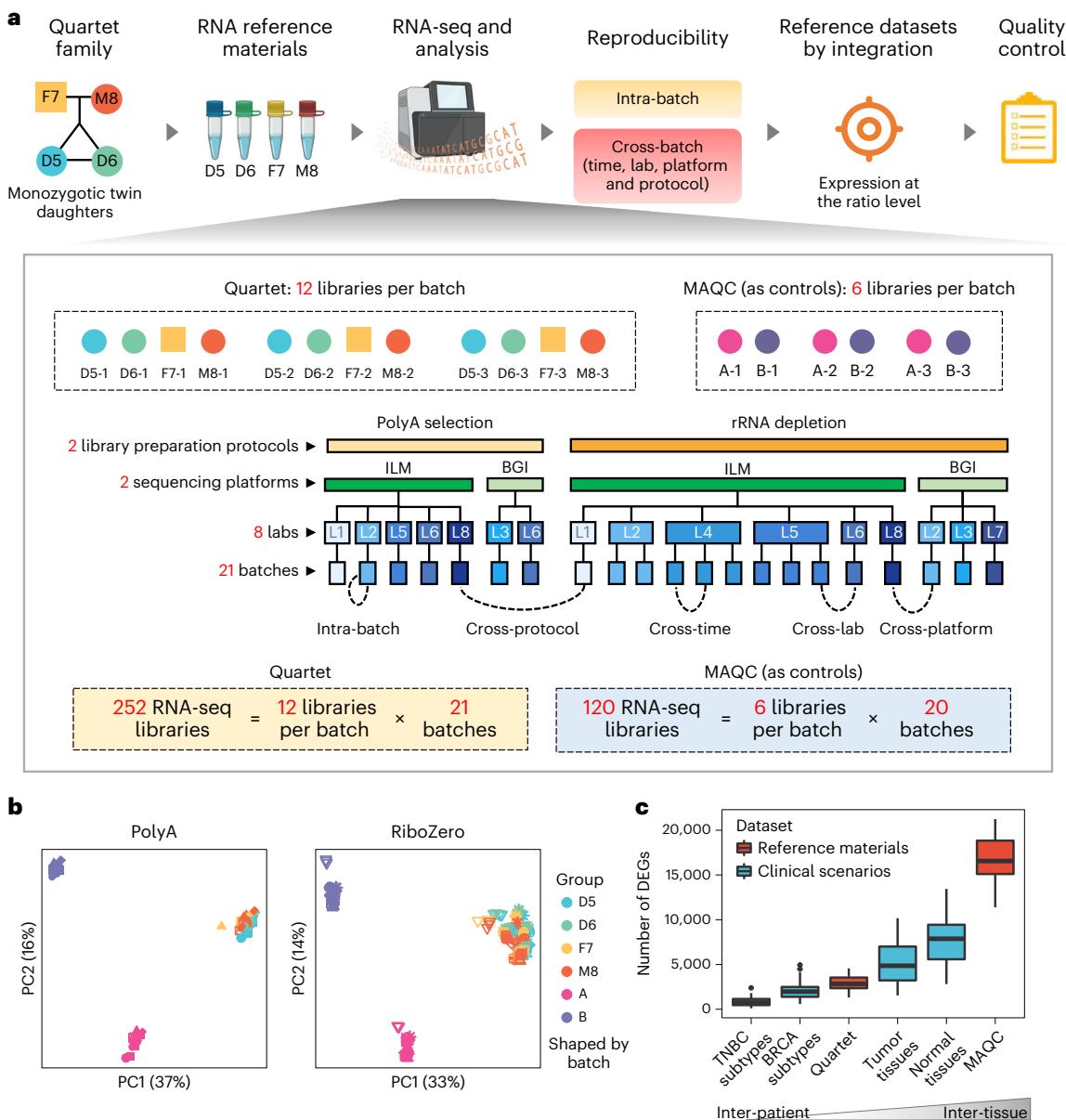
Large quantities of RNA (over 5 mg) were obtained per cell line, enabling standard RNA-seq experiments over 10,000 to 50,000 times and providing a material basis for long-term quality monitoring. RNA quality was high according to RNA integrity number (RIN) and RNA purity (Supplementary Fig. 1 and Supplementary Table 1). Moreover, the RNA reference materials showed adequate stability across 20 months of storage at  $-80^{\circ}\text{C}$  or 14 d of storage at room temperature ( $25^{\circ}\text{C}$ ) or  $4^{\circ}\text{C}$  or up to 20 times of bottle-opening and freeze–thaw cycle (Supplementary Fig. 2).

RNA-seq datasets from the Quartet RNA reference materials were then collected, consisting of 252 RNA-seq libraries from 21 batches generated in eight laboratories using two library construction protocols (PolyA selection and RiboZero) and two sequencing platforms (Illumina NovaSeq (ILM) and MGIDNBSEQ-T7 (BGI)) (Fig. 1a and Supplementary Table 2). Here, a batch is defined as 12 libraries from a standard sample set, consisting of 12 vials with each representing one of the triplicates of the Quartet RNA reference sample groups, whose library construction and sequencing experiments were conducted simultaneously. On the other hand, libraries constructed at different timepoints, in different laboratories, with different sequencing platforms, or using different library preparation protocols, are recognized broadly as cross-batch libraries (Fig. 1a). This comprehensive study design allows for objective performance assessment at multiple levels, including cross-time, cross-laboratory, cross-platform and cross-protocol. Moreover, RNA-seq experiments with the MAQC RNA reference materials (A and B) were conducted simultaneously with the Quartet reference materials in 20 of the 21 batches (Fig. 1a), enabling head-to-head comparisons between the two sources (MAQC versus Quartet) of RNA reference materials. In addition, the bioinformatic analysis pipeline was validated using published data from the MAQC RNA reference materials by comparison with previous studies<sup>15,20</sup> (Supplementary Fig. 3).

### The Quartet exhibits small intrinsic biological differences

Using principal component analysis (PCA) as an exploratory overview of data analysis, we found that multi-batch libraries of the Quartet reference materials from the same protocol (PolyA or RiboZero) were clustered together, whereas libraries of MAQC A and B samples were clustered separately into distinct groups according to protocol and sample groups (Fig. 1b and Supplementary Fig. 4). This result indicates that the intrinsic biological differences among the four groups of Quartet RNA reference materials are much smaller compared to those between the two MAQC RNA reference materials.

To investigate whether the magnitude of intrinsic biological differences or signals between the Quartet reference materials is representative of those seen in clinically relevant scenarios, we compared the extent of intrinsic biological differences between reference materials (MAQC A versus B and Quartet members) and those of four biological classification problems from published datasets ranging from four subtypes of triple-negative breast cancers (TNBCs)<sup>34</sup>, four subtypes of breast cancers<sup>35</sup>, four types of tumor tissues<sup>35</sup> and four types of normal tissues<sup>36</sup>. The number of differentially expressed genes (DEGs), previously used as a measure of 'treatment effect size'<sup>13</sup>, identified from the four biological classification problems ranged from 884 to 4,980 (mean), corresponding to an increase of intrinsic biological differences and/or decrease of within-group heterogeneity (Fig. 1c). Notably, the differences among Quartet RNA reference materials were 2,164 (mean) in terms of DEGs, which were ranked in the middle of these four clinical classification scenarios. In contrast, the differences between the two MAQC RNA reference materials were much larger (16,503, mean) than those observed in the aforementioned biological classification



**Fig. 1 | Overview of study design.** **a**, Quartet RNA reference materials were derived from immortalized EBV-infected B-LCLs from a quartet family, including monozygotic twin daughters (D5 and D6) and their father (F7) and mother (M8). Multi-batches of RNA-seq datasets were generated from independent laboratories using different library preparation protocols and sequencing platforms. Intra-batch proficiency and cross-batch reproducibility were then estimated. Based on multi-batches of RNA-seq data, we constructed ratio-based transcriptome-wide reference datasets and developed corresponding quality metrics. **b**, Scatter plots of PCs on RNA-seq data of the Quartet and MAQC RNA reference materials (marked in colors) across 20 batches (marked in shapes; see Supplementary Fig. 4 for details). log<sub>2</sub>-transformed FPKM values were used for PCA. **c**, Box plots showing the numbers of DEGs among Quartet reference materials, MAQC reference materials and four clinical/biological classification problems from published datasets. The four clinical/biological classifications used to represent clinical scenarios include four subtypes of TNBCs with different therapeutic actions (basal-like and immune-suppressed, luminal androgen

receptor, immunomodulatory subtype and mesenchymal-like subtypes)<sup>34</sup>, four subtypes of breast cancers (BRCA) with different prognosis and therapeutic actions (luminal A, luminal B, basal-like/triple negative and HER2-positive subtypes)<sup>35</sup>, four types of tumor tissues (brain, breast, kidney and lung cancers)<sup>36</sup> and four types of normal tissues (brain, breast, kidney and lung)<sup>36</sup>. The latter two types of biological classification problems are important for understanding the genetic basis of human diseases. Three samples from each clinical subtype or biological group were randomly selected for differential expression analysis to eliminate effect of number of samples used for analysis. A gene was identified as differentially expressed when satisfying the criteria of Student's *t*-test two-sided  $P < 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$  between two groups or conditions. To eliminate selection biases, this process was repeated 20 times ( $n = 20$ ). The box plots display the distribution of data with the median represented by the line inside the box and the interquartile range represented by the box. The whiskers extend from the box to the minimum and maximum values that are not outliers.

problems (Fig. 1c). These data again illustrate that the intrinsic biological differences among the Quartet reference materials are much smaller than those between MAQC RNA reference materials A and B and that such small differences are similar to those seen in clinical and biological classification scenarios.

**Signal-to-noise ratio enables assessment of data quality**  
Based on the Quartet design, a signal-to-noise ratio (SNR) metric was established to gauge the performance of a platform, a laboratory, a protocol or a batch in distinguishing the intrinsic biological differences ('signal') among the Quartet samples from variations among technical

replicates of the same sample group ('noise') (Fig. 2a). Generally, a lower SNR value indicates lower discriminating power and vice versa. An SNR value around or below zero means that the magnitude of signal is at a similar level as the noise or even lower than the noise. In this case, it is impossible to distinguish different sample groups under the high level of technical noises (Fig. 2b).

We evaluated the performance of five different methods in defining the SNR depending on whether the calculation is based on the original feature space or the dimensionality-reduced space (see Methods for details), including OriAll\_EucDist, OriSingle\_MedianEucDist, OriAll\_1-Cor, ReducedDim\_TSNE and ReducedDim\_PCA. It was found that the PCA-based SNR outperformed the other four methods in terms of its sensitivity in differentiating the quality of different datasets, as seen by a larger variability and a higher value of the SNR (Supplementary Fig. 5a). We next computed SNR using different numbers of principal components (PCs) of PCA (Supplementary Fig. 5b). SNR values based on the first component, the first two components or the first three components were highly correlated (Supplementary Fig. 5b). With the desire to maximize the range (that is, variability) of SNR over experiment batches and to match the good visual presentation of batch quality control, we chose SNR computed with the first two PCs.

SNR enables assessment of quality across the 21 batches of RNA-seq data. For most batches, the three replicates from the same sample group can be clearly distinguished from those of other sample groups (Supplementary Fig. 6). Large fluctuations of SNR values were observed across batches generated with the same protocol, the same sequencing platform or even from the same laboratory, highlighting the need for objectively assessing and monitoring the technical competency in data generation (Fig. 2c). Using an SNR cutoff of 12 (mean – s.d. across 21 batches), batches were flagged as high and low quality (Fig. 2d). It should be noted that SNR values based on different bioinformatics pipelines might differ, whereas the trend of the SNR across batches remained similar (Supplementary Fig. 7).

SNR can also be applied to diagnose potential causes of quality issues. In addition to the SNR values considering all 12 libraries in a batch, we also calculated SNR<sub>11</sub> values with any 11 of the 12 libraries in each batch (Fig. 2c). In five batches, the SNR<sub>11</sub> values increased by more than 6 dB compared to the corresponding 12-sample SNR values, indicating that the lower SNR values from these five batches might be a result of a 'random failure' of a particular technical replicate (for example, replicate M8\_1 from batch L5\_B1 and replicate F7\_2 from batch L2\_B1). In contrast, the three batches with the lowest SNR values were possibly due to systematic technical issues, because excluding any specific replicate (or potential outlier) could not greatly improve the SNR values.

Moreover, SNR enables assessment of data quality not only at gene expression level but also at alternative splicing (AS) level. Similarly, SNR values at AS level varied across batches. SNR values could be as high as 32.3, so that the three technical replicates for each sample type on the PCA plot could be loosely regarded as one dot (Supplementary Fig. 8a) or as low as 2.4 where technical replicates of one sample type were mixed with libraries from other sample types (Supplementary Fig. 8b).

Using multiple metrics, including SNR and other widely used quality metrics, with fastq, bam and expression profiles, with SNR showing the greatest differentiating power, 13 batches were flagged as high quality and were used for subsequent data integration to create the reference datasets, whereas the other eight batches were flagged as low quality and excluded from constructing the reference datasets (Fig. 2d and Supplementary Table 3).

### Ratio-based reference datasets

We next constructed transcriptome-wide reference datasets based on multi-batch and high-quality RNA-seq datasets, providing 'ground truth' for benchmarking. Ratio-based expression profiles, defined as a

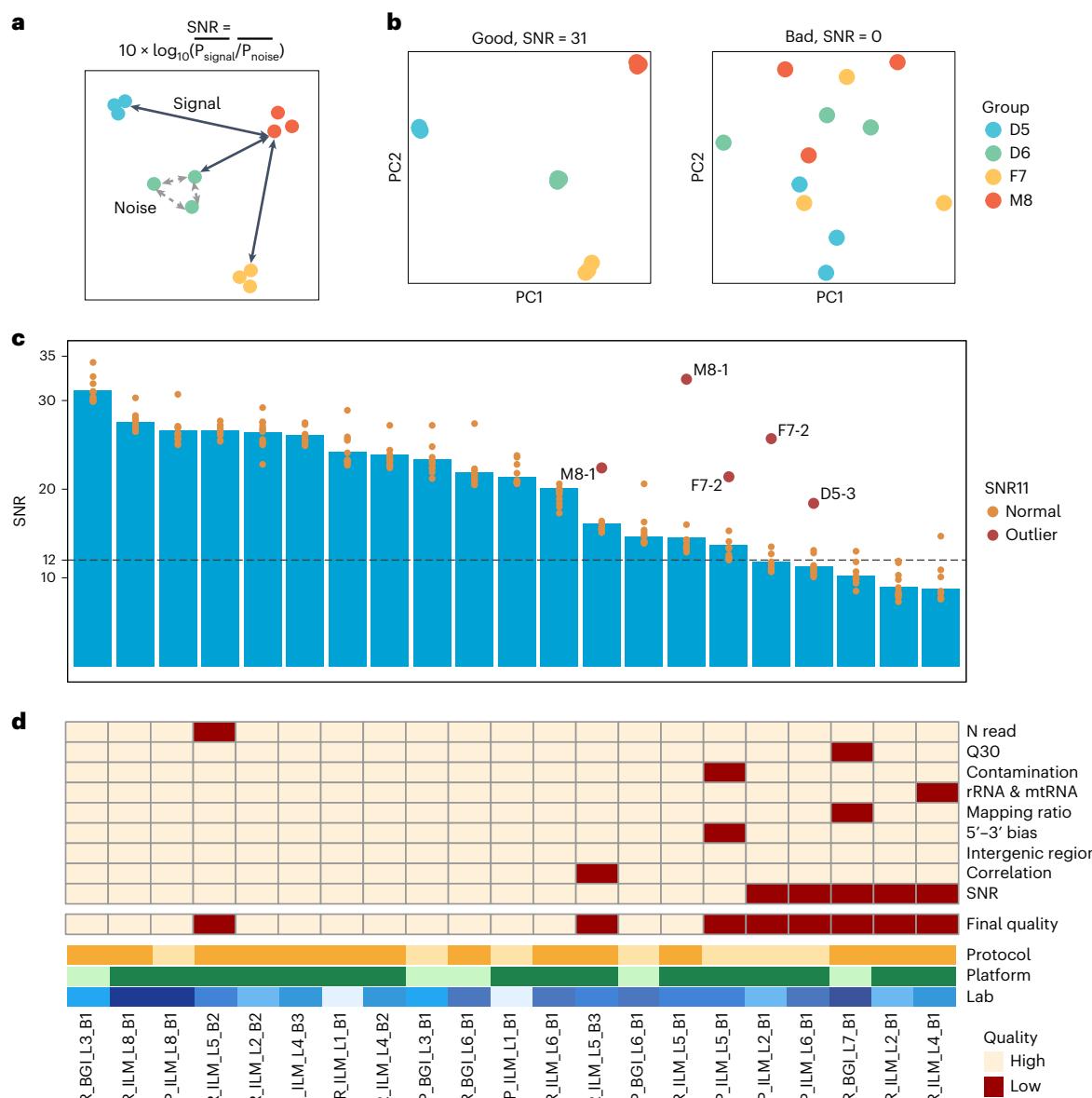
ratio or a fold change of expression levels between two sample groups for the same gene, agreed well across multiple transcriptomic technologies, including RNA-seq, microarray and quantitative polymerase chain reaction (qPCR)<sup>15,20</sup>. On the other hand, the incomparability of conventional 'absolute' expression profiles across different batches prevented meaningful cross-batch data integration<sup>15,20</sup>. Hence, we constructed the ratio-based transcriptome-wide reference datasets (Fig. 3a).

First, the detectable genes in each sample group (D5, D6, F7 or M8) were identified by consensus separately. In brief, if a gene was detected in all 13 batches in a sample group, it was considered expressed in that sample group. For the four Quartet reference materials (D5, D6, F7 and M8), 21,300, 22,161, 22,134 and 22,500 genes were expressed, respectively, representing 36.5–38.5% of the 58,395 genes annotated in GRCh38 (Fig. 3a). Moreover, around 32,104–33,937 genes (55–58%) were detected in more than four high-quality batches.

Second, ratio-based expressions (as  $\log_2$  transformed) were calculated for three pairs of sample groups using replicates of D6 as the common denominator (D5/D6, F7/D6 and M8/D6). To improve the reliability of the reference values, genes that were satisfied with thresholds of  $P < 0.05$  in each sample pair were used. Furthermore, genes that were significantly different ( $P < 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$ ) between PolyA and RiboZero protocols were removed to minimize technical variations introduced by the differences between the two distinct library preparation protocols. After these filtrations, the number of retained genes was 10,976, 9,451 and 10,728 for the three sample pairs (D5/D6, F7/D6 and M8/D6), respectively (Fig. 3a). Ratio-based reference datasets were then characterized between each pair of samples for a gene and were provided in the format of a geometric mean by summarizing from the 13 ratios calculated from each of the 13 high-quality RNA-seq datasets (Supplementary Tables 4 and 5).

Third, the homogeneity and stability of the Quartet RNA reference materials were assessed (Fig. 3a and Supplementary Fig. 9). Homogeneity and stability are two crucial characteristics of reference materials<sup>37</sup>. Homogeneity assessment aims to ensure that the previously characterized properties of reference materials are uniformly distributed across packaging units of the reference materials. Because the Quartet RNA reference materials were characterized using gene expressions, homogeneity assessment was conducted based on gene expression data. Here, we evaluated the homogeneity of reference materials by calculating within-unit ( $n = 9$ ) versus between-unit ( $n = 16$ ) variances of each gene using the analysis of variance (ANOVA) method (Supplementary Fig. 9a,b). Most (94.2–96.3%) genes performed well in homogeneity assessment (Supplementary Table 6). On the other hand, stability assessment aims to ensure that the value of the properties of the reference materials previously characterized remains unchanged over time. Here, we evaluated the stability of the Quartet RNA reference materials by calculating the slope of the regression of each gene based on the 15 batches of RNA-seq datasets that were generated from 10 timepoints over 26 months (Supplementary Fig. 9c,d). Most genes (91.9–95.1%) performed well in long-term stability assessment (Supplementary Table 6). Therefore, the Quartet RNA reference materials stored at  $-80^{\circ}\text{C}$  were homogenous and stable, as can be seen from the corresponding reference datasets.

Fourth, uncertainties of the reference materials were estimated. It is essential for identifying each source of uncertainties and to quantify the uncertainty introduced by each source. According to ISO Guide 35 (2017)<sup>37</sup>, ISO/IEC Guide 93-3 (2008)<sup>38</sup> and SAC JJF-1343 (2012)<sup>39</sup>, the source of uncertainties can be classified into characterization uncertainties ( $u_{\text{char}}$ ), sample inhomogeneities (between-bottle variation,  $u_{\text{bb}}$ ) and instabilities ( $u_{\text{i}}$ ). These values were then aggregated to form the combined uncertainties ( $u_c$ ) and expanded uncertainties ( $U$ ) with an expansion factor ( $k = 2$ , 95% confidence level) (Fig. 3a and Supplementary Table 6). As a result, most genes (83.1–88.1%) showed limited expanded uncertainties of less than 30%, demonstrating that the characterization of reference datasets was valid.



**Fig. 2 | SNR enables assessment and diagnosis of data quality.** **a**, Concept of calculating SNR. SNR was established to characterize the ability of a platform, a laboratory or a batch to distinguish the intrinsic differences among distinct biological sample groups ('signal') from variations in technical replicates of the same sample group ('noise'). **b**, Examples of good and bad batches with their SNR values and corresponding PCA scatter plots. **c**, SNR values across 21 RNA-seq batches to measure data quality. Batches were ordered by SNR values. Dots represent SNR values based on any 11 of the 12 libraries (SNR11) in each batch. A dot in dark red represents SNR11 value that increased over 6 dB compared to its standard SNR (12-sample SNR), when one library in this batch was excluded (the library ID was labeled), whereas a dot in orange represents SNR11 value

that decreased or increased less than 6 dB compared to its standard SNR. **d**, Quality flags of RNA-seq batches in terms of the number of sequencing reads (N read), percentage of Q30 (Q30), percentage of reads that were mapped to contamination species (for example, virus, bacteria and fungi) (Contamination), percentage of reads that were mapped to rRNA or mtRNA (rRNA & mtRNA), percentage of reads that were mapped to the human genome (Mapping ratio), gene body (5'-3') bias (5'-3' bias), percentage of mapped reads that were located in intergenic region in human genome (Intergenic region), Pearson correlation coefficient of technical replicates (Correlation), SNR and Final quality flag. Batches were ordered by SNR values. Protocol, Platform and Lab information of each batch is shown by the color legend.

Finally, high-confidence DEGs in the reference datasets (reference DEGs) were identified. A gene was considered as a reference DEG between two sample groups if it was concordantly discovered as an upregulated or downregulated gene ( $P < 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$ ) in more than four of the 13 high-quality batches. The number of reference DEGs was 1,863, 1,418 and 1,755 for the D5/D6, F7/D6 and M8/D6 sample pairs, respectively (Fig. 3a and Supplementary Table 7).

To verify the reliability of the reference datasets, we performed qPCR with reverse transcription (RT-qPCR) as an orthogonal validation. We selected 82 genes from the Quartet RNA reference datasets and

conducted RT-qPCR experiments on the four RNA reference materials (Supplementary Table 8). There is a high level of concordance between the Quartet reference datasets and the RT-qPCR data in terms of DEGs (92%, 91 of 99 DEGs across three sample pairs). We also compared the fold change of RT-qPCR versus that of reference datasets for the DEGs that were detected by both technologies ( $n = 91$ ) (Supplementary Table 9). We observed an expected high level of concordance to RT-qPCR ( $R = 0.85$ ), similar to what was previously reported between microarray and RT-qPCR ( $R = 0.80\text{--}1$ )<sup>13</sup> (Fig. 3b). DEGs that were identified in the reference datasets and RT-qPCR were further validated using droplet

digital PCR (ddPCR). Similar results were observed when comparing the fold changes between ddPCR and reference datasets in the aforementioned DEGs (Fig. 3c and Supplementary Table 9). Note that the level of the correlation coefficients depends on the level of the intrinsic biological differences between sample pairs under comparison. The differences among the Quartet RNA reference materials were relatively small compared to those of the MAQC samples A and B, resulting in relatively lower concordance<sup>20</sup> between the reference datasets and the RT-qPCR or ddPCR data for the Quartet reference materials.

Moreover, we used a liquid chromatography with tandem mass spectrometry (LC-MS/MS)-based proteomics dataset (batch code: NVG\_QEHFX)<sup>32,40</sup> for cross-omics validation of the RNA reference datasets (Supplementary Table 10). When all detected genes and proteins were considered, the correlation between RNA-seq and proteomics was modest ( $R = 0.45\text{--}0.57$ ) (Supplementary Fig. 10), which was similar to what was reported in previous studies ( $R = 0.36\text{--}0.60$ )<sup>41,42</sup>. However, we found that, for DEGs, there was a much higher concordance between RNA and protein data. When using DEGs that were detected by both RNA and protein measurements, the correlation increased to 0.94–0.96 (Fig. 3d and Supplementary Fig. 10). Thus, the protein-coding genes in the reference datasets that were differently expressed in the three sample pairs were successfully validated by the corresponding differential protein abundances. In addition, our findings indicated that the RNA reference datasets might also help benchmark proteomics technologies.

### Reference-dependent quality metrics

To benchmark RNA-seq data based on the aforementioned reference datasets, we developed three reference-dependent quality metrics. Specifically, we introduced the ‘relative correlation’ (RC) metric (that is, the Pearson correlation coefficient between the ratios of a test dataset for a given pair of samples and the corresponding ratio-based reference datasets, representing the trend of numerical consistency of the ratio-based expression profiles). We then introduced the ‘RMSE’ metric (that is, root mean square error (RMSE) of differences of ratios between a test dataset for a given pair of samples and the corresponding ratio-based reference datasets, representing the magnitude of average distances of ratio-based expression profiles). Moreover, we introduced the ‘MCC of DEGs’ (MCC) metric (that is, Matthews correlation coefficient (MCC) to measure the consistency of DEGs detected from a test dataset for a given pair of samples with those from the high-confidence DEGs in the reference datasets). Based on their definitions, higher values of RC and MCC of DEGs indicate a better fit between the test dataset and the reference dataset, whereas lower values of RMSE indicate a better fit. All three metrics were able to clearly demonstrate differences in data quality among the 21 batches of data, including the 13 high-quality and eight low-quality batches of data (Supplementary Fig. 11a).

One might argue that the lower RC, higher RMSE or lower MCC values of the eight pre-defined low-quality batches might have resulted from their exclusion during the construction of the reference datasets. To determine whether it was the case or not, we performed a 30-times cross-validation test. In brief, in each round, we randomly selected 13 batches from the 21 batches to ‘train’ the reference datasets. Reference-dependent quality metrics were then calculated, and

the remaining eight batches were used as a ‘validation’ set. The results showed that the ‘train’ and ‘validation’ metrics were highly correlated ( $R = 1$ ) (Supplementary Fig. 11b), demonstrating that the quality metrics were not dependent on whether the batches were included in the construction of the reference datasets or not. Instead, the three metrics objectively reflected the intrinsic quality of a dataset, indicating that they were suitable for performance evaluation of future datasets. The cutoff values of RC, RMSE and MCC values were set to 0.89, 0.38 and 0.54, respectively, which were expressed as the (mean – s.d.) of RC and MCC and the (mean + s.d.) of RMSE across validation sets in the 30-times cross-validation analysis (Supplementary Fig. 11b and Supplementary Table 3).

Furthermore, we compared characteristics between the two categories of quality metrics, including reference-independent quality metric (SNR) and reference-dependent quality metrics (RC, RMSE and MCC). In most cases, high-quality batches showed higher values of SNR, RC and MCC and lower values of RMSE, and vice versa, except for one batch (L5\_B3) (Supplementary Fig. 11a). In this batch, a high SNR value (16.1) with low reference-dependent quality metrics (RC = 0.784, RMSE = 0.735 and MCC = 0.480) was observed. In fact, a customized library preparation kit designed for removing several highly expressed RNAs (for example, *RN7S* genes) was used in this batch (L5\_B3), leading to overall differences between expression profiles from this batch and the reference datasets. Moreover, the complementarity between reference-independent and reference-dependent quality metrics was observed, indicating that both categories of quality metrics should be included in comprehensive performance assessment.

Finally, we calculated a total quality score by summarizing the two categories of quality metrics. Considering the high correlation among the three reference-dependent metrics (RC, MCC and RMSE) (absolute  $R \geq 0.92$ ) (Supplementary Fig. 11a), we used RC to represent the reference-dependent metric score for calculating the total quality score. The total quality score was expressed as the geometrical mean of SNR and RC for measuring the overall quality of a dataset for the intra-batch proficiency.

### Ratio-based expressions improve cross-batch reproducibility

In large-scale projects, expression profiles are usually measured across multiple batches and pooled together for downstream analysis. Cross-batch reproducibility is, therefore, crucial. Multi-batch RNA-seq datasets derived from the Quartet RNA reference materials allowed us for objective performance assessment of cross-batch reproducibility at multiple levels, including cross-time, cross-laboratory, cross-platform and cross-protocol.

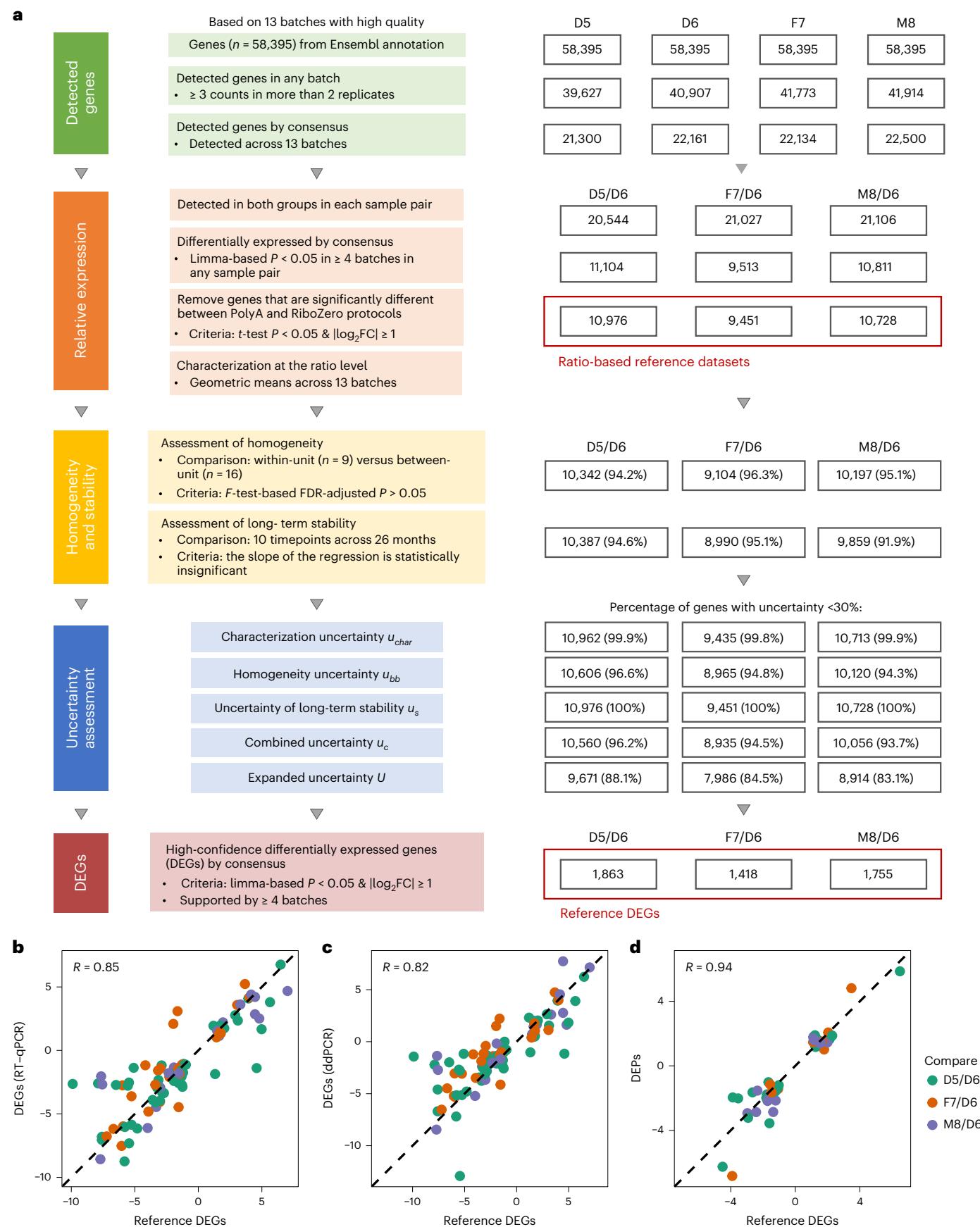
In this study, after pooling batches of data from the PolyA and/or RiboZero protocol(s) together without batch corrections, the impact of batch effects on obscuring the differentiation of biologically distinct groups could be clearly seen in a PCA plot with a diminished SNR value of below 5 (0–4.6) (Fig. 4a). Non-experimental factors, rather than intrinsic biological groups (D5, D6, F7 and M8), exhibited the largest differences. When PCA was based solely on the MAQC samples without the Quartet samples, batch effects could not be observed from PC1 due to the overwhelming biological differences between the two MAQC samples. However, indications of batch effects were appreciable from PC2 (Supplementary Fig. 12). When ratio-based expressions were

**Fig. 3 | Construction and validation of ratio-based transcriptome-wide reference datasets. a**, Workflow for constructing Quartet RNA reference datasets. Reference datasets were constructed according to the following steps: (1) identifying detectable genes; (2) calculating ratio-based expression based on reliably detectable genes that were differentially expressed; (3) assessing the homogeneity and stability of RNA reference materials; (4) assessing the uncertainty of ratio-based reference datasets; and (5) identifying high-confidence DEGs. **b–d**, Scatter plots of log<sub>2</sub> fold changes (FCs) of gene expression between reference DEGs and RT-qPCR (b), ddPCR (c) and proteomics data (d).

Pearson correlation coefficient across three sample pairs was calculated. Genes/proteins that were considered as differentially expressed in both methods shown in x axis and y axis were used for plotting. x axis: average log<sub>2</sub>FC from 13 high-quality RNA-seq batches reference DEGs. y axis: for the RT-qPCR and proteomics data, a gene or a protein was considered as a DEG/DEP when the t-test two-sided  $P < 0.05$  and FC  $\geq 2$  or  $\leq -0.5$ ; for ddPCR data, genes that were identified as DEGs based on RT-qPCR were used. Average log<sub>2</sub>FC of RT-qPCR ( $n = 3$ ), ddPCR ( $n = 2$ ) and proteomics ( $n = 3$ ) data from DEGs/DEPs were used for plotting. DEP, differentially expressed protein.

used, which referred to converting expression profiles to gene-wise relative scale within each batch using D6 as the denominator, the SNR value increased to around 20 (18.3–22.3). Meanwhile, all libraries from

the PolyA and RiboZero protocols of the same sample group were grouped together based on ratio-based expressions (Fig. 4b). Similar results were observed when an alternative gene quantification tool



(for example, RSEM) or a normalization method (for example, normalized counts) was used to quantify and compare relative expressions (Supplementary Figs. 13 and 14). These findings indicate the critical importance of detecting and correcting batch effects in multi-batch studies. Notably, ratio-based expressions were effective in mitigating such batch effects.

We then compared pairwise cross-batch performance to investigate integrability at different levels. When different batches of libraries are compared against each other, they could be classified into five different scenarios with increasing degree of differences, including intra-batch, cross-time, cross-laboratory, cross-platform of sequencing and cross-protocol levels. We compared the consistency between datasets from different levels of comparison using three quality metrics: SNR, RC and Pearson correlation coefficients.

SNR values were calculated for the five scenarios of comparisons. Compared to intra-batch SNR values (median SNR = 20.7), SNR values dropped to  $-0.4$ – $1.2$  (median SNR) at cross-time, cross-laboratory, cross-platform or cross-protocol comparisons when absolute expressions ( $\log_2\text{FPKM}$ ) of the two datasets were merged to calculate the SNR value. In this case, it is essentially impossible to distinguish different sample types under the influence of ‘batch effects’. Thus, expression profiles from two batches of libraries could not be integrated directly at the absolute expression level. However, when ratio-based expressions were used, SNR values maintained as high as  $12.3$ – $14.8$  (median SNR) (Fig. 4c). This finding again reinforced the previous notion that ratio-based expressions are much more resistant to batch effects (Fig. 4b).

Similar results were obtained for performance based on RC values. Compared to intra-batch RC (median RC: 0.946), RC values dropped to  $0.543$ – $0.772$  (median RC) when absolute expressions of two datasets were compared. However, they maintained at  $0.933$ – $0.949$  (median RC) when ratio-based expressions of the two datasets were considered (Fig. 4d).

Additionally, the median correlation of absolute expressions was as high as 0.965 for intra-batch technical replicates and 0.938 between different groups in the same batch. It dropped to  $0.814$ – $0.927$  for cross-batch technical replicates. What is worse, correlations of technical replicates for the same sample from difference batches were significantly lower than correlations between different sample groups from the same batch ( $P < 0.001$ ), highlighting the critical impact of batch effects (Fig. 4e). On the contrary, correlations of ratio-based expressions of technical replicates ( $0.319$ – $0.401$ ) were consistently higher than those of different groups ( $0.072$ – $0.093$ ) under the different levels of cross-batch comparisons (Fig. 4f), demonstrating the differentiating power at the ratio-based expression level.

Our findings support the important roles of reference materials in assessing cross-batch reproducibility and their effectiveness in removing batch effects. It should be noted that we could clearly observe/monitor batch effects based on multi-batch datasets of

Quartet RNA reference materials (Fig. 4a,b), whereas it is impossible with the MAQC reference materials due to their substantial differences (Fig. 1b). Thus, the Quartet reference materials can provide more precise assessment of measurement performance based on their small but biologically relevant intrinsic differences, highlighting their critical roles in assessing cross-batch reproducibility.

### Biological differences between the Quartet twins

It was noticed that the two LCLs corresponding to the two monozygotic twin daughters (D5 and D6) exhibited consistently large differences in gene expression in all batches of data (Supplementary Figs. 6 and 15), although one might have expected that the expression profiles from the two identical twins would show the highest similarity among all six pairs of the Quartet sample groups. Here, we used ratio-based expression profiles of the 13 high-quality batches and applied a weighted gene co-expression network analysis (WGCNA) approach<sup>43</sup> to discern the underlying biological forces behind the differences in transcriptome between the two cell lines. Genes were grouped with strong co-expression patterns across the sample set into eight modules (Fig. 5a). D5 samples were distinct from D6 samples in the PC1 space based on transcriptomic expression for most modules (seven of eight modules), including the largest module (turquoise module) with 2,368 highly co-expressed genes (Fig. 5a,b). Functional analysis showed that the turquoise module genes were enriched in Gene Ontology (GO) terms, such as cell cycle and B cell proliferation (Fig. 5b). Moreover, a 1,777-gene module (blue module), which showed dispersity between D5 and other three groups (D6, F7 and M8) in the PC1 space, was enriched in B-cell-mediated immunity. These results imply that differential processes of B cell subtype selection and effects of cell culture might have occurred among the Quartet RNA reference materials (Fig. 5b).

On the contrary, when we applied WGCNA analysis on the  $\log_2\text{FPKM}$  values (Supplementary Fig. 16), the two largest modules (turquoise and blue) were grouped according to protocols and/or batches. Only for the third (brown) and fourth (yellow) largest modules, the samples were grouped based on donors with genes enriched in B cell activation and immune responses. These results imply that biological signals of relationships among the Quartet cell lines were largely masked at the raw FPKM level, highlighting the negative impact of batch effect in absolute profile data.

To identify B cell subtypes corresponding to the Quartet RNA reference materials, we examined the expression levels of B cell surface membrane immunoglobulins (SmIg) on the Quartet cell lines. Four types of SmIg were measured, including IgD, IgM, IgG and IgA, which were biomarkers of the developmental stages of B cells. Notably, the IgA expression pattern of the immortalized cell lines from the two monozygotic twin daughters (D5 and D6) exhibited substantial differences in that IgA was highly expressed in D5 but almost undetectable in D6 (Fig. 5c). Additionally, the expression level of IgG was much higher

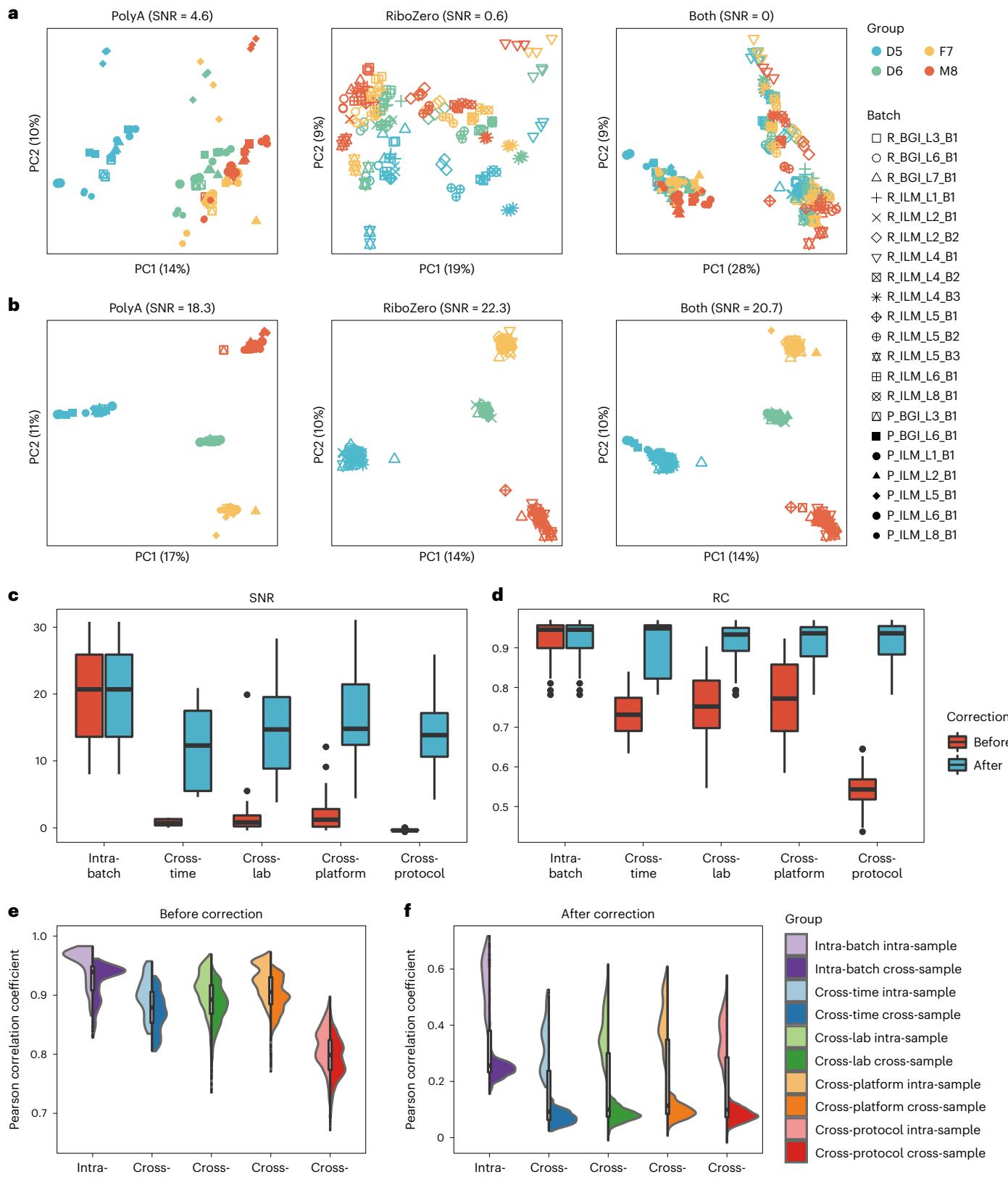
**Fig. 4 | Performance evaluation of cross-batch reproducibility.** **a,b**, Scatter plots of PCA on RNA-seq data before batch correction (a) and after correction (b) from replicates of the Quartet RNA reference materials (marked in colors) in the 21 batches (marked in shapes). Expressions in  $\log_2\text{FPKM}$  were used as before batch-correction datasets. Ratio-based expressions (which referred to converting expression profiles to gene-wise relative-scale profiles within each batch) were used to correct batch effects. Ratio-based expressions were obtained by subtracting  $\log_2\text{FPKM}$  by the mean of  $\log_2\text{FPKM}$  of the three replicates of D6 in the same batch. We used a multi-batch RNA-seq dataset, including 168 RNA-seq libraries from the RiboZero protocol and 84 RNA-seq libraries from the PolyA protocol. Plots were color-coded by sample groups and shaped by batches. **c,d**, Box plots of SNR values (c) and relative correlation with reference datasets (RC) values (d) for comparisons indicated at the x axis. When each batch of libraries was compared against each other, they could be classified into five different scenarios with increasing degree of differences, including intra-batch, cross-time, cross-laboratory, cross-platform of sequencing and cross-protocol

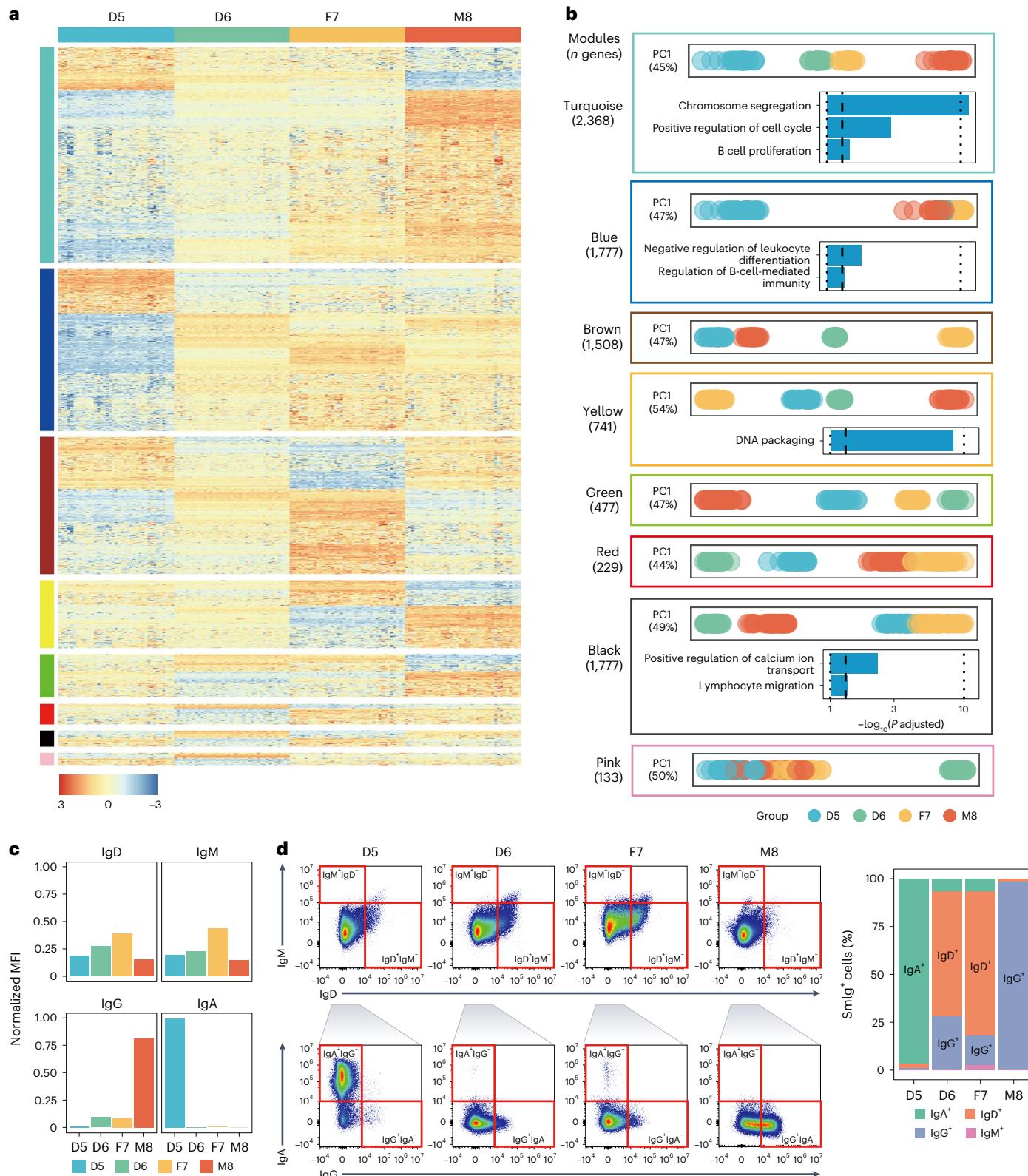
levels. Intra-batch SNR values were calculated using 12 samples in the same batch, whereas SNR values of cross-batch were calculated by combining expression data from all combinations of two batches ( $n = 24$ ). **e,f**, Violin plots of Pearson correlation coefficients based on expression profiles before (e) and after (f) batch correction for comparisons indicated at the x axis. D5, F7 and M8 samples were used to calculate pairwise correlations, whereas D6 samples were used as denominators for calculating ratio-based expressions for correcting batch effects. The number of combinations ( $n$ ) used to derive statistics in c–f in each box were as follows: **c**: intra-batch,  $n = 21$ ; cross-time,  $n = 7$ ; cross-laboratory,  $n = 62$ ; cross-platform,  $n = 43$ ; cross-protocol,  $n = 98$ ; **d**: intra-batch,  $n = 63$ ; cross-time,  $n = 21$ ; cross-laboratory,  $n = 186$ ; cross-platform,  $n = 129$ ; cross-protocol,  $n = 294$ ; **e** and **f**: intra-batch intra-sample,  $n = 189$ ; intra-batch cross-sample,  $n = 567$ ; cross-time intra-sample,  $n = 189$ ; cross-time cross-sample,  $n = 378$ ; cross-laboratory intra-sample,  $n = 1,674$ ; cross-laboratory cross-sample,  $n = 3,348$ ; cross-protocol intra-sample,  $n = 1,161$ ; cross-protocol cross-sample,  $n = 2,322$ ; cross-platform intra-sample,  $n = 2,646$ ; cross-platform cross-sample,  $n = 5,292$ .

in the M8 group compared to the other three groups (Fig. 5c). We further performed immunophenotypic analysis of the four immortalized cell lines. In agreement with the Smg findings from RNA-seq, the IgA<sup>+</sup> cells were mainly present in the cell line from D5, whereas a lower percentage of IgA<sup>+</sup> cells was found in other cell lines (Fig. 5d and

Supplementary Fig. 17). Furthermore, the percentage of IgG<sup>+</sup> cells was higher in M8 compared to the other three groups (Fig. 5d).

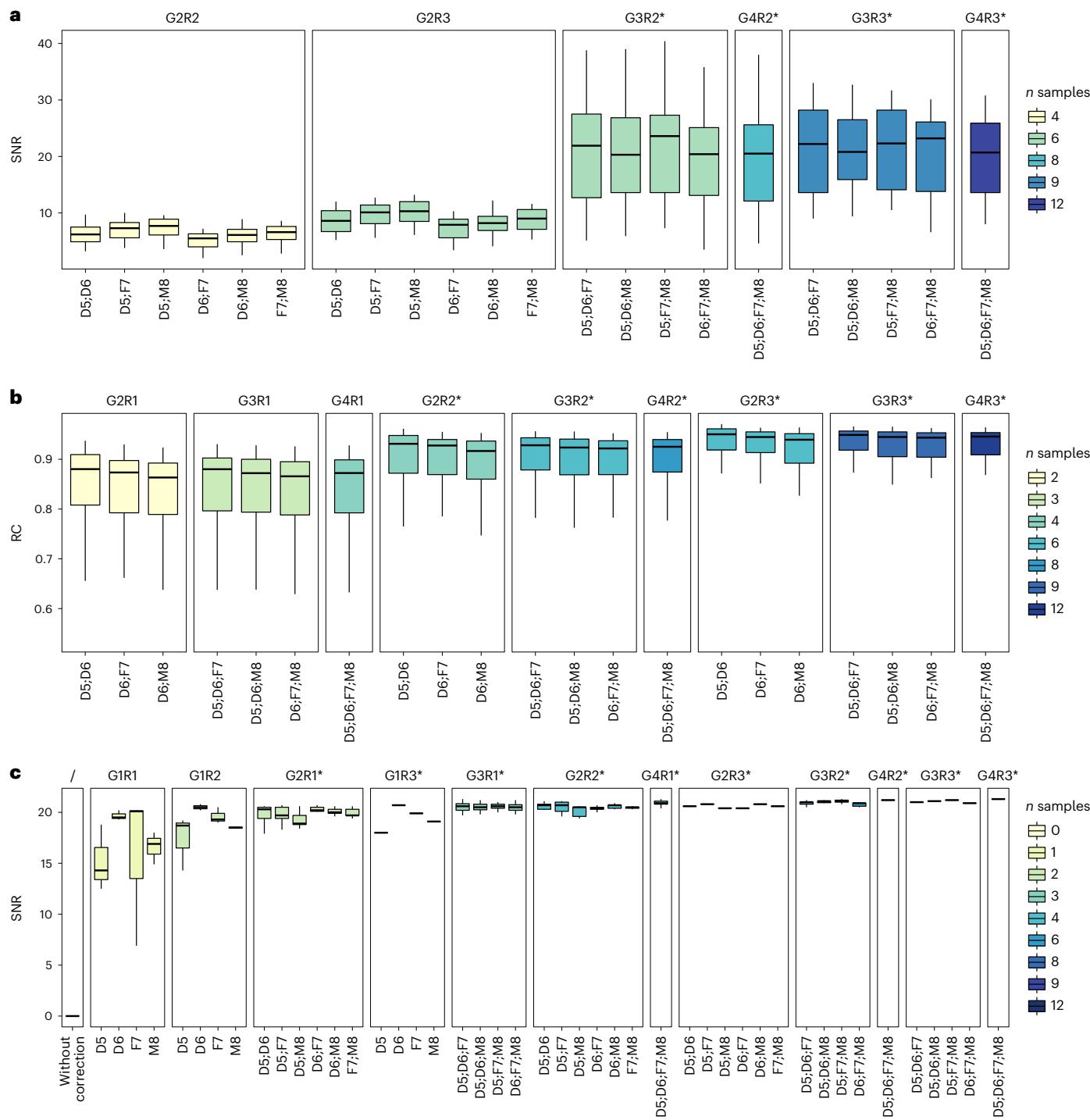
We hypothesized that the major factors driving transcriptomic expression characteristics were probably related to the processes for immortalizing cell lines (for example, B cell subtype selection during





**Fig. 5 | Biological differences between immortalized B-LCLs of the Quartet monozygotic twins.** **a**, Expression profiles from co-expression modules using data from 13 batches with high quality. Color-coded module membership was displayed in the color bars to the left of the dendrograms. Ratio-based expressions were obtained by subtracting  $\log_2$ FPKM by the mean of  $\log_2$ FPKM of the three replicates of D6 in the same batch. The heat map was colored using z-scored ratio-based expression profiles. **b**, Distances of samples in PC1 space and list of GO terms enriched with genes in each corresponding module. Enriched GO

terms were generated based on hypergeometric test using clusterProfiler<sup>64</sup>, with a Benjamini–Hochberg correction and an adjusted  $P$  value cutoff of 0.05. PC plots were colored by sample groups. Bar plots were colored based on the number of genes included in GO terms. **c**, The normalized expression level (median fluorescence intensity, MFI) of B cell surface membrane Smlg IgD, IgM, IgG and IgA in immortalized B-LCLs. **d**, Left: representative flow cytometric dot plots show the IgD<sup>+</sup> cells, IgM<sup>+</sup> cells, IgG<sup>+</sup> cells and IgA<sup>+</sup> cells in immortalized B-LCLs. Right: expression levels of IgA, IgG, IgM or IgD in the four immortalized cell lines.



**Fig. 6 | Recommended group-replicate combinations for using the Quartet RNA reference materials for quality control.** **a,b**, Distribution of SNR (**a**) and relative correlations with reference datasets (RC) (**b**) under different group-replicate combinations of Quartet RNA reference materials used for assessing intra-batch proficiency. **c**, Distribution of SNR values for ratio-based expression using different numbers of samples and/or replicates as the denominator for the calculation of the ratio-based expressions. x axis represents the enumerated

number and groups of Quartet reference materials. Titles of subpanel represent the number of sample groups (G) and replicates (R) used for calculating SNR (**a**), RC (**b**) and denominators for applying the ratio-based method (**c**). For example, 'G2R2' represents four libraries comprising two sample groups (G) with two replicates (R) per group. The recommended combinations are marked with asterisks (\*).

EBV infection and cell culture)<sup>44</sup>. To validate this hypothesis, we further conducted RNA-seq experiments based on whole-blood samples of the four donors. Expression profiles of whole-blood samples from D6 and F7 donors looked different (Supplementary Fig. 18a) and were not grouped together as what we observed based on expression

profiles from the cell lines (Supplementary Fig. 15). On the other hand, the twin daughters grouped close to each other and showed the highest similarity in expression profiles among the Quartet samples in the PCA plot (Supplementary Fig. 18b). The intrinsic biological differences between the Quartet monozygotic twins enhanced our

understanding of the Quartet RNA reference materials and could be used as another layer of built-in truth to increase the quality control utilities of the Quartet RNA reference materials<sup>45</sup>.

### Recommended group–replicate combinations

An important question is what group–replicate combinations would constitute an appropriate choice for applying Quartet RNA reference materials for quality control in routine transcriptomic profiling. Thus, the replicate number and group number of Quartet RNA reference materials that could be used were enumerated. The results revealed that a minimum of three sample groups and two replicates per batch were required for reaching SNR with high sensitivity for distinguishing data quality of different batches. The use of only two sample groups was not enough for distinguishing quality difference of datasets (Fig. 6a). Meanwhile, it was revealed that a minimum of two replicates per sample type were required for obtaining RC with high consistency with the ground truth (Fig. 6b). Under the same number of replicates and groups, the impact of group combinations (D5, D6, F7 or M8) was minor.

Furthermore, the number and groups of reference materials that could be used as a reference (denominator) in ratio-based profiling within each batch were enumerated. SNR increased markedly at the ratio (relative) level compared to the absolute level even when only one single replicate was used, with a median SNR value greater than 14 (Fig. 6c), and the SNR values further increased when more replicates and/or more sample groups were added to calculate average expression values as the denominator. Moreover, the SNR values obtained using only one sample to calculate the denominator varied greatly, whereas SNR values based on the mean of more replicates and/or sample groups as the denominator were more stable. Given the same number of samples as the denominator, a higher number of sample groups helped further increase SNR.

These results provided a solid foundation to determine the optimal number of samples and/or replicates to be used for performance assessment and ratio-based transcriptomic profiling using the Quartet RNA reference materials. When RC with the reference datasets was used for intra-batch performance assessment (Fig. 6b) and when multiple-sample groups/replicates were used as the denominator for ratio-based cross-batch effect correction (Fig. 6c), the use of two sample groups appears sufficient in many settings. However, when reference-dataset-free SNR was used for intra-batch performance assessment (Fig. 6a), multiple groups of reference materials ( $\geq 3$ ) are required. SNR has been shown to be more sensitive in assessing and diagnosing data quality issues. Moreover, SNR is a reference-independent quality metric, enabling assessment beyond the boundaries of the reference datasets. Hence, for proficiency test purposes, multi-group references are needed to implement comprehensive quality assessment.

## Discussion

We generated well-characterized, high-quality, homogenous and stable Quartet RNA reference materials and constructed corresponding reference datasets from reliable transcriptomic data, which can be a useful tool for objectively assessing data quality and improving the reliability of transcriptomic profiling, specifically within a clinical setting. Notably, the Quartet RNA reference materials have been approved by China's State Administration for Market Regulation as the First Class of National Reference Materials and are extensively being used for proficiency testing and method validation.

The Quartet RNA reference materials exhibit several advantages. First, they are a part of multi-omics reference materials, with matched DNA, RNA, proteins and metabolites generated from the same immortalized cell lines. This study design allows for cross-omics validation and will help reliably understand the biological traits of the reference materials. Second, the suite of reference materials is from a four-member Quartet family including two monozygotic twin

daughters and their father and mother. Genomic and phenotypic characteristics are involved in the four RNA samples, acting as built-in ‘truth’. The reference datasets based on intrinsic biological differences among the Quartet RNA reference materials have been constructed and can be used as ‘ground truth’ for quality assessment (Fig. 3). Expression characteristics affected by genetic relationships of the four reference materials will be further studied<sup>45</sup>. Third, the RNA reference materials are derived from cell lines from four individuals. The small intrinsic biological differences among the Quartet RNA reference materials enable precise assessment at inter-patient level, rather than inter-tissue level, which are closer to clinical scenarios of subtle differences among study groups (Fig. 1c). Fourth, the RNA samples are produced in large amounts in one batch and are renewable through cell culture. By minimizing batch effects that may be introduced during cell culture and RNA extraction, the Quartet RNA reference materials are sufficient for performing standard RNA-seq experiments over 10,000 to 50,000 times and provide a material basis for long-term quality monitoring. Based on comprehensive assessments, the Quartet RNA reference materials are homogenous and long-term stable at the storage temperature of  $-80^{\circ}\text{C}$ . The publicly available Quartet RNA reference materials can also be used for further evaluation of emerging technologies as well as new areas of interest that are beyond gene expression levels, such as AS, RNA editing, gene fusion and epitranscriptomics. In addition, the Quartet RNA reference materials comprise high-quality total RNAs, including not only full-length RNAs but also small RNA molecules such as miRNAs, enabling further quality assessment of small RNA profiling technologies.

Quality metrics derived from the Quartet RNA reference materials and reference datasets can be used for proficiency testing and external quality assessment. Previous quality metrics were focused on biases from library preparation or on detecting outliers in expression profiles<sup>4,17,21,46–49</sup>. It has been previously demonstrated that ‘lab effects’ strongly affect the detection of DEGs, highlighting the importance of assessing data quality in detecting DEGs<sup>20</sup>. In this study, we developed comprehensive quality metrics for assessing the reliability of differential expression, including discriminating power across different biological groups (SNR) and reproducibility of identifying DEGs (RC, RMSE and MCC), which are reference independent and reference dependent, respectively. In addition, distributions of these quality control measures were obtained from multiple real-world RNA-seq datasets, providing practical cutoffs to decide whether the proficiency of a test dataset is acceptable.

The Quartet RNA reference materials can be used for monitoring and correcting batch effects. Batch effects are notorious technical variations irrelevant to study factors and are challenging to deal with, especially when they are confounded with biological factors of interest<sup>50,51</sup>. Our results demonstrate that the presence of batch effects without correction can lead to misclassification of samples (Fig. 4a), but we found that these batch effects can be mitigated by using ratio-based expressions (Figs. 4b,c,f and 6c), if one or more common reference materials are profiled across batches. Our companion work found that using ratio-based data analysis by scaling the absolute feature values of study samples relative to those of concurrently measured reference sample(s) on a feature-by-feature basis could effectively mitigate the widespread problems of batch effects in epigenomics, transcriptomics, proteomics and metabolomics datasets<sup>52,53</sup>. This ratio-based method is equally effective even for study design of completely unbalanced distributions of samples in different groups between different batches. In practice, the imbalance in samples across batches is almost inevitable because of hidden biological subpopulation variabilities<sup>50,51</sup>.

In addition, the advantages of ratio-based expression profiles remain in identifying DEGs, which were extensively explored in our companion work<sup>53</sup>. A straightforward method (that is, fold change ranking with a non-stringent *t*-test *P* value cutoff) can be applied to perform differential analysis in ratio-based profiles. The effectiveness of the

DEG method was previously shown in microarray<sup>15,54</sup> and RNA-seq<sup>18,20</sup> data in absolute expressions.

The Quartet RNA reference materials can act as valuable tools for quality control in large-scale, longitudinal and multi-center projects. Many large-scale consortium projects with comprehensive and coordinated efforts help accelerate understanding of the molecular basis of transcriptome by producing RNA-seq data with a large sample size<sup>36,55–57</sup>. However, the broad variety of platforms, protocols and laboratory proficiencies<sup>58–60</sup> has created the need for comprehensive reference materials. At the starting point of a large-scale project, we recommend that researchers conducting RNA-seq experiments using the Quartet RNA reference materials in each laboratory assess and ensure intra-batch proficiency and cross-batch reproducibility before analyzing precious study samples. Meanwhile, researchers can use the Quartet RNA reference samples routinely along with study samples to monitor and correct batch effects.

The combinations of sample groups and number of replicates for the application of the Quartet RNA reference materials are context dependent. For proficiency test and external quality assessment purposes, where the frequency of reference usage could be as low as a few times per year, it is recommended to apply multiple groups of samples with multiple replicates per group. Users can apply a minimum of three sample groups and two replicates for quality assessment (Fig. 6). Users can apply a total of 12 samples, comprising the four Quartet RNA reference materials with three replicates for each RNA sample group, to implement full quality assessment mentioned in the study and remove batch effect in a robust way (Fig. 6a,b). For batch effect removal purposes in large cohort studies, where the Quartet RNA reference materials are expected to be routinely used along with study samples, and where additional cost associated with profiling reference samples becomes an issue, it is recommended to apply fewer sample groups and fewer replicates per batch. Users can even apply four sample groups or as few as two sample groups without replicates as a cost-effective choice of references for monitoring and correcting batch effects (Fig. 6c). In this case, we suggest the use of a total of four profiles from each replicate of the four Quartet RNA reference materials as the denominator per batch of 96 libraries for ratio-based expression profiling, reaching a high SNR while maintaining a reasonable additional cost ( $4 / (96 - 4) = 4.3\%$ ) per batch of 92 study samples.

To facilitate the adoption of multi-omics reference materials, reference datasets and quality metrics from the Quartet Project, we developed a Quartet Data Portal (<http://chinese-quartet.org/>) for access to the Quartet resources and for enhancing the quality consciousness of the community<sup>40</sup>. Researchers can request the multi-omics reference materials, datasets and reference datasets from the data portal. Additionally, researchers can upload RNA-seq data of their own, automatically analyze and evaluate data quality and/or share data with the community. With the growing use of the Quartet reference materials, we hope to generate and collect diverse datasets and further upgrade the reference datasets.

Although many advantages of using the Quartet RNA reference materials were obvious and are listed above, several limitations of the Quartet samples should also be noted. First, only around 55–58% of the 58,395 annotated genes were reliably detected (detected in more than four high-quality batches) in the Quartet RNA reference materials, limiting quality assessment and ratio-based scaling to these detectable genes. This is not a serious issue when using Quartet RNA reference materials for proficiency testing and external quality assessment. However, this could become a limitation when the Quartet RNA reference materials are to be used for profiling along with study samples for reporting ratio-based profiling data. Ratio-based scaling may successfully mitigate batch effects from genes when they are expressed in both the study samples and reference material(s). If a gene is expressed only in study samples but not in reference materials, its expression may not be successfully corrected. In such a scenario, a fudge factor may be used

for making the ratio calculation possible. The limitations of ratio-based scaling are also extensively discussed in our accompanying papers<sup>52,53</sup>. Second, a single analysis pipeline was used in this study, which may introduce bias in transcriptomic quantification and characterization of the reference datasets. Although previous studies compared the performance of different RNA-seq analysis tools and found overall good reproducibility for different tool combinations in terms of differential expression calls after proper filtering processes<sup>61–63</sup>, bioinformatics tools will be further evaluated and used for characterizing the reference datasets. Third, the datasets were generated by high-throughput short-read sequencing technologies. It is likely that, with further benchmarking and widespread adoption of reference materials, additional reagents, protocols and instruments will be evaluated.

In summary, the Quartet RNA reference materials and reference datasets are unique resources to improve quality of RNA-seq data. Inclusion of the Quartet RNA reference materials in RNA-seq batches coupled with reference datasets will make RNA-seq more reproducible, accurate and comparable, especially within clinical settings.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-023-01867-9>.

## References

1. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
2. Marco-Puche, G., Lois, S., Benitez, J. & Trivino, J. C. RNA-Seq perspectives to improve clinical diagnosis. *Front. Genet.* **10**, 1152 (2019).
3. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
4. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
5. Peymani, F., Farzeen, A. & Prokisch, H. RNA sequencing role and application in clinical diagnostic. *Pediatr. Investig.* **6**, 29–35 (2022).
6. Geraci, F., Saha, I. & Bianchini, M. Editorial: RNA-Seq analysis: methods, applications and challenges. *Front. Genet.* **11**, 220 (2020).
7. Wang, Y. et al. Changing technologies of RNA sequencing and their applications in clinical oncology. *Front. Oncol.* **10**, 447 (2020).
8. Orange, D. E. et al. RNA identification of PRIME cells predicting rheumatoid arthritis flares. *N. Engl. J. Med.* **383**, 218–228 (2020).
9. Mittempergher, L. et al. MammaPrint and BluePrint molecular diagnostics using targeted RNA next-generation sequencing technology. *J. Mol. Diagn.* **21**, 808–823 (2019).
10. Silvestri, G. A. et al. A bronchial genomic classifier for the diagnostic evaluation of lung cancer. *N. Engl. J. Med.* **373**, 243–251 (2015).
11. Crow, M., Lim, N., Ballouz, S., Pavlidis, P. & Gillis, J. Predictability of human differential gene expression. *Proc. Natl Acad. Sci. USA* **116**, 6491–6500 (2019).
12. Rodriguez-Esteban, R. & Jiang, X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med. Genomics* **10**, 59 (2017).
13. Wang, C. et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32**, 926–932 (2014).
14. Shi, L., Perkins, R. G., Fang, H. & Tong, W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr. Opin. Biotechnol.* **19**, 10–18 (2008).

15. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
16. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
17. Li, S. et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).
18. Salit, M. & Woodcock, J. MAQC and the era of genomic medicine. *Nat. Biotechnol.* **39**, 1066–1067 (2021).
19. Genomics informatics—reliability assessment criteria for high-throughput gene-expression data. ISO/TS 22690:2021. <https://www.iso.org/standard/73691.html> (International Organization for Standardization, 2021).
20. Su, Z. et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
21. Li, S. et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32**, 915–925 (2014).
22. Geiss, G. K. et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.* **26**, 317–325 (2008).
23. Liang, P. MAQC papers over the cracks. *Nat. Biotechnol.* **25**, 27–28; author reply 28–29 (2007).
24. Thermo Fisher Scientific. <https://www.thermofisher.cn/order/catalog/product/QS0611> (2022).
25. Fang, L. T. et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nat. Biotechnol.* **39**, 1151–1160 (2021).
26. Xiao, W. et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* **39**, 1141–1150 (2021).
27. Gong, B. et al. Cross-oncopanel study reveals high sensitivity and accuracy with overall analytical performance depending on genomic regions. *Genome Biol.* **22**, 109 (2021).
28. Foox, J. et al. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat. Biotechnol.* **39**, 1129–1140 (2021).
29. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
30. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
31. Ren, L. et al. Quartet DNA reference materials and datasets for comprehensively evaluating germline variants calling performance. Preprint at bioRxiv <https://doi.org/10.1101/2022.09.28.509844> (2022).
32. Tian, S. et al. Quartet protein reference materials and datasets for multi-platform assessment of label-free proteomics. *Genome Biol.* <https://doi.org/10.1186/s13059-023-03048-y> (2022).
33. Zhang, N. et al. Quartet metabolite reference materials for assessing inter-laboratory reliability and data integration of metabolomic profiling. Preprint at bioRxiv <https://doi.org/10.1101/2022.11.01.514762> (2022).
34. Jiang, Y. Z. et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell* **35**, 428–440 (2019).
35. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
36. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
37. Reference materials—guidance for characterization and assessment of homogeneity and stability. ISO Guide 35:2017. <https://www.iso.org/standard/60281.html> (International Organization for Standardization, 2017).
38. Uncertainty of measurement—part 3: guide to the expression of uncertainty in measurement (GUM:1995). ISO/IEC Guide 98-3:2008. <https://www.iso.org/standard/50461.html> (International Organization for Standardization, 2008).
39. General and statistical principles for characterization of reference materials. JJF 1343-2012. [http://naccrm.china-csm.org/news\\_info.aspx?id=695](http://naccrm.china-csm.org/news_info.aspx?id=695) (Standardization Administration of China, 2012).
40. Yang, J. et al. The Quartet Data Portal: integration of community-wide resources for multiomics quality control. Preprint at bioRxiv <https://doi.org/10.1101/2022.09.26.507202> (2022).
41. Liu, Y. et al. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **37**, 314–322 (2019).
42. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
43. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
44. Choy, E. et al. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* **4**, e1000287 (2008).
45. Rossi, A. & Kontarakis, Z. Beyond Mendelian inheritance: genetic buffering and phenotype variability. *Phenomics* **2**, 79–87 (2022).
46. Kumar, G., Ertel, A., Feldman, G., Kupper, J. & Fortina, P. iSeqQC: a tool for expression-based quality control in RNA sequencing. *BMC Bioinformatics* **21**, 56 (2020).
47. Mohorianu, I. et al. Comparison of alternative approaches for analysing multi-level RNA-seq data. *PLoS ONE* **12**, e0182694 (2017).
48. Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
49. 't Hoen, P. A. C. et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
50. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35**, 498–507 (2017).
51. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
52. Zheng, Y. et al. Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01934-1> (2023).
53. Yu, Y. et al. Correcting batch effects in large-scale multiomic studies using a reference-material-based ratio method. *Genome Biol.* <https://doi.org/10.1186/s13059-023-03047-z> (2022).
54. Guo, L. et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24**, 1162–1169 (2006).
55. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
56. Zhang, J. et al. The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
57. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
58. Denisenko, E. et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
59. Somekh, J., Shen-Orr, S. S. & Kohane, I. S. Batch correction evaluation framework using a-priori gene–gene associations: applied to the GTEx dataset. *BMC Bioinformatics* **20**, 268 (2019).
60. Schmidt, F. et al. An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics* **34**, i908–i916 (2018).

61. Sahraeian, S. M. E. et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* **8**, 59 (2017).
62. Labaj, P. P. & Kreil, D. P. Sensitivity, specificity, and reproducibility of RNA-Seq differential expression calls. *Biol. Direct* **11**, 66 (2016).
63. Corchete, L. A. et al. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci. Rep.* **10**, 19737 (2020).
64. Yu, G., Wang, L., Han, Y. & He, Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

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

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

© The Author(s) 2023, corrected publication 2023

<sup>1</sup>State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Shanghai Cancer Center, Fudan University, Shanghai, China. <sup>2</sup>National Institute of Metrology, Beijing, China. <sup>3</sup>Greater Bay Area Institute of Precision Medicine, Guangzhou, China. <sup>4</sup>Nexomics Biosciences Institute, Wuhan, China. <sup>5</sup>Genome Decoding Institute, Nantong, China. <sup>6</sup>Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. <sup>7</sup>EATRIS ERIC-European Infrastructure for Translational Medicine, Amsterdam, The Netherlands. <sup>8</sup>Nanjing Vazyme Biotech Co. Ltd., Nanjing, China. <sup>9</sup>MGI, BGI-Shenzhen, Shenzhen, China. <sup>10</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. <sup>11</sup>Office of Oncologic Diseases, Office of New Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA. <sup>12</sup>Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA. <sup>13</sup>National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital, Beijing, China. <sup>14</sup>National Center of Gerontology, Beijing, China. <sup>15</sup>Shanghai Public Health Clinical Center, Fudan University, Shanghai, China. <sup>16</sup>International Human Phenome Institutes, Shanghai, China. <sup>17</sup>These authors contributed equally: Ying Yu, Wanwan Hou, Yaqing Liu, Haiyan Wang, Lianhua Dong.

✉ e-mail: [joshua.xu@fda.hhs.gov](mailto:joshua.xu@fda.hhs.gov); [fengqian@fudan.edu.cn](mailto:fengqian@fudan.edu.cn); [ruihang@nccl.org.cn](mailto:ruihang@nccl.org.cn); [lemingshi@fudan.edu.cn](mailto:lemingshi@fudan.edu.cn); [zhengyuanting@fudan.edu.cn](mailto:zhengyuanting@fudan.edu.cn)

## Methods

### Cell lines

Human subjects, establishment of the EBV-transformed B-LCLs, expansion and cryopreservation of the cells, cell culture and cell quality control are described in an accompanying paper by Zheng et al.<sup>52</sup>. In brief, four healthy volunteers from a quartet family in Taizhou, Jiangsu, China were enrolled, and their peripheral blood samples were collected. The study was approved by the institutional review board of the School of Life Sciences, Fudan University (BE2050). Peripheral blood mononuclear cells were isolated; the naïve B cells were sorted and infected with EBV by centrifugation at 400g for 1 h; and the immortalized cell lines were cultured in an incubator. About  $1.0 \times 10^{11}$  cells were harvested for each cell line in the same batch to ensure that multi-omics reference materials were extracted from the same batch of cultured cells. About  $2.0 \times 10^9$  cells per cell line were used for generating Quartet RNA reference materials.

### RNA extraction and quality assessment

TRIzol reagent was added to resuspend the cells. Total RNA was extracted using an RNeasy Maxi Kit (Qiagen, cat. no. 75162) including on-column DNase-I digest, according to the manufacturer's instructions.

RIN values were obtained for assessing RNA quality with a Bioanalyzer 2100 (Agilent Technologies) using RNA 6000 Nano assay (Agilent Technologies) and a Qsep 100 system (BiOptic). RNA concentrations, OD<sub>280/260</sub>, OD<sub>260/230</sub> and 28/18S were assessed by a NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific). Over 5 mg of RNA was obtained per cell line. RNAs were then aliquoted into more than 1,000 tubes per sample group with 5 µg of RNA per tube.

As a part of the Quartet Project, multi-omics reference materials (DNA, RNA, protein and metabolite) were established simultaneously from the same batch of cultured EBV-immortalized B-LCLs from the Quartet family members. The Quartet multi-omics reference materials are available to the public. Users can request reference materials via the Quartet Data Portal (<http://chinese-quartet.org/>).

### RNA stability assessment

**Bottle-opening and freeze–thaw stability.** RNAs were stored in 0.5-ml tubes at -80 °C for over 1 h until completely frozen. Frozen samples were thawed at 4 °C for approximately 0.5 h until completely thawed (freeze–thaw 1). We then opened the tubes and took 1-µl aliquots per tube out for further analysis (bottle-opening 1). The remaining RNAs were immediately re-frozen at -80 °C. This cycle was repeated for 20 times. RIN values were assessed at the 0, 1, 2, 3, 4, 5, 6, 8, 10, 14, 16, 18 and 20 times of opening and freeze–thaw to evaluate the integrity of RNA. Three replicates per sample group were assessed during each assessment.

**Short-term stability.** The stability of Quartet RNA reference materials at room temperature (22–25 °C) and 4 °C was assessed. First, four groups of the Quartet RNA reference materials were assessed for up to 4 d. RIN values were assessed at 0 h, 6 h, 24 h and 4 d to evaluate the overall quality of RNA during storage. Second, considering the same trends and similar results across the four Quartet RNA reference materials, we used two RNA reference materials (F7 and M8) for up to 14 d. RIN values were assessed at 0, 2, 4, 5, 6, 7, 8, 10, 12 and 14 d, separately. Three replicates per sample group were assessed during each assessment.

**Long-term stability.** The stability of RNA reference materials at storage of -80 °C was monitored for up to 20 months. RIN values were assessed at 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17 and 20 months. Three replicates per sample group were assessed at each timepoint. The MAQC RNA reference materials, including A sample (Universal Human Reference RNA, Agilent Technologies) and B sample (Human Brain Reference RNA, Thermo Fisher Scientific)<sup>15</sup>, were used as controls at each timepoint.

### Library construction and sequencing

According to the Quartet Project study design, 12 tubes of RNA samples were sent to each laboratory, including four groups of the Quartet RNA reference materials with triplicates per group. Library preparation, library quality control and sequencing were conducted in a fixed order (D5-1, D6-1, F7-1, M8-1, D5-2, D6-2, F7-2, M8-2, D5-3, D6-3, F7-3 and M8-3) in each laboratory to eliminate confounding factors, such as experimental sample processing order with sample group.

RNA-seq library preparation and high-throughput sequencing were conducted by each laboratory. In brief, libraries were constructed by PolyA selection or ribosomal RNA depletion (RiboZero) methods. The libraries were sequenced on Illumina NovaSeq (ILM) or MGII DNBSEQ-T7 (BGI) platforms with paired-end (PE) reads of 100–150 base pairs (bp). A total of 252 Quartet RNA-seq libraries from 21 batches were generated. Additionally, we simultaneously generated 20 batches of RNA-seq datasets using MAQC reference materials as controls. Detailed information on RNA-seq library construction and sequencing is shown in Supplementary Table 2.

Four RNA libraries from whole blood of the Quartet donors were constructed by the RiboZero method (TruSeq RNA Library Prep Kit) and sequenced on an Illumina HiSeq 4000 platform with 150-bp PE reads.

### Alignment and gene quantification

Preliminary processing of raw fastq reads was performed using fastp version 0.19.6 to remove adapter sequences<sup>65</sup>. Read alignment and quantification were conducted using HISAT version 2.1, SAMtools version 1.3.1, StringTie version 1.3.4 and Ballgown version 2.14.1 (ref. 66). Reference human genome build 38 ([https://genome-idx.s3.amazonaws.com/hisat/grch38\\_snptran.tar.gz](https://genome-idx.s3.amazonaws.com/hisat/grch38_snptran.tar.gz)) and gene model from Ensembl ([http://ftp.ensembl.org/pub/release-93/gtf/homo\\_sapiens/Homo\\_sapiens.GRCh38.93.gtf.gz](http://ftp.ensembl.org/pub/release-93/gtf/homo_sapiens/Homo_sapiens.GRCh38.93.gtf.gz)) were used for read mapping and gene quantification. log<sub>2</sub> transformation was then conducted based on FPKM values. To avoid infinite values, a value of 0.01 was added to the FPKM value of each gene before log<sub>2</sub> transformation. Expression profiles based on detected genes were used for further analysis. A gene was considered detectable (expressed) in a biological group within a batch if ≥3 reads were mapped onto it in at least two of the three replicates. One replicate of MAQC B samples (library ID: R\_ILM\_L2\_B1\_B\_3) was not included in further analysis due to low quality.

Moreover, we applied the RSEM tool for gene quantification to evaluate the impact of pipeline on assessing the reliability of RNA-seq data by comparing with the results from StringTie. Specifically, two batches of RNA-seq datasets from Quartet RNA reference materials were used, including one high-quality batch (R\_ILM\_L8\_B1) and one low-quality batch (R\_ILM\_L4\_B1) based on prior performance evaluation. Read alignment and quantification was conducted using Bowtie2 version 2.5.1 (ref. 67) and RSEM version 1.2.28 (ref. 68). Parameters were set by default according to the recommended pipeline from RSEM (<https://github.com/deweylab/RSEM>). The same reference genome and gene model were used. Expression matrix of FPKM values was obtained.

Quality control analysis of sequencing data at pre-alignment and post-alignment level was conducted using FastQC version 0.11.5 (ref. 69), FastQ Screen version 0.12.0 (ref. 70), Qualimap version 2.0.0 (ref. 71) and MultiQC version 1.8 (ref. 72).

### Validation of analysis pipeline based on MAQC reference materials

The bioinformatics pipeline was validated using published data from the MAQC RNA reference materials<sup>15,20</sup>. Specifically, we downloaded the published RNA-seq fastq files through the Gene Expression Omnibus (GEO) ([GSE47774](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47774)), analyzed the dataset using the bioinformatics pipeline used in this study and generated an expression matrix of FPKM values as the dataset for validating the reliability of our pipeline. Meanwhile, the expression matrix of count values of the same dataset was obtained from the R/Bioconductor package seqc

version 1.28.0 (ref. 20), which could be used as positive control dataset. Data were normalized to counts per million (CPM). A value of 0.01 was added to the FPKM or CPM value of each gene, and log<sub>2</sub> transformation was then conducted. Moreover, the expression profiles of the MAQC RNA reference materials from TaqMan assay were obtained through the GEO (GSE5350)<sup>15</sup>, which were used as the ‘ground truth’. Gene ID was mapped to Ensembl gene ID using the R/Bioconductor package biomaRt version 2.50.1. A total of 725 genes were mapped to Ensembl gene ID and were detected both in the RNA-seq and TaqMan data. Pearson and Spearman correlation coefficients based on log<sub>2</sub> ratios (A/B) were further computed.

### Ratio-based expression

Ratio-based expression data were obtained within each batch on a gene-by-gene basis. Specifically, ratio-based expressions were calculated based on log<sub>2</sub>FPKM values. For each gene, the mean of expression profiles of replicates of reference sample(s) (for example, D6) was first calculated and then subtracted from the log<sub>2</sub>FPKM values of that gene in each study.

### DEGs

Differential expression analyses were implemented using the limma version 3.50.0 (ref. 73) and edgeR version 3.36.0 (ref. 74) packages according to guidelines from the limma package. A gene was considered differentially expressed in a batch between two sample groups if two-sided  $P < 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$  using the limma package for upregulation or downregulation, respectively.

### Identification and quantification of alternative splicing

The alignment results based on the HISAT2 were used to identify AS events using SplAdder toolkit<sup>75</sup> with the default parameters. Six types of AS events were quantified using percent spliced in values, including exon skip, intron retention, alternative 3' splice site, alternative 5' splice site, cassette exon and coordinated mutually exclusive exons.

### Construction of reference datasets

We constructed the reference datasets of ratio-based expression based on the following steps: (1) identifying detectable genes; (2) calculating ratio-based expression based on reliable genes that were differentially expressed and with low uncertainty; (3) assessing the homogeneity and stability; (4) assessing the uncertainty of ratio-based reference datasets; and (5) calculating high-confidence DEGs in reference datasets.

First, detectable genes were identified. A gene was considered expressed in a sample in each batch if more than three reads were mapped to it in at least two of the three replicates. If a gene was detected in all the 13 batches in a sample group (D5, D6, F7 and M8), it was considered expressed in that sample group.

Second, ratio-based expressions were calculated. We used the expression profiles of three replicates of D6 in the same batch as the denominators and derived the ratio-based expressions for the three sample pairs (D5/D6, F7/D6 and M8/D6). The reference ratio-based expressions between each pair of samples for a gene were provided in the format of an average by summarizing from the 13 fold changes calculated from each of the 13 high-quality RNA-seq datasets. To improve the reliability of the reference values, genes were included if they satisfied the following criteria: (1) detectable across the two groups of each sample pair; (2) limma-based<sup>73</sup> two-sided  $P < 0.05$  in at least four batches in each sample pair; and (3) not significantly different between PolyA and RiboZero protocols (Student's *t*-test two-sided  $P > 0.05$  or fold change  $< 2$  and  $> 0.5$ ).

Third, the homogeneity and stability were assessed using RNA-seq datasets. The Quartet RNA reference materials were considered to be homogenous and stable, as can be seen from the corresponding reference datasets. Additionally, uncertainties of reference materials were assessed.

Finally, high-confidence DEGs in the reference datasets (reference DEGs) were identified. A gene was considered as a reference DEG between two sample types if it was concordantly discovered as an upregulated or downregulated gene (two-sided  $P < 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$ ) in more than six of the 13 high-quality batches.

### Homogeneity assessment based on RNA-seq datasets

The homogeneity of the Quartet RNA reference materials was assessed using RNA-seq data. We randomly selected 17 tubes (units) of each Quartet RNA reference material and named them as N1–N17. Under the same condition, nine replicates in the N1 tube and one replicate in tubes N2–N17 of each material were assessed to represent within-unit ( $n = 9$ ) and between-unit ( $n = 16$ ) characteristics. A total of 25 RNA-seq experiments per reference material were conducted.

RNA-seq libraries were constructed by ribosomal RNA depletion methods (VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina) and sequenced on the Illumina NovaSeq platform with 150-bp PE reads. Alignment, quantification and quality control were conducted using the same analysis pipeline and parameters described above.

The within-unit and between-unit variances were then calculated using the ANOVA method<sup>37,39</sup>. Ratio-based expressions were obtained by subtracting log<sub>2</sub>FPKM by the mean of log<sub>2</sub>FPKM of the three replicates of D6 in the same batch and used. A gene was considered to be homogeneous when a cutoff of false discovery rate (FDR)-adjusted ANOVA-based two-sided  $P > 0.05$  was used. Only between-unit homogeneity is studied, because within-unit homogeneity might be negligible in the case of intrinsically homogeneous materials, such as solutions<sup>76</sup>.

### Long-term stability assessment based on RNA-seq datasets

We assessed the long-term stability of the reference materials of 15 batches of RNA-seq datasets that were generated from up to 26 months. Ratio-based expressions were obtained by subtracting the mean log<sub>2</sub>FPKM of the three replicates of D6 in the same batch from the log<sub>2</sub>FPKM values. According to ISO Guide 35 (2017)<sup>37</sup> and SAC JJF-1343 (2012)<sup>39</sup>, long-term stability assessment was conducted based on regression analysis. For each gene, the observed slope  $b_1$  and uncertainty of slope  $b_1$  ( $s(b_1)$ ) was calculated. If  $|b_1| < s(s(b_1)) \times t_{0.95,n-2}$ , the expression of the gene is stable, and vice versa, where  $t_{0.95,n-2}$  is critical *t* value for a confidence level of 95% and  $n - 2$  degrees of freedom.

### Uncertainty assessment of reference datasets

According to ISO Guide 35 (2017)<sup>37</sup>, ISO/IEC Guide 93-3 (2008)<sup>38</sup> and SAC JJF-1343 (2012)<sup>39</sup>, the source of uncertainties can be classified into characterization uncertainties ( $u_{char}$ ), sample inhomogeneities (between-bottle variation,  $u_{bb}$ ) and instabilities ( $u_s$ ). These values were then combined to form the combined uncertainties ( $u_c$ ) with a simple additive measurement model using an equal weight of the three uncertainty sources, as recommended<sup>37–39</sup>. The expanded uncertainties ( $U$ ) were further computed by multiplying  $u_c$  with an expansion factor.

First, characterization uncertainty of genes in the reference datasets was evaluated using 13 fold changes (log<sub>2</sub> scale) from each of 13 high-quality RNA-seq datasets. Relative uncertainty of characterization was used as characterization uncertainty ( $u_{char}$ ), which can be expressed as equation (1) as follows:

$$u_{char} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1) \times n}} / \bar{x} \quad (1)$$

where  $n$  is number of measurements in the sample;  $x_i$  is measurement value of  $i$ th time; and  $\bar{x}$  is average value of  $x$  across  $n$  times.

Second, sample inhomogeneity  $u_{bb}$  was evaluated using RNA-seq datasets.  $u_{bb}$  can be expressed as equations (2) and (3):

When  $s_1^2 > s_2^2$ ,

$$u_{bb} = \sqrt{\frac{s_1^2 - s_2^2}{n}} \quad (2)$$

When  $s_1^2 < s_2^2$ ,

$$u_{bb} = \sqrt{\frac{s_2^2}{n}} \times \sqrt{\frac{2}{v_{s_2^2}}} \quad (3)$$

where  $s_1^2$  is between-unit variation;  $s_2^2$  is within-unit variation;  $v_{s_2^2}$  is degree of freedom of  $s_2^2$ ; and  $n$  is number of between-unit measurements.

Third, long-term instability ( $u_s$ ) was evaluated based on RNA quality RIN across 20 months, which can be expressed as equation (4)

$$u_s = t \times \sqrt{\frac{\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2}{(n-2) \times \sum_{i=1}^n (x_i - \bar{x})^2}} \quad (4)$$

where  $b_0$  and  $b_1$  are the intercept and slope of linear regression line between  $x_i$  (month) and  $y$  (RIN);  $t$  is time (month); and  $n$  is number of observations. Short-term instability might be negligible, because reference materials are recommended to be transported using dry ice.

Fourth, a combined uncertainty ( $u_c$ ) should consider all uncertainty described above, which can be expressed as equation (5):

$$u_c = \sqrt{u_{char}^2 + u_{bb}^2 + u_s^2} \quad (5)$$

Finally, an extended uncertainty ( $U$ ) can be expressed as equation (6):

$$U = k \times u_c \quad (6)$$

where  $k$  is a constant value. Here,  $k = 2$  was applied for 95% confidence level.

## Performance metrics

Performance metrics, including SNR, RC with reference datasets, RMSE of differences with reference datasets and MCC of DEGs, were developed to evaluate the quality of RNA-seq data at expression level before a total score was calculated.

**SNR.** SNR is a measurement used in science and engineering. SNR is defined as the ratio of the power of a signal to the power of noise and is often expressed in decibels ([https://en.wikipedia.org/wiki/Signal-to-noise\\_ratio](https://en.wikipedia.org/wiki/Signal-to-noise_ratio)). In this study, the average distances representing the intrinsic ‘differences’ among distinct biological sample groups are regarded as the signal, whereas the average distances among technical replicates of the same sample group are regarded as noise.

To identify an effective way to calculate the SNR values, we evaluated the performances of SNR values calculated by five different algorithms depending on whether the sample–sample ‘distance’ (signal or noise) is calculated based on the original feature space or the dimensionality-reduced space and how the distance was calculated. For the original feature space, the distance was calculated in three different ways: Euclidean distance or (1 – Pearson correlation coefficient) with all features considered simultaneously as a vector to represent a sample (abbreviated as OriAll\_EucDist or OriAll\_1-Cor, respectively) and the median of the Euclidean distances across all features when each single feature is separately used to represent a sample (OriSingle\_MedianEucDist). For the dimensionality-reduced space, the Euclidean distance was calculated using the coordinates of

a sample in the PC space from either  $t$ -distributed stochastic neighbor embedding (tSNE) or PCA, abbreviated as ReducedDim\_tSNE and ReducedDim\_PCA, respectively. The numbers of PCs used in calculating SNR were then determined. We decided to use the first two components in PCA to calculate SNR values in correspondence with visualization in PCA plots.

Therefore, SNR is defined as equation (7):

$$SNR = 10 \times \log_{10} \left( \frac{m \times \binom{n}{2}}{\binom{m}{2} \times n \times n} \right) \times \frac{\sum_{x=1}^{m-1} \sum_{j=x+1}^m \sum_{i=1}^n \sum_{j=1}^n \sum_{p=1}^2 W_p (PC_{p,i,x} - PC_{p,j,y})^2}{\sum_{x=1}^m \sum_{i=1}^n \sum_{j=i+1}^n \sum_{p=1}^2 W_p (PC_{p,i,x} - PC_{p,j,x})^2} \quad (7)$$

where  $m$  is the number of sample groups, and  $n$  is the number of replicates in each sample group.  $W_p$  represents the  $p$ th PC of variances.  $PC_{p,i,x}$ ,  $PC_{p,j,x}$  and  $PC_{p,j,y}$  represent the  $p$ th component values of replicate  $i$  and replicate  $j$  in sample group  $x$  or sample group  $y$ , respectively.

A standard sample set consisted of 12 tubes with each representing one of the three replicates of the four RNA reference materials. Therefore, a typical SNR in the study was the ratio of the average distances between different biological groups ( $9 \times 12/2 = 54$ ) to the average distances between technical replicates of the same groups ( $2 \times 3 \times 4/2 = 12$ ). The distribution of intra-batch SNR values from 21 RNA-seq datasets was used to identify a threshold of 12 (mean – s.d.), indicative of high discriminating power.

**RC.** RC with reference datasets was calculated based on the Pearson correlation coefficient between the ratio-based expression levels of a dataset for a given pair of groups and the corresponding reference fold change values. It is referred to as the ‘relative correlation with reference datasets’ metric, representing the numerical consistency of the ratio-based expression profiles. To improve reliability, the mean of the three replicates of each sample group was calculated before performing ratio-based expression analysis. Fold changes were transformed using  $\log_2$  scaling.

**RMSE.** RMSE was calculated using fold changes between a test dataset for a given pair of samples and the corresponding ratio-based reference datasets, representing the average distances of ratio-based expression profiles. Fold changes were transformed using  $\log_2$  scaling. It was implemented using the rmse function from the Metrics package<sup>27</sup>.

**MCC.** MCC is a widely used statistic in the field of bioinformatics and machine learning, which combines test sensitivity and specificity<sup>20,78</sup>. In this study, we used MCC to measure the consistency of DEGs detected from a dataset for a given pair of samples with those from the reference DEGs or ‘MCC of DEGs’. Reference DEGs and non-DEGs as true-positive (TP) and true-negative (TN) sets were integrated by consensus voting. When DEGs and non-DEGs of a given dataset were identified, the number of TP, TN, false positive (FP) and false negative (FN) could be calculated. MCC is computed using equation (8):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

## Total quality score

The total quality score is calculated to measure the overall quality of a dataset generated from a laboratory for its effectiveness in quantifying the transcriptomic differences among the four Quartet RNA reference materials by summarizing reference dataset-independent quality measurement (SNR) and reference dataset-dependent quality measurement (RC). The total quality score is expressed as the geometrical mean of SNR and RC.

## Cross-validation of reference-based quality metrics

To examine if the lower RC, higher RMSE or lower MCC and MCC for the low-quality batches were caused by their exclusion from creating the reference datasets, we performed 30 times of cross-validation test. In brief, in one cross-validation, we randomly selected 13 batches from the 21 batches to create ('train') the reference datasets, which were then used to calculate quality measurements for all the 21 batches. Both high-quality and low-quality batches might be randomly included or excluded from 'training the reference datasets', either as training or validation sets.

## Co-expression analysis

Co-expression network was constructed using the R package WGCNA version 1.71 (ref. 43) using absolute expression and ratio-based expression profiles of 13 batches with high quality, respectively. Genes with the highest variations ( $n = 10,000$ ) were used for conducting co-expression network. Modules were then identified with a dynamic tree-cutting algorithm with a minimum module size of 50. Modules were named in color. Thirteen (13) modules were identified based on absolute expressions, including turquoise ( $n = 4,633$ ), blue ( $n = 933$ ), brown ( $n = 753$ ), yellow ( $n = 728$ ), green ( $n = 527$ ) and so on. Moreover, eight modules were identified based on ratio-based expressions, including turquoise ( $n = 2,368$ ), blue ( $n = 1,777$ ), brown ( $n = 1,508$ ), yellow ( $n = 741$ ), green ( $n = 477$ ), red ( $n = 229$ ), black ( $n = 177$ ) and pink ( $n = 133$ ). PCA and functional analysis of each module were conducted.

Functional enrichment analyses of each module were conducted based on GO terms and were conducted with the R/Bioconductor package clusterProfiler version 4.2.2, with a Benjamini–Hochberg correction and an adjusted  $P$  value cutoff of 0.05 (ref. 64).

## RT-qPCR

Primers of 83 genes were designed using online Primers-BLAST of the National Center for Biotechnology Information based on the RNA sequences, and the PCR method of reference gene (*C1orf43*) was established previously. Primers were synthesized by Beijing Liuhe Huada Gene Technology Co. Ltd. Sequences of primers are listed in Supplementary Table 8.

RT-qPCR reactions were performed in two steps. First, reverse transcription was carried out using 2  $\mu$ l of RNA mixed with 4  $\mu$ l of 5× PrimeScript IV cDNA Synthesis Mix (Takara, code no. 6215A) containing PrimeScript IV RTase, RNase Inhibitor, Oligo dT Primer, dNTP Mixture buffer and 1  $\mu$ l of random 6mers and nuclease-free water up to 20- $\mu$ l final reaction volume. This reaction mixture was incubated at 30 °C for 10 min and then for 15 min at 42 °C and finally for 5 min at 95 °C for termination. Second, cDNA obtained in the previous step was used as template for qPCR. The qPCR reactions were carried out using UltraS-YBR Mixture (Low ROX) (CW BIO, CW2601M) containing 2  $\mu$ l of cDNA and 0.4  $\mu$ l of each forward and reverse primers (final concentration of 200 nM) in a 20- $\mu$ l final volume reaction. The qPCR was performed on a Roche 480 qPCR System using the following cycling conditions: 10 min at 95 °C, followed by 45 cycles of 15 s at 95 °C and 1 min at 60 °C. Three replicates per sample per gene were conducted for eliminating random variations.

The comparative cycle threshold (Ct) method ( $\Delta\Delta\text{Ct}$  method) was used to calculate the fold differences for the three sample pairs (D5/D6, F7/D6 and M8/D6) with housekeeping gene *C1orf43* as endogenous control. For the qPCR data, a gene is called DEG when the Student's  $t$ -test  $P$  value  $< 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$ .

## ddPCR

DEGs identified in reference datasets and RT-qPCR were further validated using ddPCR. The same sequences of primers used for RT-qPCR were used for ddPCR (Supplementary Table 8). The ddPCR reaction was performed in a QX200 Droplet Digital PCR System (Bio-Rad) according to the manufacturer's instructions. Each test was prepared

in a total of 20- $\mu$ l volume of the reaction mixture, comprising 10  $\mu$ l of EvaGreen Supermix (Bio-Rad), 2  $\mu$ l of forward and reverse primers, 2  $\mu$ l of cDNA templates and 6  $\mu$ l of RNase-free ddH<sub>2</sub>O. Samples and 70  $\mu$ l of droplet generation oil were then placed into a Droplet Generator (Bio-Rad). Droplets (40  $\mu$ l) were transferred to a 96-well PCR plate. The PCR reactions were performed using the following cycling conditions: pre-denature for one cycle at 95 °C for 5 min; denature for 40 cycles at 95 °C for 30 s; and anneal and extend for 40 cycles at 60 °C for 1 min. After the cycles, a signal stabilization step of 4 °C for 5 min and 90 °C for 5 min was conducted. The signals were read by a Droplet Reader (Bio-Rad). Each reaction was performed in duplicate.

## Flow cytometry

Immortalized B-lymphoblastoid cells were centrifuged at 500g for 10 min at room temperature. Flick or aspirate to remove supernatant, and wash cells with 2 ml of PBS at 500g for 5 min at room temperature. For the sample stain, 1 × 10<sup>6</sup> cells were resuspended in 100  $\mu$ l of PBS with 2% FBS (FACS buffer) and stained with antibody cocktail for 15 min at room temperature in the dark. After surface staining, cells were washed twice with 2 ml of PBS at 500g for 5 min at room temperature. After the final wash, cells were resuspended in 250  $\mu$ l of 1% paraformaldehyde (PFA).

The following antibodies were used for cell surface staining: PE mouse anti-human IgA (Miltenyi Biotec, 130-114-002, clone IS11-8E10), PE-Cy7 mouse anti-human IgD (BD Biosciences, 561314, clone IA6-2), Alexa Fluor 700 mouse anti-human IgG (BD Biosciences, 561296, clone G18-145) and Brilliant Violet 605 (BV605) mouse anti-human IgM (BD Biosciences, 562977, clone G20-127). PE mouse anti-human IgA was verified by the vendor, Miltenyi Biotec, including specificity, sensitivity and fixation. PE-Cy7 mouse anti-human IgD, AF700 mouse anti-human IgG and BV605 mouse anti-human IgM were validated by our previous study by flow cytometry<sup>79</sup>. Flow cytometric analyses were performed on CytoFLEX LX (Beckman Counter), and data were analyzed with FlowJo version 10.7.2 software (BD Biosciences).

The representative gating strategy for flow cytometry experiments assessing LCLs is shown in Supplementary Fig. 17. For the exclusion of non-single events, cross-check the forward scatter (FSC) signal for its area (A) versus height (H) and width (W) characteristics. Immortalized B-lymphoblastoid cells were gated on the FSC-A versus SSC-A dot plot. Furthermore, IgD<sup>+</sup> cells, IgM<sup>+</sup> cells, IgG<sup>+</sup> cells and IgA<sup>+</sup> cells in LCLs were identified based on their expression levels of surface membrane immunoglobulins.

## LC-MS/MS-based proteomics

MS-based data-dependent acquisition (DDA) proteomics dataset from Quartet protein reference materials was used for cross-omics validation. Detailed description of sample preparation and data generation was provided by Zheng et al.<sup>52</sup>. In brief, large quantities of Quartet peptide reference materials (lot: 20200616) were generated from the same generation of LCLs used in Quartet RNA reference materials. LC-MS/MS-based proteomics data (4 groups × 3 replicates) were then generated in a laboratory (code: NVG) using a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific).

Peptide and protein identification and quantification were conducted using Proteome Discoverer 2.2 (PD 2.2, Thermo Fisher Scientific) based on the human reference database UniProt (<http://www.uniprot.org>). Proteins with at least one unique peptide with 1% FDR at the peptide level were retained for further analysis. Protein quantification was normalized using the fraction of total (FOT). The fraction of total was multiplied by 10<sup>5</sup> for ease of presentation.

## Biological classifications from published datasets

We used publicly available datasets to examine the extent of biological differences with four 'intrinsic' biological classification groups from published datasets<sup>34–36</sup>. Expression profiles in FPKM of four subtypes

of TNBCs with different therapeutic actions were downloaded from previous publication<sup>34</sup>, including basal-like and immune-suppressed ( $n = 124$ ), luminal androgen receptor ( $n = 75$ ), immunomodulatory subtype ( $n = 77$ ) and mesenchymal-like subtype ( $n = 60$ ). Expression profiles in FPKM of the four molecular subtypes of breast cancer were downloaded from the Genomic Data Commons (GDC) Data Portal<sup>35</sup>, including luminal A ( $n = 420$ ), luminal B ( $n = 174$ ), basal-like ( $n = 140$ ) and Her2-enriched ( $n = 65$ ). Expression profiles in FPKM from four cancer types with distinct tissue types were also downloaded from the GDC Data Portal<sup>35</sup>, including brain cancer ( $n = 74$ ), breast cancer ( $n = 77$ ), kidney cancer ( $n = 67$ ) and lung cancer ( $n = 66$ ). Expression profiles in count from four normal tissue types were obtained from GTEx version 8, including brain ( $n = 100$ ), breast ( $n = 100$ ), kidney ( $n = 89$ ) and lung ( $n = 100$ )<sup>36</sup>. Count data were normalized to CPM using the limma version 3.50.0 (ref. 73) package. Three samples from each clinical subtype or biological group were randomly selected for differential expression analysis to eliminate the effect of number of samples used for analysis. To eliminate selection biases, this process was repeated 20 times. A gene was considered as a DEG when  $t$ -test  $P < 0.05$  and fold change  $\geq 2$  or  $\leq 0.5$ .

## Statistical analysis

All statistical analyses were performed using R statistical software version 4.1.2 (<https://www.r-project.org>). PCA was conducted with the unvariance scaling, using the prcomp function. Hierarchical clustering analysis (HCA) was performed using the R package pheatmap version 1.0.12 (<https://rdrr.io/cran/pheatmap/>). Data visualization was implemented using the R package ggplot2 version 3.3.5 (<https://ggplot2.tidyverse.org/>), GGally version 2.1.2 (<http://ggobi.github.io/ggally/>) and ggsci version 2.9 (<https://github.com/nanxstats/ggsci>).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The raw sequence data and gene expression data reported in this paper have been deposited in the Genome Sequence Archive (GSA) (accession number: HRA001859)<sup>80</sup> and the Open Archive for Miscellaneous Data (OMIX) (accession number: OMIX002254)<sup>81</sup> of the China National Center for Bioinformation. Moreover, we developed the Quartet Data Portal (<http://chinese-quartet.org>) for the community to access and share the Quartet multi-omics resources.

## Code availability

The source codes for the data analyses are available at Zenodo<sup>82,83</sup>.

## References

65. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
66. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
68. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
69. Andrews, S. FastQC: a quality control tool for high throughput sequence data <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2017).
70. Wingett, S. W. & Andrews, S. FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Res.* **7**, 1338 (2018).
71. García-Alcalde, F. et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
72. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
73. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
74. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
75. Kahles, A., Ong, C. S., Zhong, Y. & Ratsch, G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**, 1840–1847 (2016).
76. Ortega-Gavilan, F., Valverde-Som, L., Rodriguez-Garcia, F. P., Cuadros-Rodriguez, L. & Bagur-Gonzalez, M. G. Homogeneity assessment of reference materials for sensory analysis of liquid foodstuffs. The virgin olive oil as case study. *Food Chem.* **322**, 126743 (2020).
77. Hamner, B., Frasco, M. & LeDell, E. Metrics: evaluation metrics for machine learning. CRAN <https://cran.r-project.org/web/packages/Metrics/index.html> (2018).
78. Shi, L. et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).
79. Gao, J. et al. Deep immunophenotyping of human whole blood by standardized multi-parametric flow cytometry analyses. *Phenomics* **3**, 309–328 (2023).
80. Quartet Project Team. Quartet Project for quality control and data integration of multiomics profiling. Genome Sequence Archive <https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001859> (2023).
81. Quartet Project Team. Expression profiles of Quartet RNA-seq dataset. OMIX <https://ngdc.cncb.ac.cn/omix/release/OMIX002254> (2023).
82. Quartet Project Team. YingYu12345/Quartet-RNA-reference-materials: Quartet RNA reference materials (v1.0.1). Zenodo <https://doi.org/10.5281/zenodo.8014734> (2023).
83. Quartet Project Team. chinese-quartet/quartet-rseqc-report: version v0.2.3. Zenodo <https://doi.org/10.5281/zenodo.8014752> (2023).

## Acknowledgements

We thank the Quartet Project team members who contributed their time and resources to the design and implementation of this project. We are grateful to CFFF (Computing for the Future at Fudan) and the Human Phenome Data Center of Fudan University for computing support. This study was supported, in part, by the National Key R&D Project of China (grant 2018YFE0201603 to Y.T.Z. and grant 2018YFE0201600 to L.S.); the National Natural Science Foundation of China (grant 31720103909 to L.S. and grant 32170657 to L.S.); the Shanghai Municipal Science and Technology Major Project; the State Key Laboratory of Genetic Engineering (grant SKLGE-2117 to J.L., L.S. and Y.T.Z.); and the 111 Project (grant B13016 to L.S.). This research was supported, in part, by the Intramural Research program of the National Library of Medicine, US National Institutes of Health. The content of the article is solely the responsibility of the authors and does not necessarily represent the official views of the US Food and Drug Administration. Some of the illustrations in this paper were created with BioRender.

## Author contributions

Y.T.Z., L.S. and Y.Y. conceived the study. Y.T.Z., L.S., R.Z., F.Q., J.X., Y.Y., L.J., X.F., J.L., W.T., H.H. and W.X. oversaw the study. Y.T.Z., W.H., H.W., S.S., Z.C., P.Z., Y.Z., R.L., S.Z., X.W., D.B. and B.Y.L. cultured the cell lines and prepared and characterized the RNA reference materials. J.G. and F.Q. performed flow cytometry assays and data interpretation. Y.T.Z., W.H., L.D., H.W., J.H. and S.S. coordinated and/or performed

NGS library preparation and sequencing. L.D., Y.T.Z., X.W., Y.P. and Y.Y. performed qPCR validation. Y.Y., W.H., Y.L., H.W., L.D., Y.M., S.S., J.Y., Z.C., Q.C., Z.L., Z.C., N.Z., J.L., L.R., H.J., J.S., T.Q., B.L., C.S., F.D., A.S., P.M., J.H., L.Z., H.J., D.T.-M., J.T.-M., W.X., H.H., W.T., J.W., J.L., X.F., L.J., J.X., F.Q., R.Z., L.S. and Y.T.Z. performed data analysis and/or interpretation. Y.Y., J.Y. and J.S. managed the datasets. Y.Y. generated most figures. Y.Y., Y.T.Z. and L.S. wrote and revised the manuscript. All authors reviewed and approved the manuscript. Dozens of participants of the Quartet Project freely donated their time and reagents for the completion and analysis of the project.

### Competing interests

J.H. is an employee of Vazyme Biotech Co. Ltd. L.Z. is the co-founder of Vazyme Biotech Co. Ltd. H.J. is an employee of MGI, BGI-Shenzhen. All other authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41587-023-01867-9>.

**Correspondence and requests for materials** should be addressed to Joshua Xu, Feng Qian, Rui Zhang, Leming Shi or Yuanting Zheng.

**Peer review information** *Nature Biotechnology* thanks Pedro Ferreira and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection Software: HISAT v2.1, SAMtools v1.3.1, StringTie v1.3.4, Ballgown v2.14.1, FastQC v0.11.5, FastQ Screen v0.12.0, Qualimap v2.0.0, MultiQC v1.8. The source codes for the data collection are available at Github (<https://doi.org/10.5281/zenodo.8014752>)

Data analysis Software: R v4.1.2. R code for data analysis and figure generation is available at Github (<https://doi.org/10.5281/zenodo.8014734>)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The raw sequence data and gene expression data reported in this paper have been deposited in the Genome Sequence Archive (GSA) (accession number: HRA001859, <https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001859>) and Open Archive for Miscellaneous Data (OMIX) (accession number: OMIX002254, <https://>)

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

### Reporting on sex and gender

We used blood samples from four participants from a Chinese Quartet family from the Fudan Taizhou Cohort, including father (F7), mother (M8) and monozygotic twin daughters (D5 and D6). Sex and gender of the participants are F7 (male), M8 (female), D5 (female) and D6 (female). Sex and gender information was determined based on self-reporting and DNA sequencing.

### Population characteristics

All four participants are adults from Han Chinese. No additional covariates-relevant population characteristics were collected for each donor due to IRB approval restrictions.

### Recruitment

The participants were recruited by advertisements in the Taizhou Longitudinal Study. We randomly selected a family with monozygotic twin daughters. No self-selection bias was expected to be introduced.

### Ethics oversight

This study was approved by the Institutional Review Board (IRB) of the School of Life Sciences, Fudan University (BE2050). It was conducted under the principles of the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](#)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

As a proficiency study, the study necessarily employed large sample sizes. A total of 252 RNA-seq libraries from 21 batches were generated in eight labs using two library construction protocols (PolyA selection and RiboZero) and two sequencing platforms (Illumina NovaSeq (ILM) and MGI DNaseQ-T7 (BGI)). In this study, a batch is defined as 12 libraries from a standard sample set, consisting of 12 vials with each representing one of the triplicates of the Quartet RNA reference sample groups, whose library construction and sequencing experiments were conducted simultaneously.

### Data exclusions

All data from planned experiments have been included.

### Replication

The reference materials were profiled within a batch in a lab in three replicates for each of the four samples (donors).

### Randomization

Aliquots of RNA from the same lot were randomly distributed to each center.

### Blinding

Each batch of samples distributed was blinded to avoid specific experimental sequences affecting the objective assessment of lab proficiency.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- n/a Involved in the study
- Antibodies
  - Eukaryotic cell lines
  - Palaeontology and archaeology
  - Animals and other organisms
  - Clinical data
  - Dual use research of concern

### Methods

- n/a Involved in the study
- ChIP-seq
  - Flow cytometry
  - MRI-based neuroimaging

## Antibodies

Antibodies used	IgA (clone IS11-8E10) (Miltenyi Biotec, cat#: 130-114-002), IgD (clone IA6-2) (BD Biosciences, cat#: 561314), IgG (clone G18-145) (BD Biosciences, cat#: 561296), and IgM (clone G20-127) (BD Biosciences, cat#: 562977).
Validation	PE Mouse Anti-Human IgA (Miltenyi Biotec, cat#: 130-114-002, clone IS11-8E10) was verified by vendor Miltenyi Biotec, including specificity, sensitivity, and fixation. PE-Cy7 Mouse Anti-Human IgD (BD Biosciences, cat#: 561314, clone IA6-2), Alexa Fluor 700 Mouse Anti-Human IgG (BD Biosciences, cat#: 561296, clone G18-145), and Brilliant Violet 605 (BV605) Mouse Anti-Human IgM (BD Biosciences, cat#: 562977, clone G20-127) were validated by our previous study by flow cytometry (Gao, J., Luo, Y., Li, H. et al. <i>Phenomics</i> , 2023, <a href="https://doi.org/10.1007/s43657-022-00092-9">https://doi.org/10.1007/s43657-022-00092-9</a> ).

## Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	The human immortalized B-lymphoblastoid cell lines of four healthy volunteers from a family Quartet, as part of the Taizhou Longitudinal Study in Taizhou, Jiangsu, China. The sex of the primary cell lines generated from human participants are: F7 (male), M8 (female), D5 (female) and D6 (female).
Authentication	The cell lines have been authenticated by STR profile, karyotype, PCR mycoplasma and sterility testing.
Mycoplasma contamination	No mycoplasma contamination found.
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No commonly misidentified cell lines were used in this study.

## Flow Cytometry

### Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation	Immortalized B-lymphoblastoid cells were centrifuged at 500× g for 10 min at RT (room temperature). Flick or aspirate to remove supernatant and wash cells with 2 mL PBS at 500× g for 5 min at RT. For the sample stain, 1×10^6 immortalized B-lymphoblastoid cells were resuspended in 100 µl PBS with 2% FBS (FACS buffer) and stained with antibody cocktail for 15 min at RT in the dark. Following surface staining using antibodies, cells were washed twice with 2 mL PBS at 500× g for 5 min at RT. After the final wash, cells were resuspended in 250 µl 1% Paraformaldehyde (PFA).
Instrument	CytoFLEX LX (Beckman Counter)
Software	FlowJo V10.7.2 software (BD Biosciences)
Cell population abundance	This study did not perform sorting.
Gating strategy	For the exclusion of non-single events, cross-check the forward scatter (FSC) signal for its area (A) versus height (H) and width (W) characteristics. Immortalized B-lymphoblastoid cells were gated on the FSC-A versus SSC-A dot plot. Furthermore, IgD+ cells, IgM+ cells, IgG+ cells, and IgA+ cells in immortalized B-lymphoblastoid cell lines were identified based on their expression levels of surface membrane immunoglobulins.

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.