nature biotechnology

Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

Peter A C 't Hoen[1,2], Marc R Friedländer[3–6,15], Jonas Almlöf[7,15], Michael Sammeth[3–5,8,14], Irina Pulyakhina[1], Seyed Yahya Anvar[1,9], Jeroen F J Laros[1,2,9], Henk P J Buermans[1,9], Olof Karlberg[7], Mathias Brännvall[7], The GEUVADIS Consortium[10], Johan T den Dunnen[1,2,9], Gert-Jan B van Ommen[1], Ivo G Gut[8], Roderic Guigó[3–5], Xavier Estivill[3–6], Ann-Christine Syvänen[7], Emmanouil T Dermitzakis[11–13] & Tuuli Lappalainen[11–13]

RNA sequencing is an increasingly popular technology for genome-wide analysis of transcript sequence and abundance. However, understanding of the sources of technical and interlaboratory variation is still limited. To address this, the GEUVADIS consortium sequenced mRNAs and small RNAs of lymphoblastoid cell lines of 465 individuals in seven sequencing centers, with a large number of replicates. The variation between laboratories appeared to be considerably smaller than the already limited biological variation. Laboratory effects were mainly seen in differences in insert size and GC content and could be adequately corrected for. In small-RNA sequencing, the microRNA (miRNA) content differed widely between samples owing to competitive sequencing of rRNA fragments. This did not affect relative quantification of miRNAs. We conclude that distributing RNA sequencing among different laboratories is feasible, given proper standardization and randomization procedures. We provide a set of quality measures and guidelines for assessing technical biases in RNA-seq data.

RNA sequencing (RNA-seq) has transformed the field of transcriptomics[1–4]. Whereas expression microarrays are limited to the detection of known transcripts and have limited capacity to differentiate between transcript variants, RNA-seq can in principle detect all coding and noncoding transcripts in the cell and determine their sequence and structure. Moreover, sequencing-based methods for expression profiling appear to be more accurate and more sensitive toward low-abundance transcripts[5–11], even if increased variability in the low-expression range has been reported[12,13]. Nevertheless, RNA-seq is not free of biases. Important biases are introduced by random hexamer priming[14], differences in fragment size and transcript length[15–17], and differences in GC content[18,19]. A systematic and large-scale analysis of the effects of such technical biases on mRNA and small-RNA (sRNA) quantification, as was performed by the MAQC consortium for expression microarrays[20–22], has not been reported yet for RNA-seq.

The GEUVADIS consortium (Genetic European Variation in Disease, a European Medical Sequencing Consortium) focuses on the standardization of next-generation sequencing technologies. The consortium initiated a large-scale RNA-seq analysis where data production was distributed across different laboratories. In this report, we evaluate the sources of technical variation in RNA-seq experiments and the feasibility and consequences of distributing sequencing. Moreover, we provide a set of essential quality measures for RNA-seq experiments and discuss possible strategies for correction of technical biases. The biological interpretation of the results is reported in a companion study[23].

RESULTS

Study design

A major objective of the current study was to evaluate the feasibility of sharing RNA-sequencing among different laboratories and subsequently combining the RNA-seq data obtained. To this end, we distributed randomly 465 RNA samples from lymphoblastoid cell lines from five populations to seven different European laboratories. Each center received 48 to 113 randomly assigned samples and strict sample preparation and sequencing guidelines (**Supplementary Note**). At these seven laboratories, the mRNA and sRNA fractions were prepared for sequencing using Illumina's TruSeq kits for RNA and small RNAs, respectively. Samples were sequenced with the Illumina HiSeq2000 platform, with paired-end, 75-bp reads for mRNA-seq and single-end 36-bp (50 bp in some laboratories) reads for sRNA-seq. Five RNA samples were prepared and sequenced at all sites to allow proper estimation of laboratory effects; 168 mRNA samples sequenced in other laboratories were prepared and sequenced twice in laboratory 1, with slightly lower numbers of reads in the repeated sequencing. The raw

[1]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [2]Netherlands Bioinformatics Centre, Leiden, The Netherlands. [3]Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain. [4]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain. [5]CRG Hospital del Mar Research Institute, Barcelona, Catalonia, Spain. [6]CIBER in Epidemiology and Public Health (CIBERESP), Barcelona, Catalonia, Spain. [7]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [8]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Catalonia, Spain. [9]Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands. [10]Full lists of members and affiliations appear at the end of the paper. [11]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [12]Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, Geneva, Switzerland. [13]Swiss Institute of Bioinformatics, Geneva, Switzerland. [14]Present address: Bioinformatics Laboratory, National Laboratory of Scientific Computing, Petropolis, Rio de Janeiro, Brazil. [15]These authors contributed equally to this work. Correspondence should be addressed to P.A.C.t.H. (p.a.c.hoen@lumc.nl) or T.L. (tuuli.e.lappalainen@gmail.com).

Received 4 January; accepted 21 August; published online 15 September 2013; doi:10.1038/nbt.2702

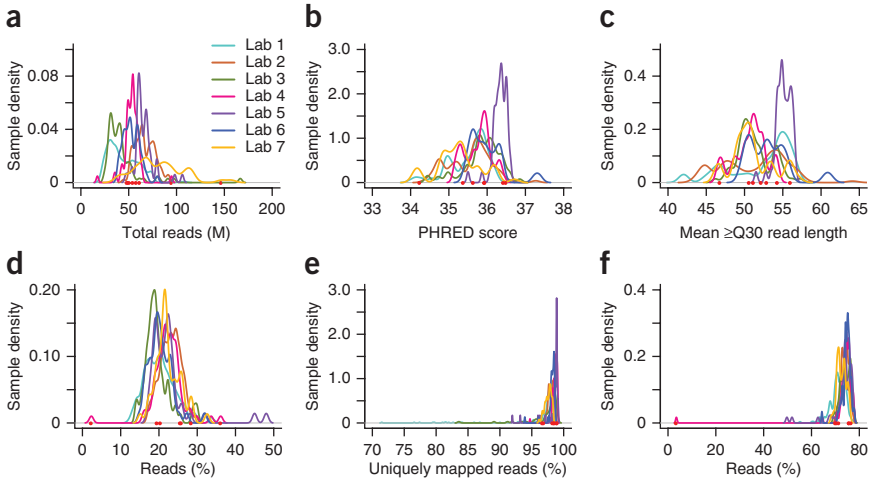

Figure 1 Basic quality statistics in mRNA sequencing across laboratories. Distribution of sequencing characteristics over 667 samples sequenced in seven different laboratories. For each feature, density plots were created to adjust for the differences in the number of samples processed by each laboratory. (**a**) Total number of reads obtained per sample. (**b**) Mean base quality (PHRED score) per sample. (**c**) Mean length of the longest continuous subsequence with quality over Q30. (**d**) Percentage of duplicate reads. (**e**) Percentage of mapped reads. (**f**) Percentage of aligned reads mapping to exons. The samples that did not pass quality control criteria in our study are shown as red dots.

data (FASTQ files) were subsequently aligned with GEM[24] (mRNA data) and miraligner[25] (sRNA data), and analyzed with a common pipeline that quantifies exon, transcript and sRNA expression levels (Online Methods).

Basic quality control steps in mRNA-seq

The laboratories were free to choose the number of samples to be pooled in one lane. Although the target number of reads was minimally 20 million (10 million paired reads), the laboratories generally decided to choose conservative pooling schemes, avoiding the need for repetition of samples whose coverage was too low. This resulted in a median of 58 million reads with a broad range of 17–167 million (**Fig. 1a**). The extent of the range was partly due to differences in the number of samples per lane and partly due to difficulties with equimolar pooling, which resulted in up to a threefold difference between the highest and lowest number of reads per sample in a lane.

We assessed the mRNA-seq data for basic quality measures, applied these metrics to compare the performance of the different laboratories (**Fig. 1**) and used several approaches to detect problematic samples (**Fig. 2**). A more extended list of all quality measures assessed is given in **Supplementary Tables 1** and **2**. All samples had similarly high mean PHRED scores, a measure for the quality of the base calling (**Fig. 1b**). The mean number of bases per read with a quality score >Q30 also indicated that the sequence quality was high (**Fig. 1c**), but this quality measure showed more variation between samples. Lower scores on this measure did not result in lower percentages of aligned reads. Sequence runs with >50% of the nucleotides having quality scores >Q30 are therefore acceptable. Duplication rates (assessed for single reads, not read pairs) centered around 20% for all laboratories (**Fig. 1d**). Higher duplication rates in RNA than in DNA sequencing are a result of genes with high expression levels. The percentage of aligned reads was generally very high except for a few samples (between 95–100%, **Fig. 1e**). Some of these outliers were associated with high duplication rates (**Fig. 1d**). Downstream analysis showed that lower mapping and higher duplication rates did not seem to affect the quantification of exons and transcripts, as the expression levels in these samples correlated strongly with all other samples (**Fig. 2b**,**c**). The percentage of aligned reads mapping to annotated exons were generally 60–80% (**Fig. 1f**). This is an important quality measure, as it collectively captures variation in enrichment for mature mRNAs, possible contamination and effectiveness of the alignment procedure. One sample (NA18861.4) had only 4% of aligned reads mapping to exons, while still having a high overall mapping rate. The extensive coverage in introns and intergenic regions of this sample suggested that it was contaminated with genomic DNA. This sample was excluded from the final set of samples used for biological interpretation. Two other samples (HG00099.5 and HG00329.5) had exonic content of only ~50% and were also characterized by high duplication and low mapping rates. These samples contained a large fraction (~20%) of rRNAs, presumably as a consequence of suboptimal polyA$^+$ RNA selection. Again, this did not affect the quantification of the exons and transcripts (**Fig. 2b**).

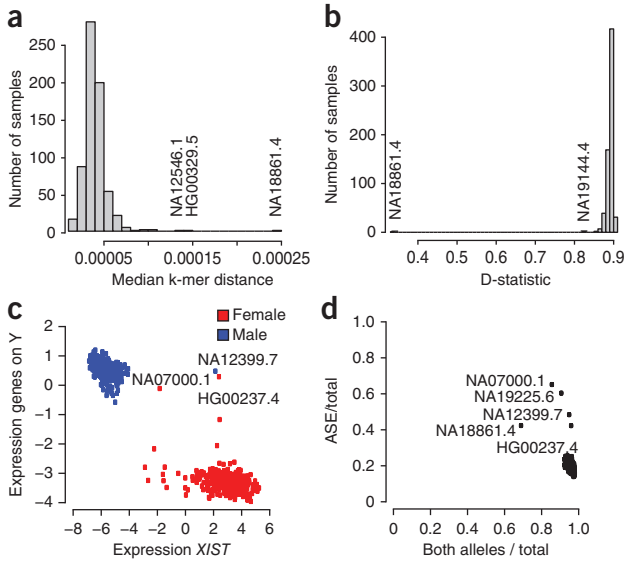

Figure 2 Detection of outliers in mRNA sequencing. (**a**) Histogram of median pairwise k-mer distances for each of the 667 samples with all other samples. (**b**) Histogram of median pairwise Pearson correlations (D-statistics) between exon expression levels after OPS transformation. (**c**) Gender-specific expression: normalized expression levels of female-specific *XIST* transcript (*x* axis) versus sum of the normalized expression levels of Y-chromosome transcripts excluding transcripts in the pseudo-autosomal regions (*y* axis). (**d**) Allele-specific expression analysis (ASE): for all heterozygous sites considered (Online Methods), the proportion of heterozygous single-nucleotide polymorphisms (SNPs) where both alleles were observed (*x* axis) was plotted against the proportion of heterozygous SNPs showing significant allelic bias in expression ($P < 0.05$, binomial test).

Detection of problematic samples

To detect whether problematic samples could be identified before alignment, we analyzed the distance between k-mer profiles. To this

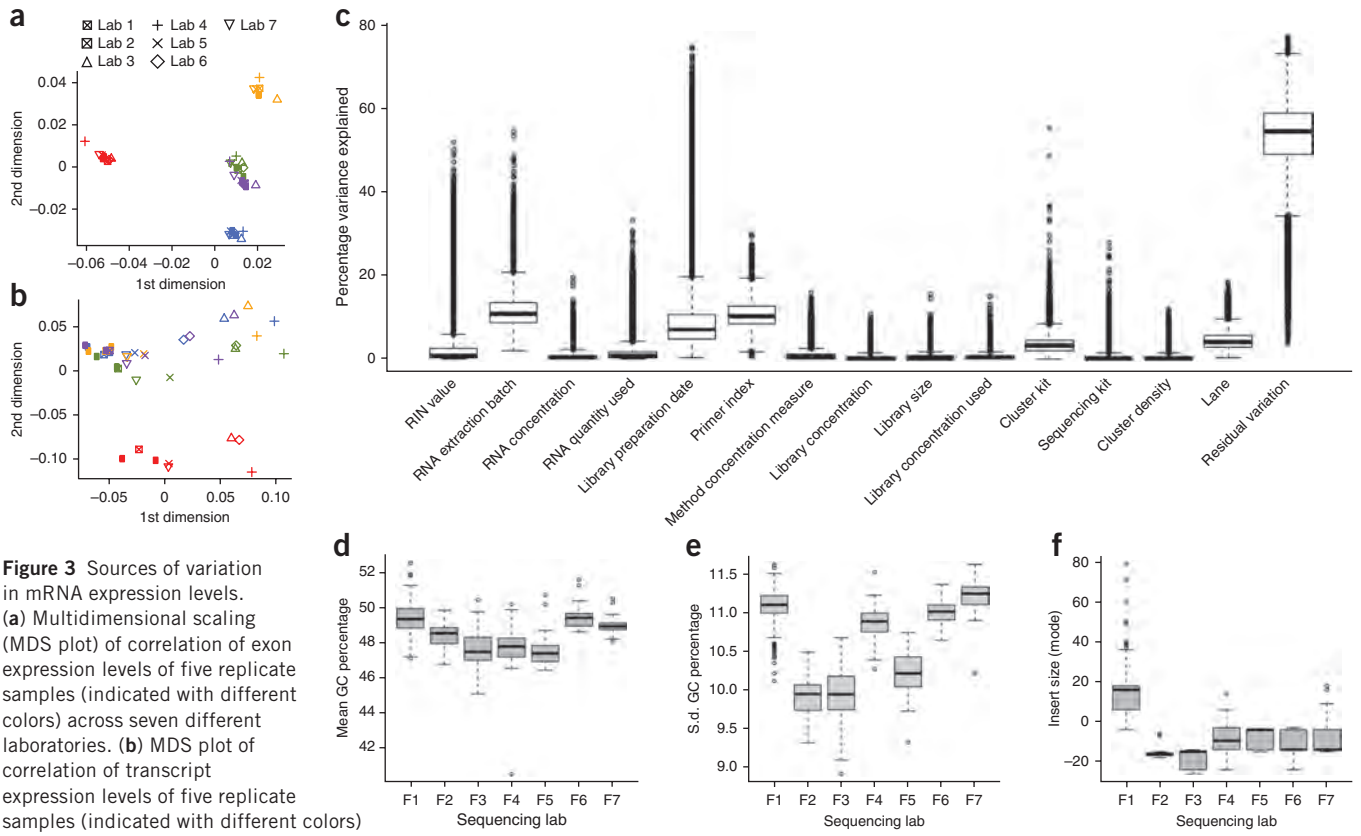

Figure 3 Sources of variation in mRNA expression levels. (**a**) Multidimensional scaling (MDS plot) of correlation of exon expression levels of five replicate samples (indicated with different colors) across seven different laboratories. (**b**) MDS plot of correlation of transcript expression levels of five replicate samples (indicated with different colors) across seven different laboratories. (**c**) Percentage variation in transcript expression levels explained by different sources: RIN (RNA integrity) value; RNA extraction batch; RNA concentration in initial sample; RNA quantity used for library preparation; library preparation date; indexing primer used; library concentration determination method (QBIT, Bioanalyzer or qPCR); library concentration obtained; mode of library size (as determined on Bioanalyzer, bp); library concentration used in sequencing; cluster kit; sequencing kit; cluster density (raw); lane of HiSeq2000 instrument in which the sample was run. Boxplots show distribution of the percentage of variance explained across all transcripts expressed in >50% of samples. Only samples with technical replicates in different laboratories passing quality control were included in this analysis ($n = 367$). (**d**) Boxplot of mean GC percentage in the reads across all 667 samples across different laboratories. (**e**) Boxplot of s.d. in GC percentage in the reads across all 667 samples across different laboratories. (**f**) Boxplot of the mode (highest frequency in the distribution) of the inferred insert size (i.e., the distance between the ends of the first and second reads; negative means overlapping sequences) of all 667 samples across different laboratories. (**c**–**f**) Boxes represent the 25–75 percentile, horizontal lines the median, whiskers the upper and lower quartile ± 1.5× the interquartile range.

end, we analyzed the abundance of all k-mers with length $k = 9$ and determined the pair-wise distance between the profiles of the different samples using a multiset distance measure[26]. The k-mer profile of NA18861.4 was clearly different from the rest (**Fig. 2a**). Some of the other samples with relatively high k-mer distances were samples with high duplication rates. The k-mer distances were strongly negatively correlated to the correlation measures obtained from the exon quantification of the samples (**Supplementary Fig. 1**), but some samples with high duplication rates and/or high rRNA content were identified only with k-mer profiling. Thus, k-mer profiling is a promising quality assessment procedure that does not require alignment to a reference genome.

After alignment, we used pair-wise correlation measures on the quantification of exons and transcripts to detect problematic samples. Given the skewness of RNA-seq data, where there are few highly expressed transcripts and many of low abundance, use of the Pearson correlation on the linear scale is not appropriate. Therefore, we first applied an optimal power space (OPS) transformation (P.R. & M.S., data not shown), ensuring that low- and high-abundance transcript outliers do not bias the correlation measure and all data points contribute equally to the computed coefficient (**Supplementary Fig. 2**). **Figure 2b** provides the distribution of the median Pearson correlations (D-statistics) of the quantification of exons for each sample. It was clear that NA18861.4 (with only 4% exonic reads) had lower correlations to the other samples. NA19144.4 was identified as an additional outlier and was removed from the analysis. Quantification of transcripts and genes identified the same outliers (**Supplementary Fig. 3**). In general, the correlations of gene expression levels were stronger than correlations of exon expression levels. Gene expression levels are more robust because of the higher number of reads in a complete gene compared to an individual exon. Transcript quantification correlated much less than gene or exon quantification due to inherent uncertainty in the deconvolution and relative quantification of transcripts from the same gene (**Supplementary Fig. 3**).

Sample mix-ups are a general problem in studies analyzing large cohorts of samples and may severely compromise their power[27]. As a first check for sample swaps, we determined male and female origin, based on the expression of the *XIST* gene (exclusively expressed in females) and Y-chromosome genes (exclusively expressed in males) (**Fig. 2c**). Clear sample swaps, where samples marked female expressed Y-chromosome genes without expression of *XIST* or vice versa, were not observed. However, in three samples, there was expression of both *XIST* and Y-chromosome genes, indicative of contamination between samples.

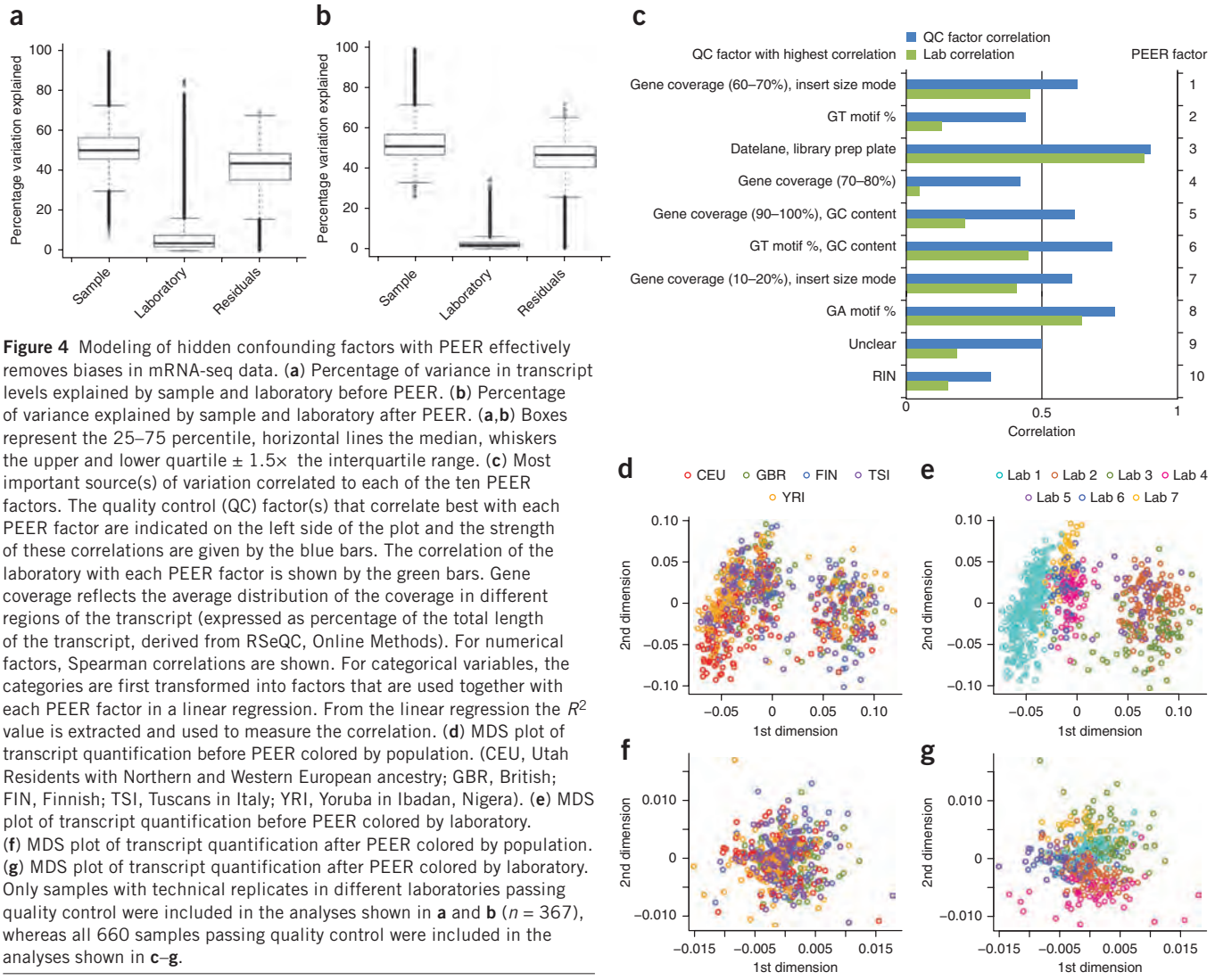

Figure 4 Modeling of hidden confounding factors with PEER effectively removes biases in mRNA-seq data. (**a**) Percentage of variance in transcript levels explained by sample and laboratory before PEER. (**b**) Percentage of variance explained by sample and laboratory after PEER. (**a**,**b**) Boxes represent the 25–75 percentile, horizontal lines the median, whiskers the upper and lower quartile ± 1.5× the interquartile range. (**c**) Most important source(s) of variation correlated to each of the ten PEER factors. The quality control (QC) factor(s) that correlate best with each PEER factor are indicated on the left side of the plot and the strength of these correlations are given by the blue bars. The correlation of the laboratory with each PEER factor is shown by the green bars. Gene coverage reflects the average distribution of the coverage in different regions of the transcript (expressed as percentage of the total length of the transcript, derived from RSeQC, Online Methods). For numerical factors, Spearman correlations are shown. For categorical variables, the categories are first transformed into factors that are used together with each PEER factor in a linear regression. From the linear regression the R^2 value is extracted and used to measure the correlation. (**d**) MDS plot of transcript quantification before PEER colored by population. (CEU, Utah Residents with Northern and Western European ancestry; GBR, British; FIN, Finnish; TSI, Tuscans in Italy; YRI, Yoruba in Ibadan, Nigera). (**e**) MDS plot of transcript quantification before PEER colored by laboratory. (**f**) MDS plot of transcript quantification after PEER colored by population. (**g**) MDS plot of transcript quantification after PEER colored by laboratory. Only samples with technical replicates in different laboratories passing quality control were included in the analyses shown in **a** and **b** ($n = 367$), whereas all 660 samples passing quality control were included in the analyses shown in **c**–**g**.

Identification of sample mix-ups in studies where DNA genotypes are available is relatively straightforward. For each individual, we evaluated the number of sites that were heterozygous in DNA genotypes for expression of both alleles in the mRNA-seq data. Owing to allele-specific expression, this is usually not 100% but is generally >90% of all heterozygous sites in expressed genes. In the case of sample mix-ups, where there is a mismatch between DNA genotypes and mRNA-seq data, this number would be considerably lower. No such samples were observed in our data set (**Fig. 2d**). We also analyzed the percentage of sites showing significant allele-specific expression, that is, imbalance between the expression of the two alleles ($P < 0.05$, binomial test). This measure is also sensitive to sample contamination, as the expected 50:50 allelic ratio over a heterozygous site in a given individual will be biased if even a small proportion of mRNA-seq reads is derived from another individual that may be homozygous for the site. The three suspected samples from the gender analysis were also found to be contaminated according to this analysis (**Fig. 2d**). In addition to these samples, another sample (NA19225.6) was identified by analysis of allele-specific expression as potentially being contaminated; the contamination probably originated from a sample with the same gender. The one problematic sample (NA18861.4), for which there was a low proportion of exonic counts, was also an outlier in this analysis.

Sources of variation in mRNA-seq

Variation in expression levels between samples originates from biological and technological sources. In this study, we were interested in quantifying the relative contribution of technical and biological variation to the total variation and in tracing the most important sources of technical variation. When comparing individual lymphoblastoid cell lines, the biological variation is limited, as the only biological difference is the individual's genetic and epigenetic background, whereas the cell type and growth conditions are the same. Nevertheless, the five samples that were sequenced in each laboratory clustered by sample and not by laboratory (**Fig. 3a** and **Supplementary Fig. 4**). The correlations between replicate samples run in the same laboratory were slightly higher than between samples run in different laboratories (on average 0.931 versus 0.925 for exon quantification). The clustering by sample was much stronger for exon quantification than for transcript quantification (compare **Fig. 3a** with **Fig. 3b** and **Supplementary Fig. 4a** with **Fig. 4b** and **Supplementary Fig. 5**).

Given the stronger effect of technical variation on transcript quantification than on exon quantification, we further investigated the sources of technical variation contributing to this variation. The RNA extraction batch was the strongest contributor to the observed technical variation (**Fig. 3c**). Slight inter-day differences between library

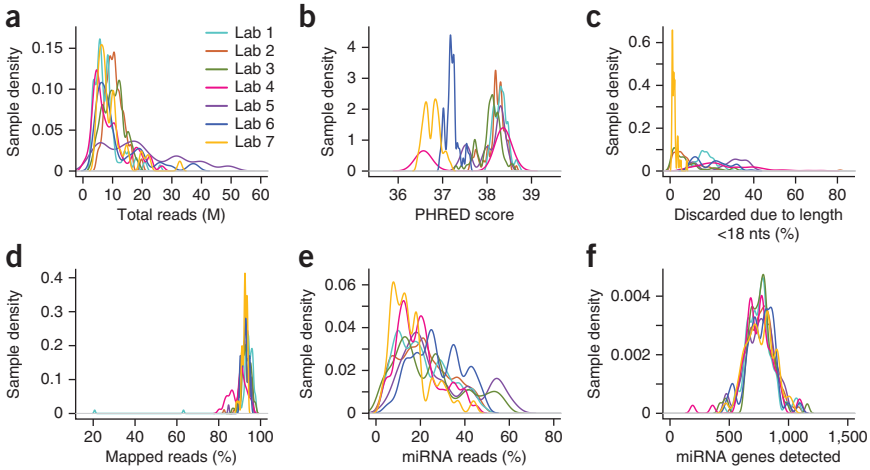

Figure 5 Basic quality statistics in sRNA sequencing across laboratories. Density plots for 492 samples sequenced in seven different laboratories. (**a**) Total number of reads obtained. (**b**) Mean base quality (PHRED score). (**c**) Percentage of reads discarded due to short length. (**d**) Percentage of mapped reads. (**e**) Percentage of mapped reads in miRNA genes. (**f**) Number of miRNA genes detected.

preparations and the effects of the different index primers were also notable (**Fig. 3c**), but these are partially confounded with the different laboratories in which the samples were processed.

Thus, despite the use of the same library preparation kits (and versions of these) and the availability of standardized protocols, slight differences in library preparations between laboratories were observed. Most notably, these were manifested in differences between the average GC percentage, the width of the distribution of GC percentages and the insert sizes (**Fig. 3d–f**). The exons with high GC content (>65%) demonstrated more variable expression levels between laboratories than exons with medium or low (<35%) GC content (**Supplementary Fig. 6a**). Relatively low representation of sequences with >65% GC content may be explained by the use of thermocyclers with high ramping speeds (**Supplementary Fig. 6b**)[28].

Although all laboratories aimed for an insert size of ~10 bp (corresponding to a fragment size of 280 bp: 10 + 2 × 75 (read length) + 120 bp (length of the adapters)), most laboratories produced slightly smaller inserts, resulting in partial overlap between the forward and the reverse read. The inferred insert size after alignment correlated well with the experimentally determined insert size (**Supplementary Fig. 7**). Differences in insert size will affect the potential to discriminate between transcript variants and consequently their relative quantification. This likely explains the stronger laboratory effects on transcript compared to exon quantification.

Other differences between laboratories included the concentration of the library obtained after sample preparation, the raw cluster density and the percentage of rRNA (**Supplementary Fig. 8**), but these did not seem to influence expression-level quantification.

Correction for variation in mRNA-seq

Next, we explored the correction of technical sources of variation. We used a recently described Bayesian framework that accounts for hidden variables in expression data (PEER[29,30]). PEER takes quantification of genes or other expression units, such as transcripts, and uses factor analysis–based methods to infer factors that explain transcriptome-wide variance components. Removing these factors from the data by regression has been shown to improve *cis*–expression quantitative trait loci (eQTL) discovery[29,30]. Before correction, the variable 'laboratory' explained 6.8% of the total variance (average across transcripts, median: 3.8%) (**Fig. 4a**). In the residuals obtained after regression with the first ten PEER factors, the laboratory effect was reduced to 2.6% (average across transcripts; median 2.0%) (**Fig. 4b**). Laboratory effects were mainly captured by PEER factors 1, 3, 6, 7 and 8 (**Fig. 4c** and **Supplementary Fig. 9b**). Moreover, the PEER factors were correlated with several of the other observed sources of technical variation: insert size (factor 1, 7), and GC content and other nucleotide biases (factor 2, 5, 6 and 8) (**Fig. 4c** and **Supplementary Table 3**). Finally, factors 1, 4, 5 and 7 were correlated with differences in the coverage in different regions of the transcript (**Fig. 4c** and **Supplementary Table 3**). This effect was not related to RNA integrity, which was captured mainly by factor 10, and may reflect biases introduced in the reverse transcription step.

After PEER correction of the transcript expression levels, clustering of samples by laboratory was less pronounced (**Fig. 4d–g**). Moreover, the number of detected eQTLs as well as the overlap with microarray-derived eQTLs increased (**Supplementary Fig. 10**). Thus, technical variation, and in particular variation that is introduced by having sequencing done in several different laboratories, can be properly accounted for and has only limited influence on quantification of exons or transcripts in such a sequencing setting.

Basic quality control steps in sRNA-seq

We analyzed 492 samples by sRNA-seq, aiming for 3–6 million mapped reads. The obtained sequencing depth varied considerably, from 0.1–50 million reads per sample, with a median of 8.6 million reads (**Fig. 5a**). The sequencing quality was uniformly high, with sample mean PHRED score in a narrow band from around 36–39 (**Fig. 5b** and **Supplementary Table 4**). After adaptor trimming and before mapping, we discarded all sequences shorter than 18 nts, because such short sequences cannot be traced to genomic loci with high confidence. The fraction of reads thus discarded differed between samples, ranging from 0.5% to 81%, the percentage strongly dependent on which laboratory did the sequencing (**Fig. 5c**). This wide range may be caused by slight differences in gel separation and purification, which were performed in the individual laboratories, or by variable degradation during library preparation.

Because many sRNAs are repetitive, we mapped the reads to the human genome build allowing for multiple mappings (Online Methods). The mapping efficiencies were uniformly high, consistent with the high sequencing quality (**Fig. 5d**). Notably, the relative miRNA content in our samples ranged from 2% to 62% of mapped reads, with a median of 19% (**Fig. 5e**). Given that some sRNA-seq studies report miRNA contents >90% (e.g., ref. 31), these numbers are overall low for reasons discussed in the next paragraph. Despite differences in sequencing depth, fraction of short sequences and miRNA content, between 500 and 900 miRNA genes were robustly detected in all samples (**Fig. 5f**). Moreover, the same miRNA genes were consistently profiled. The 500 most highly expressed miRNAs were detected, on average, in >96% of the samples.

Proportional differences do not affect quantification of miRNAs

We found that mapped sRNAs do not just originate from miRNA genes, but also from other noncoding RNA genes, in particular

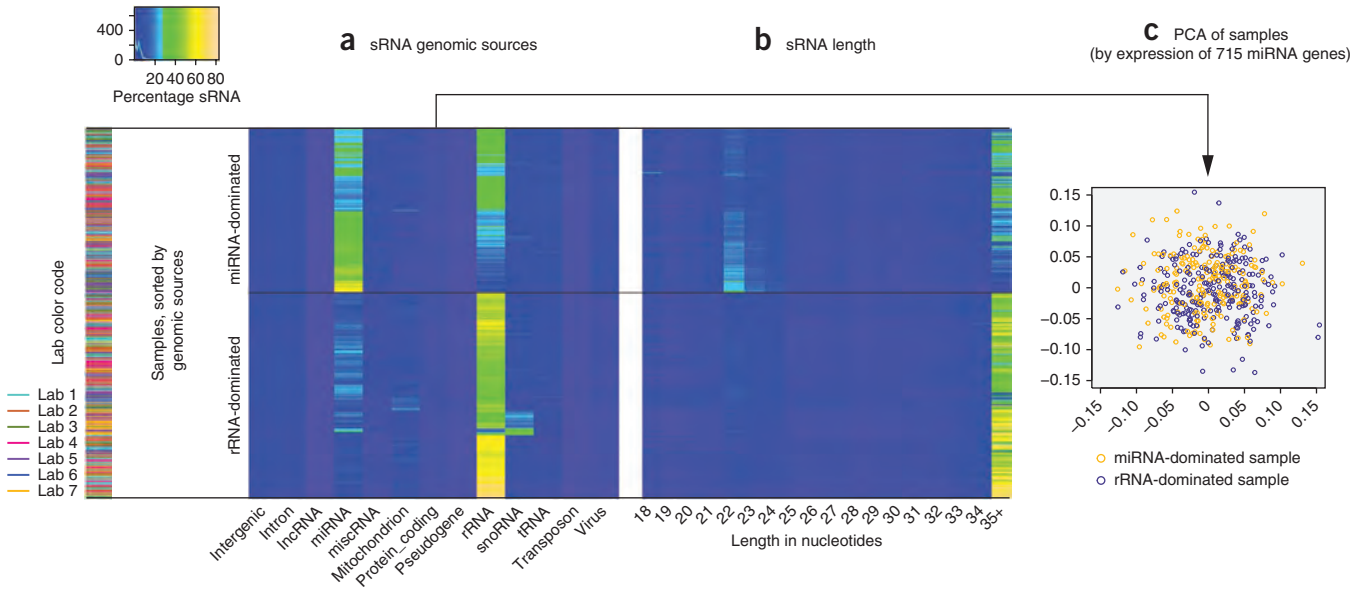

Figure 6 sRNA heterogeneity does not disturb quantification of individual miRNAs. (**a**) Heatmap of 492 sRNA samples (rows) clustered by expression of 13 types of sRNA sources (columns). The individual sources constitute from 0% (dark blue) to 82% (light orange) of total sRNA in each sample. Samples are divided into miRNA-dominated (above horizontal line) and rRNA-dominated (below horizontal line). The lab color bar to the left indicates the sequencing laboratory. (**b**) Heatmap as in **a**, but now the length of sequenced RNAs (after adaptor trimming) is shown. The same clustering is used, so samples are horizontally aligned across subfigures. (**c**) sRNA samples grouped by PCA. miRNA-dominated samples are shown in orange and rRNA-dominated samples are shown in blue. The samples do not group (are not biased) by the relative contents of miRNA and rRNA.

rRNA (**Fig. 6a**). Clustering divided the samples into those dominated by miRNA and those dominated by rRNA. The two groups were not associated with particular laboratories (**Fig. 6a**, lab color bar, left). Moreover, the replicates sequenced in all laboratories fell into groups mostly by sample and not by laboratory (**Supplementary Fig. 11**), and the miRNA and rRNA contents were more similar within samples than within laboratories (**Supplementary Fig. 12**). Importantly, the miRNA contents clearly varied between RNA extraction batches (**Supplementary Fig. 13**). Likewise, small nucleolar RNA (snoRNA) and other sRNA proportions clearly varied between samples (**Fig. 6a**). In conclusion, differences in the proportions of the different sRNAs are likely introduced during RNA extraction, before the samples were distributed across the laboratories.

Consistent with the mode of biogenesis, the reads originating from miRNA genes were typically 22 nucleotides long after adaptor clipping (**Fig. 6b**). In contrast, the reads that originated from rRNAs were 35 nucleotides long. As ~70% of these reads were mapping to the 5-kb 28S rRNA, it is likely that these reads represent rRNA fragments. To test if the heterogeneity in the contents of sRNAs biased the quantification of individual miRNAs, we calculated the expression levels of 715 miRNA genes based on their read counts. The samples did not cluster according to miRNA or rRNA content (**Fig. 6c**).

In a similar procedure as for the mRNA-seq, we calculated D-statistics for the correlation between normalized expression levels across samples, and we excluded four samples from the biological analysis that had D-statistics >0.8 (**Supplementary Fig. 14**). Again similar to mRNA-seq, we corrected miRNA expression levels by PEER and observed that GC percentage was the biggest source of variation needing correction, and that the GC percentage was correlated to the laboratory (**Supplementary Fig. 15** and **Supplementary Table 5**).

DISCUSSION

In this paper, we have demonstrated that technical variation in RNA-seq experiments is small and that results from RNA-seq experiments performed in different laboratories are consistent. This conclusion is valid as long as all participating laboratories use the exact same protocols (**Supplementary Note**) and versions of sample preparation and sequencing kits. However, even when using identical protocols, slight variations in average GC content and insert size were observed. These differences translated into variations in quantification of transcripts, whereas quantification of exons was less affected. Under less-standardized sequencing protocols, greater variation is expected. Moreover, RNA isolation and purification procedures, here performed in the same laboratory with standardized protocols, may contribute to variation in RNA-seq data.

The sources of variation contributing to differences between laboratories generally also play a role in smaller-scale experiments run in the same facility. This is, for example, true for differences in GC content, where considerable intralaboratory variation was also observed (**Fig. 3**). Based on the current study, we propose several parameters that should be assessed in any mRNA-seq data set to address the quality of the samples and/or explore the need for correction of important biases (**Table 1**). (i) The distribution of nucleotide-level quality scores

Table 1 Important quality checks in mRNA and sRNA sequencing

Quality checks common to mRNA and sRNA sequencing
Distribution of base quality scores
Average and width of the distribution of GC content
Percentage of reads mapping to the genome
Checks for sample swaps and contaminations
Outlier detection: pairwise correlations in expression quantification between samples
Quality checks specific for mRNA
The average and s.d. of insert size
Percentage of reads mapping to annotated exons
5′–3′ trends in coverage across transcripts
Quality checks specific for sRNA
Length distribution after adaptor clipping
Percentage of reads mapping to known sRNA genes

is the most basic quality measure, already implemented in nearly all sequencing centers, to address the quality of the sequencing run. (ii) The average and distribution of GC content as well as differences in proportion of reads with extreme (<35% or >65%) GC content induces biases in transcript quantification, which can be partially corrected with dedicated tools[18,32]. (iii) The average and s.d. of insert size influence mostly transcript deconvolution and quantification. A dedicated routine for correcting this bias has not been described so far. (iv) The percentage of reads mapping to annotated exons checks for genomic DNA contamination, the proportion of mature RNA in the total RNA pool, and the performance of the alignment procedure. The cut-off for this parameter depends on the sample preparation protocol and the aligner and annotation source used. As a rule of thumb derived from this and other mRNA-seq experiments, at least 60% of mapped reads should overlap with annotated exons in a good quality sample. (v) A decrease in coverage from the 3′ to the 5′ end of the transcript, for example, assessed using the *geneBody_coverage* module from the RSeQC 2.0.0 (ref. 33) package, is suggestive of RNA degradation when sequencing polyA$^+$ RNA. (vi) Sample swaps, contamination and outliers should be checked for. Procedures described in this and other papers[34,35] may be used when gender and/or genotype data are available. Alternatively, appropriate barcoding schemes are helpful to detect these artifacts.

We successfully applied the PEER algorithm to account for technical factors and to reduce their impact on expression level estimates. As for alternative methods using surrogate variables[36] or principal components[37], it is not necessary to know the sources of variation beforehand. However, this type of routine can only be used in relatively large studies, and potential removal of biological variation alongside the technical variation should be examined. For smaller studies, dedicated algorithms and standard regression methods may be applied to correct for known technical biases[14,17,18,32]. Still, minimizing technical variation by careful standardization of protocols and randomization in every experimental step (cell line handling, RNA extraction, sample preparation and sequencing) is essential.

This study focused on the quantification of both mRNA and sRNA. Although sRNA sample preparation is generally regarded as more challenging than its mRNA counterpart, technical variation introduced in the sample preparation seems limited compared to differences originating from the RNA isolation procedure. An important evaluation parameter in sRNA sequencing is the read-length distribution after adaptor clipping (**Table 1**), which is correlated with the miRNA content of the sample. We found that seemingly spurious abundances of rRNA fragments were competing with miRNAs for sequencing, in some cases resulting in low miRNA read counts. However, when we considered only miRNA reads and normalized using these counts, we found that the rRNA abundances had no major confounding impact on miRNA quantification. This indicates that sRNA-seq data should not be analyzed as a whole, but split into different sRNA fractions before normalization. Further, differences in effective sequencing depth did not strongly influence the number of miRNA genes detected. Thus, despite the heterogeneity in sRNA data, simple precautions can ensure robust results.

Our mRNA-seq study and the microarray-based MAQC studies[20] both addressed the technical variation introduced by analyzing samples in different laboratories. For both technologies, it was concluded that the intersite variability is limited when working with standardized protocols. However, it is difficult to compare the interlaboratory reproducibility of mRNA-seq in this study with the interlaboratory reproducibility of gene expression microarrays in the MAQC study, given differences in experimental design, differences in the scales on which mRNA-seq and microarray data are reported, and the much higher dynamic ranges of mRNA-seq counts compared to microarray intensities. For example, the biological variation in our experiment was orders of magnitude smaller than the differences between the tissues studied by the MAQC consortium. Where MAQC extensively validated results by confirming genes differentially expressed between a small number of tissues by quantitative PCR, we have proven the validity of our measurements by showing high power for detection of a large set of *cis*-eQTLs in a large set of 465 independent samples[23]. The small effect sizes detected with the majority of eQTLs, confirms conclusions from earlier papers that mRNA-seq technology is at least as robust as microarray technology[5–11,19,38].

In conclusion, distributing RNA-sequencing among different laboratories appears to be feasible. It is particularly attractive for large population-based and cross-biobank studies, where sequencing costs and sample logistics may require the combination of data from individual studies and laboratories.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The raw FASTQ files and BAM alignments as well as different types of quantification are available in ArrayExpress under accessions E-GEUV-1 (mRNA) and E-GEUV-2 (small RNA) for QC-passed samples and E-GEUV-3 for all sequenced samples.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This project was funded by the European Commission 7th Framework Program (FP7) (261123; GEUVADIS); the Swiss National Science Foundation (130326, 130342), the Louis Jeantet Foundation, and ERC (260927) (E.T.D.); NIH-NIMH (MH090941) (E.T.D., R.G.); Spanish Plan Nacional SAF2008–00357 (NOVADIS), the Generalitat de Catalunya AGAUR 2009 SGR-1502, and the Instituto de Salud Carlos III (FIS/FEDER PI11/00733) (X.E.); Spanish Plan Nacional (BIO2011-26205) and ERC (294653) (R.G.); ESGI, READNA (FP7 Health-F4-2008-201418), Spanish Ministry of Economy and Competitiveness (MINECO) and the Generalitat de Catalunya (I.G.G.); FP7/2007-2013, ENGAGE project, HEALTH-F4-2007-201413, and the Centre for Medical Systems Biology within the framework of The Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) (P.A.C.t.H. & G.-J.B.v.O.); The Swedish Research Council (C0524801, A028001) and the Knut and Alice Wallenberg Foundation (2011.0073) (A.-C.S.); EMBO long-term fellowship ALTF 225-2011 (M.R.F.); Emil Aaltonen Foundation and Academy of Finland fellowships (T.L.). We acknowledge the SNP&SEQ Technology Platform in Uppsala for sequencing, and the Swedish National Infrastructure for Computing (SNIC-UPPMAX) for compute resources for data analysis. The authors would like to thank P.G.M. van Overveld for help with preparation of the figures.

AUTHOR CONTRIBUTIONS

P.A.C.t.H., M.R.F., J.A., M.S., I.P., S.Y.A., J.F.J.L., H.P.J.B., M.B., O.K. and T.L. performed the analyses. P.A.C.t.H., A.-C.S., R.G., X.E., J.T.d.D., G.-J.B.v.O., I.G.G. and E.T.D. designed the study. E.T.D. and T.L. coordinated the study. P.A.C.t.H. drafted the manuscript, which was subsequently revised by all co-authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
2. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

3. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
4. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
5. 't Hoen, P.A. *et al.* Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.* **36**, e141 (2008).
6. van Iterson, M. *et al.* Relative power and sample size analysis on gene expression profiling data. *BMC Genomics* **10**, 439 (2009).
7. Sirbu, A., Kerr, G., Crane, M. & Ruskin, H.J. RNA-seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. *PLoS ONE* **7**, e50986 (2012).
8. Bradford, J.R. *et al.* A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* **11**, 282 (2010).
9. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
10. Agarwal, A. *et al.* Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**, 383 (2010).
11. Bottomly, D. *et al.* Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS ONE* **6**, e17820 (2011).
12. Raghavachari, N. *et al.* A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med. Genomics* **5**, 28 (2012).
13. Liu, S., Lin, L., Jiang, P., Wang, D. & Xing, Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res.* **39**, 578–588 (2011).
14. Hansen, K.D., Brenner, S.E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131 (2010).
15. Gao, L., Fang, Z., Zhang, K., Zhi, D. & Cui, X. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics* **27**, 662–669 (2011).
16. Oshlack, A. & Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4**, 14 (2009).
17. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).
18. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).
19. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
20. 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).
21. Canales, R.D. *et al.* Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24**, 1115–1122 (2006).
22. Patterson, T.A. *et al.* Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat. Biotechnol.* **24**, 1140–1150 (2006).
23. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* (in the press) doi:10.1038/nature12531 (2013).
24. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
25. Pantano, L., Estivill, X. & Marti, E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.* **38**, e34 (2010).
26. Kosters, W.A. & Laros, J.F.J. Metrics for mining multisets. in *Research and Development in Intelligent Systems XXIV, Proceedings of AI-2007* (Eds. Bramer, M., Coenen, F. & Petridis, M.) 293–303 (Springer, 2007).
27. Gordon, D. & Finch, S.J. Consequences of error. in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (Eds. Jorde, L., Little, P., Dunn, M. & Subramaniam, S.) (Wiley Online Library, 2006).
28. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
29. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLOS Comput. Biol.* **6**, e1000770 (2010).
30. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
31. Parts, L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genet.* **8**, e1002704 (2012).
32. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
33. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
34. Huang, J., Chen, J., Lathrop, M. & Liang, L. A tool for RNA sequencing sample identity check. *Bioinformatics* 1463–1464 (2013).
35. Westra, H.J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104–2111 (2011).
36. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161 (2007).
37. Fehrmann, R.S. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
38. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).

The GEUVADIS Consortium: Gert-Jan B van Ommen[1], Xavier Estivill[3–6], Roderic Guigó[3–5], Ann-Christine Syvänen[7], Ivo G Gut[8], Emmanouil T Dermitzakis[11–13], Stylianos E Antonorakis[11–13], Alvis Brazma[16], Paul Flicek[16], Stefan Schreiber[17], Philip Rosenstiel[17], Thomas Meitinger[18], Tim M Strom[18], Hans Lehrach[19], Ralf Sudbrak[19], Angel Carracedo[20], Peter A C 't Hoen[1,2], Irina Pulyakhina[1], Seyed Yahya Anvar[1,9], Jeroen F J Laros[1,2,9], Henk P J Buermans[1,9], Maarten van Iterson[1], Marc R Friedländer[3–6], Jean Monlong[3–6], Esther Lizano[3–6], Gabrielle Bertier[3,4], Pedro G Ferreira[3–5], Michael Sammeth[3–5,8], Jonas Almlöf[7], Olof Karlberg[7], Mathias Bränvall[7], Paolo Ribeca[8], Thasso Griebel[8], Sergi Beltran[8], Marta Gut[8], Katja Kahlem[8], Tuuli Lappalainen[11–13], Thomas Giger[11–13], Halit Ongen[11–13], Ismael Padioleau[11–13], Helena Kilpinen[11–13], Mar Gonzàlez-Porta[16], Natalja Kurbatova[16], Andrew Tikhonov[16], Liliana Greger[16], Matthias Barann[17], Daniela Esser[17], Robert Häsler[17], Thomas Wieland[18], Thomas Schwarzmayr[18], Marc Sultan[19] & Vyacheslav Amstislavskiy[19]

[16]European Bioinformatics Institute, Hinxton, United Kingdom. [17]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany. [18]Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany. [19]Max Plank Institute for Molecular Genetics, Berlin, Germany. [20]Fundacion Publica Galega de Medicina Xenomica SERGAS, Genomic Medicine Group CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

ONLINE METHODS

Samples and sequencing. Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines directly from Coriell Cell Repositories (GBR, FIN, TSI) or originally from Coriell but grown at the University of Geneva (CEU, YRI) were shipped to ECACC (European Collection of Cell Cultures) as live cultures, in batches of ~30 samples from Coriell and 2 × ~90 samples from Geneva. In ECACC, these cell lines were cultured to ~1.2×10^8 cells. These cultures were split to produce 8× cell banks of the samples, and a snap-frozen pellet of 2×10^7 cells from a proliferating culture. The cell pellets were shipped from ECACC to University of Geneva in three batches, the first batch consisting of CEU/GBR/FIN/TSI samples, and the second and third batch with YRI and the rest of CEU samples.

RNA was extracted in Geneva about 14 samples at a time. First, two-thirds of the first shipping batch was extracted with full randomization. Then, RNA was extracted from the second batch and the remaining one-third of the first batch with full randomization. Finally the third batch was extracted, again with full randomization. Total RNA was extracted from cell pellets using the TRIzol Reagent (Ambion). The pellets had been frozen at ECACC without any additives like RNAlater or TRIzol. In Geneva they were thawed, 1 ml of TRIzol was added in each sample, and the samples were transferred to Eppendorf tubes. The rest of the protocol followed the manufacturer's guidelines. RNA samples were not treated with DNase. RNA quality was assessed by Agilent Bioanalyzer RNA 6000 Nano Kit according to the manufacturer's instructions. RNA quantity was measured by Qubit 2.0 (Invitrogen) using the RNA Broad range kit according to the manufacturer's instructions.

Each of the sequencing laboratories were sent a minimum of 4 µg of total RNA of the samples allocated to them, and RNA Bioanalyzer was run for 10–20% of the RNA samples before library preparation to confirm sample quality after shipping. No further purification was done to the RNA samples other than that specified in the sequencing protocols. Library preps were done in random order in every laboratory.

mRNA sequencing was done on the Illumina HiSeq2000 platform with 75 bp paired-end sequencing with fragment size of ~280 bp—some laboratories sequenced 100-bp reads, which were trimmed to 75 bp. TruSeq RNA Sample Prep Kit v2 (the high-throughput protocol) was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired minimum of 10 M mapped and properly paired read pairs from any standard mapper, without filtering for mapping quality.

Small RNA sequencing was done on the Illumina HiSeq2000 platform with 36 bp single-end sequencing with fragment size of 145–160 bp. Some laboratories sequenced 50-bp reads which were trimmed to 36 bp. TruSeq SmRNA Sample Prep kit was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired minimum of 3 M total reads.

Extensive information of sample processing was collected from all the laboratories for both mRNA-seq and sRNA-seq in order to enable control of batch effects.

Raw data processing. Each lab submitted one demultiplexed FASTQ file per sample per mRNA and miRNAseq, produced by CASAVA 1.8 or 1.8.2, allowing one mismatch in the index. Reads failing Illumina quality filtering were removed. The FASTQ files are named as: SAMPLE_ID.SeqLabNumber. M/MI_YYMMDD_Lane_Read.fastq.gz, where M/MI stands for mRNA or miRNA sequencing, and YYMMDD is the sequencing date. All the data were submitted and initially stored in the project ftp site. Samtools was used for general data processing throughout the project.

mRNA analysis pipeline. We employed the JIP pipeline (T.G. & M.S., data not shown) to map mRNA-seq reads and to quantify mRNA transcripts. For alignment to the human reference genome sequence (GRCh37, autosomes + X + Y + M), we used the GEM mapping suite[24] (v1.349 which corresponds to publicly available pre-release 2) to first map (max. mismatches = 4%, max. edit distance = 20%, min. decoded strata = 2 and strata after best = 1) and subsequently to split-map (max.mismatches = 4%, Gencode v12 and *de novo* junctions) all reads that did not map entirely. Both mapping steps are repeated for reads trimmed 20 nucleotides from their 3′-end, and then for reads trimmed 5 nucleotides from their 5′-end in addition to earlier 3′-trimming—each time considering exclusively reads that have not been mapped in earlier iterations. Finally, all read mappings were assessed with respect to the mate pair information: valid mapping pairs are formed up to a maximum insert size of 100,000 bp, extension trigger = 0.999 and minimum decoded strata = 1. The mapping pipeline and settings are described below and can also be found in https://github.com/gemtools, where the code as well as an example pipeline are hosted.

The GEM output format was converted to BAM format, with following mapping quality scores and flags:

(1) Matches which are unique, and do not have any subdominant match: 251 ≥ MAPQ ≥ 255, XT = U
(2) Matches which are unique, and have subdominant matches but a different score: 175 ≥ MAPQ ≥ 181, XT = U
(3) Matches which are putatively unique (not unique, but distinguishable by score): 119 ≥ MAPQ ≥127, XT = U
(4) Matches which are a perfect tie: 78 ≥ MAPQ ≥ 90, XT = R.

Furthermore, the NM flag contains the number of total mismatches (read1+read2). In analysis, we used reads in categories 1 and 2 and with NM ≤ 6. The settings employed ensured that for every read at least one stratum more than the optimal mapping was assessed, to distinguish bona fide alignments of bad quality reads from mapping noise. Split mappings were detected based on the Gencode v12 annotation and additionally discovered *de novo*. Read mappings were paired and converted to BAM files, employing a scoring scheme over mismatches, quality values and uniqueness in the case of multi-maps.

Exon quantification was calculated after merging overlapping exons into meta-exons. Read counts over these meta-exons were calculated by summing the number of reads with overlapping start or end coordinates. For split reads, we counted the exon overlap of each split fragment, and added counts per read as 1/(number of overlapping exons per gene).

Flux Capacitor[39] was used for quantification of transcripts. Quantification is based on the annotation-mapped genomic mappings considering transcript structures of the Gencode transcriptome annotation: mappings of read pairs that were completely included within the annotated exon boundaries and paired in the expected orientation have been taken into account. Reads belonging to single transcripts were predicted by deconvolution according to observations of paired reads mapping across all exonic segments of a locus. Gene quantification was calculated as the sum of all transcript RPKMs (reads per kilobase per million) per gene.

sRNA analysis pipeline. Data sets with read lengths longer than 36 nts were trimmed using the FASTX suite (http://hannonlab.cshl.edu/fastx_toolkit/) and homo-polymer reads and reads with low PHRED scores were removed. Adapters were clipped using the seqBuster suite[25] and custom searches. Reads shorter than 18 nts were discarded. The remaining reads were mapped to the human genome (hg19) using bowtie and annotated with GENCODE 8, supplemented with rRNA and LINE and Alu transposon annotations from RepBase[40] and snoRNA and miRNA annotations from the UCSC table browser[41]. Annotations were first resolved so that each nucleotide on each strand had exactly one annotation. In case of nucleotides with more than one annotation, conflicts were resolved using a confidence-based floating hierarchy[42]. Each read mapping was weighted inversely to the number of genome mappings for the read, for example, a read mapping to two genomic locations would get an assigned weight of 0.5. Each mapping was counted toward the annotation of the nucleotide in the middle of the mapping. miRNA quantification was performed with the custom tool miraligner[25].

Quality control measures. A comprehensive set of quality control statistics was obtained with a combination of existing software and in-house scripts. The following programs were run on each sample whereupon relevant information in the output was extracted and collected to one quality control master file:

FastQC 0.7.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

RSeQC 2.0.0 (ref. 33) (Modules used: geneBody_coverage (using refseq release 52), bam_stat, clipping_profile, read_distribution (using refseq release 52), read_duplication, read_GC, read_NVC)

PICARD 1.59 (http://sourceforge.net/projects/picard/; Modules used: EstimateLibraryComplexity, and MarkDuplicates).

All programs were run with the default parameters except the MarkDuplicates module in Picard that needed a regular expression for read name recognition adjusted to the current data.

Additional quality control data were obtained from in-house scripts used by the Uppsala University SNP&SEQ Technology Platform and University of Geneva. To calculate the average distribution of the coverage in different regions of the transcript, we used the output from RSeQC reflecting the total number of reads that map to a position of a transcript, after scaling all transcript positions to length 100. The positions 1–100 were binned again in 10% bins and then expressed as a percentage by dividing the number of reads in each bin by the total number of mapped reads for that sample. This resulted in the Gene_coverage_perc_X columns in **Supplementary Table 2**, **Figure 4c** and **Supplementary Figure 9**. The mode of the insert size (defined as the distance between the ends of the first and second reads) is calculated from all properly paired mapped reads of chromosome 1 in the bam file. The mode is the insert size that has the highest frequency. Further details on the parameters analyzed can be found in **Supplementary Table 1**.

K-mer profiling. We counted the abundance of all k-mers ($k = 9$) within the raw sequence reads by custom python scripts (S.Y.A. *et al.*, data not shown). Subsequently, the pairwise distance between the profiles of the different samples was calculated using the multiset distance measure[26]. This metric is parametrized by a function that reflects the distance between two elements in a multiset, in this case the difference in k-mer counts for one specific k-mer. We chose the following function:

$$f(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$

To correct for differences in total number of reads, we scaled the profiles before each pairwise comparison. The scaling procedure first calculates the total amount of k-mers in both profiles and then uses the ratio to scale the values to the smallest profile.

OPS transformation and correlation measures. Because gene expression follows a power law distribution (P.R. and M.S., data not shown), it is intuitive to use a suitable exponent in order to transform data to a more normal distribution, minimizing the impact of outliers. The OPS package (http://cran.r-project.org/web/packages/ops/) dynamically optimizes the normalization power according to the distribution of data points. Supporting the general agreement of data sets produced by different laboratories, we found consistently an OPS exponent of 0.11 for all sample comparisons. To assess correlations between samples, we calculated Pearson correlations after raising the expression values to the power of 0.11. We subsequently defined the D-statistic as the median of the pairwise correlations between a sample and all other samples.

Allele-specific expression. The following heterozygous sites were considered for this analysis: (i) sites with 50-bp mappability <1; (ii) sites showing <5% difference in the mapping of simulated reads that carry the reference or non-reference allele; (iii) sites covered by ≥8 reads in each individual. We used a binomial test to compare the REF/NONREF allele counts to the expected ratio (calculated after correction for any remaining genome-wide mapping bias as well as GC bias in each individual).

Summary statistics from allele-specific expression analysis can be used to detect sample contamination and sample swaps, as such errors affect the heterozygosity over variant sites. To this end, we calculated two statistics per sample. (i) The proportion of sites where both alleles are observed in mRNA-seq reads, out of all the sites where allele-specific expression is measured. Whereas observing only one allele may sometimes be caused by true monoallelic expression, a high proportion of such sites suggests sample mislabeling, with genotype and mRNA-seq data coming from different individuals and many heterozygous sites in genotype data being actually homozygous, thus leading to only one allele observed in mRNA-seq data. (ii) Another diagnostic statistic is the proportion of sites with significant (binomial test $P < 0.05$) allele-specific expression out of all the sites. This proportion would detect sample mislabeling as well—as a very strong increase—but it can also capture more subtle sample contamination in mRNA-seq data: when analyzing allelic ratios of a heterozygous site in an individual, even a small amount of RNA from another individual who is often homozygous for the site will bias the allelic ratios and increase the probability of significant allele-specific expression.

Quantitative dissection of sources of variation. To assess the contribution of different sources of variation to transcript expression, we analyzed the expression of 74,634 transcripts that were expressed in >50% of the samples. To be able to estimate technical variation, we selected only the 376 samples coming from 173 unique RNA preparations that were analyzed more than once. Transcript quantification was normalized by using the trimmed mean of M-values (TMM) normalization method from the edgeR package[43] (v. 2.6.9), which includes scaling with respect to differences in sequencing depth after trimming of ratios. Subsequently, data were subjected to logarithmic transformation and the mean-variance trend was removed using the voom function from the limma package (v. 3.12.1; http://www.bioconductor.org/packages/2.11/bioc/html/limma.html). We subsequently analyzed the contribution of different sources of variation in the RNA sample itself or introduced during the sample preparation procedure, avoiding the inclusion of sources of variation that were confounding. Standard (nonhierarchical) linear models in R were fitted for each transcript, taking into account the weights calculated by the voom function that are based on the inverse of the variance. For each transcript, the percentage of variation explained by each factor was calculated from the resulting ANOVA tables by dividing the sum of squares by the total sum of squares. Boxplots demonstrate the distribution of the percentage of variation explained across transcripts.

PEER correction. Exon, transcript and sRNA quantification were corrected using PEER[29,30], which finds synthetic covariates from quantification data that can then be regressed out from the data. Ten and nine covariates were used for mRNA and sRNA quantification, respectively. For calculation of correlations between samples after PEER, all negative expression values were set to zero and subsequently raised to the power of 0.11 (OPS transformation).

39. Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40**, 10073–10083 (2012).
40. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
41. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
42. Berninger, P., Gaidatzis, D., van, N.E. & Zavolan, M. Computational analysis of small RNA cloning data. *Methods* **44**, 13–21 (2008).
43. 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).

doi:10.1038/nbt.2702