Reproducible mRNA and small RNA sequencing across different laboratories

Peter A.C. ’t Hoen[1,*], Marc R. Friedländer[2,¶], Jonas Almlöf[3,¶], Michael Sammeth[2,4], Irina Pulyakhina[1], S. Yahya Anvar[1,5], Jeroen F.J. Laros[1,5], Henk P.J. Buermans[1,5], Olof Karlberg[3], Mathias Brännvall[3], The GEUVADIS Consortium[1,2,3,4,6,9,10,11,12,13,#], Johan T. den Dunnen[1,5], Gert-Jan B. van Ommen[1], Ivo G. Gut[4], Roderic Guigó[2], Xavier Estivill[2], Ann-Christine Syvänen[3], Emmanouil T. Dermitzakis[6,7,8], Tuuli Lappalainen[6,7,8,*]

Affiliations

1 - Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
2 - Center for Genomic Regulation (CRG), 08003 Barcelona, Spain
3 - Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, 751 85 Uppsala, SwedenUniversity
4 - Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain
5 - Leiden Genome Technology Center, Leiden University Medical Center, Leiden, the Netherlands
6 - Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
7 - Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, 1211 Geneva, Switzerland
8 - Swiss Institute of Bioinformatics, Geneva, Switzerland
9 - European Bioinformatics Institute, Hinxton, United Kingdom
10 - Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany
11 - Institute of Human Genetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany
12 - Max Plank Institute for Molecular Genetics, 14195 Berlin, Germany
13 - Fundacion Publica Galega de Medicina Xenomica SERGAS, Genomic Medicine Group CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

*Corresponding authors
¶Equal contribution
#For the GEUVADIS Consortium: Gert-Jan B. van Ommen[1], Xavier Estivill[2], Roderic Guigó[2], Ann-Christine Syvänen[3], Ivo G. Gut[4], Emmanouil T. Dermitzakis[6,7,8], Stylianos E. Antonorakis[6,7], Alvis Brazma[9], Stefan Schreiber[10], Robert Häsler[10], Philip Rosenstiel[10], Thomas Meitinger[11], Hans Lehrach[12], Ralf Sudbrak[12], Angel Carracedo[13], Tuuli Lappalainen[6,7,8], Michael Sammeth[2,4], Marc R Friedländer[2], Peter AC ‘t Hoen[1], Jean Monlong[2], Mar Gonzàlez-Porta[9], Natalja Kurbatova[9], Thasso Griebel[4], Pedro G Ferreira[2], Matthias Barann[10], Thomas Wieland[11], Liliana Greger[9], Maarten van Iterson[1], Jonas Almlöf[3], Paolo Ribeca[4], Irina Pulyakhina[1], Daniela Esser[10], Thomas Giger[6], Andrew Tikhonov[9], Marc Sultan[12], Gabrielle Bertier[2], Esther Lizano[2], Henk PJ Buermans[1,5], Ismael Padioleau[6,7,8], Thomas Schwarzmayr[11], Olof Karlberg[3], Halit Ongen[6,7,8], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[12], Paul Flicek[9], Tim M Strom[11]

Corresponding authors:

Peter A.C. ’t Hoen
Department of Human Genetics
Leiden University Medical Center
PO Box 9600
2300 RC Leiden
The Netherlands
Email: p.a.c.hoen@lumc.nl
Telephone: +31-71-5269421

Tuuli Lappalainen
Department of Genetic Medicine and Development
University of Geneva Medical School
Rue Michel-Servet 1
1211 Geneva
Switzerland
Email: tuuli.e.lappalainen@gmail.com
Telephone: +41-22-3795551

Abstract

RNA-sequencing is an increasingly popular technology for genome-wide analysis of transcript structure and abundance. However, understanding of the sources of technical and inter-laboratory 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 mainly manifested in differences in insert size and GC content, and could be adequately corrected for. In small RNA sequencing, the miRNA content differed widely between samples due to competitive sequencing of rRNA fragments. This did not affect relative quantification of miRNAs. We conclude that distributed RNA-sequencing is well feasible, given proper standardization and randomization procedures. We further provide a set of quality measures and guidelines for assessing technical biases in RNA-seq data.

Main text

RNA-sequencing (RNA-seq) has dramatically changed the field of transcriptomics[1-4]. While expression microarrays were limited to the detection of known transcripts and had limited capacity to differentiate between transcript variants, RNA-seq is in principle able to detect all coding and non-coding transcripts in the cell and to derive the structures of these transcripts. Moreover, sequencing-based methods for expression profiling appear to be more accurate and more sensitive towards lowly abundant transcripts[5-11], even if increased variability in the low expression range has been reported[12, 13]. Nevertheless, RNA-seq is not free from 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 performed 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 distributed sequencing. Moreover, we provide a set of essential quality measures for RNA-seq experiments and a routine that adjusts for partially unknown sources of technical variation. The biological interpretation of the results is reported elsewhere (Lappalainen et al. submitted, Suppl. File).

Results

Lay out of the study

A major objective of the current study was to evaluate the feasibility of distributed RNA-sequencing. To this end, we distributed 465 total RNA samples from lymphoblastoid cell lines (LCLs) from five populations in a randomized way across seven different European laboratories. Each center received between 48 and 113 randomly assigned samples and strict sample preparation and sequencing guidelines (Suppl. File). At these seven sites, the mRNA and small RNA (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. To allow proper estimation of laboratory effects, five RNA samples were prepared and sequenced at all sites; 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 data (fastq files) were subsequently aligned with GEM[23] (mRNA data) and miraligner[24] (sRNA data), and analyzed with a common pipeline quantifying exon, transcript and sRNA expression levels (see Methods).

Basic quality control steps in mRNA-seq

The laboratories were free to choose the number of samples to be pooled in one lane. While the target number of reads was minimally 20M (10 M paired reads), the laboratories generally decided to choose conservative pooling schemes avoiding the need for repetition of samples with too low coverage. This resulted in a median of 58M reads with a broad distribution ranging from 17 - 167M (Figure 1A). This large range was partly due to differences in the number of samples per lane and partly due to difficulties with equimolar pooling, resulting in an up to 3-fold difference between the highest and lowest number of reads per sample in a lane.

Figure 1 summarizes some of the basic quality measures that were assessed and the performance of the different laboratories on these. A more extended list of all quality measures assessed is given in Suppl. Table 1 and 2. All samples had similarly high mean PHRED scores, a measure for the quality of the base calling (Figure 1B). The mean number of bases per read with quality score higher than Q30 also reflected high sequence quality (Figure 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 over Q30 are therefore acceptable. The percentage of aligned reads was generally very high except for a few samples (between 95-100%, Figure 1E). Some of these outliers were associated with high duplication rates (Figure 1D). Downstream analysis showed that lower mapping and higher duplication rates did not seem to affect the quantification of exons and transcripts, since the expression levels in these samples correlated strongly with all other samples (Figure 2B,C). The percentage of aligned reads mapping to annotated exons were generally between 60 and 80% (Figure 1F). This is an important quality measure, since it collectively captures variation in enrichment for mature mRNAs, possible contaminations and effectiveness of the alignment procedure. There was one sample (NA18861.4) with only 4% of aligned reads mapping to exons, while still having a high overall mapping rate. Given extensive coverage in introns and intergenic regions, it must have been 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 sub-optimal polyA$^+$ RNA selection. Again, this did not affect the quantification of the exons and transcripts (Figure 2B,C).

Detection of problematic samples

To detect whether problematic samples could already be identified before alignment, we analyzed the distance between k-mer profiles. To this 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[25]. The k-mer profile of NA18861.4 was clearly different from the rest (Figure 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 (Suppl. Figure 1), but some samples with high duplication rates and/or high rRNA content were only identified 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 exon and transcript quantifications to detect problematic samples. Given the skewness of RNA-seq data, where there are few highly expressed and many lowly abundant transcripts, use of the Pearson correlation on the linear scale is not appropriate. Therefore, we first applied an optimal power space (OPS) transformation (Ribeca and Sammeth, in preparation), ensuring that low and high abundant transcript outliers do not bias the correlation measure and all data points contribute equally to the computed coefficient (see Suppl. Figure 2). Figure 2B provides the distribution of the median Pearson correlations (D-statistics) of the exon quantifications for each sample. NA18861.4 (with only 4% exonic reads) had clearly lower correlations to the other samples. NA19144.4 was identified as an additional outlier and was removed from the analysis. Quantification at the transcript and gene level identified the same outliers (Suppl. Figure 3). In general, the correlations of gene expression levels were stronger than correlations of exon expression levels due to their increased robustness conferred by more read mappings. Transcript quantifications correlated much less than gene or exon quantifications due to inherent uncertainty in the deconvolution and relative quantification of transcripts from the same gene (Suppl. Figure 3).

Sample mix-ups are a general problem in studies analyzing large cohorts of samples and may severely compromise their power[26]. 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-chromosomal genes (exclusively expressed in males) (Figure 2C). Clear sample swaps, where females showed expression of Y-chromosomal genes without expression of *XIST* or vice versa, were not observed. However, in three samples, there was expression of both *XIST* and Y-chromosomal genes, indicative of contamination between samples.

Identification of sample mix-ups in studies where DNA genotypes are available, is relatively straightforward. We evaluated the number of heterozygous sites with expression of both alleles in each individual. Due to allele-specific expression, this is usually not 100% but generally >90% of all heterozygous sites in expressed genes. In 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 (Figure 2D). We also analyzed the percentage of sites showing significant allele-specific expression (ASE), *i.e.* imbalance between the expression of the two alleles (p<0.05, binomial test). This measure is also sensitive to sample contamination, since the expected 50-50 allelic ratio over a heterozygous site in a given individual will be biased if even a small proportion of RNA-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 (Figure 2D). In addition to these samples, the ASE analysis identified one more sample (NA19225.6) with potential contamination, probably originating from a sample with the same gender, and the one problematic sample (NA18861.4), which had a low proportion of exonic counts. This kind of contamination would be particularly problematic to identify in RNA-seq data sets without genome sequences.

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 to trace the most important sources of technical variation. When comparing individual LCLs, the biological variation is limited, since the only biological difference is the individual's genetic and epigenetic background, while 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 (Figure 3A, Suppl. Figure 4). The correlations for replicate samples run in the same laboratory were slightly higher than for samples run in different laboratories (on average 0.931 *vs.* 0.925 for exon quantifications). The clustering by sample was much stronger for exon quantifications than for transcript quantifications (compare Figure 3A and B and Suppl. Figure 4A and B). When considering all 667 sequence runs, some laboratory-driven clustering was observed, particularly for transcript quantifications (Figure 4D, Suppl. Figure 5B).

Given the stronger effects of technical variation on transcript than on exon quantifications, 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 (Figure 3C). Slight inter-day differences between library preparations and effects of the different index primers were also notable (Figure 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 availability of standardized protocols, slight differences in library preparations between laboratories were observed, amounting to around 5% of the total variation (Figure 4A). Most notably, these manifested in differences between the average GC-percentage, the width of the distribution of GC-percentages and the insert sizes (Figure 3D,E,F, Figure 4G). The exons with high GC content (>65%) demonstrated more variable expression levels between laboratories than exons with medium or low (<35%) GC content (Suppl. Figure 6A). Relatively low representation of sequences with >65% GC may be explained by the use of thermocyclers with high ramping speeds (Suppl. Figure 6B)[27].

While all laboratories aimed for an insert size of ~10 bp (corresponding to a fragment size of 280 bp: 10 + 2x75 (read length) + 120 bp (length of the adapters)), most laboratories achieved slightly lower insert sizes, 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 (Suppl. Figure 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 quantifications.

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

Correction for variation in mRNA-seq

Next, we explored the correction for technical sources of variation. For this correction, we used a recently described Bayesian framework that accounts for hidden variables in expression data (PEER[28, 29]). Before correction, the variable 'laboratory' explained 6.8% of the total variance (average across all genes, median: 3.8%) (Figure 4A). After partialling out the first 10 components, the laboratory effect was reduced to 2.6% (average across all genes; median 2.0%) (Figure 4B). Laboratory effects were mainly captured by PEER components 1, 3, 6, 7, and 8 (Figure 4G, Suppl. Figure 9B). Moreover, the PEER components were correlated with several of the other observed sources of technical variation: insert size (component 1, 7), and GC-content and other nucleotide biases (component 2, 5, 6, and 8) (Figure 4G, Suppl. Table 4). Finally, components 1, 4, 5 and 7 were correlated with differences in the coverage in different regions of the transcript (Figure 4G, Suppl. Table 4). This effect was not related to the RNA integrity, which was captured mainly by component 10, and may reflect biases introduced in the reverse transcription step.

After PEER correction of the transcript expression levels, samples clustered more strongly by population and less strongly by laboratory (Figure 4C-F). Thus, technical variation, and in particular variation that is introduced by distributed sequencing, can be properly accounted for and has only limited influence on exon or transcript quantifications in a distributed sequencing setting.

Basic quality control and laboratory effects in sRNA-seq

We analyzed 492 samples by sRNA-seq, targeting for 3 to 6 million mapped reads. The obtained sequencing depth varied considerably, from 0.1 to 50 million reads per sample, with a median of 8.6 million reads (Figure 5A). The sequencing quality was uniformly high, with sample mean PHRED score in a narrow band from around 36 – 39 (Figure 5B, more quality measures in Suppl. Table 3). After adapter trimming and before mapping, we discarded all sequences shorter than 18 nts, since they cannot be traced to genomic loci with high confidence. The fractions of reads thus discarded differed between samples, ranging from 0.5 to 81%, with strong dependence on the sequencing lab (Figure 5C). This wide range may be caused by slight differences in gel separations and purifications, which were performed in the individual laboratories, or by variable degradation during library preparation. Since many sRNAs are repetitive, we mapped the reads to the human genome allowing for multiple mappings (Methods). The mapping efficiencies were uniformly high, consistent with the high sequencing quality (Figure 5D). Surprisingly, the relative miRNA content in our samples ranged from 2 to 62% of mapped reads, with a median of 19% (Figure 5E). Given that some sRNA-seq studies report miRNA contents above 90% (*e.g.*[30]), 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 (Figure 5F). Moreover, the same miRNA genes were consistently profiled: the 500 most highly expressed miRNAs were detected, on average, in >96% of the samples.

Differences in relative contribution of sRNAs do not affect quantification of miRNAs

Tracing the sequenced sRNAs to their genomic sources, we found that they originate not just from miRNA genes, but also from other non-coding RNA genes, in particular rRNA (Figure 6A). Clustering divided the samples into those dominated by miRNA and those dominated by rRNA. The two groups were not associated with particular laboratories (Figure 6A, lab color bar, left). Moreover, the replicates sequenced in all laboratories grouped mostly by sample and not by laboratory (Suppl. Figure 10), and the miRNA and rRNA contents were more similar within samples than within laboratories (Suppl. Figure 11). Importantly, the miRNA contents clearly varied between RNA extraction batches (Suppl. Figure 12). Likewise, snoRNA and other small RNA proportions clearly varied between samples (Figure 6A). In conclusion, differences in the proportions of the different small RNAs are likely introduced during RNA isolation, 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 adapter clipping (Figure 6B). In contrast, the reads which originated from rRNAs were 35 nucleotides long. Since ~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 small RNA contents biased the quantification of individual miRNAs, we calculated the expression levels of 715 miRNA genes based on their read counts. The samples did not group according to miRNA or rRNA content (Figure 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 4 samples from the biological analysis that had D-statistics below 0.8 (Suppl. Figure 13). 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 (Suppl. Figure 14, Suppl. 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 under the following provisions: use of the exact same protocols (Suppl. File) and versions of sample preparation and sequencing kits in all participating laboratories. However, even when using identical protocols, slight variations in average GC content and insert size were observed. These differences translated into variations in transcript quantifications, while exon quantifications were 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 also considerable intra-laboratory variation was observed (Figure 3). Based on the current study, we propose several parameters that should be assessed in any mRNA-seq data set to assess the quality of the samples and/or explore the need for correction of important biases:

-**distribution of base quality scores:** This is the most basic quality measure, already implemented in nearly all sequencing centers, to address the quality of the sequencing run

-**average and distribution of GC-content:** Differences in average GC-content and differences in proportion of reads with extreme (<35% or >65%) GC contents induces biases in transcript quantifications, which can be partially corrected with dedicated tools[18, 31].

-**average and standard deviation of insert size:** this parameter influences mostly transcript deconvolution and quantification. A dedicated routine for correcting this bias has not been described so far.

-**percentage of reads mapping to annotated exons:** this parameter 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 RNA-seq experiments, at least 60% of mapped reads should overlap with annotated exons in a good quality sample.

In addition, checks for **sample swaps** and **contaminations** should be implemented. Procedures described in this and other papers[32, 33] 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. Like for alternative methods using surrogate variables[34] or principal components[35], it is not necessary to know the sources of variation beforehand. However, these routines can only be used in relatively large studies. For smaller studies, dedicated algorithms and standard regression methods may be applied to correct for known technical biases[14, 17, 18, 31]. Still, minimizing technical variation by careful standardization of protocols and randomization in every experimental step (cell line handling – RNA extraction – sample preparation – sequencing) is essential.

This study focused on the quantification of both mRNA and sRNA. Although sRNA sample preparation is generally regarded to be more challenging than its mRNA counterpart, technical variation introduced in the sample preparation seems limited compared to differences originating from the RNA isolation procedure. Interestingly, huge differences in the miRNA content among the sequenced small reads had no major impact on miRNA quantifications. The latter indicates that sRNA-seq data should not be analyzed as a whole, but split into different sRNA fractions before normalization.

Our RNA-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 inter-site variability is limited when working with standardized protocols. However, it is difficult to compare the inter-laboratory reproducibility of RNA-seq in this study with the inter-laboratory reproducibility of gene expression micorarrays in the MAQC study, given differences in experimental design, differences in the scales on which RNA-seq and microarray data are reported, and the much higher dynamic ranges of RNA-seq counts compared to micorarray 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 (Lappalainen et al. submitted). The small effect sizes detected with the majority of eQTLs, confirms conclusions from earlier papers that RNA-seq technology is at least as robust as microarray technology[5-11, 19, 36].

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

Acknowledgements

This project was funded by the European Commission 7th Framework Program, Project N. 261123, GEUVADIS, as well as The Swiss National Science Foundation, the Louis Jeantet Foundation, The Swedish Research Council (90559401), the Knut and Alice Wallenberg Foundation (2011.0073), the DFG Cluster of Excellence Inflammation at Interfaces, the INTERREG4A project HIT-ID, the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project, grant agreement 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). Spanish Plan Nacional funding SAF2008-00357, Academy of Finland, Emil Aaltonen Foundation, EMBO long-term fellowship ALTF 225-2011, European Community's FP7 75 CAGEKID (grant agreement 241669), the SNP&SEQ Technology Platform in Uppsala (www.sequencing.se) for sequencing, and the Swedish National Infrastructure for Computing (SNIC-UPPMAX) (www.uppmax.uu.se) for compute resources for data analysis.

The authors would like to thank Dr. P.G.M. van Overveld for help with preparation of the figures.

Author contributions

PACtH, MRF, JA, MS, IP, SYA, JFJL, HPJB, MB, OK, and TL performed the analyses. PACtH, ACS, RG, XE, JTdD, GJBvO, IGG, ETD designed the study. ETD and TL coordinated the study. PACtH drafted the manuscript which was subsequently revised by all co-authors.

Legend to Figures

<u>Figure 1:</u> Basic quality statistics in mRNA sequencing across laboratories. Distribution of sequencing characteristics over 667 samples sequenced in 7 different laboratories, colored by the indicated coloring schemes. 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.

<u>Figure 2:</u> Detection of outliers in mRNA sequencing. A. Histogram of median pairwise k-mer distances of a sample with all other samples; B. Histogram of median pairwise Pearson correlations (D-statistics) between exon quantifications after OPS transformation; C. Gender-specific expression: normalized expression levels of female-specific *XIST* transcript (x-axis) *vs.* sum of the normalized expression levels of Y-chromosomal transcripts excluding transcripts in the pseudo-autosomal regions (y-axis); D. Allele-specific expression analysis: for all heterozygous sites considered (see Methods), the proportion of heterozygous 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).

<u>Figure 3:</u> Sources of variation in mRNA expression levels. A. Multidimensional scaling (MDS plot) of correlation of exon quantifications of five replicate samples (indicated with different colors) across seven different laboratories (indicated with different symbols). B. MDS plot of correlation of transcript quantifications of five replicate samples (indicated with different colors) across seven different laboratories (indicated with different symbols) C. Percentage variation in mRNA data explained by RNA integrity value (RIN), 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, 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 Hi-seq instrument in which the sample was run. Boxplots show distribution of the percentage of variance explained across all transcripts expressed in >50% of samples; D. Boxplot of mean GC percentage in the reads across samples sequenced in different laboratories; E. Boxplot of standard deviation in GC percentage in the reads across samples sequenced in different laboratories; F. Boxplot of mean inferred length of sequence between first and second reads (negative means overlapping sequences) of samples sequenced in different laboratories.

<u>Figure 4:</u> Modeling of hidden confounding factors with PEER effectively removes biases in RNA-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; C. MDS plot of transcript quantifications before PEER colored by population (CEU=red, GBR=olive, FIN=blue, TSI=purple, YRI=orange); D. MDS plot of transcript quantifications before PEER colored by laboratory (same color code as in Figure 1); E. MDS plot of transcript quantifications after PEER colored by population (same color code as in Panel C); F. MDS plot of transcript quantifications after PEER colored by laboratory (same color code as in Figure 1); G. Most important sources of variation correlated to each PEER factor, strength of these correlations (blue bars) and the correlation of the laboratory effect to each PEER factor (green bars). 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.

<u>Figure 5:</u> Basic quality statistics in sRNA sequencing across laboratories. Density plots for 492 samples sequenced in 7 different laboratories labeled by the indicated coloring schemes. 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.

<u>Figure 6:</u> 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 purple) to 82% (light orange) of total sRNA in each sample, as indicated by the color key in the upper left corner. We have divided samples into those which are 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 panel A, but now the length of sequenced RNAs (after adapter 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 purple. The samples do not group (are not biased) by the relative contents of miRNA and rRNA.

Online Methods

Samples and sequencing

Complete details on the lay-out of the study and the quantification pipelines are described in Lappalainen et al. (submitted).

In brief, EVB transformed LCLs from Coriell Cell Repositories (GBR, FIN, TSI) and University of Geneva (CEU, YRI) were cultured at ECACC. Cell pellets were shipped to University of Geneva for total RNA extraction using the TRIzol Reagent (Ambion). 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 ug of total RNA of the samples allocated to them, and total RNA Bioanalyzer was ran for 10-20% of the RNA samples before library preparation to confirm sample quality after shipping. Library preps were done in random order in every laboratory. Guidelines provided to the different laboratories are included in the data supplement (Suppl. File).

mRNA sequencing was done on the Illumina HiSeq2000 platform with 75 bp paired-end sequencing with fragment size of 280 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. Small RNA sequencing was done on the Illumina HiSeq platform with 36-50 bp single-end sequencing with fragment size of 145-160 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.

Each lab submitted one demultiplexed fastq file per sample per mRNA and per sRNA-seq, as produced by CASAVA 1.8 or 1.8.2 allowing one mismatch in the index. CASAVA 1.8.2 quality filters were applied to fastq files processed with CASAVA 1.8 to make them comparable.

mRNA analysis pipeline

We iterated 3 incremental cycles of mapping with the GEM tool[23] employing the JIP pipeline (Griebel and Sammeth, in preparation) to adjust the region considered during alignment and the mapping parameters (Lappalainen et al. submitted). 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 quantifications were 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[37] was used for quantifications of transcripts, and were based on the annotation-mapped genomic mappings considering transcript structures of the Gencode transcriptome annotation, taking into account mappings of read pairs that were completely included within the annotated exon boundaries and paired in the expected orientation.

sRNA analysis pipeline

Datasets 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[24] 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[38] and snoRNA and miRNA annotations from the UCSC table browser[39]. 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[40]. Each read mapping was weighted inversely to the number of genome mappings for the read, *e.g.* a read mapping to two genomic locations would get an assigned weight of 0.5. Each mapping was counted towards the annotation of the nucleotide in the middle of the mapping. miRNA quantification was performed with the custom tool miraligner[24].

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[41] (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 was 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 Suppl. Table 2, Figure 4G and Suppl. Figure 9.

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 (Anvar et al., in preparation). Subsequently, the pairwise distance between the profiles of the different samples was calculated using the multiset distance measure[25]. 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

Since gene expression follows a power law distribution (Ribeca and Sammeth, in preparation), it is intuitive to use a suitable exponent in order to transform data to a more normal-like 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 datasets produced by different laboratories, we found consistently an OPS exponent of 0.11 for all sample comparisons. To assess correlations between samples, Pearson correlations were calculated after raising of 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,: 1) sites with 50bp mappability <1; 2) sites showing <5% difference in the mapping of simulated reads that carry the reference or non-reference allele (see Lappalainen et al. submitted); 3) 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, since such errors affect the heterozygosity over variant sites. To this end, we calculated two statistics per sample: (1) The proportion of sites where both alleles are observed in RNA-seq reads, out of all the sites where ASE is measured. While observing only one allele may sometimes be caused by true monoallelic expression, a high proportion of such sites suggests sample mislabeling, with genotype and RNA-seq data coming from different individuals and many heterozygous sites in genotype data being actually homozygous, thus leading to only one allele observed in RNA-seq data. (2) 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 capture also more subtle sample contamination in RNA-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 ASE.

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 which were expressed in >50% of the samples. To be able to estimate technical variation, we only selected the 376 samples coming from 173 unique RNA preparations that were analyzed more than once. Transcript quantifications were normalized by using the trimmed mean of M-values (TMM) normalization method from the edgeR package[42] (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 (non-hierarchical) 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 quantifications were corrected using PEER[28, 29], 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 quantifications, 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).

References

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. 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).

24. 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).

25. Kosters,W.A. & Laros,J.F.J. Metrics for mining multisets in *Research and Development in Intelligent Systems XXIV, Proceedings of AI-2007* 293-303 (Springer, 2007).

26. Gordon,D. & Finch,S.J. Consequences of error. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (ed. Wiley Online Library) 2006).

27. Aird,D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

28. 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).

29. 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).

30. Parts,L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS. Genet.* **8**, e1002704 (2012).

31. Benjamini,Y. & Speed,T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).

32. Huang,J., Chen,J., Lathrop,M., & Liang,L. A tool for RNA sequencing sample identity check. *Bioinformatics.*(2013).

33. 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).

34. Leek,J.T. & Storey,J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS. Genet.* **3**, 1724-1735 (2007).

35. 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).
36. Montgomery,S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777 (2010).
37. Griebel,T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40**, 10073-10083 (2012).
38. Jurka,J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462-467 (2005).
39. Karolchik,D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-D496 (2004).
40. Berninger,P., Gaidatzis,D., van,N.E., & Zavolan,M. Computational analysis of small RNA cloning data. *Methods* **44**, 13-21 (2008).
41. Wang,L., Wang,S., & Li,W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics.* **28**, 2184-2185 (2012).
42. 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).

Supplementary Figures

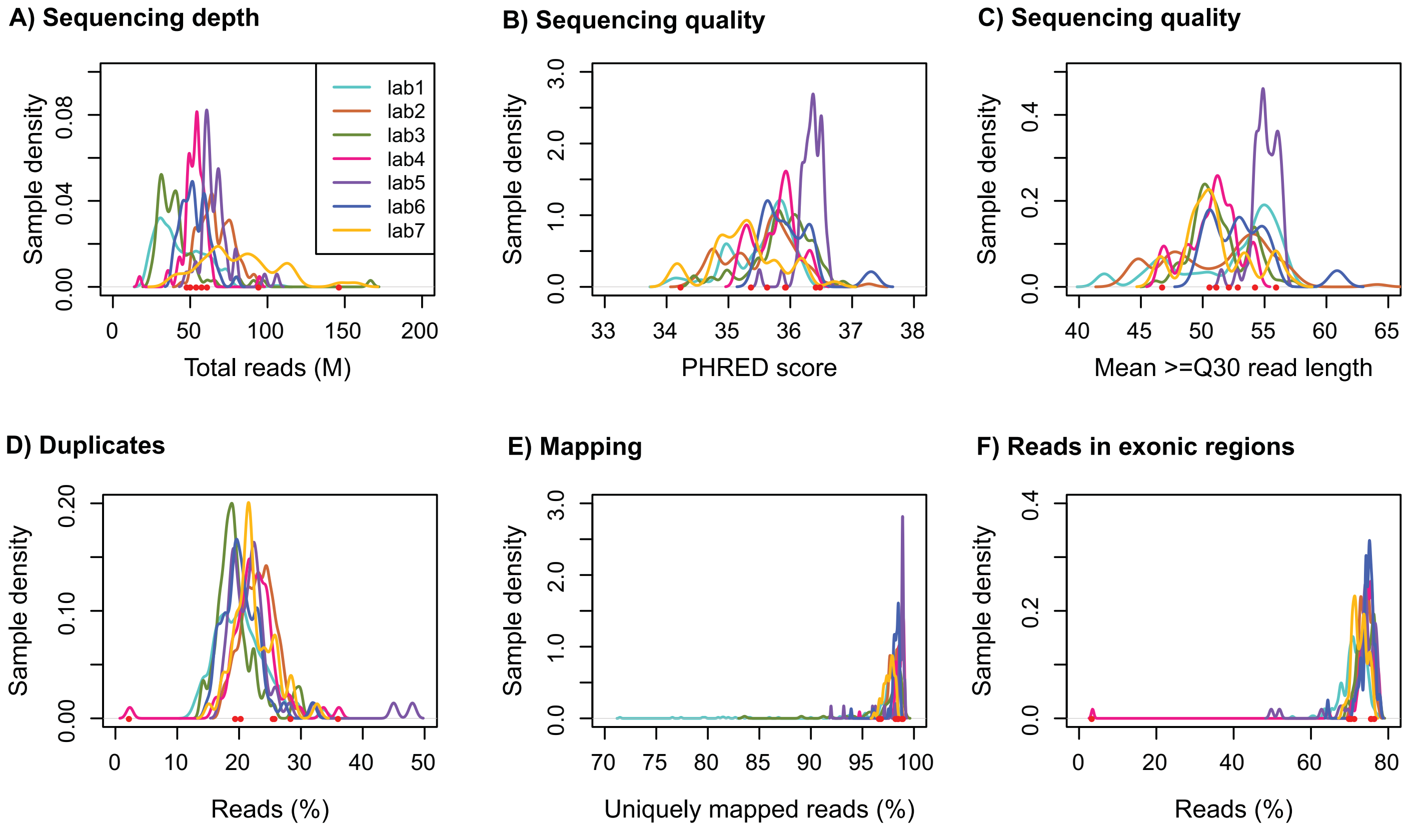
Supplementary Tables

Supplementary Files

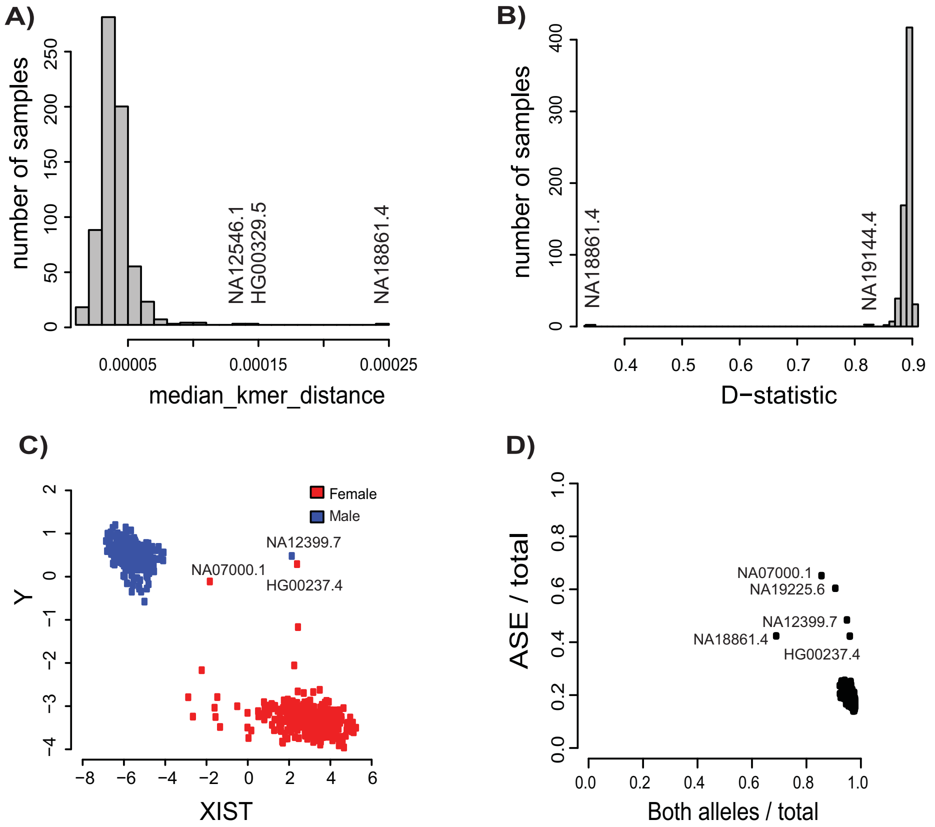
Lappalainen_submitted.pdf: manuscript submitted to Nature describing the biological interpretation of the LCL RNA-sequencing data discussed in the current manuscript

Data access

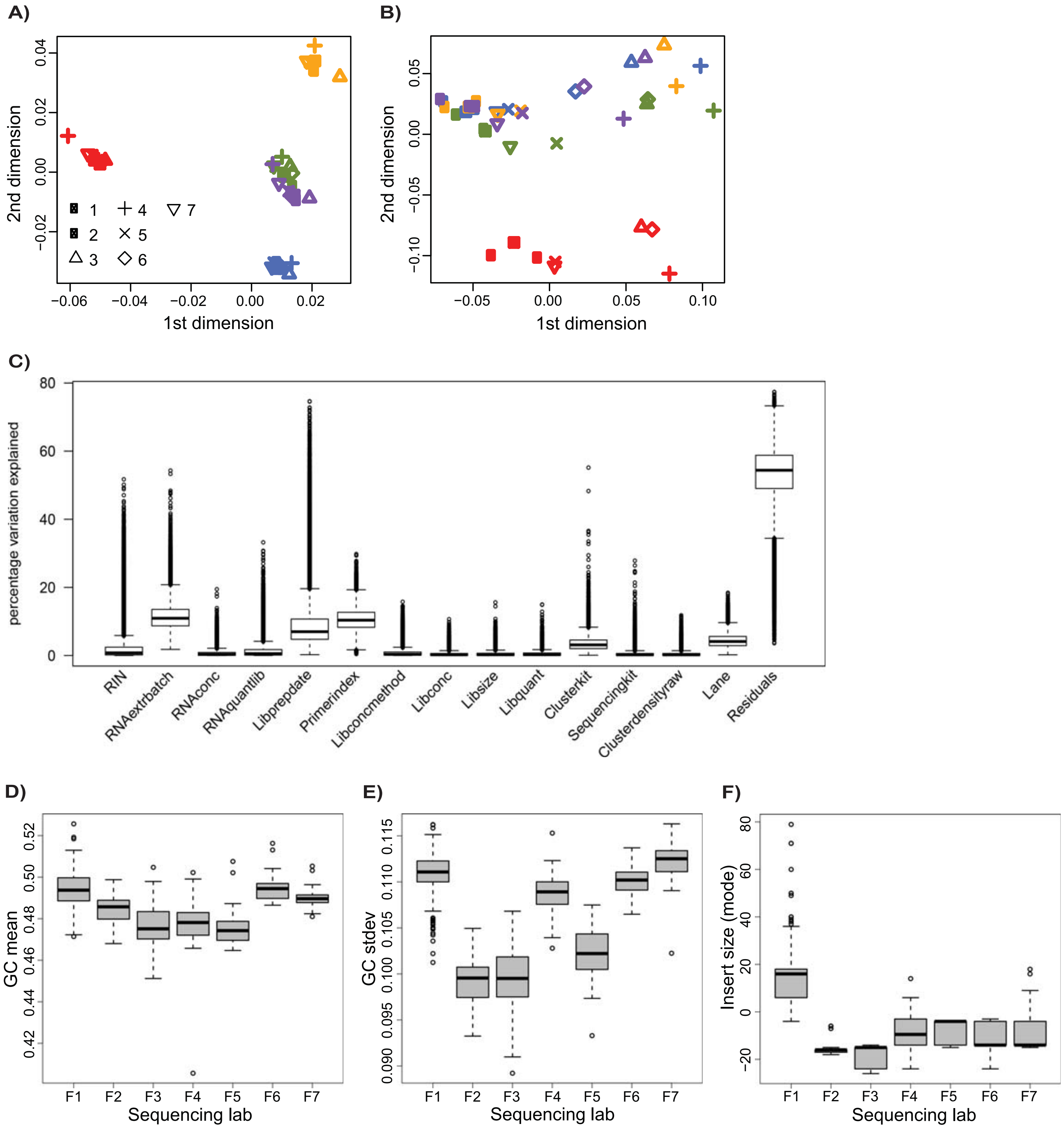
The raw fastq files and bam alignments as well as different types of quantifications are available in ArrayExpress under accessions E-GEUV-1 (mRNA) and E-GEUV-2 (small RNA) for QC-passed samples and E-GEUV-3 for all sequenced samples:
http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/
http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-2/
http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/



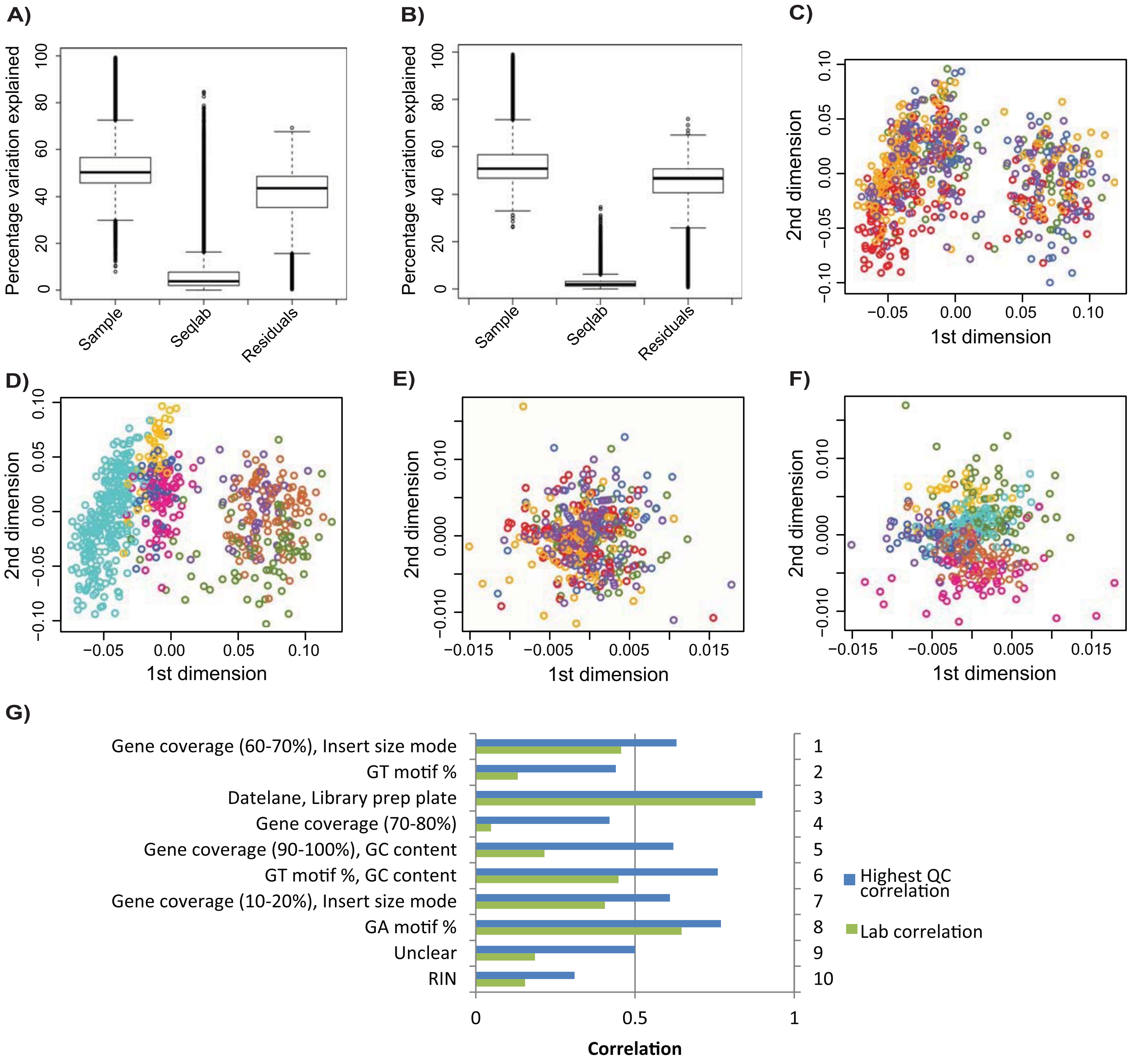
‘t Hoen et al. Figure 1



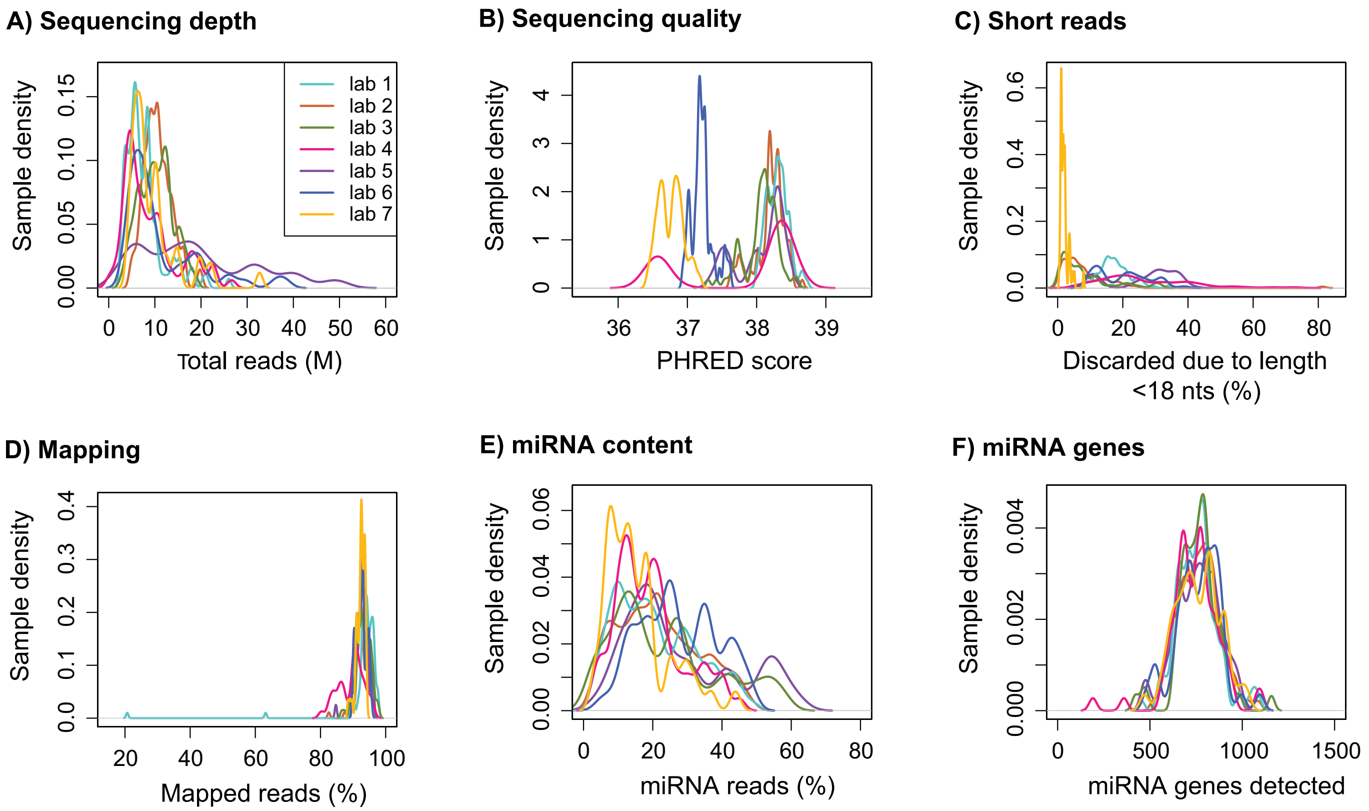
‘t Hoen et al. Figure 2



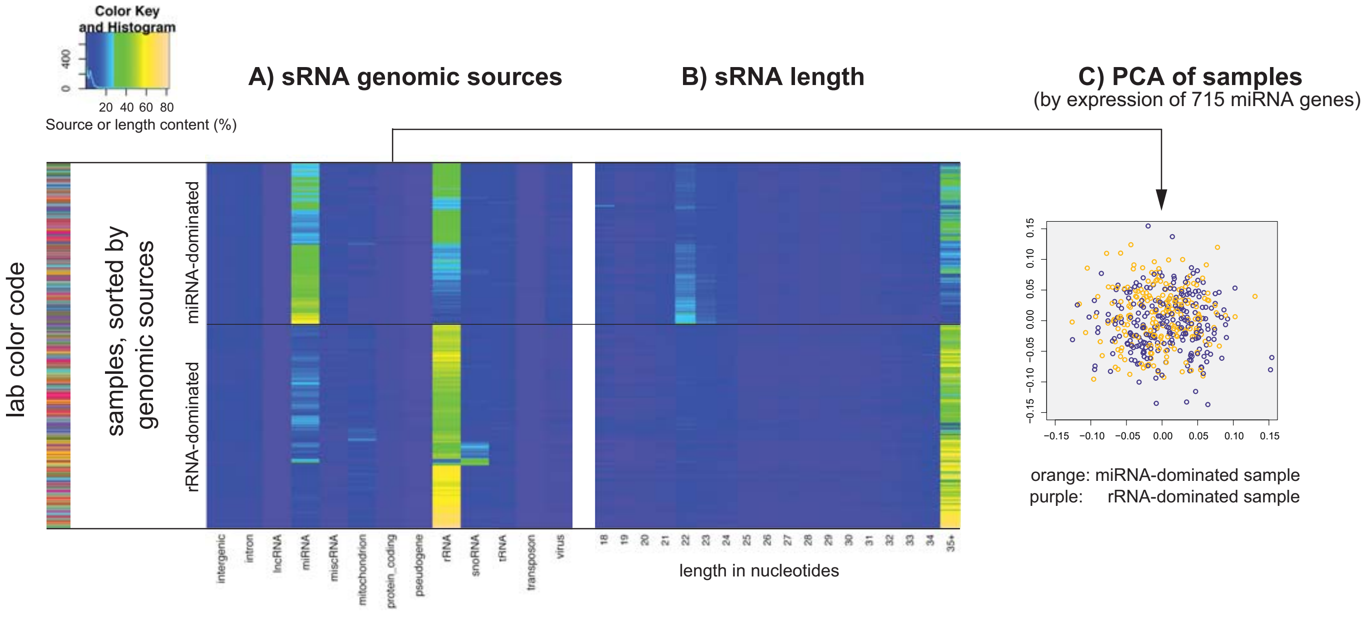
't Hoen et al. Figure 3



't Hoen et al. Figure 4



‘t Hoen et al. Figure 5



't Hoen et al. Figure 6