Characterization and Visualization of Error in Omics Technologies

Robert M Flight and Hunter NB Moseley
Department of Molecular Biology and Biochemistry, Markey Cancer Center
Resource Center for Stable Isotope-Resolved Metabolomics, University of Kentucky

Problem & Goal

Recent years have witnessed the rise of various omics technologies, including but not limited to high-throughput sequencing of DNA (genomics) and RNA (transcriptomics), DNA microarrays (transcriptomics), nuclear magnetic resonance (NMR) and direct infusion Fourier-transform mass spectrometry (metabolomics) and protein sequencing by mass spectrometry (proteomics).

Based on the underlying analytical methodology inherent in each of these technologies of counting a signal at a detector, we hypothesize that each of these omics technologies suffer from a proportional variance to different degrees. Proportional variance is detrimental to commonly used statistical methods as it breaks the assumption of variances being identically and independently distributed (iid normal).

Therefore, we have developed a novel graphical method that visualizes separate additive and proportional variance components in a given dataset, making it easier to detect and quantify both components. Based on this visualization, we have also developed a method to calculate both the additive and proportional error components. We have applied these new methods to datasets from four different omics technologies: i) DNA-seq (drosophila 1000 genomes), ii) RNA-seq (saccharomyces 48 replicate transcriptomics), iii) DNA microarray (MAQC Affymetrix®, and iv) metabolomics FTMS (RCSIRM workshop exosome lipids). Although the degree of proportional variance is different, the resulting graphs clearly show the presence of proportional variance in each dataset and in a form that is easy to quantify.

Different Components of Error

Two primary components of error in most error models are additive (Fig 1) and proportional (Fig 2). We traditionally think of an additive error component as baseline noise, and when visualized by plotting two replicates appears as deviation from the line of identity. It is constant for all values of the signal, and becomes more important at low signal (Fig 1). A proportional error component is multiplicative: as the signal increases, the error increases as well. Plotting two replicates from data that has a proportional error component will show an increasing spread of values with increasing signal, with the log-log plots collapsing to a straight line. Plotting the mean value across replicates vs the standard deviation (sd) across replicates shows increasing variance with mean signal, with constant relative standard deviation (rsd). Below we demonstrate these properties using simulated data: 1000 points drawn from a uniform distribution with values 0 - 10000, and then generating 100 replicates for each point with either additive error only (Fig 1, sd = 20), proportional error only (Fig 2, rsd = 0.1), or both (Fig 3).

Additive Error Only

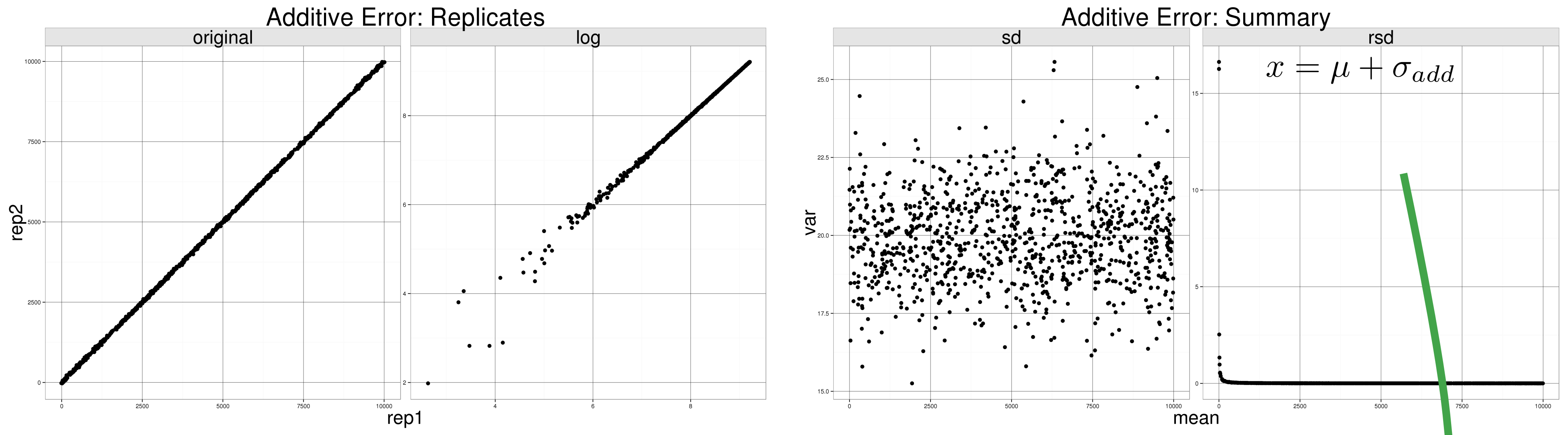


Fig 1. Additive error only. Left: two replicates plotted in "original" space and after log-transformation. Right: following calculation of the sd and rsd of each point across 100 replicates, plots of the mean vs sd and rsd, showing that additive error is constant with respect to the mean value.

Proportional Error Only

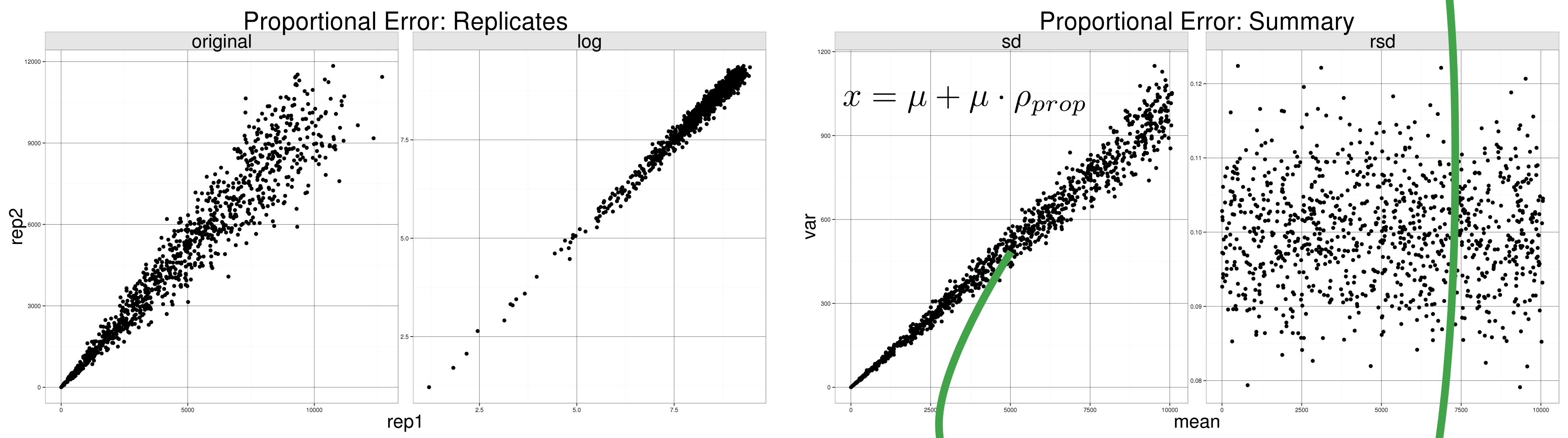


Fig 2. Proportional error only. Left: two replicates plotted in "original" space and after log-transformation. Right: following calculation of the sd and rsd of each point across 100 replicates, plots of the mean vs sd and rsd, showing that proportional error is increasing with the mean value, but constant when calculated relative to the mean value.

Mixed Error

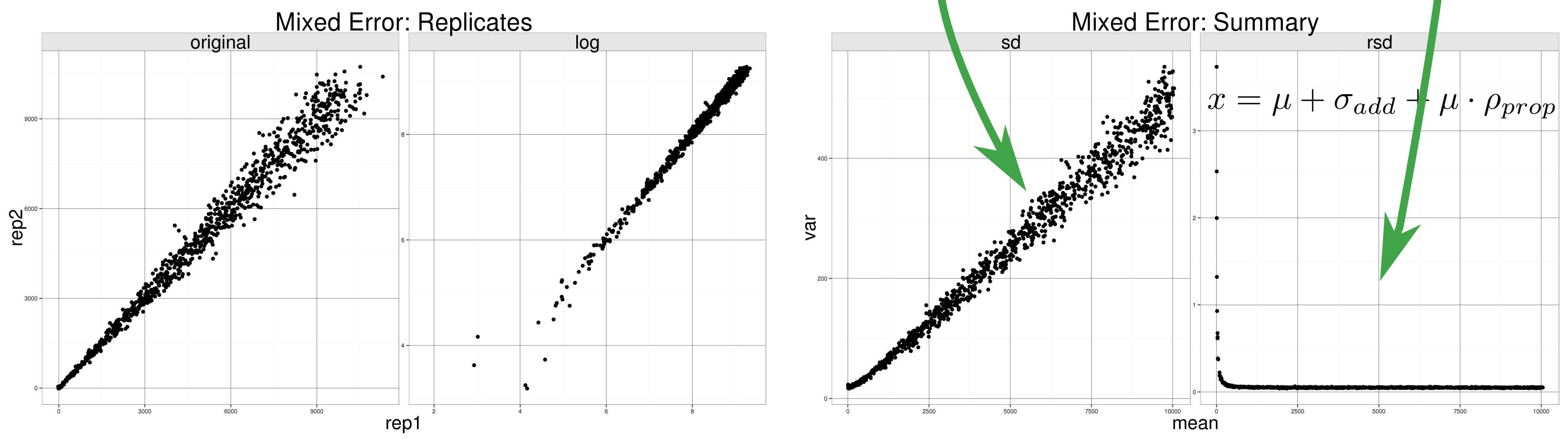


Fig 3. Additive and proportional error. Left: two replicates plotted in "original" space and after log-transformation. Right: following calculation of the sd and rsd of each point across 100 replicates, plots of the mean vs sd and rsd. All plots have characteristics from both the additive and proportional cases, showing that both additive and proportional error are present.

Estimating Component Error Values

$$sd_{rep} = \sigma_{add} + \rho_{prop} \cdot \bar{x}_{rep} \qquad rsd_{rep} = \frac{(\sigma_{add} + \rho_{prop} \cdot \bar{x}_{rep})}{\bar{x}_{rep}}$$

The two summary plots of mean vs sd and mean vs rsd lead to two equations that can be solved using non-linear least squares (see equations above). Practically, initial estimates are generated from the sd equation, and then used as starting points in the rsd equation. The ability to estimate the additive and proportional components was tested by varying the additive component value from 0 - 5000 in increments of 10, and proportional from 0 - 1 in increments of 0.01, generating 100 replicates from error free data 20 times, and estimating the errors for each. Fig 4 shows that as the additive error component is increased, the confidence in the the additive and proportional estimates decrease.

Real vs Estimates

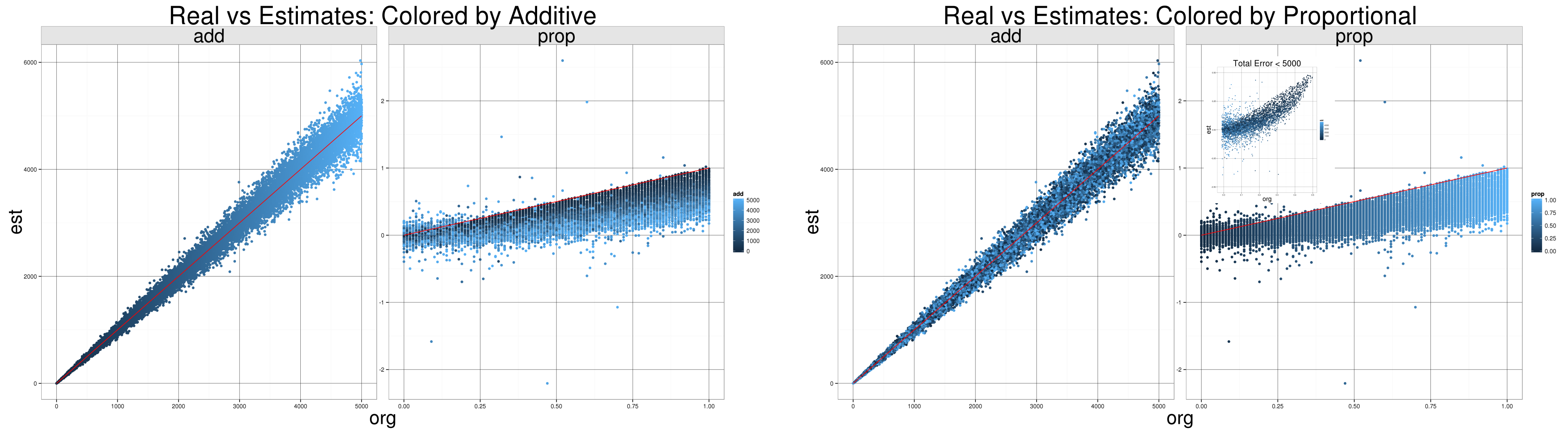


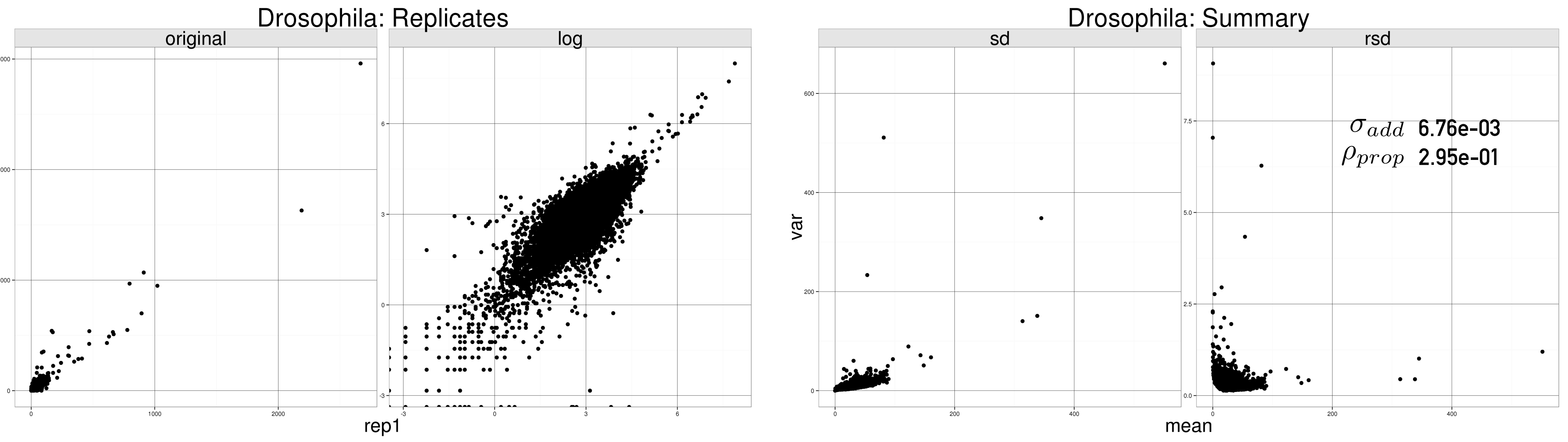
Fig 4. Actual (org) and estimated additive and proportional values colored by either the additive (left) or proportional (right) error. Red is line of identity. Inset shows proportional comparison for points with <= 5000 total error value.

Application to Omics Datasets

We applied the summarization, visualization and estimation of the error components to four different omics datasets spanning currently used technologies: i) genomics by DNA-seq; ii) transcriptomics by DNA microarray, iii) transcriptomics by RNA-seq, and iv) lipid metabolomics by Fourier-transform mass spectrometry.

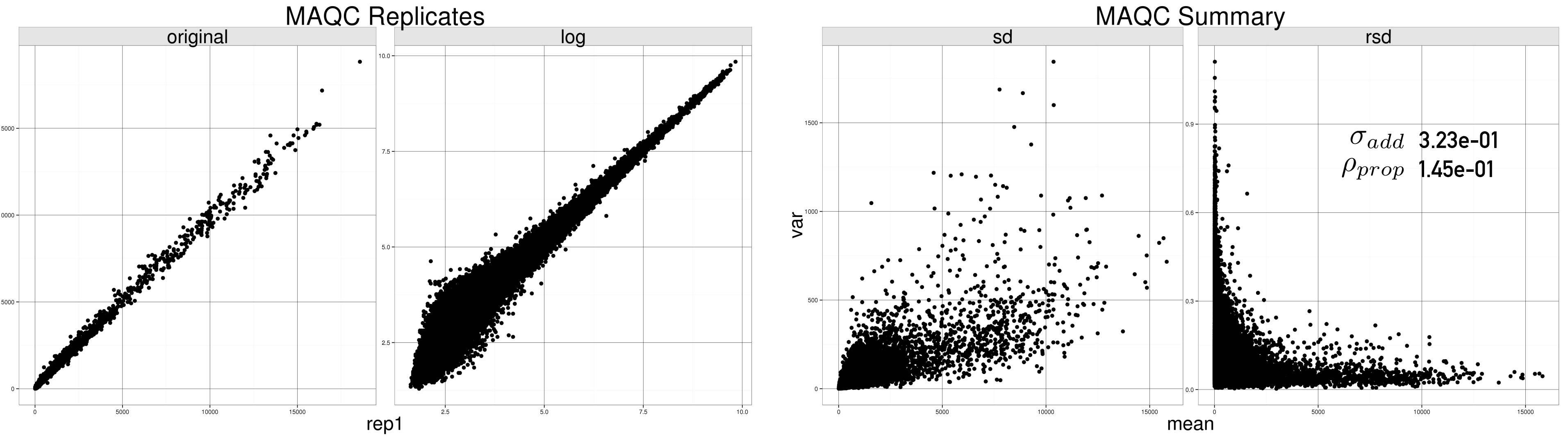
i) DNA-Seq Drosophila 1000 Genomes

Paired end reads from Illumina sequencing of 200 drosophila samples from the sequence read archive (SRA, 1000 drosophila genomes project [1], SRP006733). Aligned using bowtie v 2.0 [2] "very-fast". The drosophila dm3 genome reference was tiled into 2KB segments, and number of reads aligning to each segment counted using summarizeOverlaps [3]. Counts in each sample were normalized by the total number of reads in each sample. 5000 genomic tiles were selected randomly and mean and standard deviation calculated across the 200 replicates.



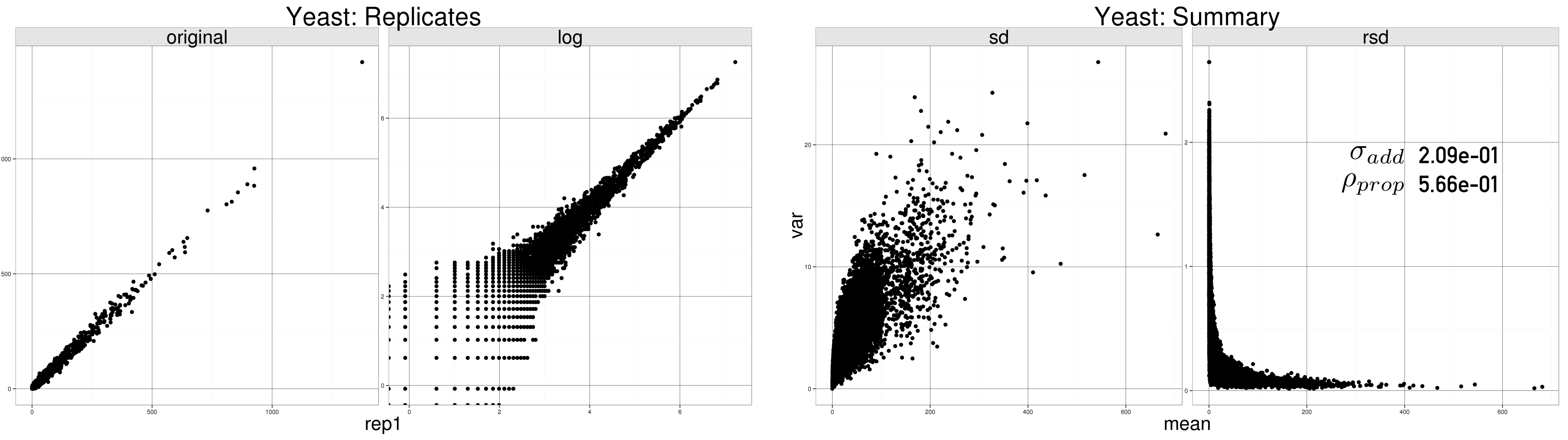
ii) DNA Microarray Quality Control Consortium Data

Affymetrix® HGU-133-plus2.0 cel files from the MicroArray Quality Control v1 (MAQC) project [4] downloaded from GEO (GSE5350). Each sample was processed by a combination of: RMA background correction [5]; constant normalization; perfect match only correction; average summarization. 54675 probe-sets across 120 samples from six sites, four samples at each site, with five replicates for each sample. 5000 probe-sets were selected randomly, and mean and standard deviation calculated across five replicates, and then pooled for visualization and estimation.



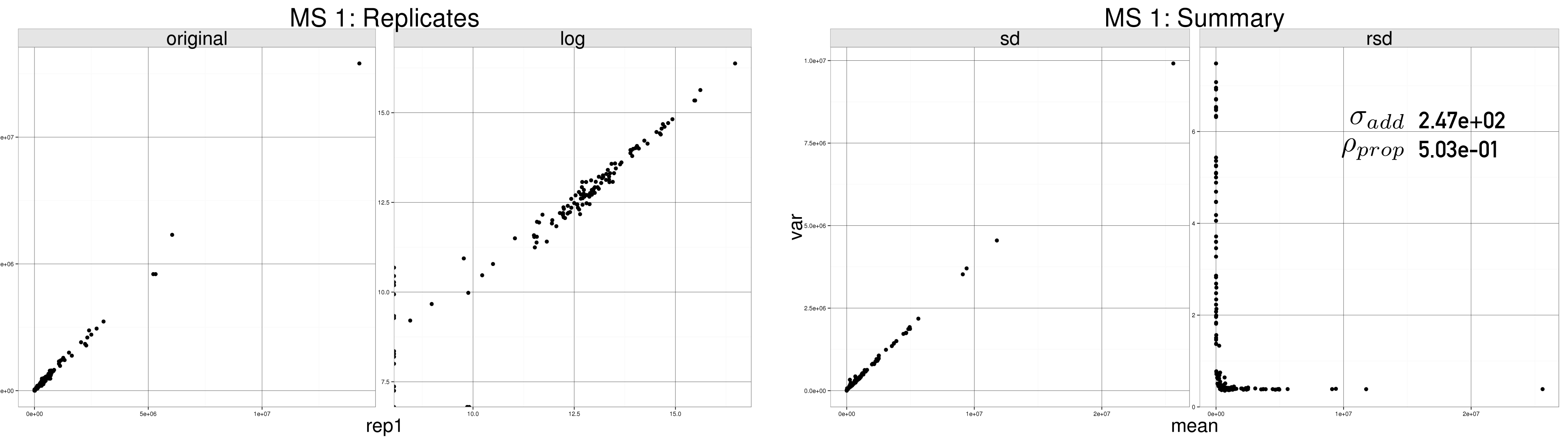
iii) RNA-Seq Yeast 7 Technical Replicates

RNA-seq transcriptomics data from 48 WT and 48 mutant Saccharomyces samples (SRA ERP004763) [6]. Each sample was barcoded, and all samples run on 7 lanes to generate 7 replicate runs for each sample (642 runs total). Sequences were aligned using TopHat [7] to the Saccharomyces reference genome (sacCer3). Exons were tiled into 100 base segments, and the number of reads aligning to each segment returned, for 94234 segments. Counts were normalized by the total number of reads in each sample. A sample of 5000 segments was chosen randomly, means and standard deviations were calculated across the seven replicates, and subsequently pooled for visualizaton and estimation.



iv) FTMS Lipids 57 Technical Replicates (scans)

Fourier-transform positive mode mass spectrometry data from one RCSIRM workshop sample of extracted lipids (2014). Peaks were determined using multi-spectral wavelets [8], and 500 null peaks added from interpeak regions. Individual scans from the run are considered as technical replicates. Means and standard deviations were calculated across the 57 scans.



Conclusions and Future Work

Based on the synthetic data with defined additve and proportional error components, the visualization and estimation accurately captures the presence and values of the error components. Applying the method to a variety of -omics datasets demonstrates that each of the different technologies contains proportional error to different degrees, observing the most effect in NGS platforms and microarrays, and the least in FTMS.

Future work will include applying the method to different levels of summarization in each of the -omics datatypes to examine how the error components vary at different summary levels, determining how many replicates are required to accurately estimate the error components. If the error components can be estimated using bootstrap sampling on small sample numbers, we should be able to examine large numbers of datasets to see how additive and proportional error components change with other measures of dataset quality.

Table 1. Raw and normalized to mean signal additive and proportional error component values for each of the -omics datasets.

	σ_{add}	ρ_{prop}	$\bar{x}_{all}$	$\sigma_{add}/\bar{x}_{all}$	$\rho_{prop}/\bar{x}_{all}$
dros	6.76e-03	2.95e-01	2.65e+01	2.55e-04	1.11e-02
maqc	2.77e-01	1.48e-01	1.73e+02	1.60e-03	8.55e-04
yeast	2.08e-01	5.69e-01	2.28e+00	9.12e-02	2.49e-01
ftms	2.47e+02	5.03e-01	2.37e+06	1.04e-04	2.12e-07

μ	Actual values
σ_{add}	Additive error component
ρ_{prop}	Proportional error component
sd_{rep}	Standard deviation across replicates
rsd_{rep}	Relative standard deviation across replicates
$\bar{x}_{rep}$	Mean across replicates
$\bar{x}_{all}$	Mean of mean values

This work was supported by funds from NIEHS 1R01ES022191-01, 1U24DK097215-01A1, NCI 1P01CA163223-01A1, NSF 1252893

1: http://www.dpgp.org/dpgp2/DPGP2.html
2: Langmead & Salzberg, Nat Methods 2012,
3: Lawrence et al., PLOS Comput Biol 2013
4: MAQC Consortium, Nat Biotechnol 2006
5: Irizarry et al., Biostatistics 2003
6: Gierliński et al., arXiv:1505.00588v1 [q-bio.GN]
7: Kim et al., Genome BIol 2013
8: Du et al., Bioinformatics 2006