Statistics
in Medicine

Research Article

Received 1 March 2016, Accepted 28 October 2016 Published online 24 November 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7175

A note on statistical repeatability and study design for high-throughput assays

George Nicholson*† and Chris Holmes

Characterizing the technical precision of measurements is a necessary stage in the planning of experiments and in the formal sample size calculation for optimal design. Instruments that measure multiple analytes simultaneously, such as in high-throughput assays arising in biomedical research, pose particular challenges from a statistical perspective. The current most popular method for assessing precision of high-throughput assays is by scatterplotting data from technical replicates. Here, we question the statistical rationale of this approach from both an empirical and theoretical perspective, illustrating our discussion using four example data sets from different genomic platforms. We demonstrate that such scatterplots convey little statistical information of relevance and are potentially highly misleading. We present an alternative framework for assessing the precision of high-throughput assays and planning biomedical experiments. Our methods are based on *repeatability*—a long-established statistical quantity also known as the intraclass correlation coefficient. We provide guidance and software for estimation and visualization of repeatability of high-throughput assays, and for its incorporation into study design. © 2016 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: high-throughput assay; technical replicate; scatterplot; study design; repeatability

1. Introduction

In the post-genome era, assays such as sequencing technologies and microarrays have underpinned major advances in biomedical genetics and form key components of recent large-scale projects in medical science, such as the Precision Medicine Initiative [1] and the 100 000 Genomes Project [2]. In recent years, the number of analytes measurable in a single experiment has increased dramatically, broadening the scope of scientific studies while raising new questions on the reproducibility of their conclusions [3–6]. While there has been extensive work on post-experimental statistical procedures for controlling false discovery rates [6–8], little guidance exists on how to assess the precision of multivariate assays and incorporate this into experimental study design and the planning of experiments. Here, we critically review the current standard practice of quantifying assay performance, which is to calculate the sample correlation of measurements across a pair of multivariate technical replicates [9–15]. We highlight important flaws in this approach and present an alternative framework based on statistical repeatability (also known as the intraclass correlation coefficient), for communicating assay precision and for integrating it into the planning of high-throughput experiments [16].

In their influential work on measuring the agreement between two medical instruments [17–19], Bland and Altman (BA) challenged the convention of scatterplotting the *univariate* data of one instrument against the other, that is, one point per patient, and of interpreting high correlation as indicating agreement between instruments. Our work can be thought of as extending these existing ideas of correlation and repeatability to a high-throughput multivariate-measurement setting, where a single instrument is used to measure multiple analytes on a set of individuals. Moreover, we pay particular attention to the issue of optimal experimental design for high-throughput assays.

Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, U.K.
**Correspondence to: George Nicholson, Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, U.K.*
†E-mail: nicholso@stats.ox.ac.uk

2. Correlation between repeated measures as an indication of assay precision.

A common means of reporting the precision of a high-throughput (also known as multiplex or high-content) assay in the literature is to compare a pair of *technical replicates*, such as those obtained by splitting a biological sample into two aliquots, and analysing each aliquot separately on the assay. The two technical replicates, each comprising measurements from multiple analytes, are plotted against each other, one point per analyte, and the corresponding sample correlation coefficient, r, is reported as a measure of experimental precision; see for example [9–15]. As illustration, Figure 1a–d displays this method applied to a pair of replicates from each of four representative high-throughput assays [20–23].

The intuition behind these plots is simple: a 'high-precision assay' has little variation in repeated measurements on the same sample, a property that is represented graphically by points lying close to the diagonal $x = y$ line, and statistically by large inter-replicate sample correlation of $r \approx 1$. This intuition is

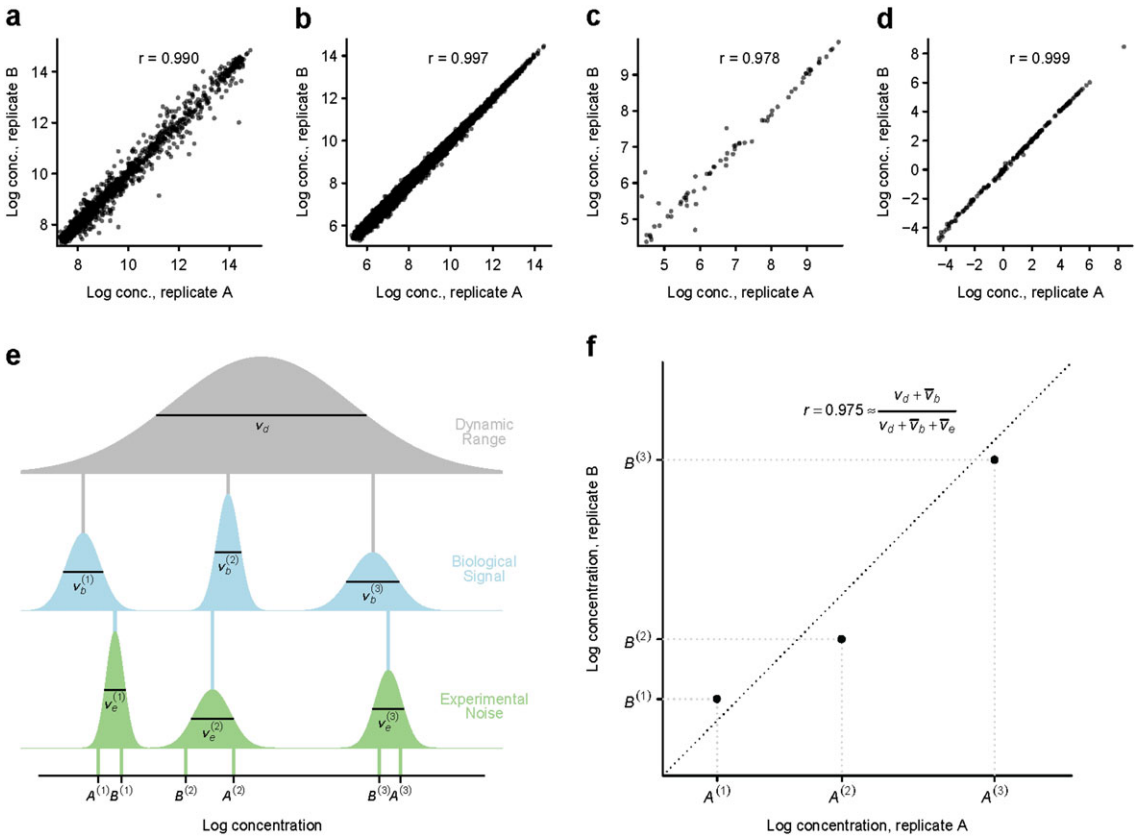

Figure 1. Scatter plots of technical replicates—examples and underlying statistical model. **(a–d)** Scatter plots of measured log concentrations from two technical replicates on each of four high-throughput assays (Table I). Each point displays the two replicate measurements of a particular analyte's concentration. Pearson's sample correlation coefficient, r, is shown. One pair of replicates was chosen at random from each data set (distribution of r across all pairs is shown in Figure S1). **(e)** Sources of variation underlying a pair of technical replicates. The grey bell-shaped distribution represents variation in concentration across analytes, spanning the entire dynamic range of the assay with *dynamic-range* variance v_d. Three analytes, labelled 1–3, are drawn from this distribution, and their population-mean concentrations are represented by vertical grey lines. The blue distributions represent variation in concentration across a population of individuals around the population's mean, represented by analyte-specific *biological signal* variances $v_b^{(1)}, v_b^{(2)}, v_b^{(3)}$, with the average of these biological variances across analytes denoted by $\bar{v}_b = \frac{1}{3}\left(v_b^{(1)} + v_b^{(2)} + v_b^{(3)}\right)$. A particular individual's concentrations at the three analytes, represented by vertical blue lines, are drawn from these distributions. The green distributions represent measurement error around the individual's true concentrations, with analyte-specific *experimental noise* variances $v_e^{(1)}, v_e^{(2)}, v_e^{(3)}$, with the average of these experimental variances across analytes denoted by $\bar{v}_e = \frac{1}{3}\left(v_e^{(1)} + v_e^{(2)} + v_e^{(3)}\right)$. A pair of technical replicates, A and B, with data labelled $(A^{(1)}, A^{(2)}, A^{(3)})$ and $(B^{(1)}, B^{(2)}, B^{(3)})$, are drawn from the green distributions and shown at the base of the plot. **(f)** Scatter plot comparing the technical replicates' data from **e**.

correct, in that extremely precise assays necessarily result in $r \approx 1$. However, the commonly employed argument that an assay exhibiting $r \approx 1$ implies an extremely precise measurement is, somewhat unintuitively, false. The reason is that the assay's dynamic range across analytes is confounded with r when considered as a measurement of experimental precision.

3. Statistical analysis using a variance components model

To understand better the phenomenon described, it is helpful to consider a multilevel statistical model for the data. We utilize a model to decompose the variation underlying concentrations of the p analytes measured in technical replicate on each of several biological samples as

$$y_{ij}^{(k)} = \mu + a^{(k)} + b_i^{(k)} + e_{ij}^{(k)}, \tag{1}$$

where $y_{ij}^{(k)}$ is the measured concentration of the kth analyte in the jth replicate of the ith biological sample, and μ is the global mean concentration. The $a^{(k)}$, $b_i^{(k)}$ and $e_{ij}^{(k)}$ are independent zero-mean random variables contributing components of variance, with $v_d \equiv V\left(a^{(k)}\right)$ as the *dynamic range* variance in concentration across analytes; $v_b^{(k)} \equiv V\left(b_i^{(k)}\right)$ as the *biological signal* variance across individuals at the kth analyte; and $v_e^{(k)} \equiv V\left(e_{ij}^{(k)}\right)$ as the *experimental noise* variance at the kth analyte.

Using the variance-component model, we are then able to relate the empirical sample correlation r to physical sources of variation. In particular, we are led to the following result,

Proposition 1

$$r \xrightarrow{Pr} \frac{v_d + \bar{v}_b}{v_d + \bar{v}_b + \bar{v}_e} \tag{2}$$

where $\xrightarrow{Pr}$ denotes convergence in probability as the number of analytes measured $p \to \infty$, and where $\bar{v}_b = \frac{1}{p}\sum_k v_b^{(k)}$, $\bar{v}_e = \frac{1}{p}\sum_k v_e^{(k)}$. The proof is contained in Supporting Information Appendix A.

To examine the finite-sample behaviour of (2), we performed a re-sampling study of the four data sets, concluding that r converges to within 1% of its final value by $p \approx 100$ (data not shown). Formula (2) reveals that r is close to 1 whenever the average noise term $\bar{v}_e$ is small relative to the sum of the dynamic range and average signal terms $v_d + \bar{v}_b$. In particular, to attain high correlation, it is not necessary for the assay's noise to be small relative to its signal, provided its noise is small relative to its dynamic range. This effect is illustrated in Figure 1e,f, where the noise variances $v_e^{(k)}$ are small relative to the dynamic range v_d, leading to high-sample correlation of $r = 0.975$, despite the noise $v_e^{(k)}$ and signal $v_b^{(k)}$ being of comparable size.

Returning to the four data sets [20–23] introduced in Figure 1a–d, we estimated their corresponding variance components directly on each full set of data (Table I). We found each assay's average noise variance $\bar{v}_e$ to be of a similar magnitude to its signal $\bar{v}_b$, but two to three orders of magnitude smaller than its dynamic range v_d. This demonstrates empirically that these assays exhibit considerable levels of noise (relative to biological signal $\bar{v}_b$) while achieving high inter-replicate correlation, as in Figure 1a–d, because their dynamic range is wide. Our advice is to avoid scatterplotting or calculating r between pairs of technical replicates, as such tools provide little statistical information on quantities of interest when correctly interpreted, and can be severely misleading when misinterpreted.

4. Repeatability of high-throughput assays and its use in study design

Instead, we suggest an approach for characterizing the precision of high-throughput assays, and for integrating that information into the planning of well powered experiments. Our recommendation is based on the repeatability, a long-established statistical quantity, also known as the intraclass correlation coefficient, reviewed in [24]. The repeatability at analyte k is defined as

$$R^{(k)} := \frac{v_b^{(k)}}{v_b^{(k)} + v_e^{(k)}} \tag{3}$$

Table I. Assay details, sample size and estimated components of variance.

Assay	Target	Number of analytes	Number of samples (**Number replicated**)	**Estimated variance** (95% CI)				Estimated $R^{(k)}$
				v_d	$\bar{v}_b$	$\bar{v}_e$	r	**Median** (IQR) across k
a	microRNAs	1624	69 (**69**)	**3.57** (3.54–3.59)	**0.05** (0.04–0.05)	**0.07** (0.06–0.07)	.990	**0.31** (0.10–0.53)
b	mRNAs	17 788	76 (**15**)	**1.95** (1.95–1.96)	**0.03** (0.02–0.03)	**0.01** (0.01–0.01)	.997	**0.59** (0.24–0.80)
c	Proteins	69	215 (**45**)	**2.47** (2.45–2.50)	**0.02** (0.02–0.03)	**0.06** (0.05–0.07)	.978	**0.31** (0.20–0.50)
d	Metabolites	163	287 (**67**)	**8.17** (8.13–8.21)	**0.04** (0.04–0.05)	**0.01** (0.01–0.01)	.999	**0.94** (0.82–0.96)

where the analyte's biological signal variance $v_b^{(k)}$ and experimental noise variance $v_e^{(k)}$ are defined in Figure 1e and its legend, and at the beginning of Section 3. The repeatability is a quantity in the interval $[0, 1]$ that records the proportion of total observed variance at an analyte that is attributable to biological sources. At the upper end of the scale, $R^{(k)} = 1$ indicates that analyte k is measured perfectly with $v_e^{(k)} = 0$ while, at the lower end, $R^{(k)} \approx 0$ signifies data that are dominated by experimental variability with $v_e^{(k)} \gg v_b^{(k)}$.

Analyte repeatabilities can be estimated directly under a standard pilot study that incorporates technical replicates (pilot design recommendations are provided in the Appendix). Potential estimation methods include analysis of variance (ANOVA), maximum likelihood and restricted maximum likelihood [24, 25]. Here, we choose ANOVA-based estimators because they are available in closed form, leading to computationally efficient implementation of the parametric bootstrap [26] used to calculate confidence intervals (Figure 2 bottom panels; Supporting Information Appendix B). ANOVA estimators for variance parameters can take negative values. In particular, it is possible that $\hat{v}_b^{(k)} < 0$, while it is known that $v_b^{(k)} \geqslant 0$. We set negative variance estimates to zero, leading to upwards bias but a net decrease in mean-squared error ([25], their Section 4.4).

Bland and Altman (BA) proposed the calculation of the 'repeatability coefficient' for a single instrument [18]. BA's repeatability coefficient ($R_{BA} \equiv 1.96\sqrt{2}\hat{v}_e$ in our notation) provides a 95% one-sided upper bound for the absolute difference between a pair of replicate readings on the instrument. R_{BA}, being on the same scale as the instrument itself, has the advantage of allowing simple clinical assessment of true biological changes [18, 27], but does not incorporate information on the biological variation across subjects, v_b. The repeatability as defined at (3) (i.e. the intraclass correlation coefficient, ICC) is a dimensionless quantity targeting the proportion of variation in an instrument's measurements that arises from non-experimental sources. We advocate the ICC for the purposes of assessing the repeatability of a high-throughput assay, for it is advantageous to have a measure of repeatability that is both scale-free (allowing direct pooling of information across analytes) and that incorporates v_b, which, together with v_e, is necessary for considerations of experimental design.

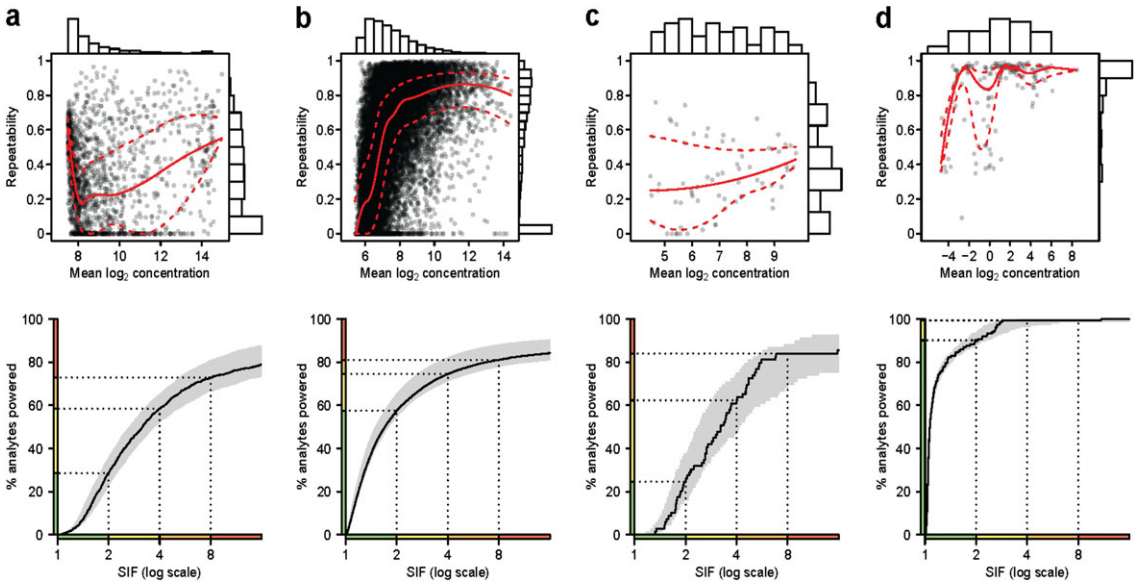

Figure 2. Proposed graphical representations of assay precision. **(a–d)** Repeatability versus concentration scatter plot (top) and plot of cumulative % of analytes powered (bottom), for four high-throughput assays (Table I). **Top panels**: Scatter plot of repeatability R against mean measured $\log_2$ concentration (one point per analyte). To visualize dependence of repeatability on concentration, median (red solid line) and quartiles (red dashed lines) of repeatability are plotted as a smooth function of concentration. The histogram at right shows the distribution of R across analytes, and the histogram at top shows the distribution of mean measured $\log_2$ concentration across analytes. **Bottom panels**: the black line shows the effect of increasing the sample size inflation factor, SIF, on the % of analytes powered to detect an effect. Grey-shaded regions are 95% bootstrap confidence intervals for the black line (details in the Supporting Information Appendix C). Intervals on the horizontal axis are coloured according to SIF and are mapped to the vertical axis for reference.

It is often the case that measurement precision shows a relationship with analyte concentration; for example, it can be relatively difficult to measure the abundance of low-concentration analytes. We recommend a scatter plot of estimated repeatability at each analyte against that analyte's average measured concentration to highlight any association (Figure 2, top panels). The distribution of repeatability estimates is visualized effectively as a histogram, as on the right edge of the top plots in Figure 2. Distributional summaries, such as median and inter-quartile range (Table I final columns), can be usefully reported when space is limited, although these particular statistics do not summarize the data distribution effectively in all cases; for example, they are not good summaries of assay b's bimodal repeatability distribution (Figure 2b top panel).

4.1. Illustrations and sample-size calculation

To illustrate the application of repeatability to study design, we first consider a sample size calculation for an experiment performed using a perfect instrument, and then show how that sample size should be increased on the basis of repeatability to ensure power is attained in the presence of measurement error.

Consider an experiment aimed at identifying differences in analyte concentration between treatment and control groups. Let μ_T denote the true underlying mean for the treatment group, and μ_C the true mean for the control group. To calculate sample size requirements, the key quantity to specify is the standardized effect size, $\Delta \equiv \frac{|\mu_T - \mu_C|}{\sqrt{v_b}}$, that is, the absolute difference between groups in units of the biological standard deviation $\sqrt{v_b}$. For a simple example, consider a user-specified targeted effect size of $\Delta = 1$, with power required to be 80% at a false-positive rate of 0.05. The resulting calculation indicates that $n_0 = 34$ participants are required, 17 in each group, to be powered to detect the specified effect on a perfect instrument (see [16] for a useful introduction to power and sample size).

In practice, instead of having a perfect instrument with repeatability 1, each analyte k on an assay is actually measured with its own particular non-zero measurement error $v_e^{(k)} > 0$ and hence repeatability $R^{(k)} < 1$. The experimenter might choose a single sample size n that applies to all analytes on the assay. It is intuitively desirable to choose n larger than the sample size for a perfect instrument, n_0, to compensate for measurement error being present. One way of characterizing the increase in chosen sample size relative to that of a perfect instrument is the ratio n/n_0 which we define as the *sample size inflation factor* (SIF),

$$\text{SIF} := \frac{n}{n_0} \equiv \frac{\text{sample size required for assay with measurement error}}{\text{sample size required by perfect instrument}} .$$

The distribution of repeatabilities across an assay provides a framework for informed choice of SIF. In particular, we are able to state the following result.

Proposition 2
The experiment is well powered to detect changes in the expected value of analyte k if

$$\text{SIF} > \frac{1}{R^{(k)}} .$$

The proof is given in the Supporting Information Appendix C.

Proposition 2 provides a basis for taking the sample size required by a perfect instrument (n_0) and inflating it to a sample size suitable for an assay with measurement error (n), so that the experiment is powered at a specified proportion of analytes. Our proposed protocol for the design of a high-throughput experiment aimed at detecting mean differences in analyte concentration between two groups is thus as follows.

(1) Estimate $R^{(k)}$ at analytes $k = 1, \ldots, p$, based on data from a pilot experiment with samples assayed in technical replicate.
(2) Select SIF large enough so that a user-specified proportion of analytes on the assay satisfy $\text{SIF} > 1/R^{(k)}$ and are hence powered. In practice, this step is best performed with reference to plots and tables based on assay-wide repeatability estimates such as Figure 2 bottom panels, and Table II.

Table II. Percentage of analytes powered for different SIF values.

	Percentage of analytes powered (95% CI)			
SIF	Assay a	Assay b	Assay c	Assay d
1.1	**1** (0–1)	**10** (9–15)	**0** (0–0)	**60** (56–64)
1.5	**11** (8–16)	**42** (39–49)	**3** (3–13)	**83** (82–86)
2	**29** (23–35)	**58** (55–65)	**25** (14–39)	**90** (88–94)
3	**48** (42–54)	**69** (67–77)	**45** (38–62)	**99** (93–99)
4	**58** (53–64)	**75** (72–82)	**62** (49–74)	**99** (96–100)
5	**65** (59–70)	**77** (75–85)	**74** (57–81)	**99** (97–100)

(3) Specify the experiment's targeted standardized effect size Δ, nominal significance level α, and power, and use them to calculate the sample size, n_0, required by a perfect instrument.[‡]
(4) Calculate the adjusted sample size as $n = \text{SIF} \times n_0$

Software in R for estimating and visualizing assay-wide repeatabilities (as per Figure 2 and Table II) from data sets with technical replicates is freely available on request.

Hence, as SIF is increased, the % of analytes that are powered increases accordingly. By quantifying and inspecting this relationship (Figure 2, bottom panels; Table II), the user can control the % of analytes at which an experiment is powered by varying SIF. For assays a, b, c, and d to be powered at approximately 60% of analytes, suitable SIFs would be 4, 2, 4, and 1.1 respectively (Table II), translating into sample sizes of 136, 68, 136 and 38 when applied to the sample-size calculation above with $n_0 = 34$. When designing a study, in addition to reporting n_0 and its calculation based on Δ, α and power, we suggest reporting the selected SIF and adjusted sample size n, along with the corresponding point estimate and confidence interval for the % of analytes powered (Table II).

It is natural to consider SIF as a form of variance inflation factor. VIFs measure collinearity amongst explanatory variables in multiple linear regression, reflecting the multiplicative increase in $V(\hat{\beta}_j)$ due to non-zero correlations between x_j and the other covariates [30]. VIFs can also be used to inflate sample sizes calculated under basic two-group designs so that they apply to more complex design settings [31]. At analyte k, the VIF

$$\frac{1}{R^{(k)}} \equiv \frac{v_b^{(k)} + v_e^{(k)}}{v_b^{(k)}} \tag{4}$$

is the multiplicative increase in $V(\hat{\beta}_j)$ (for all j) for the model $\boldsymbol{y} \sim N\left(\boldsymbol{X\beta}, (v_b^{(k)} + v_e^{(k)})\boldsymbol{I}\right)$ relative to the model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, v_b^{(k)}\boldsymbol{I})$, with Proposition 2 demonstrating that this VIF can be used to inflate sample size appropriately in the balanced two-group setting.

5. Conclusion

In conclusion, when designing high-throughput experiments, it is important to quantify those aspects of assay precision that relate directly to the study objectives. We have shown empirical and theoretical evidence that the standard approach of communicating assay precision—via correlation and scatterplotting of data from technical replicates—provides little statistical information at best and is often misleading. We have presented alternative statistical methods based on the notion of analyte repeatability, quantifying the information in an assay relative to a perfect instrument and providing a framework for adjusting sample size accordingly.

Appendix A

This appendix contains guidance on the design of pilot studies aimed at estimating repeatability and also more practical guidance on choosing SIF for main studies.

[‡]*Step 3 can be performed using any standard power software, such as* G*Power [28] *or the function* `power.t.test()` *in* `R` [29]. *Note that if statistical tests are to be performed at each of a large number of analytes then the specified significance level α should be correspondingly more stringent. For example, Bonferroni adjustment could be used to control the family-wise error rate across all analytes tested.*

A.1. Sample size for a pilot study

For choice of sample size under model (1), our suggestion is to focus on achieving effective estimation of the distribution of repeatabilities across all analytes, as opposed to the repeatability for any particular analyte. This is because it is typically unknown in advance which of the assayed analytes will be of eventual interest, and so it is natural to plan experiments based on the whole set. Also, a relatively large number—of the order of hundreds—of replicated samples is required to obtain precise repeatability estimates for individual analytes ([32], their Figure 3).

To assess what sample size is sufficient for estimating the distribution of repeatabilities, we repeatedly randomly sub-sampled and re-analysed each of the four example data sets. Each sub-data set comprised a number of samples assayed in technical duplicate, denoted by $D \in \{3, 6, 9, 12\}$, and a number of samples assayed only once, denoted by $S \in \{0, 6, 12, 18, 24\}$. The resulting plots of cumulative % of analytes powered are shown in the Figures S2–5. The feature of interest in these plots is the reduction in width of confidence interval with increasing sample size. It appears possible to reduce technical replication in the pilot study to quite a low level, for example just three replicated samples, provided that an adequate number of assays is conducted in total. Our suggestion is to perform at least 20 assays in the pilot study, with at least three samples assayed in technical duplicate (in the above notation, $D \geqslant 3$ with $2D+S \geqslant 20$).

A.2. Choice of SIF for a main study

In choosing suitable SIF, it is important to take into account the confidence intervals (CIs) for % of analytes powered, as shown in Figure 2 (bottom panels) and Table II. It is especially important in cases where the CIs are wide, for example when only a small number of pairs of replicates is assayed in the pilot study (Figures S2–5). If it is essential that a minimum % of analytes is powered, then SIF can be selected to be large enough that the lower bound of the CI exceeds the required %.

For a study in which a particular subset of analytes is of primary interest (e.g. measurements related to genes in a particular pathway), the SIF can be chosen to ensure that some proportion p_1 of the subset is powered, while a different proportion p_2 of all analytes on the array is powered. Creating such a design would involve applying our methods twice, once to the subset and once to the global set of analytes. SIF would be chosen to be the maximum of SIF_1 and SIF_2, where SIF_1 powers p_1 of the subset, and SIF_2 powers p_2 of all analytes.

Acknowledgements

The authors would like to thank Rory Bowden, Tristan Gray-Davies, Davis McCarthy, Matti Pirinen, Chris Spencer, Aimee Taylor, James Watson and Quin Wills for helpful comments on the paper and software. Chris Holmes wishes to acknowledge support from the EPSRC, ilike programme grant EP/K014463/1, and the Medical Research Council Programme Leaders award MC_UP_A390_1107.

References

1. Collins FS, Varmus H. A new initiative on precision medicine. *The New England Journal of Medicine* 2015; **372**(9): 793–795.
2. Genomics England. The 100,000 Genomes Project 2015. http://www.genomicsengland.co.uk/the-100000-genomes-project/ [Accessed on 20 February 2016].
3. Ioannidis JPA. Why most published research findings are false. *PLoS Medicine* 2005; **2**(8):0696–0701.
4. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V. Repeatability of published microarray gene expression analyses. *Nature Genetics* 2009; **41**(2):149–155.
5. Leek J.T, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* 2010; **11**(10):733–739.
6. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* 2011; **5**(3):1752–1779.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; **57**(1):289–300.
8. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(3):479–498.
9. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TMM, Bao W, Fang H, Kawasaki ES, Hager J, Tikhonova IR, Walker SJ, Zhang L, Hurban P, de Longueville F, Fuscoe JC, Tong W, Shi L, Wolfinger RD. Performance comparison of one-color and two-color platforms within the microarray quality control (MAQC) project. *Nature Biotechnology* 2006; **24**(9):1140–1150.

10. Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM, Hurban P, Phillips KL, Xu J, Deng X, Andrew Y, Tong W, Dragan YP, Shi L. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology* 2006; **24**(9):1162–1169.
11. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology* 2007; **25**(1):117–124.
12. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 2008; **5**(7):621–628.
13. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell PP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology* 2008; **26**(3):317–325.
14. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M. Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnology* 2009; **27**(7):652–658.
15. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R, Hugenholtz P. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* 2010; **7**(10): 807–812.
16. Krzywinski M, Altman N. Points of significance: power and sample size. *Nature Methods* 2013; **10**(12):1139–1140.
17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)* 1986; **1**(8476):307–310.
18. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; **8**(2):135–160.
19. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. *PLoS ONE* 2012; **7**(5):e37908.
20. Rantalainen M, Herrera BM, Nicholson G, Bowden R, Wills QF, Min JL, Neville MJ, Barrett A, Allen M, Rayner NW, Fleckner J, McCarthy MI, Zondervan KT, Karpe F, Holmes CC, Lindgren CM. MicroRNA expression in abdominal and gluteal adipose tissue is associated with mRNA expression levels and partly genetically driven. *PLoS ONE* 2011; **6**(11):e27338.
21. Min JL, Nicholson G, Halgrimsdottir I, Almstrup K, Petri A, Barrett A, Travers M, Rayner NW, Mägi R, Pettersson FH, Broxholme J, Neville MJ, Wills QF, Cheeseman J, The GIANT Consortium, The MolPAGE Consortium, Allen M, Holmes CC, Spector TD, Fleckner J, McCarthy MI, Karpe F, Lindgren CM, Zondervan KT. Coexpression network analysis in abdominal and gluteal adipose tissue reveals regulatory genetic loci for metabolic syndrome and related phenotypes. *PLoS Genetics* 2012; **8**(2):1–18.
22. Kato BS, Nicholson G, Neiman M, Rantalainen M, Holmes CC, Barrett A, Uhlén M, Nilsson P, Spector TD, Schwenk JM. Variance decomposition of protein profiles from antibody arrays using a longitudinal twin model. *Proteome Science* 2011; **9**:1–16.
23. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW, Toft H, Krestyaninova M, Viksna J, Neogi SG, Dumas ME, Sarkans U, Donnelly P, Illig T, Adamski J, Suhre K, Allen M, Zondervan KT, Spector TD, Nicholson JK, Lindon JC, Baunsgaard D, Holmes E, McCarthy MI, Holmes CC, The MolPAGE Consortium. A genome-wide metabolic QTL analysis in europeans implicates two loci shaped by recent positive selection. *PLoS Genetics* 2011; **7**(9):e1002270.
24. Nakagawa S, Schielzeth H. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society* 2010; **85**(4):935–956.
25. Searle SR, Casella G, Mcculloch CE. *Variance Components* 2nd. Wiley: Hoboken, New Jersey, 2006.
26. Davison AC, Hinkley DV. *Bootstrap Methods and their Application* 1st, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press: 32 Avenue of the Americas, New York, NY 10013-2473, USA, 1997.
27. Vaz S, Falkmer T, Passmore AE, Parsons R, Andreou P. The case for using the repeatability coefficient when calculating test–retest reliability. *PloS ONE* 2013; **8**(9):e73990.
28. Faul F, Erdfelder E, Lang AGG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 2007; **39**(2):175–191.
29. R DCT. *R: a language and environment for statistical computing*, R Foundation for Statistical Computing: Vienna, Austria, 2010. http://www.r-project.org [Accessed on 16 February 2016].
30. Fox J, Monette G. Generalized collinearity diagnostics. *Journal of the American Statistical Association* 1992; **87**(417): 178–183.
31. Hsieh FY, Lavori PW, Cohen HJ, Feussner JR. An overview of variance inflation factors for sample-size calculation. *Evaluation & The Health Professions* 2003; **26**(3):239–257.
32. Wolak ME, Fairbairn DJ, Paulsen YR. Guidelines for estimating repeatability. *Methods in Ecology and Evolution* 2012; **3**(1):129–137.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

