SUPPLEMENTARY APPENDIX

The evolution and adaptive potential of transcriptional variation in sticklebacks – signatures of selection and widespread heritability

Erica H. Leder,[1,4] R.J. Scott McCairns,[2,4,5] Tuomas Leinonen,[2] José M. Cano,[3] Heidi M. Viitaniemi,[1] Mikko Nikinmaa,[1] Craig R. Primmer,[1] and Juha Merilä[2]

[1]University of Turku
Department of Biology
Division of Genetics and Physiology
20014 TURKU
FINLAND

[2]University of Helsinki
Department of Biosciences
Ecological Genetics Research Unit
00014 HELSINKI
FINLAND

[3]University of Oviedo
Research Unit of Biodiversity (UO-CSIC-PA)
33600 MIERES
SPAIN

[4]These authors contributed equally to this work.

[5]Corresponding author
Scott McCairns
E-mail scott.mccairns@helsinki.fi

Supervised Normalization

The choice of normalization procedure is a critical step in the analysis of microarray data given that associated effects on the structure of the data can influence the validity/applicability of downstream statistical models essential for a study's objectives. Quantile normalization has become the most common method of normalization in microarray-based studies of gene expression. Whilst this has been shown to be highly satisfactory for standard comparisons of group means, such as those in case-control studies (Bolstad et al. 2003), this method is likely inappropriate for inference based on expression variances, such as the analyses which are at the heart of quantitative genetics. This is due to the fact that quantile normalization is based on equalizing the distributions of probe intensities across arrays; however, one side effect of this procedure is to imposes a uniform variance structure amongst probes (Qin et al. 2013). Yet differences in the variance amongst traits (i.e. probes/transcripts) and family groups are the parameters of interest in a quantitative genetic analysis – 'averaging' these out would be antithetical to the objectives of such a study. Supervised normalization has been proposed as an alternative method to statistically remove technical artefacts such as batch, array and dye effects (Leek et al. 2010; Mecham et al. 2010). Most importantly from a quantitative genetics perspective, supervised normalization has been shown to preserve the variance structure of expression data (Qin et al. 2013).

The procedure works by accounting for nuisance variables (e.g. array effects) as random covariates – removing them in a manner analogous to the calculation of best linear unbiased predictors conditioned on random effects – whilst leaving effects due to biologically meaningful variables unadulterated. For a formal mathematical expression of the procedure, see Mecham et al. (2010). To normalize liver mRNA transcription data for variance components analysis, we used the R/Bioconductor package 'snm' (Mecham et al. 2010). In this package, technical artefacts pertaining to each individual are coded as 'adjustment variables' in a model matrix: these included effects due to dye, array and batch (i.e. slide). Effects of interest to the study objectives are similarly coded as a design matrix of 'biological variables', with the normalization algorithm designed to leave any variation attributable to them intact. For the purpose of this study, 'biological variables' included thermal treatment, sex and family groupings – from the quantitative genetics perspective it is variation in the latter which is most essential for parameter estimation.

Evaluating effects of normalization on data structure

To evaluate the overall effect of supervised normalization on data structure, we first plotted the frequency distributions of mRNA quantitation, expressed as $\log_2$ of the fluorescence intensity, for both raw and normalized data (fig. 1). A visual comparison of the distributions suggests that supervised normalization did not impose any 'aggressive' alterations: their overall shapes are similar, and both are centered about the same range of intensity values. Likewise a scatterplot of raw and normalized expression shows that data largely follow a 1:1 relationship (fig. 2A).

To explore the relative magnitude of adjustments imposed by supervised normalization, we calculated the relative difference between pre- and post-normalization for each data point. For every data pair (pre- and post-normalization values) we expressed the absolute value of their difference, relative to their mean. Normalization affected a mean data shift of less than 10% for the vast majority of all data points (fig. 2B). We also explored average normalization effects for each individual sample. The average correlation between pre- and post-normalization data was

0.99, and none were observed less than 0.96 (fig. 2C), suggesting minimal adjustment effects of the normalization procedure. Likewise, the average adjustment effect was less than 5% for most individuals (fig. 2D). Taken together, these observations suggest that supervised normalization did not result in substantive alterations in expression data, and that subsequent inferences reflect patterns of heritability in expression, not statistical artefacts imposed on the data by normalization.

Evaluating co-hybridization

All mRNA extractions were treated with DNase and assayed for RNA quality and quantity, as well as cross-contamination with genomic DNA prior to labelling. The cRNA amplification and labeling process should preclude labelling of gDNA. Additionally, the custom array/probe design has been thoroughly validated (Leder et al. 2009), and most importantly, the array contains both negative and spike-in control features used to define a background level against which 'true' expression is discernible (Benes and Muckenthaler 2003). Thus, false signals due to co-hybridization with gDNA are extremely unlikely. However, differential signal intensity due to array (i.e. co-hybridized individual) and/or dye effects are potential sources of co-hybridization error, although supervised normalization is expected to remove these artefacts.

To evaluate signal differences due to alternate co-hybridizing 'partner' effects, 12 randomly selected individuals were replicated across different arrays, co-hybridized with a different individual on each array. Replicates were labelled with the same dye: six individuals were labelled with Cy3, and six with Cy5. Following normalization we compared the relative difference between replicate probes. Overall differences were, on average, less than 3% of mean intensity, though some outliers persisted (fig. 3A). Likewise, correlations between alternately labelled probes ranged from 0.986 to 0.996 (fig. 3B), suggesting little effect on signal intensity due to differing co-hybridization partners.

To evaluate the efficacy of removing dye effects through normalization, 14 randomly selected individuals were labelled with both Cy3 and Cy5. Following normalization we compared the relative difference between alternately labelled probes. Overall differences were generally less than 5% of mean intensity, though some outliers persisted (fig. 4A). Likewise, correlations between alternately labelled probes ranged from 0.959 to 0.994 (fig. 4B), suggesting that the majority of technical artefacts associated with dye chemistry were effectively removed by supervised normalization. Nevertheless, we also included dye as a fixed effect in subsequent analyses to statistically 'remove' any lingering dye effects from the estimation of genetic parameters. This was facilitated by the fact that assignment of dye labelling was conducted via blocked randomization: individuals were selected at random within each family-by-treatment block to ensure each family had equal numbers of individuals labelled with each dye.

As a final evaluation of the efficacy for supervised normalization to remove array effects (i.e. artefacts in measured intensity due to signal correlation with co-hybridization partner intensity), we explored patterns of sample clustering using all replicate individuals and their respective co-hybridization partners. Both samples and probes were sorted via hierarchical clustering using the 'Heatplus' package for R/Bioconductor (Ploner 2014). Clustering was performed on normalized data for all 14,955 probes prior to outlier removal and replicate averaging (fig. 5A). We reasoned that if co-hybridization artefacts were effectively removed, replicate individuals should cluster together, rather than co-hybridized partners. This was the pattern observed (fig. 5B). For example, the green asterisks flagging the co-hybridization partners for replicate individual 16 (P.16a & P.16b), are found on branches far from the clustered replicate samples (16a & 16; green brace). Likewise replicate 12b (fig. 5B, top red asterisk) clusters with the other replicate sample for individual 12 (12a & 12b_P.1a; blue brace), although it was hybridized with another replicated individual (1a_P.12b; top blue asterisk) – triplicates of this individual also cluster together (red brace), with their co-hybridization partners dispersed across the dendrogram, clustering with their own replicates (red asterisks). Taken together, patterns of replicate clustering (fig. 5), and high correlations between $\log_2$ intensity measurements from replicate samples (figs. 3&4), suggest that co-hybridization related artefacts have been effectively removed by normalization.

Outlier detection & removal

Although normalization appears to have been highly effective at removing most technical artefacts in the data, some outliers clearly persisted. Since outliers can bias variance component estimates (Gervini and Yohai 1998; Yuan and Bentler 2001; de Andrade et al. 2003), we used a systematic strategy for their detection and removal prior to final data analysis, focusing foremost on the removal of potentially problematic transcripts. We began by screening probes for individual $\log_2$ intensity values +/- 2 standard deviations from their family-by-treatment mean. This resulted in the removal of 208 individual data points (< 0.004% of total data). Next we examined the pair-wise correlations of 69 probes which had from 3-10 technical replicates each. All replicates with a correlation ≥0.9 were retained and averaged by transcript. In total, 67 transcripts were retained and 2 were deleted (r < 0.3). An additional 3,662 transcripts were represented as duplicates on the array. If the pair-wise difference of a given transcript was greater than 0.75 (i.e. >10% mean difference), those individual's data points were removed. We then removed any transcripts with a correlation < 0.9 (n = 154); individual averages were calculated for the retained transcripts.

After removal of problematic probes, the dataset was reduced to 10,711 transcripts, of which 10,303 (96%) contained no missing values. Of the 408 transcripts with missing values, 65 were missing 55 or more entries (i.e. >10% of individuals): these transcripts were also removed from the final dataset. An additional 119 transcripts were removed from the analysis due to missing values across entire family groups. The remaining 224 transcripts contained fewer than 10 missing entries, with no family-specific bias/concentration of missing values, and so were deemed suitable for inclusion in final analyses. Since the final dataset consists of less than 0.015% missing values (n = 839), it is highly unlikely that variance components estimation is biased by missing entries. The final dataset used in these analyses is available as a tab-delimited text file in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-3098 (Processed_QG_normalized.txt).

Normalization Between Experiments: Incorporating Among-Population Comparisons in mRNA Expression

For inference of selection via Q_{ST} to be valid, data should capture both genetic divergence among sampled populations, as well as within-population genetic variation (Leinonen et al. 2013). A second dataset comprising individuals from three populations, including the population used for quantitative genetic analyses, was used to infer the among-population component of genetic variance. This dataset, herein referred to as 'among-population' data, is well suited to the task given that each individual is representative of a unique, second generation family reared under identical laboratory conditions. As such, differences observed amongst population groups can be attributed to genetic, rather than environmental causes. However, this dataset alone is not sufficiently large for robust estimation of the within-population components of genetic variance (i.e. V_A). Consequently, we merged it with the larger dataset with greater power for estimation of additive genetic variance.

Although combining datasets provides an ideal solution for improved parameter estimation, it also introduces the potential for errors/artefacts in the form of batch effects. It is well known that batch effects can be a serious source of technical error, particularly for samples processed at different times or in different laboratories (Bammler et al. 2005; Leek et al. 2010). However, the supervised normalization strategy we employ has been shown to effectively remove signal artefacts associated with batch in both simulated and real data (Mecham et al. 2010).

Since not all probes present on the array used to generate the among-population dataset were present on the second array, we first removed all 'missing' probes from the quantitative genetic dataset. Next we merged the raw/non-normalized data from both sources into a single dataset containing only transcripts common to both. Data were then normalized as before, but with data source (i.e. data batch) included as the 'adjustment variable.' These data are also available via ArrayExpress (accession E-MTAB-3099; Processed_Qst_normalized.txt).

Statistical Inference: Evaluating Data & Model Assumptions

Although most modern statistical computing/analysis is robust to violations of normality assumptions, inference and estimation of variance components, the backbone of the analyses which comprise our work, can be sensitive to problems with kurtosis (DeCarlo 1997; Bonett and Seier 2002). To evaluate the degree to which data met this aspect of normality, we used the R package 'moments' to estimate Pearson's measure of kurtosis (k) for each transcript (Komsta and Novomestky 2012) – estimates were performed for each sex-by-treatment grouping, to capture the structure of downstream analyses, then averaged. Additionally, we tested whether kurtosis measured in each transcript differed significantly from normally distributed data using the Anscombe-Glynn test of kurtosis (Anscombe and Glynn 1983); significance was adjusted for multiple comparisons using a local false discovery rate.

Kurtosis estimates ranged from 2.8 to 6.8 for 95% of transcript data, with a median value of 3.6 (fig. 6A) – a value of 3 is expected for a perfectly normal dataset. Kurtosis estimates did not differ from normal expectation for the majority of transcripts (fig. 6B): fewer than 5% of transcripts (n = 506) exhibited significant leptokurtosis. As such, data largely conform to model assumptions.

Prior to modeling all 10,527 transcripts, we randomly selected 50 transcripts in order to optimize model control parameters (e.g. burn-in period; thinning interval). We began with default parameters, and tried different iterations of increasing values until trace files displayed proper mixing and model convergence. Additionally, we verified that estimates sampled from the Markov chain were sufficiently 'spaced' to avoid autocorrelation. For estimating quantitative genetic parameters, we determined that a burn-in of 50,000 iterations and a sampling interval of 200,000 iterations consistently produced good mixing and model convergence; sampling each 200th position of the Markov chain reduced autocorrelation of estimates. Models for the estimation of Q_{ST} required a longer burn-in (200,000 iterations); however, sampling parameters remained the same.

Phenotypic simulations – estimating power and false positives

We used the R package 'pedantics' to simulate phenotypic data for all individuals under the realized pedigree over a range of heritable and environmental/error variance (Morrissey et al. 2007). Simulations were conducted such that total phenotypic variance summed to one, with any variance not assigned to V_A being partitioned to residual (i.e. 'environmental' variation). We performed 100 simulations for each heritability value over the following iterative ranges: from 0 to 0.01 in 0.001 intervals; 0.01 to 0.1 in 0.01 intervals; 0.1 to 0.5 in 0.05 intervals; and 0.5 to 1 in 0.1 intervals. We conducted an additional 900 simulations under the scenario of zero heritable variation (n = 1,000 in total). Simulated data were analyzed under the same framework as the transcriptional data: DIC of a fully parameterized model was compared to a 'null' model to determine the significance of each estimate. Each simulated dataset was assigned a binomial score based on whether model selection was correct (e.g. 'null' model preference for simulated $h^2 = 0$). Statistical power was estimated via generalized linear model (binomial distribution; logit link function). Bias was estimated by modeling estimated h^2 as a function of 'true' (i.e. simulated) h^2, bounded by upper and lower 95% PDI estimates, and contrasted against an expectation of a 1:1 relationship for perfect estimation.

To estimate a 'local' false discovery rate, we calculated the proportion of correctly identified non-significant V_A estimates (n = 753) out of the 1,000 simulations for which heritability was set to zero. We then used this proportion to weight our estimate of the number of transcripts with significant additive genetic variance, and calculated the mean of these false estimates to identify values of h^2 for which significant estimates should be regarded as suspect. Unlike for V_A, the 'pedantics' package does not currently support the simulation of phenotypes with 'known' V_D. Consequently, we could not conduct a formal power analysis. However, we could estimate the 'local' false discovery rate by simulating over a range of 'known' h^2, with d^2 (V_D) implicitly zero. In 1,009 of 1,100 simulations ($h^2 = 0$ to 1 in 0.1 intervals; 100 simulations each), suggesting a high possibility of type I error. We calculated the mean of these false estimates to identify values of d^2 which should be regarded as suspect.

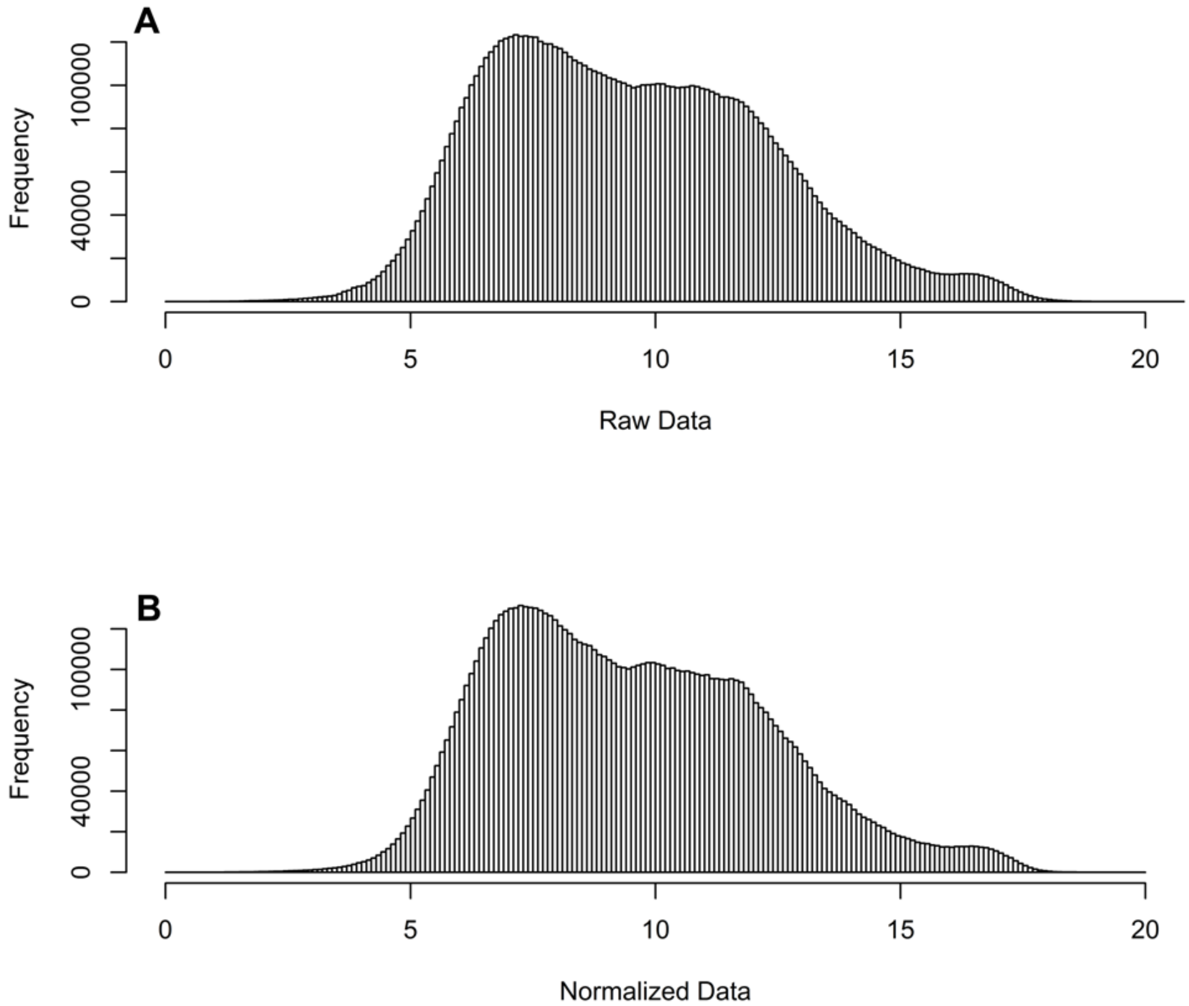

FIG. 1. Frequency distribution of $\log_2$ fluorescence intensity values before (**A**) and after (**B**) supervised normalization.

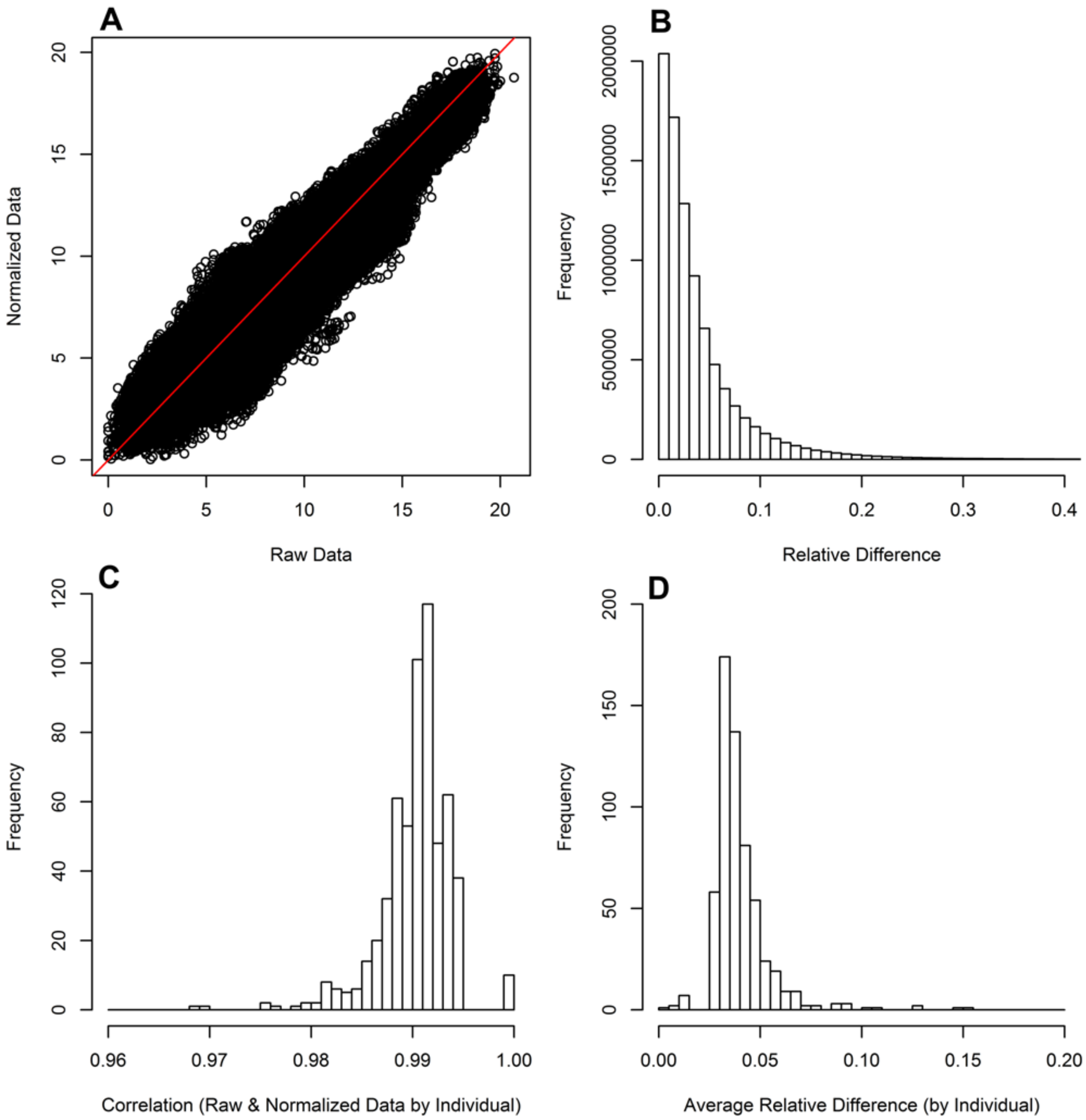

FIG. 2. Effects of supervised normalization. (**A**) Scatterplot of $\log_2$ fluorescence intensity values for all probes before (raw) and after normalization. The 1:1 line is plotted in red. (**B**) Frequency distribution of relative differences (absolute value) between pre- and post-normalization data. (**C**) Frequency distribution of correlation coefficients between pre- and post-normalized data. Correlations were calculated for each individual. (**D**) Distribution of per-individual average of relative differences (absolute value) between pre- and post-normalization data.

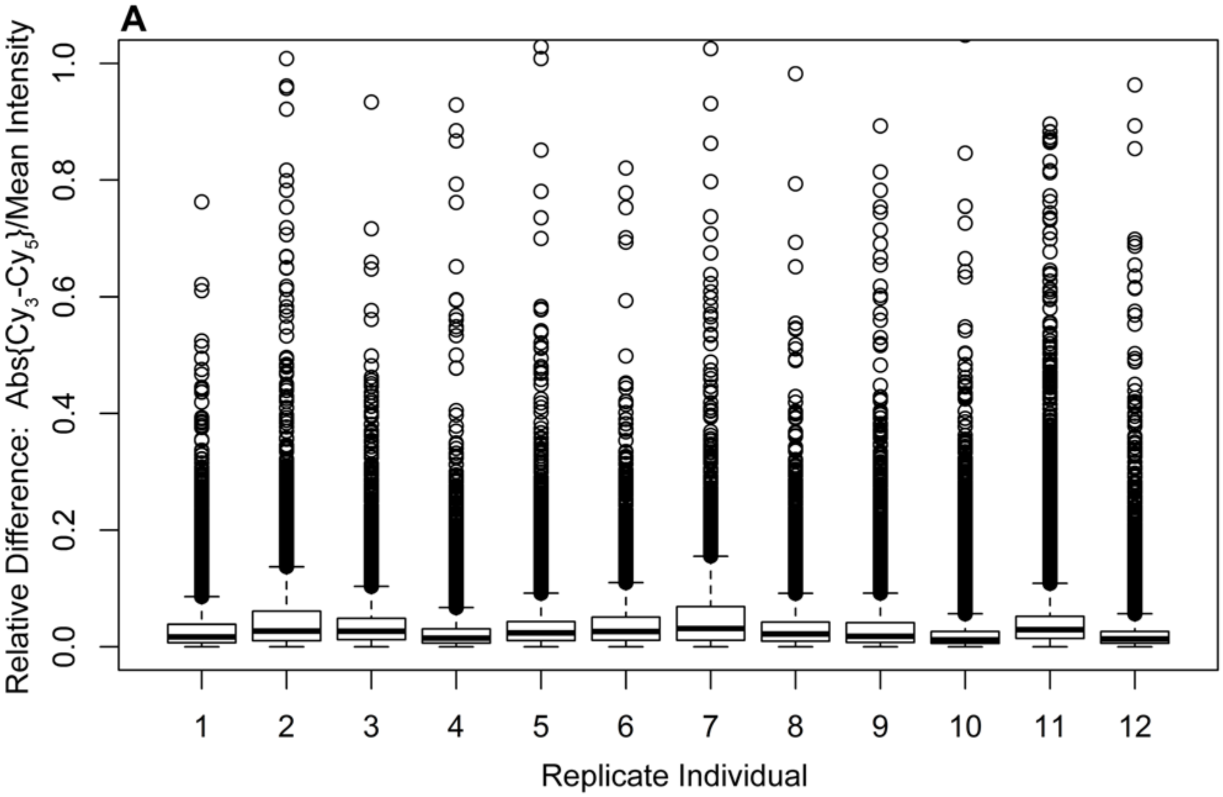

FIG. 3. Evaluating co-hybridization effects independent of dye effects. (**A**) Boxplot of relative differences in $\log_2$ fluorescence between replicate individuals co-hybridized with different individual samples across arrays – replicates are labeled using the same dye chemistry. Median difference is represented by the thick horizontal line within each box (quartile range); whiskers denote approximate 95% confidence intervals; outliers are shown as dots. (**B**) Scatterplot of $\log_2$ fluorescence intensity values for each of the 12 replicate individuals. Individuals one (B.1) to six (B.6) are labelled with Cy3; individuals seven (B.7) through twelve (B.12) are in Cy5. The 1:1 lines are plotted in red.

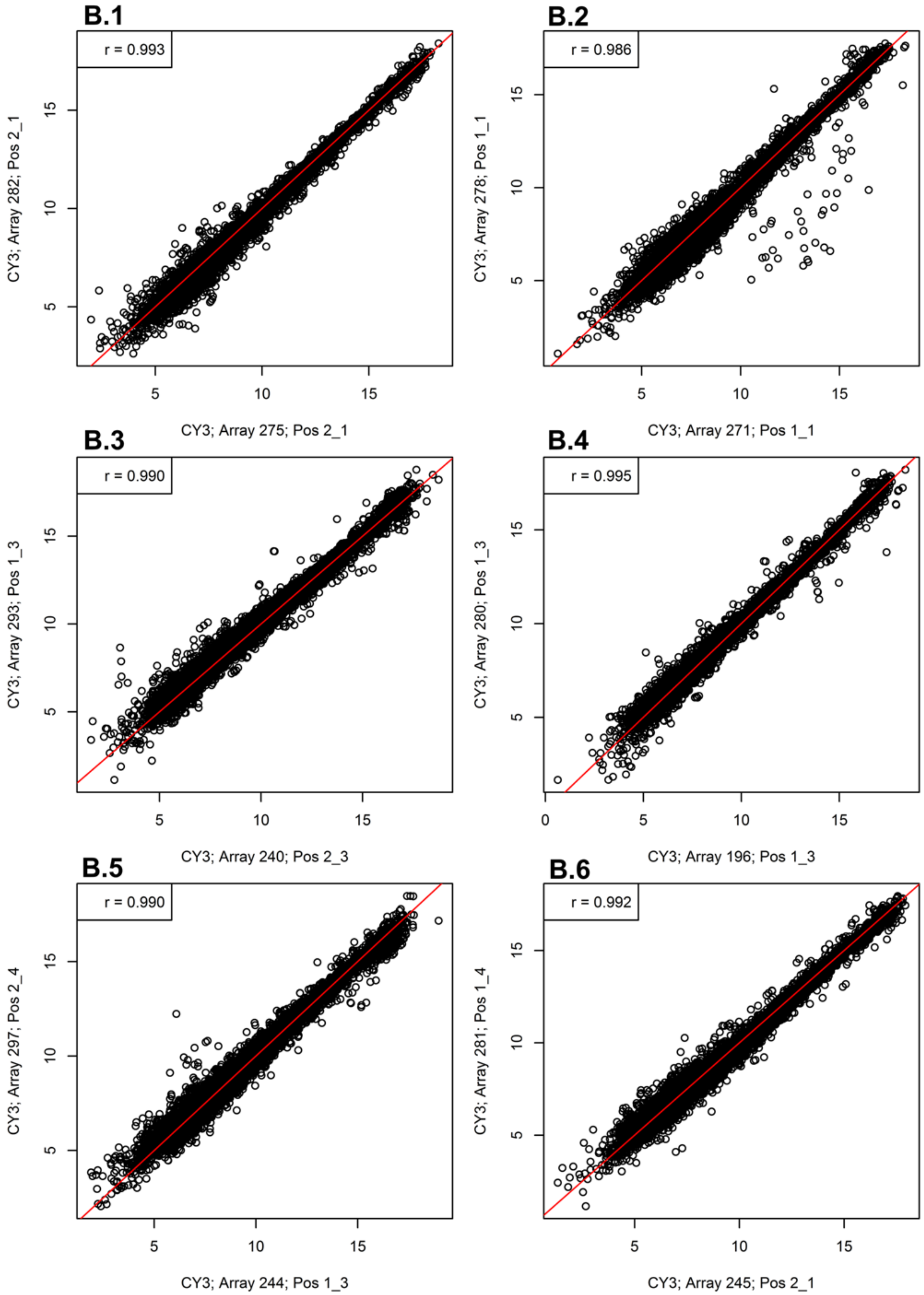

FIG. 3. *Continued*

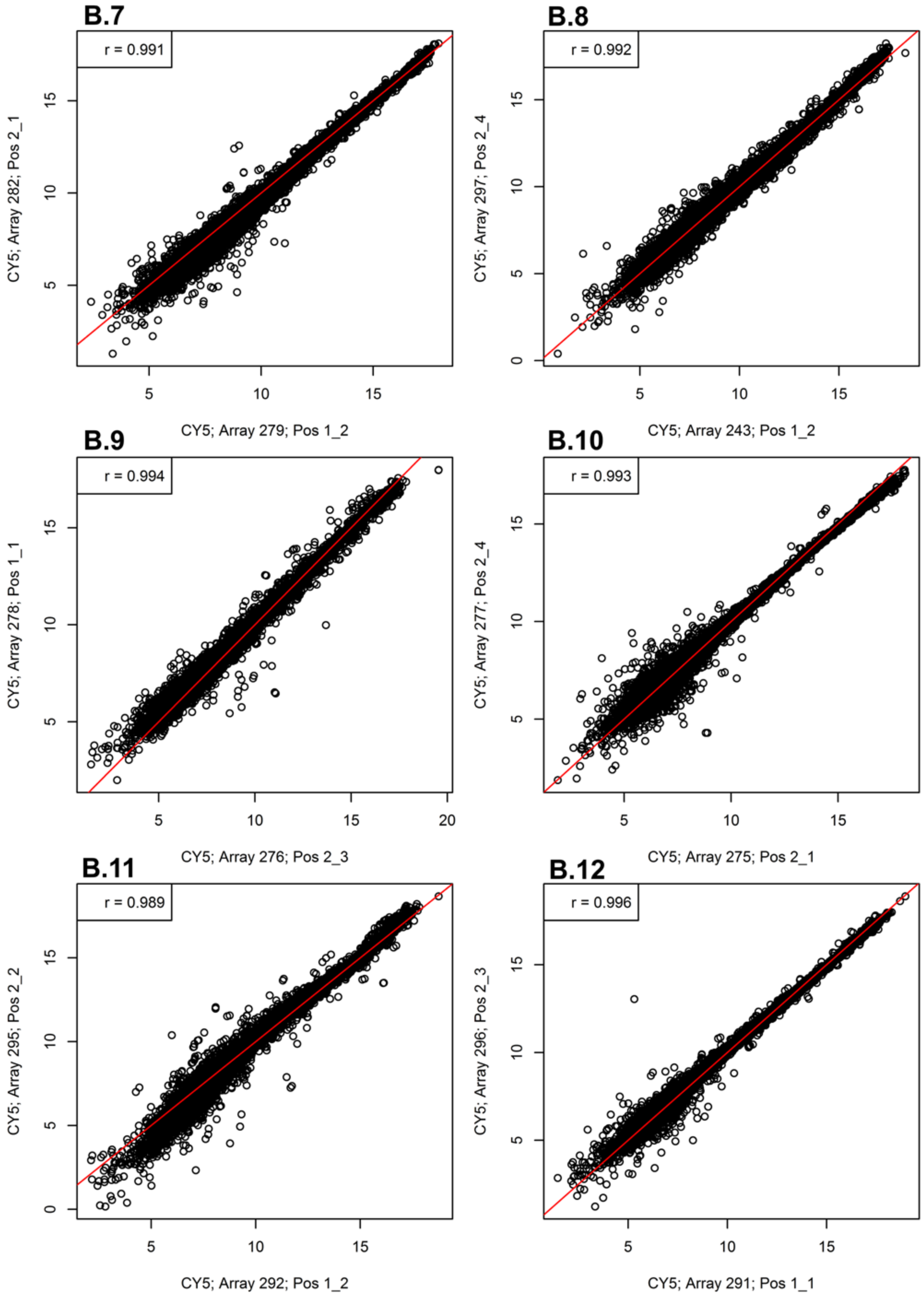

FIG. 3. *Continued*

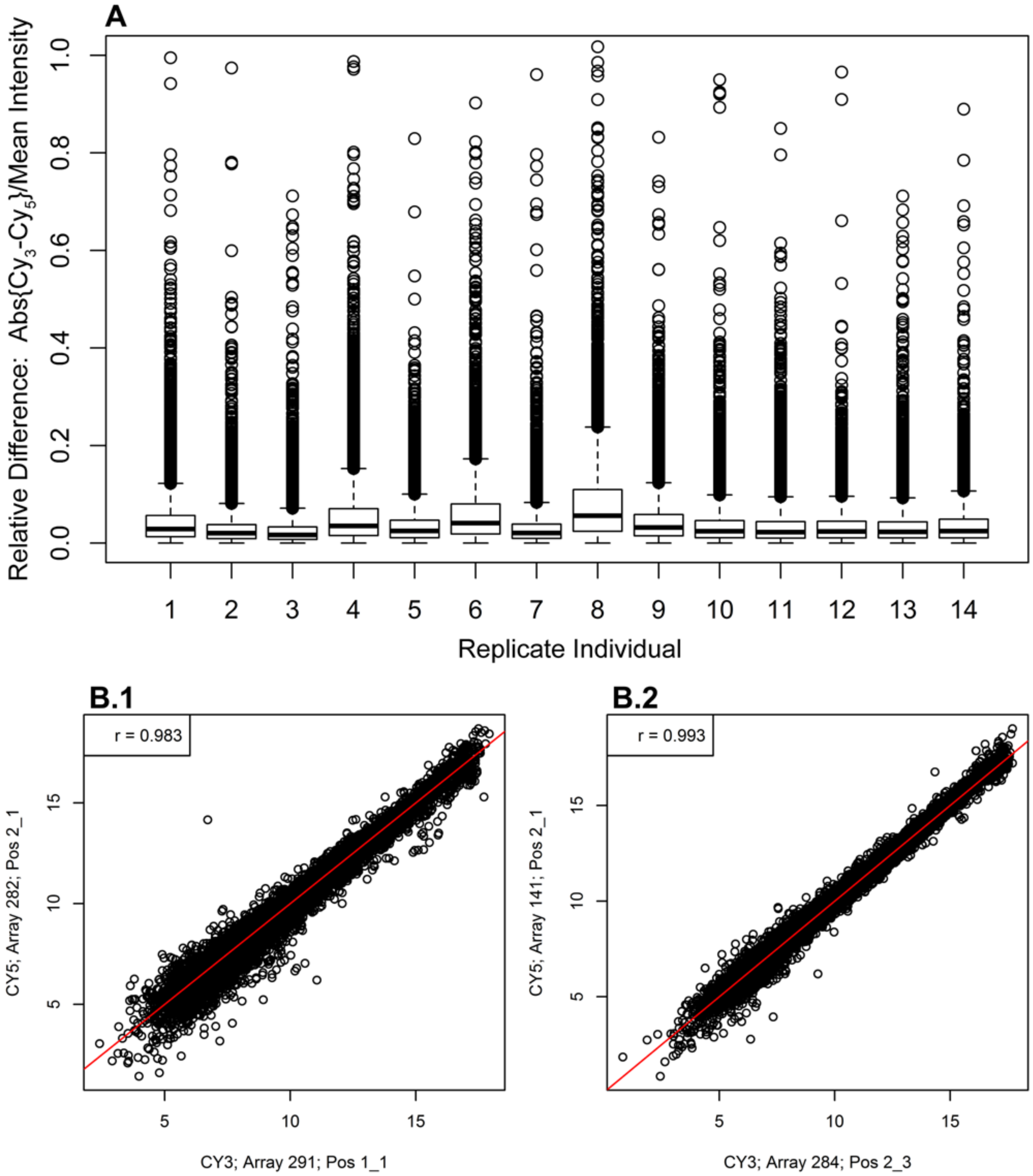

FIG. 4. Evaluating co-hybridization effects of differentially labelled probes post-normalization. (**A**) Boxplot of relative differences in $\log_2$ fluorescence between probes labelled with Cy3 and Cy5 for 14 dye-swapped individuals. Replicates are hybridized on different arrays. Median difference is represented by the thick horizontal line within each box (quartile range); whiskers denote approximate 95% confidence intervals; outliers are shown as dots. (**B**) Scatterplot of $\log_2$ fluorescence intensity values in each dye chemistry for 14 dye-swapped individuals. The 1:1 lines are plotted in red.

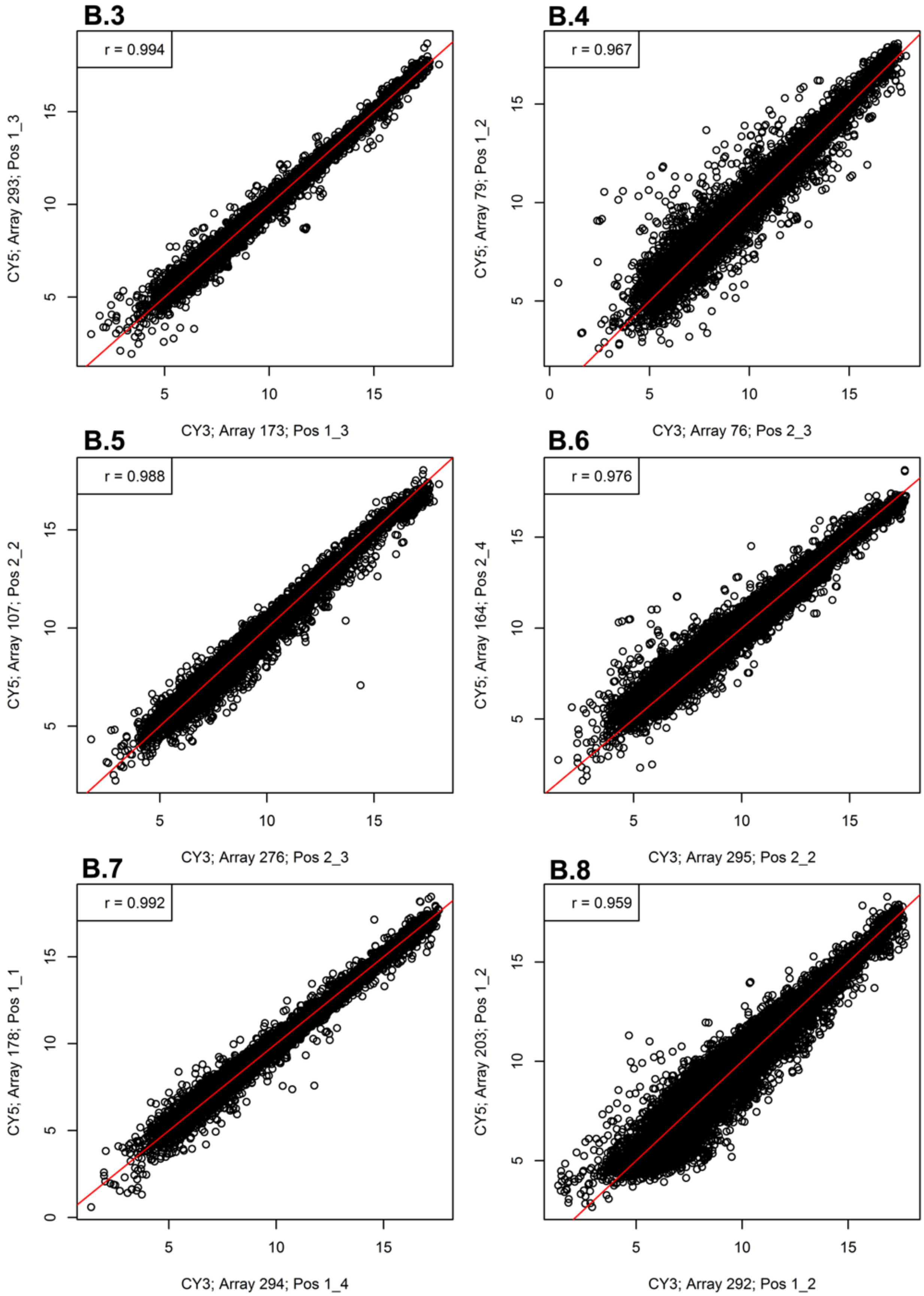

FIG. 4. *Continued*

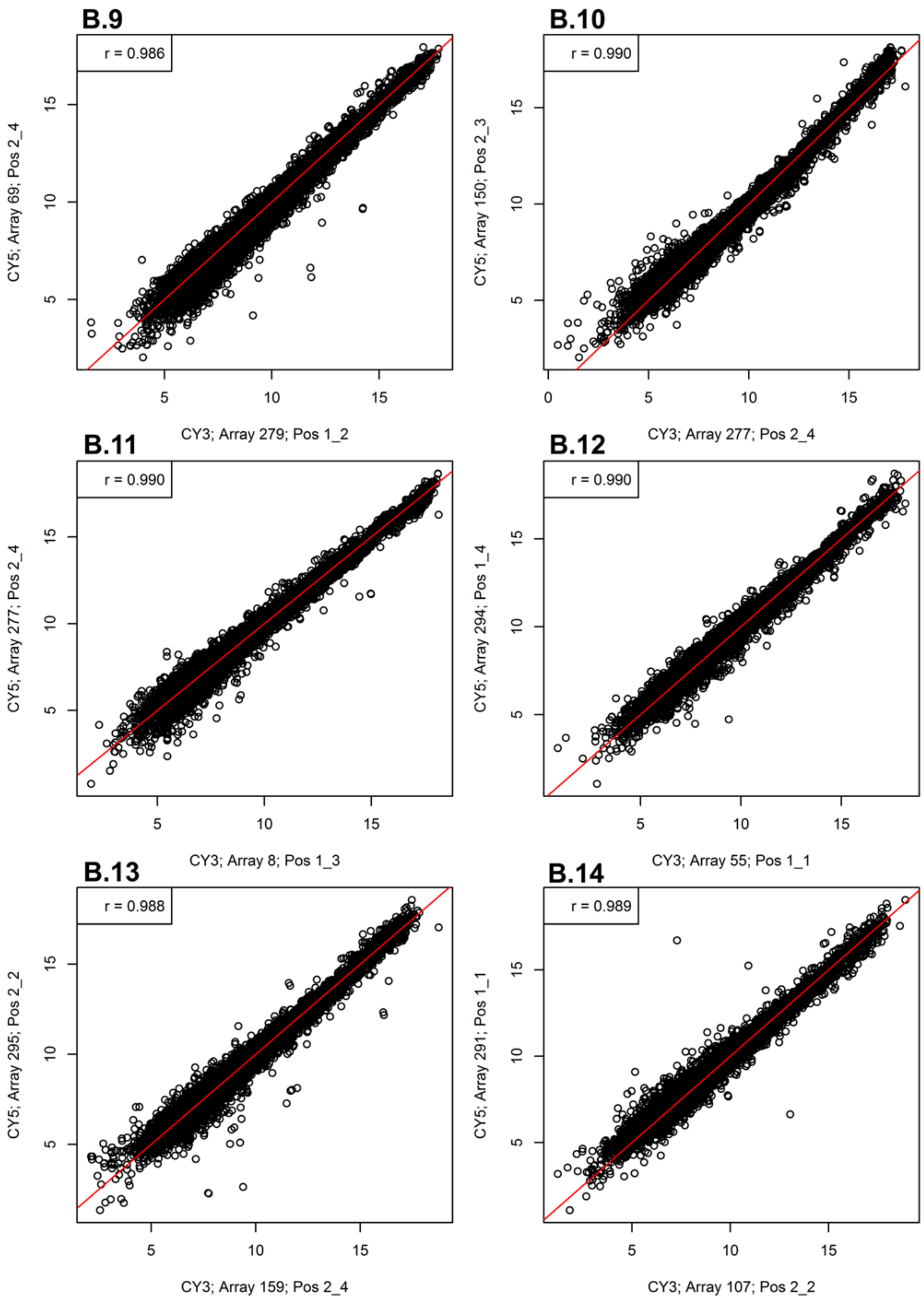

FIG. 4. *Continued*

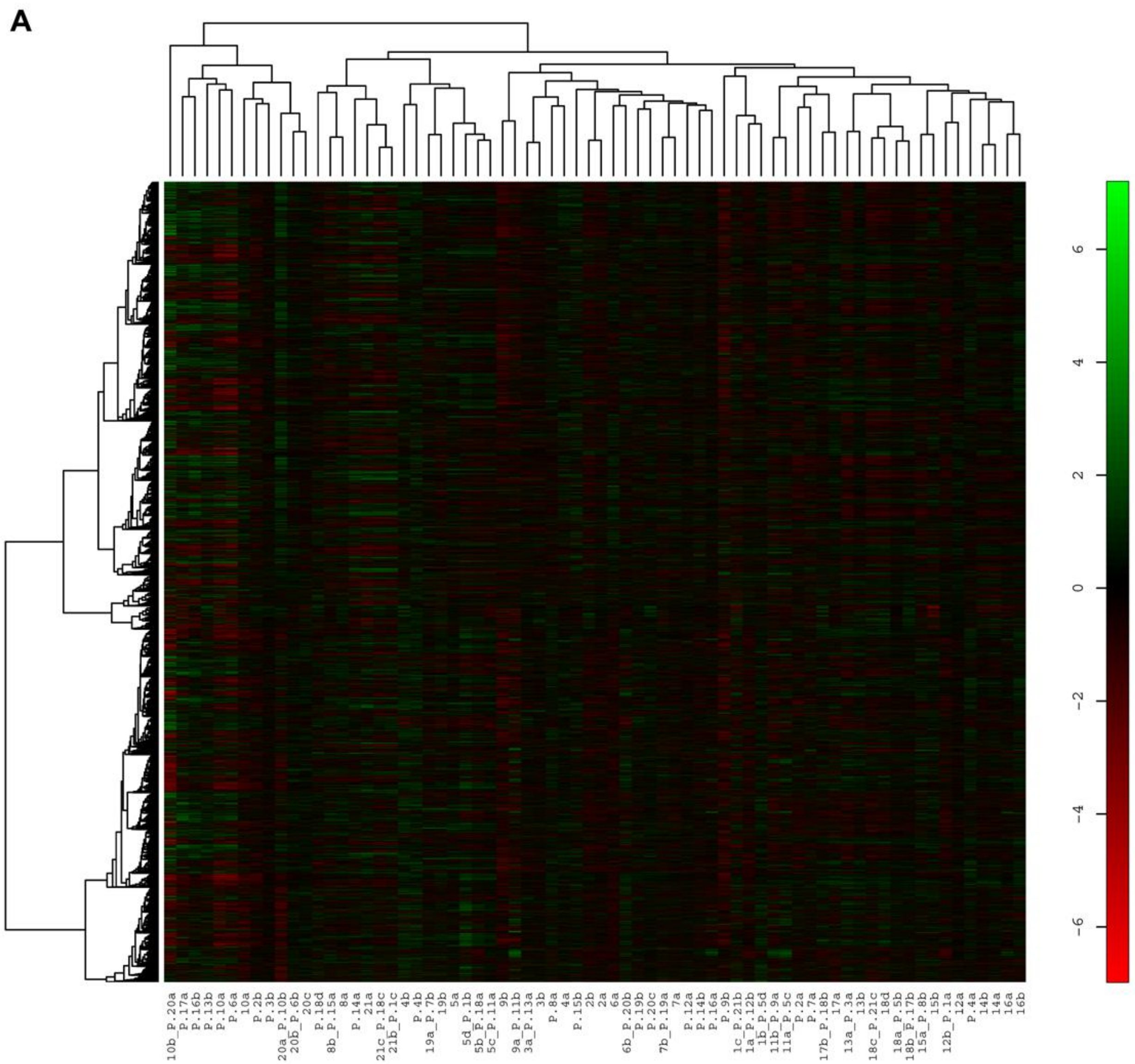

FIG. 5. Clustering of replicate individuals with their co-hybridization partners. (**A**) Heatmap showing individual/sample clusters (columns) based on overall patterns of similarity in $\log_2$ signal intensity for all 14,955 probes (rows). Data have been normalized, but outliers have not been removed, and replicate probes have not been averaged. (**B**) Expanded view of the sample dendrogram (top), rotated clockwise. Duplicates are numbered (e.g. 1a & 1b), and their co-hybridization partners are labeled in reference to the duplicate number (e.g. P.1a & P.1b). Three representative replicates are shown, highlighted with a coloured brace – co-hybridization partners are flagged with an asterisk of the same colour.

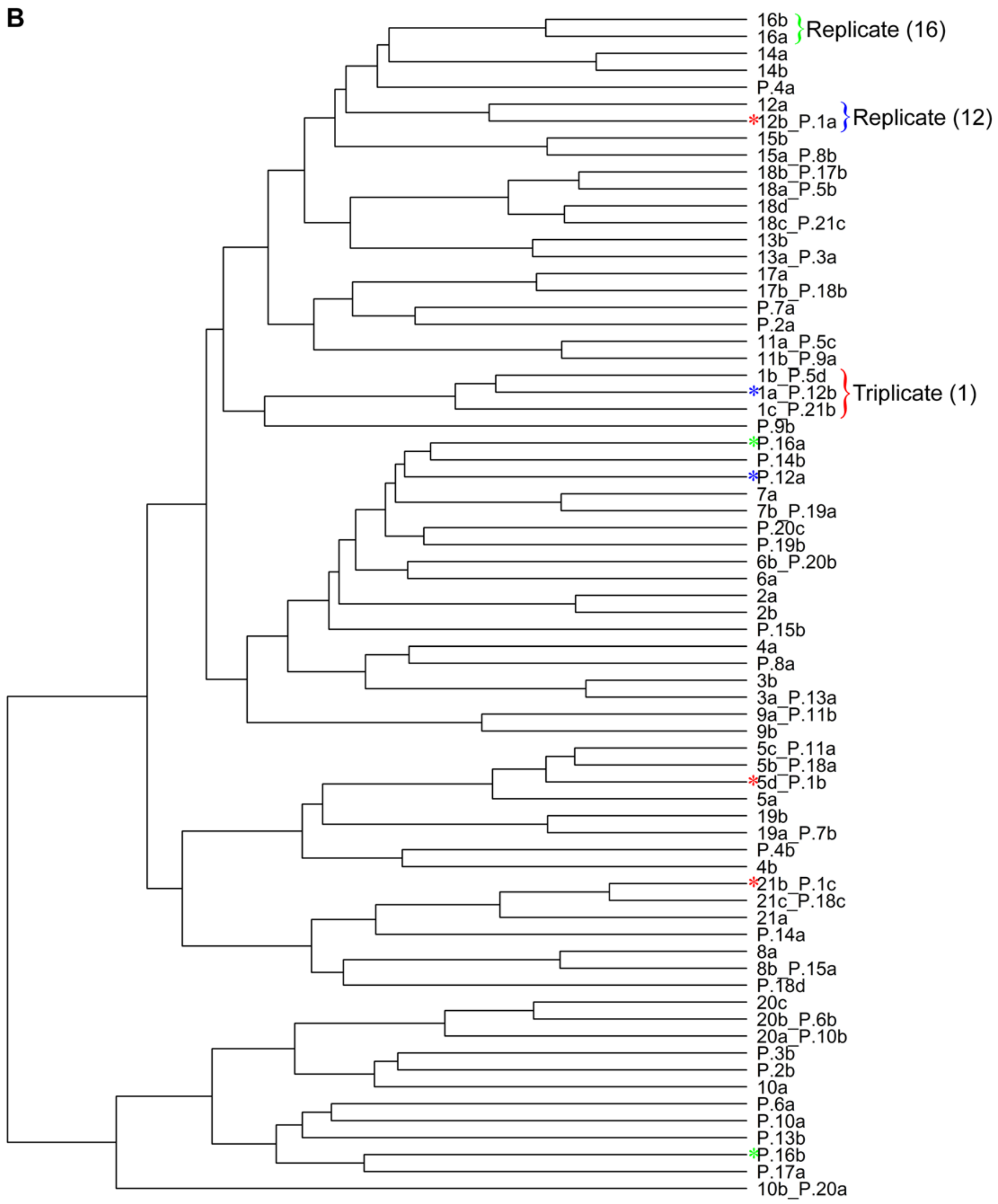

FIG. 5. *Continued*

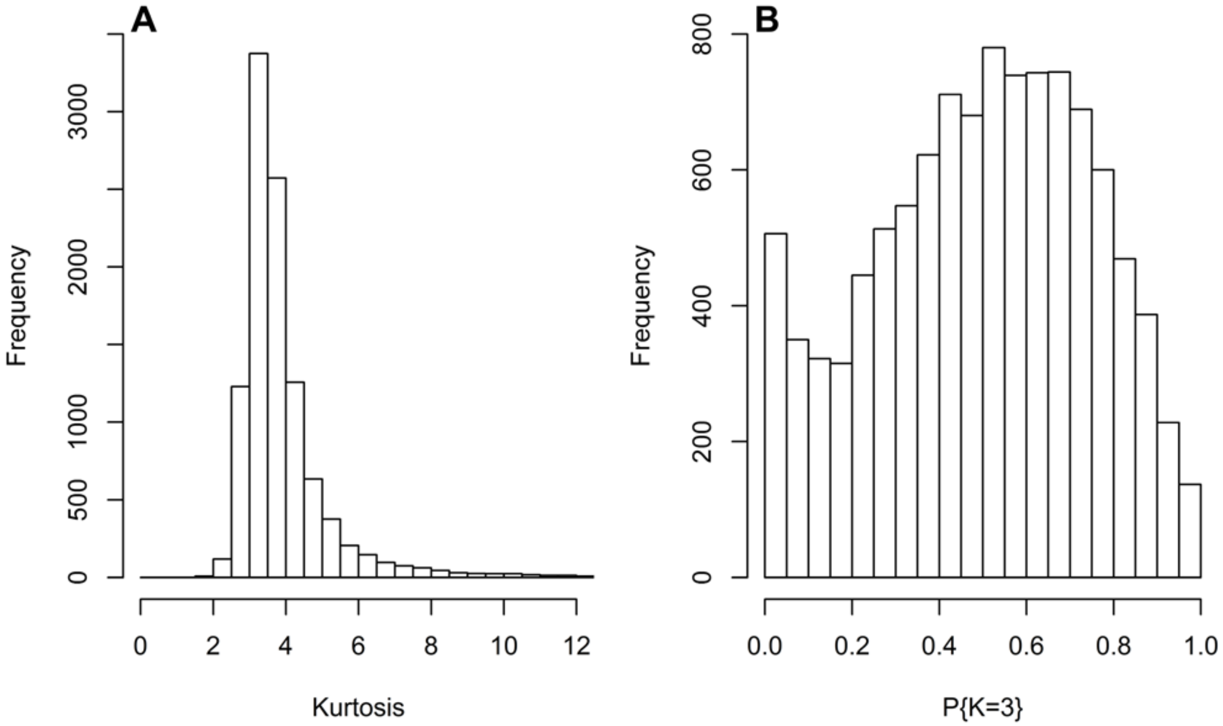

FIG. 6. Evaluating normality of $\log_2$ fluorescence intensity values for all probes. (**A**) Frequency distribution of kurtosis estimates for all transcripts. (**B**) Distribution of FDR-corrected p-values testing whether transcript kurtosis differed significantly from normal expectations (k = 3). Bars represent a bin width of 0.05; thus, the first bar corresponds to those transcripts which deviate from normality.

References

Anscombe F, Glynn WJ. 1983. Distribution of the kurtosis statistic for normal samples. *Biometrika* 70:227-234.

Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, et al. 2005. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2:351-356.

Benes V, Muckenthaler M. 2003. Standardization of protocols in cDNA microarray analysis. *Trends BiochemSci* 28:244-249.

Bolstad BM, Irizarry RA, Åstrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185-193.

Bonett DG, Seier E. 2002. A test of normality with high uniform power. *Comput Stat Data Anal* 40:435-445.

de Andrade M, Fridley B, Boerwinkle E, Turner S. 2003. Diagnostic tools in linkage analysis for quantitative traits. *Genet Epidemiol* 24:302-308.

DeCarlo LT. 1997. On the meaning and use of kurtosis. *Psychol Methods* 2:292-307.

Gervini D, Yohai VJ. 1998. Robust estimation of variance components. *Can J Stat* 26:419-430.

Komsta L, Novomestky F. 2012. *moments*: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.13: http://CRAN.R-project.org/package=moments.

Leder EH, Merilä J, Primmer CR. 2009. A flexible whole-genome microarray for transcriptomics in three-spine stickleback (*Gasterosteus aculeatus*). *BMC Genomics* 10:426.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733-739.

Leinonen T, McCairns RJS, O'Hara RB, Merilä J. 2013. Q_{ST}-F_{ST} comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nat Rev Genet* 14:179-190.

Mecham BH, Nelson PS, Storey JD. 2010. Supervised normalization of microarrays. *Bioinformatics* 26:1308-1315.

Morrissey MB, Wilson AJ, Pemberton JM, Ferguson MM. 2007. A framework for power and sensitivity analyses for quantitative genetic studies of natural populations, and case studies in Soay sheep (*Ovis aries*). *J Evol Biol* 20:2309-2321.

Ploner A. 2014. *Heatplus*: heatmaps with row and/or column covariates and colored clusters. R package version 2.10.0: http://www.bioconductor.org/packages/release/bioc/html/Heatplus.html.

Qin S, Kim J, Arafat D, Gibson G. 2013. Effect of normalization on statistical and biological interpretation of gene expression profiles. *Frontiers in Genetics* 3:160.

Yuan KH, Bentler PM. 2001. Effect of outliers on estimators and tests in covariance structure analysis. *Br J Math Stat Psychol* 54:161-175.