Tim Farrell, tmf@bu.edu
BF571: Dynamics + Evolution of Biological Networks
Project Report
April 30, 2016

Background and Motivation

It is becoming increasingly possible to compute genome-scale simulations of biological processes in both prokaryotic and eukaryotic systems (O'Brien and Palsson (2015), Thiele et al. (2013)). As these approaches become increasingly predictive of biological function, computational modeling will become highly useful in bioengineering and translational medicine (Zielinski et al. (2015)).

There are many approaches to reconstructing genome-scale cellular processes, and the scope of their techniques is expanding (O'Brien and Palsson (2015)). Among such approaches, a modeling approach (be it mechanistic, stoichiometric, contraint-based, etc.) that has the ability to incorporate high-throughput experimental data, to increase its predictive accuracy and to confer system specificity, would be most desirable. Such models are called metabolism and expression models, or ME-models, where the expression term refers to either protein or gene expression. ME-models that utilize gene expression data would be most useful due to the abundance of gene expression data and the higher throughput methods available for mRNA measurement.

However, a recent review of constraint-based ME-models that incorporate gene expression data found their predictive accuracy of metabolite levels were no better (and in some cases worse) than a stand-alone metabolic model (Machado and Herrgard (2014)). Although this performance may be due to the limitations of the modeling approaches, it is also well-known that gene expression is poorly correlated with protein expression in general. This presents a challenge to building computational models of biological function that incorporate experimental data at the mRNA level.

Here I attempt to replicate the findings of a recent paper that examined the relationship between gene and protein expression (Koussounadis et al. (2015)). The authors argue and demonstrate that although overall gene expression is poorly correlated with differential protein expression, differentially expressed genes show higher correllations with protein expression than non-differentially expressed genes. This result could provide practical insight into how gene expression data should be incorporated into ME-models.

In addition, I apply a weak causal measure, a time-series method called Granger causality (G-causality), to the gene and protein expression data in an attempt to build upon the paper's correlation analyses. Using an analogy to neuroinformatics analyses, interestingly G-causality is used to describe 'information flow' and has theoretical basis in information theory (Amblard and Michel (2012)). This methodology suggests an alternative approach to investigating the mRNA-protein relationship and to building mRNA-protein co-expression networks (Friston (2011)).

Methods

For the experimental methods used to obtain the data, as per Koussounadis et al. (2015) Methods,

> Briefly, two ovarian cancer tumour models, OV1002 and HOX42433, were implanted subcutaneously in the flanks of adult female nu/nu mice and allowed to grow to 4-6 mm

> in diameter. The mice received one of two drug treatments via intraperitoneal injection on day 0, carboplatin (50mg/kg) only or carboplatin (50mg/ kg) + paclitaxel (10mg/kg), or were left untreated as controls. Xenografts were harvested from treated mice on days 1, 2, 4, 7, and 14, and from untreated controls on days 0, 1, 2, 7, and 14.

For the data-preprocessing methods used, again as per Koussounadis et al. (2015) Methods,

> Raw mRNA expression data were background corrected, variance stabilised transformed (VST) and robust spline normalised (RSN) using Bioconductor's *lumi* package. AQUA pro- tein expression scores were log-transformed with base 2. For both mRNA and protein expression, log fold-change values for each time point in each drug treatment condition were calculated by comparing mean expression levels across biological replicates to pooled controls for that tumour model using the Bioconductor package *limma*. Both mRNA and protein expression exhibited similar dynamic ranges in log fold-change, from approximately –1 to 1. The output of limma was used to identify differentially expressed mRNAs, defined as those having FDR-adjusted p-values below 0.05. When evaluating varying FDR-cut offs, differentially expressed mRNAs were defined using FDR-adjusted p-values from 0.01 to 0.50 in steps of 0.01. The mRNA dataset has been deposited to Gene Expression Omnibus (GEO) with accession number GSE49577. The protein dataset (raw AQUA scores and limma-produced log-fold change values) is pro- vided in Supplementary Data 1.

The major difference between my analysis and the authors' (which likely somehow accounts for differences in results) is that, due to difficulties related to extracting the GSE/ GSMs as `R` `ExpressionSet` objects (via the `GEOquery` package), *lumi* log-2 rather than variance-stabilization transformation was applied during data processing. In addition, I also performed my analysis with and without *lumi* preprocessing to investigate its effects on my results.

For the Granger-causality (G-casuality) analysis, I first apply the Augmented Dickey-Fuller test to assess for each expression time series' stationarity, a property required for application of the Granger casuality test; and then perform Granger casuality tests between all mRNA-protein pairs for a given condition (e.g. "HOX424 CarboTax", "OV1002 Carbo", etc.). I use the `adf.test` and `grangertest` (from the `R` `lmtest` and `tseries` packages, respectively) for these tests.

Code, documents and data can be found here.

Results

The major results to replicate in Koussounadis et al. (2015) were Figures 2 and 4. As can be seen from the appended figures, the results of this paper could not be reproduced. Likely this was due to the difference in `lumi` preprocessing steps in my versus their analysis.

Additionally, there was some abiguity as to exactly how the correlations between differential gene and protein expression were calculated after differentially-expressed genes and non-differentially-expressed genes were identified. Specifically, since there were multiple time-points for each gene for each condition, then for each gene for each condition for each time-point, a p-value is generated from the `limma` package's log fold change (logFC) functionality. Thus, though they explicitly list an FDR cutoff of 0.05; does this mean a gene is differentially-expressed if *any* of the its time-points for

that condition are significant? Or *all* are significant? For my analyses, I used the latter and this may have skewed the results.

As far as G-causal analyses, I was able to apply ADF to all of the timeseries; and then the G-causality tests to those timeseries that showed stationarity from the ADF test. The data can be found in accompanying `data` directory.

Discussion/ Future Work

The results from my analysis show that gene and protein expression show poor correlation over time and under various conditions, even after using difference and differential transformations. Despite the differences between my and the original author's analysis, this comfirms the fact that the mRNA-protein relationship is more nuanced than mere correlation. This area requires more investigation if gene expression studies are to used for any functional arguments. Additionally, it would be highly useful to computational bioengineering applications for the gene-protein expression relationship to be resolved quantitatively.

Although I was unable to see this analysis through to its logical conclusion (due to time constraints), the idea of building (G-)causal models of gene-protein expression seems like an interesting direction for this area of research. As an analogy to neuroscientific analyses, it seems reasonable that such G-causal models could be linked with generative models to obtain a so-called "Granger predictive" models, which could be assessed through application to experimental data. See Seth, Barrett, and Barnett (2015) for review of how this is done in neuroinformatics research.

References

Amblard, Pierre-Olivier, and Olivier J. J. Michel. 2012. "The Relation Between Granger Causality and Directed Information Theory: A Review." *CoRR* abs/1211.3169. http://arxiv.org/abs/1211.3169.

Friston, Karl J. 2011. "Functional and Effective Connectivity: A Review." *Brain Connectivity* 1 (1): 13–36. doi:10.1089/brain.2011.0008.

Gustafsson, Mika, Danuta R. Gawel, Lars Alfredsson, Sergio Baranzini, Janne Björkander, Robert Blomgran, Sandra Hellberg, et al. 2015. "A Validated Gene Regulatory Network and GWAS Identifies Early Regulators of T Cell–associated Diseases." *Science Translational Medicine* 7 (313): 313ra178–78. doi:10.1126/scitranslmed.aad2722.

Koussounadis, Antonis, Simon P. Langdon, In Hwa Um, David J. Harrison, and V. Anne Smith. 2015. "Relationship Between Differentially Expressed mRNA and mRNA-Protein Correlations in a Xenograft Model System." *Scientific Reports* 5 (June): 10775. http://dx.doi.org/10.1038/srep10775.

Machado, Daniel, and Markus Herrgard. 2014. "Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism." *PLoS Comput Biol* 10 (4): e1003580. doi:10.1371/journal.pcbi.1003580.

O'Brien, Edward J, and Bernhard O Palsson. 2015. "Computing the Functional Proteome: Recent Progress and Future Prospects for Genome-Scale Models." *Current Opinion in Biotechnology*, Systems biology • nanobiotechnology, 34 (August): 125–34. doi:10.1016/j.copbio.2014.12.017.

Seth, Anil K., Adam B. Barrett, and Lionel Barnett. 2015. "Granger Causality Analysis in Neuroscience and Neuroimaging." *The Journal of Neuroscience* 35 (8): 3293–97. http://www.jneurosci.org/content/35/8/3293.short.

Thiele, Ines, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, et al. 2013. "A Community-Driven Global Reconstruction of Human Metabolism." *Nat Biotech* 31 (5): 419–25. http://dx.doi.org/10.1038/nbt.2488.

Zielinski, Daniel C., Fabian V. Filipp, Aarash Bordbar, Kasper Jensen, Jeffrey W. Smith, Markus J. Herrgard, Monica L. Mo, and Bernhard O. Palsson. 2015. "Pharmacogenomic and Clinical Data Link Non-Pharmacokinetic Metabolic Dysregulation to Drug Side Effect Pathogenesis." *Nature Communications* 6 (June): 7101. doi:10.1038/ncomms8101.

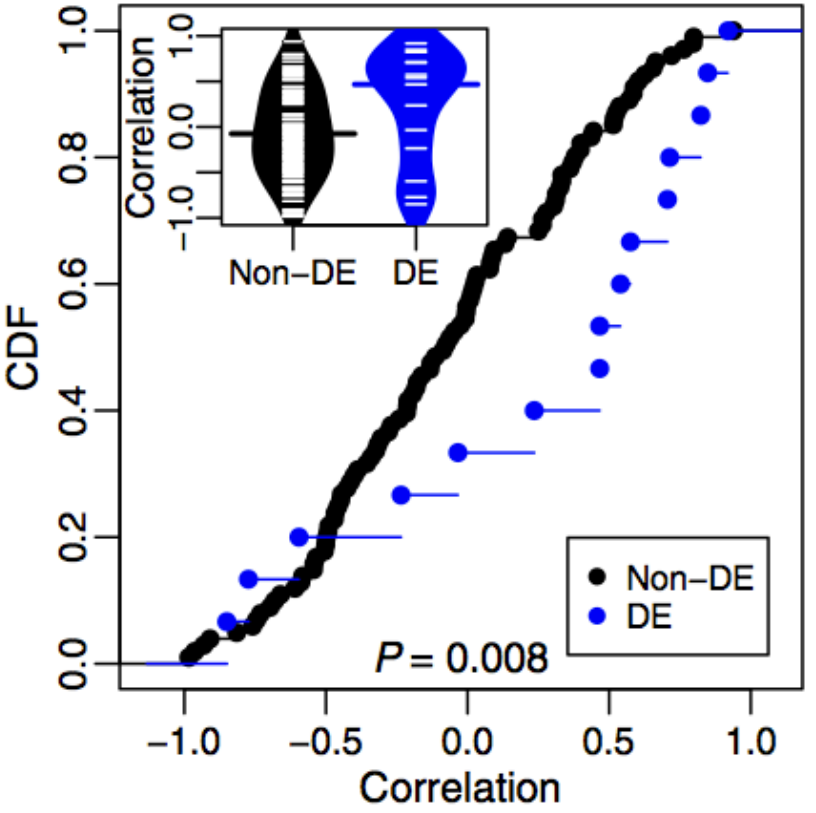
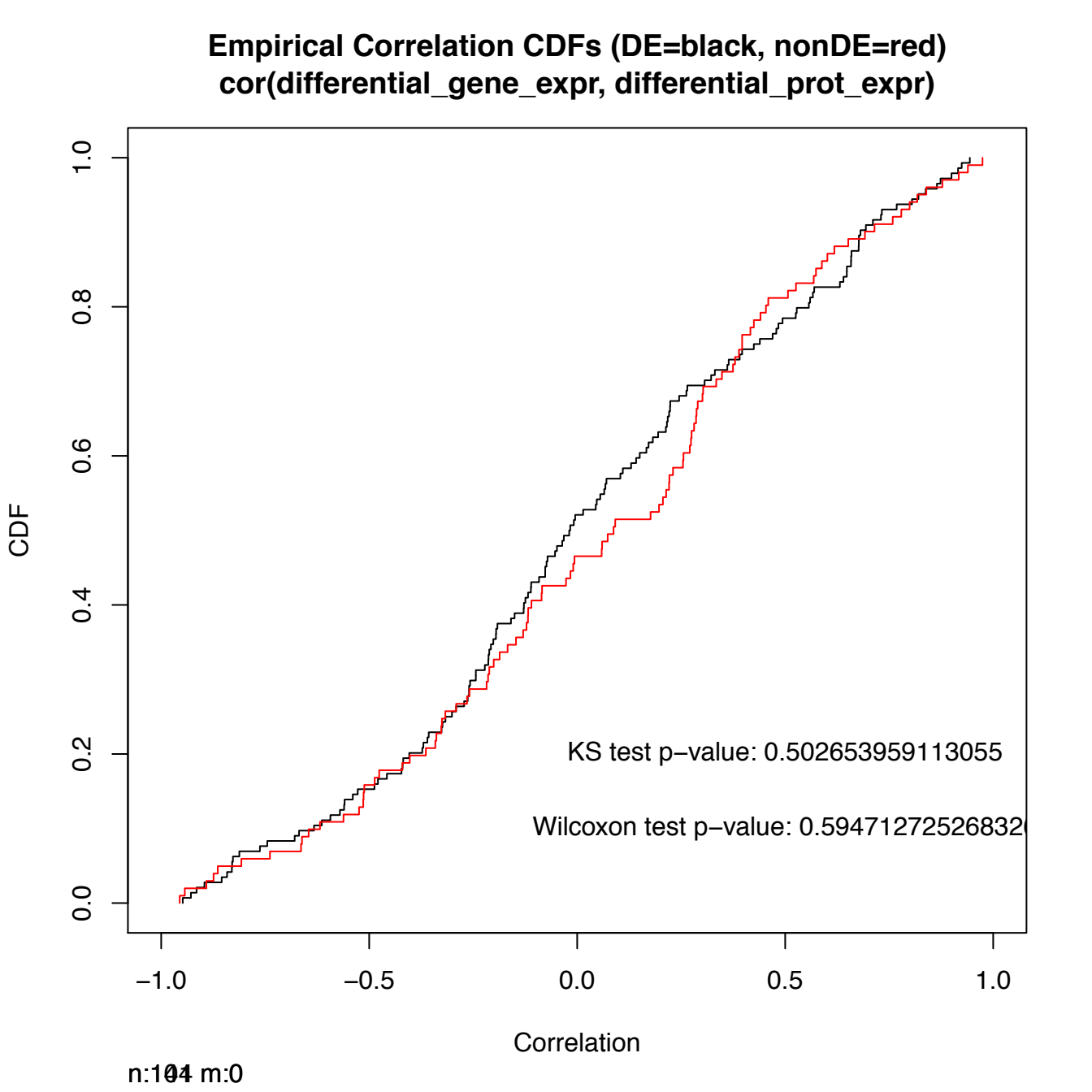
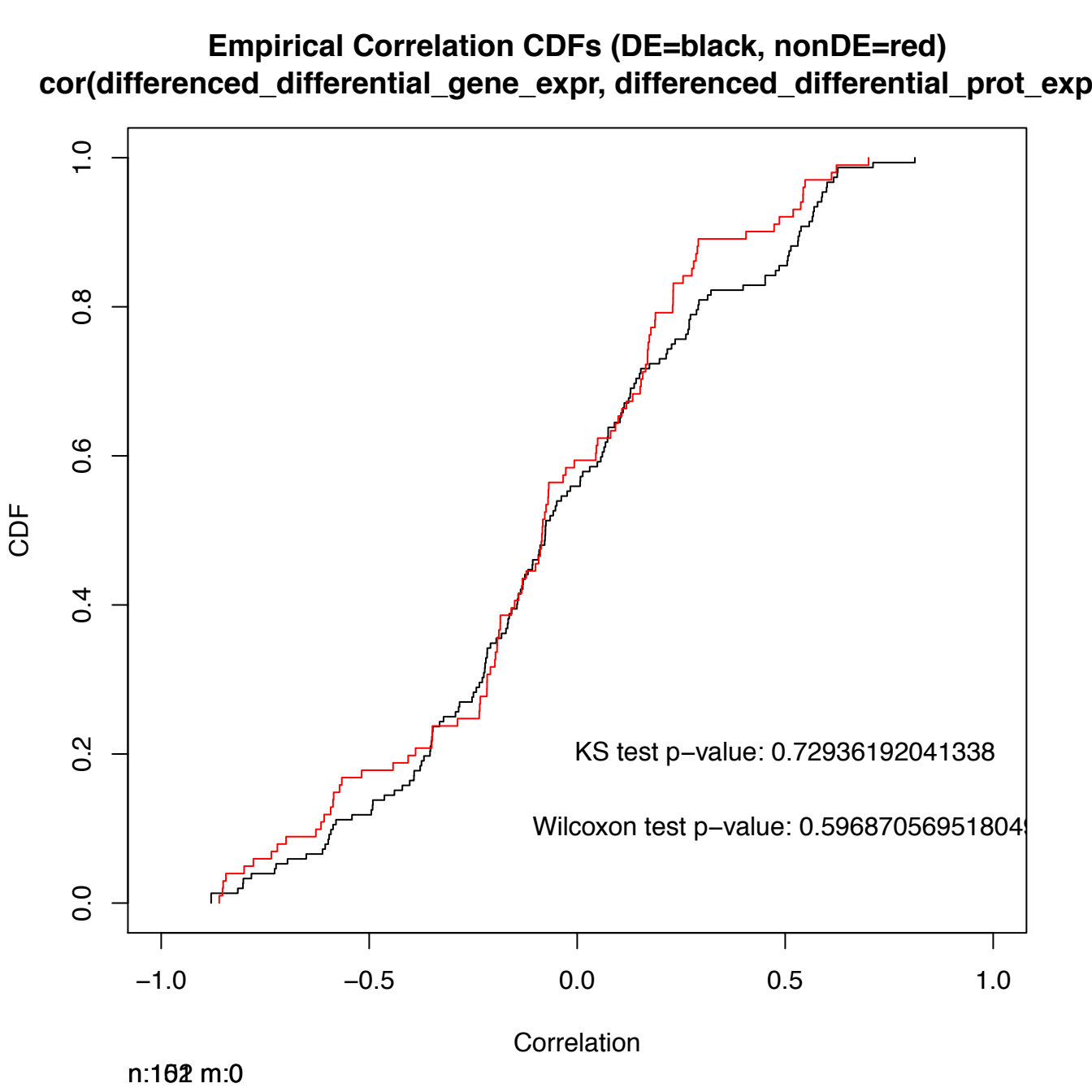
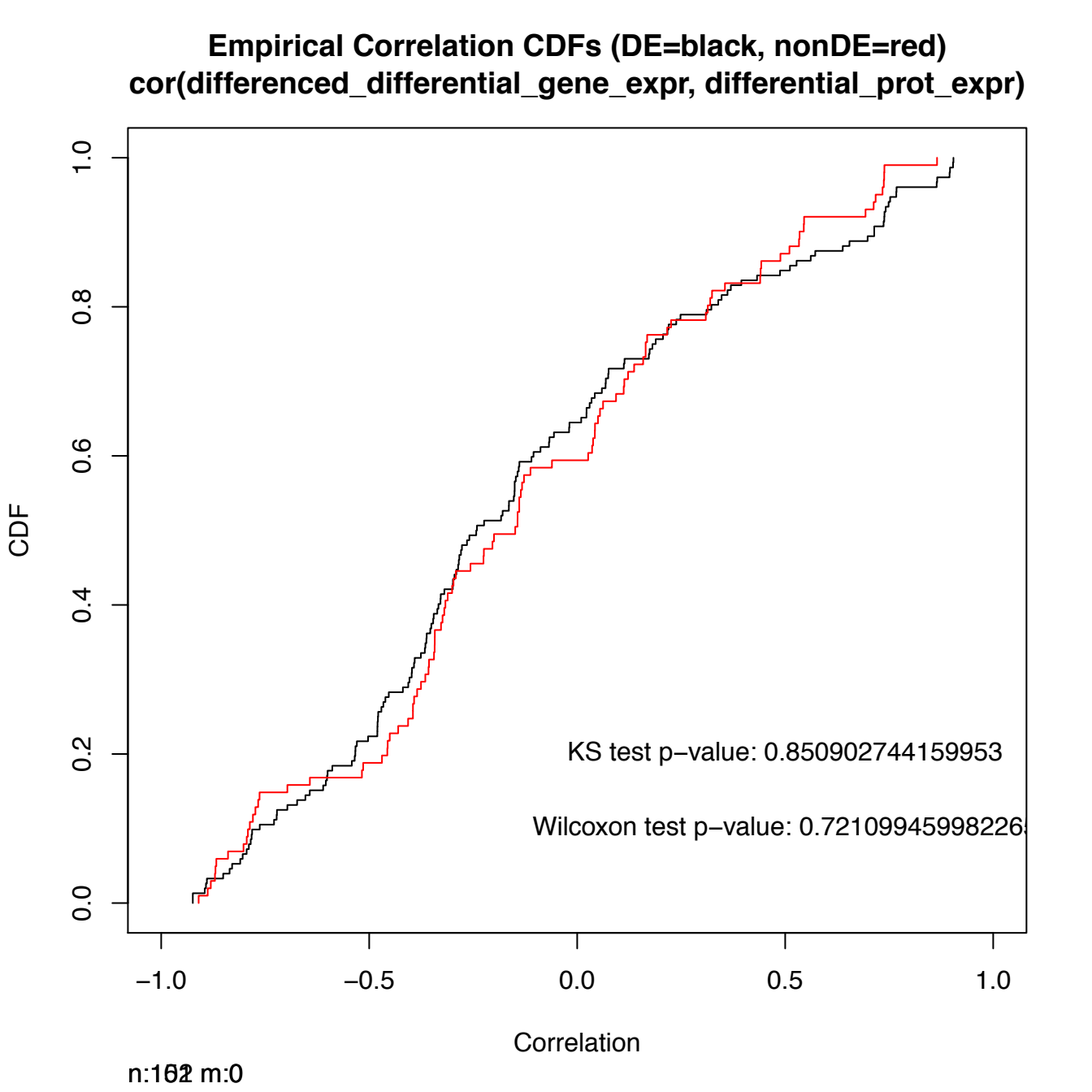
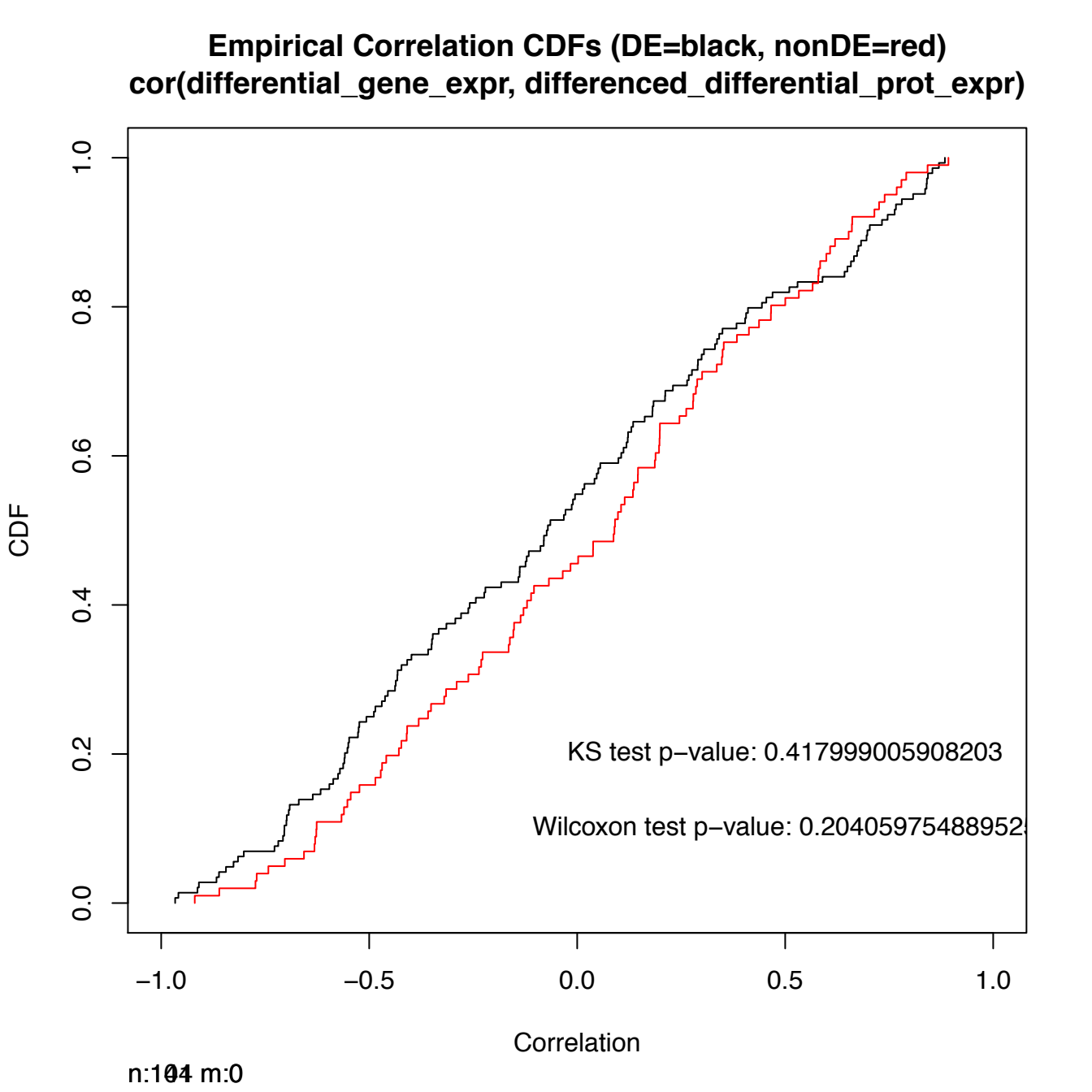


Figure 2. Distribution of mRNA-protein correlations for differentially expressed and non-differentially expressed mRNA profiles. The empirical cumulative distribution functions (CDF) over correlation coefficients between mRNA and protein values in a condition are plotted for differentially expressed mRNA profiles (DE) and non-differentially expressed mRNA profiles (Non-DE). Each individual correlation coefficient is based on n = 5 time points within a condition. Inset shows bean plots with ticks for each correlation value inside of density envelopes; bars represented medians for each group. Differentially expressed mRNA profiles are shifted significantly to higher correlations (Kolmogorov-Smirnov test, p = 0.008) and have a higher median (Wilcoxon test, p = 0.03).


Empirical Correlation CDFs (DE=black, nonDE=red)
cor(differential_gene_expr, differential_prot_expr)
CDF
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
Correlation
KS test p−value: 0.502653959113055
Wilcoxon test p−value: 0.59471272526832
n:104 m:0


Empirical Correlation CDFs (DE=black, nonDE=red)
cor(differenced_differential_gene_expr, differenced_differential_prot_exp
CDF
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
Correlation
KS test p−value: 0.72936192041338
Wilcoxon test p−value: 0.59687056951804
n:162 m:0


Empirical Correlation CDFs (DE=black, nonDE=red)
cor(differenced_differential_gene_expr, differential_prot_expr)
CDF
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
Correlation
KS test p−value: 0.850902744159953
Wilcoxon test p−value: 0.72109945998226
n:162 m:0


Empirical Correlation CDFs (DE=black, nonDE=red)
cor(differential_gene_expr, differenced_differential_prot_expr)
CDF
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
Correlation
KS test p−value: 0.417999005908203
Wilcoxon test p−value: 0.20405975488952
n:104 m:0

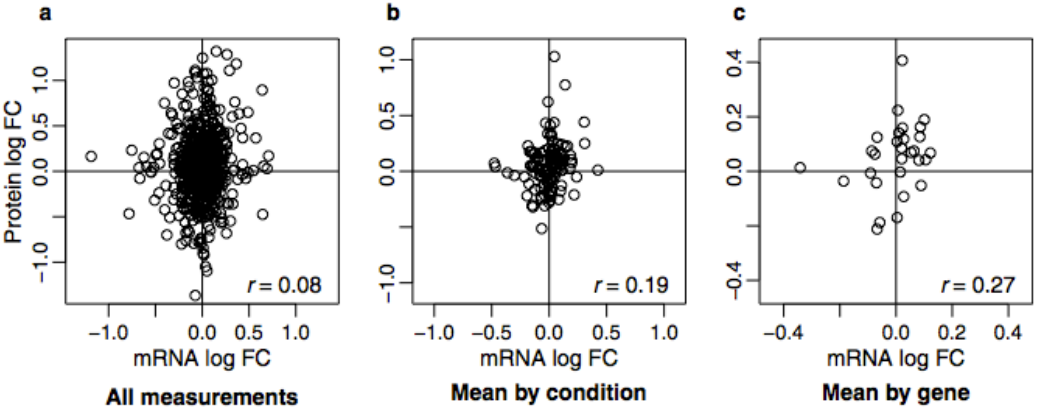
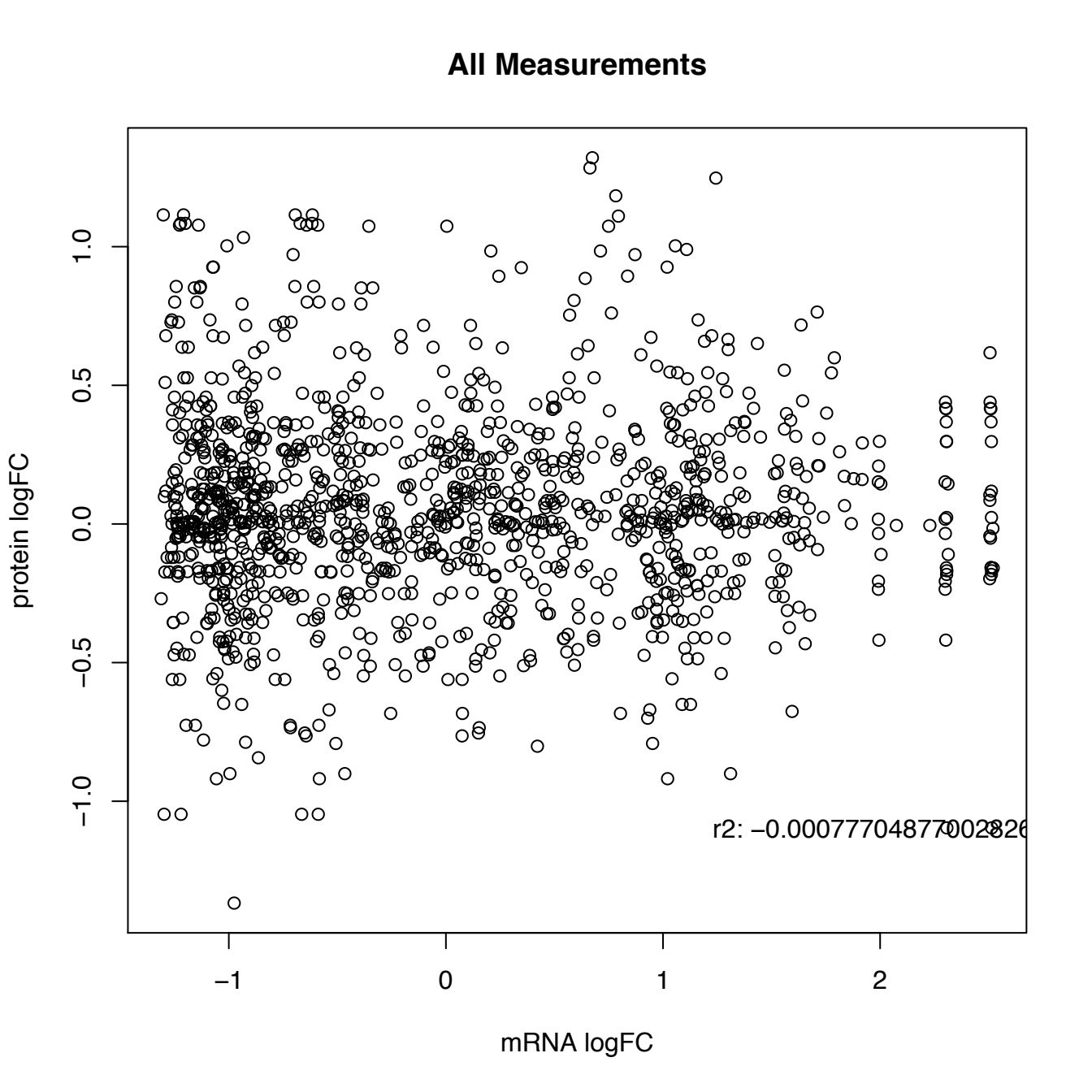
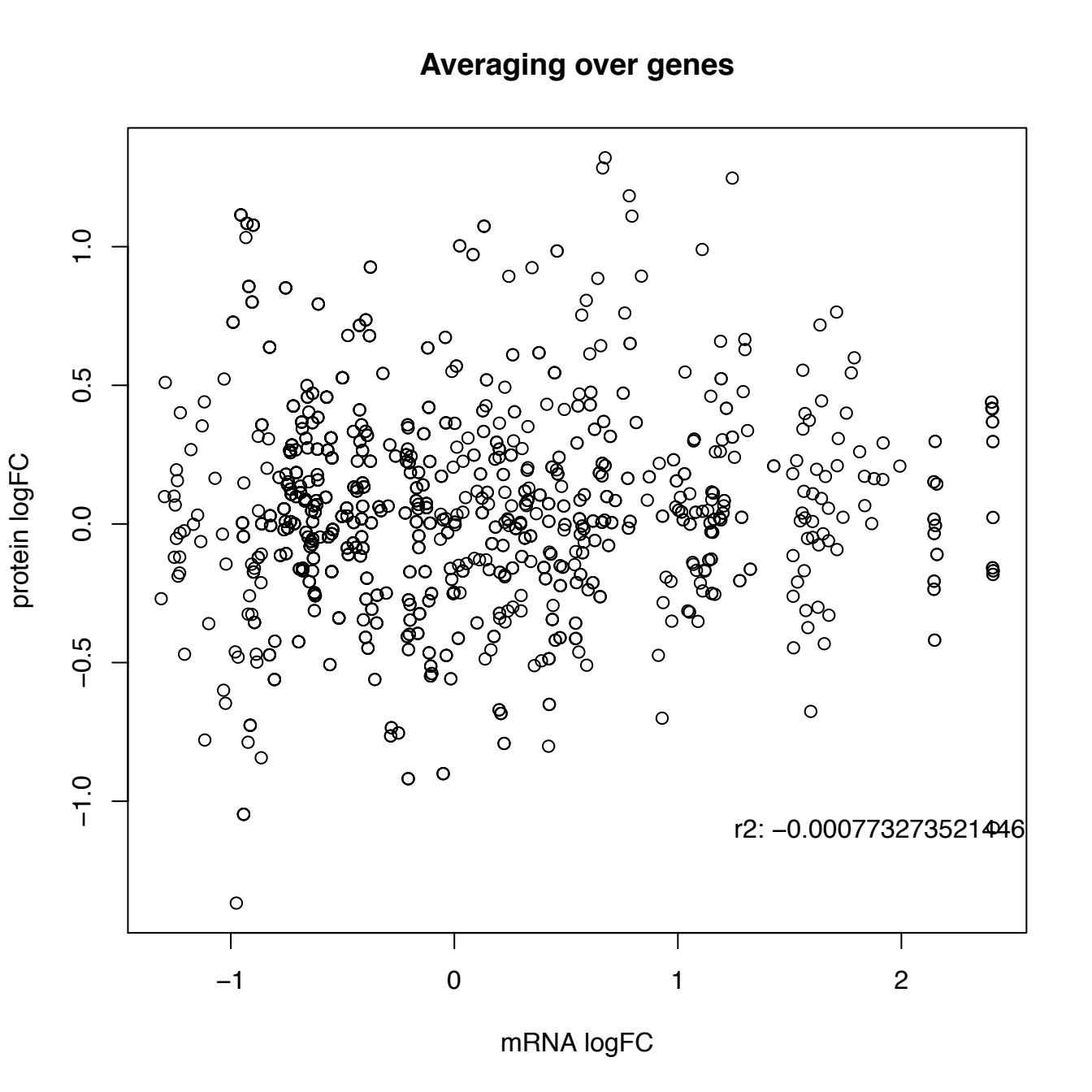


Figure 4. Genome-wide correlations for mRNA and protein expression. Scatterplots with associated correlation coefficients (r) for **(a)** all measurements taken (n = 579), **(b)** means by condition (n = 116), and **(c)** means by gene (n = 29).


All Measurements
protein logFC
mRNA logFC
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2
r2: −0.0007770487700282


Averaging over genes
protein logFC
mRNA logFC
r2: −0.000773273521446
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2