Reply to Weak effects made to appear strong by inflated correlation coefficients: The correlation between 3D genomic distance and codon usage frequencies is comparable to the maximal correlation expected

Alon Diament and Tamir Tuller

Abstract

In reply to Cherry's comments in PMID: 25510862, we further explain here why Cherry's claims were thoroughly addressed in the original manuscript, and that the methods and results were presented in a transparent manner. Most importantly, we reiterate that the relation between variables has been subject to stringent statistical tests, and that the observed signals are indeed strong with respect to expected and previously reported ones in large-scale genomic studies. In addition, we illustrate again that the reported correlations are comparable to the maximal correlation expected when comparing large scale noisy data after quantization. Finally, we show that the correlations reported in our study are similar to the correlations obtained between two Hi-C experiments; thus, if we follow Cherry's line of thought, we actually should absurdly conclude that the Hi-C protocol in general is problematic. We discuss the generality of our conclusion to systems biology analysis of Next Generation Sequencing (NGS) data.

When determining whether an observed correlation is high or low one should consider various aspects of the data such as biases and the number of points

It is important to note that determining whether an observed correlation is high or low is a relative test, and may be related to various aspects such as the signal to noise ratio and discretization of the data. For this reason, our tests for the strength of correlation – and what our conclusions are ultimately based upon – were statistically significant and passed a series of stringent control tests. P-values were estimated by computing the *binned* correlation between variables for permuted models that preserve many of the dataset's properties (a null model we termed Cyclic Chromosome Shift). It can be seen (Figure 3 in [1]) that it is not trivial at all to observe such high correlations as the ones reported between CUFS and 3DGD. Other controls included: controlling for the linear distances between genes on (unfolded) chromosomes; showing that the observed correlations are clearly related to the coding sequence (where CUFS is defined), and that they are missing from non-coding sequences that are adjacent to genes; controlling for GC content in coding and non-coding regions, for gene length, and for other sequence properties; controlling for experimental biases and resolution effects. All controls were performed for binned correlations in the same resolution and indicate that the correlations are not "inflated" but are rather strongly related to biological signals. *In addition to these controls, we validated our results using other representations of the data*, including distances between genes in 3D genome reconstructions, and depletion of CUFS between

genes with highly enriched Hi-C contact frequencies (Figure S14 in [1]). However, Cherry took no notice of these reported results.

Figure S5 in [ref 1] indeed presents a decrease in the correlation coefficient as the resolution (number of bins) increases, as expected. However, the ranking (compared with distance measures other than CUFS) and the *significance of the correlations are invariant*. Specifically, the t-test p-value estimated from Spearman's correlation on *raw data* is still $p < 10^{-323}$, and most importantly our empirical p-value is unchanged ($P_{3D}<0.01$ for almost all organisms). Thus, we showed that our conclusions are not dependent on the binning scheme, as suggested in the critical paper cited by Cherry [2]. Moreover, we selected the number of bins for each organism so that the level of averaging (number of pairs per bin) is roughly kept constant across organisms. *In fact, the number of bins that we used in 3 out of 5 organisms is higher than the number of points in the raw data analyzed in the aforementioned paper* [2].

A simulation of noisy data and quantization demonstrates how a perfect correlation between two variables decreases to be -0.08

In addition, we performed a simulation of the possible effects of noise and data quantization and included it in Figure S5 (reproduced here in **Figure 1**). This synthetic example of correlation vs. resolution shows that the profiles in Figure S5 are typically seen in any case of binning noisy measurements. The panels show 10^7 samples from two perfectly correlated variables (r=-1), that are severely corrupted by noise (additive uniform noise in [0, 3] for both variables x and y) and discretization of the data. The outcome of this process is a reduced – but still significant (two-tail t-test) – correlation of r=-0.08. The resultant curve of binned correlation vs. bin size / number of bins is similar to that observed in Figure S5 in [1]. The number of samples, number of bins, and correlation coefficients, are similar to those we observed in yeast. Binned correlation in this example indicates a true signal that is masked by noise.

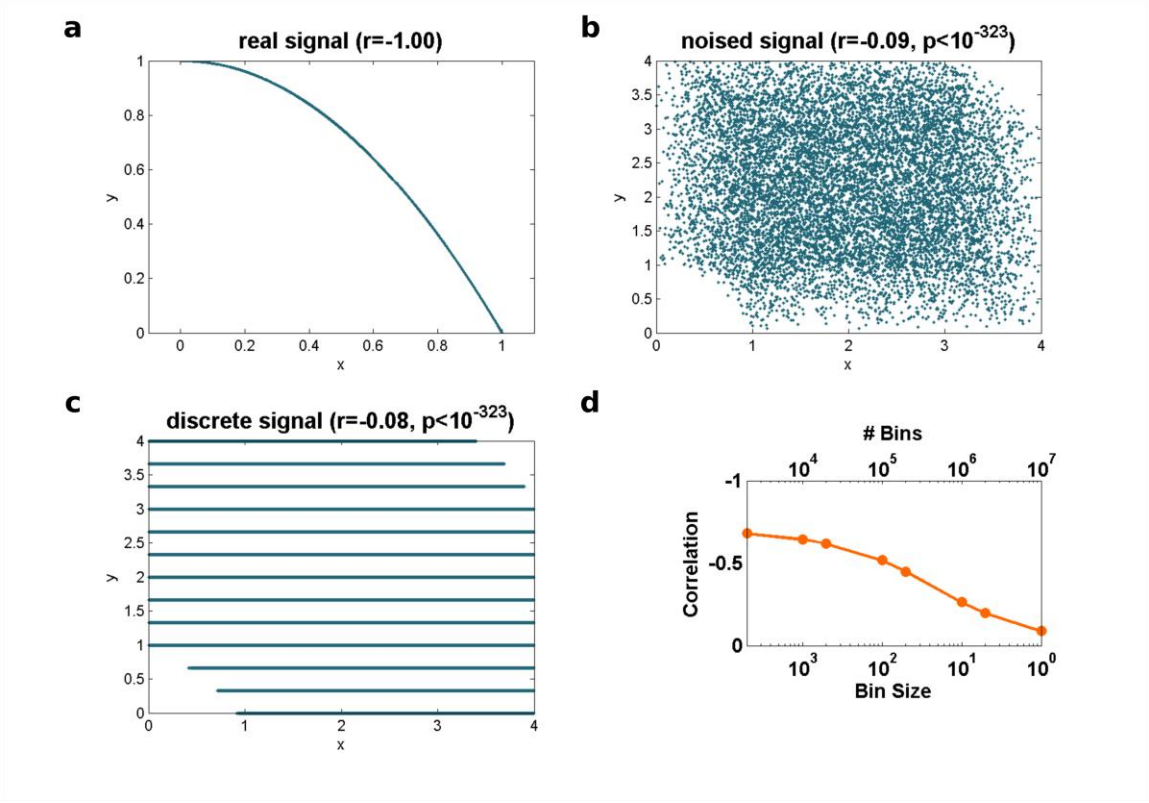

Figure 1: A synthetic example. Scatter plots: **(a)** Generated variables. **(b)** Noised variables. **(c)** Quantized variables. **(d)** Spearman's correlation coefficient between the variables in (c) for varying bin sizes / number of bins.

The number of points analyzed in the study is 4-6 orders of magnitude higher than in previous studies in the field

The number of samples in our analysis, which stems from the fact that we analyze pairwise distances between genes, is extremely high, ranging from 13 to 369 million points, orders of magnitude higher than any other systems biology study that we are aware of. For example, a recent paper used 20 bins to compare the correlation between expression profiles of pairs of interchromosomal genes as a function of the number of contacts linking the two genes [3]. Another paper used eigenvalue decomposition of a 1Mbp-binned Hi-C map to study the relation between Hi-C contact enrichment and GC content (approximately 3,000 points, and to compare human and mouse maps [4]). Naturally, various systems biology studies that dealt with a very large number of values and were published in top journals performed correlations (or, similarly, dot plots) between binned values [5–11].

The number of bins that we use ranges from 2,000 to 64,000 – still higher than typically seen in correlation analyses in the field (which usually include a few hundred points), and

for which the presented correlations are highly non-trivial. Binning was utilized as it enables comparing the reported results to other results in the field. We chose the number of bins to be comparable to previous studies in the field, and indeed the reported correlations in our study are higher than previously reported for a similar number of points [6,12,13].

It was recently shown that CUFS can be used to improve 3D genomic reconstructions based on Hi-C data

The conclusions of our paper have also been tested in a recent study [14], where we showed that 3D distances predicted by CUFS can be employed to reconstruct an improved 3D model of the yeast genome. In this study we utilized functional distances (CUFS) between gene pairs as constraints in the 3D reconstruction program (in addition to Hi-C constraints). The accuracy of the resultant models was determined by a series of 15 previously reported signals that are related to yeast genomic organization, and how strongly these signals appeared in the generated reconstruction. Reconstructions that contained additional CUFS-constraints clearly led to significant improvement on these benchmarks in most cases (12 out of 15). These results support the conjecture that CUFS is strongly and directly related to the 3D genomic organization of genes.

The correlation between 3D genomic distance and codon usage frequencies is comparable to the correlation between two different measurements of 3D genomic data

Finally, in response to Cherry's analysis without binning, *we note that his approach can be as misleading as any other, and may lead to absurd conclusions*. We repeated his confusing analysis (**Figure 2**) for yeast distances (3DGD) computed based on two different genome-wide 3C-based experimental protocols – genome conformation capture (GCC) [15] and the 4C-based high-throughput method of Duan *et al.* [16] Computing the correlations between these datasets without bins resulted, naturally, in low correlations. The correlation between all pairwise distances in the two experiments is similar to the one observed between 3DGD and CUFS without bins (r=0.05, Spearman's rho; even when we compared the *contact frequencies* of regions that were successfully monitored in both experiments, the correlation was only 0.44). Our conclusion is that this is due to the level of noise in such large-scale genomic data that is bound to distort signals related to *raw* pairwise distances.

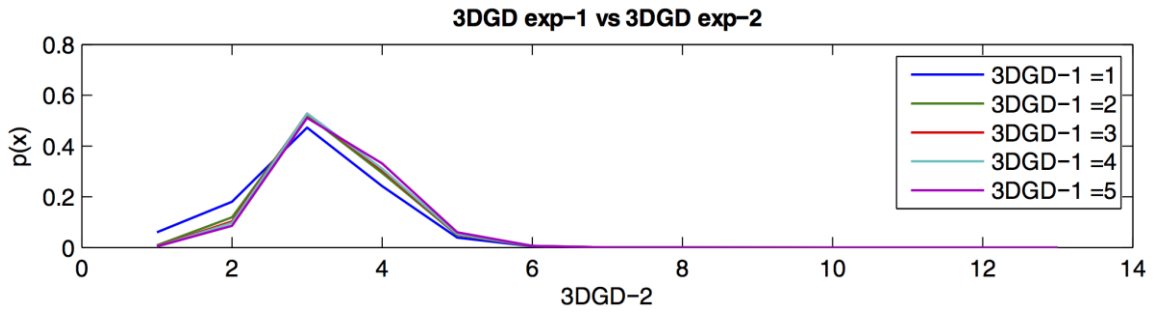

Figure 2: Comparison of Hi-C datasets using Cherry's method. Distribution of 3DGD between genes in experiment 2 (4C method) given their distance in experiment 1 (GCC method), as denoted in the legend (Spearman's rho between pairwise distances in the two datasets is r=0.05).

The results suggest that according to Cherry's method of using non-binned data, repeated experiments, and even similar variants of the Hi-C protocol, are loosely related, and therefore any attempt to study 3D genomic organization using Hi-C, including dozens of previous studies, is futile.

More generally, this example demonstrates that analyzing complex biological phenomena and large scale NGS data with over-simplified analyses (like Cherry's) is dangerous and misleading; to accurately understand and evaluate the results presented in our study the reader should read it carefully and thoroughly.

It is our view that binning is a legitimate statistical method, as long as the employed methods are clearly explained in the paper and controlled for statistical significance. As with any other paper, abstracts and titles are limited in scope and should be interpreted in the context of the complete manuscript. We have argued here that the analyses in [1] are statistically sound and, given the current resolution and quality of data, support the conjecture that 3D genomic organization and gene function and expression are strongly related.

A more general discussion

In summary, we would like to conclude with a broader discussion. We believe that Cherry's comment and our reply have a much broader scope, beyond the CUFS–3DGD relation discussed here. We believe that when evaluating the strength of the relation between variables, based on the output of a NGS experiment, the following points should be considered:

1) The number of points should be accounted for, among others, via p-value(s) computation. Binning the data usually increases the correlation; however, on the other hand, a larger number of points usually decreases the correlation. A correlation of 0.9 when there are 3 points is not significant while a correlation of 0.1 when there are 10^4

points (or bins) is most significant. Thus, reporting a correlation without considering the number of analyzed points is misleading.

2) The biases and noise in the analyzed data. Recent NGS protocols introduce various types of non-trivial biases [4,17–19]. These usually cannot be dealt with based on traditional statistical approaches (*e.g.* a t-test). Often the strength of a correlation can't be fully evaluated without tailoring specific filtering or signal processing methods related to the nature of the analyzed data. In addition, the strength of the correlation should be evaluated in light of the correlations obtained between two measurements of a certain relevant variable (which can serve as an upper bound on the possible obtained correlation).

3) The number of significant variables believed to be involved in the analyzed system. When studying complex intracellular biological processes that include dozens of relevant/central variables it is naïve to expect that the variance of one variable can be fully explained by another. The number of expected relevant variables should be considered when evaluating the strength of a correlation. For example, in the case of a system with 50 variables a correlation of 0.15 may be considered high, relevant, and biologically interesting.

References

1. Diament A, Pinter RY, Tuller T. Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function. Nat Commun. 2014;5.

2. Kenny PW, Montanari CA. Inflation of correlation in the pursuit of drug-likeness. J Comput Aided Mol Des. 2013;27:1–13.

3. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert J-P, et al. Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. 2014;24:974–88.

4. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods. 2012;9:999–1003.

5. Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotech. 2009;27:361–8.

6. Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature. 2003;425:737–41.

7. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445:168–76.

8. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. Cell. 2012;151:68–79.

9. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A. 2002;99:3695–700.

10. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, et al. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 2009;19:510–9.

11. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, et al. A Role for Codon Order in Translation Dynamics. Cell. 2010;141:355–67.

12. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008;452:423–8.

13. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet. 2006;38:1043–8.

14. Diament A, Tuller T. Improving 3D Genome Reconstructions Using Orthologous and Functional Constraints. PLoS Comput Biol. 2015;11:e1004298.

15. Rodley CDM, Bertels F, Jones B, O'Sullivan JM. Global identification of yeast chromosome interactions using Genome conformation capture. Fungal Genetics and Biology. 2009;46:879–86.

16. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional model of the yeast genome. Nature. 2010;465:363–7.

17. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43:1059–65.

18. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. Nucl. Acids Res. 2014;42:9171–81.

19. Artieri CG, Fraser HB. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res. 2014;gr.175893.114.