Briefings in Bioinformatics, 21(5), 2020, 1523–1530

doi: 10.1093/bib/bbz083
Advance Access Publication Date: 17 October 2019
Review article

Beware the Jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis

Stefania Salvatore, Knut Dagestad Rand, Ivar Grytten, Egil Ferkingstad, Diana Domanska, Lars Holden, Marius Gheorghe, Anthony Mathelier, Ingrid Glad and Geir Kjetil Sandve

Corresponding author: Stefania Salvatore, Department of Informatics, University of Oslo, Oslo, Norway. Tel.: 0047 23368970; E-mail: stefasal@ifi.uio.no

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

Abstract

The generation and systematic collection of genome-wide data is ever-increasing. This vast amount of data has enabled researchers to study relations between a variety of genomic and epigenomic features, including genetic variation, gene regulation and phenotypic traits. Such relations are typically investigated by comparatively assessing genomic co-occurrence. Technically, this corresponds to assessing the similarity of pairs of genome-wide binary vectors. A variety of **similarity measures** have been proposed for this problem in other fields like ecology. However, while several of these **measures** have been employed for assessing genomic co-occurrence, their appropriateness for the genomic setting has never been investigated. We show that the choice of **similarity measure** may strongly influence results and propose two alternative modelling assumptions that can be used to guide this choice. On both simulated and real genomic data, the Jaccard index is strongly **altered** by dataset size and should be used with caution. The Forbes coefficient (fold change) and tetrachoric correlation are less **influenced** by dataset size, but one should be aware of increased variance for small datasets. All results on simulated and real data can be inspected and reproduced at https://hyperbrowser.uio.no/sim-measure.

Key words: statistical genomics; genomic track similarity; fold enrichment; similarity measures; similarity indices.;

Introduction

A reference genome provides a unified coordinate system to locate where genomic features occur. Genomic tracks [1] indicate the base pair positions on a reference genome where a specific genomic feature is observed. Several large consortia [2–4] have contributed to the public domain with reference datasets of such genomic features such as cell-specific chromatin states and protein binding regions. Such genomic tracks are often

Stefania Salvatore is a post doctoral fellow at the Department of Informatics, University of Oslo.
Knut Dagestad Rand is a PhD fellow at the Department of Mathematics, University of Oslo.
Ivar Grytten is a PhD fellow at the Department of Informatics, University of Oslo.
Egil Ferkingstad is a researcher at the Science Institute, University of Iceland.
Diana Domanska is a post doctoral fellow at the Department of Informatics, University of Oslo.
Lars Holden is the managing director of the Norwegian Computing Center, Oslo.
Marius Gheorghe is a PhD fellow at the Centre for Molecular Medicine Norway (NCMM), University of Oslo.
Anthony Mathelier is the head of the group at the Centre for Molecular Medicine Norway (NCMM), University of Oslo.
Ingrid Glad is a professor at the Department of Mathematics, University of Oslo.
Geir Kjetil Sandve is associate professor at the Department of Informatics, University of Oslo.
Submitted: 12 March 2019; **Received (in revised form):** 13 June 2019

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

compared against reference collections to identify proteins, chromatin marks, or cell types of interest. These comparisons rely on relevant **measures** of similarity (co-occurrence), which should reflect the degree of biological association.

Many of the widely used binary similarity **measures**, such as the Jaccard index and the Forbes coefficient (also known as fold change), were initially developed in the context of species presence/absence problems within ecology [5, 6]. Within ecology, there has been a long debate on the properties of different similarity **measures [7, 8].** In the genomics field, however, these **measures** have been applied without a critical discussion of their suitedness and typically without providing any reasoning for why a particular **measure** was chosen. For instance, more than 300 papers on *bioRxiv* between 2010 and 2018, within a limited selection of genomics literature, explicitly mention the Jaccard **measure**, while the argumentation for such a choice is usually not provided. The Forbes coefficient is also commonly used, but the extent of its usage is harder to quantify. It represents a ratio of observed versus expected co-occurrence and is thus often simply termed *enrichment* or *fold change*. Existing tools for genomic track analysis typically provide only a single **measure** for the degree of co-occurrence between tracks, without any discussion of the properties and suitedness of this chosen **measure**. As an example, the widely used BEDTools [9] originally provided only raw overlap counts. Later, the Jaccard index was also added, but without any user guidance on its suitability [10].

In this report, we provide a critical evaluation of the properties of three similarity **measures** in the context of genomics: the Jaccard index, the Forbes coefficient and the tetrachoric correlation. Specifically, we focus on how these **measures** may be **altered** by experimental variability (number of genome-wide annotations per genomic feature) or technical variability (computational assessment of similarity). We introduce two alternative modelling assumptions for how an underlying biological feature may give rise to experimental datasets of varying size (number of base pairs covered by a track). According to each of these assumptions, we carefully assess the robustness of these three commonly used similarity **measures**. By connecting the choice of **similarity measure** to underlying assumptions, we enable the genomics community to make reasoned choices of **measures** in particular settings. Furthermore, we perform a large-scale and systematic study on how size variation of experimental datasets affect rankings of co-occurrence according to each of the three similarity **measures**.

Methods

In the present study, we refer to *binary vectors* or *tracks* as **datasets** of occurrences anchored to specific coordinates in a reference genome. Therefore, throughout the study, we interchangeably use the terms similarity and co-occurrence when referring to relations between genomic tracks. A reference track can be defined as a binary indicator R, where $r_s = 1$ if position s is covered by an occurrence for $s \in 1, \ldots, N$, with N being the length of a reference genome. We refer to the coverage of a track (sum across the vector) as the *size* of the track. A query track q is a separate track against which reference tracks are analysed.

We introduce two principled models for how an underlying genomic feature may give rise to experimental datasets **(tracks)** of varying size. The first model assumes that an experimental dataset of binary events is the imperfect observation of an underlying binary biological process. The second model assumes that an experimental dataset of binary events corresponds to a particular thresholding of an underlying continuous reality.

Table 1. Contingency table

	$Q_s = 1$	$Q_s = 0$
$R_s = 1$	a	b
$R_s = 0$	c	d

Model assumption 1: binary underlying reality

Depending on the context, the underlying reality of a biological process can be considered to be binary (see Supplementary Material for a formal definition). **In short, when the underlying reality of the biological process of interest is discrete (binary variable), we assume that the feature of interest is binary in its nature—i.e. formed by events either occurring or not occurring. In this situation we denote by r an imperfect observation of the underlying binary reality Y.** For instance, this could correspond to a transcription factor (TF) either binding to the DNA or not. Under this assumption, a track r represents an imperfect observation of the underlying binary reality Y. The track r could denote the locations of called peaks of a chromatin immunoprecipitation-sequencing (ChIP-seq) dataset, where some peaks reflect true binding events (true positives), some peaks do not reflect any true binding event (false positives) and some true-binding events remain undetected (false negatives) [11]. A particularly interesting case is when the rate of false positives in R is invariant to the dataset size. In the context of a ChIP-seq dataset, this occurs if the false discovery rate (FDR) has been properly controlled when calling peaks, i.e. when a given proportion of occurrences in r reflects true binding sites (the underlying reality Y), regardless of the dataset size.

Model assumption 2: continuous underlying reality

Alternatively, an observed binary **variable (r) can represent the discretisation of an underlying continuous reality (Y). As we are only observing binary values of r resulting from a thresholding of Y, we have no information on the underlying distributions for Y.** This corresponds to assuming that a TF has a varying propensity to bind and that a dataset represents the set of genomic locations where the binding propensity is above a given threshold.

When assuming a continuous underlying process, the similarity between the underlying continuous features are estimated based on the observed binary occurrences. A particular case is when two observed binary tracks q and r represent thresholded instantiations of one underlying bi-normal process [12]. Under this assumption, the correlation of this bi-normal process is a natural similarity **measure,** and the goal is to estimate the similarity between the latent continuous values based on the observed binary values.

Similarity measures

Currently, several **similarity measures** have been used to quantify co-occurrence of (similarity between) genomic tracks. As mentioned, two commonly used **measures** are the fold enrichment (Forbes coefficient) [6] and the ratio of intersection to union (Jaccard index) [5]. Using set notation, the Forbes coefficient is defined as $\frac{N*|R\cap Q|}{|R|*|Q|}$, where the notation $|\ T\ |$ indicates the number of elements of a set T. The Jaccard index is defined as $\frac{|R\cap Q|}{|R\cup Q|}$. Using the notation in Table 1, the two indexes can be calculated as $\frac{N*a}{(a+b)*(a+c)}$ and $\frac{a}{(a+b+c)}$, respectively.

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

Given our consideration of an underlying continuous reality, a third **similarity measure** of interest is the tetrachoric correlation, which is a classic measure of correlation [13]. To our knowledge, it has not previously been used to assess similarity (co-occurrence) in a genomic context. The tetrachoric correlation assumes that the two tracks Q and R are generated by thresholding two underlying continuous processes that are bi-normally distributed, and is defined as the correlation (ρ) between these two underlying processes. The thresholds and the correlation ρ can be estimated from Q and R using maximum likelihood techniques [12]. In our study, we have used the *polycor* R package [14, 15] to estimate ρ.

Tools

All plots and the information necessary for their reproduction can be found at https://hyperbrowser.uio.no/sim-measure. Plots generated by The Genomic HyperBrowser (GHB) webtools can be reproduced using the redo-functionality provided by the underlying Galaxy system. Plots generated in R are accompanied by their respective R code and the data files used to generate them. Plots generated by the University of California Santa Cruz (UCSC) Genome Browser are accompanied by URLs and form inputs required to generate similar plots in the current version of UCSC.

Results

Here, we explore the implications of the two principled models (binary and continuous underlying reality) on three different similarity **measures** (Jaccard index, Forbes coefficient and tetrachoric correlation) using analytical derivations, simulations and real data. We show that the dataset size may strongly affect the expected similarity value of a particular **measure**, as well as the uncertainty connected to this value. Further, we investigate the robustness of these **measures** by quantifying TF co-occurrence based on a large collection of ChIP-seq peak datasets of variable size [16]. We first derive the theoretical properties of the **measures in each scenario (see Supplementary Material) and present experimental results on simulated and real data.**

Theoretical properties of the measures

Model assumption 1: binary underlying reality

Assuming a binary model, we can show that if r is an imperfect observation of y with an expected FDR with respect to Y of q, i.e. $1 - E(\frac{|R \cap Y|}{|R|}) = q$, then the expected Forbes coefficient of r and q is a function of the Forbes coefficient of y and q, given by: $E(Forbes(R, Q)) = q \cdot Forbes(\bar{(}Y), Q) + (1 - q)Forbes(Y, Q)$ where $Forbes(Y, Q) = Forbes(Y = 1 \text{ and } Q = 1)$, while $Forbes(\bar{(}Y), Q) = Forbes(Y = 0 \text{ and } Q = 1)$. This allows for the Forbes similarity of two tracks r^1 and r^2 to be comparable with respect to q, as long as the FDR of y^1 and y^2 (q) is the same. If the FDR is zero ($q = 0$), $E(Forbes(r, q))$ reduces to $Forbes(y, q)$, meaning that the Forbes coefficient of r and q provides an unbiased estimate of the Forbes coefficient of y and q. We show this does not hold for the Jaccard index (see derivation given in Supplementary Material), since even for the case of no false positive $E(FDR) = 0$ **(which is a paricular case of FDR being positive though constant)**, the expected Jaccard index of r and q is dependent on the size of r (given the subsetting rate k):

$$E(Jaccard(R, Q)) = E(\frac{|R \cap Q|}{|R \cup Q|}) = \frac{|Y \cap Q|}{|Y \cup Q| + (1/k - 1)|Q|}$$

Model assumption 2: Continuous underlying reality

When binary **variables** R **are observed,** r can represent a binary discretisation of an underlying continuous reality Y^*. Correspondingly, a query track q represents the discretisation of a continuous reality X^*. Assuming that such a continuous underlying process exists, one is interested in the similarity between the latent continuous variables Y^* and X^*, given the observed binary vectors r and q, as well as the latent discretisation thresholds t and t^*. Under these conditions, the Forbes coefficient would give a systematic relative bias for small versus large observed tracks, since its values are based on the conditional probabilities $P(q_s = 1 | r_s = 1)$ (details provided in Supplementary Material). An explicit procedure of randomly subsetting occurrences (in order to scale a set of tracks R to the same size) would give rise to the same systematic bias. The reason is that the observed occurrences (r) of large tracks represent a range of continuous values (Y^*) that starts at lower values than for smaller tracks (since track size variation is assumed to be due to different thresholds). Thus, the subset of occurrences from a large track r represents (on average) lower continuous values of the underlying track (Y^*). The correlation between r and the query is therefore associated with lower continuous values of X^* and thus with lower frequency of occurrences at these locations in q. Thus, neither the use of a **similarity measure** such as Forbes or the use of explicit subsetting procedures are recommended in contexts where the observations represent a thresholded version of a continuous underlying reality. As an example, TF binding regions are typically determined by thresholding a signal that is based on the depth of mapped reads (referred to as ChIP-seq peak calling). If this read depth-based signal correlates with an underlying propensity to bind, then the variation in the dataset size will have the above characteristics.

Under such assumptions, the similarity between the tracks can alternatively be evaluated by the tetrachoric correlation, which assumes that the two tracks q **and** r are generated by thresholding two bi-normally distributed underlying (continuous) processes. The tetrachoric correlation is defined as the correlation (ρ) between these two underlying processes (details in Supplementary Material).

Experimental results

To further delineate the behaviour of the three similarity **measures** under controlled conditions, we performed a simulation study. Separately for Model assumptions 1 and 2, we generated datasets of varying sizes based on a fixed underlying degree of similarity according to a given model.

Simulated data under Model assumption 1

Figure 1 shows how the three considered **measures** behave when the datasets are simulated according to Model assumption 1. Datasets of varying sizes, but fixed (expected) FDR of 5%, were generated from the fixed base reference track y, representing the usually unknown underlying track, while a fixed query track q was used. Consistent with the analytical derivations, the Forbes coefficient seems to be less dependent on track size, while the Jaccard index is highly influenced by track size. The Jaccard index drops dramatically when the size of the reference track decreases. The size effect dominates the underlying (biological) signal when large size differences are present. The tetrachoric correlation is also **influenced by** track size, but less than the Jaccard index.

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

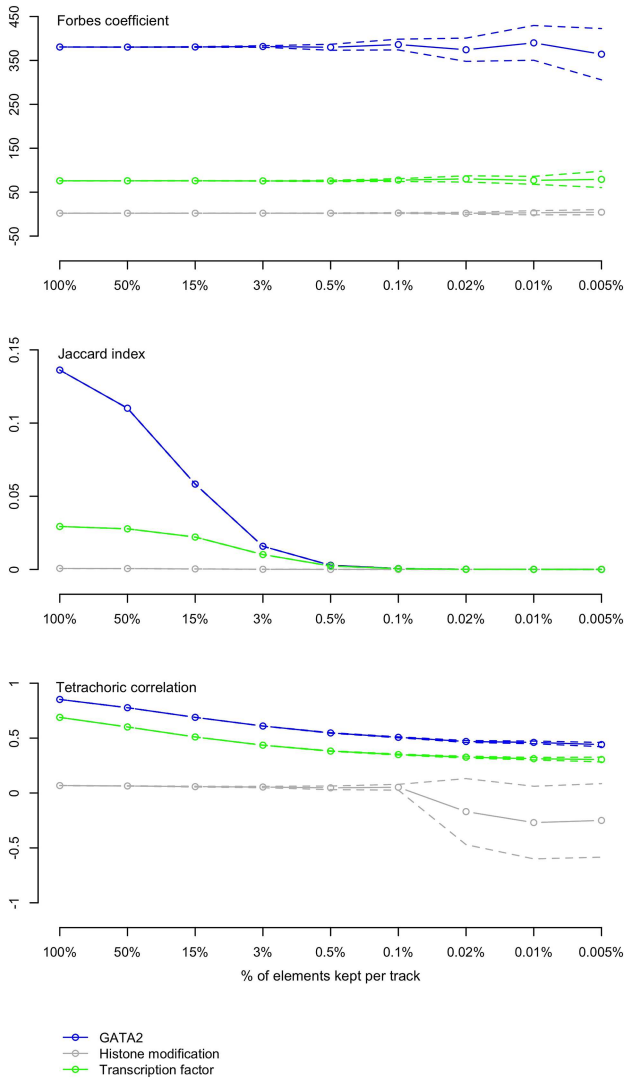

Figure 1. The behaviour of the three considered similarity measures on simulated datasets with size varying according to Model assumption 1, **and 5% FDR.** Top panel: simulation study showing the mean Forbes coefficient between a fixed query track *q* and reference tracks *rs* corresponding to varying sizes of a fixed base reference track (*y*). We set up three simulations based on a GATA1 ChIP-seq track as query track and respectively a GATA2 (blue), a transcription factor (green) and an histone modification (grey) ChIP-seq track as a base reference (details in Methods). The *x*-axis shows the proportion of occurrences in *r* compared to the reference track *y*, while the *y*-axis shows the resulting mean Forbes coefficient against the query track (continuous line) and one standard deviation from the mean Forbes coefficient (dashed lines). The middle and bottom panels show the same for the Jaccard index and the tetrachoric correlation.

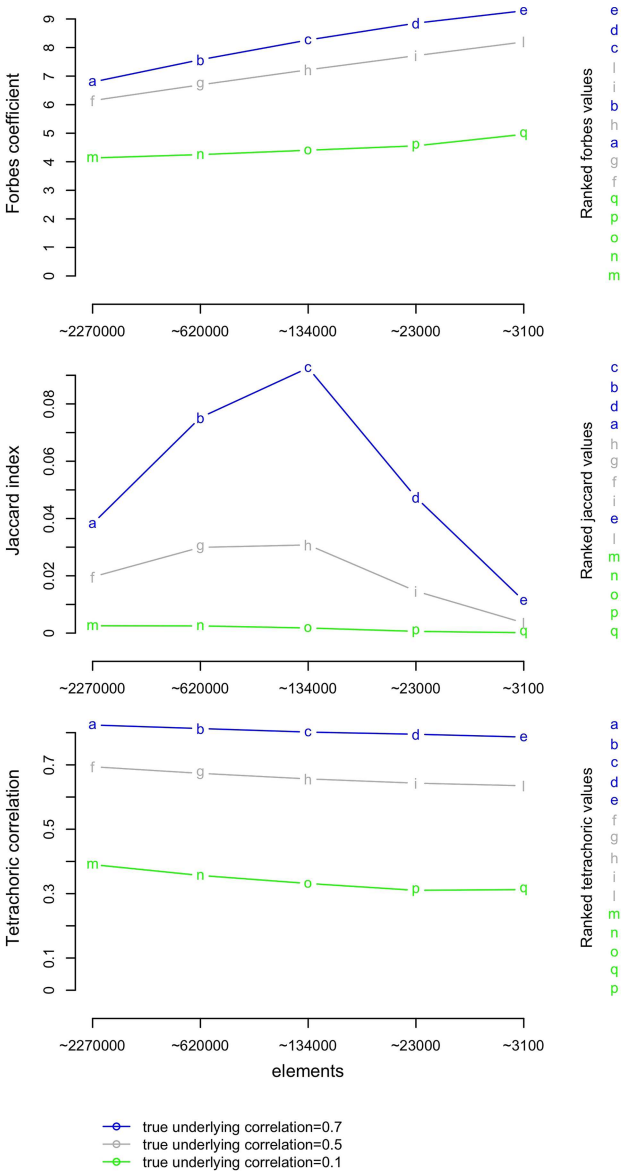

Figure 2. The behaviour of the three considered similarity **measures** on simulated datasets with size varying according to Model assumption 2. Top panel: simulation study showing the Forbes coefficient between a fixed binary query track *q* and binary reference tracks *rs* corresponding to different thresholding of a fixed, continuous reference track (*y*). We simulated under three different levels of covariance for the underlying bi-normal process giving rise to *q* and *y*, with covariance set to 0.1 (green), 0.5 (grey) and 0.7 (blue). The *x*-axis shows the number of occurrences in the resulting *rs* after applying different threshold levels for the discretisation of *y*. The *y*-axis shows the Forbes coefficient on the discretised *rs* against the fixed query track *q*. The middle and bottom panels show the same for the Jaccard index and the tetrachoric correlation.

As expected, the estimates have high uncertainty for small dataset sizes for the three **measures** (Figure 1). The expected Forbes coefficient is constant across dataset sizes and the estimated similarity may take on extreme values (either low or high) for small datasets. A similar tendency is observed for the tetrachoric correlation. Since the Jaccard index is systematically low for small datasets, the effect is less pronounced.

The same properties were found when simulating according to 50% FDR (Supplementary Figure C1a). Furthermore, Supplementary Figures C1b and c show corresponding analyses on two additional similarity measures, the Sørensen–Dice similarity index [17, 18] and the Pearson correlation [19], with both showing influence by size in a manner similar to the Jaccard Index.

Simulated data under Model assumption 2

Figure 2 shows how the three considered **measures** behave when datasets are simulated according to Model assumption 2. Datasets of varying sizes were generated by applying different thresholds on a fixed continuous base reference track (*y*), while a fixed threshold was applied to the underlying continuous

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

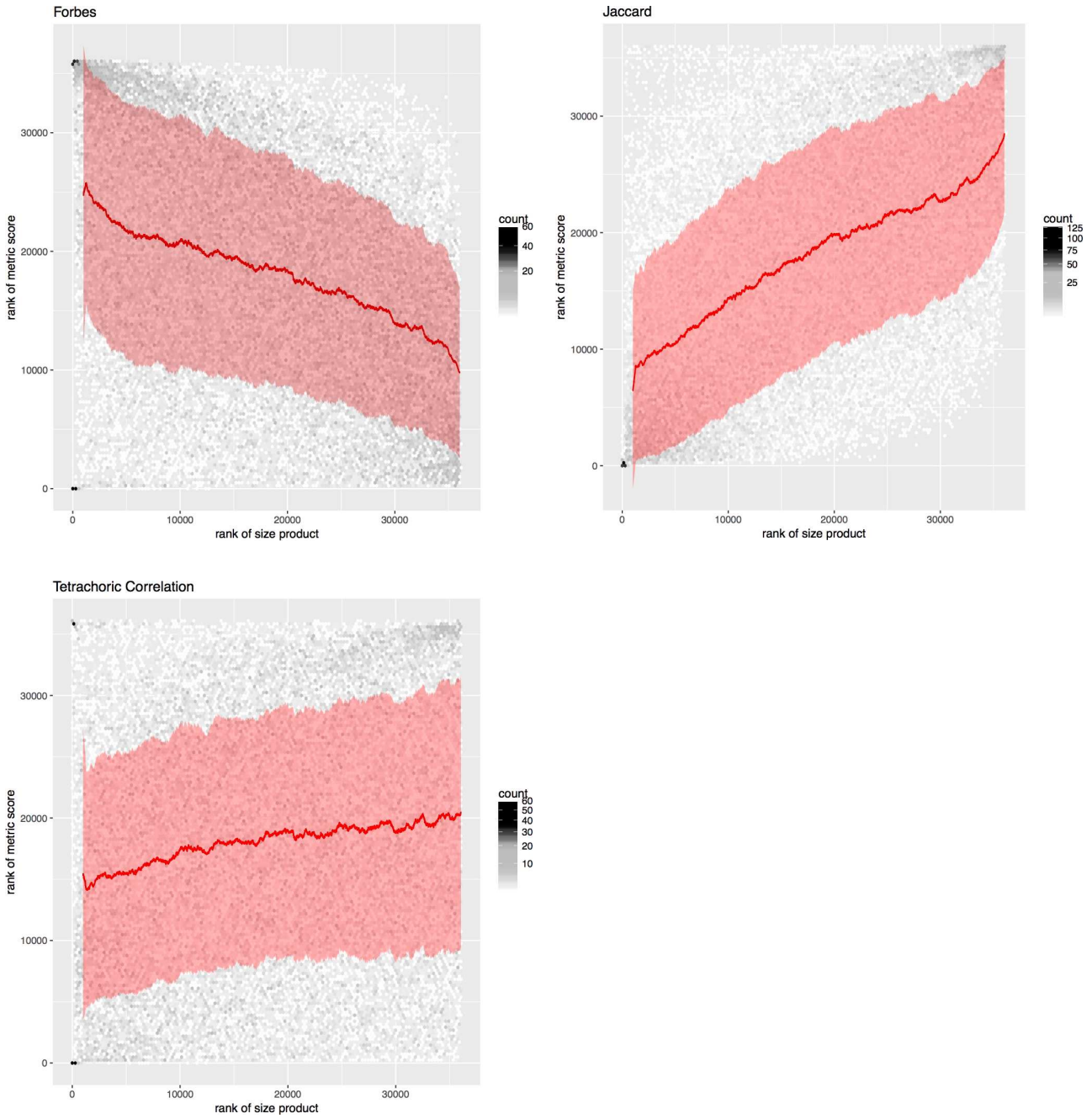

Figure 3. Large-scale empirical analysis of TF co-occurrence assessed by the three considered measures on ChIP-seq datasets of varying size. Top panel: empirical analysis of the Forbes coefficient between 36,046 pairs of experimental (ChIP-seq) datasets for 269 tracks representing binding occurrences for 15 different TFs. The *y*-axis shows the rank of Forbes coefficient, while the *x*-axis gives the rank of the product of track sizes for the two experimental datasets compared. The red line and areas show **the rolling averages, and the rolling averages +/- rolling** standard deviations with a window size of 1000. The top and bottom panels show the same for the Jaccard index and the tetrachoric correlation, respectively.

underlying reality X resulting in the query track q. The tetrachoric correlation (bottom panel) provides a stable and unbiased estimate of the underlying correlation even when applying stringent thresholds (tracks with few occurrences). In contrast, the Forbes coefficient shows a systematic increase with stringent thresholds. The Jaccard index shows a very strong systematic pattern. Interestingly, the threshold level that gives rise to the peak value of the Jaccard index is dependent on the threshold used to construct the fixed query track (details in Supplementary Material). **The same figure was made also for the Sørensen-Dice similarity index and the Pearson correlation (Supplementary Figure C2).**

Empirical analysis on a large collection of TF ChIP-seq peak datasets

In the simulation studies, we explored how the three considered **measures** behave given a particular assumption about the generation of the data. Although consideration of these alternative assumptions helps in choosing an appropriate **measure**, the appropriate underlying assumption is usually not obvious in real settings. To investigate the behaviour of the different **measures**

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

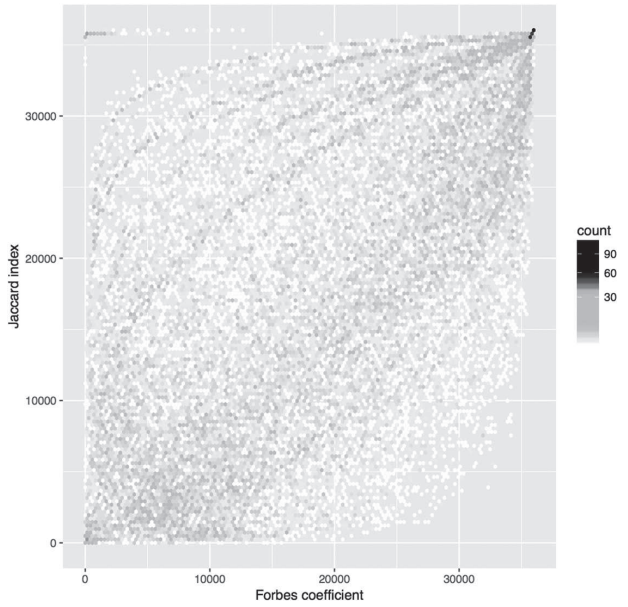

Figure 4. Scatter plot of rankings of large-scale empirical analysis of TF co-occurrence assessed by the Forbes coefficient and Jaccard index on ChIP-seq datasets. The y-axis shows the rank according to the Jaccard index, while the x-axis gives the rank (where a rank equal to one is the least similar) according to the Forbes coefficient for 36 046 pairs of experimental (ChIP-seq) datasets for 15 different TFs.

on real data, we performed a large-scale study on ChIP-seq datasets for different TFs. Specifically, we have assessed how the estimated similarity for all pairwise combinations of TF ChIP-seq experiments is affected by the size of the datasets. Figure 3 shows how each of the three considered similarity **measures** varies with respect to the size of the experimental datasets. The Jaccard index is consistently low for small datasets and increases with the size of the dataset. The Forbes coefficient and the tetrachoric correlation appear to be less influenced by the track size. The Forbes coefficient shows a downward trend while the tetrachoric correlation shows an upward trend as the track size increases. This suggests that neither Model 1 nor a binormal version of Model 2 perfectly models the size variation of the ChIP-seq datasets. Nevertheless, both model assumptions are close and the reality appears to be somewhere in between the two models. For small tracks, the Forbes coefficient shows a higher variability compared to the Jaccard index and to the tetrachoric correlation. The different behaviours with respect to dataset sizes indicate that the choice of **measure** has a substantial influence on the results. To directly address this question, we compared the ranking of the 36 046 dataset pairs from 269 datasets by co-occurrence when assessed by either the Jaccard index or the Forbes coefficient. Indeed, these two **measures** gave highly different rankings. The Spearman correlation for the ranks was 0.54, and 21 of the top 100 dataset pairs for one **measure** were among the top 100 for the other **measure** (see Figure 4 for a scatter plot of ranks according to Forbes and Jaccard). Additional figures on the agreement between Jaccard and the tetrachoric correlation, and Forbes and the tetrachoric correlation, are provided as in Supplementary Figures C4 and C5).

To further illustrate how Jaccard and Forbes can result in a different conclusion being made when applied on real data, we re-ran a previously published experiment on the co-occurrence of Crohn's disease-associated variants (genome-wide association study (GWAS) of single nucleotide polymorphisms (SNPs)) and DNase hypersensitive sites (DHSs) [20]. In the original study by Maurano *et al.* [21], Forbes was used to measure the co-occurrence of the SNPs with DHSs from different cell types, with the conclusion being that in the case of Crohn's disease, T cell subtypes (therefore, immune cells) show the most-significant GWAS variants in their DHSs (Figure 5, top panel). When we apply Jaccard on the same dataset, we find the complete opposite trend with similarity increasing with DHSs track size (Figure 5, bottom panel). **Showing that, the choice of the similarity measure strongly affects the biological conclusions that can be drawn.**

Discussion and conclusion

Here, we have performed an in depth assessment of a central, but often overlooked component of genomic co-occurrence analyses: the similarity **measure** used to evaluate and compare the degree to which different genomic features co-occur along the genome. Since the number of occurrences may vary widely between experimental datasets, the raw base pair overlap between genomic tracks must be normalised by means of a similarity **measure**. Two commonly used **measures** are the Jaccard index and the Forbes coefficient. Through modelling considerations, we propose that the tetrachoric correlation could also be employed to assess genomic co-occurrence.

We here show that the often overlooked choice of co-occurrence **measure** may lead to highly disparate rankings of similarity. To better understand what distinguishes the different **measures,** we explore their behaviour with respect to datasets size variation according to models assuming either a binary or continuous underlying reality. We show through analytical derivations and simulation that the Forbes coefficient is preferred if the underlying biological feature is assumed to be binary. If an experimental dataset is rather assumed to represent thresholded events from an underlying continuous reality, the tetrachoric correlation may be preferred. A large-scale empirical analysis of TF co-occurrence indicates that the tetrachoric correlation is the **least influenced by** size variation of experimental datasets, while the size variation is highly prone to dominate biological signal if the Jaccard index is used. Nonetheless, one should beware that estimates based on the Forbes coefficient or the tetrachoric correlation have high uncertainty for small tracks. For the tetrachoric correlation, this effect is particularly strong for tracks with mean correlation values close to zero.

While simulated datasets provide full control of the underlying truth, the resulting biases are highly dependent on the models assumed for the simulation. In contrast, the large-scale empirical study shows how the **measures** behave under real data generation processes of the considered genomic features (transcription factor binding to DNA). However, for real biological features we only have access to the imperfect experimental observations and have no way of isolating size-associated biases from the true underlying similarity. The strong correlation between the dataset size and the Jaccard index value observed in the TF study could thus, in principle, arise because large datasets mostly represented TFs that were truly more similar (in terms of their true binding locations). Although unlikely, we addressed this possibility in Supplementary Figure C3 by normalising the similarity value for a given pair of experimental datasets against the average similarity for all pairs of experimental datasets for the same pair of TFs. This would thus cancel out any effect due to larger datasets representing more similar TFs. The results after doing this correction were very well in line with those

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

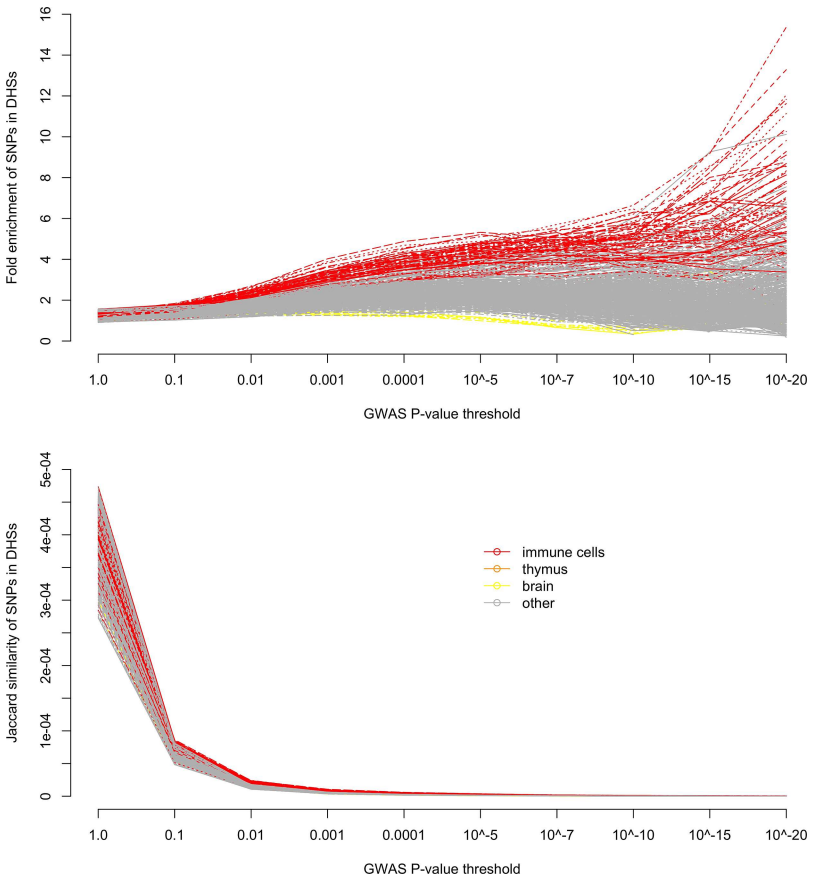

Figure 5. Co-occurrence of Crohn's disease-associated variants (GWAS SNPs) and DHSs. The top panel, shows the co-occurrence of the SNPs with DHS from different cell types by Forbes (fold enrichment), while the bottom panel shows the same analyses by Jaccard.

observed in the main figures (for all three **measures**). An even subtler point is that large experimental datasets could represent contexts (cell type and cell condition) for which the TFs were systematically behaving more similar to each other, so that even correcting for the pair of TFs would not cancel out variation in true similarity. Since such context would mainly be unspecified, it is not possible to control for. Nonetheless, there is no a priori reason to believe that there would be such systematic associations, and we consider it very unlikely that it could have a strong influence on the results. We also note that the bias we see for the Jaccard index in the empirical study is in line with what we see in the controlled simulations. Additionally, when applying the Forbes coefficient and the Jaccard index to measure the similarity between Crohn's disease GWAS SNPs and DHSs of specific cell types, as published by [20], we obtain very different results depending on which similarity **measure** is being used. When assessing colocalization by the Forbes coefficient (as in Maurano *et al.*), DHS enrichment is increasing with higher SNP significance in the presumably relevant immune cell types (Figure 5, top panel). In contrast, a corresponding assessment based on the Jaccard index (Figure 5, bottom panel) shows a strongly decreasing value of colocalization with higher SNP significance. This non-intuitive relation arises as an artifact due to the lower track size when focusing only on highly significant SNPs.

Our analysis has specifically focused on how size variation of datasets affects similarity **measures**. While there are differences between **measures** in terms of the range of values they can take and their ease of interpretation, we see the behaviour with respect to dataset size variation as their defining property. Indeed, in terms of ranking the strength of relations, neither the range nor the interpretation of a **measure may matter**; its role is to define an ordering of the discrete points in a three-dimensional space, where the size of the two datasets and their intersection make up the three dimensions. Also when a measure is used as test statistic for Monte Carlo-based statistical testing of genomic colocalization, the resulting ordering affects significance [22].

While we have here focused on the setting of genomic co-occurrence, we believe our approach to be equally valuable also for other fields in which similarity **measures** are used. This includes questions related to species presence in ecology, which is the field in which several of the considered **measures** were originally proposed. The approach could also be useful for other questions in genomics where similarity assessment is central, such as gene set enrichment analysis or various applications of clustering to gene expression or genomic location values. By considering the proposed model assumptions in light of a particular analytical setting, a reasoned choice of a **measure** can be made. As in our analysis, one should not expect a clear-cut answer, but rather a nuanced conclusion of some **measures** appearing more suitable than others.

In conclusion, with this study we did not aim to find 'the solution' to the similarity problem, since we believe that this may be unattainable, but we find that the choice of similarity measure is an important and often overlooked aspect when

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020

analysing genomic co-occurrence. Interestingly, although normalising for size variation is the primary purpose of a similarity **measure**, we find that the Jaccard index is strongly affected by dataset size, both according to underlying models of size variation and in an empirical investigation. The Forbes coefficient and the tetrachoric correlation show less systematic association with dataset size, and can each be connected to a reasonable underlying model of size variation. Both of these **measures** may work satisfactorily in a given context, though one should be aware that estimation uncertainty may lead to extreme values for small datasets. Despite its limited current usage in the field, the tetrachoric correlation indeed shows the most desirable behaviour in our analyses. Ideally, we recommend to explicitly choose between similarity **measures** based on a consideration of which of the two underlying models of size variation appear most reasonable for a given biological context. For tools that currently rely on the Jaccard index to quantify genomic co-occurrence, we recommend to include an option of using alternative **measures** like the Forbes coefficient and the tetrachoric correlation.

Key Points

- The often overlooked choice of co-occurrence **measure** in genomic colocalisation analysis may lead to highly disparate rankings of similarity;
- The commonly used Jaccard index should be avoided or used with strong caution in analyses of genomic colocalisation;
- Size variation of datasets affects different similarity **measures** with different degrees.

Acknowledgments

We would like to thank Sveinung Gundersen for his invaluable help with creating the Galaxy page and instance for this specific project.

Funding

Not applicable.

References

1. Kent WJ, Sugnet CW, Furey TS, *et al*. The human genome browser at ucsc. *Genome Res* 2002;**12**(6):996–1006.
2. The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature* 2012; **489**:57–4.
3. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al*. The nih roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010;**28**:1045–8.
4. Lonsdale J, Thomas J, Salvatore M, *et al*. The genotype-tissue expression (gtex) project. *Nat Genet* 2013;**45**:580–5.
5. Jaccard P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* 1901;**37**:547–79.
6. Forbes SA. On the Local Distribution of Certain Illinois Fishes: An Essay in Statistical Ecology, Vol. **7**. Illinois State Laboratory of Natural History, 1907.
7. Chao A, Chazdon LR, Colwell KR, *et al*. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* 2005;**8**(2): 148–59.
8. Wolda H. Similarity indices, sample size and diversity. *Oecologia* 1981;**50**(3):296–302. ISSN 1432-1939. DOI: 10.1007/BF00344966.
9. Quinlan AR. Bedtools: the swissarmy tool for genome feature analysis. *Curr Protoc Bioinformatics* 2014;**47**(1):11.12.1–11.12.34. DOI: 10.1002/0471250953.bi1112s47.
10. Quinlan AR, Kindlon N. Bedtools: jaccard.
11. Worsley Hunt R, Wasserman WW. Non-targeted transcription factors motifs are a systemic component of chip-seq datasets. *Genome Biol* 2014;**15**(7):412. DOI: 10.1186/s13059-014-0412-4.
12. Hamedani GG, Tata MN. On the determination of the bivariate normal distribution from distributions of linear combinations of the variables. *Am Math Mon* 1975;**82**(9):913–5. ISSN 00029890, 19300972. URL: http://www.jstor.org/stable/2318494.
13. Pearson K. I. mathematical contributions to the theory of evolution. –vii. on the correlation of characters not quantitatively measurable. *Philos Trans R Soc Lond A Math Phys Eng Sci* 1900;**195**(262–273):1–47. ISSN 0264-3952. DOI: 10.1098/rsta.1900.0022.
14. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria, 2016. URL: https://www.R-project.org/.
15. Fox J. *Polycor: polychoric and polyserial correlations. R package version 0.7–8*. Online [URL] http://www.cran.r-project.org/web/packages/polycor/index.html (31 August 2010, date last accessed), 2010.
16. Gheorghe M, Artufel M, Cheneby J, *et al*. Remap 2018: an updated atlas of regulatory regions from an integrative analysis of dna-binding chip-seq experiments. *Nucleic Acids Res* 2018;**46**(D1):D267–75. DOI: 10.1093/nar/gkx1092.
17. Dice RL. Measures of the amount of ecologic association between species. *Ecology* 1945;**26**(3):297–302.
18. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol Skr* 1948;**5**:1–34.
19. Pearson K. Mathematical contributions to the theory of evolution. iii. Regression, heredity, and panmixia. *Philos Trans R Soc Lond Ser A Cont Pap Math Phys Character* 1896;**187**: 253–318. ISSN 02643952.
20. Franke A, McGovern DP, Barrett JC, *et al*. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet* 2010;**42**:1118–25.
21. Maurano MT, Humbert R, Rynes E, *et al*. Systematic localization of common disease-associated variation in regulatory dna. *Science (New York, NY)*, **337**(6099):1190–5. DOI: 10.1126/science.1222794.
22. Kanduri C, Bock C, Gundersen S, *et al*. Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 2019;**35**(9):1615–24. https://doi.org/10.1093/bioinformatics/bty835.

Downloaded from https://academic.oup.com/bib/article/21/5/1523/5586919 by University of Oslo Library user on 29 October 2020