IDR analysis 101
Measuring consistency between replicates in high-throughput experiments

Qunhua Li, Ben Brown, Haiyan Huang, Peter Bickel

March 5, 2010

Abstract

Consistency analysis (IDR analysis) has been used to process ENCODE and modENCODE production dataset. This is a non-technical introduction on the statistical method of the consistency analysis (a.k.a. IDR analysis). It describes the consistency analysis on a pair of replicates.

1 Introduction

In large-scale genomic studies, it is common that a vast set of objects (e.g. genes or binding sites) is studied and a small set of objects is selected based on a significance measure (e.g. p-value, q-value or fold of enrichment) for furthur study. However, due to the complexity of the data, the significance measures often are not well-calibrated or their scales may vary across dataset, hence the choice of threshold often involves arbituary judgement. If replicates are available, one possibility is to use consistency between replicates to select signals, because genuine signals are supposed to be reproducible between replicates. In this work, we propose a statistical method to quantitatively measure the consistency between replicates and select signals with the reproducibility of signals into account.

Our method does not require the significance of signals to be well-calibrated or with a fixed scale, and only assumes they reasonably reflect the relative ranks of the significance of signals. Thus it can be used to select signals even when the significance scores are heuristic-based. As can be seen later, the quantitative summary of reproducibility estimated from our method can be used to compare reproducibility of different methods or be used as an internal measure for quality control.

The intuition behind our method is the following. If two replicates measure the same underlying biology, the significant identifications, which are likely to be genuine signals, are expected to have high self-consistency; whereas, the identifications with low significance, which are likely to be noise, are expected to have low self-consistency, due to the reshuffling of ranks at noise level. Thus, if the consistency between a pair of rank lists that contains both significant and insignificant findings is plotted along the decrease of significance, a change of consistency is expected. This change of self-consistency provides an internal indicator of the transition from signal to noise and indicates how many identifications we can trust.

Our method includes three components:

1. A correspondence curve: It is a graphical tool for visualizing and inspecting the reproducibility of the replicates and the transition from reproducible to irreproducible signals. It provides a quick graphical view of the data without making any model assumptions. It is a convenient tool for diagnosis and quality control, but not adequate for selecting signals. It is independent of the other two components.

2. An inference procedure: It quantitatively summarizes the consistency structure, including the proportion of reproducible and irreproducible signals, how strong the association between replicates is for the reproducible group, and how separate the reproducible group is from the irreproducible group. It also assigns each signal a probability to be from the irreproducible group.

3. Irreproducible Discovery Rate (IDR): It is a reproducibility criterion derived from the inference procedure (#2) in a fashion similar to FDR, and can be used to control the level of irreproducibility rate when selecting signals.

Note: Because the model for (#2 and #3) can only process signals appearing on both replicates, the signals that only appear on one replicate are removed and are not assigned the reproducible level. However, the correspondence curve (#1) can be applied to all the signals, regardless if they appear on a single or multiple replicates.

In what follows, we present a non-technical description on these methods and illustrate the methods with an example that compares nine peak callers on a CTCF ChIP-seq data from ENCODE.

2 Methods

2.1 A correspondence curve to describe the change of self-consistency

To visualize how consistency changes in the decreasing order of significance, we first define the level of consensus ($\Psi(t)$) at the top $t\%$ identifications as the proportion of common identifications that are ranked on the top $t\%$ on both replicates. We denote the derivative of $\Psi(t)$, which reflects the local change of consistency, as $\Psi'(t)$. This definition provides distinct characteristics for the two special cases that we care about, i.e. no correspondence and perfect correspondence. At no correspondence, $\Psi(t) = t^2$ and $\Psi'(t) = 2t$; at perfect correspondence $\Psi(t) = t$ and $\Psi'(t) = 1$.

Then we sequentially compute this consensus in the order of decreasing significance (i.e. t from 0 to 1), and obtain the curve by plotting the consensus along the rank list. Because good correspondence and poor correspondence have distinct characteristics in this curve, the transition of the level of correspondence is reflected from the transition of the shape of the correspondence profile. In the curve of Ψ, the transition to breakdown is shown as the transition from a line with slope 1 to a parabola; and in the curve of Ψ', the transition to breakdown is shown as the transition from a line with slope 0 to a line with positive slope.

Figure 1 shows the correspondence profile for an idealized case, where the ranks of the top 50% identifications are perfectly correlated and the ranks of bottom 50% identifications have no correlation. By comparing the empirical curve to the two special cases, one can read out consistency information in the data. For example, the closeness to perfect correspondence gives a sense on how consistent the identifications on different replicates are; and the point to reach breakdown provides guidance on how much identifications we can trust. In Figure 1, the transition to a line segment with a positive slope occurs at about 50% of the data, which agrees with the breakdown of consistency in the data.

The correspondence curve on ENCODE ChIP-seq data

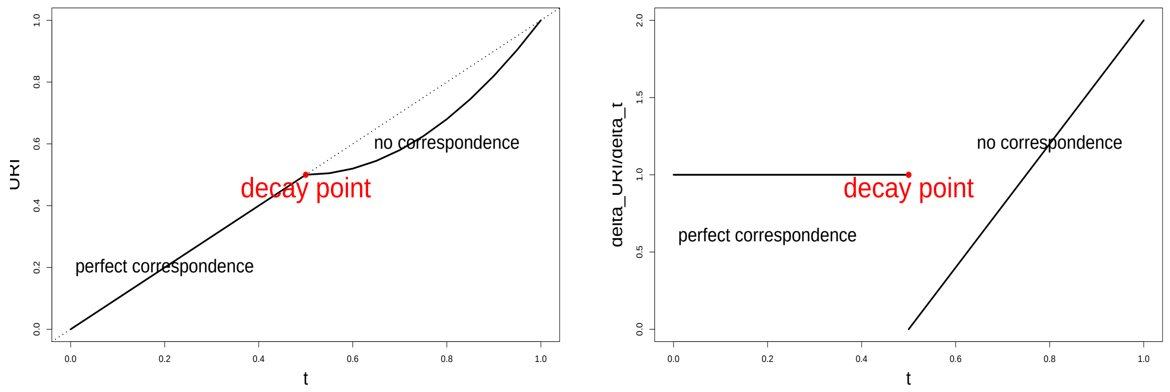
We now illustrate our approach using an example from ENCODE, where the reproducibility of several peak callers are assessed on a pair of biological replicates (CTCF ChIP-seq data). Apparently, one wishes to focus on the consistency of significant peaks here and minimize the effect due to the choices of initial cutoffs, which are not comparable for different callers. Based on the correspondence profile (Figure 2), we can rank the reproducibility on this dataset for peak callers, according to the number of peaks identified before the breakdown of consistency. For simplicity, we only consider the peaks identified on both biological replicates here. The three peak callers that report the largest number of reproducible peaks on the data are Peakseq, MACS and SPP. (Note that this only reflects the behaviors of the peak callers on *this* dataset for the selected parameter settings.)

2.2 Quantification of the consistency structure

Though we can visualize the change of consistency using the consistency curve, a quantitative summary will be helpful for pinpointing the threshold and and assessing the error associated with each threshold. To this end, we propose a copula mixture model, which aims to answer the following questions:

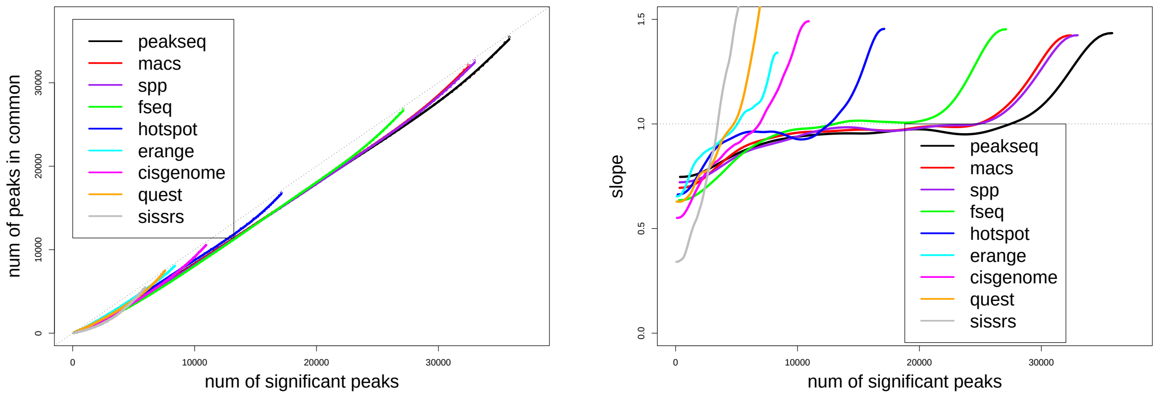
Q1. What proportion of identifications have a poor correspondence, i.e. falling into ”noise”?

Q2. How consistent are the identifications before reaching breakdown?



a. Ψ b. Ψ'

Figure 1: An illustration of the correspondence profile in an idealized case. In this case, two calling results have perfect correspondence for top 50% peaks and no correspondence for bottom 50% peaks. The transition of correspondence can be identified from the curve. a. the consistency of selected identifications; b. the change of consistency along the rank list, which is the derivative of a. In both a and b, X-axis is the proportion of selected identifications on the rank lists



Ψ Ψ'

Figure 2: The correspondence profile along the decrease of significance, plotted for 9 peak callers on a CTCF Chip-seq experiment from ENCODE. Left: the upper rank intersection at different cutoff. X-axis: the number of peaks identified on each replicate. Y-axis: the number of peaks that are commonly identified on both replicates. Right: the derivative of the upper rank intersection at different cutoff.

Q3. What is the error probability at a threshold selected according to consistency?

Our model assumes data consists of two groups, a reproducible group and an irreproducible group. In general, the signals in the reproducible group are more consistent (i.e. with larger correlation coefficient) and are ranked higher than the irreproducible group. As the significance scores are only accurate up to their ranks, we propose a copula mixture model, which allows the scales of significance scores vary across different datasets and models the heterogeneity of consistency, to cluter the signals into a reproducible group and an irreproducible group. The proportion of identifications that belongs to the ”noise” component (Q1) and the correlation of the significant component (Q2) can be estimated adaptively from the data. It also provides an idr score for each signal, which reflects the posterior probability for the signal to belong to the irreproducible group.

In essence, this model clusters signals by combining the information on the ranks of significance scores and the consistency between replicates, thus borrows strength on both replicates for selecting reliable signals.

Technical note: Copula models provide a way to model the marginal distribution nonparametrically and association structure in a parametric way. So it allows us to model the consistency between replicates without knowning the scale and distribution of significance scores.
Mixture models provide a way to combine heterogenous groups into one model and estimate proportion and membership of the groups adaptively. Here we combined the two models to separate the signals by the joint information of their ranks and consistency.

2.3 A reproducibility criterion (IDR) to select signals

To provide an error probability at a threshold selected according to consistency (Q3), we define a reproduciblity criterion based on the copula mixture model, called *Irreproducible Discovery Rate* (IDR), in a fashion analogous to the false discovery rate (FDR). This is made possible because the two-group structure in the copula mixture model, namely, reproducible group vs irreproducible group, indeed bears some similarities to the two-group setting in the context of FDR, namely, interesting finding vs uninteresting finding. Similar to FDR, IDR describes the expected probability that the selected signals come from the “error” group for a given threshold; however, the “error” group for IDR refers to the *irreproducible* group, instead of the uninteresting findings in FDR.

By plotting the IDR, which can be obtained from the estimation of the copula mixture model, at various cutoffs, one can generate an ROC-like curve to describe the tradeoff between the number of signals selected and the average probabilities that the selected signals belong to the irreproducible component. Users can read out the expected irreproducible rate at various number of selected signals from the plot, and determine the cutoff according to his or her own choice.

Using simulation studies, our method shows substantial gain on discriminative power over the FDR method that only selects signals based on information from a single replicate (Figure 3), even when the correlation between replicates are weak. We attribute this gain to our method’s effective combining information across replicates.

Remark: The selection made by IDR criterion is a combined results of ranking of the significance scores on individual replicates and consistency between replicates. It is not equivalent to thresholding the significance scores on the replicates. For example, signals that have consistent rankings on both replicates but moderately ranked may be selected before the signals that have a very high score on one replicate but low on the other.

Application on ENCODE ChIP-seq data

We then apply the copula mixture model to the same ENCODE data and estimate the parameters and membership for each signal. Figure 4 shows the irreproducible discovery rate (IDR) of the selected peaks at various cutoffs. For example, among the top 25000 peaks ranked by IDR criterion, Peakseq, MACS and SPP have about 5% irreproducible peaks. Based on the level of irreproduciblity that one is willing to accept, we may determine the number of peaks to select. This plot can also be used to compare reproducibility of different callers. As in this type of experiments one would usually hope to identify as many real calls as possible before false calls are made, we may rank the reproducibility of the

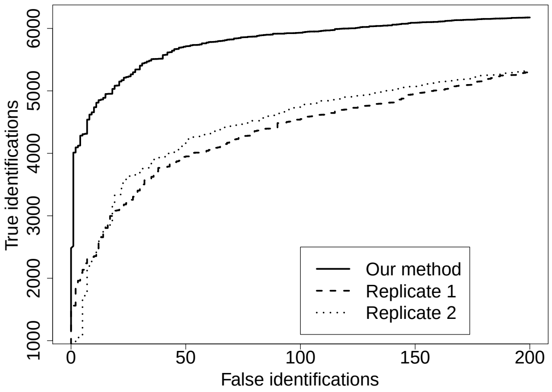


Figure 3: The number of correct and incorrect calls made at various thresholds in simulation studies. Incorrect calls: the number of spurious signals assigned posterior probabilities exceeding the thresholds (our method) or with p-values smaller than the cutoffs (individual replicates). Correct calls: the number of genuine signals assigned posterior probabilities exceeding the thresholds (our method) or with p-values smaller than the cutoffs (individual replicates). Simulation parameters are generated from the parameters estimated from a real dataset using our model.

peak callers for the peaks identified on both replicates based on Figure 4. The reproducibility ranking from Figure 4 is similar to the ranking from the correspondence profiles (Figure 2).

3 Processing production data

This method has been used to process the ENCODE production data. The processing so far includes two parts: first run the consistency analysis between replicates, then apply the parameters learned from the copula mixture model to the pooled data and select peaks on the data pooled from replicates. As the processing of pooled data is not the core part of the consistency analysis and involves heuristics that are still under development, it is not included in this writeup.

In what follows, I will use an example to show how to get a quick sense of the consistency between replicates. Figure 5 shows a series of plots generated from consistency analysis.

The first column shows the correspondence profile between *all* the peaks that are called on the replicates. The second column shows the correspondence profile between *all* the peaks that are called on both replicates. The sooner the transition occurs, the sooner the breakdown is. The closer the fragment before transition is to the diagnal line in Ψ plot or the line $y = 1$ in Ψ‘, the highly associated the two replicates are before the consistency breakdown. If two replicates have many peaks only called on one replicate, it can be reflected from the difference between the two sets of plots. The upper right plot shows IDR at different numbers of selected peaks by IDR criterion. If IDR increases quickly at a small number of peaks, it indicates the replicates have poor consistency.

As shown, Figure 5a has much better consistency than Figure **??**b. The first 2000 peaks in (a) are fairly consistent.

4 Summary

Here we summarize several features and possible uses of this method.

- This method clusters signals based on the combined information of ranks of significance scores and the consistency between the replicates.
- As it only requires the significance scores reflect reasonablely the relative ordering of significance, this method can be used for threshold selection regardless (1) if the significance scores

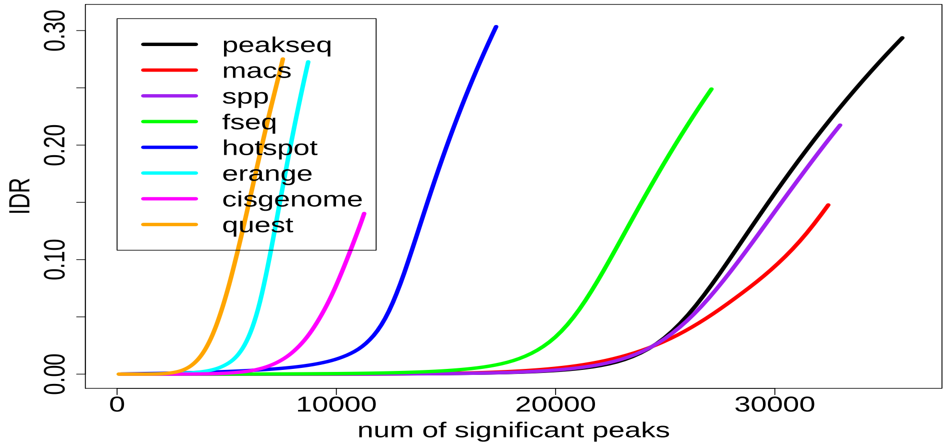


Figure 4: The average posterior probabilities that the self consistency of the identified peaks breaks down, plotted at various thresholds for 9 peak callers on a CTCF Chip-seq experiment from ENCODE. X-axis: the number of identified peaks, Y-axis: the average posterior probability that the self-consistency breaks down for the identified peaks.

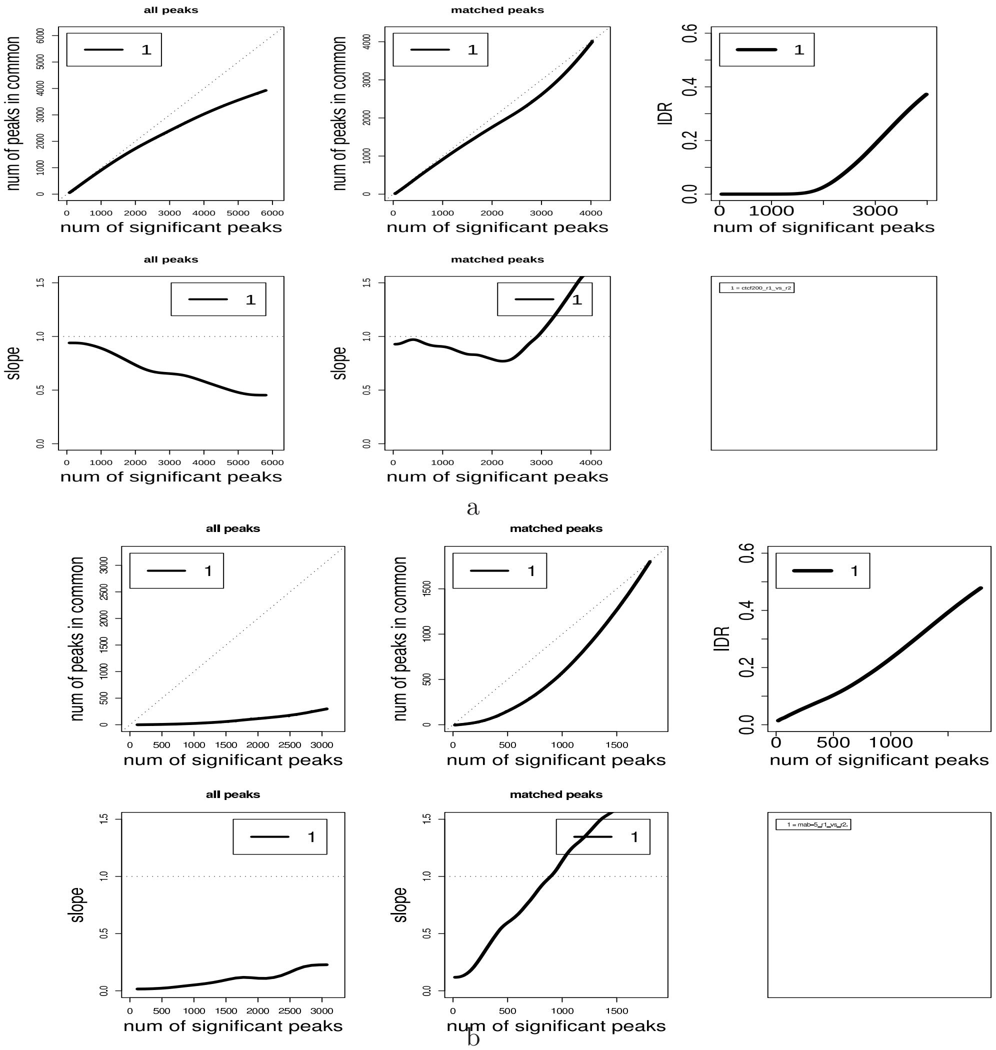
are probabilistic-based (i.e. p-value or q-value) or heuristic-based (e.g. fold of enrichment) or (2) if the scale of significance scores varies across datasets.

- Because the method estimates the proportion of ”signal” adaptively, this method is robust to the initial threshold choices to some extent.
- This method also provides a quantitative way to compare the reproducibility of different algorithms, without being sensitive to the effect due to threshold differences, as long as the breakdown point is included.
- It can be used for quality control.
- This method is not limited to the application here. It can be applied to many other high-throughput settings, as long as the significance of the signals can be ranked.

The R code for our method is available upon request.

Reference:

A longer version of this work will be available soon as a technical report (early version) in Dept of Statistics at UC Berkeley. A paper on this work will be submitted soon. For citation, you may email me at qli@stat.berkeley.edu for the technical report number.



a

b

Figure 5: The consistency plots of two datasets. a. A data set with reasonable consistency. b. A data set with poor consistency.