Journal of Next Generation Sequencing & Applications

Donald and Wilson,, Next Generat Sequenc & Applic 2016, S1:003
http://dx.doi.org/10.4172/2469-9853.S1-003

Commentary Open Access

On Evaluation of Rankings in Analysis of NGS Data

Margaret R Donald[1] **and Susan R Wilson**[*1,2]

[1]*School of Mathematics and Statistics, University of New South Wales, Kensington, NSW 2052, Australia*

[2]*Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601, Australia*

[*]**Corresponding author:** Susan R Wilson, School of Mathematics and Statistics, University of New South Wales, Kensington, NSW 2052, Australia, Tel: +61 2 6125 4460; E-mail: Sue.Wilson@anu.edu.au

Rec date: Dec 15, 2015**; Acc date:** Jan 25, 2016**; Pub date:** Jan 28, 2016

Abstract

A ranked list of genes (or proteins or regions) is a common output from analysis of NGS data. Many choices will have been made in the analysis (either explicitly or implicitly) and there is no 'correct' method to use for the analysis. So if two different and appropriate methods are used, an important question is the following: How similar are the two rankings? Allowing a looser definition of agreement than 'exact' agreement, and using a Bayesian logit model with O'Sullivan penalized splines, a useful visualisation has been developed giving the probability of agreement at each point and the credible interval at which the sequence degenerates into noise. The approach is illustrated on some typical RNA-seq data. The estimate of the point at which the agreement between the rankings degenerates into noise, as well as the credible interval, will be over-estimates of their true values. From a practical perspective, it is usually better to estimate a slightly larger set of top-ranked data than one that is smaller. Even so, the estimates found for NGS data are relatively small compared with the total length of the sequence.

Keywords: Ranked data; Comparisons of methods; Visualisation

Introduction

NGS data is collected with the aim of answering many and varied questions; see, for example Datta et al., [1]. A common aim is comparative, such as to find those genes (or proteins or regions) that are differentially expressed, or up/down regulated, between say diseased (e.g. breast cancer) and normal (control) samples. Often a ranked list of the genes (proteins, regions) is obtained. Most commonly, the ranking is with respect to p-values (increasing from the smallest). Alternatively, the ranking may be based on another list of ordered values, such as distances. A common practice is then to import the top couple of hundred genes into network analysis software (for example the actively-developed open source software Cytoscape [2]).

As part of the process of determining a ranked list, many choices have been made, either explicitly or implicitly. Two questions naturally arise. First, how stable is the ranking produced? In other words, if another set of samples were to be analysed, how similar might the ranking be? If the sample size is sufficiently large, an insight into answering this question can be obtained by using a resampling method. For example, Hall and Miller [3] considered colon microarray data, consisting of 40 tumour and 22 normal observations on 2000 genes. They give a plot of the top 30 genes (ranked by the lower tail of an estimated 90% prediction interval). They found that the 90% prediction intervals for the top 4 genes lay in the approximate range of 1 to ~50, while the 5th lay between 2 to ~115, increasing by the 15th ranked gene to about 1 to ~210, and by the 30th to the even larger range of 1 to ~380. In other words, none of the top genes are ranked exactly correctly, but the top four genes appear much more stable than the others. This is a relatively large number of observations (62), yet indicates the inherent variability in a single set of rankings. Application of this approach to much smaller sample numbers may be problematic arising from the resampling underpinning the method.

The second, somewhat related, question is the main focus of this commentary, and arises from the observation that if one had used an alternative method of analysis, say, then a different list of rankings would have been obtained. Note, that there is no 'right' way to analyse NGS data. So the question is an important one: How similar are the two rankings produced by the two methods?

A review of earlier statistical literature on ordered lists is given by Hall and Schimek [4]. Further, rather than insist on exact agreement of the ranks, they suggest (a) that we may consider two sets of ranks to be in agreement if the rankings differ by a moderate deviation. As well, they propose (b) that the rankings degenerate into noise when the probability of agreement between the rankings becomes less than 0.5. Their approach gives a point estimate. In the following we outline an alternative formulation that gives an interval around a point estimate and provides a useful visualisation of the data and the fit.

Methods

Hall and Schimek [4] suggest an algorithm for finding the point in a sequence where paired ranking agreements degenerate into noise and it has been implemented in the R-package TopKLists [5]. They make two major suggestions. First, agreement in rank need not correspond to exact agreement, instead one can allow moderate deviations. For example in a list of thousands of items, one might consider agreement less than or equal to 100, say, to be a moderate agreement in rank. So an item ranked 1234 in one list may be considered in moderate agreement (called 'agreement') with having a ranking of 1333 in the other list, but not in agreement with a ranking of 1124. In a list of 100 items, this moderate deviation parameter (or distance), δ, would be smaller, say 5 or 10. Their second major suggestion is that at some point in the sequence the agreement degenerates into noise, defined as the probability of agreement being less than 0.5.

The underpinning algorithm has several parameters, namely the distance δ; the window size, υ, for finding *pj* (the probability of agreement between rankings at the j^{th} point); and C (>0.25) that controls the moderate deviations of the probability *pj* [4].

We adopt an alternative approach based on a Bayesian logit model with O'Sullivan penalised spline bases for κ internal knots [6-9]; further details to appear elsewhere. So κ is the analogue of the window choice, υ, but we have found little sensitivity to the choice of κ. Further details, as well as R code, are available from the authors. A basic output is the figure, giving the original data, and the pointwise estimates of the median of pj and their 95% credible intervals. The medians are joined with a dashed line and the pointwise credible intervals are the shaded area. The horizontal line indicates the 95% credible interval for k, the point at which pj changes from being above 0.5 to below 0.5, that is calculated from the posterior distribution of k.

Results and Discussion

To illustrate the method we consider part of an RNA-seq data set. The data are from five patients with a rare form of blood disorder who did not respond to treatment. For each patient, a sample was taken for analysis before and after treatment. Rankings and their credible limits were found for 2974 genes using two different analysis methods, namely Mann-Whitney and negative binomial.

Figure 1 gives the plot for δ=100, and Figure 2 for δ=40 on the full sequence length of 2794. We have found our method to be most sensitive to sequence length, namely the point estimate (and credible interval) for δ=100 changed from 232 (186,280) to 224 (182,276) for length 1000 and 221(180,278) for length 750. While for δ=40, they change from 0 (0.47) to 26 (0.58) for length 1000 and 32 (0.60) for length 500. Note, the point estimate of 0 is not a function of boundary behaviour, as seen from the curve fits, and clearly indicates the need for an interval estimate besides a point estimate. Figure 3 shows the plot for length 1000. It is the point estimate that is rather sensitive to sequence length rather than the length, and position, of the credible interval.

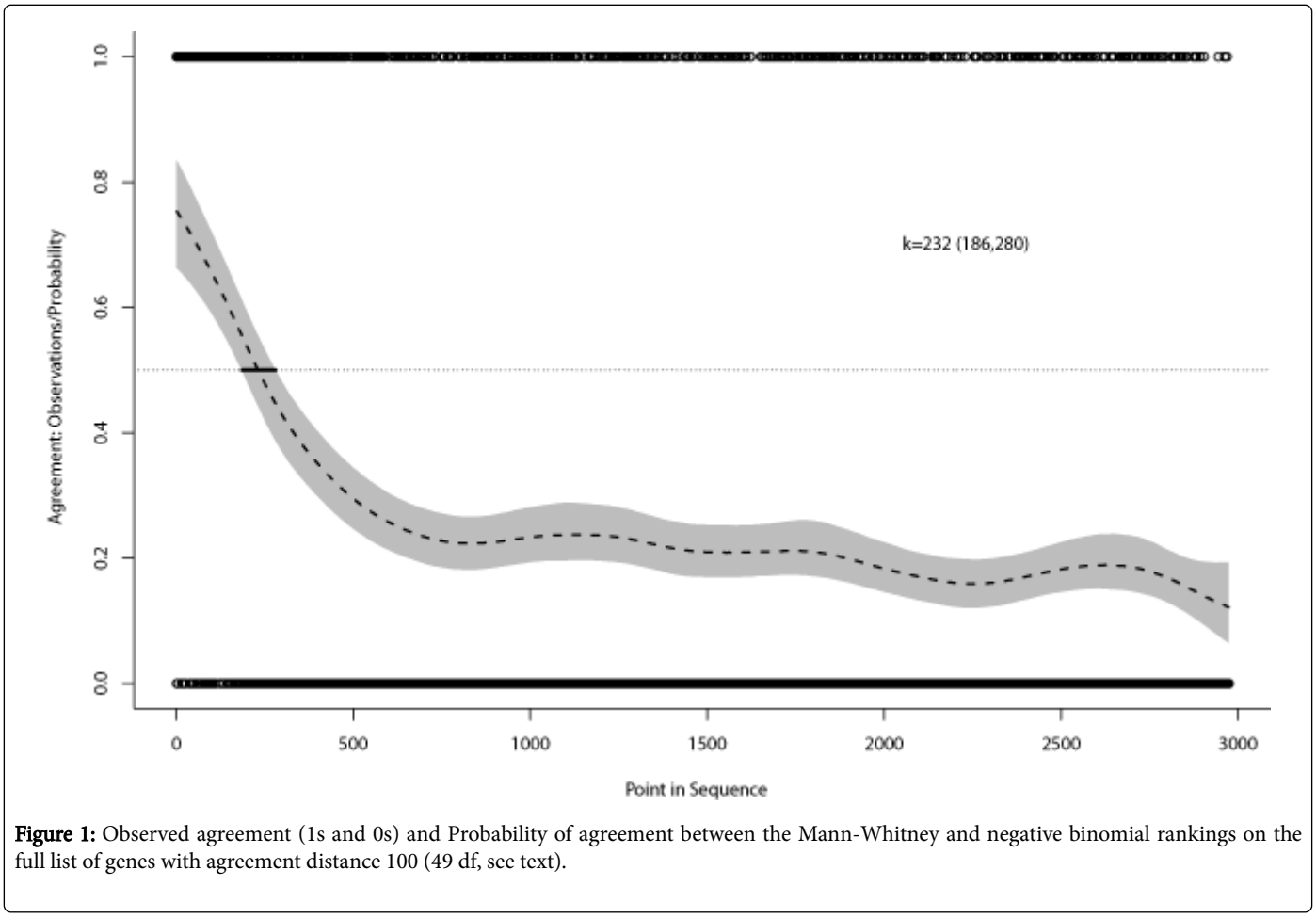

Figure 1: Observed agreement (1s and 0s) and Probability of agreement between the Mann-Whitney and negative binomial rankings on the full list of genes with agreement distance 100 (49 df, see text).

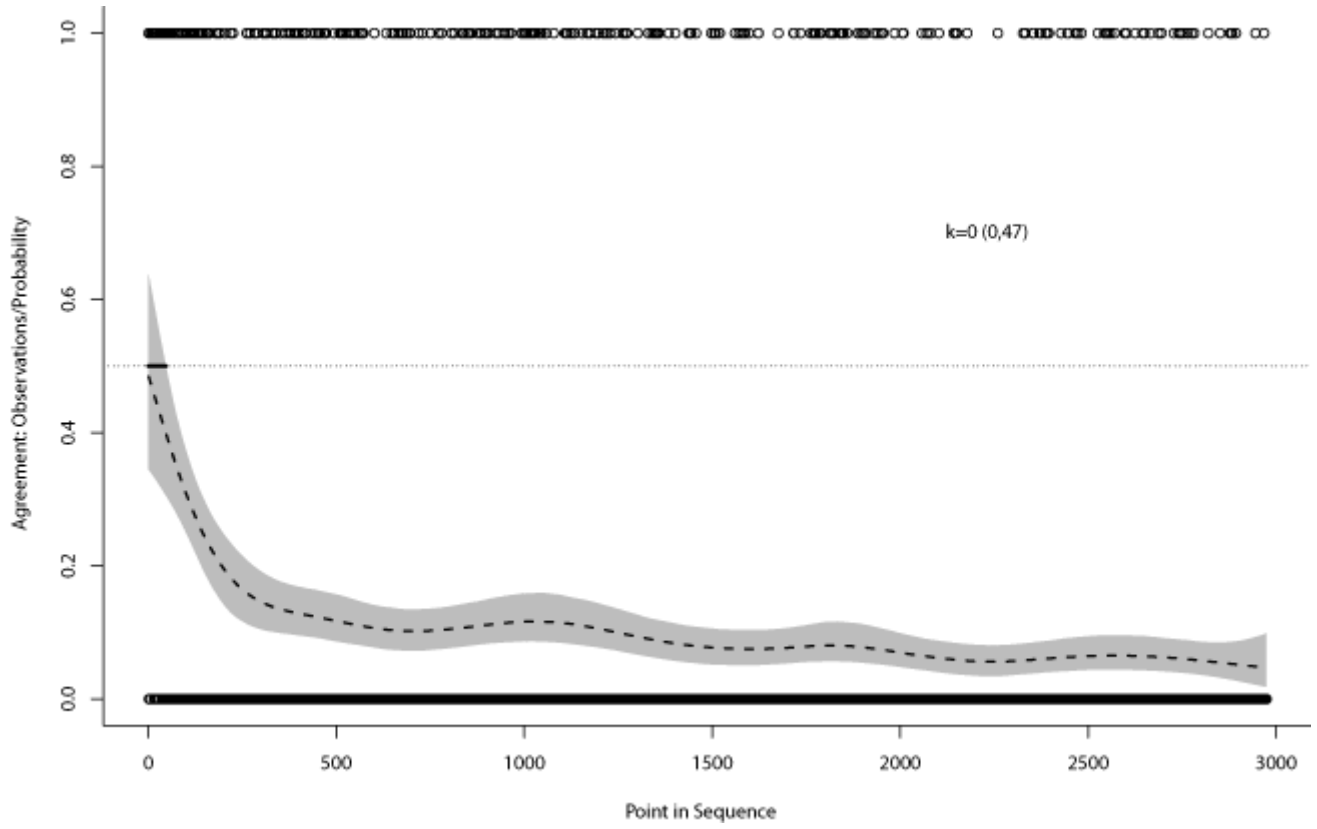

Figure 2: Observed agreement (1s and 0s) and Probability of agreement between the Mann-Whitney and negative binomial rankings on the full list of genes with agreement distance 40 (31 df, see text).

The number of knots, κ, is chosen so that there are at least five points between knots (and the degrees of freedom, df, is κ+4).

Although the point estimates given by TopKLists usually are not identical to those given by our approach, they lie within the credible interval and the difference is relatively small compared to the length of the credible interval.

The underpinning theory assumes independence, and simulations were done to evaluate behaviour assuming dependence. Overall the method was found to be robust, with a tendency to overestimate the size of the top ranked items 42.

Here, given the same data is being used and just the method of analysis is changing, one would expect the estimated size of the top set to be much larger than the true size. From a practical perspective, it is usually better to estimate a slightly larger set of top-ranked data than one that is smaller than the (unknown) true number. The estimates are actually surprisingly low, around 10% of the number of items (genes) using a relatively broad level of moderate deviation, and much less for a tighter level. Nevertheless, they are typical of other NGS data analyses we have undertaken, and are not a peculiarity of these particular data.

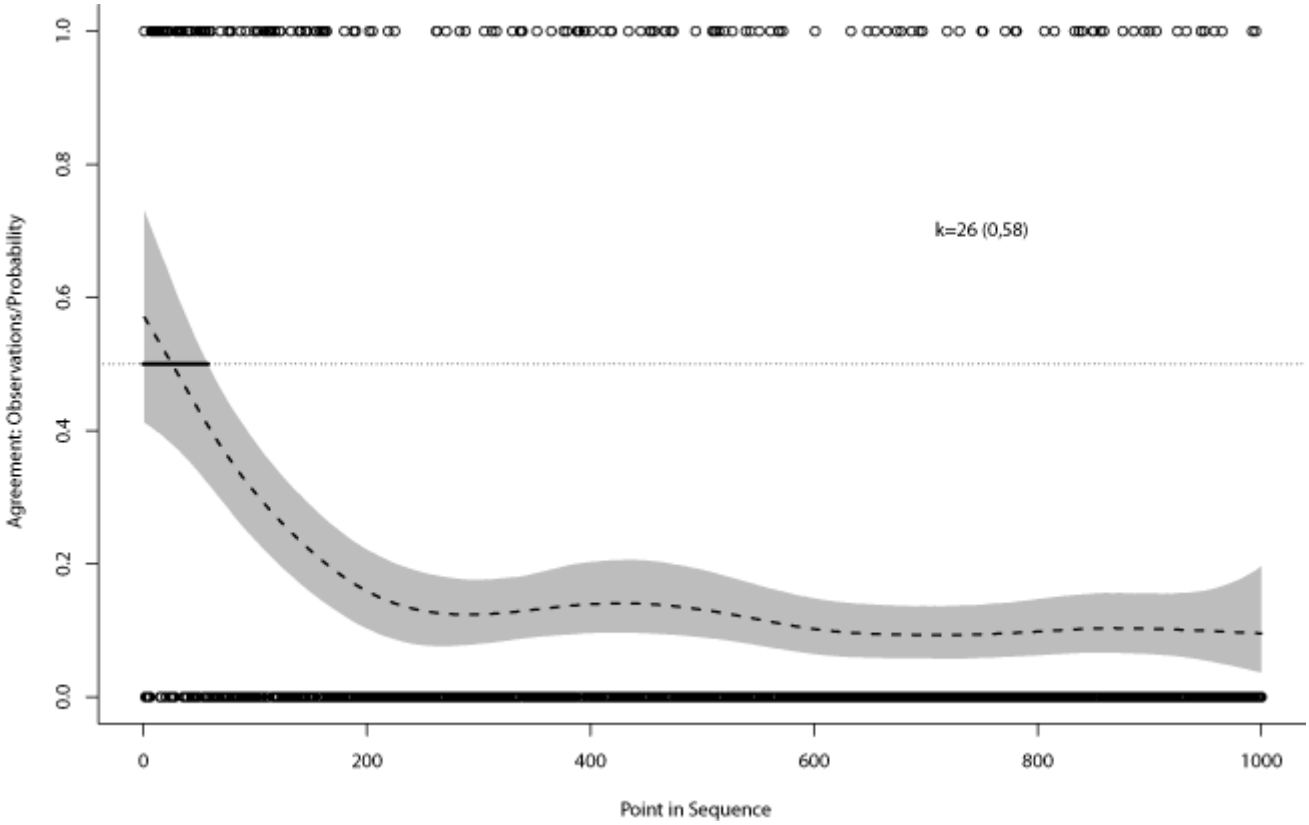

Figure 3: Observed agreement (1s and 0s) and Probability of agreement between the Mann-Whitney and negative binomial rankings on a reduced list of 1000 genes with agreement distance 40 (26 df, see text).

Acknowledgement

This research was partly supported by NHMRC project grant 525453. We thank Dr Ashwin Unni Krishnan for use of the data to illustrate our method.

References

1. Datta S, Kim S, Chakraborty S, Gill RS (2010) Statistical Analyses of Next Generation Sequence Data: A Partial Overview. Journal of Proteomics & Bioinformatics 3:183-190.
2. Crainiceanu CM, Ruppert D, Wand MP (2005) Bayesian analysis for penalized spline regression using WinBUGS. Journal of Statistical Software 14: 1-24.
3. Hall P, Miller H (2010) Modeling the variability of rankings. The Annals of Statistics 38: 2652-2677.
4. Hall P, Schimek MG (2012) Moderate-deviation-based inference for random degeneration in paired rank lists. Journal of the American Statistical Association 107: 661-672.
5. Markey JK, Wand MP (2010) Non-standard semiparametric regression via BRugs. Journal of Statistical Software 37: 1-30.
6. Schimek MG, Budinska MGS, Kugler KG, Svendova V, Ding J, et al. (2015) TopKLists: a comprehensive R package for statistical inference, stochastic aggregation and visualisation of multiple omics ranked lists. Statistical Applications in Genetics and Molecular Biology 14: 311-316.
7. Shannon P, Markiel A, Ozier NS, Wang JT, Ramage D, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13: 2498-2504.
8. Wand MP, Ormorod JT (2008) On semiparametric regression with O'Sullivan penalized splines. Australian and New Zealand Journal of Statistics 50: 179-198.
9. Wand MP (2009) Semiparametric and graphical models. Australian and New Zealand Journal of Statistics 51: 9-41.

This article was originally published in a special issue, entitled: **"Sequencing Technologies"**, Edited by Jianping Wang