Initial method for *linking* two adjacent regions using high-throughput sequencing data

Advanced Methods VI 140.756

Abstract

We propose a statistical method to determine whether two adjacent candidate differentially expressed regions should be *linked* or not. Using two chromosomes from a public data set with three groups and 25 samples, we illustrate the feasibility of the method with positive results. We further identify complex cases and propose improvements to the method.

Introduction

In the field of genomics, nowadays it is common to obtain data querying specific biological mechanisms using high-throughput sequencing. When using this technology, the data is a set of measurements along the genome axis (sometimes summarized by features such as exons). Typically, such experiments involve two or more groups where the common question is: are there differences between the groups? If so, where? Depending on the type of experiment (ChIP-seq, RNA-seq, etc), there are different solutions to find candidate regions.

A common problem is that the candidate regions can be highly fragmented. Multiple reasons provoke this behavior, such as biological reasons. For example, when the regions correspond to two exons and thus the intron creates the separation. Another source of fragmentation is the fact that measurements are far from ideal, presenting fluctuation on the data that can be hard to distinguish from the true signal.

In this work, we attempt to answer the question: Are two adjacent regions *similar*? This is a broad question and has to be refined. We start by answering the simple question: are the mean measurements similar between the two adjacent regions when adjusting for group status? A more advanced question is: is the difference between groups similar between the two adjacent regions? We were further interested in checking whether regions overlapping the same exon were more frequently similar.

In this particular work, we used `derfinder` to find candidate differentially expressed regions (DERs) from a public data set studying differential expression on brain 25 samples between control, cocaine addicts and alcohol addicts using RNA-seq. We focused on only two chromosomes as this work is a proof of concept. Figure 1 (top panel) shows a fragment of chromosome 6 where different candidate DERs are shown (green boxes). In it, we can observe region-pairs that span different exons and others contained in the same exon (black boxes); we think that the latter should be very similar.

Methods

To translate from the genomics field to terms seen in class, the outcome measured is the coverage at a given base pair of the genome. The coverage is defined as the number of sequenced reads that align at that given base pair. It is typically between 0 and several thousands, and depends on how much data you obtained for a given sample. As common in the field, we $\log_2$ transformed the data after adding a scaling factor (32). Each sample would translate to the subject, while the individual base pairs translates to the repeated visits of the subject. Note that by construction (sequenced reads are 100 base pairs in length), the data is correlated because the coverage at base i will be very similar to the one in base $i+1$. We thus considered fitting a GEE, and for comparisons purposes used independence, exchangeable and AR-1 working correlations.

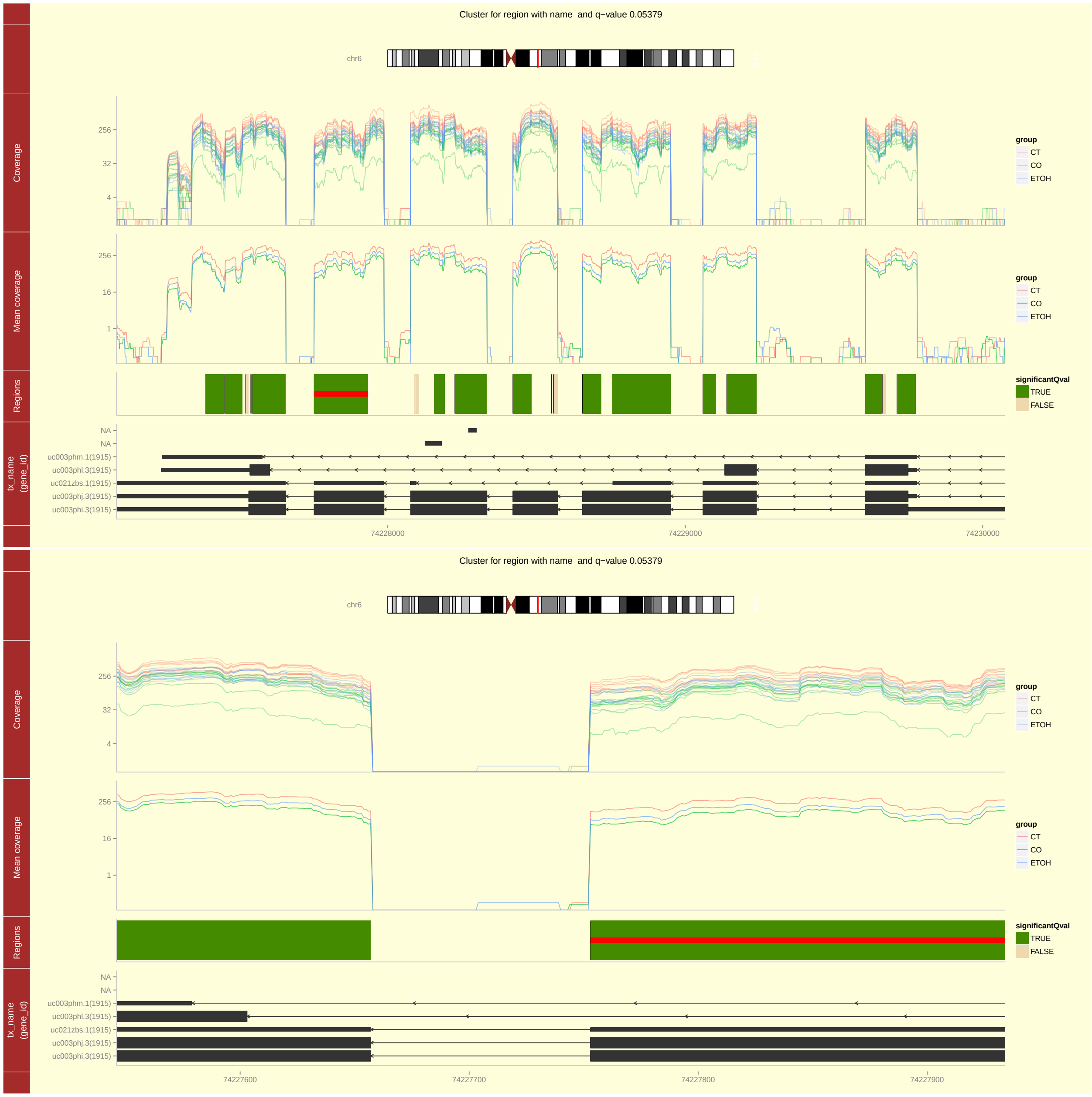

Figure 1: (Top panel) Example of a region on chromosome 6 where multiple candidate differentially expressed regions were identified (regions are less than 3000 bases apart). (Bottom panel) Zoomed in plot for the first region pair spanning two different exons.

For each pair of adjacent regions (within 3000 base pairs of each other), we built a *region-pair*. These consisted of: **region1** first region, **regionM** middle part[1], and **region2** second region.

Due to the high dimensions of some of these region-pairs[2], we only considered the region-pairs where both region 1 and 2 had a width greater than 1, but regionM had a width lower than 250. Thus, we used 300 out of 890 region-pairs ($\sim 33.7\%$) for chromosome 6 and 187 out of 573 region-pairs ($\sim 32.6\%$) for chromosome 22, for a total of 487 region-pairs.

Then for each region-pair (you can think of it as a single data set) we used the following model

$$\log_2(Y_{ijk} + 32) = \alpha + \beta_1 \text{sampleDepth}_i + \boldsymbol{\beta}_2 \text{group}_j + \boldsymbol{\beta}_3 \text{region}_k + \epsilon$$

Using region1 and the control (CT) group as the reference (for identifiability), we want to test $\boldsymbol{\beta}_3\,(\text{region}_2) = 0$. Doing so, we answer the basic question of whether region 1 and 2 are similar overall while adjusting for group.

Using the following interaction model ($i = 1, 2, \ldots, 25$ for sample)

$$\log_2(Y_{ijk} + 32) = \alpha + \beta_1 \text{sampleDepth}_i + \boldsymbol{\beta}_2 \text{group}_j + \boldsymbol{\beta}_3 \text{region}_k + \boldsymbol{\beta}_4 \text{group} \times \text{region}_{jk} + \epsilon$$

we can test whether the region2 coefficient and the region2 vs group (CO and ETOH) coefficients interactions are significantly different from 0. Doing so answers the advanced question of whether the differential expression pattern stays the same between region 1 and 2.

Note that in both models we adjust for the how much data we have per sample (called sample depth in genomics). This is required to make sure that the differences are due to the group and/or region, but not due to the amount of available data.

Results

Figure 2 shows the boxplots for the log Root Mean Squared Prediction Error (RMSPE) for the six models with the 300 region-pairs of chromosome 6. Overall, the distribution of the RMSPE is very similar for all models, with the AR-1 models having slightly higher RMSPE. Figure 3 shows a more detailed comparison of the RMSPE between the AR-1 and Exchangeable models, both with and without the interaction terms. From figure 3 (top panels), we note that the AR-1 models have slightly worse RMSPE than the Exchangeable models. Furthermore, from figure 3 (bottom panels), we can conclude that the models perform slightly better when fitting the interaction terms, although the difference is very small. Thus, judging solely in terms of RMSPE, for each question the GEE with Exchangeable working correlation is the better model.

We further investigated how the models performed in terms of the basic question when considering region-pairs where region1 and region2 overlap the same exon (*same-exon*), different exons (*diff-exon*) or at least one of the two is not overlapping an exon (*no-exons*). Because we are considering a total of 487 region-pairs, we adjusted the p-values to control the false discovery rate using the `qvalue` package available in Bioconductor. Then, we accepted (failed to reject) the null hypothesis that the region2 coefficient is 0 when the q-value was greater or equal to 0.10, and rejected otherwise. Table 1 shows the results for both chromosomes and all three types by exon status. From the table, we can observe that in the majority of the cases, the conclusion is the same between GEE with AR-1 and Exchangeable working correlations. Furthermore, for *no-exons* and

[1] Could be an intron, or a section of the same exon.

[2] Note that the total number of observations is 25 times the width of the region pair, which easily gets beyond 10 thousand and above.

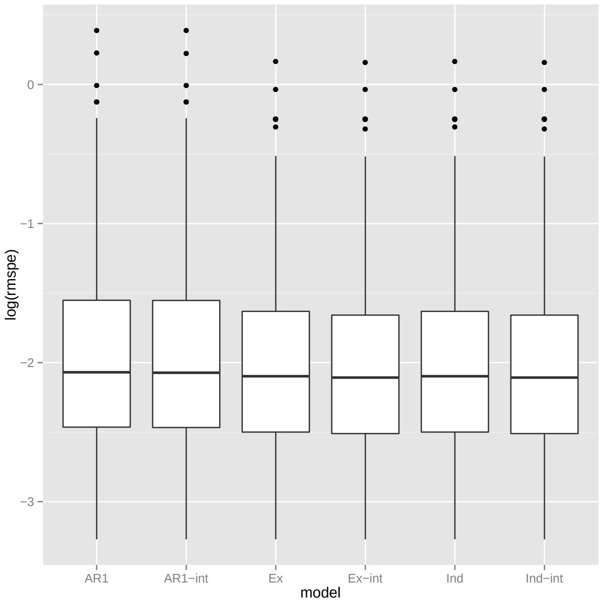

Figure 2: Boxplots of the log of the Root Mean Squared Prediction Error (RMSPE) for all six methods; only for chromsome 6.

diff-exon, any of the two models rarely accepts the null hypothesis for any region-pair, while they do so more frequently for the *same-exon* case.

BothAccept	Ar1Accept	ExAccept	BothReject	chr	ExonStatus
0	1	2	7	chr6	no-exons
0	1	0	13	chr6	diff-exon
19	24	23	210	chr6	same-exon
0	4	1	11	chr22	no-exons
0	0	0	2	chr22	diff-exon
5	15	14	135	chr22	same-exon

Table 1: Agreement between GEE with AR-1 and Exchangeable working correlations (without interaction) on accepting/rejecting the null hypothesis that the region 2 coefficient is significantly different from 0 judged by a q-value less than 0.10. Separated by exon status.

For the more advanced question, per region-pair we perform three tests (region2, region2 interaction with CO group, region2 interaction with ETOH group). Table 2 shows the results only for the *same-exons* case. From the table, we can see that chromosome 6 has every possible case at least once. In both chromosomes, the 1-1 case was the most frequent, meaning that they reject the null hypothesis for only 1 of the 3 tests. From the table, we can see that answering the advanced question of whether the differential expression by group is the same in region1 and region2 is much more complicated, even if we consider only the $0 - *$ and $* - 0$ cases versus the $3 - *$ and $* - 3$ cases.

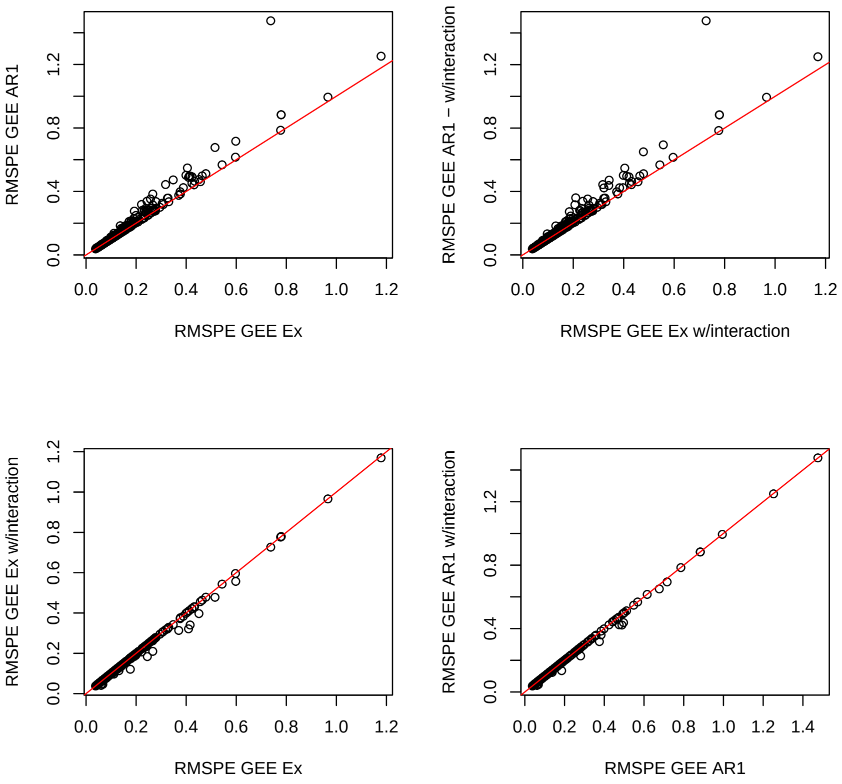

Figure 3: (Top left) GEE AR1 vs GEE Exchangeable (Ex) RMSPE, GEE AR1 performs worse. (Top right) GEE Ex with interaction vs GEE AR1 with interaction, GEE AR1 with interaction performs worse. (Bottom left) GEE Ex with interaction vs GEE Ex without interaction, the latter performs slightly worse. (Bottom right) GEE AR1 with interaction vs GEE AR1 without interaction, the latter performs slightly worse. In all panels, the red line is the diagonal line.

0-0	0-1	0-2	0-3	1-0	1-1	1-2	1-3	2-0	2-1	2-2	2-3	3-0	3-1	3-2	3-3	chr
23	26	5	3	22	110	33	11	3	20	6	7	1	1	2	3	chr6
26	24			22	73	15	2		4	2					1	chr22

Table 2: Agreement between GEE with AR-1 and Exchangeable working correlations on the region2, region2:groupCO and region2:groupETOH coefficients, shown only for the region pairs that share an exon. Columns are coded by the number of null hypothesis rejections, first for GEE AR1 and then for GEE Ex.

Conclusions

From this work, we noticed that using GEE with the longer region-pairs takes a significant amount of computing time, making the method less feasible in such cases. In these cases, it might be possible to drop regionM or take a subset of it to improve the feasibility of the method.

Furthermore, GEE with Exchangeable working correlation outperformed GEE with AR-1 working correlation in terms of RMSPE, although the difference was small for most region-pairs. When using the models without the interaction term, region-pairs sharing an exon were linked in 11-20% of the cases. The method we have proposed works as evaluated by comparing to the exon status of the region-pairs, but is restrictive in the sense that a lower proportion of the region-pairs sharing an exon were linked than what was expected. This could be due to the complexity of the data as illustrated by the models with interaction terms and the complexity of cases observed as shown in table 2.

References

GEE models were fitted using the `geepack` package. RMSPE was calculated using the `cvTools` package.

- **Project** code, results, presentation and report are available at `https://github.com/lcolladotor/756final_code`.
- A. Frazee, S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek (2013). Differential expression analysis of rna-seq data at single base resolution, Biostatistics, *recently accepted*.
- L. Collado-Torres, A. Frazee, M. Love, R. A. Irizarry, A. E. Jaffe, J. T. Leek (2013). derfinder: Software for annotation-agnostic RNA-seq differential expression analysis. Manuscript in preparation.
- `derfinder` package `https://github.com/lcolladotor/derfinder`
- Højsgaard, S., Halekoh, U. Yan J. (2006) The R Package geepack for Generalized Estimating Equations Journal of Statistical Software, 15, 2, pp1–11
- Andreas Alfons (2012). cvTools: Cross-validation tools for regression models. R package version 0.3.2.
- Alan Dabney, John D. Storey and with assistance from Gregory R. Warnes. qvalue: Q-value estimation for false discovery rate control. R package version 1.36.0.