BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 22 2012, pages 2986–2988
doi:10.1093/bioinformatics/bts545

Sequence analysis

Advance Access publication September 5, 2012

Comb-p: software for combining, analyzing, grouping and correcting spatially correlated *P*-values

Brent S. Pedersen[1,*], David A. Schwartz[1], Ivana V. Yang[1,†] and Katerina J. Kechris[2,†]

[1]Department of Medicine and [2]Department of Biostatistics and Informatics, University of Colorado, Denver, Anschutz Medical Campus, Aurora, CO 80045, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: *comb-p* is a command-line tool and a python library that manipulates BED files of possibly irregularly spaced *P*-values and (1) calculates auto-correlation, (2) combines adjacent *P*-values, (3) performs false discovery adjustment, (4) finds regions of enrichment (i.e. series of adjacent low *P*-values) and (5) assigns significance to those regions. In addition, tools are provided for visualization and assessment. We provide validation and example uses on bisulfite-seq with *P*-values from Fisher's exact test, tiled methylation probes using a linear model and Dam-ID for chromatin binding using moderated *t*-statistics. Because the library accepts input in a simple, standardized format and is unaffected by the origin of the *P*-values, it can be used for a wide variety of applications.

Availability: *comb-p* is maintained under the BSD license. The documentation and implementation are available at https://github.com/brentp/combined-pvalues.

Contact: bpederse@gmail.com

Received on March 20, 2012; revised on August 28, 2012; accepted on August 30, 2012

1 INTRODUCTION

A variety of high-throughput technologies generate genome-wide data used to study processes such as DNA-binding, methylation status and histone modifications. These technologies, including tiling arrays and sequence-based assays, generate data that are often auto-correlated across the genome, making inferences difficult. The significance of individual regions may be dampened after multiple-testing correction on potentially millions of sites. In such studies, hypotheses tests can be performed at each location to generate *P*-values for evaluating the effects of interest. For this purpose, Kechris *et al.* (2010) developed a method for combining *P*-values in sliding windows and accounting for spatial correlations across the genome. Here, we build on this approach with software that allows for uneven data structure across the genome, more general auto-correlation calculations and multiple-testing corrections for peaks (i.e. genomic regions of enrichment) with applications to a variety of different technologies.

*To whom correspondence should be addressed.
†The authors wish it to be known that in their opinion, the last two authors should be regarded as joint Last Authors.

2 APPROACH

Tiling array studies relying on two-sample comparisons may be amenable to the calculation of sliding window averages of log ratios or two-sample test statistics. However, more complex study designs often require covariates and report *P*-values from linear models or other statistical tests.

We utilize a 'moving averages' method of *P*-value correction that does not depend on the test used to generate the *P*-values. Fisher (1948) developed an approach of combining *P*-values from independent tests to get a single meta-analysis test statistic with a χ^2 distribution and degrees of freedom based on the number of tests being combined. A similar method developed by Stouffer *et al.* (1949) and Liptak (1958) first converts *P*-values to *Z*-scores which are then summed and scaled to create a single, combined *Z*-score. The Stouffer–Liptak method lends itself to the addition of weights on each *P*-value. Zaykin *et al.* (2002) introduced a method to use weights to perform a dependence correction on correlated tests. Kechris *et al.* (2010) used a sliding window correction where each *P*-value is adjusted by applying the Stouffer–Liptak method to neighboring *P*-values as weighted according to the observed auto-correlation at the appropriate lag.

Here, we provide a generic, efficient and customizable implementation of these methods with additions including handling variably spaced probes, a peak finder for dynamically sized regions, and a means to calculate a *P*-value for each peak, adjusted for multiple comparisons. We refer to this implementation as *comb-p* and illustrate its application using three different technologies.

3 IMPLEMENTATION

All programs within *comb-p* expect files in simple BED format (Kent *et al.*, 2002) sorted by chromosome and start. Additional columns contain the *P*-value(s) of interest based on the study design and generated from any software or statistical test.

comb-p first calculates the correlation at varying distance lags (here: auto-correlation—ACF). While Kechris' (Kechris *et al.*, 2010) and many ACF implementations rely on fixed offsets of adjacent probes, *comb-p* accepts a maximum distance and a single offset that segments that distance into intervals. If a given probe is 40 bases away from two (or more) other probes, then it will appear more than once in the bin containing probe pairs separated by <40 bases. This is useful in cases where the

probes generating the *P*-values are unevenly spaced, as is common with methylation arrays.

Once the ACF has been calculated, it can be used to perform the Stouffer–Liptak–Kechris correction (*slk*) where each *P*-value is adjusted according to adjacent *P*-values as weighted according to the ACF. The resulting BED file has an additional column containing the corrected *P*-value. A given *P*-value will be pulled lower if its neighbors also have low *P*-values (and little auto-correlation) and likely remain insignificant if the neighboring *P*-values are also high.

A *q*-value score based on the Benjamini–Hochberg false discovery (FDR) correction or on a null model from shuffled data may then be calculated. The peak-finding algorithm can then be used to find enrichment regions or *peaks* on the FDR *q*-value, the *slk*-corrected *P*-value or on the original *P*-value.

Once the regions are identified, a *P*-value for each region can be assigned using the Stouffer–Liptak correction. Note that the first ACF calculated above was only out to the distance specified and may not extend far enough to cover the longer regions. Therefore, this step first calculates the ACF out to a distance equal to the length of the longest region. The corrected *P*-value for each region is then calculated using the original *P*-values that fall within the peak and the full ACF. Because we use the original, uncorrected *P*-values in the calculation of significance for the peak, we side-step issues of altering the distribution in both the *slk* and FDR steps. The *region_p* program reports the *slk*-corrected *P*-value and a one-step Šidák (1967) multiple-testing correction. For a given region, the number of possible tests in the Šidák correction is the total bases covered by all input probes divided by the size of the given region. In short, we define the extents of the region using the FDR *q*-values, then we define the significance of the regions using the SLK correction of the original *P*-values.

comb-p is implemented as a single command-line application that dispatches to multiple independent sub-modules and as a set of python packages that may be used programatically. Where possible, computationally intensive algorithms are parallelized automatically—for the ACF and the *slk* steps in the analyses described below—this results in a speed-up that is nearly linear with the number of cores available. The implementation has been tuned to minimize memory use and computation time.

When run without arguments, the *comb-p* executable displays the available programs and a short description of each. Although each of these tools may be used independently, the progression of *acf*, *slk*, *fdr*, *peaks*, *region_p* works well for the examples presented. These steps can be run successively with the single command: *pipeline*.

4 APPLICATION

We extend three previously published analyses to demonstrate the utility and breadth of use of *comb-p* and to validate against the published data. The full set of commands used to run each of these analyses is available in the corresponding subdirectory at: https://github.com/brentp/combined-pvalues/tree/master/examples.

Comprehensive, high-throughput array for relative methylation is a tiling array used to quantify methylation at CpG-rich sites (Irizarry *et al*., 2008). Adjacent probes are correlated due to the regional nature of methylation and to the overlapping probes on the tiling array. The CHARM study reported in Irizarry *et al*. (2009) contains probe data for tumor and normal samples from a variety of tissues. Using these data, we fit a linear model in R (http://www.R-project.org) to obtain *P*-values for the significance of tumor or tissue-specific effects on methylation status at each probe. We then run *comb-p*, which finds 85% of the tumor-specific differentially methylated regions (C-DMRs) reported by Irizarry *et al*. (2009) and 75% of the tissue-specific T-DMRs reported. While these overlaps depend on the parameter choice, we show that the Irizarry DMRs that overlap with the *comb-p* DMRs have significantly (*t*-test, $P = 1.296e{-}145$) lower FDR values (as reported by Irizarry) than those that do not, indicating that *comb-p* is reporting the best regions. We also show that our method yields only 39 false-positive DMRs on 20 runs of shuffled of CHARM data (33 of those come from two of the sets, because those shuffles of the clinical data happen to coincide well with the case–control status of the original data). This gives a FDR rate of 0.0008 for reasonable cutoffs. We found that *comb-p* recovers more significant probes than a simple FDR correction on the original *P*-values. We also found that the process of segmenting the genome into smaller chunks and calling regions independently in each chunk generates similar regions to those derived using the auto-correlation structure for the entire genome (the default).

DamID technology followed by tiling arrays can be used to map genomic regions bound by DNA-binding proteins (van Steensel *et al*., 2001). We recreate an analysis of the DamID tiling array data for the Ci transcription factor in *Drosophila melanogaster* (Biehs *et al*., 2010). Kechris *et al*. (2010) calculated the mean log ratios of intensities between Ci experimental and control samples to obtain *P*-values from a one-sided moderated *t*-test. Of the original 878 regions, 695 (79%) are represented in the *comb-p* predicted peaks. To score each peak, the authors report $-\log_{10}$ of the smallest raw *P*-value for each peak. The published peaks overlapping the *comb-p* regions have a higher score than those that do not (*t*-test, $P = 2.98e{-}59$), indicating that *comb-p* is finding the best regions among those previously reported. In addition, *comb-p* regions also yield a higher target enrichment score based on proximity to known Ci genes, indicating that *comb-p* is selecting for known targets. The corrected *P*-value reported by *comb-p* can be used as a filter to extract regions of interest; we calculated the enrichment ratio of the number of observed to expected Ci target genes at various *comb-p*-corrected *P*-value cutoffs. For a cutoff of 0.1, the enrichment is 2.41, this enrichment increases to 3.46 and 5.29 for more stringent cutoffs of 1e−3 and 1e−4, respectively.

Bisulfite-sequencing (BS-Seq) is also used to measure methylation across the genome. As another example of the flexibility of our method, we demonstrate a possible analysis on data described in Hsieh *et al*. (2009) from *Arabidopsis thaliana* using MethylCoder (Pedersen *et al*., 2011) to map the bisulfite-treated reads to the genome. At each site, we use Fisher's exact test to obtain *P*-values for the counts of converted and un-converted cytosines between endosperm and embryo. We find DMRs between these two tissues associated with genes enriched for gene ontologies related to the ribosome ($P = 1e{-}3$).

5 CONCLUSION

The *comb-p* software is useful in contexts where auto-correlated *P*-values are generated across the genome. We have outlined our implementation and demonstrated the utility on data from three different technologies each from a different statistical test. We have also validated our method using previously published results from those studies. Future work could extend this framework to address applications with a more segmented structure such as DNA copy number and ChIP-seq. Currently, we recommend the use of *comb-p* for technologies with a more regular autocorrelation structure. In addition, we note that this method will have a sensitivity and specificity that will depend on each dataset and we recommend that users explore these tradeoffs on shuffled versions of their own data as demonstrated in the CHARM example.

Funding: The National Institutes of Health (AA016922 to K.K., HL101251 to D.A.S. and I.V.Y. and AI90052 and HL101715 to D.A.S.).

Conflict of Interest: none declared.

REFERENCES

Biehs,B. *et al.* (2010) Hedgehog targets in the *Drosophila* embryo and the mechanisms that generate tissue-specific outputs of Hedgehog signaling. *Development*, **137**, 3887–3898.

Fisher,R.A. (1948) Questions and answers #14. *Am. Stat.*, **2**, 30–31.

Hsieh,T.F. *et al.* (2009) Genome-wide demethylation of Arabidopsis endosperm. *Science*, **324**, 1451–1454.

Irizarry,R.A. *et al.* (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.*, **18**, 780–790.

Irizarry,R.A. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.

Kechris,K.J. *et al.* (2010) Generalizing moving averages for tiling arrays using combined *p*-value statistic. *Stat. Appl. Genet. Mol. Biol.*, **9**, Article 29.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1106.

Liptak,T. (1958) On the combination of independent tests. *Magyar Tudomanyos. Akademia Matematikai Kutato Intezetenek Kozlemenyei*, **3**, 171–197.

Pedersen,B.S. *et al.* (2011) MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, **27**, 2435–2436.

Šidák,Z. (1967) Rectangular confidence region for the means of multivariate normal distributions. *J. Am. Stat. Assoc.*, **62**, 626–633.

Stouffer,S.A. *et al.* (1949) *The American Soldier. Vol. 1: Adjustment during Army Life*. Princeton University Press, Princeton, NJ.

van Steensel,B. *et al.* (2001) Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.*, **27**, 304–308.

Zaykin,D.L. *et al.* (2002) Truncated product method for combining *p*-values. *Genet. Epidemiol.*, **22**, 170–185.