Bioinformatics, 38(18), 2022, 4415–4417
https://doi.org/10.1093/bioinformatics/btac542
Advance Access Publication Date: 4 August 2022
Applications Note

Genome analysis

DiNAMIC.Duo: detecting somatic DNA copy number differences without a normal reference

Vonn Walter [1,*], Hyo Young Choi[2,3], Xiaobei Zhao[4], Yan Gao [5], Jeremiah Holt[3] and D. Neil Hayes[4,5,6]

[1]Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA, [2]Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA, [3]Department of Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA, [4]Division of Hematology and Oncology, Department of Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA, [5]Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA and [6]Center for Cancer Research, University of Tennessee Health Science Center, Memphis, TN 38163, USA

*To whom correspondence should be addressed.

Associate Editor: Christina Kendziorski

Received on January 26, 2022; revised on July 18, 2022; editorial decision on July 26, 2022; accepted on August 2, 2022

Abstract

Motivation: Somatic DNA copy number alterations (CNAs) arise in tumor tissue because of underlying genomic instability. Recurrent CNAs that occur in the same genomic region across multiple independent samples are of interest to researchers because they may contain genes that contribute to the cancer phenotype. However, differences in copy number states between cancers are also commonly of interest, for example when comparing tumors with distinct morphologies in the same anatomic location. Current methodologies are limited by their inability to perform direct comparisons of CNAs between tumor cohorts, and thus they cannot formally assess the statistical significance of observed copy number differences or identify regions of the genome where these differences occur.

Results: We introduce the DiNAMIC.Duo R package that can be used to identify recurrent CNAs in a single cohort or recurrent copy number differences between two cohorts, including when neither cohort is copy neutral. The package utilizes Python scripts for computational efficiency and provides functionality for producing figures and summary output files.

Availability and implementation: The DiNAMIC.Duo R package is available from CRAN at https://cran.r-project.org/web/packages/DiNAMIC.Duo/index.html. This article uses publicly available data from the Broad Institute TCGA Genome Data Analysis Center, https://doi.org/10.7908/C11G0KM9.

Contact: vwalter1@pennstatehealth.psu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic instability is a hallmark of cancer, and it can lead to various structural variants, including DNA copy number alterations (CNAs). Some CNAs, termed recurrent, are observed in the same genomic region across multiple independent samples. It is believed that recurrent CNAs arise because they provide a selective advantage for either cancer initiation, cancer promotion or therapeutic resistance. For example, amplifications can lead to elevated expression of oncogenes, thus driving increased cell proliferation or invasiveness; similarly, losses of tumor suppressor genes may result in compromised cell cycle regulation or DNA damage repair. In contrast, sporadic CNAs that are randomly scattered across the genome are less likely to be biologically relevant. Several bioinformatics tools have been developed to identify recurrent CNAs in a single tumor cohort, including GISTIC (Beroukhim *et al.*, 2007), RUBIC (van Dyk *et al.*, 2016) and DiNAMIC (Walter *et al.*, 2011), all of which operate under the null hypothesis that there are no recurrent CNAs. These tools have been used in multi-omics tumor profiling analyses, as evidenced by the application of GISTIC in studies conducted by The Cancer Genome Atlas (TCGA) Research Network.

Remarkably, there appear to be no existing tools to identify recurrent copy number differences between two cohorts, such as when the comparator groups are two sets of tumors. In contrast, the corresponding analysis for gene expression data—namely, identifying differentially expressed genes in two conditions—is a fundamental part of many gene expression profiling studies. Instead, copy number differences are typically inferred by analyzing each cohort separately and comparing the results. For example, in the TCGA study of head and neck squamous cell carcinoma (The Cancer Genome Atlas

Network, 2015), separate GISTIC analyses of patients with and without human papillomavirus (HPV) infection showed that gains of the oncogene *EGFR* were common in HPV-negative patients but largely absent in HPV-positive subjects. While this information is useful, the analysis approach has limitations because the groups are not compared directly. Thus it is not possible to assess the statistical significance of the observed copy number differences, nor can one accurately identify regions of the genome where these differences occur. This motivated our interest in developing DiNAMIC.Duo, which builds upon and extends DiNAMIC by leveraging the theoretical studies of cyclic shift testing (Walter *et al.*, 2015). DiNAMIC.Duo provides additional functionality not present in DiNAMIC, and it achieves gains in computational efficiency by utilizing Python scripts. The DiNAMIC.Duo R package is available from CRAN.

2 DiNAMIC.Duo workflow and outputs

A DiNAMIC.Duo analysis of copy number differences starts with matrices X and Y that contain quantitative gene-level copy number measurements. Gains or losses in single cohort can be analyzed by setting $Y = \text{NULL}$. Entries of X and Y are assumed to be on the log-ratio scale and normalized so that zero corresponds to copy neutral; negative and positive values correspond to losses and gains, respectively. Rows of X and Y are indexed by a common set of genes that appear in genomic order in the autosomes; columns of X and Y correspond to independent samples. DiNAMIC.Duo includes functionality to query the biomaRt R package and reformat X and Y, if necessary.

We write $\overline{X}_i$ and $\overline{Y}_i$ for the mean DNA copy number value of gene i in X and Y, respectively. Briefly, for all genes i, observed differences in gene-level means $\overline{X}_i - \overline{Y}_i$ are calculated. Positive and negative copy number differences are analyzed separately, and here we restrict attention to positive differences; negative differences are handled similarly. Suppose the largest copy number difference, $\max_i(\overline{X}_i - \overline{Y}_i)$, is observed at gene $k = \text{argmax}(\overline{X}_i - \overline{Y}_i)$. If $\pi_{Xj}(X)$ and $\pi_{Yj}(Y)$ represent cyclic shifts of X and Y, respectively, for $j = 1,\ldots,n$, the empirical null distribution $\{\max_i(\overline{\pi_{Xj}(X)}_i - \overline{\pi_{Yj}(Y)}_i)\}_{j=1}^{n}$is used to assess the statistical significance of $\max_i(\overline{X}_i - \overline{Y}_i)$. DiNAMIC.Duo's peeling algorithm is then applied to modify entries of X and Y so that copy number differences $\overline{X}_i - \overline{Y}_i$ in a genomic region around gene k are neutral, while copy number differences in the remainder of the genome remain unchanged. Thus multiple peaks corresponding to positive copy number differences across the genome can be identified by iteratively applying the peeling algorithm. Our approach makes it possible to use the same null distribution to assess the significance of multiple positive peaks while controlling the familywise error rate. Additionally, because $\max_i(\overline{X}_i - \overline{Y}_i)$ is invariant under constant cyclic shifts, our theoretical study of cyclic shift testing (Walter *et al.*, 2015) implies that DiNAMIC.Duo's *P*-values are asymptotically consistent. Details of our approach for assessing statistical significance, the peeling algorithm and examples of summary output files and plots are provided in the Supplementary Material.

3 Application to lung cancer

Lung cancers are the leading cause of cancer-related deaths, and lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the most common types of non-small cell lung cancer. Two recent studies used gene expression data from the TCGA LUAD and LUSC cohorts (The Cancer Genome Atlas Research Network, 2014, 2012) to identify predictive biomarkers (Chen and Dhahbi, 2021) and differences in key signaling pathways (Anusewicz *et al.*, 2020). Numerous differentially expressed genes were observed, including *TP63*, *PIK3CA* and *SOX2* (all chr3q), *NKX2-1* (chr14q13) and *E2F1* (chr20q11). Figure 1 shows genome-wide mean gene-level copy number values from the TCGA LUAD (solid line) and LUSC (dashed line) cohorts, as well as differences LUAD − LUSC (dotted line). The pronounced positive and negative peaks for LUAD − LUSC suggest the presence of numerous genomic regions containing statistically significant differences. The results in Supplementary Table S1 confirm these observations, and indeed all of the genes noted above lie in regions identified by DiNAMIC.Duo. Thus underlying copy number differences may drive the observed differences in expression. Copy number gains of the *MYC* oncogene (chr8q24) leading to increased expression have been observed in both LUAD and LUSC. Interestingly, although Anusewicz *et al.* (2020) found that *MYC* was differentially expressed, we observed no statistically significant copy number differences in this region.

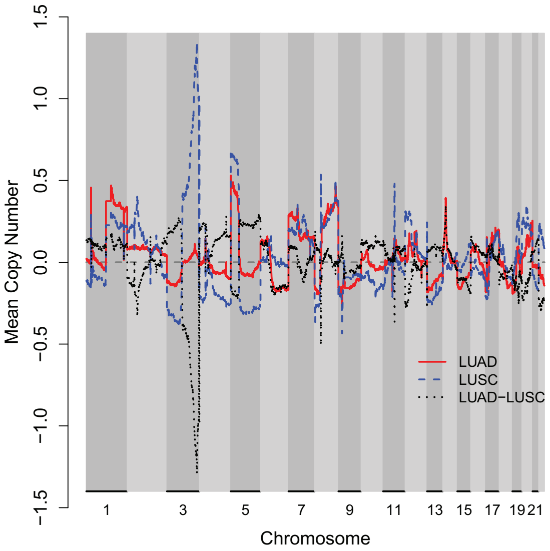

Fig. 1. DNA copy number differences in lung cancer. Genome-wide plots of mean DNA copy number values from the TCGA studies of lung adenocarcinoma (LUAD, solid line) and lung squamous cell carcinoma (LUSC, dashed line). Mean copy number differences LUAD − LUSC are shown in the dotted line, many of which are highly significant.

Both LUAD and LUSC exhibit numerous recurrent copy number gains and losses, some of which appear in the same genomic region. By directly comparing two tumor cohorts, DiNAMIC.Duo provides insight that is not possible with single cohort methods. For example, statistically significant losses of *CDKN2A* are observed in both LUAD and LUSC. The differences for *CDKN2A* in LUAD − LUSC are statistically significant and negative, thus showing that losses are more pronounced in LUSC. In contrast, even though statistically significant gains of *EGFR* are observed in both tumor types, the differences of *EGFR* in LUAD − LUSC are not significant. Single cohort methods cannot make this distinction.

4 Simulation studies

We used two different approaches to simulate DNA copy number matrices X and Y, including the method from the original DiNAMIC manuscript. Our results demonstrate that DiNAMIC.Duo detects recurrent copy number differences and that power increases as the effect size of the difference grows. Variations in tumor purity and ploidy are known to complicate DNA copy number analyses. We found that differing levels of normal contamination in X and Y increase the likelihood of identifying statistically significant differences when gains of equal effect size are present in both matrices at the same locus. Details of the simulation studies can be found in the Supplementary Material.

5 Conclusion

We introduce DiNAMIC.Duo, a novel tool for detecting recurrent DNA copy number differences between two tumor cohorts. By analyzing publicly available TCGA lung cancer data, we identify underlying copy number differences that may drive differential gene expression identified in recent studies.

Funding

This work was supported by the National Institutes of Health [R01CA211939, U24CA264021, UG1CA233333, U24CA210988 and R01DE025712 to D.N.H.].

Conflict of Interest: none declared.

References

Anusewicz,D. *et al.* (2020) Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of Notch, Hedgehog, Wnt, and ErbB signaling. *Sci. Rep.*, **10**, 21128.

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA*, **104**, 20007–20012.

Chen,J.W. and Dhahbi,J. (2021) Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci. Rep.*, **11**, 13323.

The Cancer Genome Atlas Network. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576–582.

The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous lung cancers. *Nature*, **489**, 519–525.

The Cancer Genome Atlas Research Network. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.

van Dyk,E. *et al.* (2016) RUBIC identifies driver genes by detecting recurrent DNA copy number breaks. *Nat. Commun.*, **7**, 12159.

Walter,V. *et al.* (2011) DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors. *Bioinformatics*, **27**, 678–685.

Walter,V. *et al.* (2015) Consistent testing for recurrent genomic aberrations. *Biometrika*, **102**, 783–796.