Bioinformatics Advance Access published January 26, 2014

BIOINFORMATICS APPLICATIONS NOTE 2014, pages 1–2
doi:10.1093/bioinformatics/btu011

Genome analysis

Advance Access publication January 9, 2014

CGARS: cancer genome analysis by rank sums

Xin Lu[1], Roman K. Thomas[1,2] and Martin Peifer[1,3,*]

[1]Department of Translational Genomics, [2]Department of Pathology and [3]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany

Associate Editor: John Hancock

ABSTRACT

Motivation: Cancer genomes are characterized by the accumulation of point mutations and structural alterations such as copy-number alterations and genomic rearrangements. Among structural changes, systematic analyses of copy-number alterations have provided deeper insight into the architecture of cancer genomes and had led to new potential treatment opportunities. During the course of cancer genome evolution, selection mechanisms are leading to a non-random pattern of mutational events contributing to fitness benefits of the cancer cells. We therefore developed a new method to dissect random from non-random patterns in copy-number data and thereby to assess significantly enriched somatic copy-number aberrations across a set of tumor specimens or cell lines. In contrast to existing approaches, the method is invariant to any strictly monotonous transformation of the input data which results to an insensitivity of differences in tumor purity, array saturation effects and copy-number baseline levels.

Results: We applied our approach to recently published datasets of small-cell lung cancer and squamous cell lung cancer and validated its performance by comparing the results to an orthogonal approach. In addition, we found a new deletion peak containing the *HLA-A* gene in squamous cell lung cancer.

Availability: The CGARS program package is available for download at http://www.translational-genomics.uni-koeln.de/scientific-resources/ Documentation and examples are available together with the package.

Contact: mpeifer@uni-koeln.de

Supplementary Information: Supplementary information is available at *Bioinformatics* online.

Received on August 6, 2013; revised on November 29, 2013; accepted on January 4, 2014

1 INTRODUCTION

Recently, several algorithms to identify significant copy-number alterations from array-based high-throughput data (e.g. Affymetrix SNP arrays or array CGH) have been proposed (Beroukhim *et al.*, 2007; Mermel *et al.*, 2011; Sanchez-Garcia *et al.*, 2010; Taylor *et al.*, 2008). All of these methods have in common that the occurrence of different levels of tumor purity and ploidy are barely taken into account. Especially in the case of patient-derived tumor samples, the admixture of non-neoplastic cells is an uncontrollable experimental variable, often leading to large differences in tumor purity throughout the dataset. An additional feature of cancer genomes is that some tumors exhibit diverting levels of genome ploidy (triploid, tetraploid, etc.). Both, low tumor purity and higher genome ploidy lead to a decrease of inferred copy-number amplitudes (Carter *et al.*, 2012).

By using rank sums, we propose a novel copy-number-analysis method that automatically accounts for these different levels of purity and ploidy. We applied this approach to published datasets of small-cell lung cancer (SCLC) (Peifer *et al.*, 2012) and squamous cell carcinoma (SQ) (Weiss *et al.*, 2010) and finally validated detected regions of copy-number alteration by comparing the results with those derived from a GISTIC analysis (Mermel *et al.*, 2011).

2 METHOD AND IMPLEMENTATION

The general idea behind our method is to transform raw copy numbers into ranks. Probe sets in locations of common germ line copy-number variations are removed. Rank sums for each genomic location are then computed, smoothed and statistically evaluated. These key steps are schematically shown in Figure 1A. A detailed general description of the CGARS algorithm including all mathematical details is given in the Supplementary Material.

3 RESULTS AND CONCLUSION

To validate our method and to test its performance, we analyzed published datasets of 63 SCLC (Peifer *et al.*, 2012) and 146 SQ (Weiss *et al.*, 2010) tumor samples and compared the resulting data with an analysis using GISTIC (Mermel *et al.*, 2011). We observed that CGARS consumed substantially less computation time and memory than GISTIC on the same computational infrastructure [CGARS: 101s, 2.6GB (SCLC) 264s, 5.6GB (SQ); GISTIC: 289s, 9.1GB (SCLC) 397s, 9GB (SQ)].

Almost all identified high-confidence peaks of copy-number alteration are highly consistent between the two approaches (Fig. 1B and Supplementary Fig. S1 and Tables S1 and S2). Together with previously published copy-number analyses on SCLC and SQ (Hammerman *et al.*, 2012; Rudin *et al.*, 2012) this result supports the validity of our approach.

In case of the SQ dataset, we identified seven highly significant amplification peaks (containing genes: *CCND1*, *CCNE1*, *MYC*, *FGFR1*, *EGFR*, *SOX2* and *KRAS*) and three deletion peaks (containing genes: *LRP1B*, *CDKN2A* and *PTPRD*) consistently detected by GISTIC and CGARS (Fig. 1B). Most discordant regions identified by either method are broad lesions (Supplementary Table S2). Focal copy-number alterations only identified by GISTIC include *KIAA1841*, *MTMR3* and *PARD6G*; the link of these genes to SQ is currently unclear. Among the regions identified only by CGARS are peaks

*To whom correspondence should be addressed.

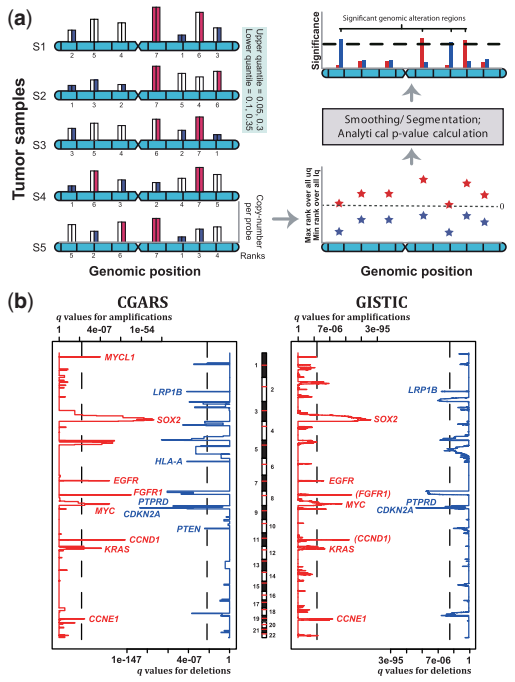

Fig. 1. (A) Overview of the key steps of the CGARS algorithm. First, raw copy numbers are transferred into rank sums. Upper and lower copy-number quantiles are then defined and ranks falling into the upper quantile (red bars) are separately analyzed from those that are situated in the lower quantile (blue bars). Each genomic location is represented by two bars indicating the multiple quantile selection. Ranks outside either quantile (white bars) are not considered for the subsequent analysis. Next, rank sums are computed for each genomic location and maximized (amplifications) or minimized (deletion) over the multiple quantiles (blue and red stars). These extremal rank sums are then subjected to a smoothing procedure. Statistics of smoothed rank sum profiles are finally computed to determine significant copy-number alterations, here shown by bars exceeding the dashed horizontal line (level of significance). (B) Results obtained from CGARS and GISTIC based on the SQ dataset. As main parameters we chose for CGARS: uq = 0.25, 0.05 (upper quantile), lq = 0.35, 0.25 (lower quantile), and for GISTIC: ta = tb = 0.4 (copy-number threshold). Vertical dashed lines indicate the significance level of 1% and putative target genes in the proximity of the copy-number peaks are shown in parentheses

containing *MYCL1* and *PTEN*; both regions appeared altered upon visual inspection of the data analyzed as well as in an independent dataset (Supplementary Fig. S2A and B).

Using CGARS we were further able to identify a previously unappreciated deletion peak region at 6p21.33 containing the *HLA-A* class I major histocompatibility gene (Supplementary Fig. S2C). In addition, *HLA-A* is frequently subjected to loss of function mutations in SQ (Hammerman *et al.*, 2012). Thus, identified deletions matched the mutation spectrum in the same tumor type and is therefore supporting the notion that *HLA-A* is a biologically relevant gene in SQ.

Finally, to demonstrate that CGARS is not over-calling copy-number aberrations, we analyzed our recently published dataset of 70 pulmonary carcinoids (Seidel *et al.*, 2013)—a tumor that is particularly silent in copy-number space. Besides a single broad deletion peak located at 11p11, none other significant copy-number aberration were detected, suggesting that CGARS is indeed not over-calling copy-number data.

In summary, we formulated a new copy-number-analysis method and tested its performance by comparing the results with an orthogonal approach. Our approach yields robust results, is computationally inexpensive, and is highly flexible. Furthermore, we identified a new deletion-peak region containing *HLA-A*; a gene that is also frequently mutated in SQ. Our methodology may thus offer a valuable addition to existing approaches since its flexibility enables the possibility of exploring datasets under various aspects.

Funding: EU-Framework Programme CURELUNG (HEALTH-F2-2010-258677 to R.K.T.); Deutsche Forschungsgemeinschaft through TH1386/3-1 and SFB832 (to R.K.T.); German Cancer Aid (3641116521 to R.K.T. and M.P.); German Ministry of Science and Education (BMBF) as part of the NGFNplus program (grant 01GS08100 to R.K.T.); Stand Up To Cancer–American Association for Cancer Research Innovative Research Grant (SU2C-AACR-IR60109 to R.K.T.).

Conflict of Interest: R.K.T. and M.P. are founder and shareholder of Blackfield AG, a company focused on cancer genome diagnostics and cancer genomics-based drug discovery. M.P. received consulting fees from Blackfield AG. R.K.T. received consulting and lecture fees (Sanofi-Aventis, Merck, Roche, Lilly, Boehringer Ingelheim, Astra- Zeneca, Atlas-Biolabs, Daiichi-Sankyo, MSD, Puma, Blackfield AG) as well as research support (Merck, EOS and AstraZeneca).

REFERENCES

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci USA*, **104**, 20007–200012.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Hammerman,P.S. *et al.* (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

Mermel,C.H. *et al.* (2011) Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.

Peifer,M. *et al.* (2012) Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.*, **44**, 1104–1110.

Rudin,C.M. *et al.* (2012) Comprehensive genomic analysis identifies sox2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.*, **44**, 1111–1116.

Sanchez-Garcia,F. *et al.* (2010) JISTIC: identification of significant targets in cancer. *BMC Bioinform.*, **11**, 189.

Seidel,D. *et al.* (2013) A genomics-based classification of human lung tumors. *Sci. Transl. Med.*, **5**, 209ra153.

Taylor,B.S. *et al.* (2008) Functional copy-number alterations in cancer. *PLos One*, **3**, e3179.

Weiss,J. *et al.* (2010) Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Sci. Transl. Med.*, **2**, 62ra93.