

CNVScope: Visually Exploring Copy Number Aberrations in Cancer Genomes

James LT Dalgleish^{id}, Yonghong Wang, Jack Zhu and Paul S Meltzer

Genetics Branch, National Cancer Institute, Center for Cancer Research, National Institutes of Health, Bethesda, MD, USA.

Cancer Informatics
Volume 18: 1–6
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935119890290



ABSTRACT

MOTIVATION: DNA copy number (CN) data are a fast-growing source of information used in basic and translational cancer research. Most CN segmentation data are presented without regard to the relationship between chromosomal regions. We offer both a toolkit to help scientists without programming experience visually explore the CN interactome and a package that constructs CN interactomes from publicly available data sets.

RESULTS: The CNVScope visualization, based on a publicly available neuroblastoma CN data set, clearly displays a distinct CN interaction in the region of the *MYCN*, a canonical frequent amplicon target in this cancer. Exploration of the data rapidly identified *cis* and *trans* events, including a strong anticorrelation between 11q loss and 17q gain with the region of 11q loss bounded by the cell cycle regulator *CCND1*.

AVAILABILITY: The shiny application is readily available for use at <http://cnvscope.nci.nih.gov/>, and the package can be downloaded from CRAN (<https://cran.r-project.org/package=CNVScope/>), where help pages and vignettes are located. A newer version is available on the GitHub site (<https://github.com/jamesdalg/CNVScope/>), which features an animated tutorial. The CNVScope package can be locally installed using instructions on the GitHub site for Windows and Macintosh systems. This CN analysis package also runs on a linux high-performance computing cluster, with options for multinode and multiprocessor analysis of CN variant data. The shiny application can be started using a single command (which will automatically install the public data package).

KEYWORDS: copy number, cancer, CNV, CNA, copy number variation, visualization, R, shiny

RECEIVED: October 14, 2019. **ACCEPTED:** October 30, 2019.

TYPE: Software or Database Review

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the Center for Cancer Research, part of the Intramural Research Program at the National Cancer Institute of the National Institutes of Health.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: All authors were employees or fellows of the National Cancer Institute.

CORRESPONDING AUTHOR: Paul S Meltzer, Genetics Branch, National Cancer Institute, Center for Cancer Research, National Institutes of Health, 37 Convent Dr., Bethesda, MD 20892-4265, USA. Email: pmeltzer@nih.gov

Introduction

Genome-wide DNA copy number (CN) data are an essential aspect of integrative cancer genome analyses directed at identifying dysregulated pathways in cancer.¹ Identification of regions and genes of interest in CN data has primarily been accomplished through the identification of consensus regions of alteration and statistically rationalized with tools such as Genomic Identification of Significant Targets in Cancer (GISTIC) that identify individual regions of recurrent CN alteration.² While useful, this approach does not address the impact of co-associations of distant genetic loci or visualize complex interactions clearly.

In contrast, Hi-C methodology maps physical interactions between chromosomes at specific loci, allowing the derivation of a matrix of chromosomal interactions of great utility in studies of 3-dimensional chromatin structure.³ Visualizations of such a matrix can show hot spots of interactions between regions, whereas edge-node graphs and CIRCOS plots can become cluttered and nearly uninterpretable. Matrices can display more interactions per unit area in a clear fashion, and the matrix display of interactions shows interaction domains with row and column order preserved in the matrix.⁴

Merging cancer CN data with a matrix-mapping approach similar to Hi-C analysis, we have developed methods to analyze

and interactively display genome-wide interactions from a CN data set, with each matrix value representing the strength of the interaction between loci. Two current trends tend to encourage investigations that benefit from this type of interaction data. One is the growing understanding of the topology and specificity of nuclear chromosome territories, and the other is the ever more widespread use of whole genome sequencing in cancer genomics, which allows unprecedented precision in mapping structural variants and local CN. Particularly in cancers with extensive genome rearrangements, there is an unmet need for tools that facilitate the discovery of genomic aberrations that depend on aspects of higher order nuclear organization. Our goal has been to develop a method that essentially precomputes and visualizes signed correlations between any 2 points in the genome using binned segmented CN values from a large set of cancer samples. We found that a recursive linear regression algorithm produces visually intuitive, interpretable results that are consistent with known aspects of chromosome structure and genome rearrangement that can also rapidly identify novel features.

Here, we report this new methodology, R package, and a suite of Web-based tools accessible to the scientific community for the exploration of complex CN data sets to generate hypotheses connecting CN phenomena and their underlying



chromosome aberrations to the pathogenesis of various cancers. The package can also accept The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) data, allowing for analysis of a broad range of cancers using existing large cohort studies.⁵ We have also provided guidance for the importation of unpublished user data. The R package is highly accessible, combining vignettes, documentation, examples, and an animated tutorial.

As an example of the application of CNVScope, we chose a sufficiently large publicly available neuroblastoma (NBL) data set. This aggressive childhood cancer has known features, notably the pattern of clustered chromosomal breakpoints in the *MYCN-NBAS* region, occurring with *MYCN* amplification, often consistent with episomal structures such as double minutes.⁵ Visualizing this data set with CNVScope reveals a distinctive signature in this region as well as several other intra- and interchromosomal features illustrated below.

Methods

Input matrix using NBL data set

From the GDC legacy archive, an NBL data set of 126 samples was obtained (see vignette). Data were aggregated into a binned sample matrix with 1 Mb bins. Each bin value corresponded to the average segmentation value for TCGA or mean relative coverage for TARGET for segments that overlapped the bin in that specific sample. Row names signified bin genomic position, whereas column names represented the sample identifiers. This input matrix was then used as the basis for the matrix of log P -values.

Linear regression, postprocessing, and matrix set formation

From this input matrix, a matrix of negative $\ln P$ -values was taken for all combinations of every genomic range bin of aggregated segmentation values against every other bin to form a matrix of values denoting the numeric association between regions based on segmentation values. The correlation sign was multiplied by each value so that those with negative associations would have a negative value and, after a numeric transform into the [0,1] color space, would be rendered as blue in color, whereas those with positive associations would be rendered as red. We also annotated each of the region pairs with genes and created 529 smaller matrices (1 for each chr1-X chromosomal pair) that could be easily viewed using the plotly R toolkit implementation in our shiny application (Figure 1). See supplementary methods, R documentation, and vignettes for more information. To view the vignettes, type `browseVignettes("CNVScope")` after installation instructions on the site (<https://github.com/jamesdalg/CNVScope/>) are followed.

Features

The CNVScope app allows the user to quickly identify hot spots and large features in a chromosomal interaction plot and

provides a clear view of the contributing samples to every single value in the matrix. Genes and expression transcript levels are identified at every combination of genomic loci. COSMIC census genes are also noted. The matrix data also can be explored using a gene search tool to provide coordinates based on ensembl-75 (hg19). With coordinates specified, users can then plot the view zoomed directly on their location pair of interest.

Controls

The application features a gene search tool to get exact gene positions, a saturation threshold slider to control the effect of outlier pixels, a heatmap height slider, dropdowns for chromosomes, and a plot button. We have provided the NBL data in complete form along with several clinical subsets. The users are also given a choice of relationship metric correlation sign*— $\log(P$ -value) or correlation. Correlation type can be selected, with users suggested to use Pearson for linear relationships, with Spearman and Kendall able to detect linear or nonlinear relationships. A P -value filter has been added, allowing users to filter out squares in the matrix that do not have a slope significantly different than 0. This P -value filter is based on false discovery rate—corrected P values being less than .05. The *Main plot* is an inter-/intrachromosomal plot of relationship values between pairs of region segmentation values. Genes and raw P values are shown on hovering, and domains/hotspots can be found while exploring the chromosomal interaction pair. Gene names can be disabled before plotting, if desired. Upon clicking, searchable lists of row/column genes appear. As expected, the strongest signal appears as the segment distance approaches 0 on the diagonal, but the local strength of this signal varies. A break in the local correlation is typically observed at centromeres related to the high frequency of centromeric chromosome rearrangements. Domain boundaries detected by image segmentation are shown on the edges. The *Mini-map* provides a close view of the main plot to see subtle details. *Census Genes* provides the COSMIC annotated genes (<https://cancer.sanger.ac.uk/census>) in the region, along with tumor type, tissue of origin, and known roles (eg, oncogene, fusion).⁶ *Sample-level information* provides 2 histograms overlaid with an opacity slider to improve visibility. A regression scatterplot shows individual segmentation values for the clicked point of the main plot, colored by sample to show the direction of the relationship as well as outliers and sample clusters. *Expression Data* provides access to the NBL expression mean and variance of all genes within the region, ordered by the expression variance percentile. *Whole Genome View* shows all chromosomes in a static map, labeled by chromosome, with a saturation slider to find regions to explore at a chromosomal level in the main plot.

Package vignettes

The package vignettes detail the process to import GDC data with images of the requisite steps, perform the relationship

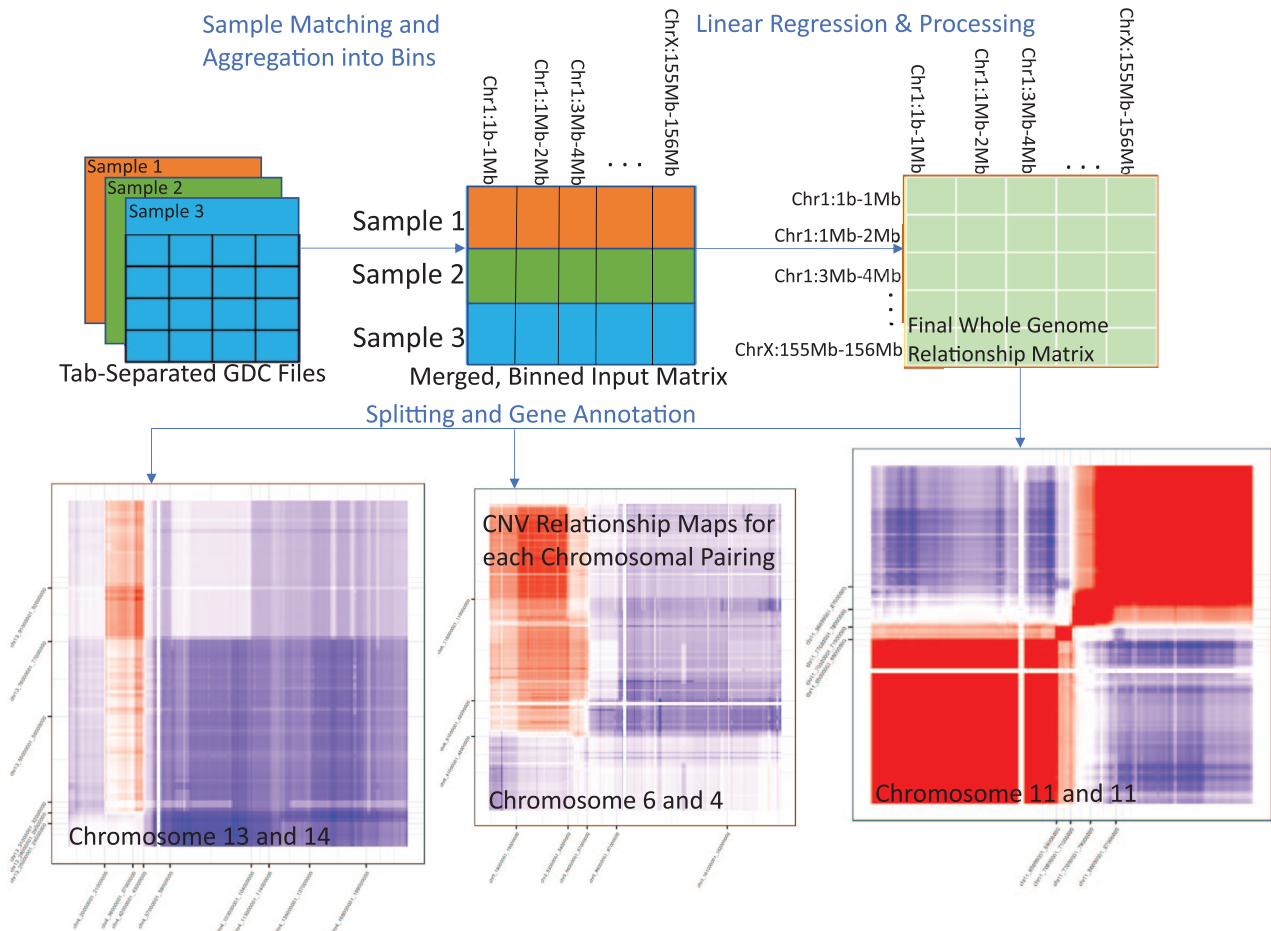


Figure 1. Workflow from GDC TARGET neuroblastoma CN data to finalized interchromosomal matrices used in the shiny application. Files are converted from GDC tab-delimited files with varying bin sizes into an input matrix of even 1 Mb bins and sample identifiers, and then into relationship metrics from linear regression (the negative log P -value). Postprocessing then sets the infinities to a high number to allow visualization and adds negative signs in regions where the strong relationship is an inverse linear relationship. Finally, the large, postprocessed matrix is converted to many small matrices with the ensembl-75 genes mapped to each genomic bin pair.

TARGET indicates Therapeutically Applicable Research to Generate Effective Treatments; CN, copy number; CNV, copy number variant.

mapping using a high-performance computing system, post-process the matrix, and briefly visualize results. A power analysis vignette is also provided, which suggests a minimum sample size of 108 individual CN cases. We also wish to note that several other cancer data sets have been demonstrated to work with the toolkit, including bladder cancer, prostate cancer, acute myeloid leukemia, and melanoma. A brief demonstration of the toolkit on these data sets is provided within the GitHub package.

Specific Observations in NBL

To understand the information carried by the CNVscope main plot, it is useful to examine the whole genome plot arising from the 126-sample NBL data set (Figure 2A). The strong correlation signal (red) on the diagonal represents the high probability of CN correlation of adjacent segments related to their chromosome topology. Note that the signal is not confined to the geometric diagonal but extends variably some distance from that line. The simplest example is the X chromosome that other than the small clearly delineated

discontinuities of the pseudoautosomal regions appears as a rather uniform block due to the fact that each sample originated from either an XX or an XY genotype. Most other chromosomes exhibit a more complex pattern, with principal blocks often demarcated at the centromeres, consistent with the known high frequency of whole chromosome arm rearrangements in cancer. For example, on the $\text{chr20} \times \text{chr20}$ plot, independent correlation blocks exist for p and q arms, with breaks in correlation at the centromere. The 20p arm correlation block ends at 26 to 27 Mb, and the p arm block begins at 29 to 30 Mb, with jointseg calling boundaries at that these loci corresponding to the centromeric ends of the alignable sequence for each arm. Remarkably, CNVscope allows these boundaries to be readily discerned against a background of high correlation for the entirety of chr20, with only 15 of 126 (11.9%) samples showing chr20 arm-specific CN aberrations. On the other chromosomes, local decreases in *cis*-association signal suggest the presence of focal CN aberrations. For example, on chr2 we observe a distinct biological signature (correlation dropping sharply off diagonal) precisely at the *MYCN* locus

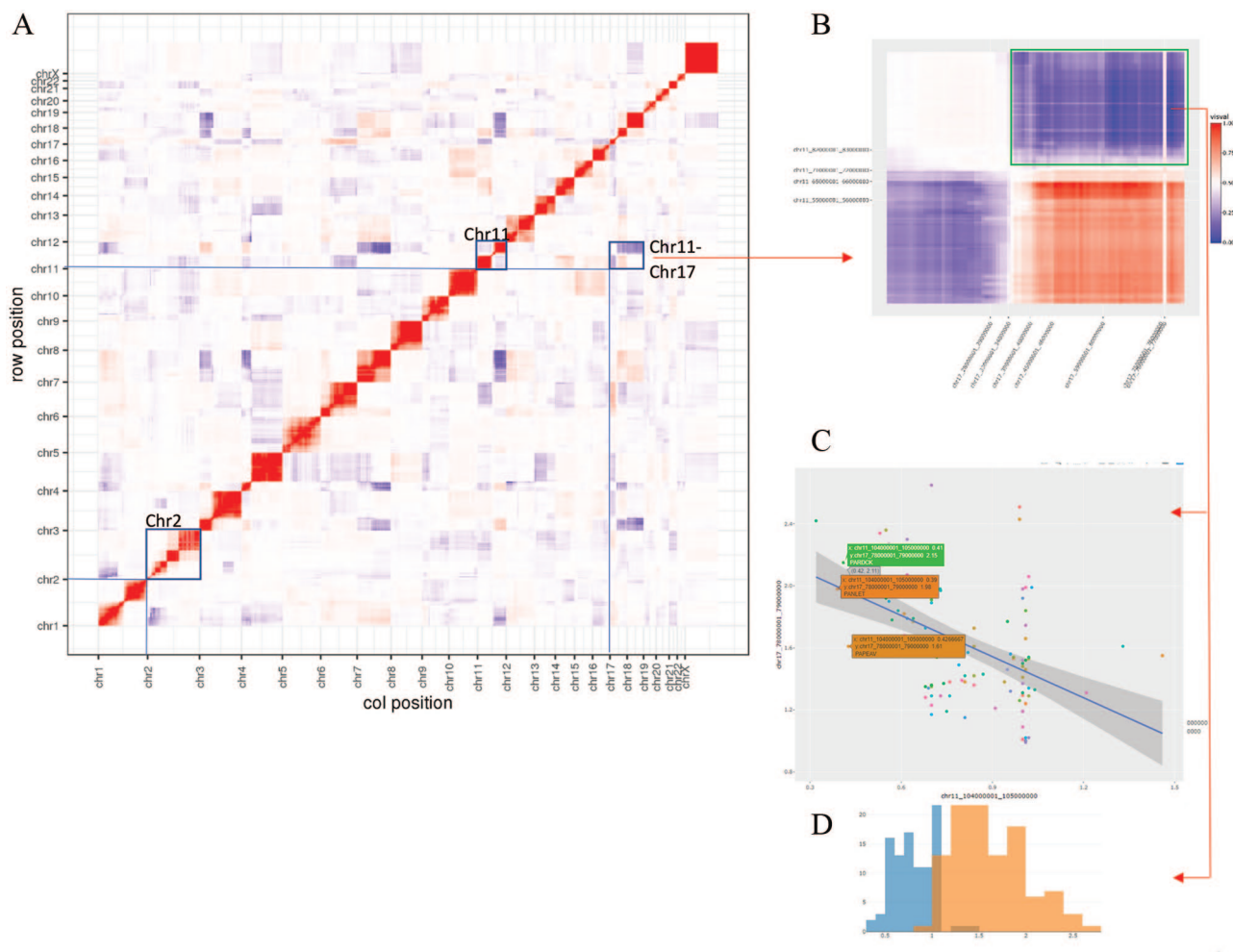


Figure 2. (A) A whole genome interaction view of neuroblastoma copy number (CN) associations (chr1-X). Boxed regions highlight chr2 (enlarged in Figure 3), chr11, and the negatively signed off-diagonal association of 11q and 17q. (B) The enlarged chr11-chr17 map illustrates the strong anticorrelated regions of 11q-17q. The lowest correlation point is highlighted ($r=-0.482$, Benjamini-Hochberg adjusted P value is .000117). The number of tests used for the adjustment is the number of bin pairs in the whole genome. (C) Representative regression plot of CN values on 11q and 17q illustrates an anticorrelation trend as well as the detailed sample-level data on 3 tooltips. (D) The data from the same coordinates as (C) are represented as a histogram showing clear separation of CN value distributions on 11q and 17q. These plots support the interpretation that there is a significant association of 11q loss with 17q gain.

(Figure 3). *MYCN* amplification has been suggested to be “the best characterized genetic marker of risk in neuroblastoma,” with approximately 25% of patients exhibiting the *MYCN* amplification associated with risk of poor outcome.⁷ CNVScope provides a new, compelling visualization of CN data illustrating the domain of the *MYCN* amplicon. This signature is apparent in either correlation or linear regression views on the application. Thus, we observe that a known marker of NBL shows a clear, visual chromosome interaction signature.

The G1/S-phase cyclin *CCND1* also shows a similar focal signature on the diagonal of the chr11 \times chr11 plot, within a microdomain of 7 Mb delineated by 2 jointseg edges (Figure 4). *CCND1* has been suggested to have a role in differentiation, proliferation, and cell cycle progression in NBL.⁸ CNVScope identifies *CCND1* as likely to be important in NBL biology. Interestingly, additional information from intrachromosomal analysis adds depth to this observation.

From the whole genome plot, there are many off-diagonal regions demonstrating significant signal. In particular, 17q and 11q show strong anticorrelated regions visible in both the whole genome and the interchromosomal views as a large block (blue) (Figure 2(B)). Histograms validate that these regions have 2 distinct distributions that are very well separated, and a linear regression view of a single sample makes clear the downward trend driving the color coding (dark blue) visually displayed in the interchromosomal view (Figure 2(C and D)). These features allow the user to drill down from the relatively abstract view of the main plot to the detailed underlying data and appreciate that the genetic phenomenon flagged in CNVScope is the significant co-occurrence of 11q loss and 17q gain. This phenomenon has been previously reported in NBL.^{8,9} Remarkably the anticorrelated portion of 11q is bounded by a jointseg edge at 71 to 72 Mb, indicating that 11q loss consistently begins telomeric to *CCND1* preserving its function as a tumor driver.

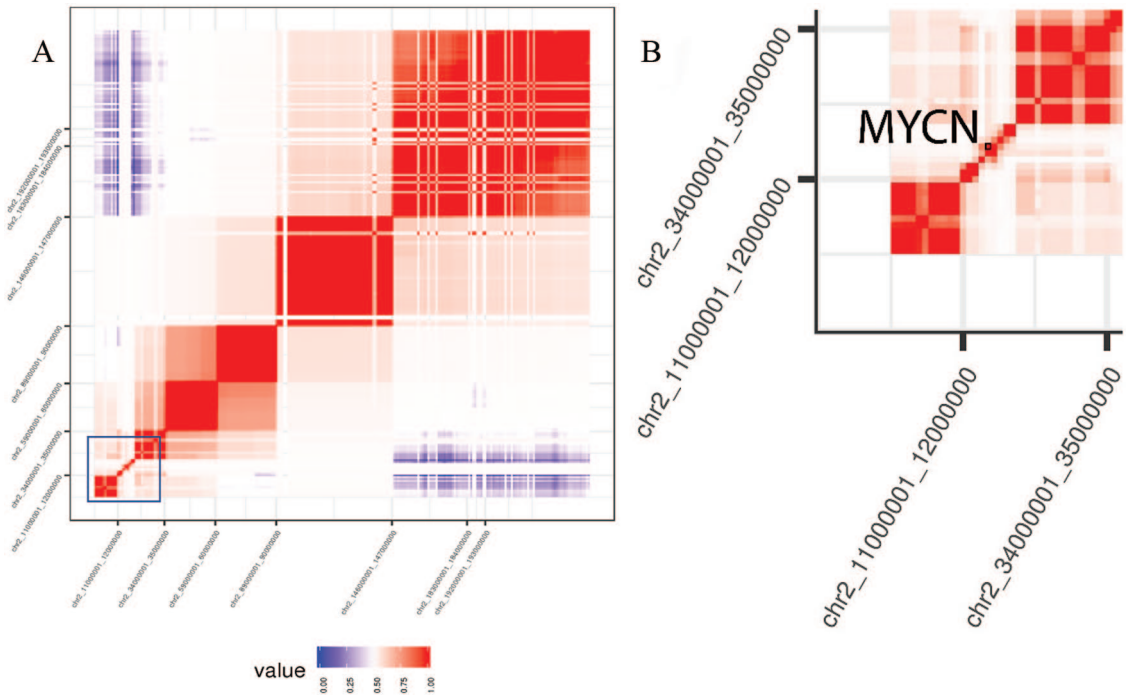


Figure 3. (A) Intrachromosomal association plot for chromosome 2. The box highlights a distinct feature on the diagonal indicating narrowing of the region of local co-association, and white lines emanating from that region show a reduction in association from the MYCN locus across all loci on the chromosome. (B) Enlarged view of the MYCN amplification domain with a break in linear regression signal near MYCN most likely due to amplification of MYCN in relatively small, often extrachromosomal, amplicons. The MYCN locus is highlighted.

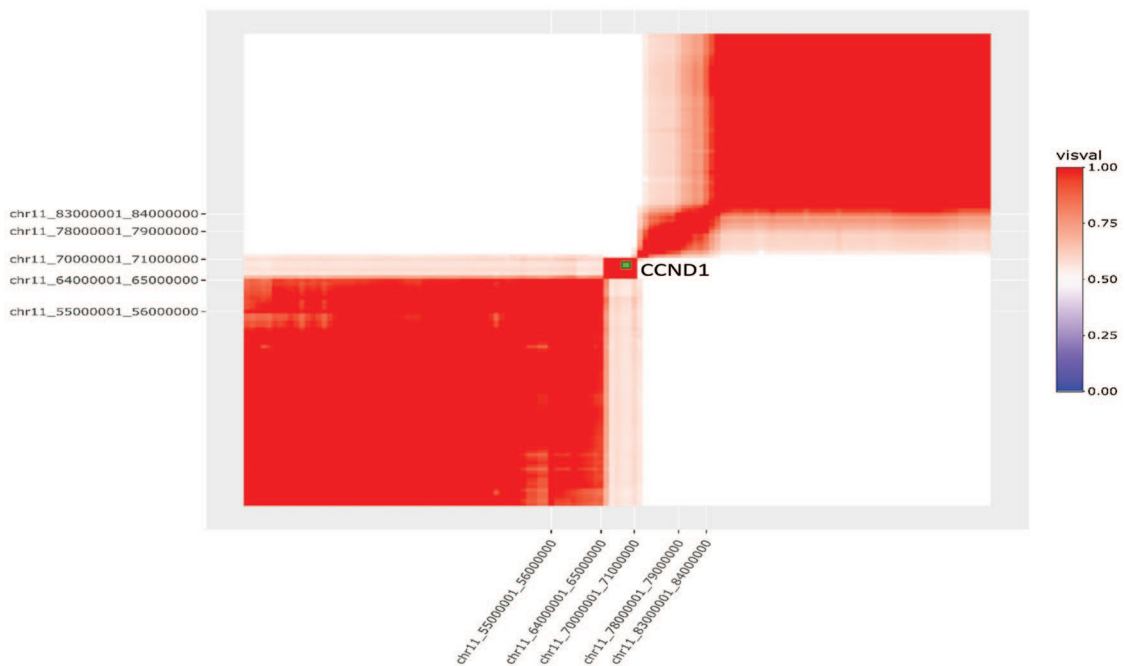


Figure 4. *CCND1*, a focally amplified cell cycle regulator, is located within a distinct association domain in 11q13. In this plot of chr11-chr11, the sharp reduction of chr11-chr11 association is likely the result of structural variants leading to *CCND1* copy number (CN) gain. False discovery rate *P*-value filtering was applied to this Pearson correlation plot of CN associations on chr11. *CCND1* is located precisely at chr11:69455855-69469242.

Many studies have examined single gene-gene associations for CN amplifications and deletions, but CNVScope does this on a whole genome scale, providing a survey view and rapid access to significant associations while also allowing access to the primary data, gene annotations, and other data types that a

user might wish to integrate with CN data.¹⁰⁻¹² In conclusion, we have described the methodology, the detailed features, and the potential of CNVScope to highlight significant genomic events such as those we have described in NBL. We invite others to explore the regions and hot spots which may be related

to functionally important aspects of NBL genome biology and to use CNVScope to explore other cancer genomics projects with available CN data.

Limitations

It is important to point out that CNVScope resolution is ultimately limited by the probe density of the input data and the bin size selected. A 1-Mb bin size was chosen for the visualization tool to allow swift and stable function of the application. We feel that this is a reasonable compromise between resolution and computational limitations. It is also consistent with many existing data sets. We also note that the toolkit allows for the use of custom data to generate the relationship matrix should users with sufficiently high-resolution data wish to create an extremely high-resolution view of a selected region. Smaller and larger bin sizes have been tested on the NBL data set (0.1 and 10Mb). Both the function and the commands for this have been listed in the input matrix vignette. The main focus of this work is to facilitate the rapid analysis of CN associations in integrative cancer genomics studies through the visualization of a precomputed association matrix.

Acknowledgements


We thank Stamen Mitev of CBIIT for his aid in bringing the server live and Mikol Ware for help in making the server public. We also thank Sean Davis for his contributed code and advice. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

Author Contributions

J.L.T.D. wrote the package, application, and the manuscript. P.S.M. gave feedback on the package, cowrote the manuscript, supervised the project as it progressed, and contributed the initial idea of 2-dimensional (2D) segmentation maps. J.Z. collaborated in these efforts and often provided supervision/aid during the project and processed expression data. Y.W. also

contributed to the development of the idea of 2D segmentation maps, made several initial matrices, and contributed intellectually during the course of the project.

ORCID iD

James LT Dalglish  <https://orcid.org/0000-0002-2053-8786>

Supplemental Material

Supplemental material for this article is available online at <http://cnvscope.nci.nih.gov/>; <https://cran.r-project.org/package=CNVScope>; and <https://github.com/jamesdalg/CNVScope/>.

REFERENCES

1. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer*. 2014;14:299.
2. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12:R41.
3. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326:289-293. doi:10.1126/science.1181369.
4. Durand NC, Robinson JT, Shamim MS, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*. 2016;3:99-101. doi:10.1016/j.cels.2015.07.012.
5. Pugh TJ, Morozova O, Attiyeh EF, et al. The genetic landscape of high-risk neuroblastoma. *Nat Genet*. 2013;45:279-284.
6. Harsha B, Creatore C, Kok CY, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2018;47:D941-D947.
7. Huang M, Weiss WA. Neuroblastoma and MYCN. *Cold Spring Harb Perspect Med*. 2013;3:a014415.
8. Mlakar V, Jurkovic Mlakar S, Lopez G, Maris JM, Ansari M, Gumy-Pause F. 11q deletion in neuroblastoma: a review of biological and clinical implications. *Mol Cancer*. 2017;16:114.
9. Ho N, Peng H, Mayoh C, et al. Delineation of the frequency and boundary of chromosomal copy number variations in paediatric neuroblastoma. *Cell Cycle*. 2018;17:749-758.
10. Cappuzzo F, Varella-Garcia M, Rossi E, et al. MYC and EIF3H Coamplification significantly improve response and survival of non-small cell lung cancer patients (NSCLC) treated with gefitinib. *J Thorac Oncol*. 2009;4:472-478.
11. Dancau A-M, Wuth L, Waschow M, et al. PPF1A1 and CCND1 are frequently coamplified in breast cancer. *Genes Chromosomes Cancer*. 2010;49:1-8.
12. Howitt BE, Sun HH, Roemer MGM, et al. Genetic basis for PD-L1 expression in squamous cell carcinomas of the cervix and vulva. *JAMA Oncol*. 2016;2:518-522.