SOFTWARE TOOL ARTICLE

An accessible GenePattern notebook for the copy number variation analysis of Illumina Infinium DNA methylation arrays [version 1; peer review: 2 approved]

Clarence K. Mah[1], Jill P. Mesirov[1,2], Lukas Chavez[1]

[1]Department of Medicine, University of California, San Diego, La Jolla, CA, 92093, USA
[2]Moores Cancer Center, University of California, San Diego, La Jolla, CA, 92093, USA

v1 **First published:** 05 Dec 2018, **7**:1897
https://doi.org/10.12688/f1000research.16338.1

Latest published: 05 Dec 2018, **7**:1897
https://doi.org/10.12688/f1000research.16338.1

Abstract

Illumina Infinium DNA methylation arrays are a cost-effective technology to measure DNA methylation at CpG sites genome-wide and across cohorts of normal and cancer samples. While copy number alterations are commonly inferred from array-CGH, SNP arrays, or whole-genome DNA sequencing, Illumina Infinium DNA methylation arrays have been shown to detect copy number alterations at comparable sensitivity. Here we present an accessible, interactive GenePattern notebook for the analysis of copy number variation using Illumina Infinium DNA methylation arrays. The notebook provides a graphical user interface to a workflow using the R/Bioconductor packages *minfi* and *conumee*. The environment allows analysis to be performed without the installation of the R software environment, the packages and dependencies, and without the need to write or manipulate code.

Keywords

Illumina Infinium methylation arrays, DNA methylation, copy number variation, pre-processing, interactive, visualization, GenePattern Notebook, Jupyter Notebook, open-source, conumee, minfi, R/Bioconductor

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 05 Dec 2018	✓ report	✓ report

1 **Robert Ivanek**, University of Basel, Basel, Switzerland

2 **Aris Floratos**, Columbia University, New York, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the GenePattern collection.

Corresponding authors: Clarence K. Mah (ckmah@ucsd.edu), Lukas Chavez (lukaschavez@ucsd.edu)

Author roles: Mah CK: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Mesirov JP**: Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Chavez L**: Conceptualization, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was funded by the National institutes of Health (grant numbers U24CA194107, U01CA184898, U41HG007517 and R01CA109467 to J.P.M.).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Mah CK, Mesirov JP and Chavez L. **An accessible GenePattern notebook for the copy number variation analysis of Illumina Infinium DNA methylation arrays [version 1; peer review: 2 approved]** F1000Research 2018, **7**:1897
https://doi.org/10.12688/f1000research.16338.1

First published: 05 Dec 2018, **7**:1897 https://doi.org/10.12688/f1000research.16338.1

Introduction

Although Illumina Infinium DNA methylation arrays, including the 450k and EPIC ("850k") BeadChips, have been designed for detecting genome-wide DNA methylation, the resulting data can also be used to analyze copy number profiles (Feber *et al.*, 2014). This feature allows the simultaneous analysis of DNA methylation and copy number variation (CNV) and reduces the quantity of material needed to perform both analyses. We have implemented an Illumina Infinium DNA methylation array-based CNV analysis workflow as an accessible, interactive GenePattern notebook, which integrates background information, workflow instructions, a graphical user interface, source code, and the results in a single electronic notebook document (Mah, 2018). Leveraging the popular GenePattern Notebook environment (Reich *et al.*, 2017), the notebook enables the sharing of reproducible analyses and results.

The workflow is initiated by a single step and performs two main analyses: loading and preprocessing the data, and copy number analysis (Figure 1). Multiple samples can be analyzed in parallel. The preprocessing step utilizes the *minfi* R package to load and process Illumina Infinium DNA methylation array data and to perform data normalization (Aryee, 2014). Copy number analysis is performed using the *conumee* R package, which compares each sample to a set of user-provided normal reference samples (Hovestadt & Zapatka, 2015). This analysis outputs a set of copy number plots for the entire genome, individual chromosomes, and for user defined gene loci of interest. Copy number profiles are described as segments along the genome and can be exported as text files for visualization with tools such as the Integrated Genome Viewer (Robinson *et al.*, 2011) and for further analysis.

Methods

Implementation

The entire workflow is implemented as a GenePattern notebook, which can be accessed at the GenePattern Notebook Repository (http://www.genepattern-notebook.org/) and run there by the user. Data preprocessing and CNV analysis steps are implemented as a GenePattern module (Reich *et al.*, 2006) and utilized by the *MethylationCNVAnalysis* notebook.

Load and preprocess data

To begin the analysis, two sets of data are required: the query sample data for which the copy number profiles are to be analyzed and appropriate control sample data used to establish baseline copy number profiles for comparison (Figure 2). The input data for this notebook (query and control samples) are raw IDAT files generated by the microarray scanner, representing two different color channels prior to normalization. As described in the *minfi* documentation, IDAT files are the most complete data types, because they include measurements on control probes, which are necessary for assessing bisulfite conversion efficiency and for normalizing technical variability.

To load the Illumina Infinium methylation array data into the notebook, the IDAT files must be combined into a single archive (.zip or .gz formats). The archive can be organized either as a flat archive where all IDAT files are packed without subfolders, or as an archive in the standard folder structure as presented in

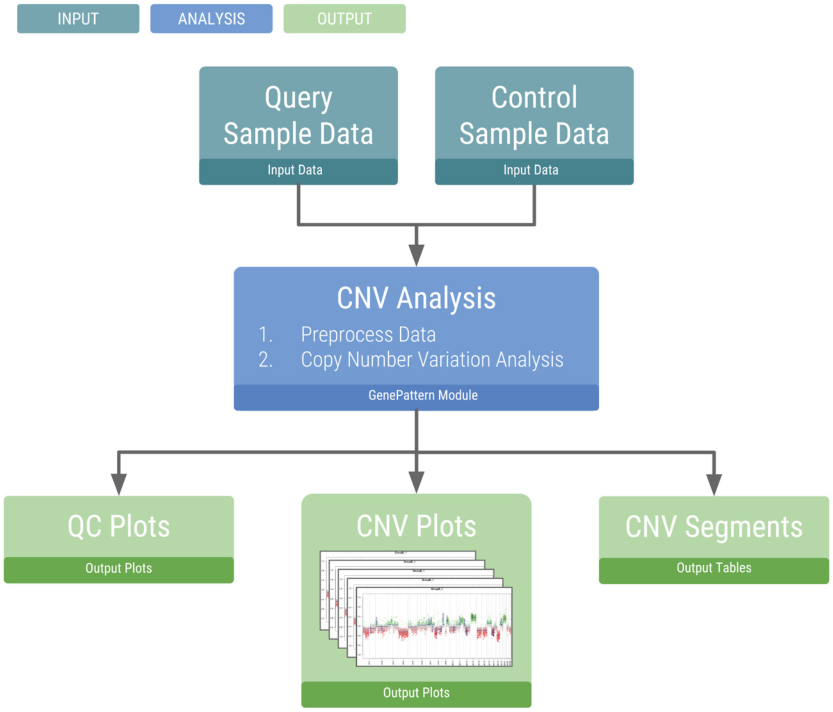
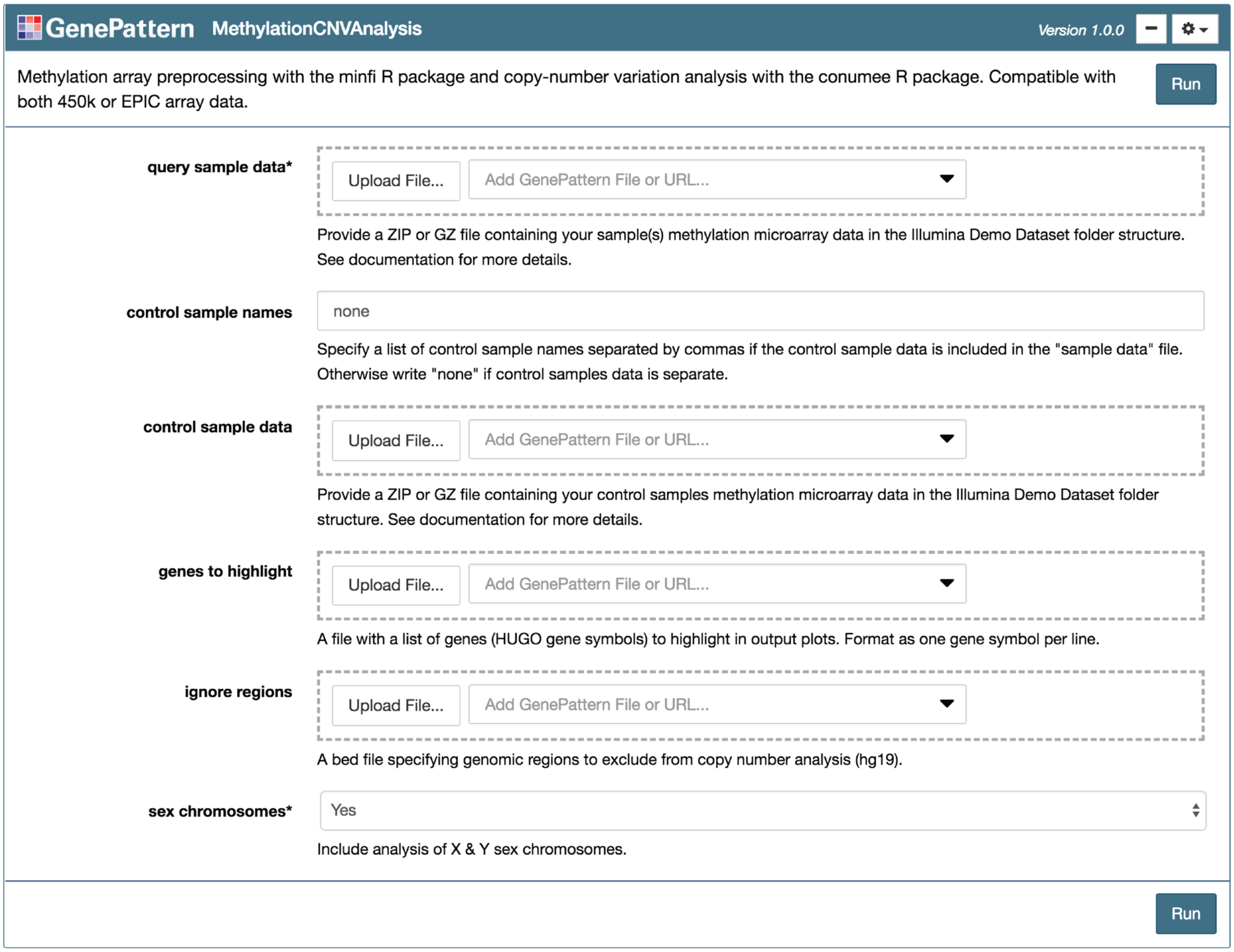


Figure 1. Analysis workflow. The flowchart shows the main inputs and outputs necessary for the copy number variation analysis.



Methylation array preprocessing with the minfi R package and copy-number variation analysis with the conumee R package. Compatible with both 450k or EPIC array data.

query sample data* — Upload File... / Add GenePattern File or URL...

Provide a ZIP or GZ file containing your sample(s) methylation microarray data in the Illumina Demo Dataset folder structure. See documentation for more details.

control sample names — none

Specify a list of control sample names separated by commas if the control sample data is included in the "sample data" file. Otherwise write "none" if control samples data is separate.

control sample data — Upload File... / Add GenePattern File or URL...

Provide a ZIP or GZ file containing your control samples methylation microarray data in the Illumina Demo Dataset folder structure. See documentation for more details.

genes to highlight — Upload File... / Add GenePattern File or URL...

A file with a list of genes (HUGO gene symbols) to highlight in output plots. Format as one gene symbol per line.

ignore regions — Upload File... / Add GenePattern File or URL...

A bed file specifying genomic regions to exclude from copy number analysis (hg19).

sex chromosomes* — Yes

Include analysis of X & Y sex chromosomes.

Figure 2. Copy number variation analysis GenePattern Notebook interface. The “MethylationCNVAnalysis” module is presented as an input form using the GenePattern Notebook graphical user interface. The user links or uploads input files and selects analysis parameters before pressing “Run” to execute the workflow.

the Illumina demo dataset. The IDAT archive can be selected and loaded through the graphical user interface of the GenePattern notebook. Both 450k array and EPIC array types are compatible as long as all samples in a single archive are of the same array type. If the query samples or control samples are of different array types, only the common set of probes between 450k and EPIC array types are evaluated across all samples.

For each sample, the data is normalized with respect to background and positive control probes on the arrays according to the implementation in Illumina’s proprietary GenomeStudio software. Upon loading the data, the notebook generates a quality control report containing two plots for identifying poor quality samples. The first plot shows the $\log_2$ median intensity of the methylated versus unmethylated channels (Figure 3A). Poor-quality samples tend to have lower median intensities and separate from the good quality samples. The second plot shows the DNA methylation levels (Beta values) of all probes on the array and for all samples as a density plot in which we expect to see a bimodal distribution with peaks at zero (no methylation) and one (100% methylation) (Figure 3B).

Control samples should be free of CNVs and have a similar methylation profile as the samples of interest. The best practice is to use control samples of the corresponding normal tissue type. If control samples are included in the query sample dataset, no separate data needs to be loaded. Instead, the control samples can be specified by providing the sample names in the CNV analysis step. Otherwise, the control data will be loaded as a separate archive of IDAT files.

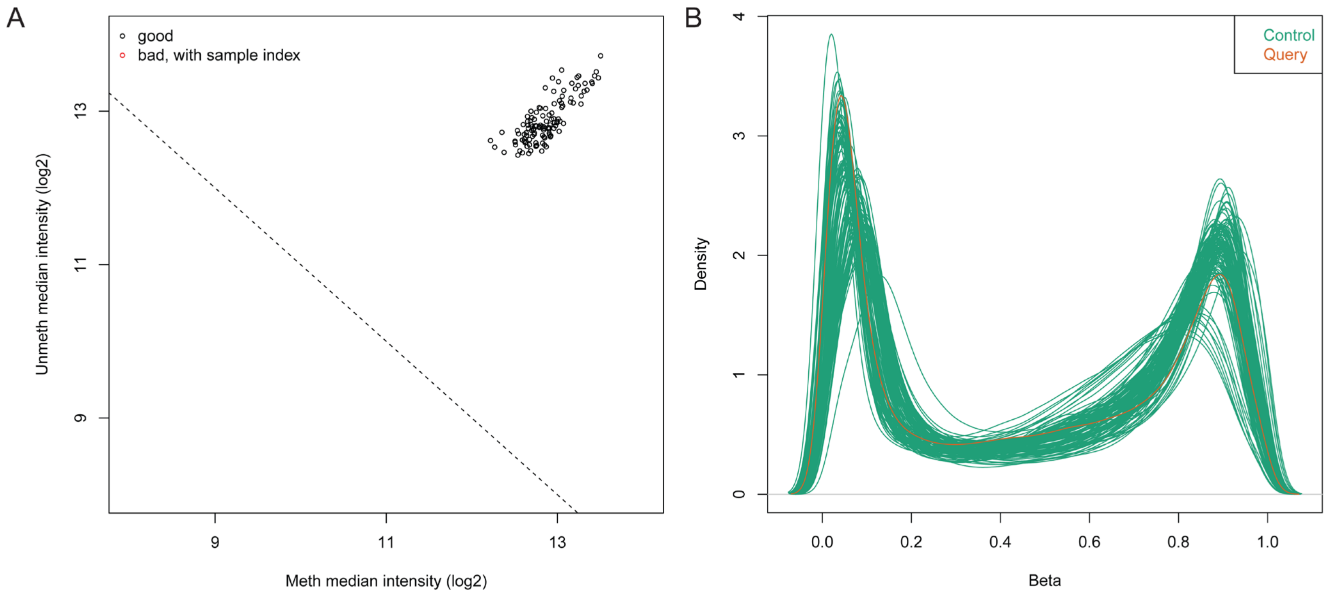


Figure 3. Plots of query and control samples. (**A**) Median intensity plot of query & control samples. Log median intensity of the methylated channel is along the x-axis and log median intensity of unmethylated channel is along the y-axis. Bad-quality samples fall under the threshold and are colored red. There is no bad quality sample in this plot. (**B**) DNA methylation (Beta-value) density plot of query & control samples. A density plot showing the distribution of beta values across each sample. Beta values should be bimodal and peak around 0 and 1.0.

CNV analysis

As outlined in the *conumee* documentation, the copy number analysis is performed as follows: each query sample is normalized to the control samples by multiple linear regression yielding the linear combination of control samples that most closely fits the intensities of the query sample. Next, the $\log_2$ ratio of probe intensities of the query sample versus the combination of control samples are calculated. Probes are then combined within predefined genomic bins. Intensity values are shifted to minimize the median absolute deviation of all bins to zero to determine the copy-number neutral state. The genome is segmented into regions of the same copy number state using the circular binary segmentation algorithm (Seshan & Olshen, 2018).

Genomic loci of genes to be highlighted in the CNV plots are retrieved from the hg19 Ensembl database using the BiomaRt R package (Durinck, 2005; Durinck, 2009). The notebook also offers an option to exclude regions from analysis, such as highly polymorphic regions that would yield inaccurate copy number calls. In addition, X and Y chromosomes can be excluded to avoid misleading results in case no appropriate control data is available.

Operation

To run the *MethylationCNVAnalysis* notebook, the user must have a GenePattern account that can be created on the GenePattern Notebook website (http://genepattern-notebook.org). After logging in, the notebook can be found in the "Community" section of the "Public Notebooks" page. The notebook can then be run from the GenePattern Notebook site, with no additional software installations needed.

Use case

The use case presented by the notebook evaluates the copy number profile of a glioblastoma tumor analyzed by an Illumina Infinium 450k DNA methylation array. This sample has been classified as an IDH wild-type midline glioblastoma according to the methylation-based classifier described by Capper *et al.* (2018). Recurrent chromosomal alterations of this tumor type are gain of chromosome 7 with or without EGFR amplification (>80%), loss of 9p21 (CDKN2A/B; >50%) and chromosome 10 loss (>70%). Amplifications of the PDGFRA oncogene are enriched in this class (present in 20–30% of cases) (Capper, 2018).

We used the 450k methylation profiles of 119 normal brain tissue samples as the corresponding control data (Capper, 2018). By inspecting the generated CNV plots, we can visually identify significant copy number loss of CDKN2A/B relative to normal brain tissues (Figure 4). Additionally, several copy number changes that are associated with glioblastoma stand out, notably MET amplification and loss of RB1.

Conclusion

The GenePattern notebook *MethylationCNVAnalysis*, hosted in the GenePattern Notebook Repository, processes Illumina Infinium DNA methylation array data and generates CNV segments and plots. Different designs of Illumina Infinium DNA

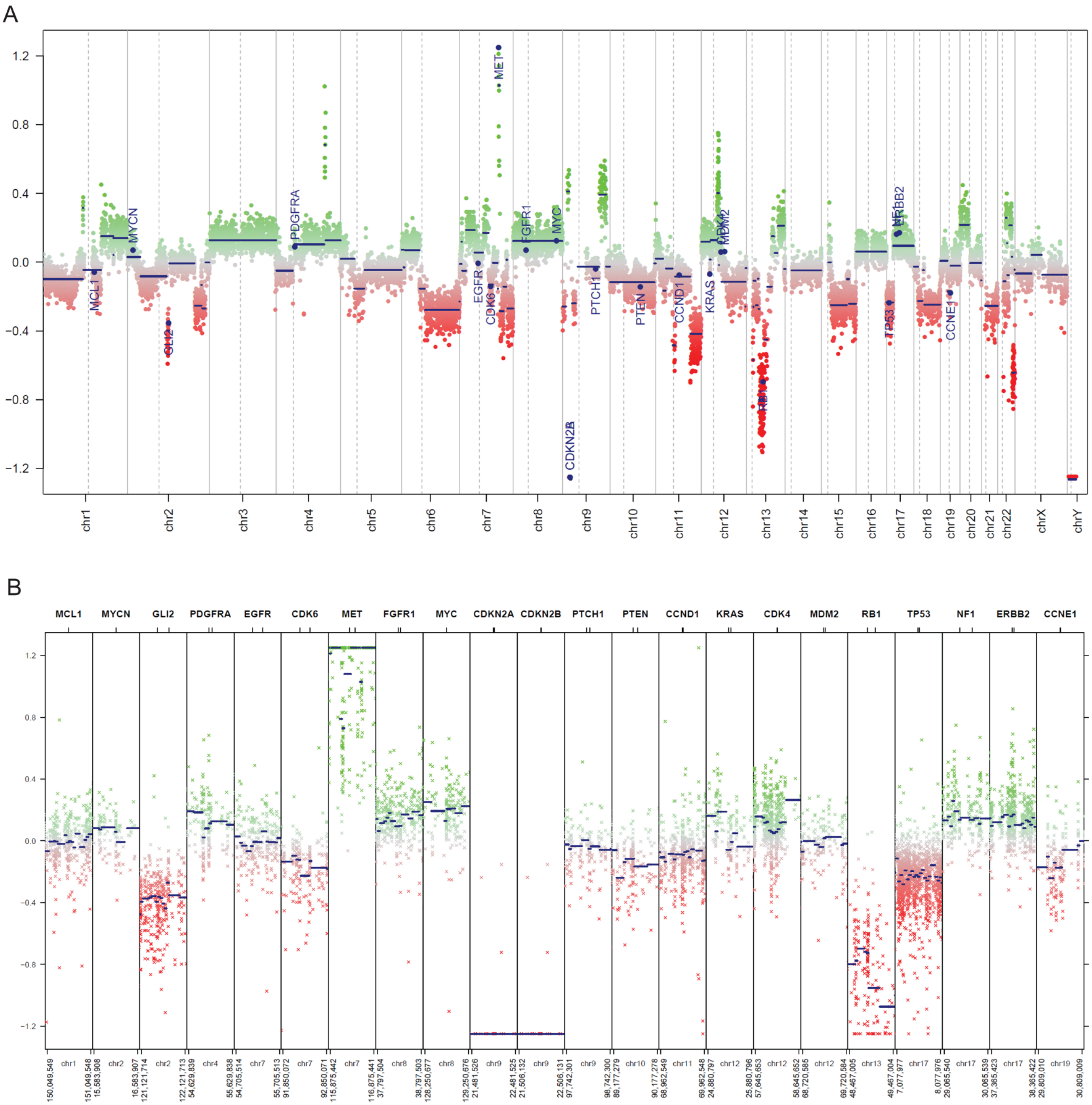


Figure 4. Plots of copy numbers. (**A**) Copy number plot of the entire genome in the example glioblastoma sample. A plot of all chromosomes across the genome. Intensity values of each bin are plotted as colored dots, green indicating above normal copy number, red indicating below normal copy number, and grey indicating close to normal copy number. Blue lines indicate the median intensity of each bin. Specified genes to be highlighted are annotated. (**B**) Copy number plot of common cancer genes in the example glioblastoma sample. An overview of the genomic loci of common cancer genes are shown in more detail. Copy number values are visualized as described in Figure 4A.

methylation arrays have been produced by the manufacturer including the 450k and EPIC arrays. Importantly, different batches of these designs can contain a variable set of probes. As a result, the GenePattern notebook requires all query samples to be of the same array design. Similarly, all control samples have to be of the same array design, which can be different from the query samples. If the query samples and the control samples are of different array designs, only the common set of probes between the array designs are evaluated for the CNV analysis. As described above, the choice of control samples is crucial for the resulting copy number profiles. The control samples should be free of CNVs and have a similar methylation profile as the samples of interest. Provided that query and corresponding control samples are available, the *MethylationCNVAnalysis* notebook in the GenePattern Notebook Repository allows the CNV analysis to be performed without the installation of software and without the need to write or manipulate code.

Data availability

The notebook includes links to the data for running the use case described above. The raw data can be found in GEO Series GSE90496: https://identifiers.org/geo/GSE90496.

Software availability

GenePattern Notebook is available from: http://genepattern-notebook.org/.

A public preview of the notebook is available from: https://notebook.genepattern.org/services/sharing/notebooks/136/preview/

GenePattern Notebook source code is available from: https://github.com/genepattern/methylation_cnv_analysis_notebook.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.1419319 (Mah, 2018).

License: BSD 3-Clause.

Grant information

This study was funded by the National institutes of Health (grant numbers U24CA194107, U01CA184898, U41HG007517 and R01CA109467 to J.P.M.).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors thank members of the Mesirov Lab, including Edwin Juarez, Ted Liefeld and the GenePattern Development team for assisting with software development. The authors also thank Briana Prager for testing the notebook on independent datasets and Stephen Mack for providing additional datasets for testing. We thank Konstantin Okonechnikov and Volker Hovestadt for providing feedback on the analysis.

References

Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al.*: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics.* 2014; **30**(10): 1363–1369.
PubMed Abstract | Publisher Full Text | Free Full Text

Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature.* 2008; **455**(7216): 1061–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Capper D, Jones DTW, Sill M, *et al.*: **DNA methylation-based classification of central nervous system tumours.** *Nature.* 2018; **555**(7697): 469–474.
PubMed Abstract | Publisher Full Text | Free Full Text

Durinck S, Moreau Y, Kasprzyk A, *et al.*: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics.* 2005; **21**(16): 3439–3440.
PubMed Abstract | Publisher Full Text

Durinck S, Spellman PT, Birney E, *et al.*: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc.* 2009; **4**(8): 1184–1191.
PubMed Abstract | Publisher Full Text | Free Full Text

Feber A, Guilhamon P, Lechner M, *et al.*: **Using high-density DNA methylation arrays to profile copy number alterations.** *Genome Biol.* 2014; **15**(2): R30.
PubMed Abstract | Publisher Full Text | Free Full Text

Hovestadt V, Zapatka M: **conumee: Enhanced copy-number variation analysis using Illumina 450k methylation arrays.** *R package version 0.99, 4.* 2015.
Reference Source

Mah C: **genepattern/methylation_cnv_analysis_notebook v1.0.1 (Version v1.0.1).** *Zenodo.* 2018;
http://www.doi.org/10.5281/zenodo.1419327

Reich M, Liefeld T, Gould J, *et al.*: **GenePattern 2.0.** *Nat Genet.* 2006; **38**(5): 500–1.
PubMed Abstract | Publisher Full Text

Reich M, Tabor T, Liefeld T, *et al.*: **The GenePattern Notebook Environment.** *Cell Syst.* 2017; **5**(2): 149–151.e1.
PubMed Abstract | Publisher Full Text | Free Full Text

Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Seshan VE, Olshen A: **DNAcopy: DNA copy number data analysis.** *R package version 1.54.0.* 2018.
Reference Source

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 02 May 2019

https://doi.org/10.5256/f1000research.17848.r47252

✓ **Aris Floratos**
Department of Systems Biology, Columbia University, New York, NY, USA

The article describes a GenePattern notebook for inferring CNVs using methylation profiling data generated by Illumina Infinium arrays. The notebook leverages the well-established Bioconductor packages minfi and conumee to implement an analysis workflow that comprises quality control, CNV calling, and results visualization. Functionality is made available through a web browser interface and requires no software installation/configuration, making it an attractive option for users with limited informatics expertise.

Some thoughts about possible improvements:

1. Given that the entire workflow can be somewhat time consuming (the notebook documentation indicates that processing a single sample takes about 2 minutes), it would be useful if there was an option to run the QC step as a stand-alone computation, not combined with the CNV calling. As things stand right now, a significant amount of time can be spent waiting for the analysis to complete, only to realize that some control samples are of low quality and, thus, need to be removed and the analysis be rerun.
2. When running the notebook without specifying values in the “genes to highlight” or “ignore regions” parameter boxes, the run fails with “getopt” error messages. It is not clear why these parameters are mandatory (e.g., in the case of "genes to highlight", it is conceivable that one may want to only inspect genome-wide patterns of aberration, without focusing on specific genes); but if so, it would be helpful to state clearly in the documentation section.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 April 2019

https://doi.org/10.5256/f1000research.17848.r45332

✓

Robert Ivanek
Department of Biomedicine, University of Basel, Basel, Switzerland

The article by CK Mah et al. describes a new GenePattern Notebook "MethylationCNVAnalysis" which allows users of GenePattern platform to run copy number analysis (CNV) of Illumina Infinium DNA methylation arrays. The tool provides graphical user interface to the CNV analysis based on Bioconductor packages `minfi` and `conumee`. Such analysis can be with this tool performed even without programming experience or without functional R installation.

Compared to analysis on the R command line, the user is asked to set only few parameters required for the analysis: set of test and control samples, gene list for detailed view, a black-list of regions excluded from the analysis and a coice to in-/exclude sex chromosomes. In my opinion it would be helpful to create a section with "Advanced settings" and allow also specification of few other parameters: `bin_minprobes` and `bin_minsize` for function `CNV.create_anno` from `conumee` package and parameters for segmentation of log2ratio (functions from `DNAcopy` package). In both cases the authors of `conumee` package warn, that they optimized parameters for 450k arrays and another array type might require further optimization (see http://bioconductor.org/packages/release/bioc/vignettes/conumee/inst/doc/conumee.html). Also in respect to general need for higher reproducibility of analyses, the tool should export short summary of software, their versions and parameters used for analysis.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.