Genes & Diseases (2019) 6, 43–46

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://ees.elsevier.com/gendis/default.asp

SHORT COMMUNICATION

AluScanCNV2: An R package for copy number variation calling and cancer risk prediction with next-generation sequencing data

Taobo Hu [a], Si Chen [a], Ata Ullah [a], Hong Xue [a,b,*]

[a] *Division of Life Science, Applied Genomics Centre and Centre for Statistical Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong*
[b] *School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, China*

Received 6 June 2018; accepted 4 September 2018
Available online 8 September 2018

KEYWORDS
AluScan;
Bioinformatics;
Cancer subtyping;
Cancer predisposition;
Machine learning

Abstract The usage of next-generation sequencing (NGS) to detect copy number variation (CNV) is widely accepted in cancer research. Based on an AluScanCNV software developed by us previously, an AluScanCNV2 software has been developed in the present study as an R package that performs CNV detection from NGS data obtained through AluScan, whole-genome sequencing or other targeted NGS platforms. Its applications would include the expedited usage of somatic CNVs for cancer subtyping, and usage of recurrent germline CNVs to perform machine learning-assisted prediction of a test subject's susceptibility to cancer.

Introduction

A major contribution to the genome variability among individuals arises from CNVs.[1,2] NGS studies produce a vast amount of raw data that could be employed to detect CNV. Read depth-based methods such as CNV-seq[3] and AluScanCNV[4] are available for calling CNVs from NGS data, and advances in understanding the role of CNVs in tumor development facilitates the prevention and treatment of tumors. In this regard, germline CNVs in white blood cell DNA have been utilized by us to successfully predict the likelihood of tumor occurrence with the assistance of machine learning.[5]

Earlier, we have developed an AluScanCNV software for calling CNVs, which comprises a collection of independent R/Perl code files.[4] In the present study, an R package designated AluScanCNV2 has been devised to implement both the CNV calling from NGS data and recurrent germline CNV-based cancer risk prediction[5] tasks, which can be incorporated into bioinformatics pipelines to expedite the search for CNV-cancer associations.

* Corresponding author. Division of Life Science, Applied Genomics Centre and Centre for Statistical Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.
E-mail address: hxue@ust.hk (H. Xue).
Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2018.09.001

Methods

Implementation

AluScanCNV2 relies on Geary-Hinkley transformation (GHT)-based comparison of the read-depth of a sequence window on the test sample with that on either a paired control sample in the case of 'paired CNV' analysis, or a reference template constructed from pooled reference samples in the case of 'unpaired CNV' analysis.[4] Functions pairedCNV(…) and unpairedCNV(…) are designed to call CNVs from paired and unpaired samples, respectively. Through correlation-based machine learning, somatic CNVs can be identified to expedite cancer classification.[4]

Operation

The AluScanCNV2 package is cross-platform effective (Windows, macOS and Linux) without any specific computer hardware requirements. Installation instructions and a list of prerequisites are provided on the package web page (https://github.com/hutaobo/AluScanCNV2).

Results

Unpaired CNV calling

For samples NGS-sequenced employing the AluScan platform,[7–9] unpaired CNV analysis is performed by comparing the test sample to a previously generated reference template[4] named 'AluScan.ref.5k.reads' in the AluScanCNV2 package. The users need to call function unpairedCNV(sample.5k.doc, window.size, seq.method, custom.ref, …); the optional parameters within function are provided on the AluScanCNV2 web page. The codes below show unpaired CNV calling of sample data sequenced by AluScan.

```
> library(AluScanCNV2)
> sample_doc_path <- system.file("extdata/Breast1_b.5k.doc", package =
"AluScanCNV2")
> unpairedCNV(sample.5k.doc = sample_doc_path, window.size = "500k",
seq.method = "AluScan", output.path = "./")
```

For CNV-based cancer risk prediction, the unpaired CNVs from the germline genomes of a group of subjects are aggregated to generate a dataset. To apply machine learning to the generated dataset, function featureSelection (…) is designed to select informative recurrent CNV features. Then, the train (…) function in the 'caret' package[6] is employed to build predictive models with 1000 iterations of two-fold cross validation based on the informative recurrent CNVs. Based on the resultant predictive models, function cancerPrediction (…) is employed to test the predictive models.

For the samples NGS-sequenced using the WGS platform, the corresponding reference template named 'WGS.ref.5k.reads' is performed in the unpaired CNV analysis by setting "seq.method = 'WGS'" in the above codes. The 'WGS.ref.5k.reads' reference template is generated from 105 pooled reference samples from various ethnic origins in the 1000 Genomes Project.[10] For CNV calling of samples sequenced by other targeted NGS platforms, users can also generate their own reference template using function doc2data (doc.list, write.file.path). However, CNV analysis of samples obtained from non-human species should be performed using specific methodology such as the copy-number analysis pipeline designed for microbiome studies.[11]

Paired CNV calling

Paired CNV analysis is performed by comparing the test sample to the control sample, by calling function pairedCNV(control.5k.doc, sample.5k.doc, window.size, …):

```
> library(AluScanCNV2)
> control_doc_path <- system.file("extdata/Breast1_b.5k.doc", package =
"AluScanCNV2")
> tumor_doc_path <- system.file("extdata/Breast1_1.5k.doc", package =
"AluScanCNV2")
> pairedCNV(control.5k.doc = control_doc_path, sample.5k.doc =
tumor_doc_path, window.size = "500k", output.path = "./")
```

The function pairedCNV() can also be applied directly to paired CNV analysis of samples sequenced by other NGS platforms.

Identification of recurrent CNVs

Function seg2CNV(seg.list, ...) is designed to aggregate unpaired CNVs from the germline genomes of a group of subjects into a training dataset. To identify informative recurrent CNVs, the users need to call function featureSelection2 (nonCancerListA, CancerListA, nonCancerListB, CancerListB, ...):

```
> alu_control <- list.files(path = 'path_to_folder_containing_seg_files',
full.names = TRUE)
> library(AluScanCNV)
> alu_control <- seg2CNV(alu_control)
> alu_control$recurrence <- alu_control$recurrence / (ncol(alu_control) - 4)
> recurr_cnv <- featureSelection2(nonCancerListA = alu_control, CancerListA
= alu_cancer, nonCancerListB = wgs_control, CancerListB = wgs_cancer, Cri =
0.33)
```

Prediction of susceptibility to cancer

After the informative CNV features are selected from the training dataset, function train (data, method, ...) in the caret package is employed to build models based on the selected CNVs. The resultant model is used in function cancerPrediction (file_path, model) to predict the CNV-based cancer-susceptibility:

```
> library(caret)
> ## Building Random Forest Model
> # fit <- train(type ~ ., data = dataset, method = "rf", metric = "Accuracy",
trControl = trainControl(method = "cv"))
> ## Prediction of susceptibility to cancer
> cancerPrediction(file_path = "./Breast1_b.local.500k.unpaired.seg", model
= fit)
```

Validation of the prediction model

For description of how consistent model probabilities are with observed event rates, function calPlot (...) is employed to create the 'calibration plot' previously described in the caret package.

```
> library(caret)
> library(ggplot2)
> # Calibration of the observed probability vs. prediction probability
> p <- calPlot(model, data, class)
> p
```

Conclusion

The AluScanCNV2 package comprises two major parts: CNV calling and CNV-based cancer risk prediction (Fig. 1). The CNV calling part is described in the previous AluScanCNV, but it is optimized and simplified in AluScanCNV2. The integration of the CNV calling and CNV-based cancer risk prediction tasks into AluScanCNV2 facilitates its incorporation into a bioinformatics pipeline to streamline analysis with reduction of analysis time.

Employing AluScanCNV2, users can complete the entire process, starting from a raw sequence file, for calling CNVs and predicting a subject's susceptibility to cancer based on the called germline CNVs. The CNVs identified may facilitate the uncovering of the underlying mechanisms in cancer genomics.

Software availability

Tool and source code are available from: https://github.com/hutaobo/AluScanCNV2.

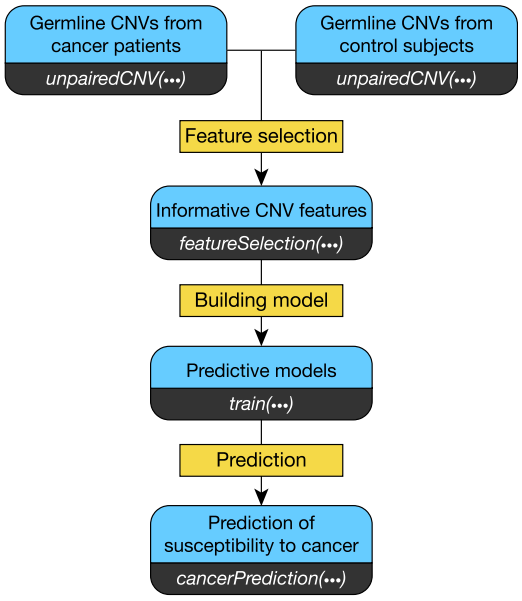

Figure 1 Schematic flow-chart of the use of AluScanCNV2 for cancer prediction. Individual functions are represented by rectangles with rounded corners divided into an upper part listing the descriptions and a lower part containing function names. The contents in the yellow background boxes are the main steps.

Archived source at time of publication: https://doi.org/10.5281/zenodo.1419652.

License: GPL-3.

Author contributions

HX conceived and initiated the study; TH, SC and AU developed the package; and HX and TH wrote the paper.

Conflict of interest

None declared.

Acknowledgements

The study was supported by grants to H. Xue from University Grants Committee (VPRDO09/10.SC08), and Innovation and Technology Fund (ITS/113/15FP) of Hong Kong SAR.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.gendis.2018.09.001.

References

1. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet*. 2015;16(3):172—183.
2. Nagao Y. Copy number variations play important roles in heredity of common diseases: a novel method to calculate heritability of a polymorphism. *Sci Rep*. 2015;5:17156. https://doi.org/10.1038/srep17156.
3. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinf*. 2009;10:80. https://doi.org/10.1186/1471-2105-10-80.
4. Yang JF, Ding XF, Chen L, et al. Copy number variation analysis based on AluScan sequences. *J Clin Bioinf*. 2014;4(1):15.
5. Ding X, Tsang SY, Ng SK, Xue H. Application of machine learning to development of copy number variation-based prediction of cancer risk. *Genomics Insights*. 2014;7(Supplementary Files 15002):1—11.
6. Kuhn M. *Caret: Classification and Regression Training*. Astrophysics Source Code Library; 2015.
7. Mei L, Ding X, Tsang SY, et al. AluScan: a method for genome-wide scanning of sequence and structure variations in the human genome. *BMC Genom*. 2011;12:564. https://doi.org/10.1186/1471-2164-12-564.
8. Kumar Y, Yang J, Hu T, et al. Massive interstitial copy-neutral loss-of-heterozygosity as evidence for cancer being a disease of the DNA-damage response. *Bmc Med Genomics*. 2015;8:42. https://doi.org/10.1186/s12920-015-0104-2.
9. Hu T, Kumar Y, Shazia I, et al. Forward-reverse mutations in stages of cancer development. *bioRxiv*. 2017:198309. https://doi.org/10.1101/198309.
10. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68—74.
11. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell*. 2015;160(4):583—594.