GRiG: A PPV-sensitive method for predicting somatic SNVs from cancer-normal paired sequencing data with greedy rule induction algorithm

Shaoping Ling[1]*, Lili Dong [1]*, Lihua Cao[1], Caiyan Jia[2,3], Xuemei Lu[1,§], Chung-I Wu[1,4,§]

[1]Beijing Institute of Genomics, Chinese Academy of Scineces, Beitucheng West 7#, Beijing, China

[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

[3]Department of Bioengineering/Bioinformatics, University of Illinois at Chicago, Chicago, IL 60612, USA

[4]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA.

*These authors contributed equally to this work
§Corresponding authors

Email addresses:
SL: spling@big.ac.cn
LD: donglili@big.ac.cn
LC: caolh@big.ac.cn
CJ: caiyan.jia@gmail.com
XL: luxm@big.ac.cn
CW: wuci@big.ac.cn

Abstract

Predicting somatic SNVs from cancer-normal paired sequencing is a key computational issue in high-throughput sequencing-driven cancer genomics. Classic methods based on statistical inference (SI) have been developed and become standard pipeline in human variation detection. However, they can not provide enough high positive prediction value (PPV) for further experimental validation and function analysis. We presented a Greedy Rule Induction alGorithm (GRiG) for predicting somatic SNVs in cancer-normal paired sequencing data, which integrates feature selection and rule inference into a machine learning frame work. We evaluated the performance of GRiG on public datasets which consist of two candidate somatic SNVs datasets from 48 breast exome capture sequencing (ECS) dataset and four whole genome sequencing (WGS) dataset for training and testing respectively. GRiG always achieved the better performance in ECS training dataset with 10x cross-validation and WGS testing dataset than both Samtools and GATK and presented comparable performance with four statistical learning algorithms including random forest, Bayesian additive regression tree, support vector machine and logistic regression in ECS training dataset with 10xcross-validation and WGS testing dataset. Especially, it always achieved better PPV than these four classifiers.

Keywords: Somatic SNV, Greedy algorithm, Feature selection, Machine learning, PPV