

ARVIN : Identifying Risk Noncoding Variants Using Disease-relevant Gene Regulatory Networks

Long Gao, Yasin Uzun, Kai Tan

October 31, 2017

Contents

1	Introcution	1
2	Network construction	2
2.1	Enhancer prediction	2
2.2	Enhancer-promoter interaction prediction	2
2.3	Obtain gene-gene interation network	2
2.3	Network scoring	2
3	Prepare features for risk variants prediction	2
3.1	Network-based features	2
3.1.1	Betweenness centrality	2
3.1.2	Closeness centrality	2
3.1.3	Pagerank centrality	2
3.1.4	Weighted degree	2
3.1.5	Module score	2
3.2	GWAVA features	2
3.3	FunSeq features	2
4	Build a classifier for prioritizing risk varints	2
4.1	Train a random forest classifier	2
4.2	Predict causal disease variants	2

1 Introcution

Identifying causalnoncoding variants remains a daunting task. Because noncoding variants exert their effects in the context of a gene regulatory network (GRN), we hypothesize that explicit use of disease-relevant GRN can significantly improve the inference accuracy of noncoding risk variants. We describe Annotation of Regulatory Variants using Integrated Networks (ARVIN), a general computational framework for predicting causal noncoding variants. For each disease, ARVIN first constructs a GRN using multi-dimensional omics data on cell/tissue-type relevant to the disease. ARVIN then uses a set of novel regulatory network-based features, combined with sequence-based features to make predictions. Using known causal variants in gene promoters and enhancers in a number of diseases, we show ARVIN outperforms state-of-the-art methods that use sequence-based features alone.

2 Network construction

2.1 Enhancer prediction

2.2 Enhancer-promoter interaction prediction

2.3 Obtain gene-gene interaction network

2.3 Network scoring

3 Prepare features for risk variants prediction

3.1 Network-based features

3.1.1 Betweenness centrality

3.1.2 Closeness centrality

3.1.3 Pagerank centrality

3.1.4 Weighted degree

3.1.5 Module score

3.2 GWAVA features

3.3 FunSeq features

4 Build a classifier for prioritizing risk variants

4.1 Train a random forest classifier

4.2 Predict causal disease variants