InCRIMP: a versatile computational model for the integrative analysis of multi-omics data

Abbas Salavaty (@mania_abbas) and Mark Pinese
Children's Cancer Institute, Lowy Cancer Centre, UNSW Sydney, Kensington, NSW, Australia

The code will be deployed on the GitHub: github.com/asalavaty

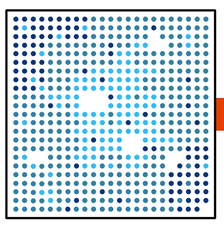
Introduction

Despite recent advancements in precision medicine, for most patients a targeted treatment cannot be identified. High-throughput studies have aimed to address our imperfect understanding of cancer biology through unbiased discovery of cancer risk and driver genes based on single omics profiles. As genes work in concert to drive cancer, we hypothesise that an integrative approach that considers multiple molecular data, in the context of multi-gene pathways, will yield the best understanding of cancer biology. Here we present InCRIMP (**In**tegratomic **C**ancer **R**isk **I**nfluential **M**odule **P**rioritization) which integrates multiple molecular measurements and state-of-the-art network analysis to achieve comprehensive molecular dissection of cancer cohorts, and unlock the true potential of molecular profiling to understand the risk genes and drivers of cancer.

Methods

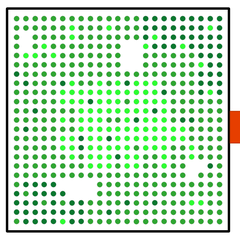
Pre-processing, filtration, and normalization of raw data. SNV data were filtered and a deleteriousness score was assigned to each SNV based on CADD scores. The transcriptomic data was TPM-normalized after filtering out the noise genes. The PPI data was obtained from the STRING database.

Germline and Somatic SNV Data



Removing low depth variants → Removing intergenic and mitochondrial variants → gnomAD-based filtration of highly frequent variants → CADD-based deleteriousness analysis → Ensembl-based imputation of missing CADD scores → GAM-based calibration of CADD scores

Cancer and Normal Expression Data



Removing noise genes → TPM normalization

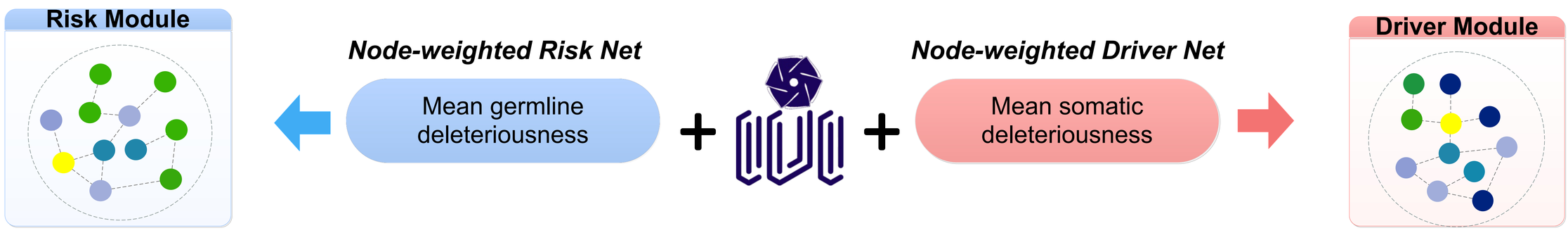
PPIs

STRING

Association Analysis and Un-weighted Multi-layer Network Reconstruction. Two multi-layer networks were reconstructed; a three-layer multi-omics Risk network and a a three-layer multi-omics driver network.

*** Risk Net ***	*** Driver Net ***
01 Germline Gene co-Deleteriousness	01 Somatic Gene co-Deleteriousness
02 Normal Gene co-Expression	02 Cancer Gene co-Expression
03 Protein-Protein Interaction (PPI)	03 Protein-Protein Interaction (PPI)

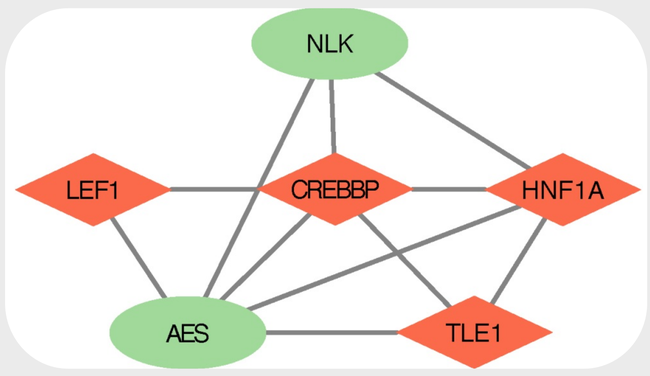
Functional module identification. Initially primitive risk and driver scores were calculated and assigned as the node weights. Then, functional modules were identified based on the Leiden algorithm. Lastly, final node scores were calculated by integrating the primitive scores and node mean neighborhood scores.



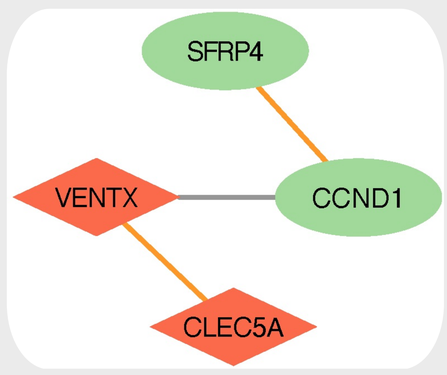
Results

An example based on HGG

First-ranked Driver Module of HGG



First-ranked Risk Module of HGG



Top 20 Candidate Drivers of HGG

Gene	Final Driver Score	Gene	Final Driver Score
TP53	1083.8597	NFKB1	513.1011
BYSL	721.0372	PNO1	503.1376
WDR12	672.4505	MPHOSPH10	490.5041
RBM28	633.7224	DDX49	485.9314
PIK3CA	613.4044	KRR1	464.0030
RPS16	605.9326	RRP9	460.6994
EP300	594.3892	NFKBIA	442.9822
NOP58	580.3518	UTP18	442.0950
UTP20	519.7518	CREBBP	439.3800
UTP6	517.2224	FBL	438.5238

Conclusion

InCRIMP has integrated multiple molecular data types in cancer to recapitulate known cancer biology, and drive the discovery of new cancer driver and risk gene networks and modules.