Context-specific Essential Genes Identification and Prediction by Learning Multi-Omics and Network Data

Maurizio Giordano[1], Lucia Maddalena[1], Mario Manzo[2], Mario Rosario Guarracino[3], Ilaria Granata[1]

1 Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy
2 Information Technology Services, University of Naples "L'Orientale", Naples, Italy
3 Department of Economics and Law, University of Cassino and Southern Lazio, Cassino, Frosinone, Italy

Introduction

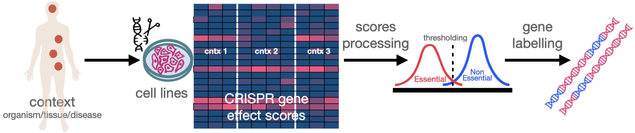
Essential genes (EGs) are generally defined as necessary genes for the growth and survival of any organism or cell.

- Some genes are essential only in specific contexts, represented by environmental and/or genetic conditions.
- The identification and contextualization of essential genes is particularly relevant in genetics, and it has many implications in system biology and precision medicine.
- CRISPR Cas9 gene silencing is the state-of-the-art technique for measuring gene essentiality in in-vitro cell lines.
- We propose the Human Gene Essentiality Labelling & Prediction (HELP) framework: a library of tools and methodologies for identification and prediction of common EGs and context-specific EGs.

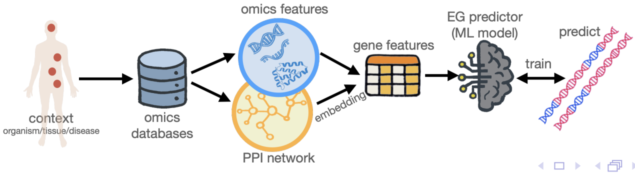
,

Proposal

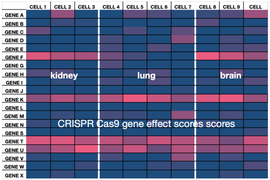
- A new identification method of context-specific EGs based on the processing of CRISPR Cas9 gene effect scores.
  - ... few identification methods based on CRISPR scores capture context-specificity of EGs (cell/tissue/disease).



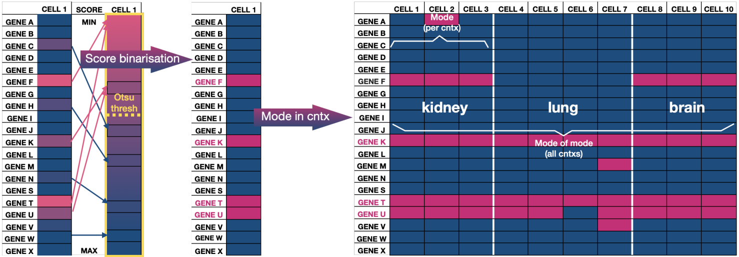
- A new ML prediction method of context-specific EGs exploiting gene multi-omics and PPI-network information.
  - ... CRISPR experiments are costly, time-intensive, and limited to in vitro models availability... prediction models compensate for these limitations.



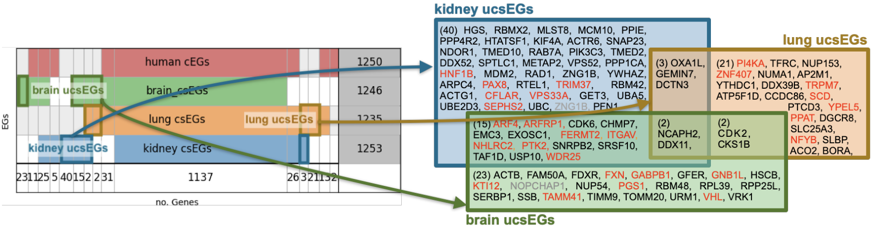
The EG identifier



- EG identification uses as input data the CRISPR Cas9 gene effect scores
- cell-specific labels: scores are binarized using a per-cell Otsu thresholding [1].
- cntx-specific labels: modes of cell-specific labels in same cntx
- organism-wide labels: modes of cntx-specific labels.



Tissue-specific EGs



(40) HGS, RBMX2, MLST8, MCM10, PPIE, PPP4R2, HTATSF1, KIF4A, ACTR6, SNAP23, NDOR1, TMED10, RAB7A, PIK3C3, TMED2, DDX52, SPTLC1, METAP2, VPS52, PPP1CA, HNF1B, MDM2, RAD1, ZNG1B, YWHAZ, ARPC4, PAX8, RTEL1, TRIM37, RBM42, ACTG1, CFLAR, VPS33A, GET3, UBA5, UBE2D3, SEPHS2, UBC, ZNG1B, PFN1

(3) OXA1L, GEMIN7, DCTN3

(21) PI4KA, TFRC, NUP153, ZNF407, NUMA1, AP2M1, YTHDC1, DDX39B, TRPM7, ATP5F1D, CCDC86, SCD, PTCD3, YPEL5, PPAT, DGCR8, SLC25A3, NFYB, SLBP, ACO2, BORA,

(15) ARF4, ARFRP1, CDK6, CHMP7, EMC3, EXOSC1, FERMT2, ITGAV, NHLRC2, PTK2, SNRPB2, SRSF10, TAF1D, USP10, WDR25

(2) NCAPH2, DDX11,

(2) CDK2, CKS1B

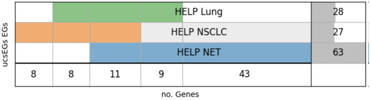
(23) ACTB, FAM50A, FDXR, FXN, GABPB1, GFER, GNB1L, HSCB, KTI12, NOPCHAP1, NUP54, PGS1, RBM48, RPL39, RPP25L, SERBP1, SSB, TAMM41, TIMM9, TOMM20, URM1, VHL, VRK1

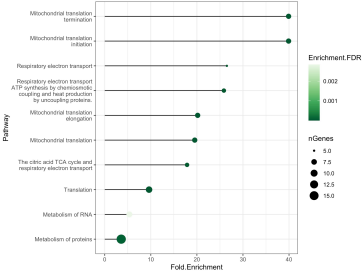
- context-specific EGs (csEGs) are genes annotated as “essential” by HELP EG identifier in a specific context (tissue/disease)
- common EGs (cEGs) are csEGs shared with the organism-wide EGs (all cell lines), while uncommon csEGs (ucsEGs) are those not shared.
- very uncommon csEGs (vucsEGs) are context-specific EGs not shared with any other context (including the organism-wide cntx).

tissue	kidney	lung	brain
TPR_{ucsEG}	80%	75%	67%
TPR_{csEG}	90%	91%	90%

Disease-specific EGs

- We identified context-specific EGs in Lung Neuroendocrine Tumor (NET) and Non-Small-Cell Lung Cancer (NSCLC).
- NET-specific EGs enrich pathways related to cellular respiration and energy production.
- NSCLC-specific EGs are significantly differentially expressed when comparing the two NSCLC subtypes (LUAD and LUSC) wrt normal samples.

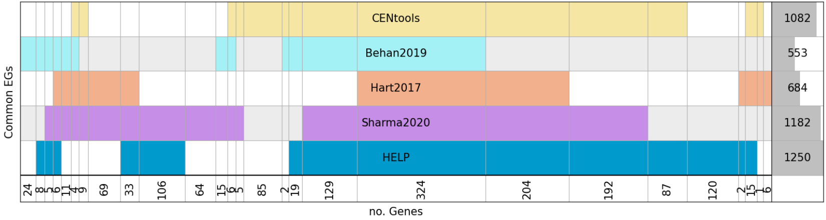




Comparison with SOTA common EGs

Overlaps of cEGs annotated by HELP with state-of-the-art common EGs (CEN-tools [2], Behan2019 [3], Hart2017 [4], and Sharma2020 [2])

- 324 shared genes.
- large coverages of HELP with more recent and less stringent cEGs annotations (CEN-tools & Sharma2020) (970/999 genes).

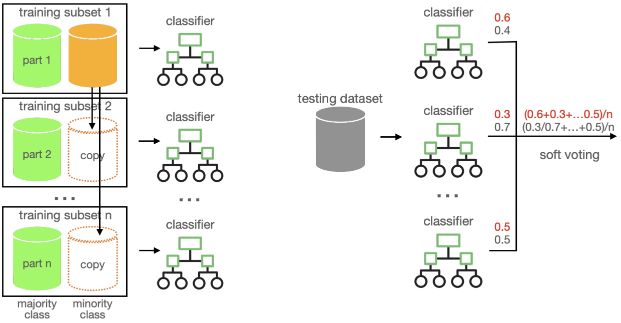


The EG predictor: gene features

- Bio : mixed attributes:
  - Structural (gene length, GC content, Transcript count)
  - Expression (GTEX_cntx, UP_Tissue, OncoDB_expr, HPA_cntx))
  - Function/Localization (GO-MF, GO-MF, GO-MF, KEGG, REACTOME)
  - Interaction (BIOGRID, UCSC_TFBS)
  - Association with disease (Driver_genes_MUT, Driver_genes_CNV, Driver_genes_MET, Gene-Disease association)
  - Conservation (Orthologs count)
- CCcfs : subcellular localisation confidence scores (COMPARTMENT[5]).
- N2V : 128 attributes (node embeddings) computed by Node2Vec [6] on Tissue-specific PPIs from Integrated Interaction Database [7].

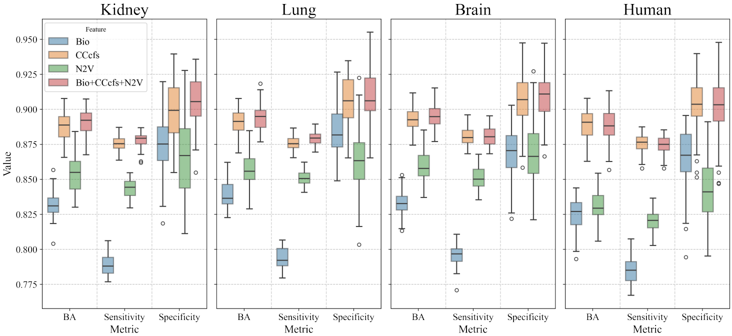
The EG predictor: the model

- We developed a new machine learning method: the Splitting Voting Ensemble learner (SVELearn) [8] to cope with unbalanced data:
  - splitting the training dataset into *n* balanced sub-datasets based on the unbalancing ratio of essential to non-essential genes
  - combining predictions (soft-voting) of a set of *n* learners trained on the balanced sub-datasets.
  - Light Gradient-Boosting Machine [9] is the base learner of the ensemble



Results: prediction varying gene attributes

- csEGs prediction performance trend with different gene sets for training (on kidney, lung, brain and organism-wide contexts).



- BA is the mean of Specificity and Sensibility
- CCcfs and Bio attributes provide highest and lowest BA
- best BA by combining all features (Bio+CCcfs+N2V).

Conclusions

- HELP is a library of tools and methodologies to address the identification and prediction of common EGs and context-specific EGs.
- We demonstrated HELP's effectivity in an organism-wide context, three human tissues and two types of lung cancer.
- EG identification & prediction of HELP validated by comparison with state-of-the-art methods:
    - identification methods: CoRe ADaM/FiPer [10], OGEE[11]
    - gold standard annotation: Sharma2020, CEN-tools [2]
    - prediction methods: DeepHE [12], CLEARER[13], EPGAT[14]
- Ongoing works:
    - exploring intermediate classes of essentiality
    - extend investigation on more diseases

Thank you ... Q & A

Citation

HELP: a computational framework for labelling and predicting human common and context-specific essential genes I. Granata, L. Maddalena, M. Manzo, M. R. Guarracino, and M. Giordano. *PLOS Computational Biology, 20(9) - September 2024*, `https://doi.org/10.1371/journal.pcbi.1012076`

HELP software

SVELearn software

References I

[1] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66, 1979.

[2] Sumana Sharma, Cansu Dincer, Paula Weidemüller, et al. CEN-tools: an integrative platform to identify the contexts of essential genes. *Mol. Syst. Biol.*, 16(10):e9698, 2020.

[3] Fiona M Behan, Francesco Iorio, Gabriele Picco, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature*, 568(7753):511–516, 2019.

[4] Traver Hart, Amy Hin Yan Tong, Katie Chan, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Genes—Genomes—Genetics*, 7(8):2719–2727, 2017.

[5] Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, et al. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

[6] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *KDD '16*, page 855–864, 2016.

References II

[7] Max Kotlyar, Chiara Pastrello, Nicholas Sheahan, and Igor Jurisica. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.*, 44(D1):D536–D541, 2016.

[8] Maurizio Giordano, Emanuele Falbo, Lucia Maddalena, Marina Piccirillo, and Ilaria Granata. Untangling the context-specificity of essential genes by means of machine learning: A constructive experience. *Biomolecules*, 14(1), 2024.

[9] Guolin Ke, Qi Meng, Thomas Finley, et al. LightGBM: A highly efficient gradient boosting decision tree. In *Proc. NIPS'17*, page 3149–3157, 2017.

[10] Alessandro Vinceti, Emre Karakoc, et al. CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. *BMC Genom.*, 22(828), 2021.

[11] Wei-Hua Chen, Guanting Lu, Xiao Chen, Xing-Ming Zhao, and Peer Bork. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.*, 45(D1):D940–D944, 10 2016.

[12] Xue Zhang, Wangxin Xiao, and Weijia Xiao. DeepHE: Accurately predicting human essential genes based on deep learning. *PLoS Comput Biol*, 16(9):e1008229, 2020.

References III

[13] Thomas Beder, Olufemi Aromolaran, Jürgen Dönitz, et al. Identifying essential genes across eukaryotes by machine learning. *NAR Genom Bioinform*, 3(4):lqab110, 2021.

[14] João Schapke et al. EPGAT: Gene essentiality prediction with graph attention networks. *IEEE/ACM Trans Comput Biol Bioinform*, 19(3):1615–1626, 2022.