Identifying and characterizing essential genes from CRISPR knockout screens

Yamini Ananth[1], Traver Hart[2]

1. Columbia University Applied Math/Computer Science, 2. University of Texas MD Anderson Cancer Biology

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

Making Cancer History®

Background

Genome-wide loss-of-function screens offer a data source for identifying core essential genes, which are required for the survival of an organism. Identifying and characterizing human essential genes is a critical step for functional genomics and cancer target-finding (1).

Identifying Essential Genes

CRISPR knockout screens for 808 mammalian cell lines across 18,111 genes were filtered for quality using Bayes factors and Cohen's D Statistic. An essentiality score was assigned to each gene based on how many cell lines in which a gene was essential.

The distribution of gene's essentiality scores (Figure 1) shows a large jump from contextual to core essential genes on the right side, suggesting a group of genes are more likely to always appear than 'almost always' appear. **11,413 never, 5,991 contextual, and 717 core essentials were identified.**

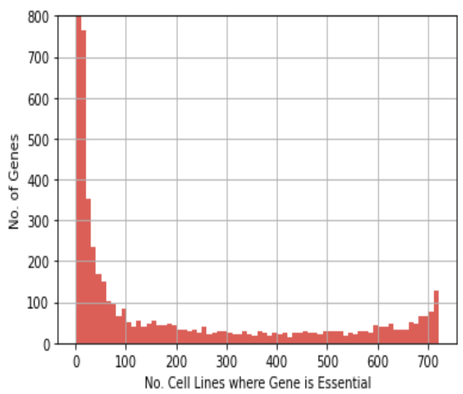

Figure 1) 73 evenly spaced bins; genes binned by the number of cell lines in which they were essential (Bayes factor >5)

Essentiality % was considered the number of cell lines in which a gene was essential divided by the total cell lines that passed filtering (727).

Gene Energetic Costs

A gene's energetic cost is the cost of biosynthesizing each amino acid it contains. Cancer cells notably reduce this cost per gene (2). Each gene's energetic cost was calculated using its UniprotKB canonical sequence and amino acid biosynthetic costs.

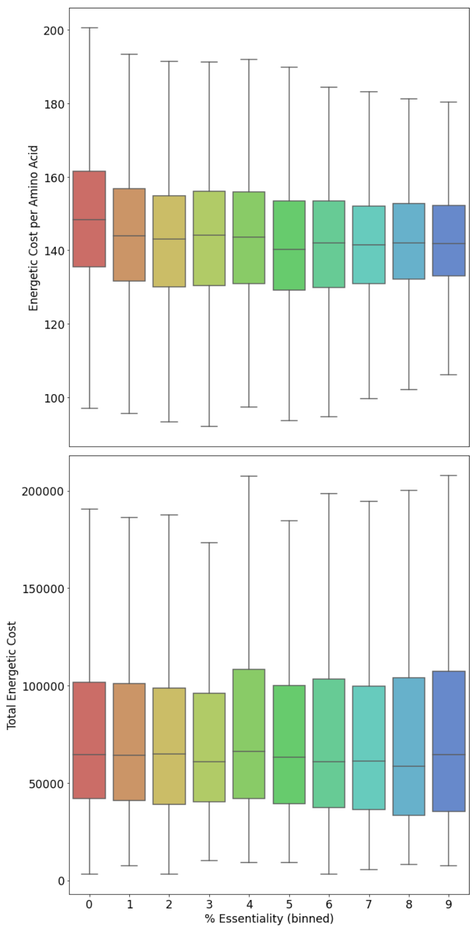

Figure 2) Distributions of total gene energetic cost (top) and cost per amino acid (bottom) for 18,111 genes binned by essentiality %, with all never-essential genes in bin 0 and core essentials in bin 9.

Naively, essential genes might have lower energetic-however, **no relationship was observed between gene essentiality and energetic cost** (Figure 2).

This suggests cells are energetically efficient enough not to have energetic cost constraints on essentiality, or cells exist in a nutrient-rich environment where energetic costs do not impose selection pressure on individual genes.

Loss of Function Association

Genes may contain variants that are predicted to result in their loss of function (lof). Using the gnomAD dataset which predicts loss of function variants for 125,000+ exomes (3), core essential genes **are less likely to contain unexpectedly high numbers of lof variants than never essential genes** (Figure 3, top). Core essential genes are being selected against for lof variants. pLI, **gene tolerance to lof based on protein truncating variant numbers, increases with essentiality** as expected (bottom).

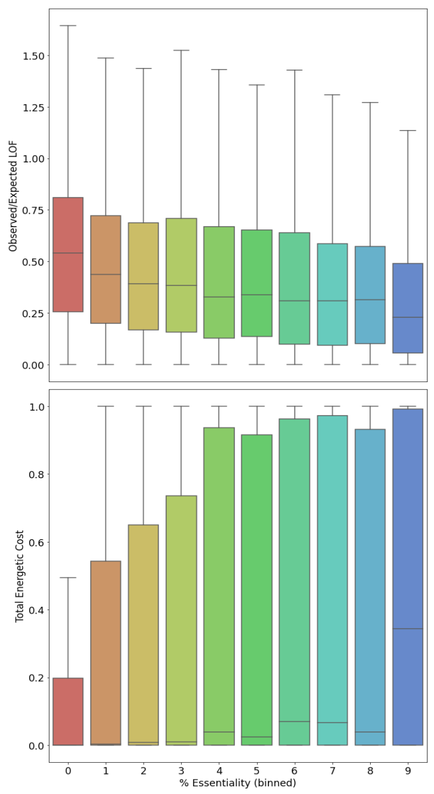

Figure 3) Fewer lof variants are observed than expected for core vs never essentials. Binomial test shows $p<0.05$ between bin 0-9 (top). pLI median increases sharply for core essentials compared to never and contextuals. $p<0.05$ for binomial tests between bin 0-9 and 8-9 (bottom).

Disease Association

Genes were analyzed for association with a disease using the OMIM Morbid Map dataset, which maps genetic variation with disease phenotypic expression (4).

Contextually essential genes are enriched for disease compared to core essential genes (Figure 4).

Previous literature supports the discovered relationship between contextually essential genes and association with disease, as contextually essentials are **more likely to show deleterious mutations** compared to core essentials.

CRISPR screens offer a more complete view of gene essentiality, adding robustness to these earlier findings.

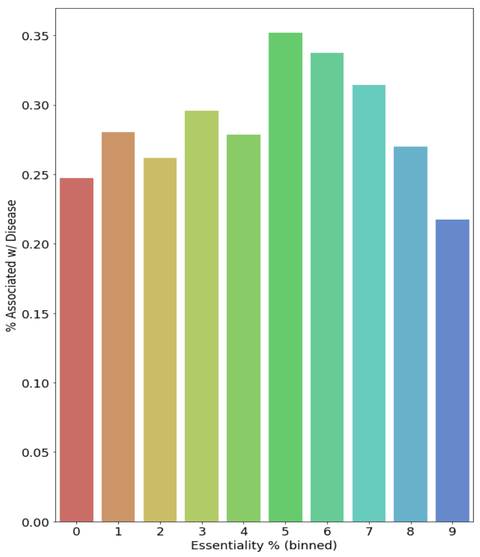

Figure 4) Core essentials (rightmost column) are 8% less associated with disease than contextually essential genes (Columns 6-8)

Phenotype Association

Genes were analyzed for association with a phenotype in the GWAS Catalog, which systematically connects genes with associated phenotypes (5). However, **no overall correlation was found between gene essentiality and phenotype expression** in the GWAS Catalog dataset, (Figure 5). Although essential genes are less associated with disease phenotypes, they are not less associated with any phenotype.

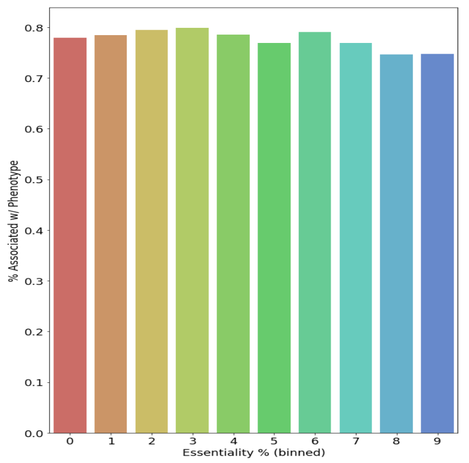

Figure 5) Approximately 74% of genes are related to at least 1 phenotype across **all** bins

Discussion

Identifying and better understanding essential genes is of broad interest to understand necessary cell functions and disease origin and mechanisms. In this exploratory analysis, gene essentiality shows no relationships with energetic costs or phenotypes in general but does relate with disease phenotypes and loss of function mutations.

References

(1) Hart T, et al. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014;10(7):733. Published 2014 Jul 1. doi:10.15252/msb.20145216

(2) Zhang, H., et al. Biosynthetic energy cost for amino acids decreases in cancer evolution. Nat Commun 9, 4124 (2018). https://doi.org/10.1038/s41467-018-06461-1

(3) Karczewski, K.J., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020)

(4) Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD)

(5) Buniello A, et al The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 2019, Vol. 47 (Database issue): D1005-D1012.