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 percentage 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.

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.

Naively, essential genes might have lower energetic costs per amino acid. However, this is clearly not the case as no relationship was observed between gene essentiality and energetic cost (Figure 2).

This suggests organisms are energetically efficient enough not to have energetic cost constraints on essentiality.

Loss of Function Association

Many genes 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 other 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.

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).

Previous literature supports the discovered relationship between peripherally essential genes and association with disease, as peripherally 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 5) Approximately 74% of genes are related to at least 1 phenotype across all bins

Phenotype Association

Genes were analyzed for association with a phenotype in the GWAS Catalog, which systemically 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.

Discussion

In this exploratory characterization/analysis, gene essentiality shows no relationships with energetic costs or phenotypes in general but does relate with disease phenotypes and loss of function mutations. Exploration of associations with essentiality were limited by quantity and quality of existing datasets linking gene/variants with phenotypes and loss of function.

References

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

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).