# Midbootcamp project

The Disease Specificity Index (DSI) is a metric used to quantify the specificity of a gene's expression pattern in relation to a particular disease or condition. It provides a numerical value that indicates how selectively a gene is expressed in a specific disease compared to its expression in other tissues or conditions. The DSI can be calculated using various approaches, but a commonly used method is the Tau statistic.

To calculate the DSI for a gene in a given disease dataset, follow these steps:

Obtain gene expression data: Gather gene expression data for the gene of interest across different tissues or conditions, including the disease dataset you are analyzing.

Normalize gene expression levels: Normalize the gene expression levels to account for variations in measurement techniques and scales. Common normalization methods include the TMM (trimmed mean of M values) or RPKM (reads per kilobase per million mapped reads) normalization.

Calculate gene expression specificity: Compute a measure of gene expression specificity for each tissue or condition. One commonly used measure is the Shannon entropy, which considers the relative expression levels of a gene in different tissues or conditions.

Calculate disease specificity: Calculate the specificity of gene expression in the disease dataset. This can be done by comparing the gene's expression levels in the disease dataset to its expression levels in other tissues or conditions. A higher expression level in the disease dataset relative to other tissues or conditions would indicate greater disease specificity.

Calculate the Disease Specificity Index: Finally, use the disease specificity value obtained in the previous step to calculate the Disease Specificity Index. The DSI is typically represented as a numerical value between 0 and 1, with higher values indicating greater disease specificity.

It's important to note that the specific implementation of the DSI calculation may vary depending on the methodology used and the characteristics of the gene expression dataset. Different studies may employ different normalization methods or employ additional statistical approaches to measure specificity. Therefore, it's essential to consult the literature or specific analysis tools for the precise calculation used in a given context.

The Disease Pleiotropy Index (DPI) is a metric used to quantify the pleiotropic nature of a gene, which refers to its involvement in multiple diseases or phenotypic traits. The DPI provides a numerical value that indicates the extent of a gene's association with diverse diseases or phenotypes.

To calculate the DPI for a gene, you can follow these steps:

Gather disease-gene association data: Collect information on the gene-disease associations from reliable sources such as public databases (e.g., Online Mendelian Inheritance in Man, OMIM) or curated disease-gene association repositories.

Define the set of diseases: Determine the set of diseases or phenotypes that will be considered for calculating the DPI. This could be a specific subset of diseases or a broader range, depending on the research question.

Calculate disease association count: Count the number of diseases or phenotypes associated with the gene of interest within the defined set. Each disease or phenotype with a documented association to the gene is considered as a pleiotropic event.

Normalize the DPI: To compare genes with different numbers of disease associations, it is common to normalize the DPI by dividing the disease association count by the total number of diseases or phenotypes in the set. This normalization step provides a DPI value between 0 and 1, where 0 indicates no pleiotropic associations and 1 indicates involvement in all diseases or phenotypes within the set.

It's important to note that the specific definition and calculation of the DPI may vary depending on the context and the disease-gene association data used. Different studies or databases may have varying inclusion criteria for diseases, and the DPI calculation might incorporate additional considerations or statistical methods. Therefore, it is recommended to consult the literature or specific analysis tools to understand the precise calculation used in a particular context.

The probability of a gene being loss-of-function (LoF) intolerant refers to the likelihood that mutations in the gene resulting in loss-of-function alterations would have a detrimental effect on an organism's phenotype. Several approaches and tools have been developed to estimate this probability based on various types of genomic and functional data.

One widely used tool for estimating the probability of a gene being LoF intolerant is the "Probability of Being Loss-of-Function Intolerant (pLI)" score provided by the Exome Aggregation Consortium (ExAC) or Genome Aggregation Database (gnomAD). This score is derived from large-scale sequencing data and quantifies the intolerance of a gene to loss-of-function mutations.

The pLI score ranges from 0 to 1, where higher values indicate a higher probability of LoF intolerance. A pLI score of 1 suggests that the gene is highly intolerant to loss-of-function mutations, while a score closer to 0 indicates a greater tolerance to such mutations.

The pLI score is typically based on various factors, including the number of observed loss-of-function mutations in the gene compared to the expected number based on the mutation rate, the functional impact of the mutations, and the gene's tolerance to sequence variation in the general population.

It's important to note that the pLI score is an estimate and should be interpreted in conjunction with other available evidence and functional studies to make robust conclusions about a gene's loss-of-function intolerance. Additionally, other tools and approaches may exist that provide similar or complementary information about a gene's LoF intolerance, such as functional genomics data or gene expression patterns in specific tissues or developmental stages.