DIFFERENTIAL EXPRESSION SEARCHES

Use this search to find genes that are differentially expressed between the samples of a microarray or RNA-Seq experiment. This search offers incredible flexibility in parameter choices as you set up your search, but your parameter choices drastically affect the subset of genes returned by the search. This document explains some parameter choices and how they affect your results.

Consider a microarray or RNA-Seq experiment that measures gene expression at 0, 1, 6, 12, 18 and 24 hours post-infection. Each time point, or sample, contains expression values for thousands of genes. This search looks for genes that are differentially expressed between time points and determines differential expression based on fold-change – which is calculated as a pairwise comparison between gene expression values in a reference and a comparison sample. When you set up the search, you define the differential expression cutoffs, sample groupings and the operations (if any) applied to sample groupings. For each gene, the search calculates differential expression as the ratio of expression in the samples that you want to compare and returns genes that meet your parameter choices.

The Basic Search: 1 reference vs. 1 comparison sample

You choose samples, fold change cutoff, and direction of change.

A one-to-one comparison between two samples is the most straightforward comparison that can be made with this search. Suppose you want to find genes that are up-regulated 2-fold or greater between the zero reference and 12 hours post-infection. For each gene, the search calculates fold change as a ratio of two numbers: expression at 12 hours post-infection vs. expression at time zero. The search returns every gene whose expression at 12 hours post-infection is greater than at time zero by at least 2-fold.

Complex search: 1 reference vs. multiple comparison samples

In addition to the Basic Search options, you choose an operation to be applied to multiple samples.

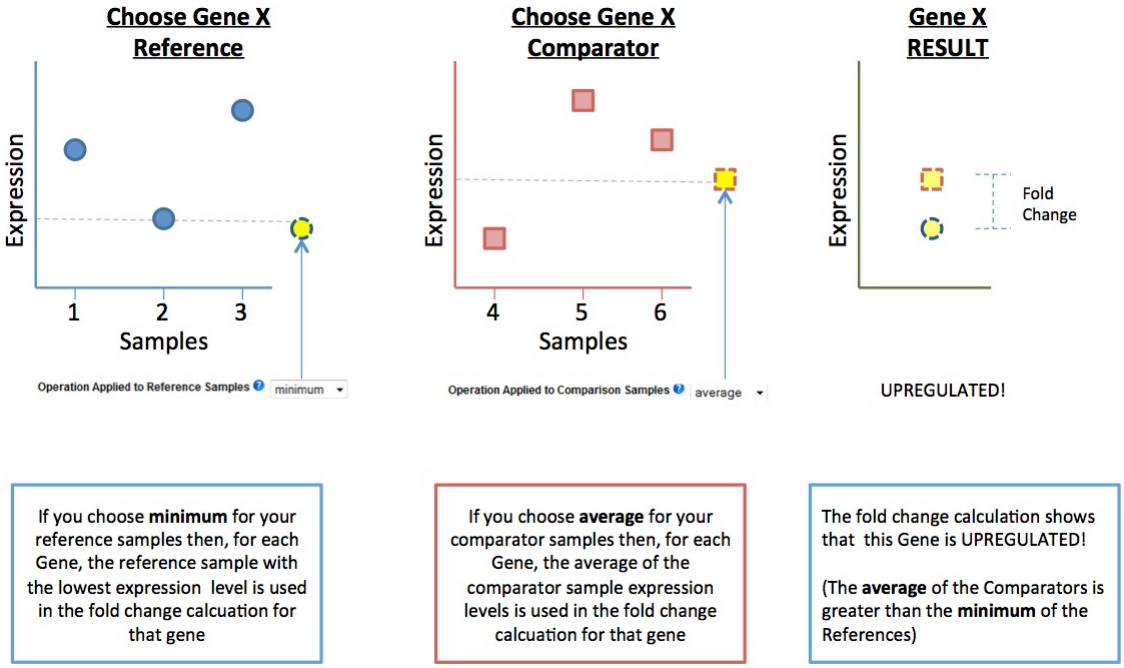
Suppose you want to know what genes were up-regulated by at least 2-fold in both the 12 and 18 hour time points, compared to time 0. In this case there are three expression values for each gene: the zero reference, and two comparisons -- 12 and 18 hours. Before calculating fold change, it is necessary to combine the expression values from the two comparison samples into a single number which can be either the maximum, minimum, median or average value for the 12 and 18 hour time points. On the search page, when you select multiple Comparison Samples, you will then be asked to define this additional parameter for the Comparison Samples. For example, if you choose to apply the minimum operator, the search will return every gene whose lowest expression value among the 12 and 18 hour time points is greater than at time zero by at least 2-fold.

Another complex search: multiple reference vs. multiple comparison samples

In addition to the Basic Search options, you choose an operation applied to multiple reference and multiple comparison samples.

Suppose you want to compare the early time points (0, 1 and 6 hours post-infection) to the late time points (12, 18 and 24 hours post-infection). In this case, there are six expression values: three from reference samples 0, 1 and 6 hours and three from comparison samples 12, 18 and 24 hours. Each group of expression values needs to be combined into a single number which can be either the maximum, minimum, median or average value across the samples in the

group. When setting up the search, you will be asked to define the operation applied to the reference and comparison samples. Applying the minimum operator to the reference and the average operator to the comparison will return every gene whose lowest expression value among the early time points is at least 2-fold lower than the average expression in the late time points.



Selecting maximum, minimum, median or average will affect the number of genes returned by the search

These four methods for combining samples into a single value will each result in a unique fold-change, and can be used to identify different patterns of expression. The broadest window for identifying up-regulated genes is created using the minimum for the Reference samples and the maximum for the Comparison samples. Using this logic, you can identify all genes with a fold-change between ANY two samples by selecting all samples for the Reference (and then choosing minimum for this group) and all samples for the Comparison (and then choosing maximum for this group). On the other hand, the most restrictive window for identifying up-regulated genes is created using the maximum of the Reference samples and the minimum of the Comparison samples. Such a search identifies genes that are up-regulated between any sample in the Reference group and any sample in the Comparison group.

Calculating fold-change in RNA-seq experiments and the importance of a floor

In RNA-seq experiments, gene expression level is measured in FPKM (Fragments Per Kilobase of transcript per Million mapped reads). To calculate fold-change for a gene, the FPKM value in the Comparison sample is divided by the FPKM value in the Reference sample. Let's say that for gene A, this is 10 / 1 = 10-fold. However, gene B is not detected in the Reference sample and its expression in the Comparison sample is 5, so the equation is 5 / 0 = Infinity. Infinity is not a reasonable fold-change. In fact, 0 is not a reasonable expression value. The value of 0 just means that the expression is below the level of detection. That leads to the question of what is the level of detection? In an RNA-seq experiment, if 1 read maps to a gene, then you may call it expressed. However, 1 is not necessarily reliable because error could cause it to be 2 or 0. We have chosen 10 reads (that map uniquely to a gene) as a reliable

indicator of gene expression. This is what we have deemed to be the “floor”; any number less than 10 reads is considered unreliable and is thus set to 10.

This number 10 can be converted to an FPKM value if you know the size of the transcript and the amount of sequencing in the experiment (i.e, the total number of mapped reads). For example, in *Plasmodium falciparum*, the average transcript size is 2.1kb. Let's say an experiment produced 10,000,000 mapped reads. Thus, 10 mapped reads for a single gene corresponds to: FPKM = 10 mapped reads / (2.1kb * 10 million reads) = 0.48. That means for this particular experiment and for the average gene in this species, 0.48 FPKM corresponds to 10 mapped reads. Thus, for this example, we use 0.48 as the default floor. Any FPKM value below this floor is raised to the floor when calculating the fold-change. Back to gene B, the new fold-change calculation will be 5 / 0.48 = 10.4 fold-change. This represents the fold-change that can be determined for this gene in this experiment. The actual biological fold-change might be greater than 10.4, but we would need greater sequencing (and more total mapped reads) to accurately determine the true fold-change. As an example of greater sequencing, this experiment could be repeated with 100,000,000 mapped reads. Thus, 10 mapped reads for a single gene corresponds to: FPKM = 10 mapped reads / (2.1kb * 100 million reads) = 0.048, which would be the floor. If gene B again produces FPKM values of 5 for one sample and 0 for the other, then the fold-change calculation is 5 / 0.048 = 104 fold-change. Thus, a greater amount of sequencing (and a correspondingly higher number of mapped reads) leads to a lower floor (and greater sensitivity for low expression).

On this website, we calculate this default floor each time based on the amount of sequencing in the experiment and the average transcript size in the organism. We recommend using this default floor, as it takes advantage of the amount of sequencing performed by the particular experiment. However, users can select from a set of reasonable floor values.
Note that in release 38 and earlier (before August 20, 2018), we were using 1 FPKM as the floor. If you want to achieve results that are consistent with your previous work at our website, choose a floor of 1.