Frequently Asked Questions for OncoMX

General FAQs

1. Where is the cancer mutation and expression data from?
   Cancer mutation and expression data is derived from multiple publically available resources such as CIViC, ClinVar, COSMIC, ICGC, IntOGen, and TCGA

2. Where can I find information about differential expression only?
   https://hive.biochemistry.gwu.edu/bioxpress/about in the readme link

3. How would you use search for a particular gene implicated in breast cancer?
   Go to https://www.oncomx.org/ then type in the gene name and press enter. The search will redirect to the **Gene Specific Overview** where interaction with different datasets for different cancers can be viewed.

4. I cannot find information on expression levels of a specific gene.
   Go to https://www.oncomx.org/, type in the gene name and search. This redirects to the gene specific overview page. Scroll down and click on the BioXpress and the table provides expression levels for the gene implicated in various cancers.

5. Where can I find portal update information for OncoMX?
   This information can be provided under the about section.

6. Where does the OncoMX data come from?
   OncoMX data is derived from different sources in Bgee, BioMuta, BioXpress, DEXTER, DiMeX, EDRN, and Reactome

BioXpress FAQs

1. Why do search results for a single cancer type report a large number of genes with significant differential expression?
   In our current analysis (BioXpress v2.0), we treated factors independently and do not impose limitations (beyond p-value) on fold change or number of patients affected, resulting in a more relaxed reporting of significant genes, especially for large datasets. Our goal for this version was to make as much information available to the user as possible, allowing users to define their own specific criteria for relevance. We are in the process of updating BioXpress with a more stringent set of default filters. We suggest the following treatment of the current version to limit the number of significant genes to higher value targets: 1) consider log2FC values - sometimes very small absolute log2FC values could be reported as significant; 2) consider patient frequencies - in our table, we provide the number of patients who have the same expression trend (up or down) as our pooled analysis; 3) consider setting a more strict adjusted p-value threshold.

2. How can users view the expression trend frequency and significance for a given gene across different types of cancer?
   Enter your gene names directly into the "Query" field and click "Search." When the search completes, you should be able to view the different charts summarizing the expression trends for your genes in cancer by toggling through the different tabs in the

chart viewer near the top of the page. Quantitative p-values and additional information are included in the table at the bottom of the page.

3. How are adjusted p-values obtained after multiple hypothesis testing?
   BioXpress v2.0 uses a slightly different method in performing differential expression analysis (DE) than v1.0. BioXpress v2.0 (2017) takes as input of DE samples for each cancer type. During our analysis, two levels were used to categorize samples pooled together for a single cancer type: one for separating cancer and adjacent non-tumorous samples, and the other one for separating samples for different patients. Adjusted p-values are automatically computed by application of the Benjamini-Hochberg procedure to adjust for multiple testing.
4. In tumor-only expression, are mRNA expression levels reported as normalized counts or raw read counts?
   We have altered the BioXpress pipeline since the initial research paper was published. In the current version (v2.0), the tumor-only expression is not normalized, but is reported as the log2(raw read counts).

5. Are pre-computed, normalized tumor expression values available for download?
   Our available data contains only the log2 fold change (and p-values) associated with a differential expression analysis of a pool of matched tumor and adjacent non-tumor samples - we do not maintain the normalized expression values in this version. All available downloads can be accessed at https://hive.biochemistry.gwu.edu/bioxpress/archive. Select the table called "BioXpress Gene Differential Expression" for the master list of all genes identified in our pipeline with associated log2FC values and p-values for all cancers (as described in the readme (https://hive.biochemistry.gwu.edu/bioxpress/readme).

6. Is the default expression trend frequency chart displaying only significant expression changes?
   In the current version of BioXpress (v2.0), all fold changes are automatically designated to be either a fold increase or a fold decrease (i.e. over-expressed or under-expressed). The chart in the expression profile is then reporting the percent of samples (patients) in each dataset that have increased or decreased fold changes without taking into account significance. The intensity of the fold change is not directly used as a criterion in determining significance of expression changes at this time. Significance is determined by p-values reported from the differential expression analysis of samples directly from the DESeq output.

7. What does the "Expression trend significance" plot show
   The "Expression trend significance" chart shows only the portion of samples (patients) deemed to be significantly over- or under-expressed by p-value.

8. Is there any possible way for me to search for a list of genes and get log2 fold change?
   You may easily download the entire table and parse that table in excel or command line. Alternately, run a search for each independently, download the tables, and then merge them.

BioMuta FAQs

1. What is the "frequency" for each position in the interim table?

The first page you land on when you search a term is an interim results page. We are currently reformatting the interim page due to lack of organizational clarity. Because the table is sorted by unique transcripts, the frequency reported corresponds to the number of mutations mapping to a specific transcript in a specific gene in a specific type of cancer.

2. Is there a way to retrieve a list of single-nucleotide variants in literature including manual curation and text-mining (along with PubMed IDs)?
   We have pilot data extracted from abstracts for such information. The method is described here (https://www.ncbi.nlm.nih.gov/pubmed/27073839).

3. What does the column labeled "Status" or "Study_type" mean in the download and results tables?
   The column label "Status" from earlier versions corresponds to the column label "Study_type" in the current version. The "SM" and "LG" represent "Small-scale study" and "Large-scale study," respectively, referring to different types of studies generating SNVs based on different sample sizes, purpose of studies, cost, type and complexity of data analysis, and generalizability of the results.

4. How were data and annotations extracted from UniProtKB?
   A complete proteome was retrieved using UniProt search "reviewed:yes AND organism:"Homo sapiens (Human) [9606]". All corresponding files were downloaded as plain file (txt format). Custom python scripts were used to automatically parse each file. For a given accession, we extracted UniProKB accession from 'AC' line; gene symbol was extracted from 'GN' line; and variation information including amino acid position, reference, variation, cancer, FT ID, cancer type, and PMID was extracted from appropriate FT line.

5. In BDNF, for example, why does the frequency in colon cancer change from 12 to 204 when you open the graphs?
   The graph is reporting all mutations for all transcripts, including the same mutation that may map to multiple transcripts. This number will likely contain some redundancy due to overlapping regions in different transcripts.

6. Why are the "specific position mutations frequencies" for a given gene different in the graph across all cancers?
   This graph is reporting, across all cancer types, the number of mutations reported at a given position in the amino acid sequence for a given protein. For example, for BDNF, position 127 has a peak height of 135 because there are 135 mutations across all cancer types reported at that position. As with the first chart, this number will almost definitely contain some redundancy due to overlap of some positions in multiple transcripts.

7. How can we search for the percentage of patients mutated at a particular gene in individual cancers?
   This information is not currently accessible in the interface.

8. How can I find the amino acid substitution at each position for a gene?
   All positions for all transcripts are mapped to the positional coordinates of the canonical isoform in UniProt.

9. Why is rs67341913 reported as –/T in the cluster report allele section but annotated as –/A in the RefSeq?
   This refSNP maps to the reference assembly on the forward strand, but the mRNA is on the reverse strand of the assembly, so the SNP allele is flipped (reverse complement) to '-/A' for reporting on mRNA.You can see this in the GeneView section of this rs number.