Introduction

This website implements the LymphGen algorithm as described in George W., et al. "A probabilistic classification tool for genetic subtypes of diffuse large B cell lymphoma with therapeutic implications." *Cancer cell* 37.4 (2020): 551-568. It takes sequence, translocation, and copy-number data to categorize diffuse large B-cell lymphoma (DLBCL) samples. Ideally, each sample should have complete exome and copy number data, as well as information as to whether the sample included BCL2 or BCL6 translocations. However, the algorithm is designed to adjust itself to the data available for each sample. The performance of the predictor will depend on what data is made available; so in addition to the prediction results, the user will be presented with estimates of how much agreement there is likely to be between the prediction under the full data vs. under the limited data set.

Input files

LymphGen accepts six files as input. The Sample Annotation file, Mutation Flat file, and Mutation Gene List file are all required for sample prediction. The Copy Number Flat file and Copy Number Gene List file are required in order for the A53 subtype to be predicted, and their inclusion will also improve the accuracy of other predictors. The Arm Flat file is not required but will improve the prediction of the A53 subtype. Descriptions of the files and their formats are listed below.

By clicking the "Load Example Data" button, a set of example data will be loaded. These are artificially constructed samples based on what one might expect to upload if one had a limited mutation gene panel, but complete copy number information and with data availability varying from sample to sample. Once the example data are loaded, links to the example input files are provided so that the user can explore them to see the proper input format.

Study name (optional): If included, this will be added as a prefix to the names of the results file; otherwise, it will default to "userStudy".

Sample Annotation file (Required): This tab-delimited text file contains information about the data availability of the samples being analyzed. Each row indicates a sample to be analyzed. It consists of the following four required columns in order:

1) **Sample name**: A text string used to identify the sample.
2) **Copy number availability**: a binary variable set to **1** if copy number data is available for that sample, set to **0** if copy number data is not available for that sample.
3) **BCL2 Translocation**: a binary variable set to 1 if that sample has a BCL2 translocation, set to 0 if that sample does not have a BCL2 translocation. If there is no information regarding BCL2 translocation, this should be set to "NA".
4) **BCL6 Translocation**: a binary variable set to 1 if that sample has a BCL6 translocation, set to 0 if that sample does not have a BCL6 translocation. If there is no information regarding BCL6 translocation this should be set to "NA".

Mutation Flat file (Required): This is a tab-delimited text file with each row indicating a single alteration on a single sample. If a sample has multiple alterations for a gene, all should be listed on separate rows. The file includes the following three required columns and a final optional column:

1) **Sample name (Required):** The sample that has the alteration. These should match the sample names given in the Sample Annotation file.
2) **ENTREZ.ID (Required):** The NCBI Entrez Gene ID number for the gene containing the alteration.
3) **Type (Required):** The type of alteration chosen from the following set of possible values:
    a. **TRUNC:** indicating a nonsense, frameshift, splice donor, splice acceptor, or start loss mutation in the coding region of the gene.
    b. **MUTATION:** indicating a missense or non-frameshift indel mutation in the coding region of the gene.
    c. **Synon:** indicating a mutation in the 5'UTR of the gene or a synonymous mutation in the coding region within 4kb of the transcription start site.
    d. **L265P:** indicating a L265P mutation of the MYD88 gene. (see below)
4) **Location (Recommended):** The chromosome location of the start of the mutation according to the HG37 genome build.

Special considerations: The LymphGen predictor treats L265P mutations differently from all other mutations of MYD88; so in order for accurate prediction, mutations of the L265P locus of MYD88 should be given the Type "L265P". Alternatively, if location is included for the MYD88 mutations, the identification of L265P will be handled automatically.

For NOTCH1 and NOTCH2 we only consider truncation mutations that affect the C-terminal PEST domain (mutation base pair position less than 139391455 for NOTCH1, and mutation base pair position less than 120459150 for NOTCH2). For CD79B, we considered only mutations that would selectively alter their C-terminal ITAM regions. Specifically, we choose truncating mutations with mutation base pair position less than 62007172 or non-truncating mutations with mutation base pair position less than 62006800. EZH2 mutations are restricted to those that targeted the catalytic domain (mutation base pair position between 148508764 and 148506238. If no location information is available, we recommend including all NOTCH1, NOTCH2, EZH2 and CD79B mutations since most fall within these their respective regions.

If a location columns is included, mutations in these gene outside of their prescribed regions will be automatically eliminate, as will Synon mutations outside of the 4kb range of the start site.

Mutation Gene List file (Required): A text file with a single column indicating the NCBI Entrez Gene IDs, for all genes that were analyzed for mutations. Every ENTREZ.ID included in the mutation flat file should be included on this list, but this list may include genes that were not found to have alternations in any of the samples. This list should apply to all samples run on the array. If the different samples have different gene lists (for example, if they are analyzed on different targeted arrays), then they should be run through the LymphGen algorithm in separate batches.

Copy Number Flat file (Required if copy number is included): This tab-delimited text file lists for each sample the genes that are overlapped by focal (30Mb or less) regions of copy-number change. If a region overlaps multiple genes, all genes overlapped by that region should be listed on separate rows. If a gene is covered by multiple regions of abnormal copy number on a sample, then include each on a separate row. Although not required, including copy number information will improve prediction accuracy, and its exclusion will prevent the prediction of the A53 subtype. The file consists of the following three columns:

1) **Sample name:** The sample that has the region of copy number change. These should match the sample names given in the Sample Annotation file
2) **ENTREZ.ID:** The NCBI Entrez Gene ID number for the gene overlapped by the region.
3) **Type :** The type of the copy number change for the region chosen from the following set of possibilities.
    a. **Gain:** Indicating a single copy increase in copy number.
    b. **Amp:** Indicating an increase in copy number of two or more.
    c. **HETLOSS:** indicating a heterozygous loss of a single copy.
    d. **HOMDEL:** indicating complete loss of both alleles in the region.

Copy Number Gene List file (Required if copy number is included): A text file with a single column indicating the NCBI Entrez Gene IDs for all genes that were analyzed for copy number. Every ENTREZ.ID included in the Mutation Flat file should be included on this list, but this list may include genes that were not found to have alternations in any of the samples. This list should apply to all samples run on the array. If the different samples have different gene lists (for example, if they are analyzed on different targeted arrays), then they should be run through the LymphGen algorithm in separate batches.

Arm Flat file (Recommended if Copy number is included): A tab-delimited text file indicating copy number changes of chromosomal arms for all samples. It is not required for the LymphGen algorithm but including this information will improve the prediction accuracy. We define an arm as having a change in copy number if more than 80% of the length of the arm has that change or one of larger effect. So, for example, an arm that has 60% of its length covered by single copy gains and an additional 30% covered by a multiple copy amplification should be counted as having a Gain for that arm. Whole chromosome changes are defined as those that have a given abnormality in both the p and q arms of the chromosome. This file has the following three columns:

1) **Sample name:** The sample that has the arm with the copy change. These should match the sample names given in the Sample Annotation file.
2) **Arm:** The arm or chromosome with altered copy number. Arms are indicated by a number followed by p or q (e.g., 6p, 18q or Xp). Whole chromosomal changes are indicated by the chromosome number followed by “chrom” (e.g., 3chrom or Ychrom).
3) **Type :** The type of the copy number change for the arm chosen from the following set of possibilities
    a. **Gain:** Indicating a single copy increase in copy number.
    b. **Amp:** Indicating an increase in copy number of two or more.
    c. **HETLOSS:** indicating a heterozygous loss of a single copy.
    d. **HOMDEL:** indicating complete loss of both alleles in the region.

Input flags

Below the data input area is a set of flags that should be used to further specify what data is available. With the exception of the individual availability of copy number as specified in the sample annotation file, it is assumed that these flags apply to all samples in the uploaded data. If different samples required different flags, then they should be analyzed in separate runs.

Copy number class

No copy number: Use this flag if no copy number data is available on any samples. This selection overrides the Sample Annotation file.

Full copy number: Use this flag if copy number data is available for some samples and is able to distinguish between single copy gains and losses vs. multiple copy amplifications and homozygous losses.

HOMDEL and AMP only: Use this flag if the copy number platform only has the sensitivity to detect homozygous deletions or multiple copy amplifications.

HETLOSS and GAIN only: Use this flag if the copy number platform only indicates an increase or decrease in copy number but is unable to determine the number of copies gained or lost.

Select Subtypes

Flag the subtypes you want to be considered for prediction. By default, all subtypes are selected. If, due to a limited set of available features, one of the subtypes has extremely poor performance, it may be helpful to unselect a given subtype so as to not adversely affect the other predictions. If no copy number is available, then the A53 subtype will be excluded; also, if NOTCH1 is not in the Mutation gene list, then the N1 subtype will be excluded even if the N1 flag is selected.

Additional flags

Has L265P: Deselect this flag if you lack mutation location information and so cannot determine whether a given MYD88 mutation hits the L265P location.

Has Trunc: Deselect this flag if you lack the ability to distinguish between truncating and non-truncating mutations.

Running prediction

Once the data has been uploaded and the flags set, click the “Submit for prediction” button to begin the prediction.

If there is a significant error in the uploaded files, this will be reported to the user. They will be returned to the input page; otherwise, the user will be presented with a confirmation of the uploaded files, and a pie graphic will indicate the progress of the prediction. When it completes its rotation, the user will be sent to the Results page.

Results

Warnings

Upon reading the uploaded files, the LymphGen algorithm may identify discrepancies that might indicate a problem with the uploaded data (for example, if a sample was indicated as having copy number data available, but no data for that sample was found in the Copy Number Flat file). It is recommended that the user review any warnings to make sure that they don't represent an error.

Model performance graphs

LymphGen algorithm adapts its prediction model to the type of data available; however, limited data may degrade performance. To determine the degree of degradation, we classify the 574 LymphGen training samples using the limited data and compare the prediction results to what they were under the complete model. We then display the within subtype sensitivity and positive predictive value with a bar graph. Multiple models may be generated to predict the samples of a single uploaded data set (for example, samples for which copy number data was not available will use a different model for prediction than those for which such data was available), in which case a performance of each model will be displayed in a separate graph.

Disclaimer: The reported statistics are estimated purely based on the availability of fusion and copy number data as indicated in the Sample Annotation file, the mutation and copy number gene lists, and the setting of the input flags. No attempt is made to model the quality of copy number or mutation calls, which could also have an important effect on model performance. The reported values should be viewed as a best-case scenario. The LymphGen algorithm has not been certified for clinical use and is intended for research use only.

Raw Result Files

Clicking the "Download Raw Results" link will allow the user to download a zip file containing the following three tab-delimited text files, prefixed by the Study name (or "userStudy", if no Study name was provided).

Warn: This includes a detailed description of the warnings reported on the Results page.

Compare: This provides a table giving the numerical values used to generate the model performance graphs as described above.

Result: This file contains the prediction results for the uploaded samples. It includes the following columns.

Sample name: The name of the sample as indicated in the sample annotation file.

Copy Number: Whether the sample was indicated as having copy number data available.

BCL2.Translocation: Whether that sample had a BCL2 translocation.

BCL6.Translocation: Whether that sample had a BCL6 translocation.

Model.Used: Which model was used to predict the sample, and so links that sample with a particular performance graph.

Confidence.BN2, Confidence.EZB, etc.: These columns give the confidence values for each predictive submodel. If a sample lacked Copy number data, the A53 prediction of the sample will be set to NA. If a subtype was not included among the possible predictions, the column representing that sample will be eliminated.

BN2.Feature.Count, EZB.Feature.Count, etc.: These columns give for each sample the number of features that that sample has which match each subtype. With the exception of the N1 subtype, the LymphGen algorithm requires that a sample have a count of at least two features associated with a given subtype in order to be predicted as that subtype.

Subtype Prediction: This column indicates the final predicted class of each sample. If the sample strongly exhibits the qualities of two or more types, then it will be predicted as a combination sample, with the multiple types separated by a "/" (e.g, "BN2/MCD").