BIRD Prediction Database User Manual

1 BIRD Prediction Database Online GUI

Directly access database online at (shinyapp link)

2 Package Installation and GUI Launching

Installation

How to launch shinyapp

3 Data Processing and Prediction

3.1 RNA-seq Data Processing and BIRD Predictions

A total of 345,118 human RNA-seq samples are collected to make chromatin accessibility predictions using BIRD. All RNA sequencing data are obtained from recount3 project R package version 1.10.2. Only human projects are used in prediction. The RNA-seq samples count matrices come from SRA, GTEx, and TCGA.

Fragments Per Kilobase of transcript per Million mapped reads (FPKM) are used as inputs to BIRD prediction model. Sample FPKMs are calculated from raw counts provided by recount3 using DESeq2. For samples with great sequencing depth such that the largest read count is larger than the integer range in R (2,147,483,647) or sparse sequencing projects for which the raw count matrix contain at least one zero read count for every gene, the sample FPKMs are calculated by method described in countToFPKM package version 1.0.

BIRD predictions are done using BIRD version 1.1.1 with the BIRD model trained on human hg38 genome (v1.4). BIRD produces a predicted chromatin accessibility value for each 200bp bin in the BIRD output genomic range. The output is log-transformed, thus all predictions downloaded from this database and all analyses made on database predictions are based on log-transformed values.

3.2 Mapping between SNP and BIRD Prediction Bins

SNP data are downloaded from GWAS Catalog all associations v1.0 on June 21, 2023. The closest BIRD output bin for each SNP is found and the distance between the output bin and the SNP is calculated. SNPs that are more than 1,000,000bp away from any BIRD prediction bins are discarded. The current rs number, which corresponds to the SNP_ID_CURRENT attribute in GWAS Catalog, for each SNP is recorded as the SNP id.

3.3 Mapping between TSS and BIRD Prediction Bins

Transcriptional start site (TSS) data for all known genes in hg38 genome is loaded from the R package TxDb.Hsapiens.UCSC.hg38.knownGene version 3.17.0. Nearest TSS within 1,000,000bp is found for all BIRD prediction bins, and the relative position of each BIRD prediction bin to its nearest TSS is

calculated (the TSS can be either downstream, upstream, or inside the bin). If a bin is of equal distance to multiple TSSs, an arbitrary TSS is selected, and the corresponding gene is assigned as its nearest gene.

4 BIRD Prediction Database GUI Overview

The BIRD Prediction Database GUI is made up of 3 sections. The first section is “Input Selection”, where users need to select samples for which the predicted chromatin accessibility can be downloaded or analysed. In this section, users can also choose to upload a text file containing chromatin accessibility values to operate on custom input. The second section is “Prediction Download”, where the selected sample predictions can be downloaded from the database in various formats or viewed in UCSC genome browser. The third section is “Prediction Visualization”, where data analysis tools can be used to extract information and visualize the selected prediction data.

5 Sample Selection

The first step is to select samples that the user is interested in viewing or analysing. There are multiple ways to do this. To select samples from the database, the user needs to select a project, either from the project table or by typing in or uploading a text file of project ids under the “Project Selection” tab. Then, the user can select the samples also from the table or by text input under the “Sample Selection” tab. After sample selection, the user can also define the prediction range under the “Sample range selection” tab, so that all successive analysis is conducted on this custom genomic range. This range can either be designated by uploading a BED file or by manually inputting the range for each chromosome.

Another way to define an input is by uploading a text file containing the desired chromatin accessibility input. The text file must be a tab-separated file with a header row. The first three columns must be Chromosome, Start, and End of the prediction bins, with the start and end indices being inclusive. The other columns should be the (predicted) chromatin accessibility of samples. The column names will be used as sample names.

6 Data Download and Custom Track Generation

Prediction data can be downloaded in txt or zip format. For txt format, the resulting file will be in tab-separated format, with first three columns indicating chromosome, genomic bin start position, and end position. The following columns are sample predictions. For zip format, the resulting folder will contain all sample predictions, each in its own txt file. The file format is the same as with the txt file download option, though each txt file in the zip folder will contain predictions for only one sample.

Aside from downloading predictions, a UCSC genome browser track hub can also be used to view the selected sample predictions. Click on the “Download session file” button to download the session file for a pre-generated UCSC genome browser hub that displays only the selected samples at the selected region. To view the tracks, the user must

7 PCA and Pseudo-Time Analysis

7.1 PCA

Principal component analysis (PCA) can be performed on selected predictions to visualize data and to enable downstream analyses. Since the prediction usually contain many genomic regions, only the top variance genomic bins will be used in PCA calculations. The mean and variance of predictions for each bin is calculated and transformed by log base 2, and a generalized additive model is fitted to relate log2 variance to log2 mean. The expected variance can be derived from the fitted model, and the ratio between the true variance and the expected variance is used to rank the genomic bins, and only the top rows are kept to perform PCA. The first two principal components of PCA result are plotted in PCA Plot, and the explained variance and cumulative explained variance are plotted against the number of principal components, while the dotted blue line in the cumulative explained variance plot indicates the computed optimal number of principal components using elbow method.

7.2 Pseudo-Time Analysis

Pseudo-time trajectory computation can be performed using a user-designated number of principal components based on the previous PCA results done with the top variance genomic bins. The user can choose a range of number of anchoring clusters for pseudo-time computation. K-means clustering is performed to cluster the samples based on chromatin accessibility prediction values to serve as the anchoring clusters. The optimal number of anchoring clusters is selected from the set range of cluster numbers using elbow method. The clustering table for the optimal number of clusters is available for both view and download. The minimum spanning tree is then found for the anchors, and the pseudo-time trajectory can be selected from a set of paths that connect all pairs of possible trajectory end points. The trajectories are ranked by decreasing path length and by decreasing number of samples included in clusters covered by the trajectory, and the top trajectory is selected to be the main trajectory. User can also choose to display other non-optimal trajectories and can download the sample ids that fall on the selected trajectory sorted by pseudo-time. The selected trajectory is plotted on the first two principal components axes. User can select points by brushing on the plot, and the selected points' project id and sample id will be displayed in a table below the plot.

Samples are ordered along a selected trajectory by the following approach. The samples belonging to the anchoring clusters along the selected trajectory are kept and projected onto the trajectory. For given anchoring clusters c_1 and c_2 that are connected by an edge e in the selected path, samples in clusters c_1 and c_2 that are closest to the cluster centers of each other are chosen to be projected onto edge e. That is, c_1 samples that are closer to c_2 cluster center than any other cluster centers are selected, and c_2 samples closest to c_1 cluster center are selected in the same way. Then, these samples are ordered based on their projection on e as well as the selected trajectory direction.

Aside form plotting the trajectory itself, users can also choose to display average predicted chromatin accessibility along selected pseudo-time trajectory together with a GAM-fitted curve. One option is to first cluster the genomic bins by k-means, and then selecting one cluster for which the average prediction value will be computed and plotted against the pseudo-time. The default number of k-means clusters is again computed by elbow method using total within-cluster sum of squares for each cluster number, and the total within-cluster sum of squares plot can be displayed by checking the "Show Total Within Cluster Sum of Squares" option. Other clustering results can be shown by checking the

“Show genomic bin k-means clustering result” box, where a series of clustering plots and tables can be chosen, including the clusters plotted by UMAP or PCA, cluster means table, as well as the clustering table. The average predicted accessibility against pseudo-time can be display on either a scatter plot or a heat map. For scatter plot, user must choose a genomic bin cluster for which the average prediction is calculated and graphed. For heat map, each row represents a genomic bin cluster, and each column is a pseudo-timepoint. An interactive heat map is available by clicking on “Show interactive heatmap” button. This interactive option is available for all heat maps.

Another option for displaying predictions along pseudo-time is to plot the prediction for a single genomic bin against pseudo-time. There are also a scatterplot or heat map option for this. The heat map rows are genomic bins, and the heat map columns are the samples sorted in pseudo-time. The heat map rows can be sorted by row clusters, decreasing row variance, or genomic position. The user can also choose to display the nearest gene for each row (genomic bin). To display the scatterplot, the user needs to choose a genomic bin from the information table, and the predicted accessibility scatterplot will be displayed below.

The last option is to display average chromatin accessibility for chosen genes. For the scatterplot option, the user must select a gene from the gene table, and the associated genomic bins average chromatin accessibility is calculated and plotted. The mapping between genes and genomic bins is done by assigning each genomic bin the nearest gene within 1000kb. The gene locations are represented by the transcriptional start sites (TSS), which a gene can have multiple. This way, each genomic bin is associated with one nearest gene, and the set of associated genomic bins for a gene are the bins that are closer to this gene than to any other genes. A selected gene’s TSSs and the set of associated bins are available in tables. For the heat map option, the user does not need to select a single gene. Rather, each row of the heat map is a gene, and the columns are the samples sorted by pseudo-time.

8 Heat Map

To display the heat map, the user needs to first select the number of top variance rows that are included in the heat map. The expected variance calculation for each genomic bin is outlined in section 7.1. The heat map columns are selected samples ordered by column clusters. Each heat map row is a genomic bin. Each heat map cell value is the predicted $\log_2$ chromatin accessibility value at a genomic bin in a selected sample. The user can choose from one of three methods for heat map rows sorting: row clusters, decreasing variance, and genomic position. User can also choose to annotate heat map rows with the nearest gene name for each genomic bin. To view an interactive heat map, the user can also click on the Show Interactive Heatmap button, which creates a pop-up modal where the user can view subsets of the heat map, search for heat map rows or columns, and adjust display dimensions of heat map.

9 Differential Analysis

9.1 Sample Grouping

To perform differential analysis, we need to first make groups from samples so that comparison can be made between groups. There are multiple ways to group samples together. One way to do this is through k-means clustering. In this method, we first reduce the prediction data dimensionality by selecting the top variance genomic bins and performing PCA. The user can choose the number of top variance rows (bins) to use in PCA computation. PCA is then performed on the top variance rows to further reduce prediction dimensionality. The user can check the PCA results by plot of the first 2 principal components, explained variance, and cumulative explained variance. Then, the user can select the number of principal components to retain for k-means clustering of samples. The default selection here is the optimal number of principal components as calculated by elbow method on cumulative explained variance. After this, k-means clustering is performed on the selected principal components to group the samples. K-means clustering is done with a range of cluster numbers from 2 to a maximum of 10, and the optimal number of clusters is chosen from the range by elbow method on total within cluster sum-of-squares for each cluster number. The user can view the total within cluster sum-of-squares plot and can choose any number of clusters within the range. The PCA plot of clusters and clustering table can be displayed for the chosen number of clusters. The user needs to click on the “Confirm groups” button to confirm group selection before moving on to differential testing.

The other three methods for sample grouping are done by manual selection. The first manual method is by text input, where the user either enters the text into the text box or uploads the text via file upload. The text indicating samples grouping has a specific format, where each line represents a group of samples, and sample ids are separated by commas in a single line. The user is required to click on the “Confirm groups” button to submit group selection.

The second manual method is by drag and drop, where groups are indicated by boxes, and the user needs to drag the sample id tags to the desired group. The user can add new groups by clicking on the “Add new group” button, or they can delete the last group by clicking on the “Delete last group” button. Clicking the “reset” button causes all samples to be cleared from groups and all new groups to be deleted. The user also needs to click on the “Confirm groups” button to check if group selection is valid.

The last selection method is by selecting sample points from a PCA plot. Details about the sample points in the PCA plot can be viewed by hovering on the points. The user can use either the Lasso Select or Box Select tool accessible on the top right corner of the plot to select points from the plot. The selected samples information is shown in a table below the plot, and the user can add selected points to a new group or to an existing group by clicking on “Add to new group” or “Add to existing group” button, respectively. The user can also add all remaining points to a new group by clicking the “Add all remaining samples to new group” button. The created groups can be viewed by clicking on the “View selected groups” button, which creates a pop-up modal showing a PCA plot of all samples colored by group. The group contents are listed below the plot as tables. To clear all existing groups and start over, the user can click on “Clear all groups” button. After grouping all samples, the user may click on “Confirm group selection” to submit the sample groups.

9.2 Differential Test Methods

User may run differential test on selected groups in the “Group differential analysis” tab. The user must first select at least two groups from all created sample groups on which the differential test is performed. At least one group in the selected groups must contain more than one sample. There are different differential test methods to select from for tests on two groups and test on more than two

groups. Tests on two groups include t-test, non-parametric Wilcoxon test, and permutation test. Tests on more than two groups include ANOVA test, non-parametric Kruskal-Wallis test, and permutation test. For two-group tests, the user will be asked to select an alternative hypothesis from 3 options: two-sided, less, or greater. For both permutation tests, the user can select the number of permutations used in the test. For ANOVA test and permutation test on more than two groups, accessibility values for some genomic bins might have 0 within-group variance for all selected groups, and these 0 variance values can result in undefined values in ANOVA test and permutation test. For these two tests, the user can choose to use shrinkage to squeeze the variances such that the total within-group variances are no longer 0 for these genomic bins so that ANOVA test and permutation tests can produce finite results. If the user chooses to not use shrinkage, the genomic bins with 0 total within-group variances are removed from the results. The user can also choose to scale the prediction values for each sample before performing the differential test.

9.3 Result Summary and Visualization

The differential test results are presented in 3 tabs: “Result”, “Summary”, and “Volcano plot”. The “Result” tab shows the result table, which displays the test statistic, p-value, and FDR for all genomic bins in user-selected genomic range. This table can be downloaded.

The “Summary” tab shows the number of significant genomic bins according to a user-defined FDR threshold. The user can download a BED file of all significant bins. This panel also shows histograms of test statistic, p-value, and FDR as well as a line plot of sorted p-values for all selected bins before and after FDR adjustment.

The “Volcano plot” tab allows the user to see volcano plots for differential accessibility between two groups, plotting $-\log_{10}$ p-value or $-\log_{10}$ FDR against $\log_2$ fold change. User can adjust the p-value or FDR and log fold change thresholds to color points outside of the threshold as red or blue, being either more or less accessible genomic bins. The user can also brush on the volcano plot to select points (genomic bins) to display a table of detailed information about the selected bins, including their nearest genes, nearest SNP, and SNP-associated trait or disease. This brushed points table can be downloaded.

9.4 Gene Ontology Analysis

With the differential test results, the user can also perform gene ontology (GO) terms enrichment analysis. First, the significant bins are selected by a user-defined FDR threshold. These significant bins mapping to nearest genes can be viewed in a table, which can also be downloaded. The genes nearest these significant bins are used as target genes in the analysis. The user needs to select a control gene group for the enrichment analysis, which can either be all genes near the user-selected genomic range or all genes within 1,000,000bp of BIRD prediction range. Then, the user selects a GO domain in which enriched terms are searched for and the number of top GO terms to return. User can click on “Find top GO terms” button to run the search.

The search will return 5 result panels that the user can choose from to visualize the results. The first one is a results table that includes information on all top GO terms returned. It shows the term ID, term description, number of genes annotated to the GO term, number of significant genes associated with the GO term, expected number of significant genes in the GO term, p-value, FDR, three odds ratios computed by different methods, and names of all significant genes annotated to the GO term. The three

odds ratios include the conditional odds ratio estimated by conditional Maximum Likelihood Estimation, the unconditional odds ratio (ratio of the odds of a significant gene being annotated to the odds of a non-significant gene being annotated in the GO term), and the unconditional odds ratio computed with a pseudo count of 0.5. This pseudo count odds ratio is calculated because for some GO terms, the conditional and unconditional odds ratio are infinite due to all annotated genes in a GO term being significant.

The user can also view the GO terms graph with either $-\log_{10}$ p-value or $-\log_{10}$ FDR plotted against GO term ids laid out on the x-axis. The user can display a significance threshold line for p-value or FDR in the respective graphs. The size of each point is proportional to the annotation size for each term. The user can hover on points to show details about the GO term and can view a table of all GO term details in a pop-up modal by clicking on a point. The significant genes list associated with the GO term can be downloaded from the pop-up. The user can also select points from the graph, and the selected points' details will be shown in a table.

A volcano plot of $-\log_{10}$ p-values plotted against odds ratios is generated, and the user can adjust the p-value significance threshold and the odds ratio threshold to display the significant GO terms in the graph. User can hover on points to show details.

The user can also view a bar plot of either $-\log_{10}$ p-value or $-\log_{10}$ FDR for all returned GO terms. The bar plot can be sorted by decreasing or increasing values or by custom order. The custom order can be inputted via a list of GO term IDs separated by line breaks or by drag sort. The user can also zoom in on the bar plot using the zoom tool on the top right corner of the bar plot.

10 Disease SNP Analysis

The disease SNP analysis computes the mean accessibility values for SNP windows selected by user. The user first selects diseases or traits that he/she wants to investigate. All SNPs that are associated with the selected traits and within 1,000,000 bp of a BIRD prediction bin can be viewed in a table by clicking on the "Show disease associated SNPs table" button. This table can be downloaded. These SNPs will be the subjects of the analysis, and they can be subsetted by selecting the desired SNPs rows in the SNPs table and clicking on "Confirm SNP subset". Then, the user can select the extension length for the SNP windows. By default, the window extends from -500 bp position of SNP to +500 bp position. The number of BIRD prediction bins inside the SNP windows are shown in a message, and the positions of these bins can be viewed in a table accessible by clicking on the "Show table of genomic bins covered by SNP windows" button. This table can be downloaded. The BED file of these covered bins can be downloaded as well. The user can click on "Run disease SNP analysis" to see the results.

The analysis result panels are different between selection of a single disease or trait and selection of multiple traits. When a single disease or trait is selected, the results panel contains three tabs: the result table, the bar plot, and the heat map. The result table rows are samples, and the columns are the mean predicted log accessibility in all bins covered by selected SNP windows for that trait as well as the normalized mean accessibility, which is the mean accessibility divided by the average accessibility over the entire BIRD output range for each sample prediction. This result table can be downloaded. The bar plot shows the mean accessibility or normalized mean accessibility of selected SNP

windows for each selected sample, depending on user's choice. The bar plot rows can be sorted by increasing or decreasing values, or they can be sorted by a custom order defined by user input text or drag sort. The heat map has the samples as columns and each covered genomic bin as a row. Only a user-selected number of top variance genomic bins are displayed in the heat map, but the total number of genomic bins covered by SNP windows is usually less than this selected number, so this usually does not influence the heat map. The user can also select to display the heat map for only genomic bins covered by a specific SNP. They can also sort the heat map rows by either row clusters (default), decreasing variance, or by genomic position order. The user can interact with the heat map by clicking on the "Show interactive heat map" button.

For selection of multiple diseases or traits, the result panel consists of two tabs: the result table and the heat map. The result table tab contains two tables, one for mean accessibility and the other for normalized mean accessibility. For each table, rows are samples and columns are selected traits or diseases. Both tables can be downloaded. The heat map also has samples as rows and diseases or traits as columns. The heat map values can be either mean accessibility or normalized mean accessibility, chosen by the user. The default sorting scheme for heat map rows is by row clusters, but the user can also choose to sort by increasing or decreasing values for one of the columns (one selected disease or trait) or by custom order inputted by text or drag sort. Interactive heat map is also available.