**DISCRIMINATOR User Manual**

**Version 1.0**

**April 2022**

----------------------------------------------------------------------------------------------------------------------------------------------------------------

**INTRODUCTION:**

DISCRIMINATOR is designed to assign to assign provisional pathogenicity classifications ('Primary', 'Secondary', 'Benign', and 'Non-Coding') to copy number variants.

**Table of Contents**

[**REQUIREMENTS:** 2](#_Toc101168161)

[**INSTALLATION:** 2](#_Toc101168162)

[**COMPLETE SETUP INSTRUCTIONS:** 2](#_Toc101168163)

[File Structure & Location for SETUP Files 2](#_Toc101168164)

[Add and Format PRIMARY Interval Files 3](#_Toc101168165)

[Add and Format COHORT Files 4](#_Toc101168166)

[(Optional): COHORT File Pre-Processing 4](#_Toc101168167)

[Add and Format Gene and Exon Boundaries files 5](#_Toc101168168)

[Modify the Set of Parameters Used by DISCRIMINATOR to Assign Provisional Classifications 6](#_Toc101168169)

[**RUNNING DISCRIMINATOR:** 6](#_Toc101168170)

[EXAMPLE 1: Single Cohort \* Primary Interval File \* Jaccard Index Run 6](#_Toc101168171)

[EXAMPLE 2: Multiple Cohort \* Primary Interval File \* Jaccard Index Run 7](#_Toc101168172)

[**COMMON ERRORS:** 8](#_Toc101168173)

[Use of Python3 instead of Python2 (2.7.X): 8](#_Toc101168174)

# **REQUIREMENTS:**

DISCRIMNATOR requires Python (2.7.X) in a Linux operating system to run. Python3 is currently not supported. A list of the required Python modules is listed below. If you want to run the pre-processing script, bedtools is also required.

Required Python Modules:

* progress
* networkx; version 1.8.1
* intervaltree

# **INSTALLATION:**

After you have downloaded the DISCRIMINATOR\_X.X.tar.gz file, move zipped file to the directory in which you want to run DISCRIMINATOR. Then, decompress the file.

|  |
| --- |
| tar -xzvf DISCRIMINATOR\_1.0.tar.gz --one-top-level=DISCRIMINATOR --strip-components 1 |

Diagram

Description automatically generated

# **COMPLETE SETUP INSTRUCTIONS:**

## File Structure & Location for SETUP Files

Diagram

Description automatically generated

Add and Format BENIGN Interval Files

The SOURCE\_DATA folder contains three subfolders: BENIGN, PRIMARY, and COHORTS. Each of these subfolders contain the set of input files for DISCRIMINATOR.

Diagram, schematic

Description automatically generated

The set of 'benign' intervals that DISCRIMINATOR will use to compute overlap with patient CNVs for assigning 'benign' provisional classifications can be found in the SOURCE\_DATA/BENIGN folder. To add a 'benign interval' file, simply add the file (in BED format) to the SOURCE\_DATA/BENIGN folder.

**BENIGN INTERVALS**: Datasets should be in standard 3 or 4-column BED format.

1. The first column contains the chromosome on which the benign interval is located, such as chr1, chr2, chr3
2. The second column contains the start coordinate for the benign interval
3. The third column contains the end coordinate for the benign interval
4. (Optional): The fourth column contains the benign interval ID

## Add and Format PRIMARY Interval Files

The SOURCE\_DATA folder contains three subfolders: BENIGN, PRIMARY, and COHORTS. Each of these subfolders contain the set of input files for DISCRIMINATOR. The set of 'primary' intervals that DISCRIMINATOR will use to compute overlap with patient CNVs for assigning 'PRIMARY' provisional classifications can be found in the SOURCE\_DATA/PRIMARY folder. To add a 'primary interval' file, simply add the file (with the following formatting requirements) to the SOURCE\_DATA/PRIMARY folder.

**PRIMARY INTERVALS**: Datasets should be in comma-delimited ('.csv'). The first line of the file should contain only the work "track". Each set of primary interval coordinates should be listed twice – once for deletions and once for duplications. Multiple coordinate sets for a single Primary CNV ID are allowed – simply list the same ID for each set of coordinates in the fourth column.

1. The first column contains the chromosome on which the primary interval is located, formatted without "chr" such as 1,2,3
2. The second column contains the start coordinate for the primary interval
3. The third column contains the end coordinate for the primary interval
4. The fourth column contains the Primary CNV ID. Spaces ARE allowed. Must end with 'DELETION' or 'DUPLICATION'. If omitted, deletion and duplication events will be grouped together in the COMPLETE\_OUTPUT file.
5. The fifth column contains the direction of Primary CNV (accepted values: CN Gain or CN Loss).

**Example File Preview:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| track |  |  |  |  |
| 1 | 145118354 | 145833117 | 1Q21.1 TAR DELETION | CN Loss |
| 1 | 145118354 | 145833117 | 1Q21.1 TAR DUPLICATION | CN Gain |
| 1 | 145413192 | 145610627 | 1Q21.1 TAR DELETION | CN Loss |
| 1 | 145413192 | 145610627 | 1Q21.1 TAR DUPLICATION | CN Gain |
| 1 | 145883119 | 148026038 | 1Q21.1 NEURO DELETION | CN Loss |
| 1 | 145883119 | 148026038 | 1Q21.1 NEURO DUPLICATION | CN Gain |
| 1 | 145118354 | 148026038 | 1Q21.1 TAR NEURO DELETION | CN Loss |
| 1 | 145118354 | 148026038 | 1Q21.1 TAR NEURO DUPLICATION | CN Gain |

## Add and Format COHORT Files

The set of patient CNVs that DISCRIMINATOR will processes can be found in the SOURCE\_DATA/COHORT folder. To add a 'cohort' file (set of patient CNVs to process), simply add the file to the SOURCE\_DATA/PRIMARY folder.

**PATIENT CNVs**: Datasets should be in tab-delimited with three columns. **NOTE: The header ('Sample', 'Chromosome Region', 'Event') is required.**

1. The first column contains the patient or sample identifier.
2. The second column contains the genomic coordinates for a single CNV in the following format (chr:start-end).
3. The third column contains the direction of the CNV (accepted values: CN Gain or CN Loss)

**Example File Preview:**

|  |  |  |
| --- | --- | --- |
| Sample | Chromosome Region | Event |
| Sample\_001 | chr10:115852177-115878450 | CN Loss |
| Sample\_001 | chr10:124334195-124354306 | CN Loss |
| Sample\_001 | chr11:18941182-18961975 | CN Loss |
| Sample\_002 | chr1:144279310-144462929 | CN Gain |
| Sample\_002 | chr11:51337076-51376963 | CN Loss |

## (Optional): COHORT File Pre-Processing

DISCRIMINATOR will natively merge all CNVs in the same direction in the same patient if they are within 30kb. A pythonic secondary merging step to account for segmental duplication content and probe design of the 1q21.1, 15q11.2, and 22q11.2 regions is currently under development. In the meantime, this optional merging step can be performed manually prior to running DISCRIMINATOR.

To manually run this optional merging step:

1. Navigate to the PREPROCESSING folder (~/PATH/DISCRIMINATOR/SOURCE\_DATA/COHORTS/PREPROCESSING)
2. Run the DISCRIMINATOR\_PreProcessing.sh script

|  |
| --- |
| ./DISCRIMINATOR\_PreProcessing.sh <INPUT FILE PATH> <OUTPUT FILE PATH> |

**Example:**

|  |
| --- |
| ./DISCRIMINATOR\_PreProcessing.sh ../COHORT1.txt ../COHORT1\_PROCESSED.txt |

## Add and Format Gene and Exon Boundaries files

DISCRIMINATOR will annotate each CNV with the gene(s) and exon(s) which are intersected. The version of these files supplied by DISCRIMINATOR consists of the GENCODE Basic Gene Set (V39lift37; last updated: 2022-01-16) which was filtered to only include known protein-coding genes (transcriptClass="coding" and geneType="protein\_coding"). Users may upload and/or replace these files (~/PATH/DISCRIMINATOR/CODE/DATA) provided they meet the following formatting requirements. **NOTE: Files must be named GENE\_BOUNDARIES\_HGXX.bed and EXON\_BOUNDARIES\_HGXX.bed**

Diagram

Description automatically generated

**GENE/EXON BOUNDARIES**: Datasets should be tab-delimited and all spaces should be converted to underscore "\_". Users are required to maintain the 4-column structure.

1. The first column contains the gene name. Duplicate names are allowed for EXON\_BOUNDARIES.
2. The second column contains the chromosome on which the gene/exon is located, such as chr1, chr2, chr3
3. The third column contains the start coordinate for the gene/exon
4. The fourth column contains the end coordinate for the gene/exon

**Example File Preview:**

|  |  |  |  |
| --- | --- | --- | --- |
| Gene Name | Chrom | Start | End |
| OR4F5 | chr1 | 65418 | 71585 |
| OR4F29 | chr1 | 367639 | 368634 |
| OR4F16 | chr1 | 621095 | 622034 |
| SAMD11 | chr1 | 859302 | 879954 |
| NOC2L | chr1 | 879582 | 894636 |
| KLHL17 | chr1 | 895963 | 901099 |
| PLEKHN1 | chr1 | 901861 | 911245 |

## Modify the Set of Parameters Used by DISCRIMINATOR to Assign Provisional Classifications

DISCRIMINATOR will assign provisional pathogenicity classifications for each cohort file \* jaccard index value \* primary interval file combination.

Diagram

Description automatically generated

**BENIGN\_FILES**: This file within the PARAMETERS folder contains the set of benign region files that DISCRIMINATOR will use to assign 'BENIGN' provisional classifications. Simply update this file to contain the file names you wish to use.

**PRIMARY\_FILES**: This file within the PARAMETERS folder contains the set of primary interval files that DISCRIMINATOR will use to assign 'PRIMARY' provisional classifications. Simply update this file to contain the file names you wish to use.

**COHORT\_FILES**: This file within the PARAMETERS folder contains the set of cohort files (or patient CNVs) that DISCRIMINATOR will be processing. Simply update this file to contain the file names you wish to use.

**JACCARD\_INDEX**: This file within the PARAMETERS folder contains the set of jaccard index values that DISCRIMINATOR will use to assign 'PRIMARY' provisional classifications. The default value of '0.40' was chosen following an optimization process, but users are free to change or add additional value(s).

# **RUNNING DISCRIMINATOR:**

Enter the DISCRIMINATOR directory, and use the following command to run DISCRIMINATOR.

|  |
| --- |
| cd ~/PATH/DISCRIMINATOR  python CODE/INTERFACE\_DISCRIMINATOR.py |

## EXAMPLE 1: Single Cohort \* Primary Interval File \* Jaccard Index Run

COHORT\_FILES Contents:

|  |
| --- |
| ReportedCNVs.txt |

PRIMARY\_FILES Contents:

|  |
| --- |
| PrimaryCNVs.csv |

JACCARD\_INDEX Contents:

|  |
| --- |
| 0.4 |

If all the required Python modules are installed DISCRIMINATOR will run and you will get the following output to the console. The overall runtime is dependent on the total number of cohort files, primary interval files, and jaccard index values that need to be processed.

|  |
| --- |
| BUILDING PATIENT MERGED DATA FILES...  Processing |################################| 100% COHORT1.txt  BUILDING BENIGN MERGED DATA FILES...  Processing |################################| 100%  ASSIGNING PROVISIONAL PATHOGENICITY CLASSIFICATIONS  Processing |################################| 100% COHORT1\_PrimaryCNVs\_0.4.txt |

DISCRIMINATOR Output Structure:

Diagram

Description automatically generated

DISCRIMINATOR will output all files to the 'OUTPUT' folder. For each Cohort File \* Primary CNV Interval File \* Jaccard Index Value combination, four files are created:

1. BENIGN\_MERGED.txt: This file contains the set of merged benign intervals from all benign interval files listed in 'BENIGN\_FILES'. Merging was direction specific.
2. COHORT\_MERGED.txt: This file contains the set of merged & processed patient/sample CNVs which DISCRIMINATOR will be evaluating.
3. COLLAPSED\_ANNOTATIONS.txt: This file contains the set of merged and processed patient/sample CNVs which were evaluated. Each CNV is annotated with the provisional classification, and the set of overlapping gene(s) and exon(s).
4. COMPLETE\_OUTPUT.txt: This file contains the total count of each primary CNV which was identified (direction specific) along with several cumulative and descriptive metrics on the CNVs within the cohort file.

## EXAMPLE 2: Multiple Cohort \* Primary Interval File \* Jaccard Index Run

COHORT\_FILES Contents:

|  |
| --- |
| COHORT1.txt  COHORT2.txt  COHORT3.txt |

PRIMARY\_FILES Contents:

|  |
| --- |
| PrimaryCNVs.csv  PrimaryCNVs2.csv |

JACCARD\_INDEX Contents:

|  |
| --- |
| 0.1  0.2 |

|  |
| --- |
| BUILDING PATIENT MERGED DATA FILES...  Processing |################################| 100% COHORT1.txt  Processing |################################| 100% COHORT2.txt  Processing |################################| 100% COHORT3.txt  BUILDING BENIGN MERGED DATA FILES...  Processing |################################| 100%  ASSIGNING PROVISIONAL PATHOGENICITY CLASSIFICATIONS  Processing |################################| 100% COHORT1\_PrimaryCNVs\_0.1.txt  Processing |################################| 100% COHORT1\_PrimaryCNVs\_0.2.txt  Processing |################################| 100% COHORT2\_PrimaryCNVs\_0.1.txt  Processing |################################| 100% COHORT2\_PrimaryCNVs\_0.2.txt  Processing |################################| 100% COHORT3\_PrimaryCNVs\_0.1.txt  Processing |################################| 100% COHORT3\_PrimaryCNVs\_0.2.txt  Processing |################################| 100% COHORT1\_PrimaryCNVs2\_0.1.txt  Processing |################################| 100% COHORT1\_PrimaryCNVs2\_0.2.txt  Processing |################################| 100% COHORT2\_PrimaryCNVs2\_0.1.txt  Processing |################################| 100% COHORT2\_PrimaryCNVs2\_0.2.txt  Processing |################################| 100% COHORT3\_PrimaryCNVs2\_0.1.txt  Processing |################################| 100% COHORT3\_PrimaryCNVs2\_0.2.txt |

# **COMMON ERRORS:**

## Use of Python3 instead of Python2 (2.7.X):

|  |
| --- |
| File "CODE/INTERFACE\_DISCRIMINATOR.py", line 6  print "\nERROR: AUTOMATION FILE NOT PROVIDED. PLEASE REVIEW README FILE."  ^  SyntaxError: Missing parentheses in call to 'print'. Did you mean print("\nERROR: AUTOMATION FILE NOT PROVIDED. PLEASE REVIEW README FILE.")? |