This package contains the following three sub-folders:

1. TCI\_GenerateGlobalDriver: Source code for a program that searches for a global driver for a phenotype using Bayesian approach.

2. TCI: Source code for a program that performs tumor-specific causal inference.

3. Data: A folder that contains a set of example training data and a shell script for running experiments.

**Compiling the code**

Install OpenMP.

Use GCC version 4.2.1 or later

Run Makefile within TCI and TCI\_GenerateGlobalDriver to compile the programs, which will yield two executable programs:

1. TCI\_GD -- This program searches for a "global driver" for a phenotype at the population level using

a Bayesian causal framework.

2. TCI -- This program performs TCI analysis.

**Preparing the data**

Both TCI\_GD and TCI take 3 matrices as input (see example shell script for options):

1. There is an N-by-D matrix, referred to as the E matrix (e.g., the DEGmatrix\_brca.csv file in the Data folder),

where N is the number of cases, and D is the number of

phenotypes of interest, such as differentially expressed genes (DEGs). Each row represents phenotypes

observed in a tumor, and rows should be matched with those in the A and P matrices, see below.

An element in a row represents whether the phenotype (indexed by D) is present ("1") or not ("0") in a given tumor. The

2. There is an N-by-G matrix, referred to as the A matrix, where N is the number of cases, and G is the number of genes. Each row represents the somatic genome alteration (SGA) data of a tumor, where a "1" indicates that the corresponding gene is altered in the current tumor, and "0" otherwise.

3. There is an N-by-G matrix for TCI, referred to as P matrix, where N and G should match those of the A matrix for the TCI analysis. Again, each row represents a tumor, where each element is the prior probability that an SGA, corresponding to a gene with "1" in A matrix, is a driver of a phenotype (e.g., a differentially expressed gene) in a tumor. The sum of the prior probabilities of a tumor (a row) should be 1. One can prepare this matrix according to available prior knowledge, e.g., the probability that TP53 is a driver gene of a tumor and normalize among SGAs in a given tumor, or use a uniform prior if no prior knowledge is available.

In TCI\_GD, there 1-by-G vector, call the P matrix, that is not case-specific; it considers every SGA ever observed in a population as a candidate cause for a phenotype.

**Performing a TCI analysis**

First, search for population-wide drivers of each phenotype by running TCI\_GD:

./TCI\_GD -p PmatrixFilePathname -f AmatrixFilePathname -d EmatrixFilePathname -o populationDriverFilePathname

TCI\_GD takes as input the 3 matrices described above. It outputs a D-by-3 comma-separated CSV file, in which each row corresponds to a phenotype. The first column is phenotype name, second column is the id its most probable driver at the population level, the third column is the probability assigned by Bayesian analysis.

Next, run TCI, which will use the results produced by TCI\_GD:

TCI:./TCI -p PmatrixFilePathname -f AmatrixFilePathname -d EmatrixFilePathname -g populationDriverFilePathname -o outputFileDirectoryName [-s startingRow -e endingRow]

TCI takes the 3 input matrices described above plus the population-wide driver information (-g). If no additional optional argument (-s and -e) are provided, it iterates through each tumor (the rows in the 3 matrices) as being a test case and uses the rest of the matrix as the training data to perform tumor-specific causal inference. For each phenotype that is present in a tumor (e.g., a DEG event indicated by a "1" in E matrix), TCI outputs the posterior probability for each SGA being the driver of that phenotype in that tumor. Thus, for each tumor, TCI outputs a D-by-G matrix in the directory designated by the "-o" option, in which each row contains the posterior probabilities of SGAs in the tumor as the cause of the phenotype indicated by the row.

If one only wants to run a selected subset of the cases as being the test cases, use "-s" and "-e" to indicate the beginning row and end row of the block of the test cases. TCI will iterate through each case within the block to perform TCI analysis, using all other cases (training and test) as a training set.

**Running an Analysis on Example Data**

There is a shell file "TCI\_GD\_TCI.sh" in the "Data" folder that uses the training data there to run the TCI system on example data.