Unraveling Epistatic Interactions Between Sites Under Drug-Dependent Selection in the *Mycobacterium tuberculosis* Genome

Technical description

Here we describe the project pipeline. To replicate analysis extract the content of the data folder with the project files from the archive into the desired location (referred further as the project root folder), enable execution of script files (\*.pl and \*.sh) and install EpiStat software versions v7.1 [https://github.com/gFedonin/EpiStat/tree/v7.1] and v.10.1 [https://github.com/gFedonin/EpiStat/tree/master ] into separate locations, according to the manuals. We assume that the user’s current folder at the beginning is the project root folder. Use the provided \*.sh scripts as examples.

The whole analysis is very time and memory consuming: to process large trees, e. g. for Isoniazid and Rifampicin GWAS, it needs HPC nodes with more than 15 cores and ~1Tb of RAM. You need to edit these scripts to adapt them for the configuration of your cluster and workload manager, e.g. SLURM. For example, it is better to rewrite the while loop in the GWAS/run\_epistat.sh, allowing for parallel execution on different HPC nodes.

## Searching for sites with mutations associated with drug phenotype changes

Select the GWAS subfolder as the current folder. Here we look for sites associated with drug phenotype changes. We analyze data for each drug separately. The GWAS folder contains subfolders with “\*.xparr” files, where mutations at genomic sites and changes in drug phenotypic states are mapped onto the tree branches for each drug.

1) Estimate the expected time of mutation occurrence for each drug tree and initialize the TAU parameters in the “\*.epistat.prm” files located in the drug subfolders. First, run the folowing command (file dirs.list contains the list of subfolders):

./estimate\_tau.sh dirs.list

This command will output “\*.xparr.tau” files, placing them in the corresponding drug subfolders. Afterwards, edit the “\*.epistat.prm” files and update the TAU parameters with the appropriate values for each drug, based on the generated “\*.xparr.tau” files.

2) Search for associations.

**Note!** We use BiRewire [https://www.bioconductor.org/packages/release/bioc/html/BiRewire.html] to randomly shuffle mutations on the tree branches. BiRewire has high RAM requirements for GWAS and you may need to constrain the number of its parallel executing jobs. Edit numbers of allowed parallel jobs in <DRUG\_NAME>/\* \_gnu\_parallel.opts files (“-j njobs” option).

Example: the file Isoniazid\_gnu\_parallel.opts contains the line “shuffle\_incidence\_matrix.R -j 15”. Increase the number of jobs (15) if you have more than 1500Gb on the node processing the Isoniazid.xparr file or decrease it proportionally otherwise.

To start the search for associations, run:

./run\_epistat.sh dirs.list

If it succeeds, the following result files will appear in each drug subfolder:

* \*.site\_pairs – the index file, which contains IDs of site pairs and the number of substitutions in each site
* \*.upper.pvalue – probabilities to obtain a value of genotype-phenotype association statistics at least as high as the observed value, derived from the null-model
* \*.lower.pvalue – probabilities to obtain a value of genotype-phenotype association statistics not greater than the observed value, derived from the null-model
* \*.lower.pvalue.pairs.fdr and \*.upper.pvalue.pairs.fdr – FDR estimated for corresponding p-values

## Searching for sites evolving under drug dependent selection

Select the DDSS subfolder as the current folder. DDSS is a shortcut for the Drug Dependent Selection Study. The DDSS folder contains subfolders with “\*.xparr” files, where mutations at genomic sites and changes in drug phenotypic states are mapped onto the tree branches for each drug. A branch node closer to the tree root is referred to as the parent node, and the other node is referred to as the daughter node. We look for sites with excess or lack of mutations occurring on the branches with resistant phenotypic states in daughter nodes.

1) **Set the same TAU parameters in “\*.epistat.prm” files** in the drug subfolders as for the GWAS.

2) Search for associations.

**Note!** We use BiRewire [https://www.bioconductor.org/packages/release/bioc/html/BiRewire.html] to randomly shuffle mutations on the tree branches. BiRewire has high RAM requirements for DDSS and you may need to constrain the number of its parallel executing jobs. Edit numbers of allowed parallel jobs in <DRUG\_NAME>/\* \_gnu\_parallel.opts files accounted by the “-j njobs” option.

Example: the file Isoniazid\_gnu\_parallel.opts contains the line “shuffle\_incidence\_matrix.R -j 15”. Increase the number of jobs – 15 if you have more than 1500Gb on the node processing the Isoniazid.xparr file or decrease it proportionally otherwise.

To start the search for associations, run:

./run\_epistat.sh dirs.list

The same result files are expected to appear in drug subfolders as were obtained for the GWAS.

## Correction of GWAS and DDSS results for correlations of drug phenotypes

Due to the high correlations between drugs phenotypes, many of the detected sites will be associated with drug resistance indirectly. We implement a procedure which allows, at least partially, disentangle direct and spurious associations. First, we convert the raw association statistics between genome sites and drugs into pseudocorrelations, and then we calculate association statistics between drugs and also convert them into pseudocorrelations (see Methods in the main text). Finally, we construct a matrix of pseudocorrelations of drugs and sites which includes three blocks: drugs-to-drugs and two drugs-to-sites blocks in direct and transposed orientations.

1. Join the results of association studies. To construct the association matrix we need to join results of all association studies for separate drugs. Select the project root folder as the current folder. Run:

./merge\_epistat\_projects.sh

After the script finishes, the initially empty subfolders merged\_GWAS and merged\_DDSS will contain the following files:

“\*.lower.pvalue”, “\*.upper.pvalue”, “\*.site\_pairs”, “\*.stat.exp”, “\*.mean”, “\*.var”.

2. Calculate association between drug phenotypes. To calculate association between drugs, select the

drugs\_vs\_drugs/ iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_GWAS as the current folder, then **estimate the TAU parameter:**

./iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_estimate\_tau.sh

**Set the estimated value for TAU** in the ‘iqtree\_snp\_drugs\_vs\_drugs\_10drugs.epistat.prm’ file and then **start the analysis**:

./iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_epistat.sh

The following result files will appear: “\*.intragene.unord\_pairs.lower.pvalue”, “\*.intragene.unord\_pairs.upper.pvalue”, “\*.site\_pairs”, “\*.mean”, “\*.var” and “\*stat.exp.\*”.

Leave the current folder and select the iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_DDSS as the current folder. Again, **estimate TAU:**

./iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_RR\_estimate\_tau.sh

**Set the estimated value for TAU** in the ‘iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_RR.epistat.prm’ file and **start the analysis**:

./iqtree\_snp\_drugs\_vs\_drugs\_10drugs\_RR\_epistat.sh

Check the results files in the current folder. In addition to “\*.intragene.unord\_pairs.lower.pvalue”, “\*.intragene.unord\_pairs.upper.pvalue”, “\*.site\_pairs”, “\*.mean”, “\*.var” and “\*stat.exp.\*”, file “\*. ord\_site\_pars.cov” will appear for DDSS.

Go to the project root folder.

3. Disentangle direct and spurious associations.

Now compile the pseudocorrelation matrices for GWAS and DDSS. Go to the drugs2vars folder. There are two parameter files describing the three blocks of each pseudo correlation matrix for GWAS and DDSS.

Before the start, please modify the content of following parameter files: ‘drugs2vars.ddss.all.mk\_coevolmtx.3.prm’ and ‘drugs2vars.gwas.mk\_coevolmtx.3.prm’, line 14. The estimated TAU parameters are parts of the names of files containing epistatic statistics generated during the analyses for associations between drug phenotypes - “\*.stat.exp.<TAU>” files. **Set an appropriate TAU value for EpiStat="" parameter in each “\*.prm” file.**

Then run the following commands:

./drugs2vars\_cor2pcor\_GWAS.sh

./drugs2vars\_cor2pcor\_DDSS.sh

For GWAS the folder will contain the following files:

* Submatrices representing blocks of the GWAS pseudo correlation matrix – ‘mtb.block.mtx’: ‘drugs.mtx’, ‘drugs2vars.mtx’ and transposed copy of the last – ‘tr(drugs2vars).mtx’.
* Submatrices representing blocks of the partial correlation matrix for GWAS: ‘drugs.l90.APC.cor2pcor.R.out, drugs2vars.l90.APC.cor2pcor.R.out’ and ‘tr(drugs2vars).l90.APC.cor2pcor.R.out’

The same results will be obtained for DDSS:

* Submatrices representing blocks of the DDSS pseudo correlation matrix – ‘mtb\_RR.all.block.mtx’: ‘drugs\_RR.all.mtx’, ‘drugs2vars\_RR.all.mtx’ and transposed copy of the last – ‘tr(drugs2vars\_RR.all).mtx’.
* Submatrices representing blocks of the partial correlation matrix for DDSS: ‘drugs\_RR.all.l90.APC.cor2pcor.R.out’, ‘drugs2vars\_RR.all.l90.APC.cor2pcor.R.out’ and ‘tr(drugs2vars\_RR.all).l90.APC.cor2pcor.R.out’

Matrices “drugs2vars\*.mtx” contain pseudocorrelation statistics, which are referred to in the main text of the paper as “association statistic” (PsCOR).

Matrices “drugs2vars\*.l90.APC.cor2pcor.R.out” contain statistics which are referred to in the main text of the paper as GnPh-PCOR.

## Calculating p-values for association of sites and drugs

We define the p-values of associations of a site and a drug as a minimum of four p-values obtained in two association tests GWAS and DDSS: two p-values, upper and lower, for each test. These p-values are used in the search for coordinated evolution of sites with control for drug phenotypic states – the main test in our study.

To calculate association p-values select the project root folder as the current folder and **run the command:**

./mk\_site2pheno.sh

File ‘min\_RR\_SR.site2pheno’ with the required statistics will appear in the directory.

**Go to the project’s root and copy this file into the ‘epistat/phen’ folder.**

## Search for concordantly and discordantly evolving sites

For the search for coordinately evolving pairs of sites, a file in the XPARR format containing the phylogenetic tree with mutations and changes of drug phenotypic states mapped on branches is required. For the version of test that takes into account phenotypic states, an additional file with p-values for association of sites and phenotypes is required (“\*.site2pheno”).

The folders epistat/phen and epistat/nophen contain the \*.xparr files with identical content. The “\*.site2pheno” file is situated in the ‘epistat/phen’ folder.

Before the analysis, it is necessary to **estimate the TAU parameter.**  Select the epistat subfolder as the current folder. Run the following command:

phen/estimate\_tau.sh

**Set the estimated value of TAU parameters in phen/epistat.phen.prm and in nophen /epistat.nophen.prm files.** The estimated TAU is included in the name of the file with epistatic statistics, which was generated during the analysis - the “\*.stat.exp.<TAU>” file. The appropriate file name has to be specified in the following parameter files in the ‘phen’ and ‘nophen’ folders before the start of the analysis: “\*.fdr.prm”, “\*mk\_summary.cfg” and “\*.mk\_coevolmtx.prm”.

To search for coordinated evolution controlling for drug phenotypic states, select the epistat/phen folder as the current folder. To start the analysis, run:

./run\_epistat.sh

If the run succeeds in the current folder the following files will appear:

* 9drugs.filtered.site\_pairs – the index file which contains IDs for all ordered pairs of sites;
* 9drugs.filtered.intragene.unord\_pairs.lower.pvalue – the file with lower p-values for epistatic statistics, site pairs with low p-values are discordantly evolving;
* 9drugs.filtered.intragene.unord\_pairs.upper.pvalue – the file with upper p-values for epistatic statistics, site pairs with low p-values are concordantly evolving;
* 9drugs.filtered.lower.pvalue.unord\_pairs.sites2.unlinked.fdr – the file with FDR estimated for lower p-values;
* 9drugs.filtered.upper.pvalue.unord\_pairs.pairs2. unlinked.fdr – the file with FDR estimated for upper p-values;
* 9drugs.all.block.mtx and 9drugs.all.mtx – the file with the pseudocorrelation matrix PsCOR;
* 9drugs.block.mtx and 9drugs.mtx – the file with positive value submatrix of the pseudocorrelation matrix;
* 9drugs.n0.l90.cor2pcor.R.out – the file with partial correlations PCOR calculated for the positive value submatrix of the pseudocorrelation matrix;

We consider site pairs as been concordantly and discordantly evolving if their upper and lower p-values meet the FDR<10% requirement, correspondingly. To calculate the FDR for upper and lower p-values, run the command:

./add\_fdr4pvalues.sh .

See results in the output files:

* 9drugs.z-scores.all.pairs2+upval\_FDR.tab
* 9drugs.z-scores.all.sites2+lpval\_FDR.tab

**Note!** To obtain the statistical summary for concordantly and discordantly evolving pairs set upper and lower p-value thresholds corresponding to the 10% FDR in ‘9drugs.pairs2.pos.FDR10.mk\_summary.cfg’ and ‘9drugs.sites2.neg.FDR10.mk\_summary.cfg’ files. To set appropriate thresholds for upper and lower p-values open each “\*. FDR10.mk\_summary.cfg” file in a text editor, e.g. notepad++ and change default threshold value in the third line of the file.

**Example**:

The third line of ‘9drugs.sites2.neg.FDR10.mk\_summary.cfg’:

“9drugs.intragene.unord\_pairs.lower.pvalue 1<0. 0146 lpval “

The “0. 0146” is the 10% FDR threshold for p-values listed in the first column of the ‘9drugs.filtered.intragene.unord\_pairs.lower.pvalue’ file.

To search coordinated evolution without control for phenotypic states on branches (the original test), select the epistat/nophen folder as the current folder and follow the instruction for the phenotype-aware test.

## Discriminating between epistatic and coordinated episodic selection

Looking for concordant and discordant evolution for a site pair, we estimate the epistatic statistic which is roughly proportional to the number of consecutive mutation pairs occurring on the tree in these sites. However, excess or lack of consecutive pairs of mutations could be caused by both epistasis and episodic selection acting on a site pair, thus epistatic statistic does not distinguish these options. In contrast to epistasis, coordinated episodic selection equally changes phylogenetic distances between nonconsecutive as well as between consecutive pairs of mutations. We use this fact to estimate whether the pair of sites evolves under episodic rather than epistatic selection. We estimate average distances between nonconsecutive pairs of mutations in a site pair. If this distance is shorter or longer than expected, sites in a pair are considered to be evolving under concordant or discordant episodic selection.

Testing for episodic selection, we also need to control for branch phenotypes.

### Looking for the site pairs evolving under coordinated episodic selection

1) First, for each site pair we need to **estimate mean distance between *consecutive* mutations** on the phylogeny.

Go to the epistat/phen folder and edit the script file 9drugs.run\_cumdist.sh specifying the EPISTAT7\_HOME variable which has to contain a path to the installed EpiStat v7.1 software:

EPISTAT7\_HOME =path/to/installed/EpiStat\_v7.1

Run:

./9drugs.run\_cumdist.sh

When it finishes, check that file ‘9drugs.filtered.stat.ident’ appeared in epistat/phen folder and files “\*.stat.ident” appeared within folders ./9drugs.filtered/pheno/<DrugID>/samples/L0I0

<DrugID> is the integer identifier of a drug. See the ‘drug\_codes.txt’ file in the project root folder to convert DrugIDs into the drug names.

Go to the ‘epistat/nophen’ folder and edit the script file ‘9drugs.run\_cumdist.sh’ specifying the EPISTAT7\_HOME variable which has to contain a path to the installed EpiStat v7.1 software:

EPISTAT7\_HOME =path/to/installed/EpiStat\_v7.1

Run:

./9drugs.run\_cumdist.sh

When it finishes, check that files ‘epistat/nophen/9drugs.filtered.stat.ident’ and “epistat/nophen/9drugs.filtered/fdr/\*.stat.ident” were produced.

Files “.stat.ident” are further used to calculate the average distance between nonconsecutive mutations which is defined as the difference of an average distance between all mutations and the average distance between consecutive mutations in a site pair.

2) To calculate distances between *non-consecutive* mutations on the phylogeny and assess if they significantly differ from the expected, go to epistat/disttest/scripts folder .

**Note!** To run this test, you need to install Bioperl if it is not already installed on your system.

For the phenotype-unaware test, check that folder epistat/phen contains files 9drugs.filtered.xparr and 9drugs.filtered.l.r.newick. At this step it should also contain files 9drugs.filtered.site\_pairs, 9drugs.filtered.stat.ident and min\_RR\_SR.site2pheno. Run

./pheno.sh disttest\_pheno.prm

If the run succeeds, file 9drugs.filtered\_dist.fdr\_results will appear in the epistat/distest/nsyn/pheno folder.

For the phenotype-unaware test, check that folder epistat/nophen contains files 9drugs.filtered.xparr and 9drugs.filtered.l.r.newick, as well as 9drugs.filtered.site\_pairs and 9drugs.filtered.stat.ident. Run

./nopheno.sh disttest\_nopheno.prm

If the run succeeds, file 9drugs.filtered\_dist.fdr\_results will appear in the epistat/distest/nsyn/nopheno folder.

### Get lists of epistatic concordantly and discordantly evolving pairs

To proceed further, we need files named ‘9drugs.filtered\_dist.fdr\_results’ for the branch phenotype-unaware test (situated in the ‘epistat/distest/nsyn/nopheno folder’) and for the test accounting for branch phenotypes (situated in the ‘epistat/distest/nsyn/pheno’ folder). Each of these two files contains estimated distances between nonconsecutive mutations in a site pair and corresponding z-scores, upper and lower p-values

To find epistatically linked pairs among the concordantly evolving ones, we need to exclude pairs with nonconsecutive mutations that are closer to each other than expected. To find epistatically evolving pairs among the discordantly evolving ones, we need to exclude pairs with nonconsecutive mutations that are more distant from each other than expected. We use Benjamini-Hochberg procedure to find pairs where distances between nonconsecutive mutations significantly differ from the expected values.

To proceed, go to the epistat/phen folder and run:

./add\_phen\_nonconseq.sh

./nonconsec\_BH.sh

The first command merges tables containing concordantly and discordantly evolving pairs with the ‘9drugs.filtered\_dist.fdr\_results’ table. The second command applies Benjamini-Hochberg procedure to the merged tables and outputs files with epistatic concordantly and discordantly evolving pairs. You can see results of the procedure in the “\*+nonconsec.BH\_greater\_FDR10.tab” files.

Finally, you need to add annotation for sites into the obtained result files. Site annotation includes gene name, site position in the corresponding gene or gene upstream, or genomic position if the site is intergenic, and WHO drug association grade. To add annotation, run the command:

./mk\_summary\_FDR10.sh

The suffices of files containing annotated tables are:

“\*+w\_gene\_names.tab” – these files contain gene names and site positions;

“\*+WHO.tab” – these files also contain WHO drug association grades for sites.

To obtain the analogous results for phenotype unaware test go to the ‘epistat/nophen’ folder and run:

./add\_nophen\_nonconseq.sh

./nonconsec\_BH.sh

./mk\_summary\_FDR10.sh

## Compare results of two tests for coordinated evolution

To compare the sets of predicted concordantly and discordantly evolving pairs of sites between methods that do or do not control for drug phenotypes, we merge the table of pairs predicted by one method with the table of pairs predicted by the other method. Note that we need to make comparisons in both directions.

To compare the sets of concordantly and discordantly evolving pairs predicted by the phenotype aware method with the predictions made by phenotype unaware method, go epistat/phen folder and run:

./add\_nophen2phen.sh

If it succeeds you’ll obtain the following result files:

* ‘9drugs.pairs2.pos.FDR10+noncosec+nophen\_FDR.tab’ – a list of concordantly evolving pairs of sites predicted by the phenotype *aware* method with statistics for nonconsecutive mutations, upper p-values estimated by the phenotype *unaware* method and the corresponding FDRs;
* ‘9drugs.sites2.neg.FDR10+noncosec+nophen\_FDR.tab’ – a list of discordantly evolving pairs of sites predicted by the phenotype *aware* method with statistics for nonconsecutive mutations, lower p-values estimated by the phenotype *unaware* method and the corresponding FDRs;
* ‘9drugs.pairs2.FDR10.pos\_epistatic.join\_nophen.tab’ - the result of the left joining operation for the set of *epistatic* concordantly evolving site pairs predicted by the phenotype *aware* method with the set of *epistatic* concordantly evolving site pairs predicted by the phenotype *unaware* method.
* ‘9drugs.sites2.FDR10.neg\_epistatic.join\_nophen.tab’ - the result of the left joining operation for the set of *epistatic* discordantly evolving site pairs predicted by the phenotype *aware* method with the set of *epistatic* discordantly evolving site pairs predicted by the phenotype *unaware* method.

To compare in the opposite direction, go to the epistat/nophen folder and run:

./add\_phen2nophen.sh

Check that the following result files are generated:

* ‘9drugs.pairs2.pos.FDR10+noncosec+phen\_FDR.tab’ – a list of concordantly evolving pairs of sites predicted by the phenotype *unaware* method with statistics for nonconsecutive mutations, upper p-values estimated by the phenotype-*aware* method and the corresponding FDRs;
* ‘9drugs.pairs2.pos.FDR10+noncosec+w\_gene\_names+phen\_FDR.tab’ - a list of concordantly evolving pairs of sites with gene names and site positions, upper p-values estimated by the phenotype-*aware* method and the corresponding FDRs;
* ‘9drugs.pairs2.pos.FDR10+noncosec+WHO+phen\_FDR.tab’ - a list of concordantly evolving pairs of sites with WHO catalog grades for sites, upper p-values estimated by the phenotype-*aware* method and the corresponding FDRs;
* ‘9drugs.pairs2.FDR10.pos\_epistatic+noncosec+phen\_FDR.tab’ – a list of *epistatic* concordantly evolving pairs predicted by the phenotype *unaware* method with upper p-values estimated by the phenotype *aware* method and the corresponding FDRs;
* ‘9drugs.pairs2.FDR10.pos\_epistatic+noncosec+w\_gene\_names+phen\_FDR.tab’ – a list of *epistatic* concordantly evolving pairs with gene names and site positions, upper p-values estimated by the phenotype-*aware* method and the corresponding FDRs;
* ‘9drugs.pairs2.FDR10.pos\_epistatic+noncosec+WHO+phen\_FDR.tab’ - *epistatic* concordantly evolving pairs with WHO catalog grades for sites, upper p-values estimated by the phenotype-*aware* method and the corresponding FDRs;
* ‘9drugs.sites2.neg.FDR10+noncosec+phen\_FDR.tab’ – a list of discordantly evolving pairs of sites predicted by the phenotype *unaware* method with statistics for nonconsecutive mutations, lower p-values estimated by the phenotype *aware* method and the corresponding FDRs;
* ‘9drugs.pairs2.FDR10.pos\_epistatic.join\_phen.tab’ - the result of the left joining operation for the set of *epistatic* concordantly evolving site pairs predicted by the phenotype *unaware* method with the set of *epistatic* concordantly evolving site pairs predicted by the phenotype *aware* method.
* ‘9drugs.sites2.FDR10.neg\_epistatic.join\_phen.tab’ - the result of the left joining operation for the set of *epistatic* discordantly evolving site pairs predicted by the phenotype

*unaware* method with the set of *epistatic* discordantly evolving site pairs predicted by the phenotype *aware* method.