Lung Function, COPD, Epigenetics and Accelerated Ageing Across the Lifespan

Coordinating team contact details

Analysts: Julieta Viglino (viglino@recerca.clinic.cat)
Sandra Casas Recasens (sacasas@recerca.clinic.cat)
Principal Investigator: Rosa Faner Canet (rfaner@recerca.clinic.cat)

Research hypothesis and specific aims

Chronic Obstructive Pulmonary Disease (COPD) is a complex, heterogeneous and prevalent disease with a high socio-economic burden whose underlying biological mechanisms are unknown. It is now well established that about half of the patients with COPD never achieved a normal lung function early in life, and there is increasing evidence that some COPD risks may derive from early life factors in this setting it is highly likely that respiratory diseases are the end result of a set of different dynamic environmental-gene interactions that can occur during the entire life span of an individual (time). Lung development and lung aging are often considered two independent phenomena, but they are tightly interrelated.

Here we propose that alterations during lung development influence age-related physiological deterioration and cause premature lung aging. We expect that the lung (lung tissue and bronchial biopsies), as affected organ in COPD, accumulates the highest burden of methylation changes, but that some of them, are reflected in the circulating blood. Subsequently we will explore if these changes are identifiable early in life. We sought to explore this hypothesis by comparing the epigenetic profile associated with FEV1<LLN and/or the severity of COPD at different age bins across the lifespan.

Our aims are to identify and compare Differentially methylated probes (DMPs) and regions (DMRs) associated to FEV1 and COPD across lifespan between age groups and calculate Methylation Risk Scores (MRS) in relation to lung function in two different context, general population, and COPD patients, to be able to classify or predict COPD and its severity. Linear models adjusted by confounding factors (such as age sex, smoking status, packs/years, and blood counts) will be used. All analyses will be done in R, scripts will be circulated, and results will be meta-analyzed in the coordinating group.

Before getting started:
We recommend to download the compressed file (**MRS_CADSET_workingDirectory.zip**) that you will find in the email and which includes following subfolders and files:

/data
/scripts/**1.create_clinical_df.R**
2.get_cohort_summary_stats.R
3.run_limma.R
4.run_robust_regressions.R
5.run_robust_regressions_per_age_bins.R
6.run_dmrcate.R
7.get_methylation_scores.R
8.analize_metthylation_scores.R
mrs/ **mrs_population_fevlt.csv**
mrs/**mrs_copd_fevlt.csv**
mrs**/mrs_copd_severity.csv**
/**results**/summary-stats
/limma
/rlm
/dmr
/mrs
/insights

The OneDrive link attached in the email will direct you to the folder: **[cohort]_MRS_cadsetproject_2023** where you should upload the folder /**results** once the analysis is done.

We also kindly ask to fill up the excel file **general_information.xlsx** attached in the email with some general information about the cohort and the methodology used for the methylation analysis.

· Analysis plan

The scripts provided have been built based on the use of Centos Linux OS. It includes R code.

1. **Creating clinical table:**
   a) Creation of a dataframe with the necessary clinical data. This table will be used for the different subsequent analysis but will not be shared [**1.create_clinical_df.R**].

2. **Getting summary statistics of the cohort:**
   a) Statistical analysis of clinical data of the cohort necessary for the final metanalysis [**2.get_cohort_summary_stats.R**].

3. **DMPs identification:**
   a) Analysis using linear model (Limma) with CpGs as response variable [**3.run_limma.R**].
   b) Analysis using robust linear model (rlm) with CpGs as independent variable [**4.run_robust_regressions.R** and **5.run_robust_regressions_per_age_bins.R**]

NOTE: Since differences in cell populations are a crucial characteristic of COPD patients, we have decided to perform the analysis with and without cell populations as covariate, so both analyses are included in the script.

4. **DMRs identification:**
   a) Identification of differentially methylated regions (DMRs) using a kernel-smoothed estimateAnalysis (DMRcate) [**6.run_dmrcate.R**].

NOTE: Since differences in cell populations are a crucial characteristic of COPD patients, we have decided to perform the analysis with and without cell populations as covariate, so both analyses are included in the script.

5. **MRS calculation:**
   a) MRS on FEV1(ml) from Lee et al.

   The most significant CpGs associated with FEV1(ml), reported by Lee *et al.* 2019 (PMID: 35536696) metanalysis have been used for the calculation of the MRS of lung function on general population. Robust regression of FEV1(L) was adjusted by age, age2, height, height2, sex, smoking status, smoking pack-years and cell populations. CpGs that had a p-value < 1e-08 were selected, which gave us a total of 64 CpGs. The CpGs are included in the provided file: **mrs_population_fevlt.csv**

   b) MRS on FEV1(ml) from COPD-BCN Blood

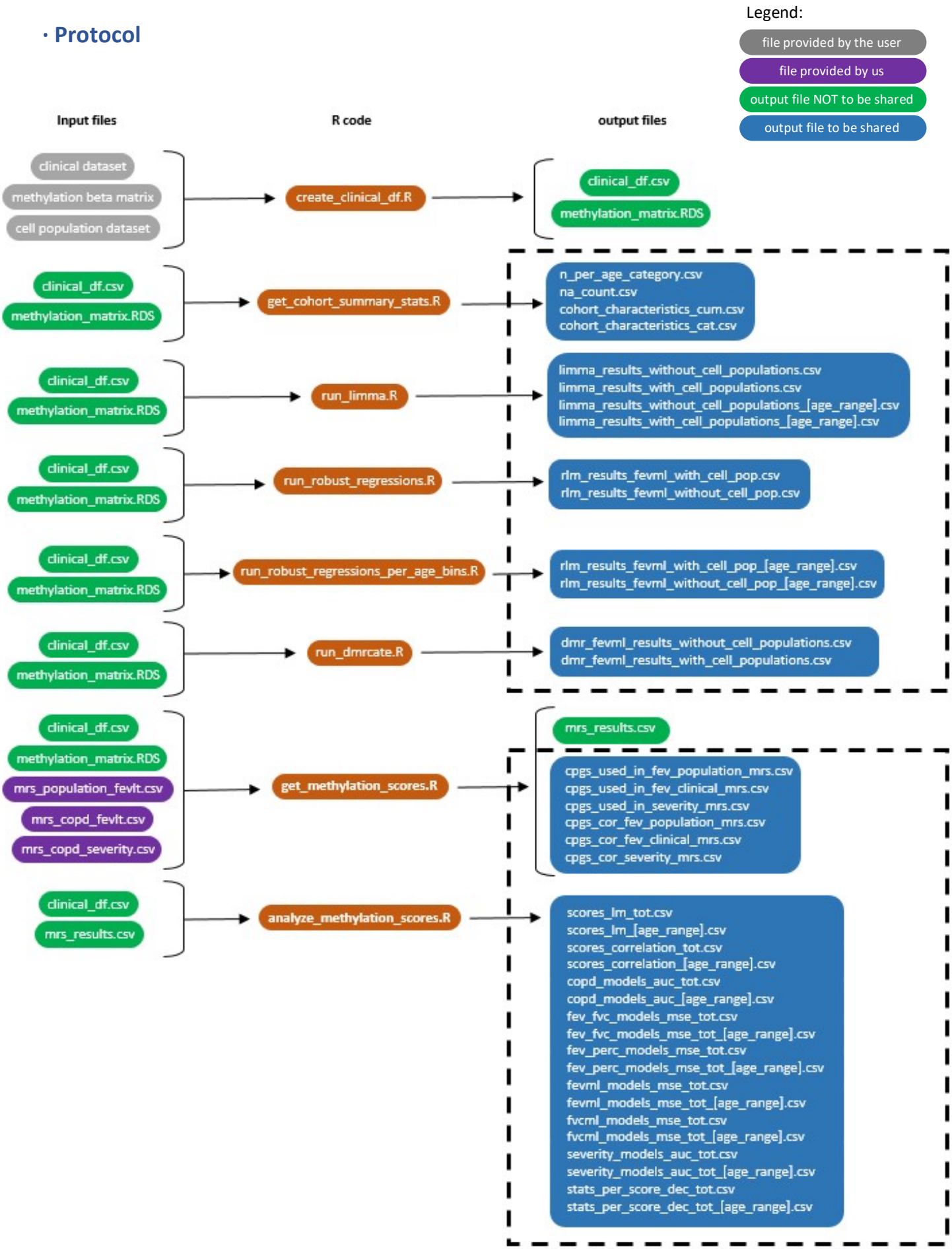
   The most significant CpGs associated with FEV1(ml), extracted from our COPD-BCN Blood cohort have been used for the calculation of the MRS of lung function on COPD patients. Robust regression of FEV1(L) was adjusted by age, age2, height, height2, sex, smoking status, smoking pack-years (and cell populations*). The top 63 CpGs were included based on model selection. The CpGs are included in the provided file: **mrs_copd_fevlt.csv**

   c) MRS on severity from COPD-BCN Blood

   The most significant differentially methylated CpGs between COPD patients GOLD Grade 1-2 and COPD patients GOLD Grade 3-4, extracted from our COPD-BCN Blood cohort have been used for the calculation of the MRS of severity on COPD patients. A binomial GLM of severity was adjusted by age, age2, height, height2, sex, smoking status, smoking pack-years (and cell populations*). The top 32 CpGs were included in the MRS based on model selection. The CpGs are included in the provided file: **mrs_copd_severity.csv**
   [**7.get_methylation_scores.R**]

6. **MRS analysis:**
   a) Analysis of the association of the MRS with different lung function parameters [**8.analize_metthylation_scores.R**].

· Protocol


Legend:
file provided by the user
file provided by us
output file NOT to be shared
output file to be shared
Input files
R code
output files
clinical dataset
methylation beta matrix
cell population dataset
create_clinical_df.R
clinical_df.csv
methylation_matrix.RDS
clinical_df.csv
methylation_matrix.RDS
get_cohort_summary_stats.R
n_per_age_category.csv
na_count.csv
cohort_characteristics_cum.csv
cohort_characteristics_cat.csv
clinical_df.csv
methylation_matrix.RDS
run_limma.R
limma_results_without_cell_populations.csv
limma_results_with_cell_populations.csv
limma_results_without_cell_populations_[age_range].csv
limma_results_with_cell_populations_[age_range].csv
clinical_df.csv
methylation_matrix.RDS
run_robust_regressions.R
rlm_results_fevml_with_cell_pop.csv
rlm_results_fevml_without_cell_pop.csv
clinical_df.csv
methylation_matrix.RDS
run_robust_regressions_per_age_bins.R
rlm_results_fevml_with_cell_pop_[age_range].csv
rlm_results_fevml_without_cell_pop_[age_range].csv
clinical_df.csv
methylation_matrix.RDS
run_dmrcate.R
dmr_fevml_results_without_cell_populations.csv
dmr_fevml_results_with_cell_populations.csv
clinical_df.csv
methylation_matrix.RDS
mrs_population_fevlt.csv
mrs_copd_fevlt.csv
mrs_copd_severity.csv
get_methylation_scores.R
mrs_results.csv
cpgs_used_in_fev_population_mrs.csv
cpgs_used_in_fev_clinical_mrs.csv
cpgs_used_in_severity_mrs.csv
cpgs_cor_fev_population_mrs.csv
cpgs_cor_fev_clinical_mrs.csv
cpgs_cor_severity_mrs.csv
clinical_df.csv
mrs_results.csv
analyze_methylation_scores.R
scores_lm_tot.csv
scores_lm_[age_range].csv
scores_correlation_tot.csv
scores_correlation_[age_range].csv
copd_models_auc_tot.csv
copd_models_auc_[age_range].csv
fev_fvc_models_mse_tot.csv
fev_fvc_models_mse_tot_[age_range].csv
fev_perc_models_mse_tot.csv
fev_perc_models_mse_tot_[age_range].csv
fevml_models_mse_tot.csv
fevml_models_mse_tot_[age_range].csv
fvcml_models_mse_tot.csv
fvcml_models_mse_tot_[age_range].csv
severity_models_auc_tot.csv
severity_models_auc_tot_[age_range].csv
stats_per_score_dec_tot.csv
stats_per_score_dec_tot_[age_range].csv

1. Creating clinical table:

Script needed for this step: **1.create_clinical_df.R**
Input files: **clinical dataset** (specific for each cohort)
cell populations dataset (specific for each cohort)
methylation data (beta matrix) (specific for each cohort)
Output file: **methylation_matrix.RDS** (To not share!!!)
clinical_df.csv (To not share!!!)

NOTE 1: If your methylation matrix contains M values instead of beta values you will have to run an extra line in the script to convert them.

NOTE 2: If your methylation data is not corrected by batch effects add in the clinical dataset the technical variance factors (ex.plate,slide…) or the top principal components (PC) inferred based on control probes.

NOTE 3: FEV1 measurements need to be in milliliter "mL" units.

2. Getting summary statistics of the cohort:

Script needed for this step: **2.get_cohort_summary_stats.R**
Input files: **clinical_df.csv**
Steps:

1. Get number of patients per age category.
2. Get number of NAs per variable.
3. Get statistics of each clinical parameter in the whole cohort.
4. Get statistics of each clinical parameter per age category.

Output file: **n_per_age_category.csv**
na_count.csv
cohort_characteristics_num.csv
cohort_characteristics_cat.csv

3. DMPs identification:

a) Linear model (limma)

Script needed for this step: **3.run_limma.R**
Input files: **clinical_df.csv**
methylation_matrix.RDS
Steps:

1. Run limma model without cell populations as covariate, using all cohort.
2. Run limma model with cell populations as covariates, using all cohort.
3. Run limma model without cell populations as covariate, in each age bin.
4. Run limma model with cell populations as covariates, in each age bin.

Output file: **limma_results_without_cell_populations.csv**
limma_results_with_cell_populations.csv
limma_results_without_cell_populations_[age_range].csv (one for each age range)
limma_results_with_cell_populations_[age_range].csv (one for each age range)

NOTE: If your methylation data is not corrected by batch effects add into the model (as independent variables) the technical variance factors(ex.plate,slide…) or the top principal components (PC) inferred based on control probes.

b) Robust linear model (rlm)

First script needed for this step: **4.run_robust_regressions.R**
Input files: **clinical_df.csv**

methylation_matrix.RDS

Steps:

1. Run rlm model with cell populations as covariate, using all cohort.
2. Run limma model without cell populations as covariates, using all cohort.

Output file: **rlm_results_fevml_with_cell_pop.csv**
rlm_results_fevml_without_cell_pop.csv

Second script needed for this step: **5.run_robust_regressions_per_age_bins.R**
Input files: **clinical_df.csv**
methylation_matrix.RDS

NOTE: If your methylation data is not corrected by batch effects add into the model (as independent variables) the top principal components (PC) inferred based on control probes.

Steps:

1. Run rlm model with cell populations as covariate, for each age range.
2. Run limma model without cell populations as covariates, for each age range.

Output file: **rlm_results_fevml_with_cell_pop_[age_range].csv** (one for each age range)
rlm_results_fevml_without_cell_pop_[age_range].csv (one for each age range)

NOTE: This step is the most computational demanding and uses parallelization to speed up computing time. Resources needed will depend on your number of subjects and array utilized (450K or EPIC). As reference, it took us around 3 hours to run a model with 300 subjects using 800K probes on a 12 cores computer consuming around 13GB of RAM. If you find any issue or are unable to use parallelization, please let us know.

4. DMRs identification:

This script needs to be run only if your cohort include individuals >12 years old (Adolescents/Adults)

Script needed for this step: **6.run_dmrcate.R**
Input files: **clinical_df.csv**
methylation_matrix.RDS

Steps:

1. Run DMRcate model without cell populations as covariates, using all cohort.
2. Run DMRcate model with cell populations as covariates, using all cohort.

Output file: **dmr_fevml_results_without_cell_populations.csv**
dmr_fevml_results_with_cell_populations.csv

NOTE: Be aware it is possible to not get any file if no significant DMRs are found.

5. MRS calculation:

The following MRS files are included in the "/**results**/mrs" folder:

mrs_population_fevlt.csv
mrs_copd_fevlt.csv
mrs_copd_severity.csv

Script needed for this step: **7.get_methylation_scores.R**
Input files: **clinical_df.csv**
methylation_matrix.RDS
mrs_population_fevlt.csv
mrs_copd_fevlt.csv
mrs_copd_severity.csv

Steps:

5. Load cpgs associated to each of the MRS
6. Calculate methylation risk scores using the 3 lists of MRS

7. Identify the CpGs of each MRS present and used in the cohort.

Output file: **mrs_results.csv**
cpgs_used_in_fev_population_mrs.csv
cpgs_used_in_fev_clinical_mrs.csv
cpgs_used_in_severity_mrs.csv
cpgs_cor_fev_population_mrs.csv
cpgs_cor_fev_clinical_mrs.csv
cpgs_cor_severity_mrs.csv

6. MRS analysis:

Script needed for this step: **8.analize_methylation_scores.R**
Input files: **clinical_df.csv**
mrs_results.csv

Steps:

1. Get the association of each MRS with the clinical variables.
2. Get the association of each MRS with the clinical variables, per age range.
3. Separate scores in deciles and get proportions of patients with COPD in each decil.
4. Compute a model for each age bin to see if adding the MRS improves COPD prediction.

Output file: **scores_lm_tot.csv**
scores_lm_[age_range].csv (one for each age range)
scores_correlation_tot.csv
scores_correlation_[age_range].csv (one for each age range)
copd_models_auc_tot.csv
copd_models_auc_[age_range].csv (one for each age range)
fev_fvc_models_mse_tot.csv
fev_fvc_models_mse_tot_[age_range].csv (one for each age range)
fev_perc_models_mse_tot.csv
fev_perc_models_mse_tot_[age_range].csv (one for each age range)
fevml_models_mse_tot.csv
fevml_models_mse_tot_[age_range].csv (one for each age range)
fvcml_models_mse_tot.csv
fvcml_models_mse_tot_[age_range].csv (one for each age range)
severity_models_auc_tot.csv
severity_models_auc_tot_[age_range].csv (one for each age range)
stats_per_score_dec_tot.csv
stats_per_score_dec_tot_[age_range].csv (one for each age range)

· Data sharing

The output files marked above in blue and stored in the **/results** folder are the ones to be shared with the coordinating team including them in the provided OneDrive folder **[cohort]_MRS_cadsetproject_2023.** Please, also include the excel file **general_information.xlsx.**

Do not hesitate to inform us if something is not clear or if you encounter any problem with the scripts, we will be happy to assist you. Once the results have been uploaded, we kindly ask you to send us an email to any of the subsequent addresses: sacasas@recerca.clinic.cat or viglino@recerca.clinic.cat

We are going to perform the following metanalysis with the results obtained from all cohorts:

Metanalysis:

- DMPs from all cohorts
- DMPs by age group

- DMRs from all cohorts

- MRS association with FEV1/FVC ratio from all cohorts
- MRS association with FEV1(ml) from all cohorts
- MRS association with FEV1% pred. from all cohorts
- MRS association with FVC(ml) from all cohorts

- MRS association with FEV1/FVC ratio by age group
- MRS association with FEV1(ml) by age group
- MRS association with FEV1% pred. by age group
- MRS association with FVC(ml) by age group

- MRS performance on FEV1/FVC ratio prediction from all cohorts
- MRS performance on FEV1(ml) prediction from all cohorts
- MRS performance on FEV1% pred. prediction from all cohorts
- MRS performance on FVC(ml) prediction from all cohorts
- MRS performance on COPD prediction from all cohorts
- MRS performance on COPD severity prediction from all cohorts

- MRS performance on FEV1/FVC ratio prediction by age group
- MRS performance on FEV1(ml) prediction by age group
- MRS performance on FEV1% pred. prediction by age group
- MRS performance on FVC(ml) prediction by age group
- MRS performance on COPD prediction by age group
- MRS performance on COPD severity prediction by age group