

Contract no. 101057548

EPIVINF

Epigenetic regulation of host factors in viral infections

D2.1 List of epigenetic markers predictive of viral load (HIV) and disease course (COVID)

ACTION: Research & Innovation Action (RIA)

CALL: HORIZON-HLTH-2021-DISEASE-04

TOPIC: HORIZON-HLTH-2021-DISEASE-04-07

Due Date of Deliverable	31/08/2025	Completion Date of Deliverable	07/08/2025
Deliverable leading partner	IRSICAIXA	Author	IRSICAIXA
WP Nº	2	WP Title	Epigenetic signatures in HIV and SARS-CoV-2 infection and impact on virus-specific host immunity and virus control

Project starting date	01/09/2022	Project Duration	60 months
-----------------------	------------	------------------	-----------

Dissemination Level	PU	Public	✓
	SEN	Sensitive	

Copyright

© Copyright IRSICAIXA

This document has been produced within the scope of the EPIVINF Project and is confidential to the Project's participants. The utilization and release of this document is subject to the conditions of the contract within the Horizon Europe Programme, contract no.101057548. The text represents the authors' views and does not necessarily represent a position of the Commission which will not be liable for the use made of such information.



This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement Nº 101057548

TABLE OF CONTENTS

1. INTRODUCTION / EXECUTIVE SUMMARY	3
2. CONTENT.....	4
2.1 HIV	4
2.1.1 HIV STUDY COHORT	4
2.1.2 RNA-SEQ IN HIV.....	6
2.1.3 DNA-METHYLATION IN HIV	8
2.1.4 Differentially expressed genes under epigenetic regulation Genes	9
2.1.5 FURTHER STEPS_ VALIDATION STUDIES	11
2.2 SARS-COV-2	12
2.2.1 SARS-COV2 STUDY COHORT.....	12
2.2.1 RNA-SEQ IN SARS-COV2	14
2.2.2 DNA-METHYLATION IN SARS-COV2	16
2.2.1 FURTHER STEPS and VALIDATION STUDIES.....	17

1. INTRODUCTION / EXECUTIVE SUMMARY

Untreated HIV infection leads to immune decline and manifestation of opportunistic infections, including cancers linked to viruses like Epstein-Barr Virus (EBV). Even HIV patients treated with antiretroviral therapy may experience lasting immune impairments and neurodegeneration, similar to post-acute effects seen in SARS-CoV-2 infection. Both viruses can cause early immune disruption and may leave long-term impacts through stable epigenetic changes, particularly DNA methylation, that impair immune control over co-pathogens, such as EBV.

In WP2, EPIVINF aims to identify and compare epigenetic changes in immune cells caused by HIV and SARS-CoV-2 infections, and how these relate to disease severity and immune dysfunction. We use genome-wide methylation analysis, transcriptomics and 10X sequencing, and proteomics to explore how these changes affect immune response to these as well as co-pathogens. The goal is to uncover shared epigenetic signatures that can be targeted by personalized therapies, including potential “epidrugs,” which may benefit a range of immune-related disorders.

Our studies in acute HIV and SARS-CoV-2 infection have identified several epigenetically dysregulated factors and pathways, and have revealed potentially drugable targets, among them host factors involved in the overall epigenetic cascade. These analyses may enable the use of some 'epidrugs' for these infections and a broader range of disorders, beyond strategies that focus solely on highly disease-specific markers.

2. CONTENT

Our studies in acute HIV and SARS-CoV-2 infection have identified multiple epigenetically dysregulated host factors and pathways, indicating significant and potentially long-lasting impacts on immune regulation. These changes, particularly in DNA methylation patterns, suggest a disruption of gene expression programs essential for effective antiviral responses and immune surveillance. Importantly, we have identified several candidate targets among host factors involved in the broader epigenetic machinery (mainly in HIV but also emerging from the studies in SARS-CoV-2 infection), providing a rationale for the use of "epidrug"—therapies that modulate epigenetic states—as a novel strategy with broad applicability beyond highly disease-specific interventions.

Here, we detail the patient cohorts used for the study of DNA methylation changes in the context of HIV and SARS-CoV-2 infection and the data obtained from these analyses. The cohorts include longitudinal samples from individuals at early stages of infection, as well as with varying degrees of disease severity (SARS-CoV-2), to capture a wide range of epigenetic alterations.

We also outline the methodology employed for identifying relevant pathways, which involves genome-wide DNA methylation profiling of peripheral blood mononuclear cell (PBMC) subsets, cell transcriptomics and bioinformatic analyses.

In addition, here we update the current progress and future directions for validation analysis and next steps, including work focussing on validating the identified targets through functional assays and cross-cohort comparisons.

2.1 HIV

2.1.1 HIV STUDY COHORT

2.1.1.1 LONGITUDINAL OMICS COHORT FOR EVALUATING HIV-ASSOCIATED EPIGENETIC CHANGES.

Taking advantage of the extensive HIV cohort at IRSI Partner, from the CHECK-EAR cohort in Barcelona with 3-monthly sampling intervals, we have currently 18 individuals with samples stored from before and post HIV infection. For a comprehensive multi-omic study, we have included 5 individuals on whom pre-HIV infection samples (Timepoint 1, T1), post-acute (Timepoint 2, T2) and after one year on ART treatment (Timepoint 3, T3) samples were stored. Clinical parameters and experimental design are shown in **Table 1** and **Figure 1**.

Table 1. Clinical data of HIV participants

PATIENT ID	Estimated HIV acquisition	1st HIV diagnose	Initi ART	Date	Rapid 4th Test	CLIA	LIA (Confirmatory)	Immunoblot bands (NOT WB)	pVL	SAMPLE ID	DataExtrac	VL at sample	CD4 at sample
1				21/10/14		Weak	Inconclusive		746.989	2013070100	2/7/13	N.A	N.A
1	24/9/14	21/10/14	29/10/14	29/10/14		Pos		gp41, p24 (undeterm)	177.894	2014100276	29/10/14	177894	244
1										2016110299	28/11/16	<40	689
2				11/2/14	Pos					2012100172	X/10/2012	N.A	N.A
2	30/10/13	11/2/14	25/7/14	14/2/14		Pos		gp120, gp41, p31, p24	74.853	2014060109	13/6/14	154816	1207
2										2015110168	13/11/15	<40	1766
3				23/12/10	Pos	Pos				2010070354	20/7/10	N.A	N.A
3	24/10/10	21/12/10	4/2/11	29/12/10		Pos		gp120, gp41, p24	98.000	2011010118	14/1/11	130615	403
3										2012100043	2/10/12	<40	472
4				10/1/12	Pos	Pos (via EHR)				2011060217	15/6/11	N.A	N.A
4	1/10/11	10/1/12	1/2/13	13/1/12				gp120, gp41, p31, p24, p17	420.000	2012020185	13/2/12	420.000	556
4										2014030416	25/3/14	<40	699
5				10-11/2/2011	Pos	Pos		gp41 (undeterm)	2.600.000	2010060044	1/6/10	N.A	N.A
5	17/1/11	10/2/11	22/2/11							2011020164	11/2/11	2.600.000	422
5										2011020420	25/2/11	2.600.000	422
5										2012060335	27/6/12	<40	551

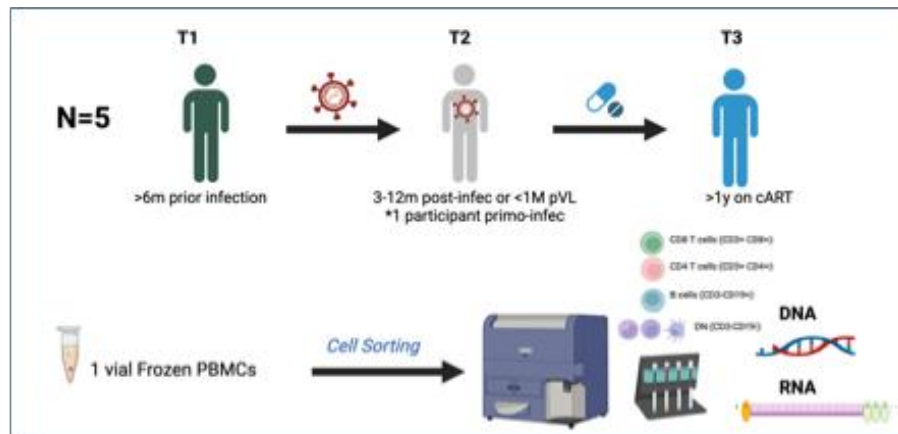


Figure 1. Study and experimental design

2.1.1.2 CELL SORTING

After thawing PBMCs, cells were stained for surface markers to identify T cells (CD3⁺CD4⁺ and CD3⁺CD8⁺), B cells (CD3⁺CD19⁺), and non-T/B cells (CD3⁻CD19⁻) for subsequent sorting using a FACS Aria sorter (**Figure 1**). The median cell yields obtained after sorting across the three timepoints and five participants were as follows: 6.75×10^5 CD4⁺ T cells, 1.13×10^6 CD8⁺ T cells, 2.34×10^5 B cells, and 7.12×10^5 non-T/B cells.

2.1.1.3 RNA/DNA EXTRACTIONS

From dry cell pellets of the sorted populations, simultaneous RNA and DNA extractions were performed using RNA/DNA purification kit (Norgene Biotek®, Canada), for subsequent DNAmethylation and RNAseq analysis.

RNA concentration was measured using Qubit RNA HS Assay, and RNA quality and fragment distribution were assessed using the Fragment Analyzer system with the High Sensitivity RNA 15 nt reagent kit, suitable for detecting small or low-abundance RNA species.

DNA concentration was measured using the Quant-iT High Sensitivity dsDNA Assay Kit (Thermo Fisher Scientific) on a Qubit fluorometer. DNA integrity was assessed using the Genomic DNA 50 kb Kit (DNF-467) on the Fragment Analyzer (Agilent), and DNA Integrity Numbers (DINs) were calculated automatically.

The quantity and quality of nucleic acids obtained from each sorted cell population were as follows. For **DNA**, median yields and integrity (DIN) were: CD4⁺ T cells – 1.0 µg (DIN: 8.7), CD8⁺ T cells – 1.1 µg (DIN: 9.1), B cells – 0.3 µg (DIN: 8.45), and non-T/B cells – 0.75 µg (DIN: 8.55). For **RNA**, median yields and integrity (RIN) were: CD4⁺ T cells – 0.44 µg (RIN: 9.6), CD8⁺ T cells – 0.7 µg (RIN: 9.8), B cells – 0.16 µg (RIN: 9.2), and non-T/B cells – 0.4 µg (RIN: 5.7).

These values indicate that the isolation and sorting of immune cell subsets yielded sufficient quantities of high-quality DNA and RNA for downstream epigenomic and transcriptomic analyses. CD4⁺ and CD8⁺ T cells consistently provided the highest nucleic acid yields with excellent integrity scores (DIN and RIN > 8), ensuring robust data generation for both methylation and gene expression studies. While B cells and non-T/B cells yielded lower amounts of nucleic acids, their quality remained within acceptable ranges, except for RNA from non-T/B cells, which showed slightly reduced integrity (median RIN: 5.7), potentially reflecting biological heterogeneity or increased RNA degradation susceptibility in this cell population specially after thawing and sorting cryopreserved cells. All in all, these results confirm the suitability of the sorted cell populations for integrative multi-omics analyses, mainly in adaptive immune cells, and support the reliability of the data derived from each subset.

2.1.2 RNA-SEQ IN HIV

Stranded total RNA sequencing . For each sample, a range between of 50–200 ng of total RNA was used as input for library preparation using the Illumina TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold for ribosomal RNA depletion. Libraries were prepared according to the manufacturer’s protocol, maintaining strand specificity. Paired-end sequencing (2×150 bp) was carried out on the Illumina NovaSeq 6000 platform (S4 flow cell), aiming for a sequencing depth of ~30 million reads per sample.

2.1.2.1 RESULTS RNA-SEQ

After normalization and quality control, MDS (multidimensional scaling) plots revealed clear differences among cellular fractions. As expected, CD4⁺ and CD8⁺ T cells clustered more closely together (both being T cell subsets), whereas B cells and non-T/B cell populations appeared more distinct and segregated (**Figure 2**).

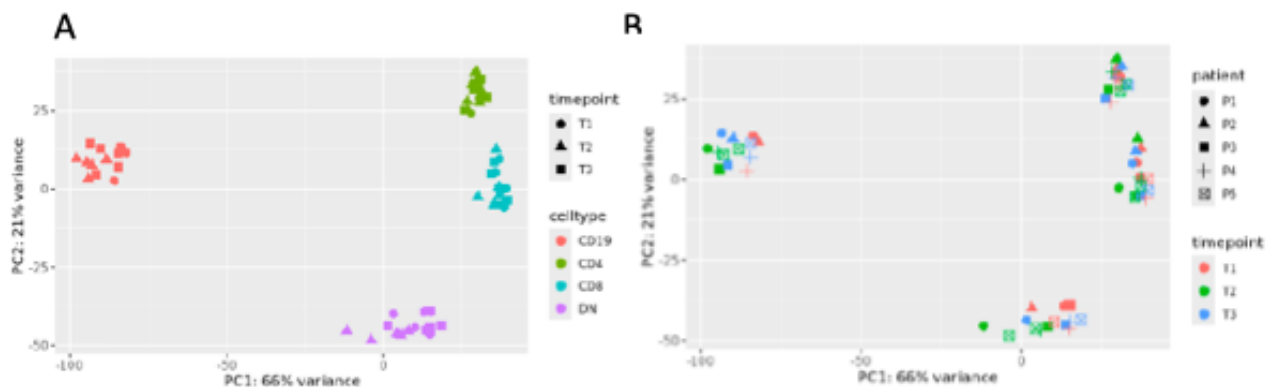


Figure 2. Multidimensional scaling (MDS) plots illustrating RNA-seq variability across samples. Samples are color-coded according to: A) cell type (CD4⁺ T cells, CD8⁺ T cells, B cells, and double-negative non-T/non-B cells [DN]) and time point (pre-infection, T1; post-infection, T2; and post-ART, T3), and B) time point and per patient.

Subsequent analysis using limma-voom pipeline was conducted. Pairwise comparisons were performed between sequential timepoints (T2 vs. T1, T3 vs. T2) and the overall comparison (T3 vs. T1) within each sorted cellular fraction (CD4⁺ T cells, CD8⁺ T cells, B cells, and non-T/B cells; **Figure 3**). Genes with low counts across all samples were filtered prior to analysis. Differentially expressed genes (DEGs) were identified using an empirical Bayes moderated t-test, and significance was defined as FDR < 0.2 with nominal $p < 0.05$ (Benjamini-Hochberg correction). Overall, a substantial number of genes were upregulated following infection (T1), with expression levels contracting after ART initiation, particularly in CD8⁺ T cells and B cells (Figure 3). Interestingly, although CD4⁺ T cells displayed the largest fold changes in the T2 vs. T1 comparison, they exhibited fewer DEGs overall, suggesting a more targeted transcriptional response.

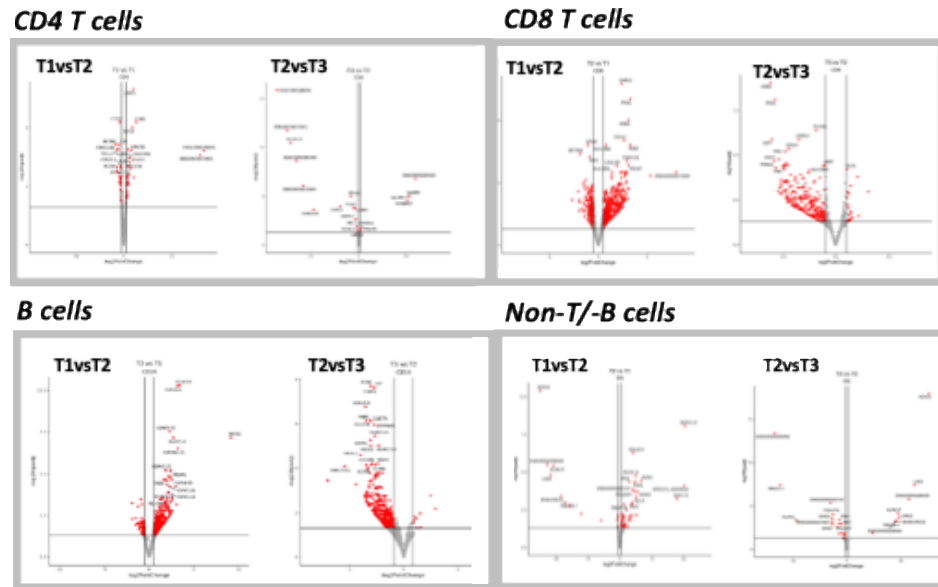


Figure 3. Volcano plots showing differential gene expression in CD4⁺ T cells (upper left), CD8⁺ T cells (upper right), B cells (lower left), and non-T/-B cells (lower right) for T2 vs T1 and T3 vs T2 comparisons. Differentially expressed genes (DEGs) were defined as $p < 0.05$. Gene upregulation was most pronounced after infection (T2 vs T1) and decreased following ART (T3 vs T2), particularly in CD8⁺ T cells and B cells.

Furthermore, CD4⁺ T cells showed the least restoration of gene expression profiles after ART initiation post-infection. Even after one year of virologic suppression under ART, many dysregulated pathways in CD4⁺ T cells remained uncorrected, suggesting persistent immune dysregulation in this compartment.

Then, Gene Set Enrichment Analysis (GSEA) was performed to identify dysregulated biological processes and pathways following infection, using Gene Ontology (GO) terms and the KEGG and Reactome databases. The analysis revealed that changes in CD4⁺ T cells were enriched for pathways associated with viral response, viral life cycle, interferon signalling, and Toll-like receptor signalling. CD8⁺ T cells showed enrichment for cell cycle and interferon-related pathways, while B cells were enriched for pathways related to B cell-mediated immunity and immunoglobulin production (Figure 4), all in all reflecting a strong adaptive immune reactivity to acute HIV infection.



Figure 4. Gene Set Enrichment Analysis (GSEA) of dysregulated biological processes and pathways following infection. Shown are the Top 10 upregulated pathways for the T2 vs T1 comparison (top panels) and the Top 10 downregulated pathways (bottom panels), assessed using Gene Ontology (GO) in CD4⁺ T cells (left), CD8⁺ T cells (middle), and B cells (right). In the dot plots, the x-axis indicates the gene ratio for each GO category (y-axis), the dot size reflects the number of genes per pathway, and the dot color represents the adjusted p-value.

Our findings reveal **cell type-specific transcriptional responses** to infection and upon ART initiation. While CD8⁺ T cells and B cells partially normalized after viral suppression, **CD4⁺ T cells remained transcriptionally dysregulated**. These results suggest that **incomplete restoration of CD4⁺ T cell transcriptional profiles underlies persistent immune dysfunction despite virologically effective and fully suppressive ART**.

2.1.3 DNA-METHYLATION IN HIV

Whole-genome enzymatic methyl-sequencing (WG-EM-seq) was performed on genomic DNA isolated from bulk-sorted cells using the NEBNext® Enzymatic Methyl-seq Kit (New England Biolabs), following the manufacturer's protocol. Briefly, approximately 50–200 ng of high-quality genomic DNA was enzymatically fragmented to ~300 bp, end-repaired, A-tailed, and ligated to methylated Illumina-compatible adapters. Unmethylated cytosines were enzymatically converted using TET2 and APOBEC enzymes, preserving methylated cytosines. After PCR amplification, libraries were quantified and quality-checked. Paired-end sequencing (2 × 150 bp) was performed on the Illumina NovaSeq 6000 platform (S4 flow cell), targeting ~99 Gb per sample.

2.1.3.1 RESULTS WG-EM-SEQ

After normalization and quality control, differential methylation analysis was performed to identify Differentially Methylated Regions (DMRs) across timepoints (T2 vs T1, T3 vs T2, and T3 vs T1) within each cellular fraction,

following the same comparative design used for the RNA-seq analysis. As shown in **Figure 5**, the frequency of DMRs across the different cell populations was relatively low, with only a limited number of regions showing significant alterations in hypomethylation or hypermethylation across time points (DMR defined as B-value difference < -0.2 or > 0.2 , with $p < 0.05$).

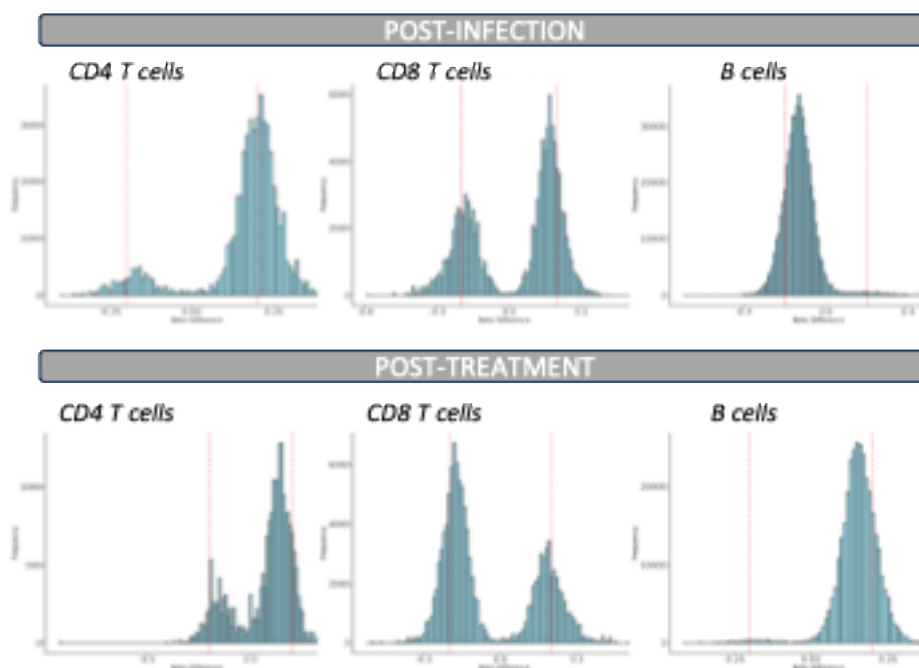


Figure 5. Distribution of Differentially Methylated Regions (DMRs) across cell types. Top panels show methylation changes post-infection (T2 vs T1) in CD4⁺ T cells (left), CD8⁺ T cells (middle), and B cells (right), while bottom panels show changes post-treatment (T3 vs T2). Histograms represent the frequency of significant regions (p -value < 0.05) and dotted red line indicates the hypomethylation or hypermethylation DMR showing a B-value difference < -0.2 or > 0.2 .

Interestingly, while the CD8⁺ T cell fraction displayed several genes with an inverse relationship between methylation status and gene expression—suggesting potential epigenetic regulatory effects at specific loci—in the B cell population, we detected a distinct shift in DNA methylation dynamics, characterized by gene-associated hypomethylation following infection and subsequent hypermethylation upon ART treatment. This opposing methylation pattern suggests a more robust and coordinated epigenetic response to infection and therapeutic intervention in CD8 T and B cells compared to the other cell types.

2.1.4 DIFFERENTIALLY EXPRESSED GENES UNDER EPIGENETIC REGULATION GENES

To explore the relationship between methylation changes and gene expression, we integrated RNAseq and WG-EM-seq datasets through correlation analysis. Basically, for each cell type and comparison, DMRs were mapped to proximal genes, linking epigenetic modifications to potential regulatory targets. Only those genes that were DEGs with an associated DMR were considered for correlation analyses performed between the level of the methylation change ($\Delta\beta$ -values) and gene expression change ($\log_2$ fold change) (**Figure 6**).

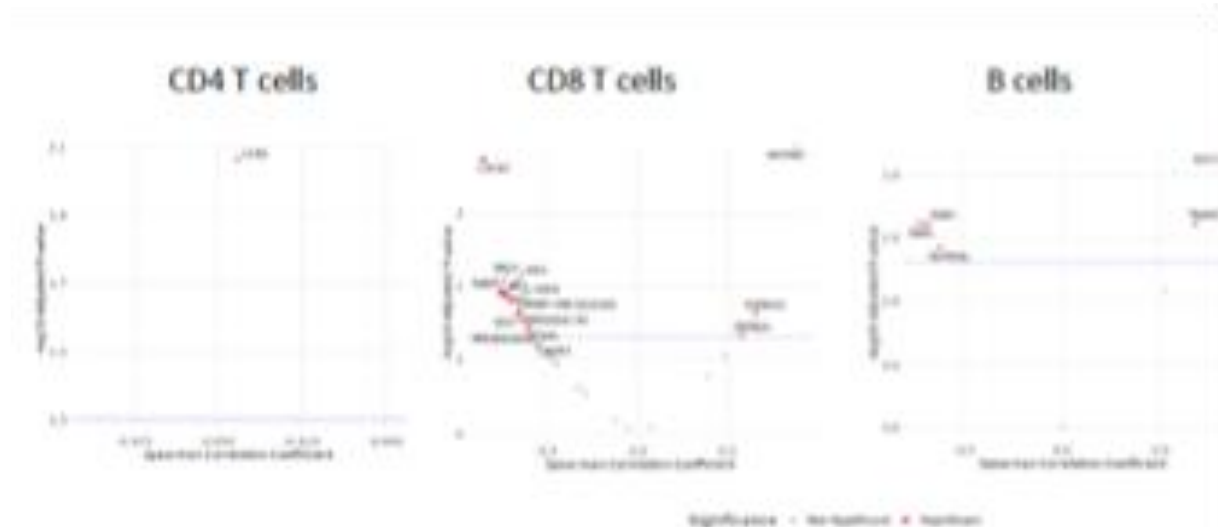


Figure 6. Integration of DNA methylation and gene expression changes across sorted immune cell populations. Scatter plots display the correlation between DMR-associated methylation changes ($\Delta\beta$) and gene expression changes ($\log_2$ fold change) across all timepoints in CD4⁺ T cells, CD8⁺ T cells, and B cells. The x-axis represents the Spearman correlation coefficient (Rho value), and the y-axis shows the $-\log_{10}$ adjusted p-value. Positive Rho values indicate that hypomethylation is associated with upregulation (or hypermethylation with downregulation), while negative Rho values suggest an inverse relationship.

Across the integrated analysis of DNA methylation and gene expression, CD8⁺ T cells exhibited the highest number of differentially expressed genes (DEGs) under apparent epigenetic regulation, as evidenced by a strong inverse correlation between promoter or proximal to the specific methylation sites close to the promoter and transcriptional activity. Specifically, genes showing hypomethylation were frequently upregulated, while hypermethylated regions were associated with gene downregulation, suggesting a dominant repressive role of DNA methylation in shaping the transcriptional response of CD8⁺ T cells to HIV infection and subsequent ART. In contrast, CD4⁺ T cells and B cells displayed fewer DEGs with significant methylation-expression correlations, indicating more limited epigenetic control over their transcriptional dynamics at this level.

Moreover, this study identifies key dysregulated genes whose expression changes are tightly linked to DNA methylation alterations in their promoter and flanking regions, underscoring the role of epigenetic regulation in modulating gene activity. Figure 6 illustrates these findings through heatmaps that integrate per-patient gene expression (RNA-seq) and DNA methylation ($\Delta\beta$) data, highlighting significant methylation-expression correlations in promoter and 1–5 kb flanking regions. The unsupervised clustering of samples further reveals distinct epigenetic and transcriptional patterns over multiple time points (T1, T2, T3), providing insights into the dynamic regulatory landscape. Together, these results confirm our hypothesis that methylation changes upon HIV infection influence gene dysregulation in immune cells, offering potential biomarkers and therapeutic targets for further investigation.

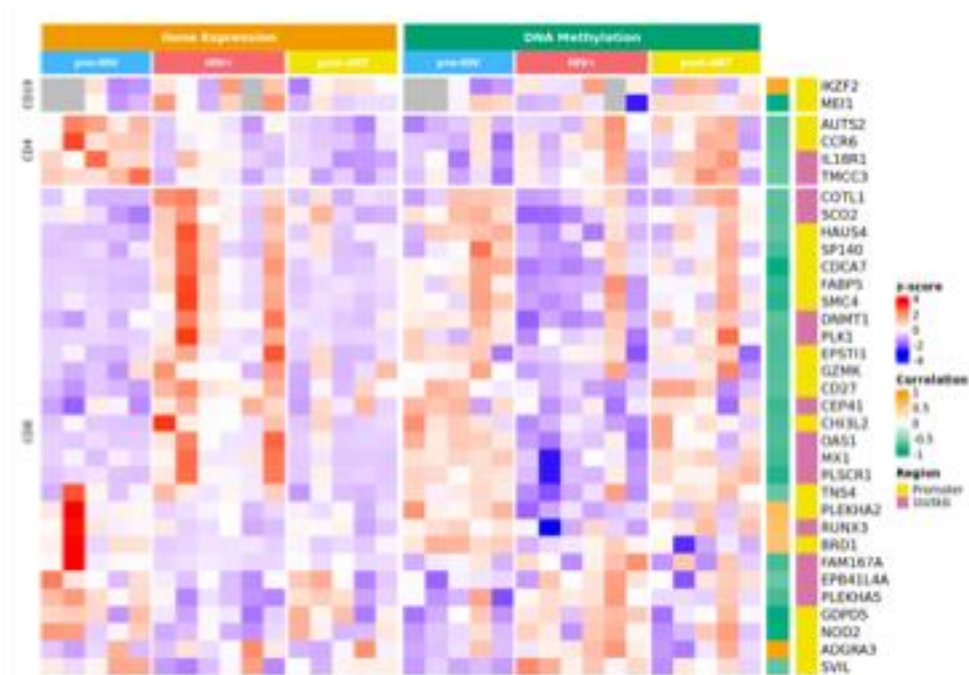


Figure 7. Dysregulated genes with correlated DNA methylation patterns in promoter and flanking regions. Supervised Heatmap depict per-patient gene expression levels and DNA methylation for selected genes (z-score) exhibiting significant methylation–expression correlations in promoter or 1–5 kb flanking regions. Samples are arranged by RNA-seq data from timepoints T1, T2, and T3 displayed first, followed by corresponding DNA methylation data (T1, T2, T3). An annotation column provides the z-score of expression, Spearman correlation coefficient (ρ), and the genomic region where the methylation correlation was identified (promoter or 1–5 kb flanking).

Integrating DNA methylation and transcriptomic data poses significant computational challenges, especially in pilot studies like this and clinical cohorts with limited sample sizes. Traditionally, differential methylation analyses that we have applied here, often require large statistical power to account for multiple testing across numerous CpG sites, which can limit the detection of meaningful signals. Aside from this approach, **Deliverable 6.1** details our developed framework overcoming these challenges by leveraging a targeted **Regulatory Region Set Enrichment Analysis** combined with a robust **Multi-Modal Data Integration Strategy**. This approach enhances the identification of biologically relevant methylation changes correlated with gene expression alterations by focusing on functionally relevant genomic regions and integrating complementary data types. Overall, our analysis methodology provides a scalable and sensitive framework for epigenetic-transcriptomic integration, paving the way for improved biomarker discovery and mechanistic insights in clinical research settings.

2.1.5 FURTHER STEPS_VALIDATION STUDIES

To ensure the robustness and clinical relevance of our findings, we have planned a multi-tiered validation strategy that combines independent cohort testing with orthogonal experimental approaches.

2.1.5.1 INTEGRATION OF 10x ATAC-SEQ DATA

Ongoing analyses focus on integrating **10x Genomics ATAC-seq** data to map chromatin accessibility profiles in the context of the identified methylation–expression correlations in same participants, timepoints and blood collection sample. This will enable the characterization of regulatory elements associated with our lead candidate genes and the identification of potential enhancer–promoter interactions driving dysregulation.

2.1.5.2 VALIDATION IN INDEPENDENT COHORTS

For external validation, we will analyse two independent and unrelated cohorts (the SABES and Merlin Studies [Lama et al., 2018](#); [Lama et al., 2020](#)) and the Peru High-Risk Cohort at IrsiCaixa) which represent a clinically enriched populations for testing candidate biomarkers. Lead genes identified from our integrative framework (**Figure 7 and D6.1**) will undergo experimental validation using pyrosequencing, enabling precise quantification of methylation levels at selected CpG sites. Current efforts focus on finalizing the list of top candidates for validation, after which we will perform methylation screening across both cohorts and evaluate the concordance with gene expression and ATAC-seq chromatin accessibility profiles. Together, these validation studies will confirm the reproducibility of our findings and strengthen the translational potential of our multi-modal epigenomic framework.

2.2 SARS-COV-2

2.2.1 SARS-COV2 STUDY COHORT

2.2.1.1 COHORT FOR LONGITUDINAL EVALUATION OF SARS-COV-2 IMPACT ON EPIGENETICALLY REGULATED FACTOR

A pilot longitudinal study was conducted in five age- and gender-matched participants to evaluate the impact of acute SARS-CoV-2 infection on host epigenetics. Participants and samples were sourced from Partner IRSI's healthcare worker cohort, originally established before 2019 and continuously followed during the COVID-19 pandemic as part of the KING cohort extension. This cohort provides precious pre-/post-infection samples, which are increasingly rare due to the rapid evolution of the pandemic, the emergence of viral variants, and the widespread impact of vaccination. Samples were collected at two time points: pre-2020 (pre-infection timepoint, T1) and one-month post-symptom onset (post-infection timepoint, T2). All cases corresponded to mild COVID-19 from the first wave of the pandemic, ensuring analysis in a context free from the confounding effects of vaccination or the pathophysiological diversity seen with infection by different strains. The data highlight the unique value of early native samples and demonstrates the feasibility of longitudinal multi-omics integration to capture epigenetic alterations associated with SARS-CoV-2 infection (**Table 2 and Figure 8**).

Table 2. Clinical data of SARS-COV-2 participants

Irsi_Name	Pre-Infection	Post-Infection	Age	Gender	Remdesivir	Tocilizumab	Other antivirals	Corticosteroids	Hydroxychloroquine	Antibiotics	Enoxaparin	Interferon B
Mild1	y	y	52	Male	N	N	N	N	N	N	N	N
Mild2	y	y	39	Female	N	N	N	N	N	N	N	N
Mild3	y	y	53	Female	N	N	N	N	N	N	N	N
Mild4	y	y	45	Male	N	N	N	N	N	N	N	N
Mild5	y	y	40	Male	N	N	N	N	N	N	N	N
Mild6	n	y	66	Female	N	N	N	N	N	N	N	N
Mild7	n	y	48	Male	N	N	N	N	N	N	N	N
Mild8	n	y	55	Female	N	N	N	N	N	N	N	N
Mild9	n	y	55	Female	Y	N	N	N	Y	Y	N	N
Severe2	n	y	56	Male	Y	N	N	N	Y	Y	Y	N
Severe4	n	y	29	Male	N	N	N	Y	Y	Y	Y	N
Severe6	n	y	64	Male	Y	N	Y	N	Y	Y	N	Y
Severe8	n	y	62	Male	Y	N	N	N	Y	Y	N	N
Severe9	n	y	73	Male	N	N	N	N	N	N	N	N
Severe11	n	y	76	Male	Y	Y	Y	N	Y	Y	Y	Y
Severe12	n	y	62	Male	Y	N	N	Y	N	N	Y	N
Severe13	n	y	59	Male	Y	N	Y	Y	Y	Y	Y	N
Severe14	n	y	47	Male	Y	Y	N	Y	N	N	Y	N
Severe15	n	y	44	Male	Y	N	N	Y	N	N	Y	N
Severe16	n	y	70	Male	Y	N	N	N	Y	Y	Y	N

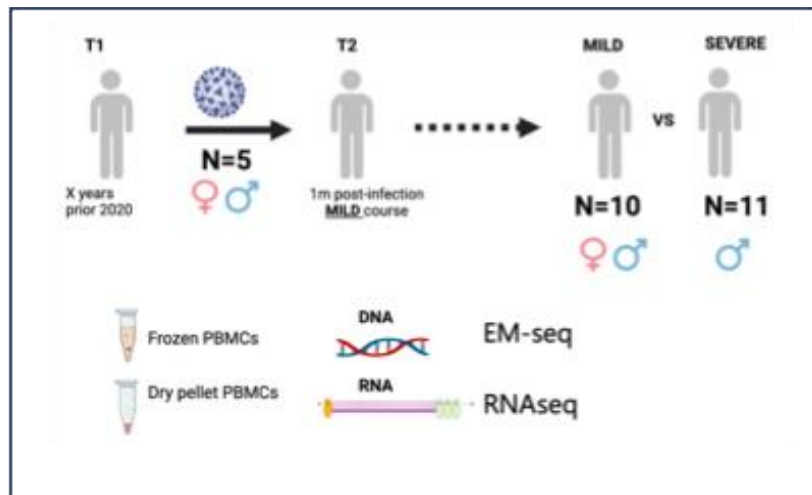


Figure 8. Study and experimental design

2.2.1.2 COHORT FOR CROSS-SECTIONAL OMICS ANALYSIS OF EPIGENETIC REGULATION IN COVID-19 SEVERITY

For the cross-sectional analysis of epigenetic regulation in COVID-19 severity, the initial group of five participants with mild disease course described above was expanded to include a total of 9 mild cases (MILD, 5 females and 4 males), which were compared with 11 severe cases (SEVERE, all male) (**Table 2 and Figure 8**). An inherent limitation of this study, reflecting the early pandemic context, is that all severe cases were male and slightly older. Although no statistically significant age differences were found between groups (MILD median age 52 years vs SEVERE median age 62 years, p -value=0.09 Mann Whitney), this demographic imbalance is important to consider when interpreting the results. This comparison provides valuable insights into the epigenetic distinctions associated with COVID-19 severity, while also highlighting challenges related to cohort composition during the initial pandemic wave. These learnings also influenced the further cohort compositions at the two other clinical sites with Partner SAAR and OSR.

2.2.1.3 SAMPLE SOURCE AND DNA/RNA EXTRACTION

From available stored 2 million dry-pellet and cryopreserve PBMCs, simultaneous RNA and DNA extractions were performed using RNA/DNA purification kit (Norgene Biotek®, Canada) for further DNA-methylation and RNAseq analysis.

RNA concentration, quality, and integrity were assessed using the Agilent BioAnalyzer RNA 6000 Nano kit. This assay provides RNA quantification and evaluates fragment distribution, including the 18S and 28S rRNA peaks, generating a RNA Integrity Number (RIN) suitable for standard RNA samples with moderate to high abundance.

DNA concentration was measured using the Qubit dsDNA Broad Range (BR) Assay Kit (Thermo Fisher Scientific) on a Qubit fluorometer. DNA integrity was evaluated by agarose gel electrophoresis and confirmed by qPCR amplification of human GAPDH and AMLX/Y targets.

Bulk PBMC samples provided sufficient quantities of high-quality DNA and RNA suitable for downstream epigenomic and transcriptomic applications. Overall, nucleic acids extracted from these samples displayed robust integrity, supporting reliable data generation for both methylation and gene expression studies. While some variability in RNA quality was observed, the overall integrity remained within acceptable ranges for multi-omics workflows. Collectively, these findings confirm that bulk PBMCs (cryopreserved or dry-pellet) are appropriate for integrative analyses and ensure the generation of reproducible and biologically meaningful data.

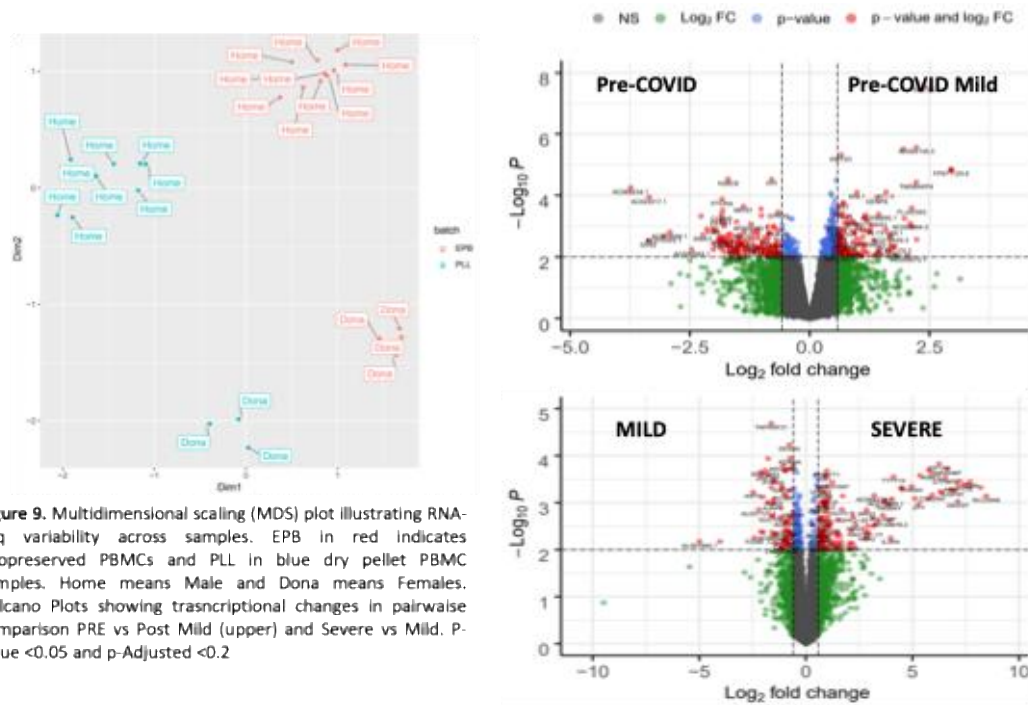
2.2.1 RNA-SEQ IN SARS-COV2

Stranded total RNA sequencing was performed on high-quality total RNA extracted from cryopreserved PBMCs and dry pellet PBMCs. RNA-Seq libraries were prepared using the KAPA Stranded mRNA-Seq Illumina® Platforms Kit (Roche), starting from 500 ng of total RNA. The poly-A RNA fraction was enriched by oligo-dT magnetic bead selection, followed by fragmentation of the mRNA. Strand specificity was maintained by performing second strand synthesis in the presence of dUTP instead of dTTP, ensuring strand-specific cDNA synthesis. The resulting blunt-ended double-stranded cDNA was adenylated at the 3' ends and ligated with Illumina-compatible adaptors containing unique dual indexes (UDIs) and unique molecular identifiers (UMIs) from Integrated DNA Technologies. The ligation products were amplified by PCR with 15 cycles to enrich the library. Libraries were sequenced on the Illumina NovaSeq 6000 platform using paired-end 2 × 50 bp reads, achieving a sequencing depth of over 30 million paired-end reads per sample.

RNA-seq data were processed and analyzed using the **edgeR** and **limma** R packages. Raw gene count matrices were first filtered to retain only genes with sufficient expression (total counts > 500 across all samples) and normalized using the trimmed mean of M-values (TMM) method implemented in **edgeR** to account for library size differences. Exploratory data analysis, including multidimensional scaling (MDS) plots, was performed with **ggplot2** and **ggrepel** to visualize sample clustering by biological and technical covariates such as batch, sex, and experimental group. A linear modeling framework was implemented using **limma-voom**, with duplicate correlation applied to account for repeated measures from the same individual. The design matrix included group and batch effects, and contrast matrices were constructed to test specific pairwise comparisons of interest (e.g., *Mild_post* vs. *Mild_pre* and *Severe* vs. *Mild*). Differential expression was assessed using moderated t-statistics with empirical Bayes shrinkage, and genes were ranked by adjusted p-values to identify significant expression changes. Normalized expression values and full differential expression results for each contrast were exported for downstream visualization and functional enrichment analyses.

2.2.1.1 RESULTS RNA-SEQ

After normalization and quality control, multidimensional scaling (MDS) plots revealed clear differences between dry pellet and cryopreserved PBMC samples, calling for the inclusion of this variable as a covariate alongside sex (**Figure 9**). Based on these identified variables, differential expression analyses were performed through pairwise comparisons, specifically POST vs. PRE infection within the Mild group and for SEVERE vs. MILD disease manifestation. Volcano plots were generated to visualize the transcriptional profiles, and differentially expressed genes (DEGs) were defined as those meeting $p < 0.05$ and $FDR < 0.2$ (**Figure 9**). These analyses demonstrated pronounced transcriptional changes in both comparisons, with the most substantial fold changes observed when comparing severe cases to the group with milder symptoms, highlighting the enhanced transcriptional dysregulation associated with severe COVID-19.



Gene Set Enrichment Analysis (GSEA) was performed to uncover significantly dysregulated biological pathways and processes following SARS-CoV-2 infection, using Gene Ontology (GO) annotations alongside curated pathway databases such as KEGG and Reactome. Differentially expressed genes (DEGs) comparing pre- and post-infection samples revealed significant enrichment in pathways central to cell cycle regulation, mitotic processes, cilium assembly, SUMOylation of DNA replication proteins, and viral mRNA translation (**Figure 10 Left**). These results indicate that SARS-CoV-2 infection disrupts fundamental cellular mechanisms responsible for cell division and genome integrity, while also hijacking host translational machinery to facilitate viral replication. The involvement of cilium assembly pathways may reflect viral impact on cellular structures critical for respiratory epithelial function and host defence.

When comparing Severe versus Mild COVID-19 cases, distinct enrichment patterns emerged in immune-related and signalling pathways, including CD22-mediated B cell receptor signalling, classical complement activation (notably through C4 and C2 activators), initial complement triggering, CDC20 regulation, and Notch signalling pathways regulated by RUNX3. The heightened activation of complement pathways and immune receptor signalling in severe cases suggests an exacerbated inflammatory and immune response that may contribute to immunopathology. Additionally, pathways linked to neuronal signalling, SUMOylation, cell cycle control, and NGF-stimulated transcription indicate broader impacts on cellular homeostasis and immune regulation during severe disease (**Figure 10 Right**). Together, these findings highlight key molecular mechanisms underlying pathophysiological outcomes of SARS-CoV-2 infection and provide insight into the differential host responses that may drive disease progression and severity.

PRE- and POST- SARS-COV-2

SEVERE vs MILD SARS-COV-2

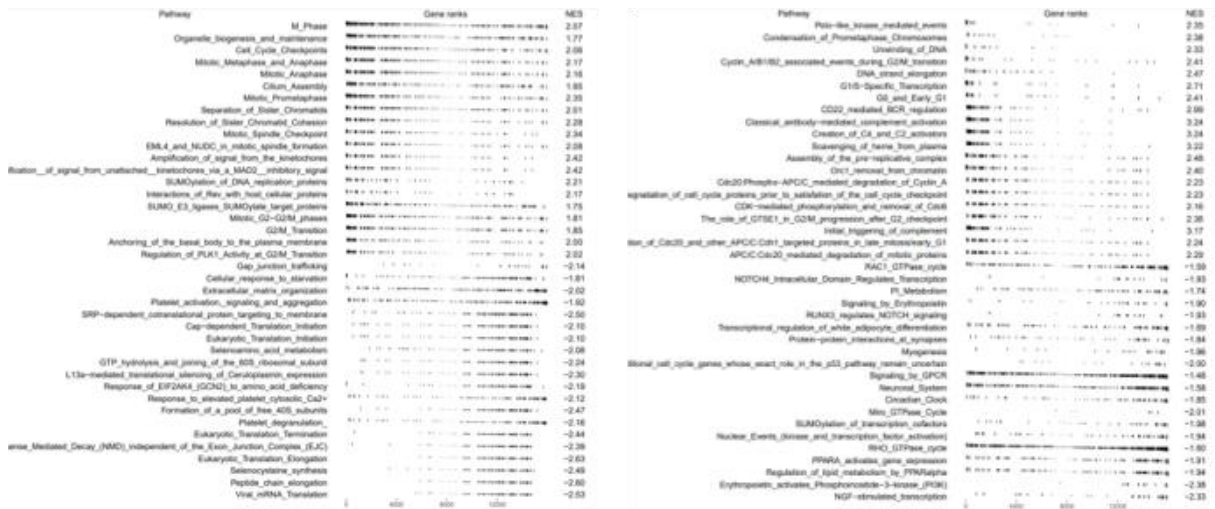


Figure 10. Gene Set Enrichment Analysis (GSEA) of dysregulated biological processes and pathways following infection (LEFT: pre- vs. post-infection timepoints during mild SARS-CoV-2 disease) and during the acute phase of the disease (RIGHT: severe vs. mild symptomatology).

2.2.2 DNA-METHYLATION IN SARS-COV2

Whole-genome enzymatic methyl-sequencing (WG-EM-seq) was performed on genomic DNA isolated from PBMCs and dry pellet PBMCs. Library preparation for comprehensive methylome profiling was performed using the NEBNext® Enzymatic Methyl-seq (EM-seq™) Kit, which preserves the integrity of methylated cytosines while enabling accurate detection across the genome. In this workflow, 10–200 ng of input DNA undergoes enzymatic conversion in which TET2 oxidizes 5-methylcytosine (5mC) to 5-carboxylcytosine (5caC) and the Oxidation Enhancer converts 5-hydroxymethylcytosine (5hmC) to 5-glycosylhydroxymethylcytosine (5ghmC). Unmethylated cytosines are subsequently deaminated to uracil by APOBEC and sequenced as thymines, while 5mC and 5hmC remain as cytosines, ensuring faithful methylation calling. Libraries were sequenced on the Illumina NovaSeq 6000 platform using 2 × 150 bp paired-end reads to achieve approximately 30× genome coverage (~99 Gb per sample), providing high-resolution methylation maps suitable for downstream epigenomic analyses.

Raw sequencing reads were first quality-checked and aligned to the reference genome using standard WGBS/EM-seq workflows. Methylation calls at CpG sites were extracted and compiled into BSseq objects using the bsseq R package, followed by filtering to remove low-coverage or low-quality sites, generating the *.bsseq.filter.rds datasets for downstream analyses. Smoothed methylation estimates were calculated to improve signal-to-noise ratios across the genome. Differential methylation analysis was then performed using the DSS package, which applies a Bayesian hierarchical model with dispersion shrinkage to identify differentially methylated loci (DMLs) and regions (DMRs). Filtered DSS objects (*.dss.filter.rds) were generated to store high-confidence results, which were subsequently used for visualization and integration with downstream epigenomic analyses.

2.2.2.1 RESULTS WG-EM-SEQ

Gene Set Enrichment Analysis revealed significant epigenetic and metabolic pathway dysregulation following SARS-CoV-2 infection. Notably, pathways involved in SUMOylation of DNA methylation proteins and NOTCH

signalling were dysregulated, indicating alterations in epigenetic regulation and cell fate control. Concurrently, metabolic processes including insulin secretion and lipid metabolism were markedly affected, reflecting systemic metabolic disruptions. Additionally, signalling pathways related to immune activation, platelet function, and extracellular matrix remodelling were enriched, consistent with inflammatory and tissue damage responses. In contrast, pathways linked to neuronal function and synaptic signalling present a negative NES (Normalized Enrichment Score), suggesting downregulation of potentially underlying neurological symptoms observed post-infection (**Figure 11 Left**). Together, these results suggest SARS-CoV-2 infection drives widespread epigenetic and metabolic reprogramming that impacts immune responses, cellular homeostasis, and neurobiology.

In comparing severe versus mild COVID-19 cases, key epigenetic and signalling pathways show notable dysregulation. NOTCH signalling pathways (NOTCH1, NOTCH3, NOTCH4) and their transcriptional regulators, including RUNX3, are enriched in severe cases, suggesting enhanced epigenetic control linked to altered cell differentiation and immune modulation. Concurrent activation of RHO GTPase pathways and mitotic processes indicates increased cytoskeletal remodelling and cell proliferation. In contrast, pathways related to B cell receptor signalling and anti-inflammatory IL-10 synthesis showed a negative NES suggesting a downregulation of gene expression, implying compromised adaptive immune responses in severe disease. Additionally, suppression of neuronal signalling pathways such as NMDA receptor activation and beta-catenin-mediated transcription may relate to neurological symptoms seen in severe COVID-19 (**Figure 11 Left**). Together, these results reflect a shift toward heightened cellular signalling and epigenetic remodelling alongside impaired immune regulation, potentially driving disease severity.



Figure 11. Gene Set Enrichment Analysis (GSEA) of EPIGENETICALLY dysregulated biological processes and pathways following infection (LEFT: pre- vs. post-infection timepoints during mild SARS-CoV-2 disease) and during the acute phase of the disease (RIGHT: severe vs. mild symptomatology).

2.2.1 FURTHER STEPS AND VALIDATION STUDIES

Our analyses show that integrating DNA methylation and transcriptomic data presents substantial computational challenges, particularly in clinical cohorts with limited sample sizes. Standard differential methylation analyses require considerable statistical power to correct for multiple testing across hundreds of thousands of CpG sites, which can hinder the detection of subtle yet biologically meaningful signals. To overcome these limitations, as detailed in Deliverable 6.1, we developed a targeted **Regulatory Region Set Enrichment Analysis** combined with a **robust Multi-Modal Data Integration Strategy**, initially validated in HIV studies and subsequently applied to

SARS-CoV-2 infection samples. By prioritizing functionally relevant regulatory regions and leveraging the complementary nature of methylation and expression data, our framework enhanced the identification of epigenetic alterations directly linked to transcriptional changes. This scalable and sensitive approach enables more reliable biomarker discovery and provides mechanistic insights into host-pathogen interactions in clinical research contexts for the identification of druggable therapeutic targets.

To ensure the robustness and clinical relevance of the findings, we will validate our results combining independent cohort testing with different experimental approaches as mentioned above in the context of Long-COVID (WP4), additional cohorts at geographically distinct sites (SAAR and OSR) and with animal models that have been established in WP5.