ID-DarkMatter-NCD deliverable report

D2.1 Genetic data

Lead beneficiary	7- BRC	Due Date	30th June 2025
WP no	2	New due date (if delay)	
Task no	2.1	Actual Delivery Date	30th June 2025
Dissemination level	Please specify: PU – Public	Status	FINAL DRAFT

Authors

Authors	Partner no	Partner organisation	Name of author
Main author	7	BRC	Máté Manczinger
Contributing author(s)	8	KI	Gunilla Karlsson Hedestam
	6	UKSH	Hesham El Abd

Review

Authors	Partner no	Partner organisation	Name of author
Technical review	1	MUW	Thomas Vogl
Language review – *if applicable*	3	EUTEMA	Mikael Muegge

Document history

Date	Version	Chapters affected	Description of change	Author	Document status
20th June	0.1	all	Draft version	BRC	DRAFT
25th June	0.2	all	Updated draft	BRC, KI, UKSH	DRAFT
30th June	1.0	all	Final Draft	BRC, KI, UKSH	FINAL

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

Table of Contents

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

1. Executive Summary

Objective

The overarching aim of this task (T2.1, lead: BRC) is to curate and assemble existing HLA and whole genome sequencing (WGS) datasets across project cohorts (participants: VHIR, KI, UKSH/CAU, LUMC, UMCG). The goal is to make these datasets interoperable and readily exchangeable to enable uniform comparison between cohorts and ensure compatibility with newly generated data. This includes the establishment of standardized data formats and pipelines for downstream analyses, such as HLA promiscuity and clustering. The work is also coordinated with T2.2 (HLA genotyping, lead: UKSH/CAU), which focuses on generating new genotyping data where needed, as well as with T7.6, which addresses ethical and legal aspects of data sharing.

Work conducted

1. **Assessment of available genotype data**

   Participating cohorts were evaluated for the availability and quality of HLA genotype data. For many cohorts, HLA genotyping had already been completed using SNP arrays or WGS/WES technologies (the latter used primarily for PGS and not for HLA calling). These data, typically at a two-field resolution, were deemed compatible with the project's analytical requirements.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

2. **Identification of cohorts requiring genotyping**

   We also identified cohorts lacking HLA genotype data or requiring additional genotyping. These cohorts will undergo genotyping using SNP arrays and imputation, a cost-effective and accurate method, especially in Caucasian populations.

3. **Definition of essential HLA data format**

   A standardized HLA genotype format was established, containing two alleles per individual at each of the following loci: HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, and -DRB1. A two-field resolution is considered the minimal threshold for inclusion in the downstream analytical pipeline.

4. **Pipeline integration and analysis capabilities**

   The pipeline developed by BRC processes this standardized data format efficiently. It constructs stable HLA-DQ alpha-beta complexes, identifies cis and trans variants, and determines all HLA-DP complexes per individual. The pipeline also computes HLA promiscuity at both locus and class levels and categorizes alleles and complexes into clusters based on binding preference similarities.

5. **Data exchange mechanisms and standardization**

   Mechanisms for data sharing among partners were established, focusing on secure, anonymized data transfer. In most cases, data exchange can be standardized across cohorts. However, several deviations were identified:

   - Some cohorts require that analyses be conducted via restricted online platforms due to data access policies.
   - A few datasets are subject to ethical or legal restrictions preventing anonymization, necessitating the use of protected, non-standardized formats and protocols.

Major Deviations

No major deviations have occurred. While the original plan assumed full standardization of data formats and exchange mechanisms, the actual implementation revealed several smaller exceptions. In particular, legal and ethical restrictions in some cohorts prevented full data harmonization. These deviations necessitate tailored solutions for specific cohorts, such as local

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

analysis or alternative data-sharing agreements, but do not impede overall project progress, and hence only represent minor issues.

Conclusion

The curation and harmonization of HLA/WGS data under T2.1 has made significant progress. The data from existing cohorts have largely been standardized, and protocols for new genotyping (T2.2) are in place to close the remaining gaps. The established genotype format and analysis pipeline enable efficient HLA promiscuity calculation and clustering. Despite a few necessary deviations in data handling, the overall structure supports consistent, comparative immunogenomic research across cohorts, fully aligned with the objectives stated in the original project proposal. This work lays the foundation for identifying risk alleles and immunological patterns relevant to infection-related non-communicable diseases (IR-NCDs), as envisioned in the HORIZON-HLTH-2023-DISEASE-03-07 call.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

2. Assessment of available genotype data / Identification of cohorts requiring genotyping

Disease	Cohort	N	HLA available	HLA needed
PCC	Linz		Yes	Yes
PCC	MUW (150 + controls)	~150 (+healthy controls)	No	Not yet
PCC	MUW (5000)	~5000	No	Not yet
ME/CFS	UMCG	~200	Partly	Partly
MS	UNIBAS	>1000	No	Yes
IBD	SU/Paris	>4000	No	Not yet
IBD	UMCG	~1000–1200	Yes	Yes
IBD	UMCG	~100 (Longitudinal fecal)	No	Not yet
IBD	UMCG	~4500	Yes	Yes
IBD	UKSH	~1400	Yes	No
IBD	UKSH	~700	Yes	No
IBD	UKSH	~2886 index + relatives	Yes	No
Healthy	UMCG	~1,000	Yes	Yes
Healthy	MUW	~5000	No	No
RA	Leiden	~3000	707 with HLAII and GWAS	300
RA	Barcelona	~1500	1250 (Imputed)	TBD
RA	Vienna	~1000	No (WGS for 400 RA)	No
SLE	Leiden	~550	No	TBD
SLE	Barcelona	~700	700	TBD
SLE	Vienna	235 confirmed	No	TBD
SSc	Basel	~1000	No	No
SSc	Leiden	~750	No	TBD
Myositis	Vienna	~123	No	TBD

Table 1. List of cohorts with information on HLA genotype availability. TBD: to be decided

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

2.1. Post-COVID Condition (PCC)

The Linz cohort already possesses HLA genotyping data, yet additional genotyping is still needed to complete the dataset. In contrast, the MUW cohort has two distinct sample groups: one comprising 150 patients plus healthy controls, and another consisting of 5000 samples. For both groups, HLA data is not currently available.

2.2. Myalgic Encephalomyelitis / Chronic Fatigue Syndrome (ME/CFS)

The UMCG cohort, with approximately 200 samples, has partial HLA genotyping data available. Correspondingly, additional genotyping is partially needed to complete the dataset.

2.3. Multiple Sclerosis (MS)

The UNIBAS MS cohort includes over 1000 participants. HLA genotyping is being carried out using HLA imputation. For the VHIR MS cohort, no data on HLA availability or genotyping needs has been provided at this stage.

2.4. Inflammatory Bowel Disease (IBD)

The SU/Paris cohort includes more than 4000 samples, but HLA genotyping is not available and no additional genotyping is planned currently. The UMCG hosts several IBD-related sub-cohorts. Among these, one with 1000–1200 samples already has HLA data available, and additional genotyping is also planned. Another sub-cohort focused on longitudinal weekly faecal sampling over one year (n=100) lacks HLA data and will not undergo genotyping yet. However, a larger UMCG IBD cohort comprising 4500 individuals does have HLA data and is also scheduled for further genotyping. At UKSH, three groups—one with 1400 samples, one with 700, and a larger cohort of 2886 index patients plus 2–4 relatives each—have HLA genotypes imputed from GWAS data. No additional genotyping is required because imputation from GWAS is considered sufficient.

2.5. Healthy Controls

The UMCG healthy cohort, estimated at around 1000 individuals, has existing HLA genotyping data and will also be subject to further genotyping efforts. In contrast, the MUW healthy cohort, comprising 5000 individuals, lacks HLA data and the need for HLA genotyping is to be determined.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

2.6. Rheumatoid Arthritis (RA)

In Leiden, up to 3000 RA samples are available. Of these, 707 already have HLA genotype data, and the dataset will be expanded to include 1000 genotyped individuals. An additional 300 samples will be genotyped to meet project needs. The Barcelona cohort has over 1500 RA patients, with HLA data imputed from GWAS for 1250 individuals. At Vienna, while more than 1000 samples are part of the RA cohort, no HLA data is currently available. However, whole genome sequencing has been performed on 400 patients, and the need for HLA genotyping is being determined.

2.7. Systemic Lupus Erythematosus (SLE)

The Leiden SLE cohort includes approximately 550 participants and currently lacks HLA genotyping data. Additional genotyping is considered desirable but has not yet been confirmed. In Barcelona, more than 700 individuals are included, with HLA data reportedly available for most individuals. The Vienna cohort has 235 confirmed cases with more expected; while HLA data is not yet available, the need for genotyping is to be determined.

2.8. Systemic Sclerosis (SSc)

The Basel cohort consists of around 1000 individuals. HLA genotyping is not yet carried out and its need will be determined later. At Leiden, roughly 750 samples are available, but no HLA genotyping has been performed. Additional genotyping is to be determined.

2.9. Myositis

The Vienna cohort for myositis includes 123 samples. No HLA genotyping has been performed to date and its need is being determined soon.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

3. HLA genotype imputation

HLA imputation is a computational technique used to predict classical HLA alleles from single nucleotide polymorphism (SNP) data obtained through whole genome genotyping arrays. Given the high polymorphism and dense linkage disequilibrium (LD) structure of the HLA region on chromosome 6, direct HLA typing is often expensive and technically challenging. Imputation offers a cost-effective, scalable alternative, enabling researchers to infer high-resolution HLA genotypes from commonly available SNP datasets.

HLA imputation is most accurate for common alleles and in populations that closely match the reference panel. Class I genes (HLA-A, -B, -C) and class II genes (HLA-DRB1, -DQA1, -DQB1, etc.) can be imputed at two-field resolution with >90% accuracy in European populations. Accuracy declines for rare alleles and admixed populations without suitable reference panels.

We used the HIBAG algorithm for HLA imputation. HIBAG is an R package designed to impute HLA genotypes using a machine learning method called attribute bagging. It requires SNP Genotype Data in PLINK or VCF format and pre-trained HIBAG models, which are locus-specific classifiers trained on a reference panel with known HLA alleles.

The output of HIBAG is an R object or table listing imputed HLA genotypes along with posterior probabilities. Below is an example output for imputation at the HLA-A locus:

SAMPLEID	HLA-A_1	HLA-A_2	PROBABILITY
S1	A*02:01	A*03:01	0.98
S2	A*01:01	A*02:01	0.96
S3	A*24:02	A*24:02	0.91

HLA-A_1 / HLA-A_2: The two imputed alleles per individual (diploid genotype).

Probability: Confidence score for the imputed genotype, based on the posterior probability computed by the bagging classifier.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

4. Definition of essential HLA data format

A standardized HLA genotype format was established, containing two alleles per individual at each of the following loci**:** HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, and -DRB1**.** A two-field resolution is considered the minimal threshold for inclusion in the downstream analytical pipeline.

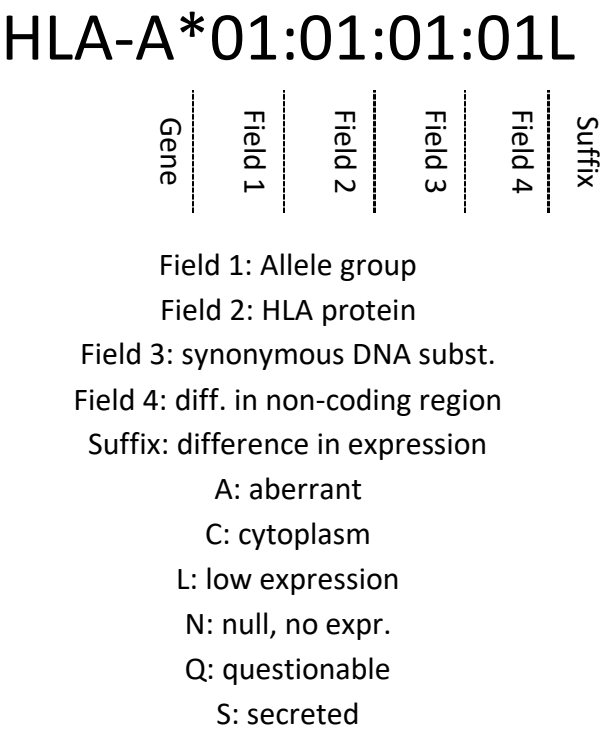

Figure 1. The nomenclature of HLA alleles.

These loci were selected because they represent the most functionally relevant class I and class II HLA genes with direct roles in antigen presentation and immune response modulation. HLA-A, -B, and -C are classical class I genes that present intracellular peptides to CD8+ cytotoxic T cells, making them critical in viral immunity, tumour surveillance, and transplant compatibility. HLA-B is notably the most polymorphic HLA locus and strongly associated with many disease susceptibilities and adverse drug reactions.

On the class II side, HLA-DRB1, DQA1, and DQB1 encode proteins forming the HLA-DR and HLA-DQ heterodimers, which present extracellular peptides to CD4+ helper T cells. These loci are central to the pathogenesis of many autoimmune diseases and infections, and HLA-DRB1 in particular carries strong disease associations due to its high polymorphism and expression.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Similarly, HLA-DPA1 and DPB1 encode the HLA-DP heterodimer, which also presents antigens to CD4+ T cells. Though somewhat less polymorphic, the HLA-DP genes contribute to the overall breadth of the antigen-presenting landscape and are implicated in certain infections, autoimmune conditions, and transplant outcomes.

Including all these loci ensures comprehensive coverage of both class I and class II HLA functionality. This is critical for analyses of HLA promiscuity, binding motif clustering, and associations with immune-mediated diseases across diverse cohorts.

5. Pipeline integration and analysis capabilities

The pipeline developed by BRC processes the standardized HLA genotype format efficiently and supports comprehensive immunogenetic analysis. It constructs stable HLA-DQ alpha-beta complexes, identifies cis and trans configurations, and determines all HLA-DP complexes per individual. In addition to structural assembly, the pipeline computes HLA promiscuity at both individual locus and broader class levels. It further classifies alleles and allele complexes into functional clusters based on similarities in peptide-binding preferences, which are critical for understanding antigen presentation diversity and immune responsiveness.

To quantify binding specificity, the pipeline incorporates a computational framework that evaluates peptide–HLA interactions using data curated from immunopeptidomics studies. For accurate estimation, only HLA alleles with data on a sufficient number of peptides are analyzed.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

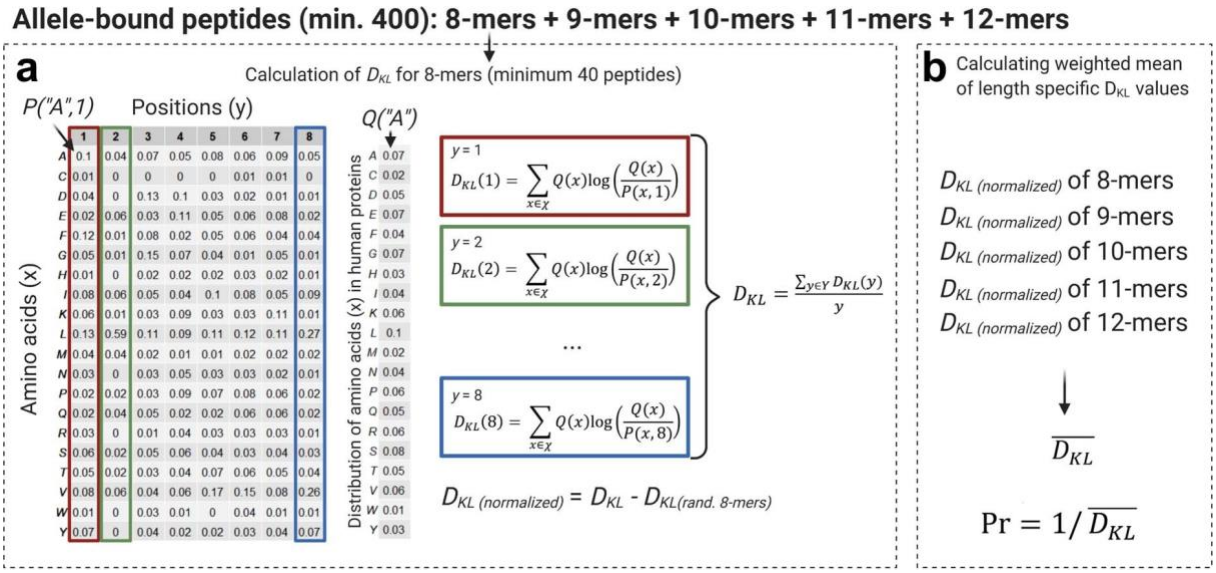

Figure 2. The calculation approach for HLA-I promiscuity. See text for details.

Peptide-binding promiscuity is calculated using Kullback–Leibler divergence (D_{KL}), a metric that quantifies the difference between amino acid frequency distributions in peptide-binding repertoires versus the background distribution in the human proteome (obtained from UniProt). For each peptide length and binding position, the amino acid frequencies are compared, and DKL is computed. A smoothing factor (10^{-7}) is applied to avoid division by zero. These position-specific D_{KL} values are then averaged to obtain peptide-length-specific scores, which are normalized against randomly sampled peptides (D_{KL_rand}) to eliminate bias due to peptide count.

The final promiscuity metric, denoted as *Pr*, is calculated as the reciprocal of the normalized average D_{KL} across peptide lengths. Lower D_{KL} indicates broader, less selective binding (i.e., higher promiscuity), while higher D_{KL} reflects more selective binding patterns. Importantly, this measure of promiscuity is robust across different reference proteomes (e.g., bacterial, viral) and is not significantly influenced by sample size. Furthermore, Pr values calculated using D_{KL} show a strong positive correlation with those derived from the Shannon entropy index, reinforcing their conceptual validity (Spearman's $\rho = 0.86$, $P < 2.2 \times 10^{-16}$).

Through these analyses, the pipeline provides a functional perspective on HLA diversity, enabling downstream tasks such as clustering alleles based on similar peptide-binding behaviors, identifying immunodominant motifs, and investigating associations with immune-mediated diseases and vaccine responses.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

6. High throughput genomic analysis of BCR and TCR germline genes – ImmuneDiscover

The ImmuneDiscover genotyping technique is based on targeted genomic sequencing of approximately 350 non-rearranged BCR and TCR V, D and J genes. ImmuneDiscover functions in a high throughput manner to enable the personalized immune gene profiling of large numbers of individuals (>1000) from disease cohorts where DNA samples are available. The aim is to identify potential associations between germline-encoded polymorphisms in adaptive immune receptor genes and clinical outcomes, such as in patients diagnosed with autoimmune conditions.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

7. Conclusions

The work conducted under Task 2.1 has successfully contributed to several key objectives of the ID-DarkMatter-NCD project. First, we completed a comprehensive assessment and curation of HLA and WGS datasets across all participating cohorts. We identified which cohorts already possess usable HLA genotyping data, as well as those requiring additional genotyping. This evaluation ensures efficient targeting of resources for new data generation under Task 2.2, thereby strengthening overall data harmonization and project cohesion.

A major achievement of this task was the definition and adoption of a standardized HLA genotype format, incorporating eight key loci at two-field resolution. This ensures that all genotyping, whether existing or newly generated, meets a consistent threshold suitable for downstream analysis. The inclusion of both class I and class II loci—HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, and -DRB1—provides comprehensive immunogenetic coverage necessary to explore disease associations and antigen presentation dynamics across a broad spectrum of immune-mediated diseases and post-infectious conditions.

We also implemented a robust pipeline developed at BRC that integrates this standardized data. This pipeline calculates HLA promiscuity and functional similarity among alleles using metrics such as Kullback–Leibler divergence. It enables functional classification and clustering of HLA alleles, which will directly feed into disease association models and help interpret patterns of immune activation or tolerance.

In addition, we ensured compatibility and secure exchange of data between partners. While most datasets can be shared in a fully anonymized and standardized format, we have addressed cohort-specific constraints—such as ethical limitations or platform-specific access rules—through alternative solutions such as localized analysis and non-standardized exchange protocols.

Collectively, these results provide a strong foundation for the project's future work packages focused on linking infection histories, HLA genotypes, and immune responses to non-communicable disease risk. The curated data and analytical pipelines developed here ensure that all partners can contribute to and benefit from harmonized, high-resolution immunogenetic data in support of the overarching objectives of the ID-DarkMatter-NCD project.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

8. Disclaimer

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

9. Acknowledgement of funding

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by