Phage display sequencing reveals that genetic, environmental, and intrinsic factors influence variation of human antibody epitope repertoire

Lifelines Cohort Study; Andreu-Sánchez, Sergio; Bourgonje, Arno R.; Vogl, Thomas; Kurilshchikov, Aleksandr; Leviatan, Sigal; Ruiz-Moreno, Angel J.; Hu, Shixian; Sinha, Trishla; Vich Vila, Arnau

Published in:
Immunity

DOI:
10.1016/j.immuni.2023.04.003

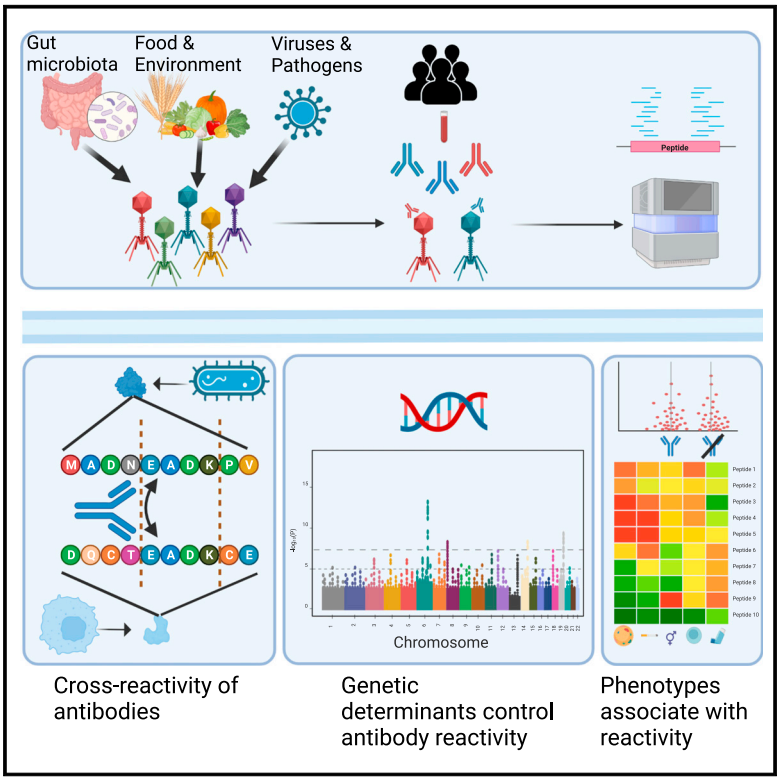
Publication date:
2023

Resource

Immunity

Phage display sequencing reveals that genetic, environmental, and intrinsic factors influence variation of human antibody epitope repertoire

Graphical abstract



Highlights

- PhIP-seq libraries from microbial and environmental antigens in a population cohort
- Antibody-bound peptide co-reactivity highlights bacterial mimicry
- Common genetic variants in HLA, IGHV, and FUT2 determinate antibody reactivity
- Antibody reactivity is associated with widespread phenotypical factors

Authors

Sergio Andreu-Sánchez, Arno R. Bourgonje, Thomas Vogl, ..., Rinse K. Weersma, Jingyuan Fu, Alexandra Zhernakova

Correspondence

thomas.vogl@meduniwien.ac.at (T.V.), sashazhernakova@gmail.com (A.Z.)

In brief

In this study, Andreu-Sánchez et al. utilize phage-displayed immunoprecipitation sequencing to investigate the environmental and genetic determinants shaping human adaptive immunity. The results suggest that both genetics and environmental exposures shape human antibody epitope repertoires, with specific signatures of distinct phenotypes and genotypes. Furthermore, co-occurring antibody responses suggest a link between bacterial immunity and the development of allergies or autoimmunity.

Andreu-Sánchez et al., 2023, Immunity *56*, 1376–1392
June 13, 2023
https://doi.org/10.1016/j.immuni.2023.04.003

CellPress

CellPress

Immunity

Resource

Phage display sequencing reveals that genetic, environmental, and intrinsic factors influence variation of human antibody epitope repertoire

Sergio Andreu-Sánchez,[1,2,9] Arno R. Bourgonje,[3,9] Thomas Vogl,[4,5,6,7,*] Alexander Kurilshikov,[1] Sigal Leviatan,[4,5] Angel J. Ruiz-Moreno,[1] Shixian Hu,[1,3] Trishla Sinha,[1] Arnau Vich Vila,[1,3] Shelley Klompus,[4,5] Iris N. Kalka,[4] Karina de Leeuw,[8] Suzanne Arends,[8] Iris Jonkers,[1] Sebo Withoff,[1] Lifelines Cohort Study,[1] Elisabeth Brouwer,[8] Adina Weinberger,[4,5] Cisca Wijmenga,[1] Eran Segal,[4,5] Rinse K. Weersma,[3,9] Jingyuan Fu,[1,2,9] and Alexandra Zhernakova[1,10,*]

[1]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[2]Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[3]Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[4]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
[5]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel
[6]Diagnostic and Research Institute of Hygiene, Microbiology and Environmental Medicine, Medical University Graz, Graz, Austria
[7]Center for Cancer Research, Medical University of Vienna, Wien, Austria
[8]Department of Rheumatology and Clinical Immunology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[9]These authors contributed equally
[10]Lead contact
*Correspondence: thomas.vogl@meduniwien.ac.at (T.V.), sashazhernakova@gmail.com (A.Z.)
https://doi.org/10.1016/j.immuni.2023.04.003

SUMMARY

Phage-displayed immunoprecipitation sequencing (PhIP-seq) has enabled high-throughput profiling of human antibody repertoires. However, a comprehensive overview of environmental and genetic determinants shaping human adaptive immunity is lacking. In this study, we investigated the effects of genetic, environmental, and intrinsic factors on the variation in human antibody repertoires. We characterized serological antibody repertoires against 344,000 peptides using PhIP-seq libraries from a wide range of microbial and environmental antigens in 1,443 participants from a population cohort. We detected individual-specificity, temporal consistency, and co-housing similarities in antibody repertoires. Genetic analyses showed the involvement of the *HLA*, *IGHV*, and *FUT2* gene regions in antibody-bound peptide reactivity. Furthermore, we uncovered associations between phenotypic factors (including age, cell counts, sex, smoking behavior, and allergies, among others) and particular antibody-bound peptides. Our results indicate that human antibody epitope repertoires are shaped by both genetics and environmental exposures and highlight specific signatures of distinct phenotypes and genotypes.

INTRODUCTION

The adaptive immune system encompasses an extremely complex group of biological processes that orchestrate responses to invading pathogens in all jawed vertebrates (*gnathostomes*), including humans.[1] Its ability to recognize, adapt to, and remember threats relies on polymorphic genetic structures that encode receptors for antigens, which are typically amino acid sequences.[1] Antibodies are the primary effector molecules responsible for humoral immunity and are highly adaptable and influenced by individual's genetics and environmental factors. Antibody repertoires determine the fate of the immune response against pathogens and the development of autoimmunity or allergies, and they have garnered special attention because they can be used to study herd immunity acquisition.[2] In an adult human, there are around 10^{10}–10^{11} B-lymphocytes, each expressing a unique B cell receptor (BCR) (a non-soluble antibody form) that identifies a molecular pattern.[3] The diversity of BCRs results from somatic rearrangements of gene segments, insertion and deletion of nucleotides, and somatic hypermutation.[4]

To gain more insights into antibody-antigen interaction, efforts have been made to directly sequence the BCR[5,6] and to directly infer it from single-cell transcriptomic sequencing.[7] Although this

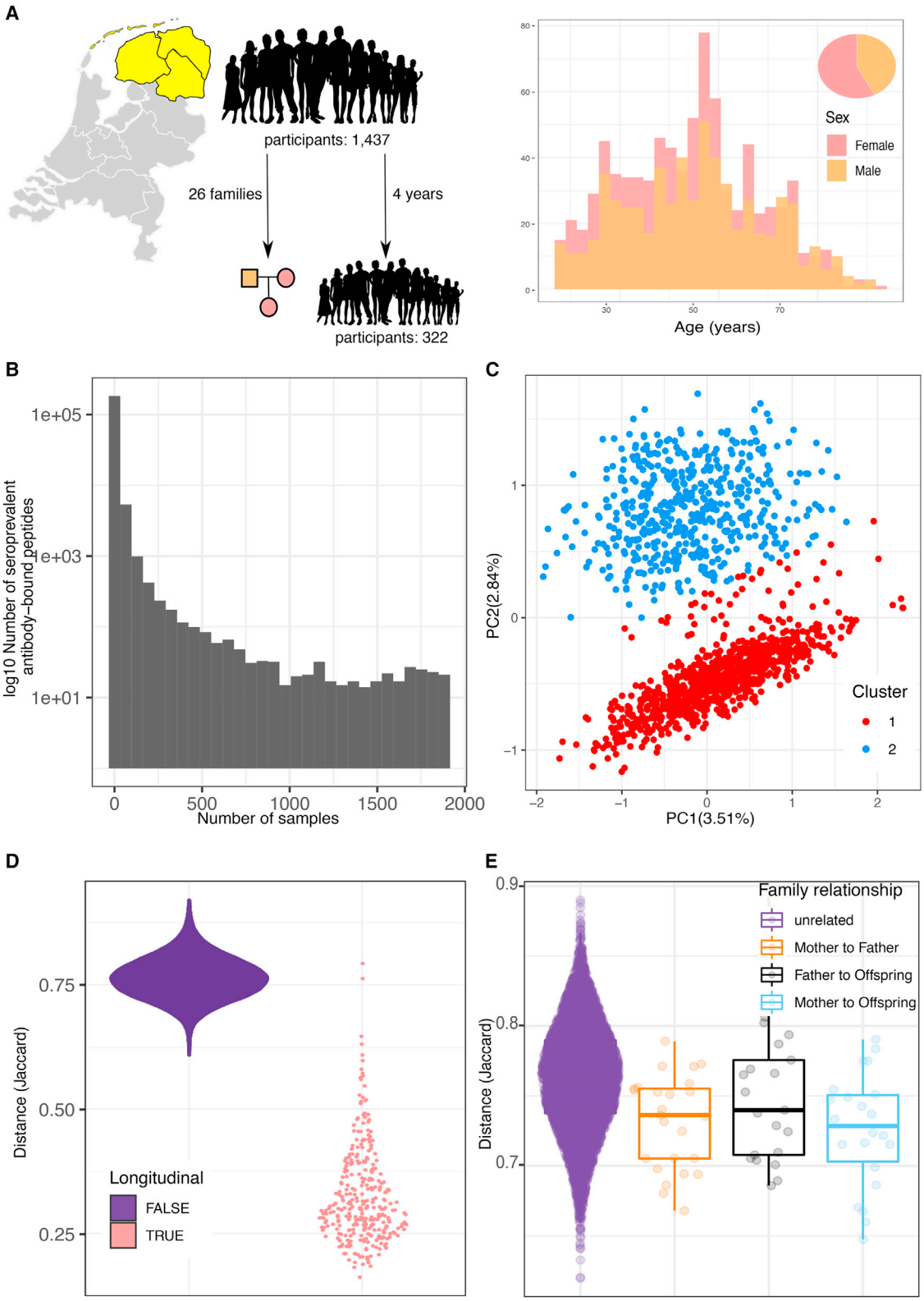


Figure 1. PhIP-seq antibody-bound peptide profiles of 1,443 individuals representative of the Dutch population show temporal stability and family similarity

(A) Cohort characteristics. Lifelines-Deep (LLD) is a population cohort from Northern Netherlands. In this work, we performed PhIP-seq in 1,443 participants (including 26 trio families), 322 of whom have data from a second time point after 4 years. Other data layers include phenotypes (questionnaires and clinical

(legend continued on next page)

methodology provides information on the potential for generation of immune responses against yet unknown antigens, it does not directly link BCR sequences to the exact nature of antigenic epitopes. In addition, in terms of scaling, it is limited to just a small proportion of the immense number of these receptors.[8] On the other hand, antibody-binding analysis, such as peptide microarrays[9,10] or enzyme-linked immunosorbent assay (ELISA), enables the determination of antibody seroprevalence against selected antigens. Although easily implemented for a limited set of antigens, these methodologies have been difficult to scale up to thousands of antigens in a large population. Phage-displayed immunoprecipitation sequencing (PhIP-seq) allows for the comprehensive determination of the interactions of the human antibody epitope repertoire with rich libraries of potential antigens. Briefly, a group of antigenic peptides is integrated and displayed on bacteriophages that are incubated with blood samples. Subsequently, all the reactive antibodies present in a sample will bind to their corresponding antigens, with bound phages then extracted by immunoprecipitation and sequenced to obtain an ''immunological fingerprint'' of the individual's antibody repertoire. PhIP-seq has been described previously[11,12] and successfully applied to characterize autoimmune antibody prevalence in patients with multiple sclerosis, type 1 diabetes, and rheumatoid arthritis,[13,14] the human virome,[15–19] and the widespread presence of antibodies against virulence factors[20,21] and the gut microbiome.[21] In addition, previous studies characterized environmental and genetic contributors to immunological traits other than PhIP-seq, such as cytokine responses,[22] blood cell composition,[23] T cell receptor repertoire,[24] and BCRs.[25,26] However, to date, no comprehensive study has been carried out that identifies the environmental, intrinsic, lifestyle, and genetic factors associated with antibody generation against a wide array of antigen exposures in the general population.

In this work, we set out to uncover the antibody epitope repertoire in a deeply phenotyped population cohort from the northern part of the Netherlands, Lifelines-DEEP (LLD).[27] We used two PhIP-seq libraries previously described[21,28] to characterize 344,000 peptide antigens related to: (1) microbes (including human gut microbiota, probiotic strains, pathobionts, antibody-coated species, and virulence factors from the virulence factors database [VFDB]), (2) the Immune Epitope Database (IEDB),[29] (3) proteins from allergen databases, and (4) bacteriophages. Leveraging the rich metadata available for this deeply phenotyped cohort (including imputed genotypes, gut microbiota shotgun sequencing, clinical blood tests [immune, metabolic, and autoimmune markers], family information, lifestyle and self-reported diseases, and allergy questionnaires) alongside the PhIP-seq data allowed us to establish key genetic and environmental factors shaping the human antibody epitope repertoires.

RESULTS

Antibody-bound peptide repertoires are personalized, linked to shared environments (co-housing) and time-dependent

We interrogated a total of 344,000 peptides in 1,778 samples from 1,437 individuals (for 341 of whom we had data at two time points 4 years apart) from a northern Dutch population cohort (LLD) (Figure 1A).

After immunoprecipitation with protein A/G, binding primarily IgG antibodies,[21] and sequencing, we detected an enrichment of sequenced reads (see STAR Methods) of 175,242 (antibody-bound) peptides in at least one participant (average number of peptides bound per person = 1,168, range = 3–3,161) (see STAR Methods). Peptide seropositivity was defined as a presence/absence binary score (enriched/not enriched) that was used for all subsequent analyses. Most antibody-bound peptides showed low seroprevalence, indicating the individual-specificity of the antibody epitope repertoire (Figure 1B). Based on peptide sequence identity and prevalence (see STAR Methods for details), we chose 2,815 peptides for further analyses (Table S1.1).

The large variability in the antibody-bound peptide enrichment profile could be seen through a principal component analysis (PCA), where the amount of variability recovered by the first 10 principal components (PCs) was just 15.5% and 709 components were needed to retrieve 90% of the total antibody-bound peptide variability (Figure 1C). Despite the relatively low variability accounted for by the first two PCs (6.3%), we observed two clusters of samples in PC2 that were driven by cytomegalovirus (CMV)-related antibody-bound peptides (K-medoids, k = 2) (Figure 1A). Removal of these peptides resolved PC2 clustering (Figure S1A), although the effect of CMV could still be detected shaping interindividual antibody differences. This is consistent with a previous observation that nearly 50% of the Dutch adult population are seropositive for this herpesvirus.[30] These CMV-related antibody-bound peptides tended to increase with age, suggesting a gain in antibodies against this virus with viral reactivations over the course of life (Figure S1B). On the other hand, PC1 was highly related to the number of seropositive peptides (affine linear model R^2 = 0.72). In a permutational multivariate analysis of variance (PERMANOVA) (adjusted for age, sex, and sequencing plate), person-to-person antibody-bound peptide repertoire dissimilarity showed effects (2,000 permutations, $p < 5 \times 10^{-4}$) of age (R^2 = 0.14), smoking (R^2 = 0.018), blood measurements (e.g., cholesterol R^2 = 0.012), and blood cell counts (lymphocyte relative abundance, R^2 = 0.016), among many other phenotypes (Table S2.1).

In agreement with previous reports, we observed temporal consistency in the antibody-bound peptide repertoire[20,21] for the 322 participants who were followed up after 4 years. We

measurements), genetics (imputed microarrays), and microbiome (bacterial taxonomic quantification). There is a higher proportion of females within the participants (57%). The age distribution is slightly left skewed, with a mean of 44.5 years (female effect on age = −1, p = 0.16).

(B) Prevalence of antibody-bound peptides in the population. x axis depicts seroprevalence. y axis is the number of antibody-bound peptides with a given seroprevalence.

(C) Principal component analysis identified two clusters (color represents cluster labels after 2-medoids clustering).

(D) Jaccard distance between antibody repertories of 322 samples longitudinally followed 4 years apart and between unrelated samples.

(E) Jaccard distance between antibody repertories of 26 family trios and between unrelated participants.

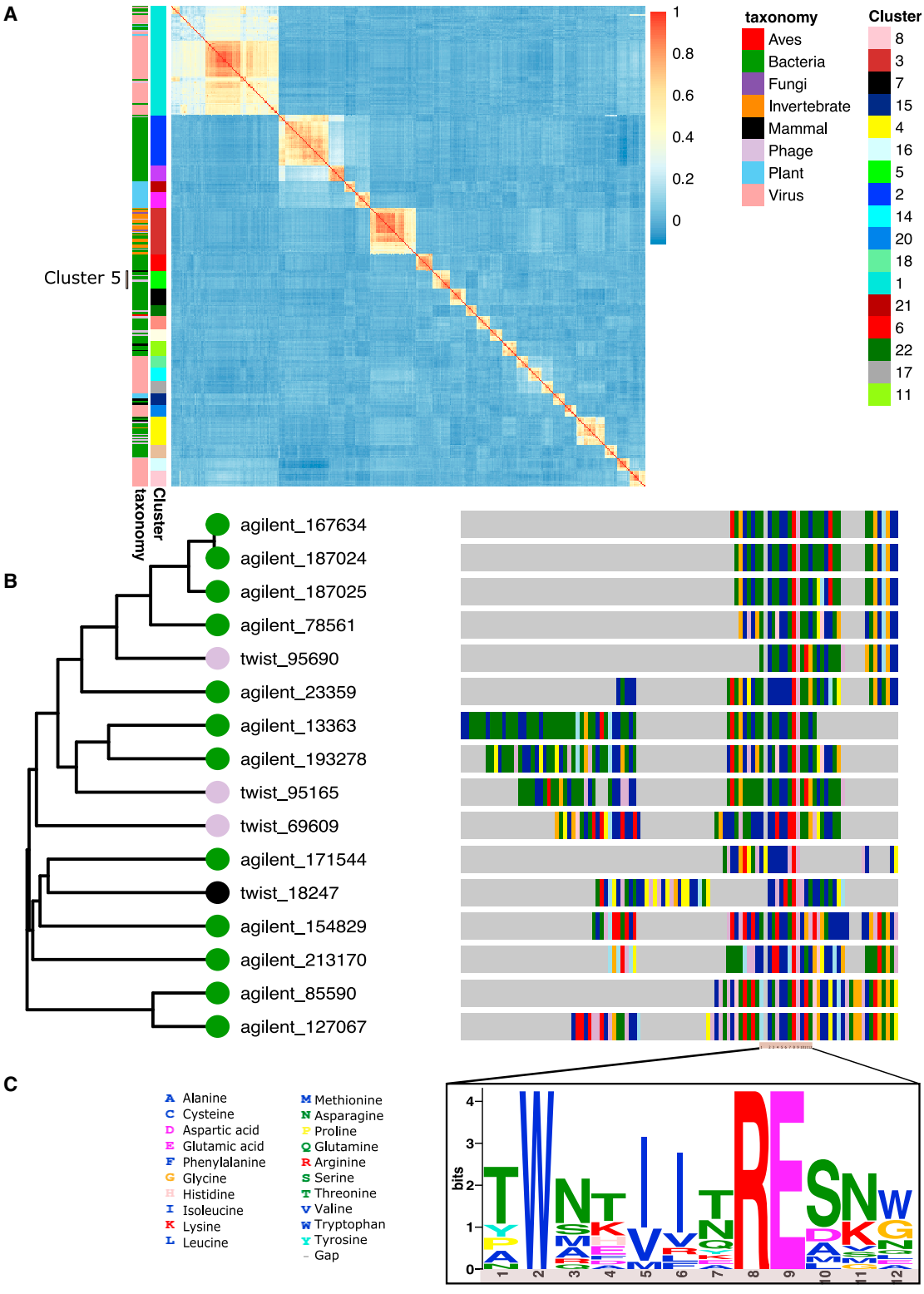


Figure 2. Peptide co-occurrence highlights the presence of motifs driving antibody cross-reactivity

(A) Correlation heatmap between peptides that belonged to co-occurrence modules of at least 10 peptides using 1,443 individuals. Annotation displays the taxonomic origin of each peptide and the cluster assigned by WGCNA. Module 5 is highlighted.

(legend continued on next page)

observed that the distance between samples taken from the same individuals 4 years apart was on average lower than the distance of unrelated individuals ($p < 5 \times 10^{-4}$; 2,000 label permutations) (Figure 1D), and this was independent of the antibody-bound peptides used for calculating the distance, as similar results were observed using subsets of 20%, 40%, 60%, or 80% of the antibody-bound peptides. Overall, the distance between baseline and follow-up was not associated with baseline age or sex. The temporal consistency of antibody-bound peptides showed a bimodal distribution, with most peptides consistent between time points and only a subset that tended to change (Figure S1D; Table S1.1). This change was more often a loss of enrichment rather than a gain, and this difference could not be directly attributed to a batch effect (Wilcoxon test, $p = 0.45$). This highlights that the time elapsed since antigen encounter might be a determining factor for the detection of antibody-bound peptide enrichment, which agrees with humoral studies showing that the prevalence of antibodies fades over time.[31–33]

Next, we studied whether genetically related individuals or those living in similar environments (co-housing) would show more similarity in antibody-bound peptide enrichment compared with unrelated individuals. To explore this, we used 26 family trios from the LLD population[34] (note that most offspring are unlikely to currently cohouse with their parents as their average age was 37 ± 10.1 years old). Mother-offspring, father-offspring, and father-mother antibody-bound peptide distances were significantly lower than those between unrelated individuals ($p < 5 \times 10^{-4}$, $p = 0.013$, and $p < 5 \times 10^{-4}$, respectively; 2,000 label permutations). However, no significant differences were found between pairs of family members, although father-offspring pairs were, on average, more distant (Figure 1D). The role of the common environment in shaping antibody repertoires is supported by the decreased father-mother distance, whereas offspring associations could indicate an important role of the environment during early life, a common lifestyle, the effect of genetics, or all, to some degree.

Co-occurrence of peptides identifies multiple epitopes for the same antigen, antibody cross-reactivity in related structures, and co-occurrence of antibodies against unrelated proteins

To understand the relation between antibody-bound peptides, we computed their correlation and built a network using weighted gene co-expression network analysis (WGCNA) by computing correlation coefficients from the binary profile of all selected peptides without missing values. 435 peptides could be assigned to 22 modules of at least 10 highly correlated peptides (denoted by the number of peptides per module, 1–22) (Figures 2A and S2; Table S1.1). A bootstrapping consistency analysis identified high consistency in all but one module (module 17). After assessment of the antibody-bound peptides within each of the modules and the sequence similarity between them, we identified three main types of modules: class I—modules driven by antigens from the same biological source, class II—modules driven by homologous antigenic sequences, and class III—modules that include peptides that are not taxonomically or structurally related but do correlate strongly with each other (Table S1.3).

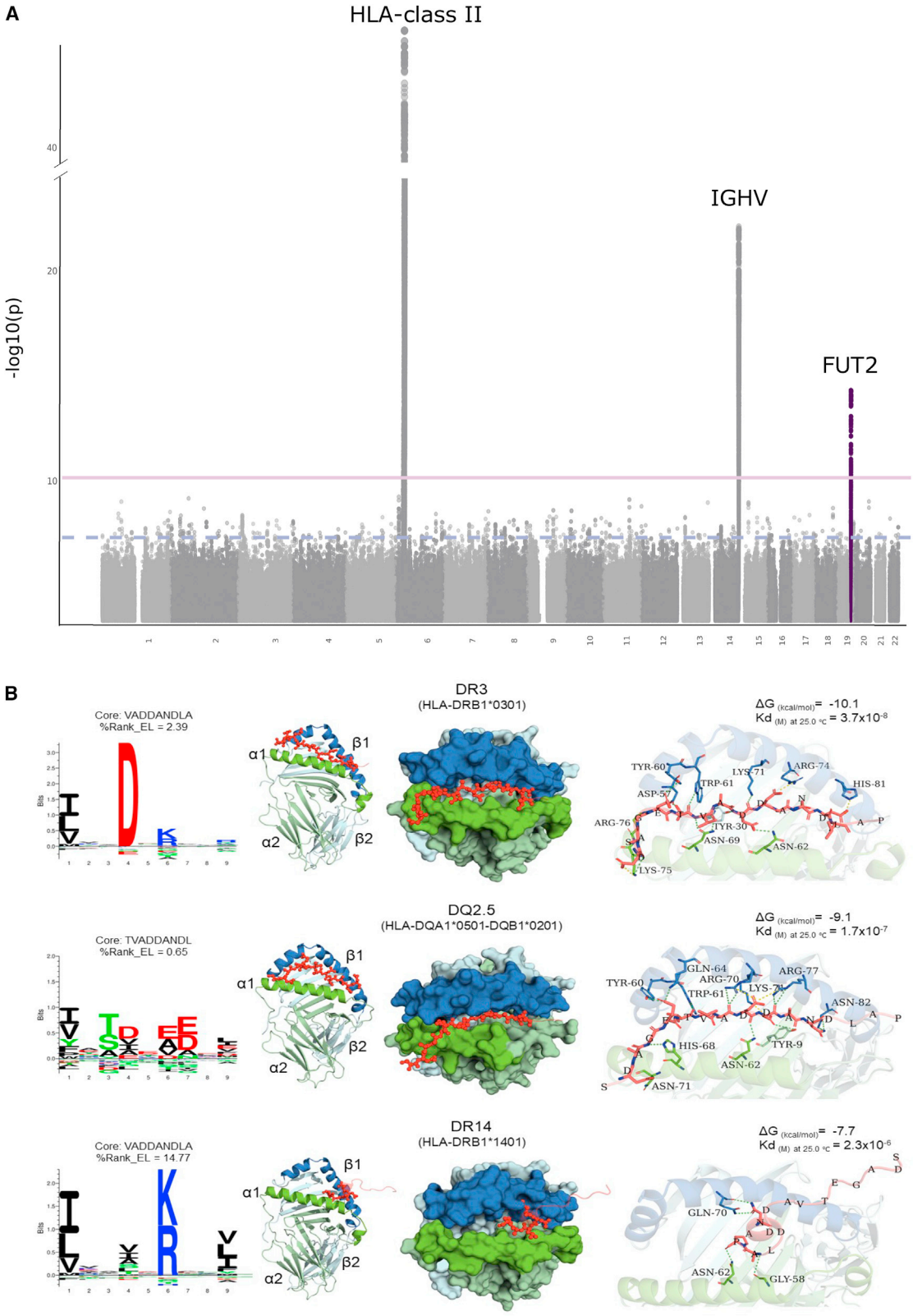
We observed five category I modules (Figure S2). For example, module 16 was composed of two different Epstein-Barr virus (EBV) proteins, including capsid protein VP26 and nuclear antigen 1 (EBNA-1); module 20 was composed of high-identity peptides belonging to different strains of influenza B viruses, and module 1 was mainly driven by CMV peptides, although also including some EBV and other peptides. Category II modules, driven by similar sequences in different peptides, highlight the cross-reactivity of the antibody response (Figure S2). For example, module 21 was composed of plant thionins, small cytotoxic plant compounds produced by many species, but here mainly derived from common wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and rye (*Secale cereale*). Module 9 contained related antigens from wheat, Asian rice (*Oryza sativa*), rye, barley, and grass (*Setaria italica*) that represent plant granule-bound starch synthase peptides. Modules 14, 17, and 18 were characterized by antibody-bound peptides representing genome polyproteins from a series of viruses, including Enterovirus A71, B, and C; rhinovirus B and serotype 2; coxsackievirus (type A9), and poliovirus. Module 3 was dominated by allergen peptides, including antigens involved in common insect and seafood allergies, e.g., *Artemia franciscana* (shrimp), *Octopus vulgaris* (octopus), *Blattella germanica* (German cockroach), *Dermatophagoides farinae* (house dust mite), *Portunus trituberculatus* (gazami crab), *Bombus hypocrita* (bumble bee), and *Ctenocephalides felis* (cat flea).

Examples from category III, where no structural or taxonomic relation is seen, were harder to interpret (Figure S2). Although some members in this category had a majority of peptides belonging to category I or II, others did not show major structural relations and were mainly composed of bacterial peptides or bacterial and autoimmune peptides that clustered together. Although no overall homology was observed in these modules, a detailed analysis of their sequence similarity identified common motifs that appeared in most modules (4, 5, 11, 13, 15, 19, and 22) (Table S1.2). The presence of these common motifs could imply recognition by a common antibody, causing cross-reactivity. One such module (module 5) (Figure 2B) linked the presence of a common motif (TWNTIITRESNW, E value = 7.10×10^{-60}) in different bacterial proteins (from *Lactobacillus*, *Prevotella*, or *Dorea*), peptides belonging to *Lactobacillus* phages and human idursulfase. Human idursulfase is commonly used during enzyme replacement therapy in patients with Hunter syndrome. Allergic reactions to idursulfase have been reported in some patients, but no clear risk factors or sequence similarity to common allergens have been reported.[35] Our result might

(B) Module 5 motif discovery. At left, a hierarchical clustering (average method) based on sequence similarity between the peptides belonging to module 5. At right, their multiple sequence alignment (each colored line represents an amino acid, and gray indicates an alignment gap). Peptides' colors indicate their taxonomic origin.

(C) Logo of the most significant motif from the module 5 sequences (MEME, E value 7.1×10^{-60}). y axis represents bits of information for each position and amino acid.

(B and C) Amino acid residues are colored according to their chemical properties represented in the same legend.



(legend on next page)

point to a role for the gut microbiome in sensitization against this drug through bacterial mimicry.

In addition, we built a second network using logistic regression coefficients instead of correlation values (STAR Methods). This second network identified a total of 12 modules (with at least 10 peptides each). Of those, eight were homologous to the findings in the correlation-based network (modules 9, 11, 15, 16, 19, 20, 21, and 22). The four additional modules mainly belonged to bacterial proteins, and we found common sequence motifs in two of them (Table S1.2).

Peptide enrichment is associated to HLA, FUT2, and IGHV genetic regions

Our observation that both common environments and genetic relations within families affect the antibody-bound peptide repertoire (Figure 1E) made us wonder about the specific drivers of repertoire variability. Genetics are known to influence antibody repertoires,[36–39] but the exact contribution of genetic and environmental factors to bacterial and, especially, commensal gut microbiota immune-reactivity is incompletely characterized.

We estimated the proportion of antibody-bound peptide presence/absence variability accounted for by common genetic variation, i.e., its heritability (H^2), using common genetic variants in 1,255 unrelated individuals. We saw an overall moderate genetic contribution to the variability of antibody-bound peptide enrichment (mean $H^2 = 0.1$, median = 0.06, min = 0, and max = 0.96) (Table S1.1). A total of 35/2,814 antibody-bound peptides showed very high heritability ($H^2 \geq 0.5$), whereas a substantial number (597/2,814) had a relatively high heritability ($H^2 \geq 0.2$). Using the highly heritable antibody-bound peptides ($H^2 \geq 0.5$), we then computed genetic correlations to determine similar genetic signals across antibody-bound peptide presence. We found a correlation of 0.47 between the matrices of presence/absence and genetic correlations (Mantel test, $p < 1 \times 10^{-4}$, 999 permutations) (Figure S1E). We also observed hubs of highly genetically correlated groups of peptides in which the genetic signatures are more correlated than antibody-bound peptide presence itself (Figure S1E). This indicates the existence of a common genetic architecture explaining the presence of antibody-bound peptides.

Next, we set out to uncover specific loci contributing to the observed heritability. We ran a genome-wide association study (GWAS) on 4,546,708 genotyped and imputed SNPs in 2,815 peptides. To reduce the false discovery rate (FDR) and increase the power of the analysis, we meta-analyzed the results of our LLD GWAS with those of a dataset that used the same PhIP-seq libraries in the context of inflammatory bowel disease (IBD) (490 participants),[40] bringing us up to a total of 1,745 individuals (Table S2.2). At the study-wide significance threshold ($p < 5.67 \times 10^{-11}$), we identified multiple signals in three genetic loci associated with 149 antibody-bound peptides. These were located in chromosome 6 (human leukocyte antigen [*HLA*] locus), chromosome 14 (Immunoglobulin heavy-chain variable [*IGHV*] region), and chromosome 19 (fucosyltransferase 2 [*FUT2*] gene) (Figure 3A).

The strongest genetic signal belonged to the HLA-class II region in chromosome 6, where we found 130 peptides associated with 134 different leading SNPs. Most of the associated peptides belonged to *Streptococcus* and *Staphylococcus* species, but we also found several peptides belonging to human viruses (adenoviruses or herpesviruses) and phages, as well as some related to allergens (ovomucoid, barley, casein, and wheat, among others) and gut microbiota. Focusing on this genomic region, we conducted a specific imputation of HLA SNPs, indels, amino acids, and gene isoforms and performed an association analysis with all peptides (see STAR Methods and Table S2.4). This analysis substantially increased the number of associated peptides. We discovered that a large number of peptides (530/2,813) had at least one significant ($p < 1 \times 10^{-6}$, after correction for the number of independent tests; see STAR Methods) association with HLA variants (amino acids, insertions, SNPs, or genes). At the HLA gene level, we identified 1,192 statistically significant peptide-gene associations with 276 different peptides. Most of those associations (and the strongest) belonged to allelic variants of HLA-II (1,070 associations to 271 different peptides) in comparison to variants of HLA-I (122 associations to 41 different peptides). Within the HLA-II variations, most associations were observed for various alleles in DQ and DR beta chain genes.

To determine whether these associations were due to the capacity of a specific HLA complex to present the peptide, we performed computational modeling of the HLA-peptide complex using some of our top associations. This modeling was done for: (1) streptococcal C5A peptidase and DR3, which was the top association for the DR3/DQ2 haplotype relevant for several autoimmune diseases,[42,43] (2) *Lactobacillus* phage hypothetical protein LfeINF_097 and DR15, which was the strongest association observed in our association analysis with HLA genes (odds ratio [OR] = 13.3, $p = 1.44 \times 10^{-47}$), and (3) *Human mastadenovirus* minor core protein and DR4/DQ8 haplotype, which is also linked to autoimmunity.

Here, we identified that the predicted residues that are recognized from the peptide by a specific HLA complex[41] can form stable structures with their associated HLA complexes.

The streptococcal C5a peptidase (*TPSDAGETVADDANDL APQAPAKTADTPATSKATIRDLNDPSQVKTLQEKAGKGAGTVV AVIDA*) is highly associated with DRB1*0301 (always bound to the alpha chain DRA*01, DR3 haplotype) (OR = 3.78,

Figure 3. Genetics contribute to antibody-bound peptide variability

(A) Manhattan plot from genome-wide association study meta-analysis of 2,798 antibody-bound peptides in 1,745 participants (490 IBD). Genome-wide association threshold (5×10^{-8}, blue) and study-wide significance (7×10^{-11}, red) are shown as horizontal lines. Labels indicate the three major loci identified. Colored dots represent a recessive model. Gray dots represent additive models.

(B) Peptide motif deconvolution maps of DR3, DQ2.5, and DR14 (amino acids code: negatively charged, red; positively charged, blue; polar uncharged, green; hydrophobic, black) compared with the *Streptococcus agalactiae* C5a peptidase peptide core and percentage of elution score (%Rank_EL: strong binding ≤ 2.0, weak binding 2.0–10.0, no binding > 10) predicted by NetMHCIIpan-4.0.[41] Predicted binding mode, polar molecular interactions (dashes, hydrogen bonds: green, and salt bridges: yellow), binding energy, and dissociation constant (K_D) of the *Streptococcus agalactiae* C5a peptidase peptide core (red cartoon and sticks) into HLA-II receptors (chain A in green and chain B in blue).

$p = 1.65 \times 10^{-31}$) and with DQB1*0201 (OR = 3.75, $p = 5.16 \times 10^{-31}$) and the alpha chain DQA1*0501 (OR = 1.91, $p = 4.80 \times 10^{-13}$), which together form the haplotype DQ2.5 that is highly linked to DR3. The predicted core recognized by the HLA complex (STAR Methods) was nearly identical for both DR3 and DQ2.5 (*VADDANDL*) and has a high similarity to the amino acid composition identified from HLA ligand elution experiments.[44] Additionally, we employed the predicted binding metric (percentage of elution score, %Rank_EL; STAR Methods) to assess the binding of the core peptide to the selected alleles. This analysis found a favorable binding prediction of the core to DR3 and DQ2.5 complexes, with a higher binding for DQ2.5 (%Rank_EL 2.39 and 0.65, respectively). We further compared the binding prediction for this epitope with a non-associated negative control (DR14), which was predicted to be non-binding (%Rank_EL 14.77). Additionally, structural modeling and analysis of binding mode showed that the computed dissociation constant (K_D) had an order of magnitude less affinity for the non-associated allele (2.3×10^{-6} M) compared with DR3 (3.7×10^{-8} M) and DQ2.5 (1.7×10^{-7} M) (Figure 3B). As a result, the peptide core exhibited similar behavior and key stabilizing polar interactions when binding into the binding sites of DR3 and DQ2.5. For example, the hydrogen bonds occurring between the Tyrosine 60 (Tyr60) and Tryptophan 61 (Trp61) present in the beta chain of both DR3 and DQ2.5 interact with glutamic acid (Glu) and threonine (Thr) in the peptide core. By contrast, although we could model the peptide binding into the negative control DR14, the majority of the peptide's amino acids are located outside of the binding site and in the opposite direction compared with DR3 and DQ2 (Figure 3B).

Next, we focused on the other two highly associated HLA-peptide complexes: (1) the combination of the peptide *Lactococcus* phage (YP_009222335.1 hypothetical protein LfeInf_097) with the DR15 haplotype (DRB1*0301), which showed the strongest study-wide association (OR = 13.3, $p = 1.44 \times 10^{-47}$) (Figure S3A), and (2) a combination of a peptide from the *Human mastadenovirus* minor core protein with the associated DR4-DQ8 haplotype (encoded by the DRB1*0401 and DQA1*0301-DQB1*0302 genes) (DRB1*0401, OR = 5.69, $p = 4.45 \times 10^{-15}$; DQA1*0301, OR = 2.55, $p = 2.12 \times 10^{-18}$; and DQB1*0302, OR = 3.14, $p = 4.17 \times 10^{-20}$) (Figure S3B). We observed a positive identification of the peptide core matching known deconvolution motifs, as well as a favorable binding prediction for the *Lactococcus* phage peptide to DR15 and the *Human mastadenovirus* peptide to DR4-DQ8 haplotypes. Similarly, the binding mode modeling of the peptide cores to the HLA-II complexes resulted in energetically favorable binding energy calculations and K_D in the nanomolar range (*Lactococcus* phage-DR15, 1.6×10^{-7} M; *Human mastadenovirus*-DR4/DQ8, 1.2×10^{-7} and 1.3×10^{-7} M, respectively). These results suggest that the identified HLA-peptide associations point to biologically relevant processes in which a specific HLA complex can preferentially bind and display the specific peptide sequence.

A second study-wide significant signal in our GWAS pointed to the *IGHV* region in chromosome 14 that encodes the IGHV domain. Here, we found 16 associated peptides in 11 leading loci within the region. The majority of SNPs (11/16) were located in non-coding regions around the *IGHV* gene, whereas *Ovis aries* casein protein (representing the primary sheep's milk allergy food allergen) was associated with a missense variant that changes glycine, a non-polar amino acid, for arginine, a positively charged amino acid. Next to the *Ovis aries* casein peptide, the top peptides associated with this region are bacteria-related (*Bacteroides uniformis*, *Blautia producta*, and *Lactobacillus plantarum*) or viral (influenza A, *Lactobacillus* phage, and Norwalk virus). The strongest association was observed in *Lactobacillus plantarum* (aggregation-promoting factor) and *Lactobacillus* phage (endolysin).

We found a third study-wide significant signal in the *FUT2* gene in chromosome 19. This gene status controls the secretion or non-secretion (homozygous for loss of function) of the H-antigen, an oligosaccharide. Thus, we subsequently ran the analysis in a dominant/recessive model to increase power and detected three study-wide significant peptides, all of which originally belonged to Norwalk virus polyproteins and were negatively associated with the same leading variant, *rs2251034* (A>G, 3′ UTR). This variant is in high linkage with an early-stop variant in *FUT2* that is known to stop the secretion of the H-antigen, rs601338 (A>G, $R^2 = 0.85$, 1000G, CEU population). FUT2 secretor status has been previously associated with multiple phenotypes, including infection susceptibility,[45] gut microbiome,[46,47] human milk oligosaccharides,[48] and cardiovascular traits.[49] Our finding supports the previously reported association between Norwalk virus susceptibility and FUT2 secretor status,[50] since this virus requires the H type 1 oligosaccharide ligand for successful attachment in the cell surface.

Although not reaching study-wide statistical significance, many other loci reached genome-wide significance ($5 \times 10^{-8} > p > 5.67 \times 10^{-11}$). We identified a total of 158 clumped variants associated with antibody-bound peptide profiles. From those, most polymorphisms were in intergenic regions (91), whereas 67 were annotated to their closest gene. Although no polymorphism was present in exons, they were present upstream, downstream, and in UTR and intronic regions. All 67 genes were uniquely associated with a single antibody-bound peptide. Some of the top associations include *MAML2* gene association to a *Ruminococcus* unknown protein ($p = 7.82 \times 10^{-10}$) and *ANKRD13C* association to *Blautia producta* ABC transporter ($p = 9.79 \times 10^{-10}$) or *Lactobacillus plantarum* WCFS1 and *TIGAR* ($p = 1.64 \times 10^{-9}$).

Similarly, we performed a GWAS meta-analysis at the co-occurrence module level (Table S2.4). As seen at the antibody-bound peptide level, two major GWAS signals were identified. IGHV was strongly associated with module 5, a class III module with a common motif in all peptides. Meanwhile, HLA-II was found to be associated with module 21 (a high similarity module of plant allergens), module 19 (a category III module with a highly conserved module), and module 5. Other genome-wide results that did not reach study-wide significance ($p > 2.27 \times 10^{-9}$) include associations between module 10 (characterized for bacterial flagellins) and the *GALNT13* gene ($p = 2.629 \times 10^{-8}$). This gene codes for a galactosyltransferase linked to host adaptation to pathogenic interactions.[51] In addition, module 9 (characterized by pollen allergens) was associated with *ESRP1* ($p = 4.03 \times 10^{-8}$), a gene implicated in proper skin barrier function, where defects have been linked with allergen response in respiratory tracts.[52] A subsequent HLA-imputed analysis not only supported the strong association of specific HLA variants

and module 21 ($p = 1.12 \times 10^{-16}$) but also showed (Bonferroni) significant ($p < 3.6 \times 10^{-6}$) associations to modules 13 (top $p = 2.17 \times 10^{-10}$), 14 (top $p = 2.05 \times 10^{-6}$), 19 (top $p = 8.41 \times 10^{-10}$), and 5 (top $p = 2.07 \times 10^{-10}$) (Table S2.5). Of these modules, 13, 19, and 5 all present a common sequence motif, whereas modules 21 and 14 are composed of highly conserved homologous sequences. This highlights that the presence of common motifs allows the binding of co-occurring proteins to the same HLA and *IGHV* variants.

Phenotypic and environmental effects on antibody-bound peptide enrichment

More than 200,000 bacterial antigens, including proteins originating from pathogenic, probiotic, and commensal gut microbiota species, were included in the peptide libraries. We therefore explored the relations between gut microbiome composition, analyzed by metagenomics sequencing, and the presence of antibody responses. To increase the power of the study, we performed taxonomic abundance-peptide associations in 1,051 LLD participants and then ran the meta-analysis including 137 IBD participants.[40] Neither the cohort-specific analysis nor the meta-analysis strongly supported taxonomy metagenomic associations with antibody-bound peptides (minimum FDR 0.52) (Table S2.6). These results were also in line with previous observations.[21] In addition, we quantified the abundance of a subset of 647 microbiome-derived peptides included in our PhIP-seq library in the available metagenomes (STAR Methods), we again did not find any strong association between the microbial abundance of those peptides and the presence or absence of the antibody-bound peptide.

To uncover the relationships of lifestyle and environmental factors with the antibody-bound peptide repertoire, we associated 84 available phenotypes (Table S1.2) with the presence/absence of antibody-bound peptide profiles in 1,437 LLD participants. Here, we uncovered 837 strongly supported associations between the presence of antibody-bound peptides and lifestyle and environmental factors (FDR < 0.05), covering 544 peptides and 48 different phenotypes (Figure 4A; Table S2.7). Phenotypic factors that were associated (after age, sex, and sequencing plate adjustment) with antibody-bound peptides included age (386 associations and not controlled for age), lymphocyte counts (101 associations and both absolute counts and cell proportions), neutrophil counts (86 associations and absolute counts and cell proportions), smoking (84 associations and both former and current smoking), sex (43 associations and not controlled for sex), allergies (35 associations, including any pollen, dust, or animals), autoantibodies (40 associations), and blood cholesterol concentrations (13 associations and both total cholesterol and LDL-cholesterol).

Of the 386 significant associations with age, 199 were positive and 187 were negative. Older age was associated with a higher prevalence of antibody-bound peptides from several herpes viruses (including CMV, EBV, and herpes simplex virus [HSV] 1 and 2), *Streptococcus* bacteria (in particular *S. pyogenes* and *S. dysgalactiae*), and several pathogenic bacteria (including *Shigella flexneri*, *Yersinia enterocolitica*, *Campylobacter* genus, and *Helicobacter pylori*). Younger individuals had higher frequencies of antibody-bound peptides related to particular viruses (including human rhinovirus serotype 2, influenza A virus, and enteroviruses), and bacteria, mainly *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Mycoplasma pneumoniae*, *Haemophilus influenzae*, and *Escherichia coli* (particularly antigens from the type III secretion system [T3SS] of serotype O157:H7). Younger individuals also showed more frequent antibody responses against alpha S1 casein proteins.

Sex demonstrated 43 significant enrichments (24 for males and 19 for females). Females exhibited more frequent antibody-bound peptides from *Lactobacillus acidophilus* and *Lactobacillus johnsonii*, both known inhabitants of the vaginal microbiome.[53,54] Antibody-bound peptide responses were particularly directed against *Lactobacillus* surface proteins, including S-layer proteins (SLPs, e.g., SlpA and SlpX proteins) and the peptidoglycan lysozyme N-acetylmuramidase, reproducing previous findings.[21] Females also demonstrated increased enrichment of EBV and CMV peptides. Males showed a higher prevalence of antibody-bound peptides from *Haemophilus influenzae* bacteria (e.g., serotype Rd KW20 or strain 3179B), also as previously described,[55,56] and of several peptides derived from *Streptococcus*, *Staphylococcus*, *Bacteroides*, and alphaherpesviruses (including HSV-1 and varicella zoster virus).

Associations between antibody-bound peptides and laboratory cell counts included both cell proportions and absolute cell quantifications, both of which appeared to be largely driven by antibody-bound peptides from CMV. Lymphocyte counts not only showed almost exclusively positive associations with CMV but also some with EBV, whereas the same antibody-bound peptides demonstrated many negative associations with neutrophil counts.

Smoking associations included associations to the current smoking status (41) (Figure 4B), ever smoking for at least a year (43), and parental smoking (7). Most associations were related to the higher prevalence of peptides belonging to enteroviruses, both rhinovirus and poliovirus. The relationship between smoking and rhinovirus infection has been previously described,[57] and thus, associations to other viral peptides belonging to enteroviruses could be due to cross-reactivity to homologous proteins. We also observed a consistently higher seroprevalence of EBV in smokers, which might be reactivated by smoking, as shown by an *in vitro* model.[58] In addition, there were increased antibody responses against miscellaneous respiratory pathogens, including several *Streptococcus* spp. On the other hand, flagellin antibody-bound peptides (*Roseburia*, *Lachnospiraceae*, *Eubacterium*, and *Clostridiales*) show a lower prevalence in smokers, as do *Escherichia* virulence factors (Figure 3B).

We used serological information about the presence of autoantibodies to identify bacterial and allergen peptides linked to the presence of these autoimmune antibodies (Figure 4C). Anti-cyclic citrullinated peptide (anti-CCP) antibody U/mL, a marker for rheumatoid arthritis, was positively associated with 23 antibody-bound peptides, including peptides derived from *Bacteroides*, *Parabacteroides*, *Prevotella*, *Streptococcus*, *Lactobacilli*, and *Porphyromonas gingivalis* bacteria. These findings correspond well with bacterial genera that are known to be altered in the microbiome of patients with anti-CCP-positive rheumatoid arthritis.[59] For instance, *Prevotella* might mimic autoantigens typical of rheumatoid arthritis,[60] an oral

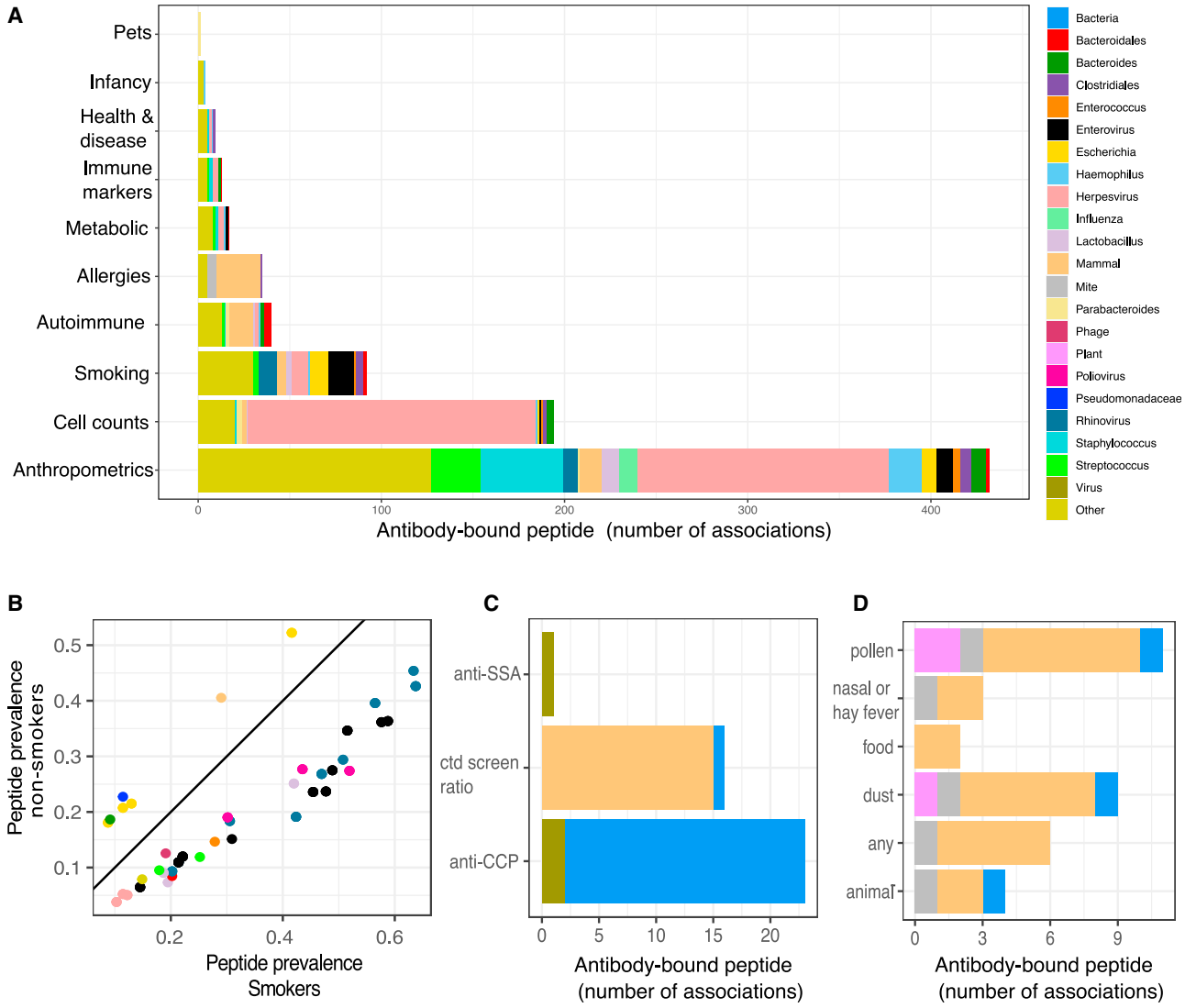


Figure 4. Phenotype-antibody-bound peptide associations

(A) Bar plot displaying the number of associations per phenotype (FDR < 0.05). Phenotypes are grouped in categories. Peptides associated with >5 phenotypes are grouped. Peptides associated with < 5 phenotypes are labeled "other."

(B) Smoking-linked antibody-bound peptide prevalence. x axis shows prevalence of peptides in smokers. y axis shows the prevalence in non-smokers. Colors of dots depict peptide taxonomy.

(C and D) Autoimmune- and allergy-specific association counts of antibody-bound peptides, per category. Bacterial peptides are binned as "bacteria." Viral peptides are binned as "virus." Autoantigens or antigens to casein are binned as "mammal." Plant peptides are binned as "plant." Anti-SSA, anti-Sjögren's-syndrome-related antigen A autoantibodies; anti-CTD, anti-connective tissue diseases screening ratio; anti-CCP, anti-cyclic citrullinated peptide.

Streptococcus bacteria isolate was seen to induce arthritis in arthritis-prone mice,[61] gut *Lactobacilli* are associated with rheumatoid arthritis dysbiosis,[62] and *P. gingivalis* can catalyze citrullination.[63] On the other hand, the connective tissue disease (CTD) screen panel, in which the total reactivity to a mixture of antigens associated with several autoimmune diseases is measured, was almost exclusively associated with increased antibody-bound peptide frequencies of alpha-S1-casein or kappa casein belonging to *Bos taurus* (cow), *Ovis aries* (sheep), *Bubalus bubalis* (buffalo), and *Capra hircus* (goat). Indeed, several autoimmune diseases such as celiac disease, juvenile idiopathic arthritis, and Ehlers-Danlos syndrome have been associated with mucosal reactivity against milk allergy, where the casein protein seems to be a regulator of the inflammatory response.[64,65] Anti-Sjögren's-syndrome-related antigen A antibodies (anti-SSA/anti-Ro), which are typical anti-nuclear antibodies associated to autoimmunity, were positively associated with an antibody-bound peptide representing thymidine kinase of EBV. This association has previously been described in the context of Sjögren's syndrome, in which anti-SSA autoantibodies and higher frequencies of serological EBV reactivation[66] are more frequently observed.

The strongest association to total cholesterol (mmol/L) was with an antibody-bound peptide of *Haemophilus parainfluenzae* strain

T3T1. Other bacterial peptides are also enriched with higher cholesterol concentrations, including *Streptococcus* or *Pseudomonadaceae*. We also observed an enrichment of viral peptides, such as rubeola, *Pneumoviridae*, HSV, and EBV. Many intracellular pathogens are known to use cholesterol drafts to successfully infect cells and to impair the regular cholesterol metabolism and the immune system.[67] We observed three associations between body-mass index (BMI) and antibody-bound peptides, all of which represented glycoprotein D of human alphaherpesviruses (HSV-1/HSV-2). Indeed, obesity has previously been associated with a higher prevalence of herpesvirus infections, in particular HSV-1, by promoting human adipogenesis.[68]

Finally, participants having any allergy (44.5% of participants) showed associations with six different antibody-bound peptides (Figure 4D). Using more-detailed questionnaires with information about different allergies such as dust, pollen, food, and others (Table S1.3), we identified 13 different peptides associated with at least one phenotype. As expected, the strongest association was observed for dust allergy, showing associations with antibody-bound peptides from the house dust mite *Dermatophagoides pteronyssinus* ($p = 2.93 \times 10^{-8}$). In addition, the most common associations were observed between casein proteins derived from cow, sheep, and buffalo milk, which were linked not only with food allergies but with almost all allergy types. Wheat allergens were linked with self-reported dust and pollen allergies. Additionally, we identified a couple of associations with influenza (higher prevalence with pollen allergy), bacterial flagellin associations with animal allergies, and *Shigella flexneri* with dust allergy. Previous analyses have linked dust mites with bacterial sensitization, although not for these specific lineages.[69] Importantly, several of these significant associations represent a linkage between common aeroallergens (e.g., pollen and dust) and food allergy (e.g., *Triticum aestivum* [wheat] and casein), recapitulating the frequent co-occurrence of allergen cross-reactivity.[70]

In addition to this analysis, and given the complexity of the data, we also used the PCs of the antibody-bound peptides as summaries of common antibody trends in the population. Looking at the top 100 antibody-bound peptide PCs, we identified 28 significant associations (FDR < 0.05) (Table S2.8). Cholesterol (both total and LDL concentrations) was positively associated with PC1, which is negatively loaded by many bacterial pathogens. Anti-CCP (U/mL) was positively associated with PC6 (loaded by several bacteria). Anti-CTD (U/mL) was negatively associated with PC12 (negatively loaded by casein) and PC45 (loaded by influenza and *H. pylori*). Several allergies were negatively associated with PC12 (negatively loaded by casein). Pet history was negatively associated with PC75 (negatively loaded by enteroviruses and positively loaded by *N. meningitidis*). Smoking was associated with enterovirus-loaded PCs and with PC33, loaded by the airway pathogen *P. aeruginosa*. The latter associations confirm the observed smoking-enterovirus relation and highlight another known association between *P. aeruginosa* and smokers.[71] Similarly, we also again saw associations of cell counts with CMV and allergies with casein. In addition, we observed a negative relation between bacterial infections and cholesterol concentrations, in line with a previous report.[72]

Common lifestyle and anthropometric parameters might help explain the co-occurrence of antibody-bound peptides. Thus, we additionally associated the co-occurrence modules (represented as eigengenes, see STAR Methods) with phenotypic information available for study participants (Table S2.9). This identified 21 significant associations (FDR < 0.05). The strongest positive associations were between smoking phenotypes and module 14 (characterized for enterovirus proteins), although other positive smoking associations were found with module 16 (EBV), as were negative associations with the flagellin module 2. Cell counts were associated with module 1, which is mainly enriched in CMV proteins, as expected. The presence of anti-CCP was positively associated with the presence of module 7, which is characterized by uncharacterized bacterial proteins with high similarity, whereas the CTD ratio was negatively associated with this module. Plant allergens from module 21 were associated with self-reported pollen allergy. C-reactive protein (CRP) concentration, a marker of inflammation, was positively linked with the presence of the herpes-enriched module 8. Female sex was associated with the EBV module 16. Finally, age was positively associated with modules 1, 8, and 16 (CMV, herpes simplex, and EBV, respectively), module 12 (*H. pylori*), and module 4 (mix of bacteria and self-antigens with the same motif) and negatively associated with module 14 (enteroviruses).

DISCUSSION

In this study, we aimed to characterize the antibody repertoire in the blood of a Dutch population and reveal which factors contribute to its variation. In particular, the factors that contribute to the generation of antibodies against microbiota and different allergens remain elusive. Here, we combined phenotypic and genetic information together with the immune interrogation of 2,815 common peptides from microbes, viruses, allergens, and self-peptides to study this variability. Using population, family, and longitudinal samples, we identified the antibody profile in the general population, assessed the stability of antibodies after 4 years, and investigated the effect of genetic and environmental factors on individual immune profiles.

The relation between genetics and antibody repertoire has been extensively described[36–39] but has been limited to a relatively small number of antibodies until now. PhIP-seq has enabled the investigation of the genetic contribution to antibody variability on a much broader scale, although it has so far mainly been investigated for viruses, toxins, and virulence factors[20,39] and not for other antigens such as allergens and gut microbiota-derived proteins. Here, we identified three genomic regions highly associated with the variability of antibody-bound peptide repertoires. As expected, we replicated the relation between HLA loci and antibody-bound peptide prevalence.[20,39,73,74] Through imputation of HLA alleles, amino acids, and structural variants, we also set out to uncover the specific HLA variations that allow the peptide to be displayed. Our structural simulations of the HLA alleles agree with the observed association patterns, supporting the hypothesis that the strong associations are due to HLA-display capabilities. We report specific HLA associations to more than 500 peptides at a high confidence level. These association data will be used in the future to further understand HLA-peptide interactions by modeling possible residue interactions. Similar to our findings, TCR variants have also been postulated to be selected by HLA

haplotypes.[75] Our findings also support previous observations, such as the association of *FUT2* and Norwalk virus peptides[50] that is explained by the attachment of the viral particle to the epithelia of *FUT2*-secretor cells.[76] We also observed association in the *IGHV* locus that was not previously reported in relation to antibody profiles. This association is in a complex genetic region as several genes with multiple isoforms coexist in the genome that are hard to address with microarrays.[77] In addition, we lack information about the rearrangements that this gene undergoes during B cell maturation. Nevertheless, although we cannot directly interpret the relation between variation and peptide recognition, this is a genetic region that is expected to contribute to antibody-bound peptide variability. However, our study did not identify the previously reported association of the nucleoredoxin gene (*NXN*) with *S. pyogenes'* M3 Streptolysin O (*SLO*) protein,[20] although we do find a weak positive association between *rs4968063* and the prevalence of this antibody-bound peptide in the combined LLD and IBD cohort ($p = 0.01$).

In the present study, we observe a lack of concordance between fecal microbial composition and PhIP-seq-based epitope repertoires, which is in line with findings from studies using the exact same library of antigens in a healthy population-based Israeli cohort and a disease-cohort consisting of patients with IBD.[21,40] The top associations do not present clear relationships between specific microbial taxa and antibody-bound peptides, which could be explained in various ways. First, this apparent lack of association might point to past events, such as microbial translocation, that may have triggered long-lasting immunity that was captured by PhIP-seq profiling,[78] whereas the respective bacteria have been cleared from the gut. This agrees with previous observations,[75] where IgG responses have been seen to occur predominantly for translocating bacteria, whereas IgA governs mucosal bacterial homeostasis. Second, there may have been a lack of resolution in the microbiome data. For example, some bacterial species commonly detected by metagenomics may have been accompanied by higher detection thresholds in PhIP-seq, whereas highly immunogenic antigen peptides may not be frequently detected by metagenomics sequencing.[21] In addition, the use of fecal microbiota as a proxy for the gut microbiota limits the characterization of local immune-microbiota interactions. Profiling mucosa-attached microbiota rather than fecal microbiome could have improved the antibody-bacteria concordance as locally residing (mucosal) microbial communities may elicit stronger immune responses that may also depend on the anatomical location within the intestines.[79] The coexistence of bacterial communities in different niches (luminal and mucosal) has been previously reported,[80] and it has been suggested that mucosal-associated bacteria might be a reservoir of bacteria that evolve to acquire translocating capabilities.

We also explored the relationship between peptide prevalence and various morphological, biochemical, and lifestyle factors. We observed that EBV and CMV were associated with lymphocyte and neutrophil counts. These findings are in accordance with observations of absolute lymphocytosis and neutropenia that constitute characteristic laboratory findings in individuals affected by EBV (infectious mononucleosis)[81,82] or CMV infections,[83,84] which may translate into altered immune cell proportions in the longer term. Antibody-bound peptides from EBV and a group of peptides identified to co-occur with EBV were also seen to be more prevalent in females than in males, which might be attributed to higher disease prevalence[85–87] or higher antibody titers.[88] We also identified a series of associations of allergies and allergens. Allergies are normally triggered by the epitope interaction with IgE antibodies. However, in this study, we mainly used IgG for immunoprecipitation since IgE are found in small amounts in serum and bind with relatively low affinity to the protein A/G coated magnetic beads employed for the immunoprecipitation. Previous studies have shown that allergens have the chance to bind both to IgG and IgE, although they might have different epitope preferences.[89] Thus, the allergen associations presented here should be interpreted with caution as they may differ from the classical pathway involved in allergy.

Using co-occurrence networks, we identified different peptide groups that normally belonged to the same taxa or orthologous structures in different taxa. In the context of the gut microbiome, a recent study highlights that T cell interactions with gut bacteria are largely strain-specific and that common epitopes tend to be recognized in multiple strains, which might be seen in our analysis through the lens of antibody-bound peptide co-occurrence.[90] However, the existence of modules with apparently unrelated peptides may indicate either a biological phenomenon or technical factors that we are not accounting for. Most of these co-occurrences of unrelated peptides could be attributed to the presence of common sequence motifs that might be recognized by the immune system. In modules including peptides belonging to bacteria, humans, and allergens, this might indicate a mechanism linking bacterial infections with the development of immune disorders through bacterial mimicry. We saw some examples of this in module 15, where a common motif is found in a human Chromodomain helicase DNA-binding protein, Ribosomal RNA-processing protein 8, and pollen allergens; in module 4, where a consistent motif is seen in bacteria and human junctional protein associated with coronary artery disease; and in module 5, where it links the presence of antibodies against idursulfase, a drug used in the treatment of Hunter syndrome, with bacteria and phages. Module 5 was also associated with variants in the *IGHV* gene, which might predispose carriers to the recognition of this motif and idursulfase allergy. On the other hand, phenotypic associations also allow us to conjecture about observed cryptic peptide co-occurrence. For instance, CMV peptides were seen to co-occur with several bacterial and plant peptides. Most of those peptides were associated with the same phenotypes, mainly blood cell leukocyte and granulocyte counts, age, and sex, meaning that the co-occurrence could be driven by those factors or that those phenotypes may mediate their co-occurrence.

All in all, although earlier individual studies found some of the associations we report, our large, widespread analysis represents a valuable resource for subsequent studies. MHC-peptide associations might clear up the complex HLA-peptide interactions for thousands of different peptides. Associations between phenotypes and antibody-bound peptides range from the expected (smoking with rhinovirus infection) to potentially relevant but unknown associations that warrant future studies (bacterial associations with autoimmunity markers or cholesterol associations with bacterial infections). Finally, the co-occurrence of, to all appearances, unrelated peptides comprising allergens,

pathogens, self-antigens, and commensal microbiota, and the ostensibly shared motifs among them, are findings that require further investigation and validation and might help elucidate the development of allergies[91] and autoimmunity.[92]

Limitations of the study

PhIP-seq is currently limited to linear epitopes and lacks post-translational modification information, and thus, new technologies or improvements of the current method (e.g., as previously shown[14]) are still to be developed. Similarly, the nature of the assay will also miss tridimensional structure information from the antigens that might be recognized by the antibodies. In addition to these technological issues, our relatively small sample size for genetic studies hampers an accurate estimation of antibody-bound peptide heritability and genetic correlation. It is also important to acknowledge that the antibody-bound peptides we identified mainly correspond to circulating IgG and may overlook other types of immunoglobulins or immunoglobulins not in systemic circulation. Finally, due to the mostly cross-sectional nature of the experimental design, it is hard to draw causal links from the associations we present, and further studies are needed to establish causality and dependence. We could not attempt to replicate most of the phenotypic associations here presented, since other cohorts lack the phenotypic breadth and antibody panels tested in this study. Validation of peptide presence with ELISA showed significant but imperfect correlations with antibody presence defined with PhIP-seq. Orthogonal assays are necessary to further support the observed correlations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human samples
- METHOD DETAILS
  - PhIP-Seq library design, preparation, sequencing and processing
  - Composition of the antigen library
  - Peptide antibody-binding enrichment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Antibody-bound peptides exploratory analysis
  - Antibody-bound peptide selection
  - Principal component analysis
  - Time and family distance analysis
  - Network analysis
  - Peptide similarity
  - Multiple sequence alignment information content
  - Motif discovery
  - CMV analysis
  - PhIP-Seq validation
  - Phenotype association analysis
  - Genotyping and imputation
  - Genetic preprocessing
  - Heritability and genetic correlation
  - Genome-wide association
  - Genetic meta-analysis
  - HLA imputation and association
  - Modeling of peptide presentation in HLA complexes
  - Metagenomic sequencing
  - Metagenomic processing
  - Microbiome-peptide association analysis
  - Microbiome meta-analysis
  - Microbial peptide quantification

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.immuni.2023.04.003.

CONSORTIA

The members of Lifelines DEEP Cohort are Lude Franke, Sasha Zhernakova, Jingyuan Fu, Morris Swertz, and Cisca Wijmenga.

ACKNOWLEDGMENTS

We thank K. Mc Intyre for English editing. The Lifelines Biobank initiative has been made possible by a subsidy from the Dutch Ministry of Health, Welfare and Sport; the Dutch Ministry of Economic Affairs; the University Medical Center Groningen; the University of Groningen; and the Northern Provinces of the Netherlands. The authors wish to acknowledge the services of the Lifelines Cohort Study, the contributing research centers delivering data to Lifelines, and all the study participants. Funding: the researchers participating in this project are supported by several funding agencies. J.F. and A.Z. are supported by the Netherlands Heart Foundation (IN-CONTROL CVON grants 2012-03 and 2018-27, respectively). J.F., S.W., and C.W. are supported by the Netherlands Organ-on-Chip Initiative, a Netherlands Organization for Scientific Research (NWO) Gravitation project (024.003.001) funded by the Ministry of Education, Culture, and Science of the Government of The Netherlands. J.F., A.K., and A.Z. are supported by the Gravitation Exposome-NL, a NWO Gravitational project (024.004.017), funded by the Ministry of Education, Culture, and Science of the Government of The Netherlands. The Seerave Foundation and the Netherlands Organization for Scientific Research support R.K.W. J.F. is supported by both NWO-VIDI (864.13.013) and NWO-VICI (VI.C.202.022). A.Z. is supported by NWO-VIDI (016.178.056). I.J. supported by the NWO-VIDI (016.171.047). C.W. is supported by the NWO Spinoza Prize SPI 92-266 and the European Research Council (ERC) (FP7/2007-2013/ERC Advanced Grant 2012-322698). ERC Starting Grant 715772 supports A.Z.; ERC Consolidator Grant (grant agreement no. 101001678) supports J.F.; the RuG Investment Agenda Grant Personalized Health supports C.W.; A.R.B. (grant no. 17-57) and T.S. (grant no. 17-34) hold scholarships from the Junior Scientific Masterclass, University of Groningen. E.S. is supported by grants from the European Research Council, the Israel Science Foundation and by the Seerave Foundation. T.V. gratefully acknowledges support from the Austrian Science Fund (FWF, Erwin Schrödinger fellowship J4256). I.J. and A.Z. were supported by a Rosalind Franklin fellowship from the University of Groningen. A.Z. is supported by the EU Horizon Europe Program grant INITIALISE (101094099).

AUTHOR CONTRIBUTIONS

Conceptualization, S.A.-S., A.R.B., A.Z., J.F., R.K.W., I.J., T.V., and S.L.; methodology, S.A.-S., A.K., S.H., A.J.R.-M., A.V.V., A.R.B., T.S., S.L., T.V., S.K., and I.N.K.; investigation, S.A.-S., A.R.B., A.K., S.H., A.R., A.V.V., T.S., T.V., A.Z., J.F., and I.J.; funding acquisition, A.Z., J.F., R.K.W., and C.W.; supervision, A.Z., J.F., and R.K.W.; writing – original draft, S.A.-S. and A.R.B.; writing – review & editing, all coauthors.

DECLARATION OF INTERESTS

R.K.W. acted as consultant for Takeda, received unrestricted research grants from Takeda, Johnson & Johnson, Tramedico, and Ferring, and received speaker fees from MSD, Abbvie, and Janssen Pharmaceuticals.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: December 7, 2021
Revised: December 13, 2022
Accepted: April 6, 2023
Published: May 9, 2023

REFERENCES

1. Cooper, M.D., and Alder, M.N. (2006). The evolution of adaptive immune systems. Cell *124*, 815–822.
2. Burkholder, W.F., Newell, E.W., Poidinger, M., Chen, S., and Fink, K. (2017). Deep sequencing in infectious diseases: immune and pathogen repertoires for the improvement of patient outcomes. Front. Immunol. *8*, 593.
3. Ganusov, V.V., and De Boer, R.J. (2007). Do most lymphocytes in humans really reside in the gut? Trends Immunol. *28*, 514–518.
4. Hoehn, K.B., Fowler, A., Lunter, G., and Pybus, O.G. (2016). The diversity and molecular evolution of B-cell receptors during infection. Mol. Biol. Evol. *33*, 1147–1157.
5. Galson, J.D., Schaetzle, S., Bashford-Rogers, R.J.M., Raybould, M.I.J., Kovaltsuk, A., Kilpatrick, G.J., Minter, R., Finch, D.K., Dias, J., James, L.K., et al. (2020). Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. Front. Immunol. *11*, 605170.
6. Goldstein, L.D., Chen, Y.-J.J., Wu, J., Chaudhuri, S., Hsiao, Y.-C., Schneider, K., Hoi, K.H., Lin, Z., Guerrero, S., Jaiswal, B.S., et al. (2019). Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. Commun. Biol. *2*, 304.
7. Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S.-W., Sollid, L.M., Teichmann, S.A., and Stubbington, M.J.T. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat. Methods *15*, 563–565.
8. Kim, D., and Park, D. (2019). Deep sequencing of B cell receptor repertoire. BMB Rep. *52*, 540–547.
9. Atak, A., Mukherjee, S., Jain, R., Gupta, S., Singh, V.A., Gahoi, N., K P, M., and Srivastava, S. (2016). Protein microarray applications: autoantibody detection and posttranslational modification. Proteomics *16*, 2557–2569.
10. Yu, X., Song, L., Petritis, B., Bian, X., Wang, H., Viloria, J., Park, J., Bui, H., Li, H., Wang, J., et al. (2017). Multiplexed nucleic acid programmable protein arrays. Theranostics *7*, 4057–4070.
11. Larman, H.B., Zhao, Z., Laserson, U., Li, M.Z., Ciccia, A., Gakidis, M.A.M., Church, G.M., Kesari, S., Leproust, E.M., Solimini, N.L., et al. (2011). Autoantigen discovery with a synthetic human peptidome. Nat. Biotechnol. *29*, 535–541.
12. Mohan, D., Wansley, D.L., Sie, B.M., Noon, M.S., Baer, A.N., Laserson, U., and Larman, H.B. (2018). PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. Nat. Protoc. *13*, 1958–1978.
13. Larman, H.B., Laserson, U., Querol, L., Verhaeghen, K., Solimini, N.L., Xu, G.J., Klarenbeek, P.L., Church, G.M., Hafler, D.A., Plenge, R.M., et al. (2013). PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. J. Autoimmun. *43*, 1–9.
14. Román-Meléndez, G.D., Monaco, D.R., Montagne, J.M., Quizon, R.S., Konig, M.F., Astatke, M., Darrah, E., and Larman, H.B. (2021). Citrullination of a phage displayed human peptidome library reveals the fine specificities of rheumatoid arthritis-associated autoantibodies. EBioMedicine *71*, 103506.
15. Eshleman, S.H., Laeyendecker, O., Kammers, K., Chen, A., Sivay, M.V., Kottapalli, S., Sie, B.M., Yuan, T., Monaco, D.R., Mohan, D., et al. (2019). Comprehensive profiling of HIV antibody evolution. Cell Rep. *27*, 1422–1433.e4.
16. Finton, K.A.K., Friend, D., Jaffe, J., Gewe, M., Holmes, M.A., Larman, H.B., Stuart, A., Larimore, K., Greenberg, P.D., Elledge, S.J., et al. (2014). Ontogeny of recognition specificity and functionality for the broadly neutralizing anti-HIV antibody 4E10. PLoS Pathog. *10*, e1004403.
17. Mina, M.J., Kula, T., Leng, Y., Mamie, L., de Vries, R.D., Knip, M., Siljander, H., Rewers, M., Choy, D.F., Wilson, M.S., et al. (2019). Measles virus infection diminishes preexisting antibodies that offer protection from other pathogens. Science *366*, 599–606.
18. Shrock, E., Fujimura, E., Kula, T., Timms, R.T., Lee, I.H., Leng, Y., Robinson Matthew, L., Sie, B.M., Li, M.Z., Chen, Y., et al. (2020). Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. Science *370*, eabd4250.
19. Xu, G.J., Kula, T., Xu, Q., Li, M.Z., Vernon, S.D., Ndung'u, T., Ruxrungtham, K., Sanchez, J., Brander, C., Chung, R.T., et al. (2015). Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. Science *348*, aaa0698.
20. Angkeow, J.W., Monaco, D.R., Chen, A., Venkataraman, T., Jayaraman, S., Valencia, C., Sie, B.M., Liechti, T., Farhadi, P.N., Funez-dePagnier, G., et al. (2022). Phage display of environmental protein toxins and virulence factors reveals the prevalence, persistence, and genetics of antibody responses. Immunity *55*, 1051–1066.e4.
21. Vogl, T., Klompus, S., Leviatan, S., Kalka, I.N., Weinberger, A., Wijmenga, C., Fu, J., Zhernakova, A., Weersma, R.K., and Segal, E. (2021). Population-wide diversity and stability of serum antibody epitope repertoires against human microbiota. Nat. Med. *27*, 1442–1450.
22. Ter Horst, R., Jaeger, M., Smeekens, S.P., Oosting, M., Swertz, M.A., Li, Y., Kumar, V., Diavatopoulos, D.A., Jansen, A.F.M., Lemmers, H., et al. (2016). Host and environmental factors influencing individual human cytokine responses. Cell *167*, 1111–1124.e13.
23. Aguirre-Gamboa, R., Joosten, I., Urbano, P.C.M., van der Molen, R.G., van Rijssen, E., van Cranenbroek, B., Oosting, M., Smeekens, S., Jaeger, M., Zorro, M., et al. (2016). Differential effects of environmental and genetic factors on T and B cell immune traits. Cell Rep. *17*, 2474–2487.
24. Krishna, C., Chowell, D., Gönen, M., Elhanati, Y., and Chan, T.A. (2020). Genetic and environmental determinants of human TCR repertoire diversity. Immun. Ageing *17*, 26.
25. Nielsen, S.C.A., Roskin, K.M., Jackson, K.J.L., Joshi, S.A., Nejad, P., Lee, J.Y., Wagar, L.E., Pham, T.D., Hoh, R.A., Nguyen, K.D., et al. (2019). Shaping of infant B cell receptor repertoires by environmental factors and infectious disease. Sci. Transl. Med. *11*, eaat2004.
26. de Bourcy, C.F.A., Angel, C.J.L., Vollmers, C., Dekker, C.L., Davis, M.M., and Quake, S.R. (2017). Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. Proc. Natl. Acad. Sci. USA *114*, 1105–1110.
27. Tigchelaar, E.F., Zhernakova, A., Dekens, J.A.M., Hermes, G., Baranska, A., Mujagic, Z., Swertz, M.A., Muñoz, A.M., Deelen, P., Cénit, M.C., et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open *5*, e006772.
28. Leviatan, S., Vogl, T., Klompus, S., Kalka, I.N., Weinberger, A., and Segal, E. (2022). Allergenic food protein consumption is associated with systemic IgG antibody responses in non-allergic individuals. Immunity *55*, 2454–2469.e6.
29. Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A., et al.

(2015). The immune epitope database (IEDB) 3.0. Nucleic Acids Res. *43*, D405–D412.

30. Korndewal, M.J., Mollema, L., Tcherniaeva, I., van der Klis, F., Kroes, A.C.M., Oudesluys-Murphy, A.M., Vossen, A.C.T.M., and de Melker, H.E. (2015). Cytomegalovirus infection in the Netherlands: seroprevalence, risk factors, and implications. J. Clin. Virol. *63*, 53–58.
31. Erles, K., Sebökovà, P., and Schlehofer, J.R. (1999). Update on the prevalence of serum antibodies (IgG and IgM) to adeno-associated virus (AAV). J. Med. Virol. *59*, 406–411.
32. Hendrikx, L.H., Oztürk, K., de Rond, L.G.H., Veenhoven, R.H., Sanders, E.A.M., Berbers, G.A.M., and Buisman, A.-M. (2011). Identifying long-term memory B-cells in vaccinated children despite waning antibody levels specific for Bordetella pertussis proteins. Vaccine *29*, 1431–1437.
33. Kontio, M., Jokinen, S., Paunio, M., Peltola, H., and Davidkin, I. (2012). Waning antibody levels and avidity: implications for MMR vaccine-induced protection. J. Infect. Dis. *206*, 1542–1548.
34. Genome; Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.
35. Kim, J., Park, M.R., Kim, D.S., Lee, J.O., Maeng, S.H., Cho, S.Y., Han, Y., Ahn, K., and Jin, D.K. (2013). IgE-mediated anaphylaxis and allergic reactions to idursulfase in patients with Hunter syndrome. Allergy *68*, 796–802.
36. Grundbacher, F.J. (1974). Heritability estimates and genetic and environmental correlations for the human immunoglobulins G, M, and A. Am. J. Hum. Genet. *26*, 1–12.
37. Kalff, M.W., and Hijmans, W. (1969). Serum immunoglobulin levels in twins. Clin. Exp. Immunol. *5*, 469–477.
38. Rowe, D.S., Boyle, J.A., and Buchanan, W.W. (1968). Plasma immunoglobulin concentrations in twins. Clin. Exp. Immunol. *3*, 233–244.
39. Venkataraman, T., Valencia, C., Mangino, M., Morgenlander, W., Clipman, S.J., Liechti, T., Valencia, A., Christofidou, P., Spector, T., Roederer, M., et al. (2022). Analysis of antibody binding specificities in twin and SNP-genotyped cohorts reveals that antiviral antibody epitope selection is a heritable trait. Immunity *55*, 174–184.e5.
40. Bourgonje, A.R., Andreu-Sánchez, S., Thomas, V., Hu, S., Vila, A.V., Gacesa, R., Leviatan, S., Kurilshikov, A., Shelley, K., Kalka, I.N., et al. (2023). Phage-display immunoprecipitation sequencing of the antibody epitope repertoire in inflammatory bowel disease reveals distinct antibody. Immunity *56*. https://doi.org/10.1016/j.immuni.2023.04.017.
41. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.
42. Lázár-Molnár, E., and Snyder, M. (2018). The role of human leukocyte antigen in celiac disease diagnostics. Clin. Lab. Med. *38*, 655–668.
43. Noble, J.A., and Valdes, A.M. (2011). Genetics of the HLA region in the prediction of type 1 diabetes. Curr. Diab. Rep. *11*, 533–542.
44. Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W.H., Peters, B., and Nielsen, M. (2020). Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. J. Proteome Res. *19*, 2304–2315.
45. Tian, C., Hromatka, B.S., Kiefer, A.K., Eriksson, N., Noble, S.M., Tung, J.Y., and Hinds, D.A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat. Commun. *8*, 599.
46. Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., Le Roy, C.I., Raygoza Garay, J.A., Finnicum, C.T., Liu, X., et al. (2021). Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat. Genet. *53*, 156–165.
47. Lopera-Maya, E.A., Kurilshikov, A., van der Graaf, A., Hu, S., Andreu-Sánchez, S., Chen, L., Vila, A.V., Gacesa, R., Sinha, T., Collij, V., et al. (2022). Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. Nat. Genet. *54*, 143–151.
48. Williams, J.E., McGuire, M.K., Meehan, C.L., McGuire, M.A., Brooker, S.L., Kamau-Mbuthia, E.W., Kamundia, E.W., Mbugua, S., Moore, S.E., Prentice, A.M., et al. (2021). Key genetic variants associated with variation of milk oligosaccharides from diverse human populations. Genomics *113*, 1867–1875.
49. Zhernakova, D.V., Le, T.H., Kurilshikov, A., Atanasovska, B., Bonder, M.J., Sanna, S., Claringbould, A., Võsa, U., Deelen, P., Franke, L., et al. (2018). Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. Nat. Genet. *50*, 1524–1532.
50. Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., and Baric, R. (2003). Human susceptibility and resistance to Norwalk virus infection. Nat. Med. *9*, 548–553.
51. Gagneux, P., and Varki, A. (1999). Evolutionary considerations in relating oligosaccharide diversity to biological function. Glycobiology *9*, 747–755.
52. Bebee, T.W., Park, J.W., Sheridan, K.I., Warzecha, C.C., Cieply, B.W., Rohacek, A.M., Xing, Y., and Carstens, R.P. (2015). The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. eLife *4*, e08954.
53. Davoren, M.J., Liu, J., Castellanos, J., Rodríguez-Malavé, N.I., and Schiestl, R.H. (2019). A novel probiotic, Lactobacillus johnsonii 456, resists acid and can persist in the human gut beyond the initial ingestion period. Gut Microbes *10*, 458–480.
54. Integrative HMP (iHMP) Research Network Consortium (2019). The integrative human microbiome project. Nature *569*, 641–648.
55. Angkeow, J.W., Monaco, D.R., Chen, A., Venkataraman, T., Jayaraman, S., Valencia, C., Sie, B.M., Liechti, T., Farhadi, P.N., Funez-dePagnier, G., et al. (2021). Prevalence, persistence, and genetics of antibody responses to protein toxins and virulence factors. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.01.462481.
56. Kurtti, P., Isoaho, R., Von Hertzen, L., Keistinen, T., Kivelâ, S.-L., and Leinonen, M. (1997). Influence of age, gender and smoking on Streptococcus pneumoniae, Haemophilus influenzae and Moraxella (Branhamella) catarrhalis antibody titres in an elderly population. Scand. J. Infect. Dis. *29*, 485–489.
57. Cohen, S., Tyrrell, D.A., Russell, M.A., Jarvis, M.J., and Smith, A.P. (1993). Smoking, alcohol consumption, and susceptibility to the common cold. Am. J. Public Health *83*, 1277–1283.
58. Xu, F.-H., Xiong, D., Xu, Y.-F., Cao, S.-M., Xue, W.-Q., Qin, H.-D., Liu, W.-S., Cao, J.-Y., Zhang, Y., Feng, Q.-S., et al. (2012). An epidemiological and molecular study of the relationship between smoking, risk of nasopharyngeal carcinoma, and Epstein-Barr virus activation. J. Natl. Cancer Inst. *104*, 1396–1410.
59. Bodkhe, R., Balakrishnan, B., and Taneja, V. (2019). The role of microbiome in rheumatoid arthritis treatment. Ther. Adv. Musculoskelet. Dis. *11*, 1759720X19844632.
60. Pianta, A., Chiumento, G., Ramsden, K., Wang, Q., Strle, K., Arvikar, S., Costello, C.E., and Steere, A.C. (2021). Identification of novel, immunogenic HLA–DR-presented Prevotella copri peptides in patients with rheumatoid arthritis. Arthritis Rheumatol. *73*, 2200–2205.
61. Moentadj, R., Wang, Y., Bowerman, K., Rehaume, L., Nel, H., O Cuiv, P., Stephens, J., Baharom, A., Maradana, M., Lakis, V., et al. (2021). Streptococcus species enriched in the oral cavity of patients with RA are a source of peptidoglycan-polysaccharide polymers that can induce arthritis in mice. Ann. Rheum. Dis. *80*, 573–581.
62. Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X., Li, J., Tang, L., Li, Y., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat. Med. *21*, 895–905.
63. Lundberg, K., Wegner, N., Yucel-Lindberg, T., and Venables, P.J. (2010). Periodontitis in RA—the citrullinated enolase connection. Nat. Rev. Rheumatol. *6*, 727–730.

64. Cutts, R.M., Meyer, R., Thapar, N., Rigby, K., Schwarz, C., Mailliard, S., and Shah, N. (2012). Gastrointestinal food allergies in children with Ehlers Danlos type 3 syndrome. J. Allergy Clin. Immunol. *129*, AB34.

65. Kristjánsson, G., Venge, P., and Hällgren, R. (2007). Mucosal reactivity to cow's milk protein in coeliac disease. Clin. Exp. Immunol. *147*, 449–455.

66. Fox, R.I., Luppi, M., Kang, H.I., and Pisa, P. (1991). Reactivation of Epstein-Barr virus in Sjögren's syndrome. Springer Semin. Immunopathol. *13*, 217–231.

67. Sviridov, D., and Bukrinsky, M. (2014). Interaction of pathogens with host cholesterol metabolism. Curr. Opin. Lipidol. *25*, 333–338.

68. Hasan, M.R., Rahman, M., Khan, T., Saeed, A., Sundararaju, S., Flores, A., Hawken, P., Rawat, A., Elkum, N., Hussain, K., et al. (2021). Virome-wide serological profiling reveals association of herpesviruses with obesity. Sci. Rep. *11*, 2562.

69. Dzoro, S., Mittermann, I., Resch-Marat, Y., Vrtala, S., Nehr, M., Hirschl, A.M., Wikberg, G., Lundeberg, L., Johansson, C., Scheynius, A., et al. (2018). House dust mites as potential carriers for IgE sensitization to bacterial antigens. Allergy *73*, 115–124.

70. Popescu, F.-D. (2015). Cross-reactivity between aeroallergens and food allergens. World J. Methodol. *5*, 31–50.

71. Chien, J., Hwang, J.H., Nilaad, S., Masso-Silva, J.A., Jeong Ahn, S., McEachern Elisa, K., Moshensky, A., Byun, M.K., and Crotty Alexander, L.E. (2020). Cigarette smoke exposure promotes virulence of Pseudomonas aeruginosa and induces resistance to neutrophil killing. Infect. Immun. *88*. e00527–e00520.

72. Bartlett, S., Gemiarto, A.T., Ngo, M.D., Sajiir, H., Hailu, S., Sinha, R., Foo, C.X., Kleynhans, L., Tshivhula, H., Webber, T., et al. (2020). GPR183 regulates interferons, autophagy, and bacterial growth during Mycobacterium tuberculosis infection and is associated with TB disease severity. Front. Immunol. *11*, 601534.

73. Kachuri, L., Francis, S.S., Morrison, M.L., Wendt, G.A., Bossé, Y., Cavazos, T.B., Rashkin, S.R., Ziv, E., and Witte, J.S. (2020). The landscape of host genetic factors involved in immune response to common viral infections. Genome Med. *12*, 93.

74. Scepanovic, P., Alanio, C., Hammer, C., Hodel, F., Bergstedt, J., Patin, E., Thorball, C.W., Chaturvedi, N., Charbit, B., Abel, L., et al. (2018). Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. Genome Med. *10*, 59.

75. Ishigaki, K., Lagattuta, K.A., Luo, Y., James, E.A., Buckner, J.H., and Raychaudhuri, S. (2022). HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. Nat. Genet. *54*, 393–402.

76. Marionneau, S., Ruvoën, N., Le Moullac-Vaidye, B., Clement, M., Cailleau-Thomas, A., Ruiz-Palacois, G., Huang, P., Jiang, X., and Le Pendu, J. (2002). Norwalk virus binds to histo-blood group antigens present on gastroduodenal epithelial cells of secretor individuals. Gastroenterology *122*, 1967–1977.

77. Watson, C.T., and Breden, F. (2012). The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. *13*, 363–373.

78. Marchix, J., Goddard, G., and Helmrath, M.A. (2018). Host-Gut Microbiota Crosstalk in Intestinal Adaptation. Cell. Mol. Gastroenterol. Hepatol. *6*, 149–162.

79. Christmann, B.S., Abrahamsson, T.R., Bernstein, C.N., Duck, L.W., Mannon, P.J., Berg, G., Björkstén, B., Jenmalm, M.C., and Elson, C.O. (2015). Human seroreactivity to gut microbiota antigens. J. Allergy Clin. Immunol. *136*, 1378–1386.e1.

80. Yang, Y., Nguyen, M., Khetrapal, V., Sonnert, N.D., Martin, A.L., Chen, H., Kriegel, M.A., and Palm, N.W. (2022). Within-host evolution of a gut pathobiont facilitates liver translocation. Nature *607*, 563–570.

81. Fisher, B.D. (1973). Neutropenia in infectious mononucleosis. N. Engl. J. Med. *288*, 633.

82. Hudnall, S.D., David Hudnall, S., Patel, J., Schwab, H., and Martinez, J. (2003). Comparative immunophenotypic features of EBV-positive and EBV-negative atypical lymphocytosis. Cytometry *55b*, 22–28.

83. Lima, C.S.P., Paula, E.V., Takahashi, T., Saad, S.T.O., Lorand-Metze, I., and Costa, F.F. (2006). Causes of incidental neutropenia in adulthood. Ann. Hematol. *85*, 705–709.

84. Solana, R., Tarazona, R., Aiello, A.E., Akbar, A.N., Appay, V., Beswick, M., Bosch, J.A., Campos, C., Cantisán, S., Cicin-Sain, L., et al. (2012). CMV and immunosenescence: from basics to clinics. Immun. Ageing *9*, 23.

85. Kuri, A., Jacobs, B.M., Vickaryous, N., Pakpoor, J., Middeldorp, J., Giovannoni, G., and Dobson, R. (2020). Epidemiology of Epstein-Barr virus infection and infectious mononucleosis in the United Kingdom. BMC Public Health *20*, 912.

86. Crawford, D.H., Swerdlow, A.J., Higgins, C., McAulay, K., Harrison, N., Williams, H., Britton, K., and Macsween, K.F. (2002). Sexual history and Epstein-Barr virus infection. J. Infect. Dis. *186*, 731–736.

87. Winter, J.R., Taylor, G.S., Thomas, O.G., Jackson, C., Lewis, J.E.A., and Stagg, H.R. (2020). Factors associated with cytomegalovirus serostatus in young people in England: a cross-sectional study. BMC Infect. Dis. *20*, 875.

88. Keane, J.T., Afrasiabi, A., Schibeci, S.D., Fewings, N., Parnell, G.P., Swaminathan, S., and Booth, D.R. (2021). Gender and the sex hormone estradiol affect multiple sclerosis risk gene expression in Epstein-Barr virus-infected B cells. Front. Immunol. *12*, 732694.

89. Monaco, D.R., Sie, B.M., Nirschl, T.R., Knight, A.C., Sampson, H.A., Nowak-Wegrzyn, A., Wood, R.A., Hamilton, R.G., Frischmeyer-Guerrerio, P.A., and Larman, H.B. (2021). Profiling serum antibodies with a pan allergen phage library identifies key wheat allergy epitopes. Nat. Commun. *12*, 379.

90. Nagashima, K., Zhao, A., Atabakhsh, K., Weakley, A., Jain, S., Meng, X., Cheng, A.G., Wang, M., Higginbottom, S., Alex, D., et al. (2022). Mapping the T cell repertoire to a complex gut bacterial community. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.04.490632.

91. Kearney, J.F., Patel, P., Stefanov, E.K., and King, R.G. (2015). Natural antibody repertoires: development and functional role in inhibiting allergic airway disease. Annu. Rev. Immunol. *33*, 475–504.

92. Elkon, K., and Casali, P. (2008). Nature and functions of autoantibodies. Nat. Clin. Pract. Rheumatol. *4*, 491–498.

93. Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Gwen, F., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science *352*, 565–569.

94. Imhann, F., Bonder, M.J., Vich Vila, A., Fu, J., Mujagic, Z., Vork, L., Tigchelaar, E.F., Jankipersadsing, S.A., Cenit, M.C., Harmsen, H.J., et al. (2016). Proton pump inhibitors affect the gut microbiome. Gut *65*, 740–748.

95. Hu, S., Vich Vila, A., Gacesa, R., Collij, V., Stevens, C., Fu, J.M., Wong, I., Talkowski, M.E., Rivas, M.A., Imhann, F., et al. (2021). Whole exome sequencing analyses reveal gene-microbiota interactions in the context of IBD. Gut *70*, 285–296.

96. Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J.L., Dotinga, A., Vonk, J.M., van Dijk, F., van Zon, S.K.R., Wijmenga, C., Wolffenbuttel, B.H.R., et al. (2015). Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int. J. Epidemiol. *44*, 1172–1180.

97. Lambers, W., Arends, S., Roozendaal, C., Brouwer, E., Bootsma, H., Westra, J., and de Leeuw, K. (2021). Prevalence of systemic lupus erythematosus-related symptoms assessed by using the Connective Tissue Disease Screening Questionnaire in a large population-based cohort. Lupus Sci. Med. *8*, e000555.

98. van Zanten, A., Arends, S., Roozendaal, C., Limburg, P.C., Maas, F., Trouw, L.A., Toes, R.E.M., Huizinga, T.W.J., Bootsma, H., and Brouwer, E. (2017). Presence of anticitrullinated protein antibodies in a

large population-based cohort from the Netherlands. Ann. Rheum. Dis. *76*, 1184–1190.

99. Imhann, F., Van der Velde, K.J., Barbieri, R., Alberts, R., Voskuil, M.D., Vich Vila, A., Collij, V., Spekhorst, L.M., Van der Sloot, K.W.J., Peters, V., et al. (2019). The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. BMC Gastroenterol. *19*, 5.
100. Dixon, P. (2003). VEGAN, a package of R functions for community ecology. J. Veg. Sci. *14*, 927–930.
101. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal. Complex Syst. *1695*, 1–9.
102. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.
103. Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. *4*, 17.
104. Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biol. Rev. *57*, 669–689.
105. van Borkulo, C.D., Borsboom, D., Epskamp, S., Blanken, T.F., Boschloo, L., Schoevers, R.A., and Waldorp, L.J. (2014). A new method for constructing networks from binary data. Sci. Rep. *4*, 5918.
106. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.
107. Wilbur, W.J., and Lipman, D.J. (1983). Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA *80*, 726–730.
108. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
109. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.
110. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. *57*, 289–300.
111. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
112. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
113. Deelen, P., Bonder, M.J., van der Velde, K.J., Westra, H.-J., Winder, E., Hendriksen, D., Franke, L., and Swertz, M.A. (2014). Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res. Notes *7*, 901.
114. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
115. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
116. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
117. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540–2542.
118. Imhann, F., Vich Vila, A., Bonder, M.J., Fu, J., Gevers, D., Visschedijk, M.C., Spekhorst, L.M., Alberts, R., Franke, L., van Dullemen, H.M., et al. (2018). Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut *67*, 108–119.
119. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
120. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. Genome Biol. *17*, 122.
121. Alexander, T.A., and Machiela, M.J. (2020). LDpop: an interactive online tool to calculate and visualize geographic LD patterns. BMC Bioinformatics *21*, 14.
122. Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics *31*, 3555–3557.
123. Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.-M., Concannon, P.J., Rich, S.S., Raychaudhuri, S., and de Bakker, P.I.W. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One *8*, e64683.
124. Robinson, J., Barker, D.J., Georgiou, X., Cooper, M.A., Flicek, P., and Marsh, S.G.E. (2020). IPD-IMGT/HLA database. Nucleic Acids Res. *48*, D948–D955.
125. Zhou, P., Jin, B., Li, H., and Huang, S.-Y. (2018). HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. Nucleic Acids Res. *46*, W443–W450.
126. Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). Haddock: a protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. *125*, 1731–1737.
127. Adasme, M.F., Linnemann, K.L., Bolz, S.N., Kaiser, F., Salentin, S., Haupt, V.J., and Schroeder, M. (2021). PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Res. *49*, W530–W534.
128. Honorato, R.V., Koukos, P.I., Jiménez-García, B., Tsaregorodtsev, A., Verlato, M., Giachetti, A., Rosato, A., and Bonvin, A.M.J.J. (2021). Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. Front. Mol. Biosci. *8*, 729513.
129. Vangone, A., and Bonvin, A.M. (2015). Contacts-based prediction of binding affinity in protein-protein complexes. eLife *4*, e07454.
130. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
131. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
132. Beghini, F., McIver, L.J., Blanco Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife *10*, e65088.
133. Schwarzer, G. (2007). meta: an R package for meta-analysis. R News *7*, 40–45.
134. Kaminski, J., Gibson, M.K., Franzosa, E.A., Segata, N., Dantas, G., and Huttenhower, C. (2015). High-specificity targeted functional profiling in microbial communities with ShortBRED. PLoS Comput. Biol. *11*, e1004557.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
HRP conjugated anti human IgG antibody	Southern Biotech	Cat#2042-05; AB_2795660
anti human IgE antibody	Southern Biotech	Cat#9250-05; AB_2796719
mouse anti-human IgG Fc-BIOT	Southern Biotech	Cat#9040-08; AB_2796600
goat anti-human IgA-BIOT	Southern Biotech	Cat# 2050-08; AB_2795706
Bacterial and Virus Strains		
T7Select 10-3 cloning kit	Merck	Cat#70550-3
Biological Samples		
1,778 serum samples of 1,437 individuals	Tigchelaar et al.[27]	N/A
Chemicals, Peptides, and Recombinant Proteins		
IPEGAL CA 630	Sigma-Aldrich	Cat#I3021
Protein A magnetic beads	Thermo Fisher Scientific	Cat#10008D
Protein G magnetic beads	Thermo Fisher Scientific	Cat#10009D
1-Step™ Turbo TMB-ELISA Substrate Solution	Rhenium	Cat#TS-34022
Q5 polymerase	New England Biolabs	Cat#M0493L
Bovine Serum Albumin, heat shock fraction, pH 7, ≥98%	Sigma-Aldrich / Merck	Cat#A7906-100G
Pierce Streptavidin Magnetic Beads	ThermoFisher	Cat#88817
Critical Commercial Assays		
QIAquick gel extraction kit	Qiagen	Cat#28704
QIAquick PCR purification kit	Qiagen	Cat#28104
Deposited Data		
Raw data for the PhIP-Seq experiments	This paper	EGA: EGAS00001006999
Raw data for PhIP-Seq experiment in IBD	Bourgonje et al.[40]	EGA: EGAD00001010118
Fecal shot-gun sequencing	Zhernakova et al.[93]	EGA: EGAD00001001991
Fecal shot-gun sequencing IBD	Imhann et al.[94]	EGA: EGAD00001004194
Genetics IBD	Hu et al.[95]	EGA: EGAD00010001495
Oligonucleotides		
library amplification primer fwd	GATGCGCCGTGGGAATTCT	N/A
library amplification primer rev	GTCGGGTGGCAAGCTTTCA	N/A
Recombinant DNA		
Oligo pool (200 mers)	Twist Bioscience	N/A
Oligo pool (230 mers)	Agilent Technologies	N/A
Software and Algorithms		
Peptide quantification and enrichment determination	Vogl et al.[21] and Leviatan et al.[28]	https://zenodo.org/record/7307894
Descriptive stats, GWAS, network and assocations	This paper	https://zenodo.org/record/7773433
Other		
Nunc™ Immobilizer™ Streptavidin Plates	Thermo Scientific	Cat#436014
BioTides™ Peptides	JPT Peptide Technologies (Berlin, Germany)	N/A
Freedom Evo liquid handling robot	Tecan	N/A
MASTERBLOCK, 96w, PP, 2ml, Natural, 50/case	Danyel biotech	Cat#60-780270
Corning Axygen® AM-2ML-SQ AxyMat™	Biolab Ltd	Cat#AXY-AM-2ML-SQ

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Alexandra Zhernakova (a.zhernakova@umcg.nl).

Materials availability

Antibody-bound peptides generated for this study are available in the European Genome-Phenome Archive (EGA) and are publicly available from the date of publication. Accession number is listed in the key resources table.

Data and code availability

- The data presented here belongs to Lifelines. Lifelines is specifically organized to make assessment results available for (re)use by third parties genetics and phenotypic data can be requested through Lifelines. A research proposal must be submitted for evaluation by the Lifelines Research Office.
  - o LLD PhIP-Seq: Raw and processed PhIP-Seq data generated for this study are available in the European Genome-Phenome Archive (EGA) and are publicly available from the date of publication. Accession number is listed in the key resources table.
  - o LLD Phenotypic data: Researchers must submit a data order (i.e. a selection of variables) and research proposal in the Lifelines online catalog.
  - o LLD Genetics used for GWAS: Genotyping data is not publicly available to protect participants' privacy, and neither can be deposited in public repositories to respect the research agreements in the informed consent. The data can be accessed by all bona-fide researchers with a scientific proposal by contacting the LifeLines Biobank (instructions at https://www.lifelines.nl/researcher/how-to-apply). Researchers will need to fill in an application form that will be reviewed within 2 weeks. If the proposed research complies with LifeLines regulations, such as noncommercial use and warranty of participants' privacy, then researchers will receive a financial offer and a data and material transfer agreement to sign.
  - o LLD raw fecal metagenomics can be accessed from EGA and are publicly available from the date of publication. Accession number is listed in the key resources table.
    In addition to Lifelines data, we used data belonging to the 1000IBD cohort study for meta-analysis.
  - o IBD PhIP-Seq data from the IBD cohort used for meta-analysis, IBD Genetics data used for GWAS meta-analysis, and IBD raw fecal metagenomics are available in EGA Accession numbers are listed in the key resources table.
  - o Supplementary material includes summary statistics from most analysis described. In addition, intermediate files and additional material can be accessed online in: Mendeley Data: https://doi.org/10.17632/4wzz7d9yf6.1.

- All original code, which was used for performing data analysis, has been deposited Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human samples

Lifelines is a multi-disciplinary prospective population-based cohort study examining, in a three-generation design, the health and health-related behaviors of 167,729 individuals living in the North of the Netherlands. It employs a broad range of investigative procedures to assess the biomedical, socio-demographic, behavioral, physical and psychological factors that contribute to the health and disease of the general population, with a special focus on multi-morbidity and complex genetics.[96] We collected data from the subcohort LLD[27] (58% female, mean age 45.04 years, mean BMI 25.26, 12% obese participants with BMI > 30). Approval from institutional ethics review is available under reference number M12.113965. In this study, we used a subset of LLD (n = 1,437, 57% female, mean age 44.5 years) with available information including anthropometrics, blood parameters and self-assessed questionnaires about health and lifestyle. These questionnaires included questions about allergies in which we identified abnormally high numbers of self-reported allergies, mainly driven by the category ''other allergies'', which might include other conditions such as food intolerances. Previous works described the autoantibody panels for anti-CCP and CTD ratio[97] and anti-SSA.[98]

The 1000IBD cohort is a large, prospective observational cohort study based in Groningen, the Netherlands, aiming to biologically and clinically characterize patients with IBD who are included at the outpatient IBD clinic of the University Medical Center Groningen (UMCG).[99] Detailed phenotypic data and multi-omics profiles have been generated for over 1,000 included patients with IBD, enrolled from 2,007 onwards. Antibody-bound peptide repertoires (PhIP-Seq profiles) were generated for 497 patients included in the 1000IBD cohort (median age 39 years, 63% females, median BMI 24.7 kg/m^2), of which 256 patients were diagnosed with Crohn's disease, 207 with ulcerative colitis and 34 with an undetermined type of IBD (IBD-U). Ethical approval for participation in the 1000IBD cohort has been granted by the Institutional Review Board of the UMCG (in Dutch: ''Medisch Ethische Toetsingscommissie'', METc) under registration number 2008/338 and the study has been conducted in accordance with the principles of the

Declaration of Helsinki (2013). Patients provided written informed consent for their participation in the study. Further details on the subcohort of 1000IBD of which PhIP-Seq profiles were generated can be found elsewhere.[40]

METHOD DETAILS

PhIP-Seq library design, preparation, sequencing and processing

Microbial library description[21] and the allergen, IEDB and phages library[28] have been previously presented. The general PhIP-Seq protocol was initially described by Larman et al.[13] and was performed with minor modifications as previously outlined.[21] In short, PCR plates in contact with phage/antibody mixtures were blocked with bovine serum albumin (BSA) solution (used concentration were previously described[21]). BSA was supplemented into phage-buffer mixtures for immunoprecipitations (IPs). Phage wash buffer for IPs contained 0.1% (wt/vol) IPEGAL® CA 630 (Sigma-Aldrich cat. no.I3021). Phage and antibody amounts for IPs were used as previously optimized[21] at 3 μg of serum IgG antibodies (measured by ELISA) and phage library at 4,000-fold coverage of phages per library variant. As technical replicates of the same sample agreed (average Pearson's $\rho = 0.96$[21]), measurements were performed in single reactions. The microbial libraries[21] (230 nt, 244,000 variants) were mixed in a 2:1 ratio with the phage, immune and allergen library (200 nt, 100,000 variants).[28] Phage–antibody mixtures mixed with overhead mixing at 4°C. A 50%-50% mix of protein A and G magnetic beads (total 40 μl; ThermoFisher Scientific, cat. nos. 10008D and 10009D, prepared according to the manufacturer's recommendations) was added after overnight incubation and further rotated at 4°C for 4 h, then the beads were transferred to PCR plates and washed twice, as previously reported.[21] Therefore, a Tecan Freedom Evo liquid-handling robot with filter tips was used.

PCR amplifications (pooled Illumina amplicon sequencing) were run with Q5 polymerase (New England Biolabs, cat. no. M0493L) according to the manufacturer's recommendations (primer pairs as previously outlined[21]).

Composition of the antigen library

This work uses two previously developed peptide libraries: a microbial library[21] and an allergen library.[28] The microbial library contains 244,000 peptide sequences from 28,668 different proteins, from which 27,837 proteins were derived from microbial antigens, while the rest are controls. This contains genes predicted from metagenome-assembled genomes (147,061 peptides), known pathogenic bacterial species (61,250 peptides), bacteria known to be coated with antibodies (22,050 peptides), probiotic bacteria (14,700 peptides), virulence factors extracted from the virulence factor database (VFDB) (24,164 peptides) and controls (11,525 oligos). Antigens were selected giving priority to known immunogenic antigens and focusing on secreted, membrane and motility proteins. The second library contained 5,527 peptides from five different allergen databases,[28] 31,436 peptides from the Immune Epitope Database (IEDB)[29] and approximately 40,000 bacteriophage peptides.

Peptide antibody-binding enrichment

All sequencing samples were rarefied to the same sequencing read-depth prior to statistical testing. Samples were subsampled to 1.25 million paired-end reads. In previous experiments, we found this number of reads sufficient to reproduce all enriched antibodies found using the dataset, with no subsampling, whereas more exhaustive subsampling results in a loss of significant enrichment hits.

Antibody-binding against peptide (seropositivity) was defined as previously described.[21] In brief, null distributions per input level (number of reads per clone without IP) were generated in each sample. A two-parameter generalized Poisson model was fitted to the null distribution, and the P-value to obtain the coverage level after IP for a given clone is estimated. Model parameters were estimated for each null distribution using maximum likelihood or directly interpolated.[11] A strict Bonferroni cut-off at $P_{Bonferroni} < 0.05$ was then used to define seropositivity. A total of 175,242 peptides were seropositive in at least one participant.

QUANTIFICATION AND STATISTICAL ANALYSIS

Antibody-bound peptides exploratory analysis

Data analysis was performed in R v4.0.3 using the packages tidyverse, stats, vegan,[100] corrplot, igraph,[101] WGCNA,[102] readxl, pheatmap, cairo and patchwork.

Antibody-bound peptide selection

Peptides were selected for the analysis based on two filters. First, we chose peptides with a prevalence of at least 5% and less than 95% in either 1000IBD or LLD (excluding follow-up samples). Second, several available peptides had the same amino acid sequence, which may arise from different nucleotide sequences. For these antibody-bound peptides with identical sequences, we chose the most prevalent. Applying these two filters left 2,815 antibody-bound peptides for subsequent analyses.

Principal component analysis

We used 2,815 peptides to compute perform a component analysis (PCA). Eigenvalues were used to produce a scree plot and eigenvectors to identify top peptides contributing to the first components. A K-medoids algorithm (k = 2) was performed on the dimensionally reduced dataset (PC1 and 2) to label observed clusters (PAM, cluster R package). This analysis was reproduced after removal of the 90 peptides belonging to CMV. The top 100 PCs (43% of total variability) were used for association to phenotypes. PCA regression was carried out using the phenotypes as the dependent variable and all 100 PCs, sex, age and sequencing plate

as covariates. For all continuous phenotypes, a linear model was performed. For binary outcomes, a logistic regression was performed. For ordered factors, an ordered logistic regression was performed.

Time and family distance analysis

322 LLD samples belonging to two different time points were used for a time consistency analysis. Jaccard distance was used as the dissimilarity metric between samples. P-value of longitudinal effect of mean distance was estimated by computing the P-value of the mean pairwise difference of longitudinal samples in a null distribution of mean distances of pairwise differences of 2,000 label swaps. Interrogation of factors that might affect the degree of change in longitudinal samples was performed using pairwise distances from longitudinal samples as dependent variable and age and sex as covariates in a linear model. Antibody-bound peptide consistency was computed by averaging the number of changes in the enrichment profile of a peptide among all samples with longitudinal data points. To check whether antibody-bound peptide enrichment changes seen in follow-up are due to a different reactivity of the plates used for baseline and follow-up samples, we ran a Wilcoxon test comparing the number of enriched antibody-bound peptide of participants profiles from plates with follow-up samples vs plates with no follow-up samples.

We then selected samples belonging to the same family[34] with three members (26 families). We computed pairwise distances (Jaccard) between family members (father to offspring, mother to offspring and father to mother). For each of the comparisons, we estimated a P-value comparing the mean distance with a random distribution of means from 2,000 label permutations.

To study the influence of the number of peptides used on the conclusions based on Jaccard distances, we subsampled the set of 2,815 peptides in 4 subsets comprising 20%, 40%, 60% and 80% of the peptides and repeated the time and family similarity analyses. After obtaining permutation P-values, we reached largely the same conclusions. In addition, we observed that the distance matrices using different data subsets were largely correlated (Mantel test), finding a median ρ of 0.88 (max = 0.95, min = 0.82). This supports that the results are largely independent from the number of peptides used for distance calculation.

In addition to these analyses, we reproduced the findings using a Manhattan distance matrix instead of Jaccard.

Network analysis

We used a weighted gene co-expression network analysis[102] in the context of antibody-bound peptide presence/absence to identify modules of peptide co-occurrence. We used all LLD samples (1,784) and the subset of selected peptides with no missing values (2,787) to build the network. The soft thresholding power was chosen by visually inspecting the model fit of powers from 1 to 20. To test the network assumption of scale invariance, WGCNA reports the R^2 between log(k) and log(p(k)), where k is the number of edges from each of the nodes of the network and the function p is the power function.[103] Values close to 1 indicate strong evidence of scale invariance. A power of 7 identified the highest R^2 value (0.94), and thus, we decided to use this power. An unsigned adjacency matrix was computed using Pearson correlation between antibody presence/absence profiles. This matrix was further processed into a topological overlap distance matrix (TOM). Hierarchical clustering (method = average) of the TOM distance was followed by a dynamic tree cut algorithm to identify clusters of at least 10 peptides. Cluster eigengenes were estimated and used to merge similar modules together (mergeCloseModules, cutHeight = 0.5) to produce the final set of modules. Peptides belonging to a module of at least 10 peptides were used to build a visual network graph (igraph). A maximum spanning tree algorithm was used to build the network.

The peptide identity from the identified modules was checked and a sequence similarity analysis was run. Module eigengenes were extracted and correlated between modules. Strong module correlation was defined on the basis of achieving a $P_{Bonferroni} < 0.05$.

We carried out further investigations to ensure module consistency. First, we checked two other distance metrics to define the adjacency matrix used by WGCNA that may be better-suited to binary data, namely Jaccard and Kulczynski.[104,105] However, WGCNA's checks on scale invariance failed (maximum $R^2 < 0.8$). Therefore, we decided to use a different approach to build a network of binary traits. The R package IsingFit implements the method described in this paper, which consists of determining network adjacency based on logistic regression with an l1 penalty (lasso). The regularization strength hyperparameter λ is selected using an information criteria metric. The resulting adjacency matrix was normalized to a 1-0 range and transformed into a distance matrix. Clustering was performed as in the WGCNA matrix by hierarchical clustering of the samples (method = average) and identifying modules with a dynamic tree cut. Most of the identified modules (8/12) were defined to be homologous to the WGCNA-defined ones (eigengene's Pearson's $r > 0.95$). The four extra modules were analyzed to identify peptide similarity, as previously described. Binary-matrix modules are available at Table S1.2.

We performed a bootstrapping analysis to estimate the consistency of WGCNA modules. Sampling with replacement of 20%, 40%, 60% and 80% of samples was carried out 50 times. A WGCNA network was built in each of those subsets as previously defined. We defined homologous modules by computing Jaccard distances between binary peptide labels (assigned to module/not assigned), and picked the module with highest similarity to the complete set as its homologous for each data subset (if similarity was not above 0.5, no module was picked as homologous). Finally, per peptide, we quantified the percentage of times it was assigned to a homologous cluster.

Additionally, we performed a combined network analysis between the IBD and LLD cohorts. Once again, we used eigengene correlations to define clusters that are homologous in the combined analysis to the ones defined using LLD only ($\rho > 0.95$), which identified 21/22 clusters to be consistent between both analyses.

To check if co-occurrence modules might be driven by batch effects (due to PhIP-Seq sequencing plate), we computed the prevalence of each peptide within a module. If a common batch effect was present in all peptides of a module, we would expect to see a

significant batch effect adding variation to the mean prevalence within all modules (Null hypothesis, Prevalence ~ Peptide + Batch). If this batch effect was different per peptide, then the batch effect would show a significant interaction with the peptide (Alternative hypothesis, Prevalence ~ Peptide + Batch + Peptide*Batch). If the alternative hypothesis was true, the batch would have a different effect per peptide, and thus it is not the only explanation to observe high co-occurrence between antibody-bound peptides. We fitted the null and alternative hypothesis in two linear models, and computed a P-value for the peptide–batch integration by computing a likelihood ratio test between both models. All tested models showed a significant interaction effect, indicating that batch most likely has a different effect per peptide.

To associate the presence of co-occurrence modules with genetic, environmental and lifestyle variables, we used WGCNA to compute and extract eigengenes. Eigengenes from all modules (except the low-consistency module 17) were included in a GWAS analysis (see STAR Methods section genome-wide association) and associated with all available phenotypes (see STAR Methods section phenotype association analysis).

Peptide similarity

Sequence similarity between peptide groups was estimated using Clustal Omega.[106] Clustal Omega uses this distance matrix to build guiding trees for the progressive multiple sequence alignment algorithm. This distance is internally calculated using the k-tuple method.[107]

Multiple sequence alignment information content

We ran MAFFT v7.487[108] to obtain multiple sequence alignments.

Using the distance matrix obtained from Clustal, we performed hierarchical clustering (average method) to visualize sequences in a dendrogram. Multiple sequence alignments were attached to the dendrogram to visualize sequence similarity.

Information content per position in each multiple sequence alignment was obtained by calculating Shannon entropy (Equation 1) and then applying (Equation 2). Gaps were included in the information content computation as one more character.

$$H^2 = -\Sigma_{aa} P(aa) \times log_2(P(aa)) \quad \text{(Equation 1)}$$

Shannon entropy of a position in a sequence alignment. aa stands for amino acid, which could take the value of any of the 20 common amino acids and gap. H^2 stands for entropy. The probability of each amino acid was estimated as its frequency per position.

$$I = log_2(22) - H^2 \quad \text{(Equation 2)}$$

Information content of a position in a sequence alignment. 'I' stands for information. H^2 stands for entropy and is obtained in Equation 1.

Motif discovery

Groups of peptides of interest were subject to motif discovery using MEME.[109] MEME is an expectation maximization framework that allows for identification of enriched kmers in a group of unaligned sequences. We ran MEME v5.05 with the following parameters: zoops as distribution of motifs, since we expected either no motif or only one motif per sequence; number of motifs to find = 3; minimal motif width = 3 amino acids (maximum of 50); the classic objective function; Markov order = 0; and a minimum of 7 sequences containing the motif.

CMV analysis

We interrogated whether CMV antibody-bound peptide breadth increased with age. To do so, we clustered samples in three groups depending on the number of CMV peptides detected (0, from 0 to average number 16, and above the average number 16). We then performed ANOVA and an ad-hoc Tukey test to determine whether the age of the different groups differed.

We also tested CMV and EBV as a factor that might determine differences in antibody consistency after 4 years. With that aim, we performed a linear model in which the Jaccard distance of an individual between baseline and follow-up was used as a dependent variable, including baseline age, sex and CMV status (defined as the 2-medoids clustering determined using PC1 and 2) and EBV status (defined as a local minimum in the EBV peptide breadth distribution) as covariates.

PhIP-Seq validation

To validate the antibody-bound peptide signals used throughout this paper, we performed two analyses.

First, 294 participants from the IBD cohort had available CMV IgG measurements in addition to PhIP-Seq. Since CMV peptides are the major pattern of variability among our two Dutch cohorts, a 2-means clustering was performed in the whole IBD and LLD antibody-bound peptide dataset. We explored the association between belonging to a given cluster and IgG seropositivity by means of a logistic regression (log-odds of being IgG positive if PhIP-Seq clustering was positive, 6.72, $p < 2x10^{-16}$). Only two false positives were seen by defining clustering belonging as CMV seropositivity and there were 11 false negatives.

Second, we chose 8 peptides for ELISA validation, which included a human gamma herpesvirus 4 (EBV) as positive control (80%–90% prevalence) and human SAPK4/MAPK13 as a negative control (0% prevalence). We validated the other 6 peptides available in the PhIP-Seq profile (see Table S1.4 for sequence and taxonomy). All peptides used for ELISA consisted of 20 amino acid peptide

sections, since the full-length sequences could not be chemically synthesized due to technical limitations (increasing impurity). These sections were selected based on the presence of sequence motifs identified in the network analysis or on the overlap of adjacent PhIP-Seq peptides that showed high correlation and belonged to the same protein. Oligo synthesis was carried out at JPT Peptide Technologies (Berlin, Germany). Subsequent ELISAs were performed following supplier's instructions (Protocols BioTides TM Peptides Revision 1.0; Peptide ELISA Revision 1.2). Peptides were bound to streptavidin-coated microtiter plates (ThermoFisher Scientific, Nunc Immobilizer Streptavidin Plates, cat. no. 436014) and incubated with 100 μL of 1,000-fold diluted blood samples from 40 population controls and 54 patients with IBD (27 CD, 27 UC). Detection of Antibody-binding was assessed using horseradish peroxidase-conjugated anti-human IgG antibody (Southern Biotech, cat. no. 204205), 3,3',5,5'-tetramethylbenzidine (TMB) as substrate and 25% sulfuric acid as stop solution (ThermoFisher Scientific, Stop Solution for TMB Substrates, cat. no. N600).

Resulting antibody absorbances were compared between the groups of samples predicted to be antibody negative and positive based on PhIP-Seq data using a non-parametric Wilcoxon test.

Phenotype association analysis

Jaccard distances between all baseline samples were used as the dependent variable in a PERMANOVA (R vegan package, adonis2) against, sex, age and PhIP-Seq plate in order to identify covariates (2,000 permutations). To associate individual enrichment profiles to available phenotypes, we performed a logistic regression on the presence/absence of antibody-bound peptides using the phenotype, PhIP-Seq plate, age and sex as covariates on 1,437 baseline participants. We controlled the FDR at 0.05 using the Benjamini-Hochberg procedure.[110] We reproduced the analysis in three more scenarios. First, removing a total of five participants where the number of enriched antibody-bound peptides was below an interquartile range from the 25th quartile (200 enriched antibody-bound peptides), since they might have failed PhIP-Seq for an undetermined reason (Table S2.7). A second analysis was carried out while including absolute abundances of blood counts as covariates (Table S2.7). In addition, we also included CMV status (as defined based on PCA clustering analysis) as a covariate in the model, since it has a major impact on interindividual antibody-bound peptide differences (Table S2.7). We observed good correspondence in the results from all three additional models and our standard model.

Genotyping and imputation

Genome-wide genotyping data was generated previously generated[27] and processed.[49] Briefly, microarray data were generated on CytoSNP and ImmunoSNP platforms and processed on the Michigan Imputation Server.[111] Haplotype phasing was carried out using SHAPEIT and imputation was done using the HRC version R1 as reference [112].

Genetic preprocessing

We used GenotypeHarmonizer[113] for imputation quality (minimum posterior probability of 0.4), call rate (minimal call rate of 95% of samples), Hardy-Weinberg equilibrium (minimal P-value allowed of $1x10^{-6}$) and SNP ambiguity filtering. We then computed identity by descent among samples using PLINK v1.9[114] on linkage disequilibrium (LD)–pruned genotypes (window size 50 Kb, variance inflation threshold 5 and maximum R^2 between variants 0.2). We estimated identity by descent between all samples using PLINK and randomly selected a sample from the pairs with a PI_hat value > 0.2, which resulted in the removal of 14 samples from subsequent analysis (total of 1,255 available samples).

Heritability and genetic correlation

GCTA[115] was used to compute a genomic relationship matrix (GRM) using genotyped SNPs with a minor allele frequency (MAF) of at least 0.05. The GRM was used to estimate antibody-bound peptide heritability using a linear mixed model between unrelated individuals (GREML approach)[115,116] while controlling for age, sex and PhIP-Seq sequencing plate. Similarly, genetic correlations between peptides were estimated using GCTA.[117]

Genome-wide association

For each of the available antibody-bound peptides, we conducted an association analysis between genotypes (MAF > 0.05) and presence/absence profile. PLINK v1.9[114] logistic mode was run while controlling for age and sex and using the genotype in an additive model. This analysis was reproduced in a recessive model between 49.1 and 49.3 Mb in chromosome 19. Additionally, co-occurrence module's eigengenes were also associated with genotypes using a linear model in PLINK v1.9.

Genetic meta-analysis

A second study using the same PhIP-Seq library panel and protocol has been conducted in an IBD cohort from the Netherlands.[40,118] Genotyping information is also available for this cohort.[95] The same quality control steps and analysis methods were used as described above, while the disease subtype (Crohn's disease or ulcerative colitis) was also added as an extra covariate in the logistic regression.

Summary statistics from both the LLD and 1000IBD cohorts were meta-analyzed using METAL.[119] We performed a *P*-value–based fixed-effects meta-analysis. A study-wide significance threshold was estimated by dividing the genome-wide significance threshold of $5x10^{-8}$ by the number of independent peptides included in the GWAS. The number of PCs needed to reach 90% of antibody-bound peptide repertoire variability in LLD was used as a number of independent tests (708 components), obtaining a study-wide threshold of $5.67x10^{-11}$. For each peptide's summary statistics we extracted genome-wide significant associations ($p<5x10^{-8}$) for clumping. We clumped variants in windows of 1,000 Kb if they had a minimal R^2 (computed from LLD genotypes) of at least 0.1 using

PLINK. Leading variants of each clump were then annotated using the Ensembl Variant Effect Predictor and the grCh37 human build.[120] LD between our identified leading variants and other publicly reported variants was estimated in the CEU population from the 1,000 genomes using the LDlink webtool.[121,122]

HLA imputation and association

The chromosome 6 region with 25–34 Mb that contains the MHC genes was extracted. Imputation of the HLA region, including HLA alleles, polymorphic amino acids, SNP variants and indels, was then performed using SNP2HLA (v2) with the Type 1 Diabetes Genetics Consortium (T1DGC) reference panel (2,767 unrelated European descent individuals) HLA Reference Panel.[123] Next, we combined both imputed and genotyped SNPs, HLA alleles and amino acid variants, resulting in a total of 8,926 variants. Variants with MAF < 0.05 and imputation quality score (INFO) < 0.5 were removed before association.

HLA to peptide association was performed using linear models in 1,175 participants, while controlling for age, sex, PhIP-Seq plate and disease subtypes (Crohn's disease/ulcerative colitis, only specific to IBD cohort). Summary statistics from both datasets were further meta-analyzed using a fixed-effects model in PLINK v1.9. The statistical significance threshold was determined by dividing the usual P-value 0.05 threshold by the number of independent features tested (66 PCs were needed to reach 90% of HLA feature variability in LLD, while 708 PCs were needed to capture 90% of the peptide variability, resulting in 46,728 independent tests), resulting in a threshold of $1x10^{-6}$. FDR was estimated using the Benjamini-Hochberg method.

Modeling of peptide presentation in HLA complexes

To explore whether HLA–peptide associations potentially point to HLA-II ability to display a specific peptide, we performed computational modeling of the complex–peptide interaction.

The protein sequences of DR3, DR4, DR14, DR15 and DQ2 were obtained from the IPD-IMGT/HLA database[124] and aligned against the entire Protein Data Bank database using pBLAST. Protein structures displaying 100% amino acid identity with the HLA-II database sequences were chosen to build the peptide binding modes. Those structures correspond to the HLA complexes DR3:7N19, DR4:1D5M, DR14:6ATF, DR15:1YMM, DQ2:6PX6 and DQ8:2NNA. Proteins other than HLA-II, water molecules and heteroatoms were removed from the structures prior to modeling. The NetMHCIIpan-4.0[41] server was then used to predict peptide binding to the corresponding associated HLA alleles: DRB1*1501 for *Lactobacillus* phage LfeInf; DRB1*0301, DQA1*0501-DQB1*0201 and DRB1*1401 for *Streptococcus agalactiae* C5a peptidase; and DRB1*0401 and DQA1*03-DQB1*0302 for Human mastadenovirus minor core protein. The DRB1*1401 for *Streptococcus agalactiae* C5a peptidase was selected as a no binding negative control for these experiments. Following the identification of the peptide core by NetMHCIIpan-4.0, we used %Rank_EL as a representative metric indicating predicted binding strength. %Rank_EL is calculated as the percentile of the predicted binding affinity compared to the distribution of affinities calculated on a set of random natural peptides (%Rank_EL; strong binding: $\leq$ 2.0, weak binding: 2.0–10.0, no binding: > 10). The protein structures and identified peptide core were submitted to HPEPDOCK Server for peptide–protein molecular docking.[125] In brief, cleaned protein structures were used as receptors, and the peptide core sequence was used to generate 100 different conformers and a global sampling of binding orientations into the peptide binding domain of HLA-II receptors. Following docking, the peptide-HLA-II complexes with the highest complementarity were selected for receptor–peptide refinement in the HADDOCK Refinement Interface.[126] Finally, the peptide-HLA complexes were analyzed for the formation of molecular interactions and binding energy using PLIP[127] and PRODIGY.[128,129]

Metagenomic sequencing

Metagenomic collection and sequencing has previously been detailed.[93] In brief, participants collected fecal samples at home and directly stored then in the freezer. Fecal samples were collected on dry ice and transferred to the laboratory. Aliquots were stored at -80°C until further processing. The allPrep DNA/RNA Mini Kit (Qiagen; cat. 80204) was used for DNA isolation. DNA was sent to the Broad Institute (Cambridge, Massachusetts, USA) where library preparation and shotgun metagenomic sequencing were performed on Illumina HiSeq.

Metagenomic processing

Low-quality reads were discarded by the sequencing facility. Reads aligning to the human genome or to Illumina sequencing adapters were removed using default parameters of the KneadData pipeline (version 0.39). In short, this software uses Trimmomatic[130] for adapter removal and quality trimming of reads and Bowtie2[131] for mapping and removal of reads mapped against the human genome (hg19). Taxonomy abundance estimation was then performed using MetaPhlan3 and default parameters.[132] Next, microbial relative abundance was transformed using log-ratios on the relative abundance table (adding ½ of minimal non-zero relative abundance to each cell in the table), with species geometric mean as denominator (centered-log ratio). Bacteria not present in at least 10% of samples were discarded.

Microbiome-peptide association analysis

Co-occurrence between fecal microbiota and blood antibody–bound peptides was assessed using logistic regression analysis in 1,051 participants. In total, we analyzed the relation between 284 bacteria and 2,815 antibodies. Each antibody-bound peptide was modeled in generalized linear models as a response variable in a model including age, sex, PhIP-Seq plate and transformed bacterial abundance as predictors.

Microbiome meta-analysis

To increase the statistical power to detect associations between gut microbiota and blood antibodies, we combined the results of our cohort with the results derived from the 1000IBD cohort (n = 137, blood and fecal samples collected with <1 year difference) by performing a meta-analysis. We filtered out peptides not seen in at least 10 samples in both IBD and LLD cohorts. Heterogeneity coefficients (I^2 and Cochran's Q) were estimated per association. Meta-analysis was conducted by pooling summary statistics for both cohorts and under random and fixed-effects assumptions using the *meta* R package (v4.19-0)[133]. FDR was estimated from the resulting associations.

Microbial peptide quantification

To quantify the presence of the exact antibody-peptide sequences in the microbiome, we selected 647 peptide-bound peptides with origins in the human microbiome that we found to be associated with at least one phenotype. We used ShortBRED v0.9.5[134] to generate a database of the peptide sequences using UniRef90 as a reference, and quantified all available LLD metagenomes. Each antibody-bound peptide presence/absence profile was associated with its gut microbiome quantification while controlling for age, sex and PhIP-Seq sequencing plate. Benjamini-Hochberg FDR was estimated.