Population-wide diversity and stability of serum antibody epitope repertoires against human microbiota

Population-wide diversity and stability of serum antibody epitope repertoires against human microbiota

Thomas Vogl[1,2#], Shelley Klompus[1,2#], Sigal Leviatan[1,2#], Iris N. Kalka[1,2], Adina Weinberger[1,2,3]*, Cisca Wijmenga[4], Jingyuan Fu[4,5], Alexandra Zhernakova[4], Rinse K. Weersma[6], Eran Segal[1,2]*

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Israel
[2]Department of Molecular Cell Biology, Weizmann Institute of Science, Israel
[3]Louis H. Sackin Research Fellow Chair in Computer Science
[4] University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands
[5] University of Groningen, University Medical Center Groningen, Department of Pediatrics, Groningen, the Netherlands
[6] University of Groningen, University Medical Center Groningen, Department of Gastroenterology and Hepatology, Groningen, the Netherlands

[#] These authors contributed equally
* email: adina.weinberger@weizmann.ac.il, eran.segal@weizmann.ac.il

ORCIDs
T.V.: 0000-0002-3892-1740
C.W.: 0000-0002-5635-1614
J.F.: 0000-0001-5578-1236
A.Z.: 0000-0002-4574-0841
R.K.W.: 0000-0001-7928-7371

Abstract

Serum antibodies can recognize both pathogens and commensal gut microbiota. However, our current understanding of antibody repertoires is largely based on DNA sequencing of the corresponding B-cell receptor genes, and actual bacterial antigen targets remain incompletely characterized. Here, we have profiled the serum antibody responses of 997 healthy individuals against 244,000 rationally selected peptide antigens derived from gut microbiota, pathogenic, and probiotic bacteria. Leveraging phage immunoprecipitation sequencing (PhIP-Seq) based on phage-displayed synthetic oligo libraries, we detected a wide breadth of individual-specific as well as shared antibody responses against microbiota that associate with age and gender. We also demonstrate that these antibody epitope repertoires are more longitudinally stable than gut microbiome species abundances. Serum samples of >200 individuals collected five years apart could be accurately matched, and could serve as an immunologic fingerprint. Overall, our results suggest that systemic antibody responses provide a non-redundant layer of information about microbiota beyond gut microbial species composition.

Introduction

Humans are covered with approximately the same number of bacteria as body cells [1] and the gut microbiota in particular influence various aspects of human health [2]. Intestinal bacteria elicit a multitude of innate and adaptive immune responses to prevent overgrowth and their passage into the bloodstream, where they could potentially cause sepsis [3]. Mucosal IgA antibodies play a pivotal role in exerting this immune system-microbiota equilibrium by protecting the host from pathogens and maintaining intestinal homeostasis [4]. A growing number of studies have demonstrated that gut microbiota antigens also elicit systemic IgG [5–7] and IgA [5,8] responses in blood [9] and described how mucosal and systemic antibody responses relate to each other [10]. Despite considerable progress in understanding microbiota-driven antibody responses in animal models [5–8,10], it is incompletely understood which microbial antigens/epitopes are targeted by human antibodies and how these responses associate with health and disease [11].

Recent B cell receptor sequencing (BCR-seq) studies have provided unprecedented insights into antibodies' role in the adaptive immune system[12,13] relating to gut microbiota [10,14] as well as pointing towards connections of immune mediated diseases and microbial antigens [15]. While unraveling the clonal diversity [16], and changes caused by microbiota [10] of the underlying Ig-epitope repertoires, their functional capacity towards antigen recognition in humans has remained largely elusive.

The complexity of human microbiota represents a key challenge for systematic investigations of antibody-antigen interactions. Humans bear thousands of bacterial species [17] with each species' genome encoding thousands of genes, representing an enormous space for potential protein antigens. Conventional methods for studying antibody binding of microbiota such as ELISAs and peptide arrays are limited to testing hundreds to thousands of antigens in parallel. Concomitantly, bacterial flow cytometry combined with microbiome sequencing [9,18–20] informs on antibody coated species, but not the exact antigens bound. Hence, there is a lack in our understanding of the 'dark matter' of the antigenic space represented by human microbiota.

Phage immunoprecipitation sequencing (PhIP-Seq) [21] allows the evaluation of antibody responses to hundreds of thousands of antigens in parallel, as successfully demonstrated primarily with autoimmune diseases [22–24] and viruses [25–27]. As the chemical synthesis of peptide antigens is limited by short lengths and high costs, PhIP-Seq relies on antigen libraries encoded by synthetic DNA oligonucleotides. These libraries are cloned into, and displayed on the surface of T7 phages. Antibody bound phages are enriched by immunoprecipitation and identified by next generation sequencing (Fig. 1a) [21].

Here, we have created a PhIP-Seq library representing 244,000 peptide antigens of microbiota to profile population wide systemic immunoglobulin (Ig) epitope repertories in 997 healthy individuals. We correlated these antibody profiles with meta data available for this cohort [28], including clinical data as well as gut metagenomic data, to evaluate associations with age, gender, and high longitudinal stability.

Results

Microbiota peptide Library Design

We designed a library including commensal, pathogenic, and probiotic bacterial species as well as positive and negative controls (Fig. 1b, Methods). Potential gut microbiota antigens were selected from metagenomics sequencing of 953 stool samples of individuals [28], for whom serum samples for antibody testing were also available. In addition to proteins of several gut pathogens, we also included the entire virulence factor database (VFDB) [29] and pathogens causing infectious diseases as well as human autoantigens from the immune epitope database (IEDB) [30]. Furthermore, proteins from bacterial strains commonly applied as probiotics [31], and of strains previously reported to be coated by

antibodies [19] were included (Supplementary table 1). Approximately 28,000 proteins were split into 244,000 peptides (length of 64 amino acids with overlaps of 20 amino acids), allowing for high resolution, epitope resolved analysis of antibody targeted protein segments. Each peptide was encoded with *E. coli* codon usage and barcoded within the coding sequence for identification (Methods). We enriched the library for secreted and surface proteins, which are more likely to be bound by antibodies.

After optimizing the experimental PhIP-Seq workflow [21] for our liquid handling robots (Extended Data Fig. 1a), we assessed assay performance by a series of controls: Technical (Extended Data Fig. 1b,c) and biological replicates (Extended Data Fig. 1d,e) showed high reproducibility (average Pearson R^2=0.96, Extended Data Fig. 1b) and also the employed barcoding strategy proved reliable, introducing no systematic bias (Supplementary figure 1). We observed little to no potential unspecific binding against random peptides or negative controls of human proteins (that should not elicit auto-antibody responses in our healthy cohort, Extended Data Fig. 2a), while antibody responses against positive controls of common viral epitopes [25] were reliably reproduced (Extended Data Fig. 2b). We also tested seven antibodies generated with immunogens of full-length proteins and bacterial cells as well as two antibodies specific for human self-antigens as negative controls. We robustly detected binding of these antibody preparations to peptide representations of the respective immunogens included within the library in six out of seven cases and little evidence for cross-reactivity (Extended Data Fig. *3*, Supplementary table 2). Taken together, these results validate the accuracy and reproducibility of our PhIP-Seq assay implementation.

Population wide antibody profiling

We measured serum antibody responses of 997 individuals (including the 953 individuals of whom metagenomic data had been employed to select the gut microbiota antigens of the library, see methods) [28]. This healthy cohort spanned a range of 17 to 70 years of age with clinical metadata such as blood tests available (Fig. 1c). In total we assayed for over 200 million antibody-peptide interactions (244,000 epitopes in each of the 997 individuals). Employing strict Bonferroni correction, on average ca. 800 peptides were significantly enriched (after scoring against input reads, see Methods) per individual (Fig. 1d). Tens of thousands of peptides were enriched in less than 1% of the population indicating individual specific, private antibody responses. We also detected overlapping antibody responses against microbiota antigens between individuals, as 10,750 bound peptides were shared in >1% of individuals, 1,566 in >5%, 130 in >50%, and 39 peptides in >90% (Fig. 1e). Public epitopes were not limited to pathogens such as *Staphylococcus*/*Streptococcus* species [26] and viruses [25,26], but also extended to antigens and strains from other subgroups of the library (Fig. 1f, *Extended Data Fig. 4*) including common gut microbiota such as *Bacteroidales* incl. *Prevotella copri* eliciting antibody responses in >95% of individuals. Antigens of *Blautia producta* (80%), *Parabacteroides merdae* (75%), *Eubacterium rectale* (60%), *Enterococcus faecalis* (43%), *Lactobacillus plantarum* (41%, a common probiotic), *Dorea formicigenerans* (35%) were also frequently detected (see Supplementary table 3 for a detailed list of antigens), supporting gut microbiota to commonly elicit systemic antibody responses beyond the mucosa [10] in humans.

Similar public epitopes have been reported for viruses [25], which were included as controls in our library (Extended Data Fig. 2b). Our results demonstrate that antibody responses shared between healthy individuals extend to gut microbiota antigens and probiotics, suggesting that population wide convergent antibody recognition is not limited to pathogens. The vast majority of microbiota antigens were bound by IgG isotypes, as illustrated by probing antibody binding separately with protein A and G coated magnetic beads Extended Data Fig. 5a-c, Supplementary table 4). We also measured a subset of samples with IgG and IgA specific capture antibodies [27] (Extended Data Fig. 5d), suggesting that some peptides are more frequently bound by IgG or IgA, whereas for other peptides the two antibody

classes overlap to varying extents. Amongst functional groups of bound peptides, flagella and secreted proteins were significantly overrepresented (Supplementary figure 2), in line with their known role as dominant bacterial antigens. A few antigens bound nearly universally appeared to represent antibody binding proteins such as *Staphylococcus* protein A and a homolog of a recently reported antibody binding protein from gut microbiota [32], as inferred from isotype control experiments (Fig. *2*a-c). The binding peptides of protein A cover B-domains known to bind the Fc region of IgG [33] (Fig. *2*d). When phages displayed protein A peptides covering a complete B-domain (i.e. peptides #221096 and #133222), we observed the strongest binding to IgG in our PhIP-Seq assay. Weaker interactions were observed when the phage displayed peptides contained shortened/permutated B domains. These results are in full agreement with expected binding behavior of B-domains [33].

Associations of antibody responses and metagenomics data

Systemic antibody responses against commensal microbiota have been reported in mouse models and cohorts of dozens of humans [5–10]. These studies have specifically detected serum antibodies against certain gut microbiota species. However, the degree to which serum Ig-epitope repertoires correlate with gut microbiota present in an individual has to the best of our knowledge not been investigated with large human cohorts. We previously generated metagenomics sequencing data for >900 individuals [28], for whom we had profiled serum Ig-epitope repertoires (Fig. 1c). The metagenomics reads were mapped to species-level genome bins (SGBs) [17] representing a large reference database of bacterial species. To test for similarities between antibody responses and gut microbiome compositions, we compared the Hamming distances of serum antibody responses of different individuals and the Bray-Curtis distances of corresponding gut microbiota abundances in metagenomics sequencing (Fig. 3a). We also tested for specific associations between peptides significantly bound by antibodies and bacterial SGBs (Fig. 3b, Extended Data Fig. 6, Supplementary table 5). While there was no general association between Ig-epitope repertoire and metagenomics abundances on an individual specific level (Fig. 3a), we found 1,706 significant population-scale associations between pairs of bound peptides and SGBs (after FDR correction for ~4.7 million tests, Fig. 3b, Extended Data Fig. 6). Some of the most significant associations include common commensal gut microbiota such as *Clostridiaceae* but also pathogens (*Staphylococcus*, *Streptococcus*). Some of the SGBs are correlated with antibody binding of up to 23 peptides per species. These SGBs include common gut microbiota from the *Firmicutes* phylum such as *Clostridiales* and *Ruminococcaceae* (Fig. 3b, Extended Data Fig. 6) as well as unknown species.

Ig-epitope repertoires associate with age and gender

We next leveraged metadata previously collected [28,34] for the 997 individuals profiled in this study to mine for possible associations to their Ig-epitope repertoires. Abundances of antibody responses (*i.e.* population wide presence or absence of antibodies against specific peptides) showed some age (Fig. 4a) and gender (Fig. 4b) related differences. Antibody responses against several proteins of *Shigella* species that are part of a type III secretion system (T3SS) [35] were approximately 10-fold overrepresented in elderly individuals (Fig. 4c, Supplementary table 6). Up to six different peptides per *Shigella* protein were bound with detectable antibody responses in up to 78% of individuals older than 61 years but only up to 9% of individuals less than 28 years of age (representing approximately the youngest and oldest decile of the studied cohort, passing multiple hypothesis testing, raw correlations shown in Supplementary figure *3*). The peptides bound significantly more frequently in older individuals included effector proteins such as ipaC and ipB, which are required for binding to human host cells, as well as well as the autotransporter icsA (Fig. 4c). Antibody binding against ipaC and icsA as well as other peptides detected with PhIP-Seq was significantly associated with binding in peptide ELISAs (Extended Data Fig. 7). Elderly individuals also showed more frequent antibody

responses against proteins of commensal bacteria as *Bacteroidales* and *Clostridiales*. Younger individuals showed more frequent antibody responses against antigens of *Staphylococcus aureus* and *Streptococcus* species, although differences to older individuals were less pronounced (~1.5-fold opposed to ~10-fold differences for *Shigella* antigens). We also detected age related differences in the binding of viral antigens that had been included as controls including proteins of Influenza and Herpes viruses. Independent of age, women showed significantly increased binding against antigens of *Lactobacillus acidophilus* and *L. johnsonii* strains (Fig. 4b,d) suggesting also gender differences in the Ig-epitope repertoires against bacteria. While cell wall associated proteins such as S-layer proteins [36] or an N-acetylmuramidase were bound in up to 6% of males, binding of these antigens was detected in up to 40% of females (Fig. 4d).

Machine learning predictions from Ig-epitope repertoires

Next, we examined whether machine learning algorithms can uncover any additional associations. Gradient boosting decision trees [37] based on the serum Ig-epitope repertoires showed associations with age (R^2=0.56, Fig. *5a*) and gender (AUC=0.77, Fig. *5b*). A significant association, albeit with low predictive power, was also observed for the inflammation marker C-reactive protein (CRP, *Extended Data Fig. 8*). These results point towards even broader associations with human health, that may be predicted with greater accuracy leveraging larger antigen libraries. Furthermore, Ig-epitope repertoire based associations of age and gender exceeded the accuracy of models trained on metagenomics microbiome sequencing data of the same group of individuals (Fig. *5c,e*). In contrast, antibody responses against human self-proteins and random peptides carried virtually no predictive power (*Extended Data Fig. 9*), precluding that self-reactivity or potential cross-reactivity against random peptides underlie these strong associations. Machine learning predictions from subgroups of the microbiota library (such as antigens selected from metagenomics data alone *etc*. Supplementary figure *4*) also yielded high accuracy for age (Fig. *5d*) and gender(Fig. *5f*), demonstrating that the predictive power from the measured Ig-epitope repertoires is not limited to pathogens but includes antigens of commensal gut microbiota of healthy individuals.

Temporal stability of Ig-epitope repertoires

The stability of the serological response to infection or vaccination is well known. Although antibody secreting plasma cells have been shown to persist in the human intestines for decades [38], the stability of systemic antibody responses to gut microbiota antigens is unclear. For 213 individuals of the cohort follow-up blood samples were collected after approximately five years. We measured their Ig-epitope repertoires and noticed high individual specific stability compared to the baseline sample (Fig. 6a). Sample pairs of the same individual showed a higher average correlation than random pairs (Pearson correlations of log. fold changes of 0.78 vs. 0.27, Fig. 6b). All except one follow up sample could be accurately matched to individuals' baseline serum samples collected five years apart (by simply picking the closest matching sample). Employing a greedy matching algorithm (taking the closest match for every sample) yielded perfect matches for all samples. The longitudinal stability of these Ig-epitope repertoires was not limited to pathogens, indicating that gut microbiota can also elicit lasting systemic antibody responses: Matching individuals' samples on antigens from the entire microbiota library correlated with an average Pearson correlation coefficient (using the log of fold changes) of R=0.78, antigens only from gut microbiota sequencing matched with R=0.76 and antigens of pathogens (VFDB) with R=0.83 (Fig. 6c, Extended Data Fig. 10). Yet, VFDB antigens also showed a greater correlation between unmatched samples (R=0.38) than antigens selected from microbiome sequencing (R=0.23) or the complete microbiota library (R=0.27), suggesting a higher convergence of individuals' antibody responses against pathogens and less discriminatory power than commensal antigens.

For 188 of the 213 individuals with longitudinal antibody data we also obtained longitudinal microbiome sequencing data. We compared the longitudinal stability of gut microbiome composition

derived from metagenomics sequencing of the same individuals five years apart (Fig. 6d) and we observed lower correlations over time than with the matched Ig-epitope repertoires. The average Bray Curtis metagenomic distance was 0.34 between two samples of the same individual 5 years apart, and 0.19 between two samples of two different individuals. A greedy algorithm could only match 38% of individuals' longitudinal samples (71/188) based on metagenomics data based on abundances. As relative abundances may show higher fluctuation than presence/absence of bacterial species, we also evaluated stability of the existence of genes in metagenomics data (Extended Data Fig. 10h). In this case, a greedy algorithm could match 49% of individuals' longitudinal samples (92/188), representing an improvement over the use of relative abundances. However, using Ig-epitope repertoire data allowed us to match 100% of individuals' longitudinal samples. Microbiome stability may change considerably depending on the region of the gut sampled. Thus, the stool samples analyzed in these experiments may be less stable on a per individual level than serum antibody repertoires.

Discussion

By measuring functional serum Ig-epitope repertoires against 244,000 peptides in 997 individuals, we detected a multitude of private and public antibody responses against antigens of gut microbiota. Our work offers a population scale perspective on anti-microbiota Ig-epitope repertoires, while previous studies had focused on smaller cohorts of dozens of individuals [9,19,20] and had not included the analysis of clinical metadata [28] integrated within this study.

We have not detected clear individual specific associations between serum antibody responses and abundances of corresponding gut microbiota species in metagenomics sequencing (Fig. 3a), although antibody responses against ca. 1,700 peptides were significantly associated with abundance of bacterial species in metagenomic data on the population scale (Fig. 3b). Most of these associations were detected between species and peptides that appear in a small fraction (2-5%) of individuals. Despite SGBs from metagenomics sequencing associating significantly with antibody-bound peptides, these associations are sparse and not sufficient to match individuals' metagenomics abundances to antibody responses (which could be demonstrated for longitudinal metagenomics/ Ig-epitope repertoire data, Fig. 6). These results are limited by detection thresholds. Small amounts of bacteria, that are present in the body but not detectable in microbiome sequencing, may elicit weak antibody responses that could show associations in a larger fraction of individuals. In our experiments, microbiota commonly detected in metagenomics sequencing of stool samples do not elicit strong serum antibody responses and vice versa, possibly owing to eradication of bacterial species whose products reach the blood stream in parallel by the mucosal immune system. Due to the higher temporal stability of Ig-epitope repertoires targeting microbiota than microbiome abundances suggested by metagenomics sequencing data of stool samples (Fig. 6), potential translocation of transient gut microbiota could provoke lasting systemic responses detected with our assay, but missed by metagenomics sequencing. Overall, changes in the gut microbiome may not directly be reflected by the serum Ig-epitope repertoire against gut microbiota-specific antigens. Serum antibody responses could also be affected by factors beyond the gut microbiome (such as exposure from other body sites).

Population wide serum Ig-epitope repertoires of this study were strongly associated with age and gender, suggesting that that they could carry a wealth of biological information related to human health. Antibody responses against antigens of *Lactobacillus* species were overrepresented in women. *L. acidophilus* and *L. johnsonii* are common in the intestinal and vaginal microbiome, pointing towards a gender specific impact of the urogenital tract or a difference in the consumption of probiotics. Also, for other bacterial species such as *Prevotella*, exposure at other body sites beyond the gut, as well as

cross-reactivity, may potentially contribute to the observed systemic antibody responses. Antibody responses in the peripheral blood of healthy individuals against gut microbes may putatively originate by different mechanisms. While the healthy individuals profiled in our study are expected to have an intact intestinal barrier, small amounts of gut microbial products may nonetheless reach the blood stream and elicit antibody production by systemic B cells. The dominance of the IgG isotype in the detected antibodies supports this notion, although we also detected IgA responses when analyzing a subset of samples at greater depth (Extended Data Fig. 5). Systemic IgA responses potentially originate from gut-derived plasmablasts and plasma cells secreting IgA or IgM [39] that may re-circulate back to the effector site of the lamina propria. IgG responses against the antigens detected with PhIP-Seq may originate from class-switching, from peripheral exposure to antigens, or gut-derived B cells, that eventually home to the bone marrow [40]. Further studies will be required to elucidate these aspects , with potentially multiple mechanisms being at work in parallel

Elderly individuals more frequently exhibited antibody responses against *Shigella* species (gut pathogens causing diarrhea) [35] as well as various gut microbiota. These differences could be explained by older individuals having encountered more antigens throughout their lifetime or by increased gut permeability and potential translocation of microbiota products in the elderly [41,42]. Another possible explanation is that a change in environment and habits or the year of birth could be linked to different exposures to microbiota (e.g. if all individuals born in the 1960s were exposed to an outbreak of a certain pathogen). Epidemiological data on Shigellosis in Israel [43] is in this respect somewhat inconclusive: while cases peaked in the 1980s, infections have remained higher after this peak than before it. A combination of exposures throughout an individual's lifespan together with aforementioned factors such as increased gut permeability and potential translocation in elderly may account for the antibody responses observed.

PhIP-Seq Ig-epitope repertoire data represents a unique layer of information compared to other methods of studying antibody responses against gut microbiota. FACS sorting and DNA sequencing based methods to elucidate antibody coating of resident gut microbiota [9,18–20] capture a snapshot of microbes currently present or panels of cultivatable organisms. Our approach offers a complementary strategy to study protein-based antigens and their epitopes at high resolution, including immunological memory of antigens previously encountered. This temporal aspect provides an additional layer of information beyond microbiome DNA sequencing (which is also limited to detection of bacteria present at the time of sample collection) and could inform on lasting immune effects of microbiota [44].

Our study is limited by technical characteristics of PhIP-Seq previously discussed in depth [21,22,25,26], most notably length constraints of the presented peptides (64 amino acids [aa] for our library). This length is expected to adequately represent linear epitopes, whereas conformational epitopes requiring correct folding of larger protein regions may be missed, impacting the sensitivity of our assay. The ratio of linear/conformational epitopes recognized by human antibodies is not exactly known and experimentally challenging to determine. Furthermore, the length distribution of conformational epitopes is unknown and it is unclear which percentage of conformational epitopes will be covered by 64 aa peptides. Previous use of the PhIP-Seq workflow has relied on similar peptides lengths [22,25,26,45] and yielded reliable results primarily related to autoimmunity [22,45] and viruses [25–27]. Yet, even if our PhIP-Seq approach could only detect 10% of antibody-antigen interactions targeting bacteria, the fact that our library covers more than 28,000 proteins would still surpass the throughput of current ELISA or peptide array-based approaches by an order of magnitude. Interestingly, peptides originating from known antibody binding proteins (such as protein A) within our library interacted

with the Fc region of antibodies (Fig. *2*) suggesting correct folding, despite incomplete length. Antibody binding events of single peptides identified by PhIP-Seq need to be interpreted with care and should be validated with orthogonal methods (*e.g.* Extended Data Fig. 7). However, the associations reported in this study are corroborated by binding against multiple proteins per species or even multiple peptides per protein (Fig. 4c,d), making random associations highly unlikely.

Several other antibody profiling methods have been used to generate serological classifiers of disease and assessing other biological parameters [46]. In our opinion, PhIP-Seq provides a good compromise of peptide lengths, library size, amenability for parallelized measurements, and cost, while allowing for rational selection of the presented peptides (*i.e.* not requiring the use of random peptides).

Our study is also limited to protein antigens. Microbiota antigens also include glycans, lipids, and post-translational modifications. Non-protein products such as LPS can exert powerful immune-modulatory effects on innate and adaptive immune cells[47]. Protein antigens are thought to elicit T cell dependent antibodies of high specificity, whereas non-protein antigens are generally targeted by low affinity, high avidity, T cell independent antibodies [4]. Therefore, the protein antigens used in our study may allow for more sensitive detection and could represent more promising biomarkers than low affinity antibodies against non-protein antigens. While our experimental approach informs on the functional antigens recognized by antibodies, linking these to the associated BCR sequences or demonstrating causality necessitates alternative experimental approaches (*e.g.* [48] or [10]).

It is increasingly appreciated that gut microbiota affect the immune system beyond the intestines and antibody responses against microbes have been implicated in several immune-mediated diseases beyond inflammatory bowel diseases [11,15], yet the actual antigens bound remain unknown. The microbiota antigen library created here could represent a powerful, broadly applicable tool to mine for systemic biomarkers and targets in these settings

Acknowledgments

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 J 4256). R.K.W. is supported by the Seerave Foundation and the Netherlands Organization for Scientific Research. A.Z. is supported by the ERC Starting Grant 715772, Netherlands Organization for Scientific Research NWO-VIDI grant 016.178.056, the Netherlands Heart Foundation CVON grant 2018-27, and the NWO Gravitation grant ExposomeNL 024.004.017. J.F. is supported by NWO Gravitation Netherlands Organ-on-Chip Initiative (024.003.001), the ERC Consolidator grant 101001678, and the Netherlands Heart Foundation CVON grant 2018-27. C.W. is supported by NWO Gravitation grant 024.003.001 and NWO Spinoza Prize SPI 92-266. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript

Author Contributions Statement

T.V., S.K. conceived the project and designed the library. S.L. designed and implemented the coding of the library. T.V., S.K. planning and calibration of experimental system, and performing biological experiments. S.L. designed and implemented computational pipeline. S.L., I.N.K performed high-throughput data analysis, T.V. analyzed additional data and wrote the manuscript. E.S. and A.W. conceived and directed the project. T.V., S.K., S.L., I.N.K., A.W., C.W., J.F., A.Z., R.K.W., E.S. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Data availability statement

The data generated or analyzed during this study is included within the manuscript, its supplementary information files, and public repositories. Detailed information on the cohort, library content, and PhIP-Seq data are available via the *Nature Medicine* website (files: cohort_info.csv , MB_composition.csv, library_content_info.csv, and PhIP-Seq_data.zip). Patient-related data not included in the paper may be subject to patient confidentiality. The following figures Extended Data Fig. 3, Fig. 1, Extended Data Fig. 5, Fig. 3a,b/Extended Data Fig. 6, and Fig. 4a,c have associated raw data provided in the respective supporting files Supplementary table 2, Supplementary table 3, Supplementary table 4, Supplementary table 5, and Supplementary table 6. Raw data of the PhIP-Seq experiments is deposited in the Harvard Dataverse public repository: https://doi.org/10.7910/DVN/3SOZCQ. Antigens included in the PhIP-Seq library were obtained from the immune epitope database (IEDB, https://www.iedb.org/) and virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/), as well as other sources outlined in the Methods.

Code availability statement

Custom code used for analyzing the PhIP-Seq data is publicly available: https://github.com/erans99/PhIPSeq_external The code repository is sub-divided into two sub-folders: 1.) Analyse_Fastq - Code to analyze a NextGen Sequencing plate, containing 96 wells, of which 80 are data wells, and 16 are different types of controls of well quality (4 negative controls, 8 mocks, and 4 positive control ('anchor') samples). The output of this is a file, per data well, of fold change and -log10(p-value)'s. 2.) Analysis - Code for executing different tests and analyses on the results of the PhIP-Seq output (as cached from files like the ones in the PhIPSeq_data directory).

Figures

Fig. 1

Fig. 1: PhIP-seq ~~profiling~~ of microbiota directed antibody epitope repertoires. a, Phage-display immunoprecipitation sequencing (PhIP-Seq) [21] workflow applied to measure serum antibody epitope repertoires. **b**, Content of the 244,000 variant antigen library. IEDB – immune epitope database [30], see the methods section, Supplementary table **1**, Extended Data Fig. 1 to Extended Data Fig. 3 for lists of the exact strains included and details on controls. **c**, Antibody epitope repertoire measurements were performed on a cohort of 997 individuals with diverse metadata available (see methods section for details on the blood tests). **d**, On average ca. 800 peptides of the microbiota library are significantly enriched per individual. The center line shows the median; box limits indicate the 25th and 75th percentiles as determined by R software [49]; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, *all data points are plotted*, n = 997 individuals. **e**, Antibody epitope repertoires of 997 individuals recognize private (occurring in single individuals) and public (shared in up to 99% of the cohort) microbiota antigens. **f**, Public microbiota antigens are not limited to pathogens but extend to diverse microbiota including commensal and probiotic bacteria (controls and IEDB [30] epitopes are not included in panels e,f).

Fig. 2

Fig. 2: Functional antibody binding proteins within the phage displayed microbiota antigen library. **a,b** Canonical binding of phage displayed antigens with the Fab of antibodies compared to interactions of potential phage displayed antibody binding proteins with the Fc part of antibodies. **c** Testing antibodies with known specificity suggest functional antibody binding proteins present in the phage displayed antigen library. Two monoclonal antibodies (IgG1-HER2 and IgG1-TNFa) and an IgG-Fc preparation were mixed with the phage displayed antigen library and processed in the same way as serum samples. Reaction mixtures were set up in triplicates (IgG-FC) or quadruplicates (IgG1-HER, IgG1-TNFa). Significantly bound peptides occurring in at least three reactions overall are listed. **d** For two variants of the Staphylococcal protein A antibody binding protein, we compared biochemical and structural information to the binding peptides, indicating that binding is mediated by B-domains known to bind the Fc region of IgG [33]. An alignment of the two highly similar proteins with the accession numbers WP_000728751 and WP_000728715 is shown (details on the proteins is provided in panel c). The dark line next to the accession number represents the protein sequence indicating gaps in the consensus alignment where applicable. The number right next to the dark line is the protein length in amino acids. B-domains are highlighted according to the information deposited in the NCBI entries of the respective accession numbers. The peptides binding to the antibodies listed in panel c are marked with their identifying number.

Fig. 3

Fig. 3: Comparison of antibody epitope repertoires against the microbiota library with microbiome composition inferred from metagenomics gut microbiome sequencing.. **a**, The hamming distances between pairs of serum antibody profiles (each of 996 individuals vs. all others) measured with this antigen library do not associate with Bray Curtis distances computed between the same pairs from metagenomics gut microbiome sequencing data derived of stool samples of the same individuals (Mantel test p-value 0.514). The same test was run for other distance measures for antibody profiles (Pearson correlation and Jaccard distance) as well as presence/absence of bacterial species/antibody responses (alternatively to fold change and abundances shown in panel a), yielding similar results. **b,** Histogram of the number of antibody bound peptides significantly correlated with the metagenomics abundance of bacterial species. We tested for correlations of antibody bound peptides vs. species abundance only for those appearing in >2% of individuals (Extended Data Fig. 6, Supplementary table 5). The histogram is showing the 1,706 pairs passing FDR correction for multiple hypothesis testing (approximately 4.5 million tests). Species abundances were computed on SGBs [50]. SGBs with ≥14 significantly bound peptides are marked with species names if annotated. 'u.s.': unknown species, see Supplementary table 5 for a full list and details for SGBs.

Fig. 4

Fig. 4:a,b, Serum antibody epitope repertoires associate with age and gender. a,b Each dot represents a peptide with its abundance in the respective cohort plotted on the x/y axes. The cut-offs of <28 and >61 years of age, were applied as they represent approximately the youngest and oldest decile of the studied cohort. Peptides bound significantly different between groups (Chi-Square test with one degree of freedom, FDR correction) are highlighted and listed in Supplementary table 6. **c,** Many of the peptides marked in panel a, that significantly associate with age, originate from a *Shigella* type III secretion system (T3SS). The illustration on the left side shows how *Shigella* cells can secrete effector proteins through their own inner and out membranes (IM/OM) into through the host membrane (Host M.) into the human target cells in the intestines. This illustration is derived from ref.[35]. On the right side, antibody binding against multiple peptides from the same T3SS proteins is illustrated. '*' denotes significantly differently bound peptides between older and younger individuals. Peptides marked with '**' were encoded twice in nearly identical form within the PhIP-Seq library and both were detected as significantly different. Additional peptides not passing the significance threshold (such as originating from ipaA) are also shown and would likely pass FDR correction with a larger sample number. The percentages of antibody binding in young and old individuals per peptide is indicated above each peptide. Only peptides bound in >5 individuals are shown. See Supplementary table 6 for details on the peptides. **d,** Most antibody bound peptides associated with gender (shown in panel b) originate from surface proteins of *Lactobacillus* sp., including S-layer (SL) proteins [36]. For these peptides, the frequency of antibody binding in women and men is listed as well as p-values for the significance (Chi-Square test with one degree of freedom, FDR correction; details in Supplementary table 6). Overlapping peptides from 2 proteins significantly differently bound between women and men are shown at the right side of the panel. See Supplementary table 6 for details on the peptides.

Fig. 5

Fig. 5: Serum antibody epitope repertoires predict age and gender by machine learning better than metagenomics gut microbiome sequencing. a,b Machine learning based predictions of age and gender from population wide antibody epitope repertoires were performed with XGBoost with 10-fold cross validation. **c-f,** Machine learning based predictions of age (c) and gender (e) from metagenomics gut microbiome abundances are surpassed by antibody repertoire-based predictions. Combined machine learning based predictions were either performed by averaging results of separate predictions of antibody epitope repertoire and metagenomics abundances (ensemble) or using antibody and metagenomics data together as features for building predictors. Average and standard deviation derived by 10 repeats of XGBoost with 10-fold cross validation. The predictive power of antibody epitope repertoires for age (d) and gender (f) is not limited to antigens of pathogens (either from VFDB or pathogenic strains) but extends to antigens selected from metagenomics sequencing (strains and genes outlined in Supplementary table 1 and the methods section). See Supplementary figure 4 for extended machine learning based predictions on subgroups of the antigen library. Color legend: antibody epitope repertoire associations/machine learning based predictions with age (blue) and gender (purple), analyses involving metagenomics gut microbiome abundances data (yellow). Antigen groups sizes for panels d,f: All microbiota – 231,975 peptides, VFDB – 24,164 peptides, Library excluding VFDB – 207,811 peptides, Metagenomics antigens - 147,061 peptides.

Fig. 6

Fig. 6: The longitudinal stability of antibody epitope repertoires over five years. a, Antibody epitope repertoires of 213 individuals over 5 years against the entire library of microbiota antigens show high stability. Pearson correlations of log fold changes of all baseline (t=0) and follow up (t=5 years) samples compared with each other are shown. **b**, Correlation coefficients of panel 'a' are shown as histogram for comparisons of random pairs of samples (left y-axis, 213^2-213 comparisons) and individuals' matched samples (right y-axis, 213 comparisons) collected five years apart. **c**, Average correlation coefficients of the stability of antibody epitope repertoires against antigen subgroups of pathogens (VFDB) and antigens selected from metagenomics sequencing data of this cohort [28] are compared to the antigens of the complete library. Mean values and standard deviations of n=213, see Extended Data Fig. 10 for diagrams and correlations of additional antigen sub groups. Same sample size as outlined in panel b.. Antigen groups sizes for panels: All microbiota – 231,975 peptides, VFDB – 24,164 peptides, Metagenomics antigens - 147,061 peptides. **d**, Gut microbiome stability inferred from abundances of metagenomics sequencing of stool samples collected five years apart of 188 individuals. The Bray Curtis distances for all baseline (t=0) and follow up (t=5 years) samples compared with each other are shown.

Methods

Serum samples, clinical data, and metagenomics

1,051 serum samples of 1,07 individuals had been collected in Israel in 2013/2014 for previous studies [28,34] along with clinical and metagenomics data. Various phenotypes and blood test results were available for most (>900 for phenotype/blood test) individuals [9], with results of few tests missing in some individuals. We focused most of the antibody epitope repertoire analysis on baseline samples (1st sample collected per individual). Ten samples did not pass the threshold of >200 peptides significantly bound and were excluded from analyses (see section 'Data analysis' below), leaving data of 997 individuals for the analysis. The 213 longitudinal serum samples and 188 stool samples for metagenomics sequencing had been obtained from participants of one of the previous studies [28] after ca. five years in 2019/2020. Research with these samples has been approved by the Tel Aviv Sourasky Medical Center (#0658-12-TLV) and the Weizmann Institute of Science's institutional review board (#1079-1) and the participants had consented to using the samples.

Processing of antigen sequences and cloning of the phage library

See the section below for a detailed description of the content of the antigen phage library. The final list of proteins were cut to peptides of 64 amino acids (aa) with 20 aa overlaps (to cover all possible epitopes of the maximal length of linear epitope [depending on the definition between five to nine up to 20 aa [52–54]]) between adjacent peptides. The peptide aa sequences were reverse translated to DNA using the *Escherichia coli* codon usage (of highly expressed proteins), aiming to preserve the original codon usage frequencies, excluding restriction sites for cloning (*Eco*RI and *Hind*III) within the coding sequence (CDS). The coding was re-performed, if needed, so that two possible barcodes were formed in the CDS, by the 44/75 nt at the 3' end of each oligo. Every such barcode is a unique sequence at Hamming distance three (with a 44 nt read, or five with a 75 nt read) from all prior sequences in the library, which allows for correcting of a single read error in sequencing the barcode with a 44 nt read (reading 75 nt continuously would allow to correct two read errors). Eventually, we used the 44 nt read option and sequenced also a section of the 5' end (see below, in order to verify matching 5' and 3' sequences and exclude potential presence of multiple inserts). For similar peptide sequences, alternative codons were used following *E. coli* codon usage, to achieve discrimination. Including the sequencing barcode as part of the CDS, rather than a separate barcode, allowed the use of the entire oligo for encoding a peptide (and as opposed to completely omitting a barcode, it did not require sequencing the complete CDS). For encoding peptides shorter than 64 aa, a random sequence was added after the stop codon with addition of the restriction site *Swa*I (allowing to remove short peptides by restriction enzyme digestion on the oligo level in case they would take over the signal [which was eventually not observed and digestion hence not required]). After finalizing the peptide sequence, the *Eco*RI and *Hind*III restriction sites, stop codon, and annealing sequences for library amplification were added and ordered from Agilent Technologies as 230mer pool (library amplification primers, fwd: GATGCGCCGTGGGAATTCT, rev: GTCGGGTGGCAAGCTTTCA) and cloned into T7 phages following the manufacturers recommendations (Merck, T7Select®10-3 Cloning Kit, product number 70550-3).

Controls for the effect of different DNA encodings of the same amino acid sequence

Employing different DNA sequences to encode the same amino acid sequence yielded generally good agreement both when comparing fold change (top of panels a and b of Supplementary figure 1) or population wide abundance (Supplementary figure 1, bottom of panels a,b). The vast majority of peptides were reproducibly not bound in any individuals (95% of comparisons, Supplementary figure 1c). For peptides bound in at least one individual (Supplementary figure 1d), all triplicates were in

agreement in 71% of cases. Comparing the calculated p-values for each encoding of a peptide with the other two encodings in all individuals (3 encodings of 347 peptides in 997 individuals representing ca. 1 million comparisons) yielded good agreement (R^2=0.77). There were a few peptides for which one DNA encoding strongly differed from the other two. For example, the 11th peptide of the human gamma herpesvirus 4 EBNA 1 protein appeared in 2/3 encodings in ca. 30% of individuals, but the third encoding was not detectable at all (panel a, bottom) with direct effects on the observed fold changes (Supplementary figure 1a, top).
These results are not solely impacted by the DNA encodings, but also different abundances of DNA oligoes within the manufacturing process. Hence, care should be applied when comparing different oligoes (as the absolute values can be impact by DNA encodings or oligo manufacturing).
As expected, encoding viral and bacterial peptides with different DNA sequences yielded rather frequent antibody binding (Supplementary figure 1a), while peptides originating of human proteins displayed very little binding (indicating that different encodings do not represent a major source for false positives). These triplicate encoding results also confirm results from single encoding controls (Extended Data Fig. 2a).
Overall, the variability between different encodings was surpassed by the variability in antibody binding between individuals (standard deviations in the top of Supplementary figure 1a,b), indicating little bias for the population scale analysis performed in this work.

Content and design of the PhIP-seq microbiota antigen library

Given the enormous complexity of potential antigens from human microbiota (for example, the Integrated Reference Catalog of the human gut microbiome (IGC) is composed of 10^7 genes [55]), it is with current DNA synthesis technologies not possible to represent the entire human microbiome. We aimed to broadly cover both potential uncharacterized antigens as well as previously reported bacterial strains and proteins eliciting antibody responses by rationally choosing potential antigens (section 'Library content' below). Antibody binding of live bacteria is focused on exposed surface or secreted proteins; hence we enriched the library for these protein groups (section 'Selection of protein targets' below). Moreover, due to current limits of DNA oligo synthesis (230 nt for this library), most proteins were split into peptides. These peptides' amino acid sequences were reverse translated to *Escherichia coli* codon usage (Material and Methods section). Ultimately, we generated a library representing 244,000 peptides derived from 28,668 proteins (thereof 27,837 microbiota proteins [excluding proteins from the IEBD and controls, see below]).

Library content

Bacterial species and databases

About 60% (147,061 oligoes) of the library content (Fig. 1b) was dedicated in an unbiased manner to potential antigens from the microbiome of healthy individuals. For this, we used gene and species abundances from metagenomics data of 953 stool samples of the same cohort (personalized nutrition project, PNP [28]) on whom we eventually performed the antibody epitope repertoire profiling. Another 25% (61,250 oligoes) were dedicated to pathogenic bacteria, probiotic bacteria, and gut microbiota previously reported to be coated by antibodies [19]. We also included the entire virulence factor database (VFDB, [29]) making up 10% (24,164 oligoes) of the library and left 5% (11,525 oligoes) of the library for various controls (such as infectious disease and auto-immune human proteins from the immune epitope database [IEDB] [30] and technical controls).

Metagenomics data of the cohort: Selection of genes and species (MetaPhlAn2)

Metagenomics data from shotgun sequencing of healthy individuals of our cohort was processed in two ways to select antigens: First, mapping to the IGC database and the calculation of the relative abundance of each gene was performed as previously described [28,56,57]. The genes data of the PNP

cohort contained approximately $4*10^6$ different genes that were mapped to the IGC [55]. Fifty percent of the library content was filled with peptides derived from the proteins encoded by these genes (see below for the exact selection criteria). Second, in addition to this gene database, we dedicated another 10% of the library to abundant strains identified by MetaPhlAn2 (MPA), a computational tool for profiling the phylogenetic composition of microbial communities from metagenomic shotgun sequencing data [58]. We included this strain-based approach, to mimic the selection process of pathogenic, probiotic and antibody coated strains described below. After sorting for the ten most abundant bacterial strains using MetaPhlAn2, the fasta files of the bacteria's proteins were downloaded from the NCBI (Supplementary table 1) and processed as outlined below to select potential antigens.

Pathogenic, probiotic, and antibody coated bacterial species

In addition to commensal bacteria of healthy individuals, we added three more groups of bacterial species: gut pathogens, probiotic strains, and bacteria reported to be coated with IgA in previous studies [19] accounting together for 25% of the library content (Fig. 1b).

Seventeen bacterial species known to be (gut) pathogens were chosen based on their likelihood to have been encountered by our Israeli cohort. We focused on gut pathogens and chose the most prevalent ones (e.g. *Campylobacter*, *Shigella* and *Salmonella*) according to a report of the central laboratories of the Israeli ministry of health from 2015. In addition, we added also *Listeria* which can cause serious illness in pregnant women, newborns, adults with weakened immune systems and the elderly (Supplementary table 1).

Probiotic strains (Supplementary table 1) were chosen based on a recent review by Lebeer *et al.* [31].

Bacterial species coated by antibodies were chosen based on the work of Palm *et al*. [19], who examined the microbiota coated by IgA in healthy individuals, Crohn's disease (CD) and ulcerative colitis (UC) patients. Bacteria passing the threshold of relative abundance of greater than 10^{-6} and IgA coating index >10 in at least three patients were chosen. All together, nine such bacterial species were selected, five species that were abundantly bound in healthy individuals, two from CD patients and two from UC patients.

All the genomes, from pathogenic, probiotic and IgA coated bacteria, were downloaded from the NCBI and are summarized in Supplementary table 1 including accession numbers.

Virulence factor database

In addition to these bacterial species, we included the virulence factor database (VFDB, [29]) to represent pathogenic species at greater depth, accounting for 10% of the library. The proliferation of pathogenic bacteria in their host depends on their ability to deploy virulence factors (VFs) to establish infections, survive in the hostile host environment and as a result cause disease. We included the entire 'set A' of the VFDB, which covers genes associated with experimentally verified virulence factors representing 2,624 gene sequences.

Positive/negative controls

We benchmarked and validated the antibody reactivities against microbiota proteins (described above) with several control antigens. We therefore included 12,025 oligoes covering proteins from the following groups: 1.) proteins of various infectious diseases 2.) human proteins known as targets in autoimmune diseases and 3.) technical controls (such as identical amino acid sequences coded by differently codon optimized DNA sequences and random amino acid sequences).

Positive and negative controls of infectious diseases and human proteins

We have included subsets of B cell antigens from the IEDB (the immune epitope database), the most comprehensive repository covering various antigens reported in the literature [30]. As positive controls, we have selected all antigen epitopes from B cell assays labelled as infectious diseases (excluding

parasites) with human host. These 290 proteins have been reported in the literature to be targets of antibody responses and were covered with 4,250 oligoes.
As negative controls, antigens from B cell assays of human autoimmune diseases were included (as these proteins should not lead to a strong response in our healthy cohort) representing 430 proteins and 7,700 oligoes. Not only the exact epitopes reported in the IEDB, but the full-length protein sequences (obtained from UniProt by the accession numbers listed in the IEDB) were used and divided into overlapping oligoes as described below.
In addition to these IEDB positive/negative controls, we have included additional control antigens. We have added viral proteins, that have previously been reported to elicit recurrent antibody responses in 47.9 to 97.2% of humans using a similar phage display approach (original Table S2 by Xu *et al*. [25]). Both full length proteins divided into overlapping oligoes and the exact short peptides reported by Xu *et al*. [25] were included.
Furthermore, we included negative controls that should not have been encountered by our cohort and hence not elicit antibody responses, such as several Ebola proteins. In addition to human proteins from the IEDB (with known auto-reactivities), we have also included several other abundant human proteins that should not evoke antibody reactivities in healthy individuals (such as serum albumin, histone proteins, glycolysis enzymes and ribosomal proteins). These sequences are represented by 300 oligos. Results of positive and negative controls are shown and discussed in Extended Data Fig. 2a,b.

Technical controls

In addition to these biological positive and negative controls with expectation towards antibody binding, we also included 450 control oligoes to assess technical aspects of the experimental system, and 100 oligos encoding random amino acid sequences (without internal stop codons), that should not be recognized by antibodies (results shown and discussed in Extended Data Fig. 2a).
Furthermore, we included codon optimization replicate controls (350 oligoes) to test for biases of representing the same amino acid sequence with different DNA sequences. Oligoes from both the microbiota library and the positive and negative controls were chosen and encoded by three different codon optimized sequences coding for the same amino acid sequence (results shown in Supplementary figure 1). Additionally, 50 oligoes representing short peptides (<45 aa) were included to test for additional effects of varying the random sequence at the 3' end (see detailed explanation below).

Selection of microbiota protein targets

The pool of microbiota genes derived from metagenomics (ca. four million) and all proteins of the selected pathogenic, probiotic, and antibody coated strains (Supplementary table 1) exceeded the library size of 244,000 variants. Hence, we enriched the library for proteins expected to elicit more frequent binding (such as highly abundant genes, and bacterial genes identified as flagella, membrane or secreted proteins that are more likely to be exposed to antibody binding than intracellular proteins).

Selection by abundance and annotation

Using the metagenomics data of relative abundance of genes, subsets of sequences were chosen solely based on abundances in the cohort, starting with a cut-off of 10^{-6} relative abundance (RA) as criterion for presence in our cohort. Three percent of the library was dedicated to the most abundant genes occurring in >95% of the cohort (highly abundant), 3% of the library was dedicated to genes that appear in half of the cohort (moderately abundant) and 3% was dedicated to genes that appear in less than 1 percent of the cohort (rarely abundant).

Another set of genes was selected based on annotations and cellular localization predictions focusing on proteins that have higher chance to be exposed to the host's immune system. We started with genes that were present in more than 20% of our cohort resulting in a list of ca. 140k genes. We focused on three groups: membrane proteins, secreted proteins and motility proteins/flagella, as these proteins are surface exposed [59] and have previously been reported to be bound by antibodies in small scale studies [7].

To assign these functionalities/localizations to gene sequences (to select membrane/secreted/motility proteins), we applied Blast2GO, a bioinformatics platform for the high-throughput and automatic functional annotation of DNA or protein sequences based on the Gene Ontology database [60]. The BLAST step was done locally against the NCBI non redundant protein database with up to 10 hits per sequence. The analysis of the GO was done locally (Database that was updated to 01.2017) using the 2.8 version of Blast2GO. Proteins that were assigned GO terms of membrane localization or extracellular localization or secretion or motility were filtered out. This step resulted a list of ca. 34,000 membrane proteins, 461 secreted proteins and ca. 100 motility proteins.

Membrane protein selection

Membrane proteins contain three distinct parts: transmembrane domains, extracellular domains, and intracellular domains. We have focused on the extracellular domains as they are more likely to be bound by antibodies and we have avoided hydrophobic transmembrane domains. We used TopGraph for the prediction of intracellular, membrane, and extracellular sequences of the membrane proteins. Extracellular domains with a length of >20 amino acids were included in the library (alongside a control set of full-length membrane proteins representing ca. 600 proteins).

Secreted protein selection

In addition to the Blast2GO approach, SignalP 4.0 was used for the prediction of signal peptides (SPs). The 140K genes (appearing in >20% of the cohort) were analyzed by SignalP 4.0 for both gram-positive and gram-negative signal peptides. The sequences that were predicted to have SPs were filtered out and the mature sequences (without SPs) included in the library (ca. 7,000 proteins).

Not all secretory proteins carry signal peptides. Some proteins, including various virulence factors, enter a non-classical secretory pathway without any currently known sequence motif. In Gram-negative bacteria, type I, III, IV and VI secretion systems function without signal peptides. As another approach to select for secreted proteins, we used DIAMOND which is an alignment algorithm potentially more than 20,000 times faster than BLASTX while maintaining a similar sensitivity. First, reference databases were created by searching the UniProt website (http://www.uniprot.org/) for bacterial toxins, and flagella proteins (only reviewed sequences were chosen). Then we searched for hits between the entire IGC database of human gut microbiome genes and these well characterized reference databases of bacterial toxins and flagella using DIAMOND. Genes in the IGC with at least one match with an E value $<10^{-6}$ were filtered out. This approach resulted in additional 324 predicted toxins and 1,265 predicted flagella proteins.

The same approaches for selecting membrane and secreted proteins applied to the metagenomics data, were also applied to the pathogenic, probiotic, and antibody coated strains etc. (Supplementary table 1) to enrich for proteins potentially targeted by antibodies.

Clustering by CDhit

In order to avoid redundancy due to sequences that are highly similar in the library we used CDhit for clustering. All the metagenomics data (genes and strains) were concatenated in two groups, TopGraph sequences (membrane proteins) and the rest. Sequences of pathogenic, probiotic, and antibody coated strains were treated in the same manner. Each group was clustered to 70% homology and the cluster representatives were chosen for the next step. Membrane and secreted proteins from

metagenomics data were selected based on the original abundances of the genes. All predicted secreted proteins from the genomes of selected bacteria were included, but membrane protein sequences were randomly selected from a subset of the strains (indicated in Supplementary table 1).

Immunoprecipitation and sequencing

The PhIP-Seq experiments were performed as outlined in a published protocol [21] with the following modifications: PCR plates for the transfer of beads and washing were blocked with 150 µL BSA (30 g/L in DPBS buffer, incubation overnight at 4°C) and BSA was added to diluted phage/buffer mixtures for immunoprecipitations (IPs) to 2 g/L. Phage wash buffer for IPs was prepared as outlined [21] with 0.1% (w/v) IPEGAL CA 630 (Sigma-Aldrich cat. # I3021). To determine the optimal ratio of phages and antibodies per reaction, we mixed phage amounts ranging from a 2,000- to a 16,000-fold coverage per variant with antibody amounts ranging from 0 to 16 µg. The optimal concentrations appeared to be a 4,000-fold coverage of phages per variant and between 2 to 4 µg of antibodies. While the optimal antibody amount used is similar to previous PhIP-Seq applications (2 µg recommended by Mohan *et al*. [21]), the number of phages per library variant is lower (10^5 phages per variant recommended by Mohan *et al*.). This difference may be due to different binding potential of this novel microbiota antigen library or additional blocking steps performed (we added BSA to the diluted reaction mixtures and blocked also the PCR plate used for the washing steps with BSA). After optimizing phage and antibody amounts for IPs (Extended Data Fig. 1a), 3 µg of serum IgG antibodies (measured by ELISA) were mixed with the phage library (4,000-fold coverage of phages per library variant). . As technical replicates of the same sample were in excellent agreement (average Pearson R^2=0.96, n=191, Extended Data Fig. 1b), measurements were performed in single reactions. The microbiota library was mixed in a 2:1 ratio with a 200mer 100,000 variant pool (manuscript in preparation).

The phage library and antibody mixtures were incubated in 96 deep well plates at 4°C with overhead mixing on a rotator. Forty microliters of a 1:1 mixture of protein A and G magnetic beads (Thermo Fisher Scientific, catalog numbers 10008D and 10009D, washed according to the manufacturers recommendations) were added after overnight incubation and incubated on a rotator for at 4°C. After four hours, the beads were transferred to PCR plates and washed twice as previously reported [21] using a Tecan Freedom Evo liquid handling robot with filter tips. The following PCR amplifications for pooled Illumina amplicon sequencing were performed with Q5 polymerase (New England Biolabs, catalog number M0493L) according to the manufacturers recommendations (primer pairs PCR1: tcgtcggcagcgtcagatgtgtataagagacagGTTACTCGAGTGCGGCCGCAAGC and gtctcgtgggctcggagatgtgtataagagacagATGCTCGGGGATCCGAATTC, PCR2: Illumina Nextera combinatorial dual index primers, PCR3 [of PCR2 pools]: AATGATACGGCGACCACCGA and CAAGCAGAAGACGGCATACGA [21]). PCR3 products were cut from agarose gel and purified twice (1x QIAquick Gel Extraction Kit, 1x QIAquick PCR purification kit; Qiagen catalog numbers 28704/28104) and sequenced on an Illumina NextSeq machine (custom primers for R1: ttactcgagtgcggccgcaagctttca, and for R2: tgtgtataagagacagatgctcggggatccgaattct, R1/R2 44/31 nts). Paired end reads were processed as described below.

Data analysis

All analysis code was written in Python (3.7.4), using the libraries sklearn (0.23.2), scipy (1.5.4), statsmodels (0.12.1), pandas (1.1.5), numpy (1.18.5), matplotlib (3.3.3), and seaborn (0.11.0). Also xgboost (1.18.5) and shap (0.37.0) implementations were used. Additional data analysis software: BoxPlotR/R software (Spitzer et al., 2014), MetaPhlAn2 (Truong et al., 2015). DNA sequencing (performed on the Illumina NextSeq platform) reads of IPs were down-sampled to 1.25 million IDable reads per sample, *i.e.* reads with a barcode within one error of the set of possible barcodes of the two mixed libraries for which the paired end matched the IDed oligo. When not enough reads were obtained, a minimal threshold of 750,000 reads was enforced for data analysis. Enriched

peptides were calculated by comparing the number of reads per oligo to that of input coverage (library sequencing of phages before IPs). Scoring was done assuming each input level creates an output level distribution which is a Generalized Poisson distribution. Parameters for this Generalized Poisson distribution were estimated for each input level of each sample separately, and then fitted to three parameters for the whole samples, extrapolated for each input level and scored [22]. Derived p-values were subject to Bonferroni correction (p-value 0.05) for multiple hypothesis testing, and log-fold-change (number of reads of bound peptides vs. baseline sequencing of phages not undergoing IPs) was computed for all peptides which passed the threshold p-value, all other peptides were given a log-fold-change value of 0. Samples, for which less than 200 peptides significantly bound, were excluded from analyses. The input sequencing of the phage library was before IPs was performed at >100-fold coverage. For the calculation fold changes, input reads were set to a minimum of 25 reads.

We used the gradient boosting trees regressor from Xgboost [37] as the algorithm for the regression predictive model for different phenotypes. We used the gradient boosting trees classifier from Xgboost as the algorithm for the classification predictive model for phenotypes with binary values.
The parameters of the predictors when using microbiome features were: colsample_bylevel=0.075, max_depth=6, learning_rate=0.0025, n_estimators=4000, subsample=0.6, min_child_weight=20. These parameters were used for regression as well as classification. The rest of the parameters had the default values of Xgboost.
All analysis was performed by 10-fold cross validation, so that any overfitting would only worsen prediction accuracy.
In general, adding irrelevant features to an Xgboost model will inevitably worsen predictions, as some of the trees will not have any relevant features in them, which will add noise to the prediction. This effect is stronger the larger the proportion of irrelevant features, so it is expected that prediction of any phenotype by age and gender alone would be much better than the same prediction with many extra features (in our case log fold changes of peptides), if they do not have a significant contribution to the phenotype.

Raw data and code

Raw data files

library_content_info.csv

This file is directly available via the *Nature medicine* website. Details on the 244,000 peptides contained within the PhIP-Seq microbiota library. Every line represents a phage displayed peptide numbered consecutively (column 'peptide_number'). 'pos' refers to the starting position of the peptide within the originating protein. 'len_seq' indicates the full length of the originating protein. 'aa_seq': amino acid sequence of the peptide. The following columns indicate the origin of the selected proteins including the Immune epitope database (is_IEDB), positive controls (is_pos_cntrl), negative controls (is_neg_cntrl), random peptides (is_rand_cntrl), the virulence factor database (is_VFDB), sequences selected from gut microbiota metagenomics sequencing (is_gut_microbiome), pathogenic strains (is_patho_strain), antibody coated strains (is_IgA_coated_strain), probiotic strains (is_probio_strain), and if applicable the bacterial strain of origin (bac_src). Proteins functions were annotated by mapping to the UniRef90 database (uniref and uniref_func).

cohort_info.csv

This file is directly available via the *Nature medicine* website. Details on the individuals and serum samples that were analyzed (including longitudinal samples of the same individual). The first column contains information on the individual and sample number in the format "individuals' number" _X_ "sample number". 'yob' – year of birth, gender: 0 = female, 1 = male, 'bmi' – body mass index,

'bt__crp_hs' - C-Reactive Protein blood test, "bt__hba1c" - Hemoglobin A1C. 'old_RegistrationCode' is used for matching the 213 longitudinal samples with their counterparts. 'num_passed_total' and 'num_passed_microbiota' number of peptides significantly bound by antibodies in the PhIP-Seq assay (all peptides from the two mixed libraries vs. only peptides from the microbiota library, see Materials and Methods section). When computing age ranges, the main text (e.g. Fig. 1c) only reports the age range for the 997 baseline samples, which is 17-70 years. For 213 individuals, we had collected follow up samples after approximately 5 years. Metadata for these follow up samples is also provided in the cohort_info.csv file. By chance, one of the oldest individuals of the baseline cohort (70 years), was amongst the follow up samples collected, so the total age range increases to 75 years when looking at the baseline + follow up samples.

MB_composition.csv

Microbiome composition inferred from metagenomics sequencing of stool samples of the respective individuals [17,28,61]. The same identifier as used in the information on the cohort can be used to match the antibody and metagenomics datasets.

PhIP-Seq_data directory

This .zip file is directly available via the *Nature medicine* website. PhIP-Seq results of each sample measured are provided by applying the same identifiers given in 'cohort_info.csv'. Every line represents a peptide significantly bound by antibodies (with the same identifiers as in the file 'library_content_info.csv'). 'fold_change' (from base input levels) and 'p_value' (-log10 of the p-value, based on input and output levels) metrics were computed with the Generalized Poisson distribution approach as outlined in the Materials and Methods section. Raw data of the Illumina NextSeq reads is provided in https://www.dropbox.com/sh/uqjttdpu2ws34lk/AAAOv_WbpXzzf__-i6jdpsE_a?dl=0 [we are in parallel also in the process of depositing this data in a publicly hosted repository].

Code repository

Custom code used for analyzing the PhIP-Seq data is publicly available: https://github.com/erans99/PhIPSeq_external

The code repository is sub-divided into two sub-folders:

Analyse_Fastq

Code to analyze a NextGen Sequencing plate, containing 96 wells, of which 80 are data wells, and 16 are different types of controls of well quality (4 negative controls, 8 mocks, and 4 positive control ('anchor') samples).

The output of this is a file, per data well, of fold change and -log10(p-value)'s.

Analysis

Code for executing different tests and analyses on the results of the PhIP-Seq output (as cached from files like the ones in the PhIPSeq_data directory).

Validation experiments

Detection of epitopes recognized by antibody preparations generated against immunogens of full-length proteins and bacterial cells

We obtained antibody preparations generated by immunizing rabbits/goats either with single bacterial proteins or with inactivated bacterial strains (see the first sheet of supporting file .xlsx file Supplementary table 2 for details on the antibodies and immunogens). These samples were processed following our standard PhIP-Seq workflow also applied to human samples (Extended Data Fig. 3).

Antibody binding with protein A and protein G coated beads separately and antibody coated beads capturing IgA and IgG separately

To gain understanding by which antibody classes the antigens of our library are bound, we performed an experiment with altered immunoprecipitation conditions (Fig. 1a). In addition to using a mixture of protein A and protein G coated beads (which binds all antibody classes), we mixed the same serum samples separately with protein A alone and protein G alone. According to the manufacturer's specifications of the superparamagnetic beads used in these experiments (Thermo Fisher Scientific, catalog numbers 10008D [protein A] and 10009D [protein G]), protein A binds strongly to human IgG1,2,4 and weakly/moderately to IgG3, IgA, IgM, and IgE, while it does not bind IgD. In contrast, protein G binds strongly to human IgG1,2,4 as well as IgG3, but does not bind to IgA, IgM, IgE or IgD. We processed serum samples of 78 individuals each with a mixture of protein A and G (equivalent to the standard protocol used for serum measurements shown in this work), protein A alone, and protein G alone.

Hence, antigens detected with both protein A and protein G indicate binding of IgG subclasses, whereas antigens bound by IgA, IgM, and IgE can be identified by only binding to protein A beads (Extended Data Fig. 5a-c).In addition to the experiments with protein A and G separately, we also verified the same set of 80 samples with beads covered with IgG and IgA capture antibodies (following a published PhIP-Seq protocol [27]). Rather than mixing the phage/antibody complexes with protein A+G, we mixed them with IgA and IgG specific biotinylated capture antibodies (Mouse Anti-Human IgG Fc-BIOT and Goat Anti-Human IgA-BIOT, Southern Biotech) by adding 6 µg of each capture antibody (in a separate reaction) prior to the overnight incubation step (outlined in the methods section). Sample IgG concentrations (3 µg used per reaction) were determined as outlined in the methods section, for IgA concentration measurements we applied a Human IgA ELISA kit (abcam, ab196263) and also used 3 µg per reaction. For the pulldown in the immunoprecipitation step, 25 µL of Pierce Streptavidin Magnetic Beads (ThermoFisher Scientific) were added per reaction (washed according to the manufacturer's recommendations). The following incubation/washing steps were performed identical to when using a mixture of protein A and G.

Following this protocol, we measured serum samples of 80 individuals with this IgA and IgG specific workflow (Extended Data Fig. 5d). These were the same 80 samples, that we had also measured with the standard protein A+G workflow and protein A and protein G separately (Extended Data Fig. 5a-c).

Isotype control experiments on antibody binding proteins

We performed Isotype control experiments (**Fig. *2***), indicating the presence of antibody binding proteins within the microbiota antigen library. When studying antigens occurring nearly universally in our cohort (Fig. 1e), we noticed the frequent appearance of *Staphylococcus* protein A. Protein A (as well as and *Streptococcus* protein G) is an antibody binding protein interacting with the Fc region of antibodies. The magnetic beads used in this study are for example coated with proteins A and G to carry out the immunoprecipitation and washing steps (**Fig. *2***a/Fig. 1a). We hypothesized that the frequent binding of these peptides may not be due to interactions of antibodies Fab region (**Fig. *2***a), but rather their nature as antibody binding proteins and interactions with the Fc part (**Fig. *2***b).

Therefore, we performed isotype control experiments with commercial antibodies/fragments to probe for interactions of our phage library beyond canonical Fab dependent binding. We have included two IgG monoclonal antibodies (mAbs) with different specificities (IgG1 anti human HER2 [R&D systems, catalog number MAB9589], and IgG1 anti human TNFα [R&D systems, catalog number MAB9677]). Additionally, an Fc preparation of IgG from human blood was included (Novus, catalog number NBP2-47132). The two IgG mAbs could allow to detect cross-reactivities of single Fabs, whereas the Fc preparation should completely eliminate any contribution of the Fab to the detected binding.

The mABs and the Fc preparation were mixed with the phage library and treated in the same way as regular serum samples (using also the identical amount of 3 µg per reaction).

Peptide ELISAs

To validate the PhIP-Seq results, we selected 6 peptides included within our PhIP-Seq library for analysis in a peptide ELISA (results shown in Extended Data Fig. 7). We included a positive control of a viral peptide (Epstein-Barr virus [EBV] nuclear antigen 1) with frequent population scale antibody responses [25] (corresponding peptide in the PhIP-Seq library: #3387) as well as a negative control of a human protein (SAPK4/MAPK13) that was expected not to elicit antibody binding in sera of healthy individuals (corresponding PhIP-Seq peptides: #1575, #1576, #1577 [the identical peptide was encoded as negative controls 3x within the library, with neither DNA encoding of the peptide eliciting binding, see Supplementary figure 1b for details]).

We also included peptides of two *Shigella* proteins ipaC and icsA/virG associated with age (Fig. 4c, Supplementary table 6) as well as peptides of *Staphylococcus* (extracellular matrix protein-binding adhesin, Emp, WP_000728052.1) and *Streptococcus* proteins (CHAP domain-containing protein, WP_020916184.1) frequently bound in PhIP-Seq. As chemical synthesis of 64 aa peptides displayed on the phages is costly, we aimed to reduce the peptide length. Therefore, we selected 20 aa sections representing the overlap of adjacent peptides of the same protein bound in PhIP-Seq. This overlap can for example be observed in Fig. 4c/Supplementary table 6 for the Shigella ipaC peptides #226014 and #232269. Likewise, also 20 aa from the overlap of peptides of icsA (#221918 and #235092), Staph. Emp (#180309 and #24623), and Strep. Chap (#110572 and #169922) were selected. As both of these peptides were bound in PhIP-Seq, they may share the same epitope covered by the overlap between them. The following aa sequences were selected: EBV - PPPGRRPFFHPVAEADYFEY, SAPK4 - KIMGMEFSEEKIQYLVYQML, Shig. ipaC - GKNPVLTTTLNDDQLLKLSE, Shig. icsA/virG - NGGDSITGSDLSIINQGMIL, Staph Emp - ASEDKLNKIADPSAASKIVD, Strep Chap - SATSYINTILNSKSVSDAIN.

These aa sequences were ordered from JPT Peptide Technologies (Berlin, Germany) as biotinylated chemically synthesized peptides and the peptide ELISA was performed according to the manufacturer's guidelines with the recommended concentrations (Protocols BioTides™ Peptides Revision 1.0, and Peptide ELISA Revision 1.2). In short, the peptides were bound to Streptavidin coated plates (Thermo Scientific™ Nunc™ Immobilizer™ Streptavidin Plates, cat. no. 436014) and incubated with serum samples (diluted 1:1,000 fold). Antibody binding was detected with an HRP conjugated anti human IgG antibody (Southern Biotech, cat. no. 204205) and TMB as substrate. Sera of 80 individuals (for whom also PhIP-Seq data was available and the protein A/G, IgG/IgA experiments had been performed Extended Data Fig. 5) were tested with each of the 6 peptides.

References

1. Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164**, 337–40 (2016).

2. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).

3. Levy, M., Kolodziejczyk, A. A., Thaiss, C. A. & Elinav, E. Dysbiosis and the immune system. *Nat. Rev. Immunol.* **17**, 219–232 (2017).

4. Bunker, J. J. & Bendelac, A. IgA Responses to Microbiota. *Immunity* **49**, 211–224 (2018).

5. Koch, M. A. *et al.* Maternal IgG and IgA Antibodies Dampen Mucosal T Helper Cell Responses in Early Life. *Cell* **165**, 827–841 (2016).

6. Gomez de Agüero, M. *et al.* The maternal microbiota drives early postnatal innate immune development. *Science* **351**, 1296–302 (2016).

7. Zeng, M. Y. *et al.* Gut Microbiota-Induced Immunoglobulin G Controls Systemic Infection by Symbiotic Bacteria and Pathogens. *Immunity* **44**, 647–58 (2016).

8. Wilmore, J. R. *et al.* Commensal Microbes Induce Serum IgA Responses that Protect against Polymicrobial Sepsis. *Cell Host Microbe* **0**, 1–10 (2018).

9. Fadlallah, J. *et al.* Synergistic convergence of microbiota-specific systemic IgG and secretory IgA. *J. Allergy Clin. Immunol.* (2018). doi:10.1016/j.jaci.2018.09.036

10. Li, H. *et al.* Mucosal or systemic microbiota exposures shape the B cell repertoire. *Nature* (2020). doi:10.1038/s41586-020-2564-6

11. Sterlin, D., Fadlallah, J., Slack, E. & Gorochov, G. The antibody / microbiota interface in health and disease. *Mucosal Immunol.* 1–9 (2019). doi:10.1038/s41385-019-0192-y

12. Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).

13. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).

14. Lindner, C. *et al.* Diversification of memory B cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nat. Immunol.* **16**, 880–888 (2015).

15. Bashford-Rogers, R. J. M. *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).

16. Meng, W. *et al.* An atlas of B-cell clonal distribution in the human body. *Nat. Biotechnol.* (2017). doi:10.1038/nbt.3942

17. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).

18. Moor, K. *et al.* Analysis of bacterial-surface-specific antibodies in body fluids using bacterial flow cytometry. *Nat. Protoc.* **11**, 1531–1553 (2016).

19. Palm, N. W. *et al.* Immunoglobulin A coating identifies colitogenic bacteria in inflammatory bowel disease. *Cell* **158**, 1000–1010 (2014).

20. Bunker, J. J. *et al.* Innate and Adaptive Humoral Responses Coat Distinct Commensal Bacteria with Immunoglobulin A. *Immunity* **43**, 541–553 (2015).

21. Mohan, D. *et al.* PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. *Nat. Protoc.* **13**, 1958–1978 (2018).

22. Larman, H. B. *et al.* Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.* **29**, 535–41 (2011).

23. Larman, H. B. *et al.* PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. *J. Autoimmun.* **43**, 1–9 (2013).

24. Vazquez, S. E. *et al.* Identification of novel, clinically correlated autoantigens in the monogenic autoimmune syndrome APS1 by proteome-wide PhIP-Seq. *Elife* **9**, 1–25 (2020).

25. Xu, G. J. *et al.* Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).

26. Mina, M. J. *et al.* Measles virus infection diminishes preexisting antibodies that offer protection from other pathogens. *Science* **366**, 599–606 (2019).

27. Shrock, E. *et al.* Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science* **370**, 1–23 (2020).

28. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–94 (2015).

29. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res.* **44**, D694-7 (2016).

30. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–D412 (2015).

31. Lebeer, S. *et al.* Identification of probiotic effector molecules: present state and future perspectives. *Curr. Opin. Biotechnol.* **49**, 217–223 (2018).

32. Bunker, J. J. *et al.* B cell superantigens in the human intestinal microbiota. *Sci. Transl. Med.* **11**, eaau9356 (2019).

33. Ultsch, M., Braisted, A., Maun, H. R. & Eigenbrot, C. 3-2-1: Structural insights from stepwise shrinkage of a three-helix Fc-binding domain to a single helix. *Protein Eng. Des. Sel.* **30**, 619–625 (2017).

34. Korem, T. *et al.* Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses. *Cell Metab.* **25**, 1243-1253.e5 (2017).

35. Mattock, E. & Blocker, A. J. How do the virulence factors of shigella work together to cause disease? *Front. Cell. Infect. Microbiol.* **7**, 1–24 (2017).

36. Klotz, C., Goh, Y. J., O'Flaherty, S. & Barrangou, R. S-layer associated proteins contribute to the adhesive and immunomodulatory properties of Lactobacillus acidophilus NCFM. *BMC Microbiol.* **20**, 248 (2020).

37. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785

38. Landsverk, O. J. B. *et al.* Antibody-secreting plasma cells persist for decades in human intestine. *J. Exp. Med.* **214**, 309–317 (2017).

39. Magri, G. *et al.* Human Secretory IgM Emerges from Plasma Cells Clonally Related to Gut Memory B Cells and Targets Highly Diverse Commensals. *Immunity* 1–17 (2017). doi:10.1016/j.immuni.2017.06.013

40. Chen, K., Magri, G., Grasset, E. K. & Cerutti, A. Rethinking mucosal antibody responses: IgM, IgG and IgD join IgA. *Nat. Rev. Immunol.* (2020). doi:10.1038/s41577-019-0261-1

41. Wilms, E. *et al.* Intestinal barrier function is maintained with aging – a comprehensive study in healthy subjects and irritable bowel syndrome patients. *Sci. Rep.* **10**, 1–10 (2020).

42. Thevaranjan, N. *et al.* Age-Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and Macrophage Dysfunction. *Cell Host Microbe* **21**, 455-466.e4 (2017).

43. Cohen, D. *et al.* Recent trends in the epidemiology of shigellosis in Israel. *Epidemiol. Infect.* **142**, 2583–2594 (2014).

44. McCoy, K. D., Burkhard, R. & Geuking, M. B. The microbiome and immune memory formation. *Immunol. Cell Biol.* **97**, 625–635 (2019).

45. Xu, G. J. *et al.* Systematic autoantigen analysis identifies a distinct subtype of scleroderma with coincident cancer. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7526–E7534 (2016).

46. Paull, M. L. & Daugherty, P. S. Mapping serum antibody repertoires using peptide libraries. *Curr. Opin. Chem. Eng.* **19**, 21–26 (2018).

47. Puga, I. *et al.* B cell-helper neutrophils stimulate the diversification and production of immunoglobulin in the marginal zone of the spleen. *Nat. Immunol.* **13**, 170–180 (2012).

48. Setliff, I. *et al.* High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell* **179**, 1636-1646.e15 (2019).

49. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat. Methods* **11**, 121–2 (2014).

50. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).

51. Wozniak, J. M. *et al.* Mortality Risk Profiling of Staphylococcus aureus Bacteremia by Multi-omic Serum Analysis Reveals Early Predictive and Pathogenic Signatures. *Cell* **182**, 1311-1327.e14 (2020).

52. Forsström, B. *et al.* Dissecting antibodies with regards to linear and conformational epitopes. *PLoS One* **10**, 1–11 (2015).

53. Berglund, L., Andrade, J., Odeberg, J. & Uhlén, M. The epitope space of the human proteome. *Protein Sci.* **17**, 606–13 (2008).

54. Forsström, B. *et al.* Proteome-wide Epitope Mapping of Antibodies Using Ultra-dense Peptide Arrays. *Mol. Cell. Proteomics* **13**, 1585–1597 (2014).

55. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

56. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* 150540 (2018). doi:10.1038/nature25973

57. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nat. 2019* 1 (2019). doi:10.1038/s41586-019-1065-y

58. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–3 (2015).

59. Babu, M. *et al.* Global landscape of cell envelope protein complexes in Escherichia coli. *Nat. Biotechnol.* (2017). doi:10.1038/nbt.4024

60. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).

61. Rothschild, D. *et al.* An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *bioRxiv* (2020). doi:10.1101/2020.05.28.122325