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

Vogl, Thomas; Klompus, Shelley; Leviatan, Sigal; Kalka, Iris N.; Weinberger, Adina; Wijmenga, Cisca; Fu, Jingyuan; Zhernakova, Alexandra; Weersma, Rinse K.; Segal, Eran

Published in:
Nature Medicine

DOI:
10.1038/s41591-021-01409-3

Publication date:
2021

Citation for published version (APA):
Vogl, T., Klompus, S., Leviatan, S., Kalka, I. N., Weinberger, A., Wijmenga, C., Fu, J., Zhernakova, A., Weersma, R. K., & Segal, E. (2021). Population-wide diversity and stability of serum antibody epitope repertoires against human microbiota. *Nature Medicine*, *27*(8), 1442-1450. https://doi.org/10.1038/s41591-021-01409-3

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-021-01409-3

Check for updates

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

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

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 and pathogenic and probiotic bacteria. Leveraging phage immunoprecipitation sequencing (PhIP-Seq) based on phage-displayed synthetic oligo libraries, we detect 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 more than 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.

Humans are covered with approximately the same number of bacteria as body cells[1], and the gut microbiota, in particular, influences 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 immunoglobulin-A (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 describe 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 studies have provided unprecedented insights into the role of antibodies in the adaptive immune system[12,13] relating to gut microbiota[10,14], as well as pointing toward connections of immune-mediated diseases and microbial antigens[15]. While unraveling the clonal diversity[16] and the changes caused by microbiota[10] of the underlying Ig-epitope repertoires, their functional capacity toward antigen recognition in humans has remained largely elusive.

The complexity of human microbiota is 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 enzyme-linked immunosorbent assay (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.

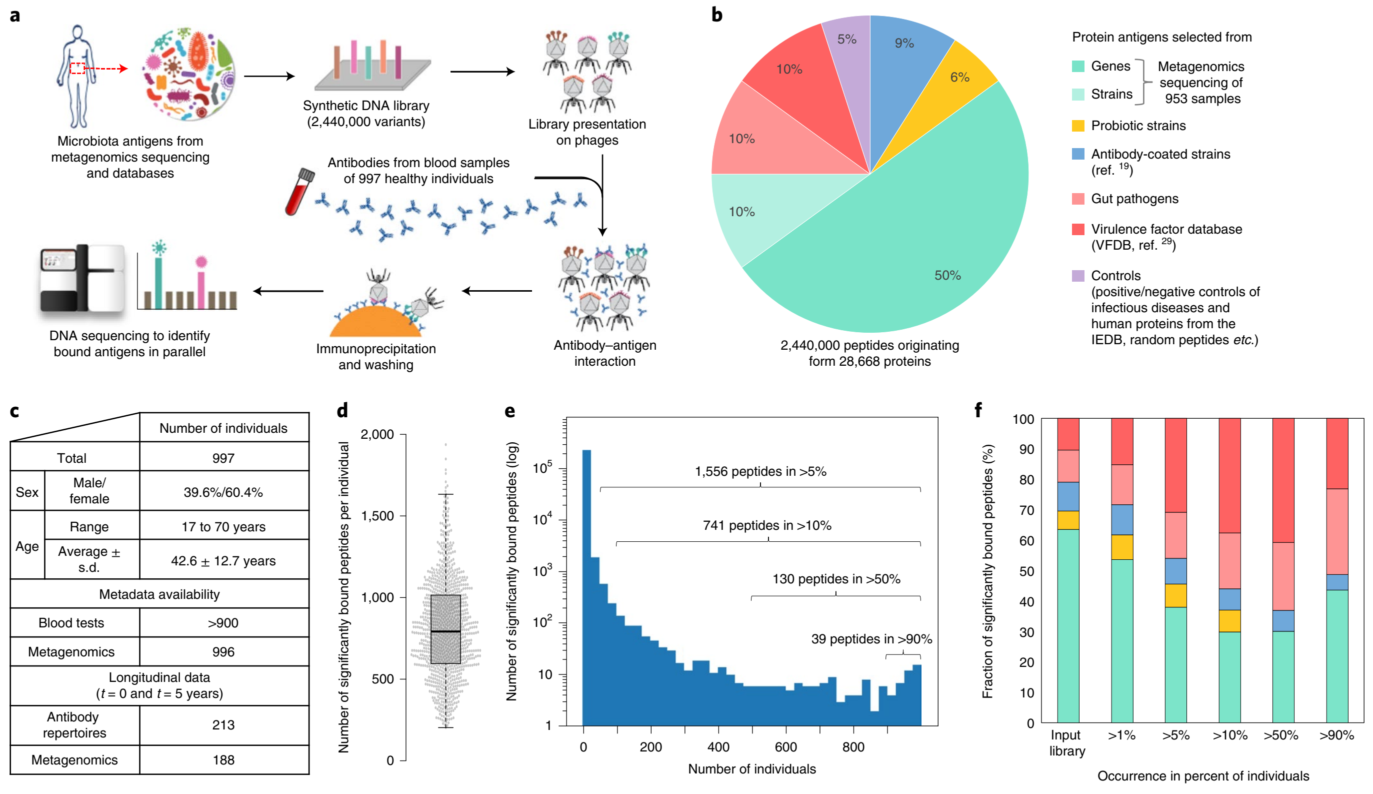
Phage immunoprecipitation sequencing (PhIP-Seq)[21] allows the evaluation of antibody responses to hundreds of thousands of antigens in parallel, as successfully demonstrated primarily with autoimmune diseases[22–24] and viruses[25–27]. As the chemical synthesis of peptide antigens is limited by short 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].

In this article, we have created a PhIP-Seq library representing 244,000 peptide antigens of the microbiota to profile population-wide systemic immunoglobulin epitope repertoires in 997 healthy individuals. We have correlated these antibody profiles with metadata 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 and Methods). Potential

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [4]Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [5]Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [6]These authors contributed equally: Thomas Vogl, Shelley Klompus, Sigal Leviatan. ✉e-mail: adina.weinberger@weizmann.ac.il; eran.segal@weizmann.ac.il



		Number of individuals
Total		997
Sex	Male/ female	39.6%/60.4%
Age	Range	17 to 70 years
	Average ± s.d.	42.6 ± 12.7 years
Metadata availability		
Blood tests		>900
Metagenomics		996
Longitudinal data ($t = 0$ and $t = 5$ years)		
Antibody repertoires		213
Metagenomics		188

Fig. 1 | PhIP-Seq of microbiota directed antibody epitope repertoires. a, The 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 Methods, Supplementary Table 1 and Extended Data Figs. 1–3 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 for details on the blood tests). **d**, On average, ~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 to 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 **e** and **f**). The coloring in **f** is the same as in **b**.

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 of 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 *Escherichia 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 with a series of controls. Technical (Extended Data Fig. 1b,c) and biological (Extended Data Fig. 1d,e) replicates showed high reproducibility (average Pearson $R^2 = 0.96$, Extended Data Fig. 1b) and the employed barcoding strategy also proved reliable, introducing no systematic bias (Supplementary Fig. 1). We observed little to no potential unspecific binding against random peptides or negative controls of human proteins (which 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, as well as little evidence for cross-reactivity (Extended Data Fig. 3 and Supplementary Table 2). Taken together, these results validate the accuracy and reproducibility of our PhIP-Seq assay implementation.

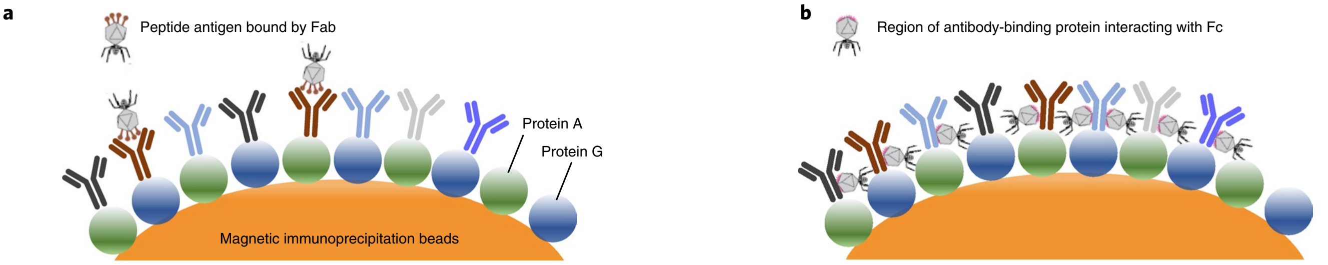
Population-wide antibody profiling. We measured the serum antibody responses of 997 individuals (including the 953 individuals of whom metagenomic data had been employed to select the gut microbiota antigens of the library; Methods)[28]. This healthy cohort spanned a range of 17–70 years of age and had 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 ~800 peptides were significantly enriched (after scoring against input reads; 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, 1,556 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 and Extended Data Fig. 4), including common gut microbiota such as *Bacteroidales* incl. *Prevotella copri* eliciting antibody responses in more than 95% of individuals. Antigens of *Blautia producta* (80%), *Parabacteroides merdae* (75%), *Eubacterium rectale* (60%), *Enterococcus faecalis* (43%), *Lactobacillus plantarum* (41%, a common probiotic) and *Dorea formicigenerans* (35%) were also frequently detected (Supplementary Table 3 provides a detailed list of antigens), supporting that gut microbiota commonly elicits 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 among 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 magnetic beads coated with protein A and G (Extended Data Fig. 5a–c and 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. Among the functional groups of bound peptides, flagella and secreted proteins were significantly over-represented (Supplementary Fig. 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. 2a–c). The binding peptides of protein A cover B-domains known to bind the Fc region of IgG[33] (Fig. 2d). When phages displayed protein A peptides covering a complete B-domain (that is, 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 the 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 the 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 more than 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 and Supplementary Table 5). Although 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 false discovery rate (FDR) correction for ~4.7 million tests; Fig. 3b and Extended Data Fig. 6). Some of the most significant associations include common commensal gut microbiota such as Clostridiaceae but also pathogens (*Staphylococcus* and *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 and 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 (that is, 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 over-represented in elderly individuals (Fig. 4c and 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 in only up to 9% of individuals less than 28 years of age (representing approximately the youngest and oldest deciles of the studied cohort, passing multiple hypothesis testing; raw correlations are shown in Supplementary Fig. 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 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 such 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 *Lactobacillus johnsonii* strains (Fig. 4b,d), suggesting also gender differences in the Ig-epitope repertoires against bacteria. Although 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 (area under the curve (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 toward even broader associations with human health, which 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). By 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 underlies these strong associations. Machine learning predictions from subgroups


a
Peptide antigen bound by Fab
Protein A
Protein G
Magnetic immunoprecipitation beads
b
Region of antibody-binding protein interacting with Fc

c

Protein details				Binding with isotype control antibodies			
Peptide number	Protein origin	Protein length (aa)	Starting position of peptide (aa)	IgG-FC	IgG1-HER2	IgG1-TNFa	All
221096	IgG-binding protein A precursor (*Staphylococcus aureus* subsp. aureus MW2) NP_644899 (WP_000728751)	492	264	3/3	4/4	4/4	11/11
236864			88	3/3	4/4	4/4	11/11
237349			220	2/3	4/4	0/4	6/11
234840			44	0/3	4/4	0/4	4/11
236639			132	2/3	1/4	0/4	3/11
171770	Immunoglobulin-binding protein SBI (*Staphylococcus*) WP_000792567	408	0	3/3	4/4	4/4	11/11
133222	Staphylococcal protein A (*Staphylococcus aureus*) WP_000728715	480	226	3/3	4/4	4/4	11/11
53727			182	3/3	4/4	3/4	10/11
23320			0	3/3	4/4	2/4	9/11
60579			88	1/3	1/4	0/4	2/11
224194	IgG-binding protein SBI (*Staphylococcus aureus* subsp. aureus MW2) NP_647158	437	88	3/3	3/4	1/4	7/11
235563			44	0/3	3/4	1/4	4/11
129495	Hypothetical protein ((*Ruminococcus*) *gnavus*) WP_024853058	330	0	0/3	4/4	0/4	4/11
221185	Type III secretion system ATPase (*Chlamydia trachomatis* D/UW-3/CX) NP_220188	442	378	0/3	0/4	4/4	4/11
199945	*Eubacterium rectale*, flagellar M-ring portein (UniRef derived annotation)	398	0	0/3	3/4	0/4	3/11
46413	Hypothetical protein (*Akkermansia muciniphila*) WP_042447749	59	0	0/3	0/4	3/4	3/11

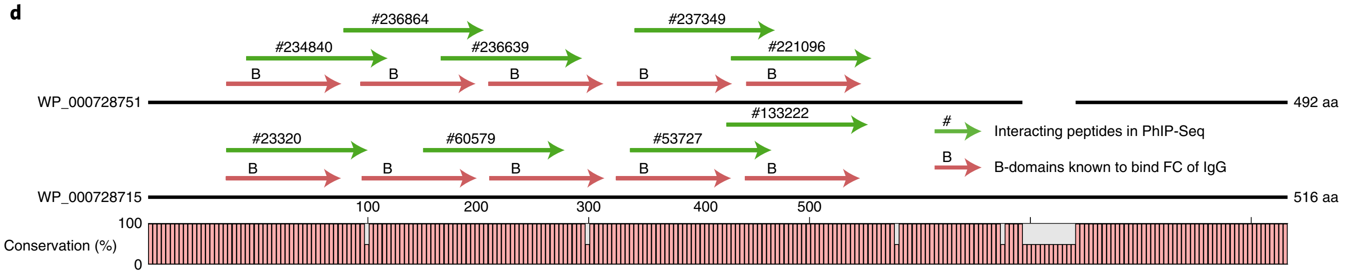


Fig. 2 | Functional antibody-binding proteins within the phage-displayed microbiota antigen library. a,**b**, Canonical binding of phage-displayed antigens with the Fab of antibodies compared to interactions of potential phage-displayed antibody-binding proteins with the Fc part of antibodies. **c**, 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 triplicate (IgG-Fc) or quadruplicate (IgG1-HER and IgG1-TNFa). Significantly bound peptides occurring in at least three reactions overall are listed. aa, amino acid. SBI, second binding protein for immunoglobulins. **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]. The alignment of two highly similar proteins with accession numbers WP_000728751 and WP_000728715 is shown (details on the proteins are provided in **c**). The dark lines to the right of the accession numbers represent the protein sequences, showing gaps in the consensus alignment where applicable. The numbers to the right of the dark lines are the protein lengths in terms of 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 **c** are marked with their identifying number.

of the microbiota library (such as antigens selected from metagenomics data alone and so on; Supplementary Fig. 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 the 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 change) of 0.78 versus 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

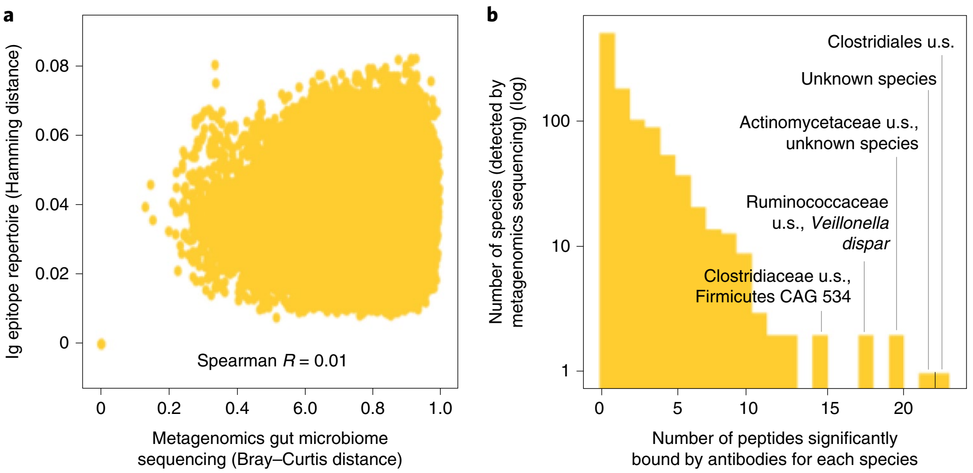


Fig. 3 | Comparison of antibody epitope repertoires against the microbiota library with microbiome composition inferred from metagenomics gut microbiome sequencing. a, The Hamming distances between pairs of serum antibody profiles (each of 996 individuals versus all others) measured with this antigen library do not associate with Bray–Curtis distances computed between the same pairs from metagenomics gut microbiome sequencing data derived from stool samples of the same individuals (Mantel test, $P = 0.514$). The same test was run for other distance measures for antibody profiles (Pearson correlation and Jaccard distance) as well as the presence/absence of bacterial species/antibody responses (alternately to the fold change and abundances shown in **a**), yielding similar results (not shown). **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 versus species abundance only for those appearing in more than 2% of individuals (Extended Data Fig. 6 and Supplementary Table 5). The histogram shows the 1,706 pairs passing FDR correction for multiple hypothesis testing (~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.

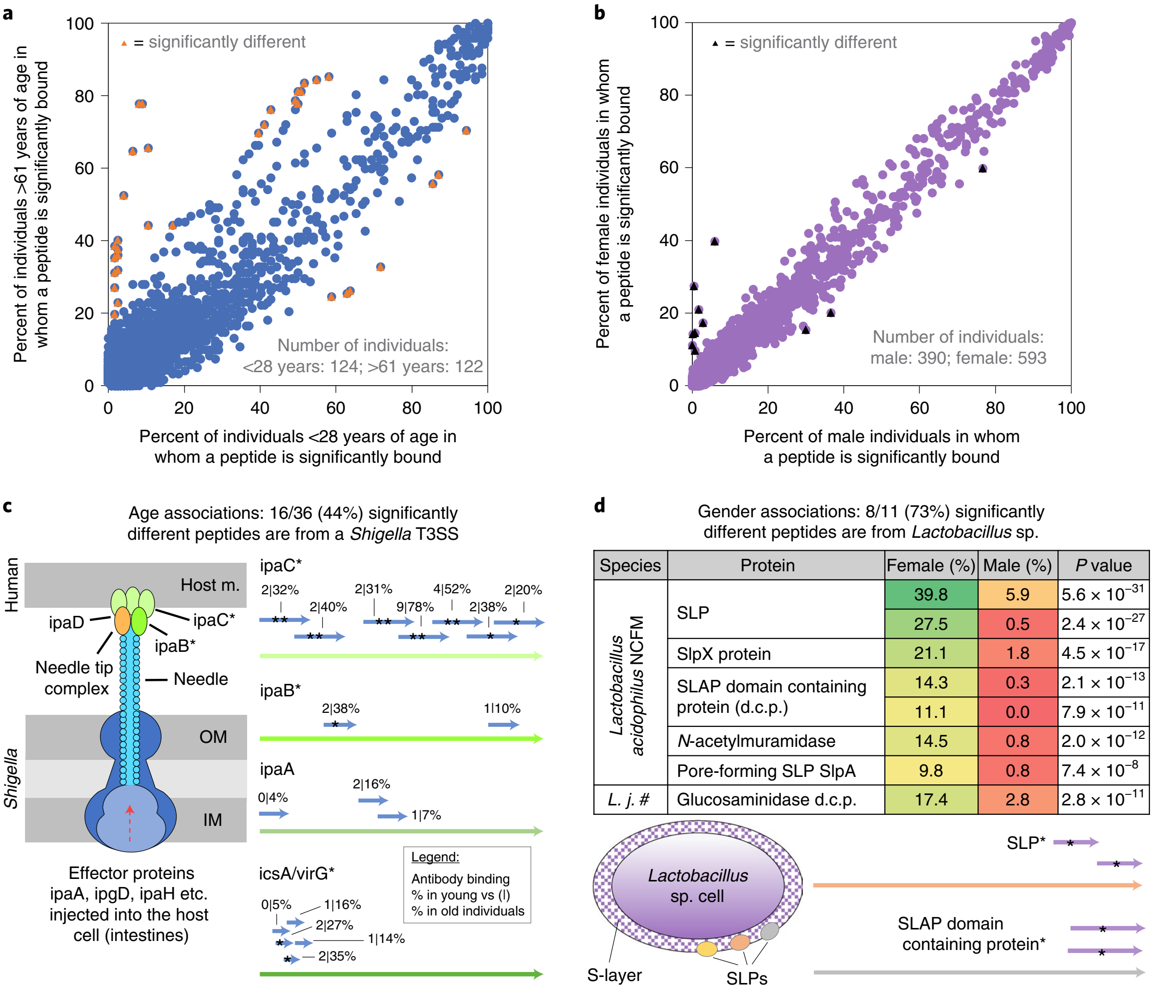
entire microbiota library correlated with an average Pearson correlation coefficient (using log(fold change)) of $R = 0.78$, antigens only from gut microbiota sequencing matched with $R = 0.76$ and antigens of pathogens (VFDB) with $R = 0.83$ (Fig. 6c and Extended Data Fig. 10). However, VFDB antigens 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 five years apart, and 0.19 between two samples of two different individuals. A greedy algorithm could only match 38% of individuals' longitudinal samples (71 of 188) based on metagenomics data based on abundances. As relative abundances may show higher fluctuation than presence/absence of bacterial species, we also evaluated the 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 of 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, whereas previous studies focused on smaller cohorts of dozens of individuals[9,19,20] and did not include 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 ~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 due to eradication of bacterial species whose products reach the bloodstream in parallel by the mucosal immune system. Owing 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 be directly 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).



Species	Protein	Female (%)	Male (%)	*P* value
Lactobacillus acidophilus NCFM	SLP	39.8	5.9	5.6×10^{-31}
		27.5	0.5	2.4×10^{-27}
	SlpX protein	21.1	1.8	4.5×10^{-17}
	SLAP domain containing protein (d.c.p.)	14.3	0.3	2.1×10^{-13}
		11.1	0.0	7.9×10^{-11}
	N-acetylmuramidase	14.5	0.8	2.0×10^{-12}
	Pore-forming SLP SlpA	9.8	0.8	7.4×10^{-8}
L. j. #	Glucosaminidase d.c.p.	17.4	2.8	2.8×10^{-11}

Fig. 4 | Serum antibody epitope repertoires associate with age and gender. a,**b**, Serum antibody epitope repertoires associate with age (**a**) and gender (**b**). Each dot represents a peptide, with its abundance in the respective cohort plotted on the *x* and *y* axes. Cutoffs of <28 and >61 years of age were applied as they represent approximately the youngest and oldest deciles of the studied cohort. Peptides bound significantly differently between groups (chi-square test with one degree of freedom, FDR correction) are highlighted and are listed in Supplementary Table 6. **c**, Many of the peptides marked in **a** that significantly associate with age originate from a *Shigella* T3SS. The illustration on the left side shows how *Shigella* cells can secrete effector proteins through their own inner and outer membranes (IM/OM) into and through the host membrane (Host m.) into the human target cells in the intestines. Illustration adapted from ref. [35], Frontiers Media SA. On the right side, antibody binding against multiple peptides from the same T3SS proteins is illustrated. * indicate 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 probably pass FDR correction with a larger sample number. The percentages of antibody binding in young and old individuals per peptide are indicated above each peptide. Only peptides bound in more than five individuals are shown. See Supplementary Table 6 for details on the peptides. **d**, Most antibody-bound peptides associated with gender (**b**) originate from surface proteins of *Lactobacillus* sp., including S-layer proteins (SLPs)[36]. For these peptides, the frequency of antibody binding in women and men is listed as well as *P* values (chi-square test with one degree of freedom, FDR correction; Supplementary Table 6). NCFM is a strain designation. SLAP, S-layer associated proteins. Overlapping peptides from two proteins significantly differently bound between women and men are shown on the right. See Supplementary Table 6 for details on the peptides.

The 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 over-represented in women. *L. acidophilus* and *L. johnsonii* are common in the intestinal and vaginal microbiome, pointing toward 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 from different mechanisms. Although 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 bloodstream 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 recirculate 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

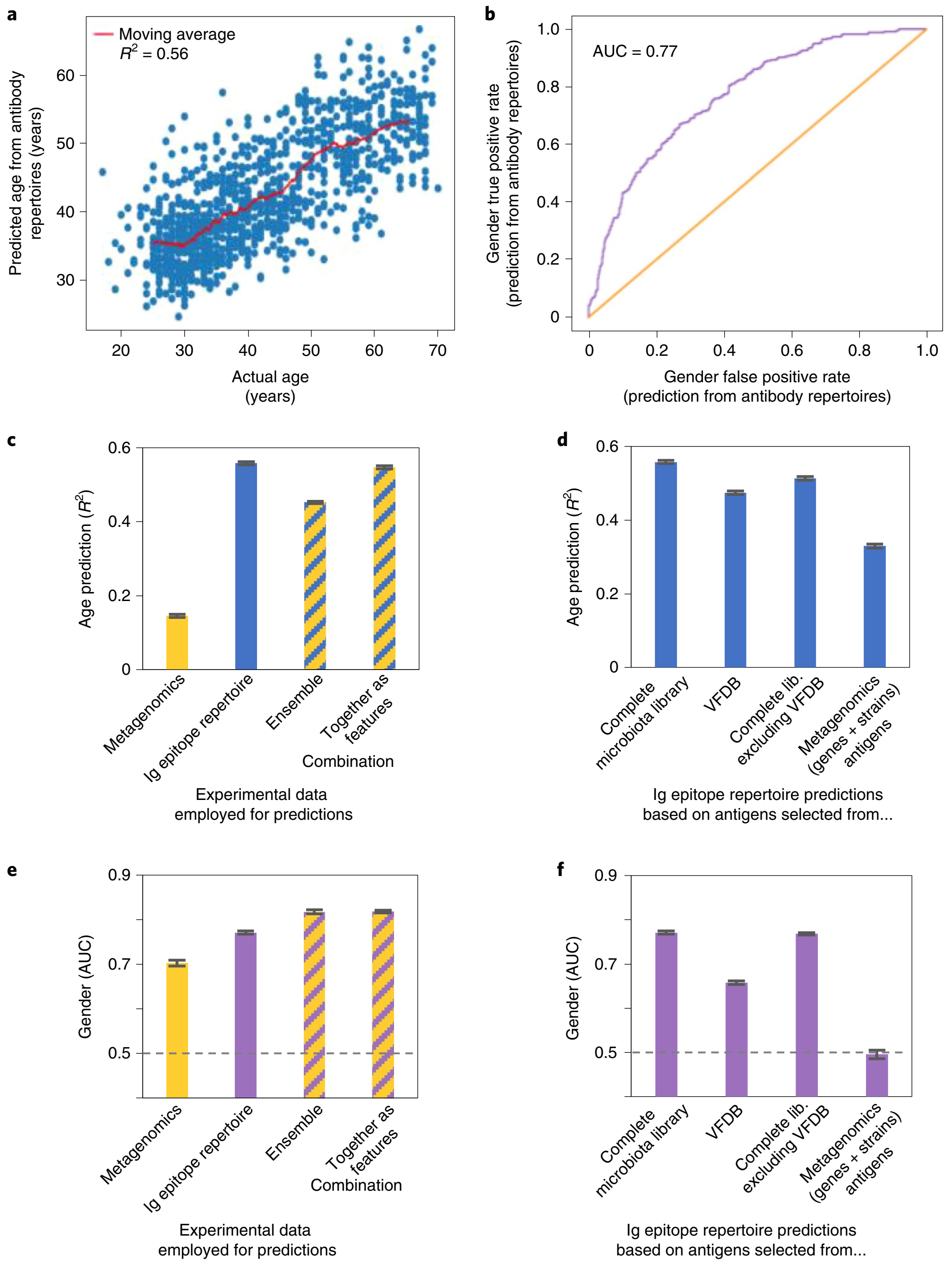


Fig. 5 | Serum antibody epitope repertoires predict age and gender by machine learning better than metagenomics gut microbiome sequencing. **a**,**b**, Machine learning-based predictions of age (**a**) and gender (**b**) from population-wide antibody epitope repertoires were performed with XGBoost with 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. Averages and standard deviations are shown and were 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). See Supplementary Fig. 4 for extended machine learning-based predictions on subgroups of the antigen library. Colors: 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 **d** and **f**: all microbiota, 231,975 peptides; VFDB, 24,164 peptides; library excluding VFDB, 207,811 peptides; metagenomics antigens, 147,061 peptides.

B cells, which 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

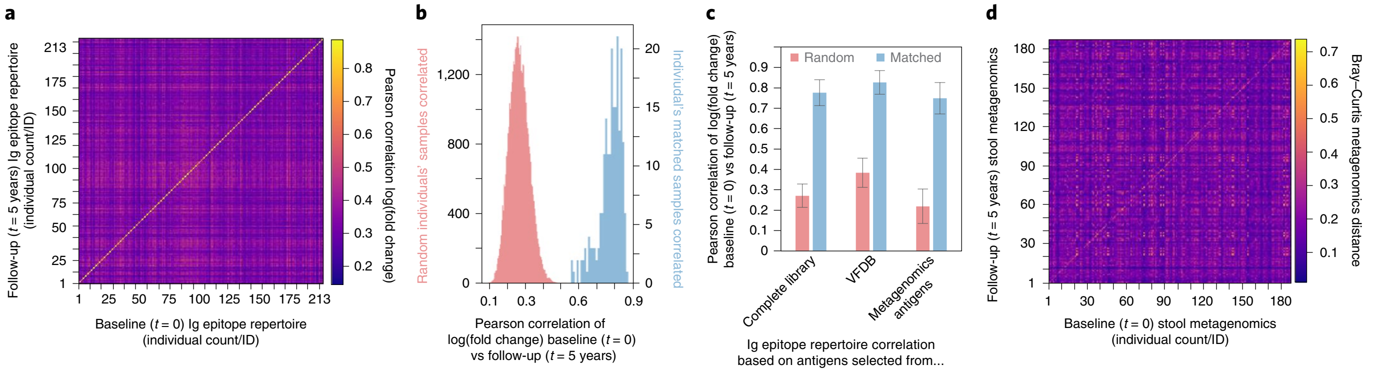


Fig. 6 | The longitudinal stability of antibody epitope repertoires over five years. a, Antibody epitope repertoires of 213 individuals over five years against the entire library of microbiota antigens show high stability. Pearson correlations of log(fold change) of all baseline ($t=0$) and follow-up ($t=5$ years) samples compared with each other are shown. **b**, Correlation coefficients of **a** are shown as a 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. Data are shown as mean values and standard deviations of $n=213$; Extended Data Fig. 10 provides diagrams and correlations of additional antigen subgroups. The sample size is the same as outlined in **b**. Antigen group sizes: 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.

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 (for example 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 the elderly may account for the antibody responses observed.

PhIP-Seq Ig-epitope repertoire data represent a unique layer of information compared to other methods of studying antibody responses against gut microbiota. Fluorescence-activated cell sorting (FACS) and DNA sequencing-based methods to elucidate the 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, as well as the immunological memory of antigens previously encountered. This temporal aspect provides an additional layer of information beyond microbiome DNA sequencing (which is also limited to the detection of bacteria present at the time of sample collection) and could inform on the lasting immune effects of microbiota[44].

Our study is limited by the technical characteristics of PhIP-Seq, as previously discussed in depth[21,22,25,26], most notably the length constraints of the presented peptides (64 amino acids 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 is 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-amino-acid peptides. Previous use of the PhIP-Seq workflow has relied on similar peptide 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 the 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 (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 assess other biological parameters[46]. In our opinion, PhIP-Seq provides a good compromise among peptide length, library size, amenability for parallelized measurements and cost, while allowing for rational selection of the presented peptides (that is, 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 lipopolysaccharides 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. Although our experimental approach informs on the functional antigens recognized by antibodies, linking these to the associated B-cell receptor sequences or demonstrating causality necessitates alternate experimental approaches (for example, ref. [48] or ref. [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 other than 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-021-01409-3.

Received: 6 January 2021; Accepted: 25 May 2021;
Published online: 19 July 2021

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**, 337–340 (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**, 1296–1302 (2016).
7. Zeng, M. Y. et al. Gut microbiota-induced immunoglobulin G controls systemic infection by symbiotic bacteria and pathogens. *Immunity* **44**, 647–658 (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. Synergistic convergence of microbiota-specific systemic IgG and secretory IgA. *J. Allergy Clin. Immunol.* **143**, 1575–1585 (2019).
10. Li, H. et al. Mucosal or systemic microbiota exposures shape the B cell repertoire. *Nature* **584**, 274–278 (2020).
11. Sterlin, D., Fadlallah, J., Slack, E. & Gorochov, G. The antibody/microbiota interface in health and disease. *Mucosal Immunol.* **13**, 3–11 (2020).
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.* **35**, 879–884 (2017).
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 (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**, 535–541 (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**, e55053 (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**, 1079–1094 (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–D697 (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 (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 *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016); https://doi.org/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* **47**, 118–134 (2017).
40. Chen, K., Magri, G., Grasset, E. K. & Cerutti, A. Rethinking mucosal antibody responses: IgM, IgG and IgD join IgA. *Nat. Rev. Immunol.* **20**, 427–441 (2020).
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**, 475 (2020).
42. Thevaranjan, N. et al. Age-associated microbial dysbiosis promotes intestinal permeability, systemic inflammation and macrophage dysfunction. *Cell Host Microbe* **21**, 455–466 (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. USA* **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 (2019).
49. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat. Methods* **11**, 121–122 (2014).
50. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Serum samples, clinical data and metagenomics. A total of 1,051 serum samples of 1,007 individuals were collected in Israel in 2013 and 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[28], with the results of a few tests missing in some individuals. We focused most of the antibody epitope repertoire analysis on baseline samples (the first sample collected per individual). Ten samples did not pass the threshold of >200 peptides significantly bound and were excluded from analyses (section 'Data analysis'), leaving data of 997 individuals for analysis. The 213 longitudinal serum samples and 188 stool samples for metagenomics sequencing were obtained from participants of one of the previous studies[28] after approximately five years, in 2019 and 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 'Content and design of the PhIP-Seq microbiota antigen library' for a detailed description of the content of the antigen phage library. The final list of proteins was cut to peptides of 64 amino acids (aa) with 20-aa overlaps (to cover all possible epitopes of the maximal length of the linear epitope, depending on the definition, between 5 to 9 up to 20 aa[51–53]) between adjacent peptides. The peptide amino-acid sequences were reverse-translated to DNA using *E. 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 also sequenced a section of the 5′ end (to verify matching 5′ and 3′ sequences and exclude the potential presence of multiple inserts). For similar peptide sequences, alternate 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 of 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 removal of 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 was thus 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 a 230-mer pool (library amplification primers: fwd, GATGCGCCGTGGGAATTCT; rev, GTCGGGTGGCAAGCTTTCA) and cloned into T7 phages following the manufacturers recommendations (Merck, T7Select 10-3 cloning kit, product no. 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 (Supplementary Fig. 1a,b, top) and population-wide abundance (Supplementary Fig. 1a,b, bottom). The vast majority of peptides were reproducibly not bound in any individuals (95% of comparisons, Supplementary Fig. 1c). For peptides bound in at least one individual (Supplementary Fig. 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 (three encodings of 347 peptides in 997 individuals representing ~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 two of three encodings in ~30% of individuals, but the third encoding was not detectable at all (Supplementary Fig. 1a, bottom), with direct effects on the observed fold changes (Supplementary Fig. 1a, top).

These results are not solely impacted by the DNA encodings, but also by the different abundances of DNA oligos within the manufacturing process. Care should be applied when comparing different oligos (as the absolute values can be impacted by DNA encodings or oligo manufacturing).

As expected, encoding viral and bacterial peptides with different DNA sequences yielded rather frequent antibody binding (Supplementary Fig. 1a), while peptides originating from 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 the 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 are shown at the top of Supplementary Fig. 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 the potential antigens from human microbiota (for example, the integrated reference catalog of the human gut microbiome (IGC) is composed of 10^7 genes[54]), 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'). Antibody binding of live bacteria is focused on the exposed surface or secreted proteins, so we enriched the library for these protein groups (section 'Selection of microbiota protein targets'). Moreover, because of the 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 *E. coli* codon usage (Methods). 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).

Library content. Bacterial species and databases. About 60% (147,061 oligos) of the library content (Fig. 1b) was dedicated in an unbiased manner to potential antigens from the microbiome of healthy individuals. We used gene and species abundances from the 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 oligos) were dedicated to pathogenic bacteria, probiotic bacteria and gut microbiota previously reported to be coated by antibodies[19]. We also included the entire VFDB[29], making up 10% (24,164 oligos) of the library, and left 5% (11,525 oligos) of the library for various controls (such as infectious disease and autoimmune human proteins from the IEDB[30] and technical controls).

Metagenomics data of the cohort and selection of genes and species (MetaPhlAn2). The metagenomics data from shotgun sequencing of healthy individuals of our cohort were processed in two ways to select antigens. First, mapping to the IGC database and calculation of the relative abundance of each gene was performed as previously described[28,55,56]. The genes data of the PNP cohort contained $\sim 4\times 10^6$ different genes that were mapped to the IGC[54]. Fifty percent of the library content was filled with peptides derived from the proteins encoded by these genes (exact selection criteria are described in the following). 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[57]. We included this strain-based approach to mimic the selection process of pathogenic, probiotic and antibody-coated strains described in the following. After sorting for the 10 most abundant bacterial strains using MetaPhlAn2, 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 (for example, *Campylobacter*, *Shigella* and *Salmonella*) according to a report of the central laboratories of the Israeli Ministry of Health from 2015. In addition, we added *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 and others[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 and patients with Crohn's disease and ulcerative colitis. Bacteria passing a threshold of relative abundance of greater than 10^{-6} and IgA coating index >10 in at least three patients were chosen. In total, nine such bacterial species were selected, five species that were abundantly bound in healthy individuals, two from patients with Crohn's disease and two from patients with ulcerative colitis.

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 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 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 and negative controls. We benchmarked and validated the antibody reactivities against microbiota proteins (described above) with several control antigens. We therefore included 12,025 oligos 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 most comprehensive repository covering various antigens reported in the literature[30]. As positive controls, we selected all antigen epitopes from B-cell assays labeled as infectious diseases (excluding parasites) with a human host. These 290 proteins have been reported in the literature to be targets of antibody responses and were covered with 4,250 oligos.

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 oligos. As well as the exact epitopes reported in the IEDB, the full-length protein sequences (obtained from UniProt by the accession numbers listed in the IEDB) were used and divided into overlapping oligos as described in the following.

In addition to these IEDB positive and negative controls, we included additional control antigens. We added viral proteins that have previously been reported to elicit recurrent antibody responses in 47.9–97.2% of humans using a similar phage display approach (table S2 in ref. [25]). Both full-length proteins divided into overlapping oligos and the exact short peptides reported by Xu et al.[25] were included.

We also 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 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 toward antibody binding, we also included 450 control oligos to assess the 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 (the results are shown and discussed in Extended Data Fig. 2a).

Furthermore, we included codon optimization replicate controls (350 oligos) to test for biases of representing the same amino-acid sequence with different DNA sequences. Oligos 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 are shown in Supplementary Fig. 1). Additionally, 50 oligos representing short peptides (<45 aa) were included to test for additional effects of varying the random sequence at the 3′ end (a detailed explanation is given in the following).

Selection of microbiota protein targets. The pool of microbiota genes derived from metagenomics (approximately 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. We thus 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 cutoff of 10^{-6} relative abundance as the 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 appeared in half of the cohort (moderately abundant) and 3% was dedicated to genes that appeared in less than 1% of the cohort (rarely abundant).

Another set of genes was selected based on annotations and cellular localization predictions focusing on proteins with a 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 ~140,000 genes. We focused on three groups: membrane proteins, secreted proteins and motility proteins/flagella, as these proteins are surface exposed[58] 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[59]. The BLAST step was done locally against the NCBI non-redundant protein database with up to 10 hits per sequence. Analysis of the GO was done locally (database updated to January 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 in a list of ~34,000 membrane proteins, 461 secreted proteins and ~100 motility proteins.

Membrane protein selection. Membrane proteins contain three distinct parts: transmembrane domains, extracellular domains and intracellular domains. We focused on the extracellular domains, as these are more likely to be bound by antibodies, and we 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 ~600 proteins).

Secreted protein selection. In addition to the Blast2GO approach, SignalP 4.0 was used for the prediction of signal peptides. The 140,000 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 signal peptides were filtered out and the mature sequences (without signal peptides) were included in the library (~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, an alignment algorithm that is potentially more than 20,000 times faster than BLASTX, but which maintains 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). 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 of $<10^{-6}$ were filtered out. This approach resulted in an 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, and so on (Supplementary Table 1), to enrich for proteins potentially targeted by antibodies.

Clustering by CDhit. 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: polymerase chain reaction (PCR) plates for the transfer of beads and washing were blocked with 150 µl of BSA (30 g l^{-1} in Dulbecco's phosphate-buffered saline (DPBS) buffer, incubated overnight at 4 °C) and BSA was added to diluted phage/buffer mixtures for immunoprecipitations (IPs) to 2 g l^{-1}. Phage wash buffer for IPs was prepared as outlined in ref. [21] with 0.1% (wt/vol) IPEGAL CA 630 (Sigma-Aldrich cat. no. 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 and 4 µg of antibodies. Although 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 the different binding potential of this novel microbiota antigen library or additional blocking steps performed (we added BSA to the diluted reaction mixtures and also blocked the PCR plate used for the washing steps with BSA). After optimizing the 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 200-mer 100,000 variant pool (S.L., 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. A 1:1 mixture of protein A and G magnetic beads (40 µl; Thermo Fisher Scientific, cat. nos. 10008D and 10009D, washed according to the manufacturer's recommendations) was added after overnight incubation and incubated on a rotator at 4 °C. After 4 h, 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, cat. no. M0493L) according to the manufacturer's 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 (1× QIAquick gel extraction kit, 1× QIAquick PCR purification kit; Qiagen cat. nos. 28704 and 28104) and sequenced on an Illumina NextSeq machine (custom primers for R1: ttactcgagtgcggccgcaagctttca; for R2: tgtgtataagagacagatgctcggggatccgaattc; R1/R2 44/31 nt). Paired-end reads were processed as described in the following.

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 included BoxPlotR/R software[49] and MetaPhlAn2[57]. DNA sequencing (performed on the Illumina NextSeq platform) reads of IPs were downsampled to 1.25 million ID-able reads per sample, that is, 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 ID-ed 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 that is a generalized Poisson distribution. Parameters for this generalized Poisson distribution were estimated for each input level of each sample separately, 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 ($P = 0.05$) for multiple hypothesis testing, and log(fold change) (number of reads of bound peptides versus baseline sequencing of phages not undergoing IPs) was computed for all peptides that passed the threshold P value, and all other peptides were given a log(fold change) value of 0. Samples for which fewer than 200 peptides significantly bound were excluded from analyses. The input sequencing of the phage library was before IP was performed at >100-fold coverage. For the calculation of 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 online at *Nature Medicine*, as well as 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' is the amino-acid sequence of the peptide. The subsequent 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 online at *Nature Medicine*, as well as 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' are the number of peptides significantly bound by antibodies in the PhIP-Seq assay (all peptides from the two mixed libraries versus only peptides from the microbiota library; Methods). When computing age ranges, the main text (for example Fig. 1c) only reports the age range for the 997 baseline samples, which is 17–70 years. For 213 individuals, we collected follow-up samples after ~5 years. Metadata for these follow-up samples are also provided in the cohort_info.csv file. By chance, one of the oldest individuals of the baseline cohort (70 years) was among 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 was inferred from metagenomics sequencing of stool samples of the respective individuals[17,28,60]. 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 online at *Nature Medicine*. The 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' ($-\log_{10}$ of the P value, based on input and output levels) metrics were computed with the generalized Poisson distribution approach as outlined in the Methods. Raw data of the PhIP-Seq experiments are deposited in the Harvard Dataverse public repository: https://doi.org/10.7910/DVN/3SOZCQ.

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

The code repository is subdivided into two subfolders as follows.

Analyse_Fastq. The first subfolder contains 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 (four negative controls, eight mocks and four positive control ('anchor') samples). The output of this is a file, per data well, of fold change and $-\log_{10}$(P value).

Analysis. The second subfolder contains code for executing different tests and analyses on the results of the PhIP-Seq output (as cached from files such as those 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 or goats either with single bacterial proteins or with inactivated bacterial strains (see the first sheet of Supplementary Table 2 for details on the antibodies and immunogens). These samples were processed following our standard PhIP-Seq workflow as 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 of by which antibody classes the antigens of our library are bound, we performed an experiment with altered IP conditions (Fig. 1a). In addition to using a mixture of protein A- and protein G-coated beads (which bind 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, cat. nos. 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. By 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). Sample IgG concentrations (3 µg used per reaction) were determined as outlined in the Methods, and 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 IP step, 25 µl of Pierce streptavidin magnetic beads (Thermo Fisher Scientific) were added per reaction (washed according to the manufacturer's recommendations). The subsequent incubation/washing steps were performed identically 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 same 80 samples were 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) that indicated 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 *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 IP and washing steps (Figs. 2a and 1a). We hypothesized that the frequent binding of these peptides may not be due to interactions of the antibodies' Fab region (Fig. 2a), but rather their nature as antibody-binding proteins and interactions with the Fc part (Fig. 2b).

We thus performed isotype control experiments with commercial antibodies/fragments to probe for interactions of our phage library beyond canonical Fab-dependent binding. We included two IgG monoclonal antibodies (mAbs) with different specificities (IgG1 anti-human HER2, R&D Systems, cat. no. MAB9589; IgG1 anti-human tumor necrosis factor-α, R&D Systems, cat. no. MAB9677). Additionally, an Fc preparation of IgG from human blood was included (Novus, cat. no. NBP2-47132). The two IgG mAbs could allow detection of the 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 six peptides included within our PhIP-Seq library for analysis in a peptide ELISA (results are 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 control three times within the library, with neither DNA encoding of the peptide eliciting binding; see Supplementary Fig. 1b for details)).

We also included peptides of two *Shigella* proteins, ipaC and icsA/virG, associated with age (Fig. 4c and 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 the 64-aa peptides displayed on the phages is costly, we aimed to reduce the peptide length. We thus 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 and Supplementary Table 6 for *Shigella* ipaC peptides #226014 and #232269. Likewise, 20 aa from the overlap of peptides of icsA (#221918 and #235092), *Staphylococcus* Emp (#180309 and #24623) and *Streptococcus* 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 amino-acid sequences were selected: EBV, PPPGRRPFFHPVAEADYFEY; SAPK4, KIMGMEFSEEKIQYLVYQML; *Shig.* ipaC, GKNPVLTTTLNDDQLLKLSE; *Shig.* icsA/virG, NGGDSITGSDLSIINQGMIL; *Staph.* Emp, ASEDKLNKIADPSAASKIVD; *Strep.* CHAP, SATSYINTILNSKSVSDAIN.

These amino-acid sequences were ordered from JPT Peptide Technologies 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 a horseradish peroxidase-conjugated anti-human IgG antibody (Southern Biotech, cat. no. 204205) and 3,3′,5,5′-tetramethylbenzidine (TMB) as substrate. Sera of 80 individuals (for whom PhIP-Seq data were also available and the protein A/G and IgG/IgA experiments had been performed; Extended Data Fig. 5) were tested with each of the six peptides.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data generated or analyzed during this study are included within the paper, its Supplementary Information files and public repositories. Detailed information on the cohort, library content and PhIP-Seq data are available in the Supplementary Data files (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. 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, respectively, in Supplementary Table 2, Supplementary Table 3, Supplementary Table 4, Supplementary Table 5 and Supplementary Table 6. Raw data for the PhIP-Seq experiments are deposited in the Harvard Dataverse public repository at 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

Custom code used for analyzing the PhIP-Seq data is publicly available at https://github.com/erans99/PhIPSeq_external. The code repository is subdivided into two subfolders: (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 (four negative controls, eight mocks and four positive control ('anchor') samples). The output of this is a file, per data well, of fold change and $-\log_{10}(P \text{ value})$; (2) Analysis, code for executing different tests and analyses on the results of the PhIP-Seq output (as cached from files like those in the PhIPSeq_data directory).

References

51. Forsström, B. et al. Dissecting antibodies with regards to linear and conformational epitopes. *PLoS ONE* **10**, e0121673 (2015).
52. Berglund, L., Andrade, J., Odeberg, J. & Uhlén, M. The epitope space of the human proteome. *Protein Sci.* **17**, 606–613 (2008).
53. Forsström, B. et al. Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays. *Mol. Cell. Proteom.* **13**, 1585–1597 (2014).
54. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
55. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
56. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
57. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
58. Babu, M. et al. Global landscape of cell envelope protein complexes in *Escherichia coli*. *Nat. Biotechnol.* **36**, 103–112 (2018).
59. Götz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
60. Rothschild, D. et al. An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.28.122325 (2020).
61. 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 (2020).

Acknowledgements

A.W. is the Louis H. Sackin Research Fellow Chair in Computer Science. E.S. is supported by grants from the European Research Council, the Israel Science Foundation and by the Seerave Foundation. T.V. gratefully acknowledges support from the Austrian Science Fund (FWF, Erwin Schrödinger fellowship J4256). R.K.W. is supported by the Seerave Foundation and the Netherlands Organization for Scientific Research. A.Z. is supported by ERC Starting Grant 715772, the 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), 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

T.V. and S.K. conceived the project and designed the library. S.L. designed and implemented the coding of the library. T.V. and S.K. planned and calibrated the experimental system and performed biological experiments. S.L. designed and implemented the computational pipeline. S.L. and 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. and E.S. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

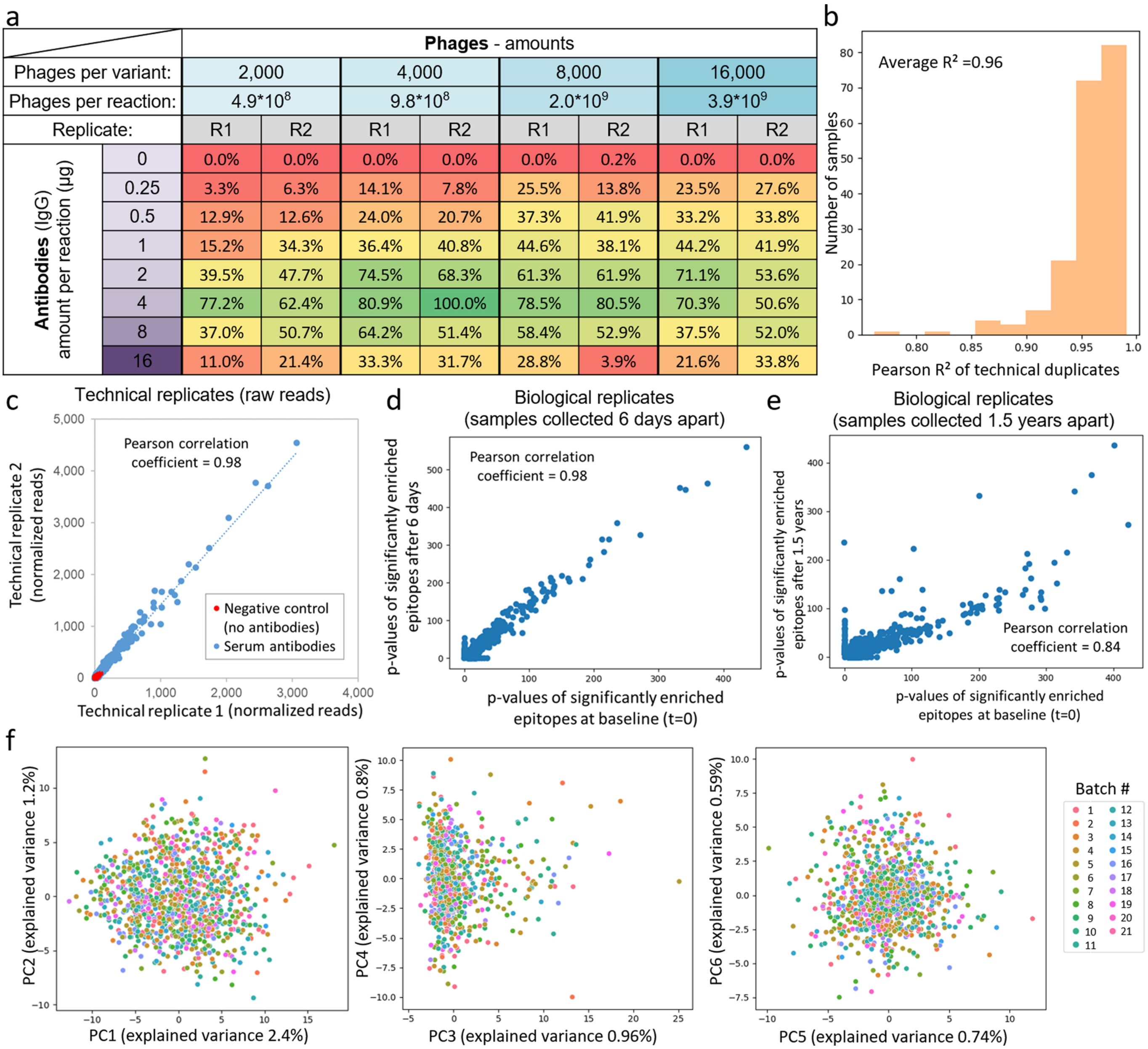
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-021-01409-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-021-01409-3.

Correspondence and requests for materials should be addressed to A.W. or E.S.

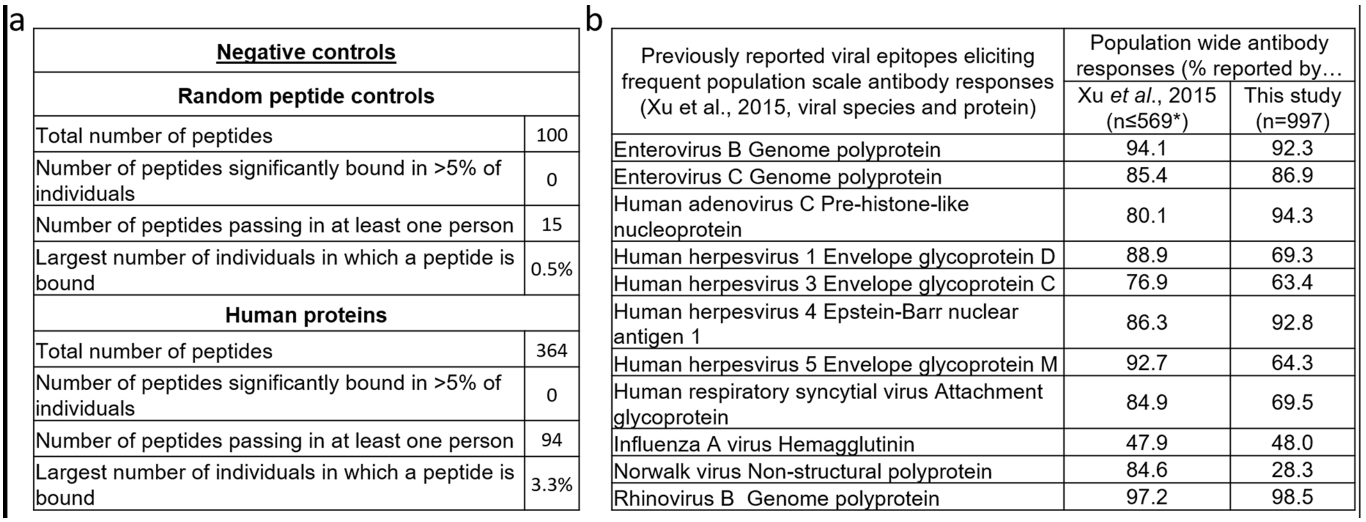
Peer review information *Nature Medicine* thanks Rachael Bashford-Rogers, George Georgiou, Andrea Cerutti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Saheli Sadanand was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



		Phages - amounts							
Phages per variant:		2,000		4,000		8,000		16,000	
Phages per reaction:		$4.9*10^8$		$9.8*10^8$		$2.0*10^9$		$3.9*10^9$	
Replicate:		R1	R2	R1	R2	R1	R2	R1	R2
Antibodies (IgG) amount per reaction (μg)	0	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%
	0.25	3.3%	6.3%	14.1%	7.8%	25.5%	13.8%	23.5%	27.6%
	0.5	12.9%	12.6%	24.0%	20.7%	37.3%	41.9%	33.2%	33.8%
	1	15.2%	34.3%	36.4%	40.8%	44.6%	38.1%	44.2%	41.9%
	2	39.5%	47.7%	74.5%	68.3%	61.3%	61.9%	71.1%	53.6%
	4	77.2%	62.4%	80.9%	100.0%	78.5%	80.5%	70.3%	50.6%
	8	37.0%	50.7%	64.2%	51.4%	58.4%	52.9%	37.5%	52.0%
	16	11.0%	21.4%	33.3%	31.7%	28.8%	3.9%	21.6%	33.8%

a
b
Average R² =0.96
Number of samples
0
10
20
30
40
50
60
70
80
0.80
0.85
0.90
0.95
1.0
Pearson R² of technical duplicates
c
Technical replicates (raw reads)
Pearson correlation coefficient = 0.98
Negative control (no antibodies)
Serum antibodies
Technical replicate 2 (normalized reads)
0
1,000
2,000
3,000
4,000
5,000
Technical replicate 1 (normalized reads)
d
Biological replicates (samples collected 6 days apart)
Pearson correlation coefficient = 0.98
p-values of significantly enriched epitopes after 6 days
0
100
200
300
400
500
p-values of significantly enriched epitopes at baseline (t=0)
e
Biological replicates (samples collected 1.5 years apart)
Pearson correlation coefficient = 0.84
p-values of significantly enriched epitopes after 1.5 years
0
100
200
300
400
p-values of significantly enriched epitopes at baseline (t=0)
f
PC2 (explained variance 1.2%)
PC1 (explained variance 2.4%)
−10
−5
0
5
10
15
PC4 (explained variance 0.8%)
PC3 (explained variance 0.96%)
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
20
25
PC6 (explained variance 0.59%)
PC5 (explained variance 0.74%)
Batch #
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

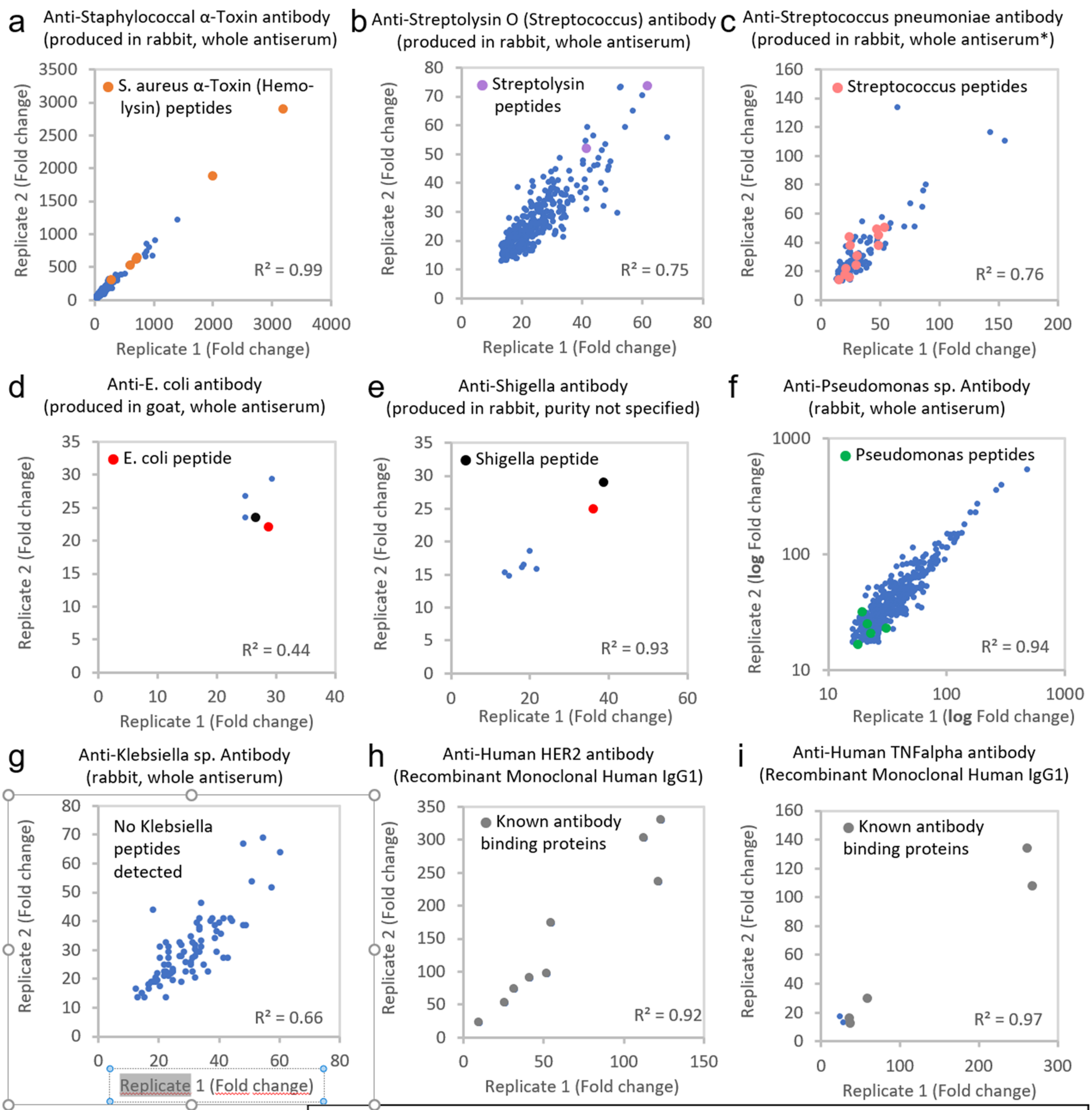
Extended Data Fig. 1 | Control experiments of optimizing the ratio of phage/antibody amounts in IPs (a), the reproducibility of technical duplicates (b), examples of high technical (c) as well as biological reproducibility (d,e), and the > 1,000 samples reported in the main manuscript were processed in batches of 96 well plates, that were not biased by batch effects (f). a, 'Phages per variant' refers to the number of phages per library variant, 'Phages per reaction' refers to the number of total phages in a reaction mixture of the microbiota library (244,000 variants times the number of phages per variant). IP reactions were performed in duplicates (R1, R2), the numbers of significantly bound peptides are shown normalized as percent of the highest binding phage/antibody combination (4,000-fold phage coverage and 4 μg of IgG antibodies). A mixed pool of human serum samples was used as antibody material for this calibration. **b**, Technical replicates (n = 191 samples measured in duplicates) were in excellent agreement with an average Pearson R^2 (of FCs) of 0.96 between duplicates. 95% of duplicates correlated with R^2 greater than 0.90 (181/191) and 78% of duplicates even with an R^2 greater than 0.95 (149/191). Given this high reproducibility and little added information gained from duplicates, the exploratory experiments reported in this manuscript were carried out in single reactions. For potential diagnostic applications of PhIP-Seq technical replicates may be valuable to validate results. **c**-**e**, Examples of high technical reproducibility and low background binding (**e**) as well as biological reproducibility of samples collected 6 days (**f**) and 1.5 years apart (**g**). In red in panel e low background binding of a negative control without antibodies ('Mock IP'[21]) is illustrated. Samples collected days (**f**) or years (**g**) apart and processed in different PhIP-Seq runs show excellent reproducibility. **f**, Principle component analysis (PCA) of samples measured in different batches of PhIP-Seq experiments. PCs were computed on signals (log FC) against bound peptides of the entire antigen, the first six PCs are shown. Samples measured in the same batch do not cluster separately from other batches indicating no clear bias of batch effects for these samples.



Negative controls	
Random peptide controls	
Total number of peptides	100
Number of peptides significantly bound in >5% of individuals	0
Number of peptides passing in at least one person	15
Largest number of individuals in which a peptide is bound	0.5%
Human proteins	
Total number of peptides	364
Number of peptides significantly bound in >5% of individuals	0
Number of peptides passing in at least one person	94
Largest number of individuals in which a peptide is bound	3.3%

Previously reported viral epitopes eliciting frequent population scale antibody responses (Xu et al., 2015, viral species and protein)	Population wide antibody responses (% reported by…	
	Xu *et al*., 2015 (n≤569*)	This study (n=997)
Enterovirus B Genome polyprotein	94.1	92.3
Enterovirus C Genome polyprotein	85.4	86.9
Human adenovirus C Pre-histone-like nucleoprotein	80.1	94.3
Human herpesvirus 1 Envelope glycoprotein D	88.9	69.3
Human herpesvirus 3 Envelope glycoprotein C	76.9	63.4
Human herpesvirus 4 Epstein-Barr nuclear antigen 1	86.3	92.8
Human herpesvirus 5 Envelope glycoprotein M	92.7	64.3
Human respiratory syncytial virus Attachment glycoprotein	84.9	69.5
Influenza A virus Hemagglutinin	47.9	48.0
Norwalk virus Non-structural polyprotein	84.6	28.3
Rhinovirus B Genome polyprotein	97.2	98.5

Extended Data Fig. 2 | Analysis of negative controls for estimating nonspecific background signal (a) and comparison of viral positive controls of this study to population wide responses previously reported[25] (b). a, Negative controls indicate little nonspecific background-binding impacting population-scale interpretation of the measured antibody epitope repertoires. We had included negative controls of proteins that were expected to elicit little binding in healthy individuals. These included random amino acid sequences (100 peptides), as well as human proteins (autoimmune disease targets reported in the IEDB[30] and various abundant housekeeping genes such as histones and glycolytic enzymes represented as 364 peptides). Analyzing binding to these negative controls in the cohort showed that a few random peptides were significantly enriched in up to 0.5% (5/997 individuals), indicating a low background of unspecific binding (or cross-reactivity) which can be eliminated by using a threshold for peptides bound in >1% of individuals. Peptides of human proteins were bound in up to 3.3% (33/997) of individuals, similar to results previously reported using PhIP-Seq[23]. It has been speculated that such antibody binding against human proteins may arise from cross-reactivity and are unlikely to have detrimental consequences in healthy individuals[23]. The following machine learning based predictions in this work were limited to peptides bound in at least 2% of the population. **b**, Controls of viral epitopes measured on our cohort match previously reported seroprevalences from a population scale study[25]. Xu *et al*[25]. employed a PhIP-Seq workflow with a library covering viral antigens ('VirScan') and detected near universal population wide targeting of certain viral peptides. They had reported a list of 11 viral peptides including the amino acid sequences and seroprevalences (supporting information, Table S2 of their publication[25], % seroprevalences are reproduced from this table). We had included the exact same peptides and detected similar rates of seroprevalence, demonstrating the reproducibility of the PhIP-Seq workflow and sensitivity of our implementation. *: Xu *et al*. have analyzed in total sera of 569 individuals, although the exact number of individuals for calculating the seroprevalence is not specified in the caption of their supporting table S2.



j

		Antibody tested in PhIP-Seq								
		Anti-Staph. α-Toxin antibody	Anti-Streptolysin O antibody	Anti-S. pneumoniae antibody	Anti-E. coli antibody	Anti-Shigella antibody	Anti-Pseudomonas sp. antibody	Anti-Klebsiella sp. antibody	Anti-Human HER2 antibody	Anti-Human TNFalpha antibody
	Number of antigen specific peptides bound by the antibody itself	**6**	**2**	**12**	**1**	**1**	**5**	**0**	**0(9)****	**0(5)*****
Number of specific peptides bound by other antibodies	Anti-Staph. α-Toxin antibody	**6**	0	0	0	0	0	0	7****	4****
	Anti-Streptolysin O antibody	0	**2**	0	0	0	0	0	4****	3****
	Anti-S. pneumoniae antibody	0	0	**12**	0	0	0	0	2****	2****
	Anti-E. coli antibody	0	0	0	**1**	1*	0	0	0****	0****
	Anti-Shigella antibody	0	0	0	1*	**1**	0	0	1****	1****
	Anti-Pseudomonas sp. antibody	0	0	0	0	0	**5**	0	0****	0****
	Anti-Klebsiella sp. antibody	0	0	0	0	0	0	**0**	0****	0****
	Anti-Human HER2 antibody	0	0	0	0	0	0	0	**0(9)****	5****
	Anti-Human TNFalpha antibody	0	0	0	0	0	0	0	5****	**0(5)*****

* potential cross-reactivity between highly similar peptides, see discussion in the text

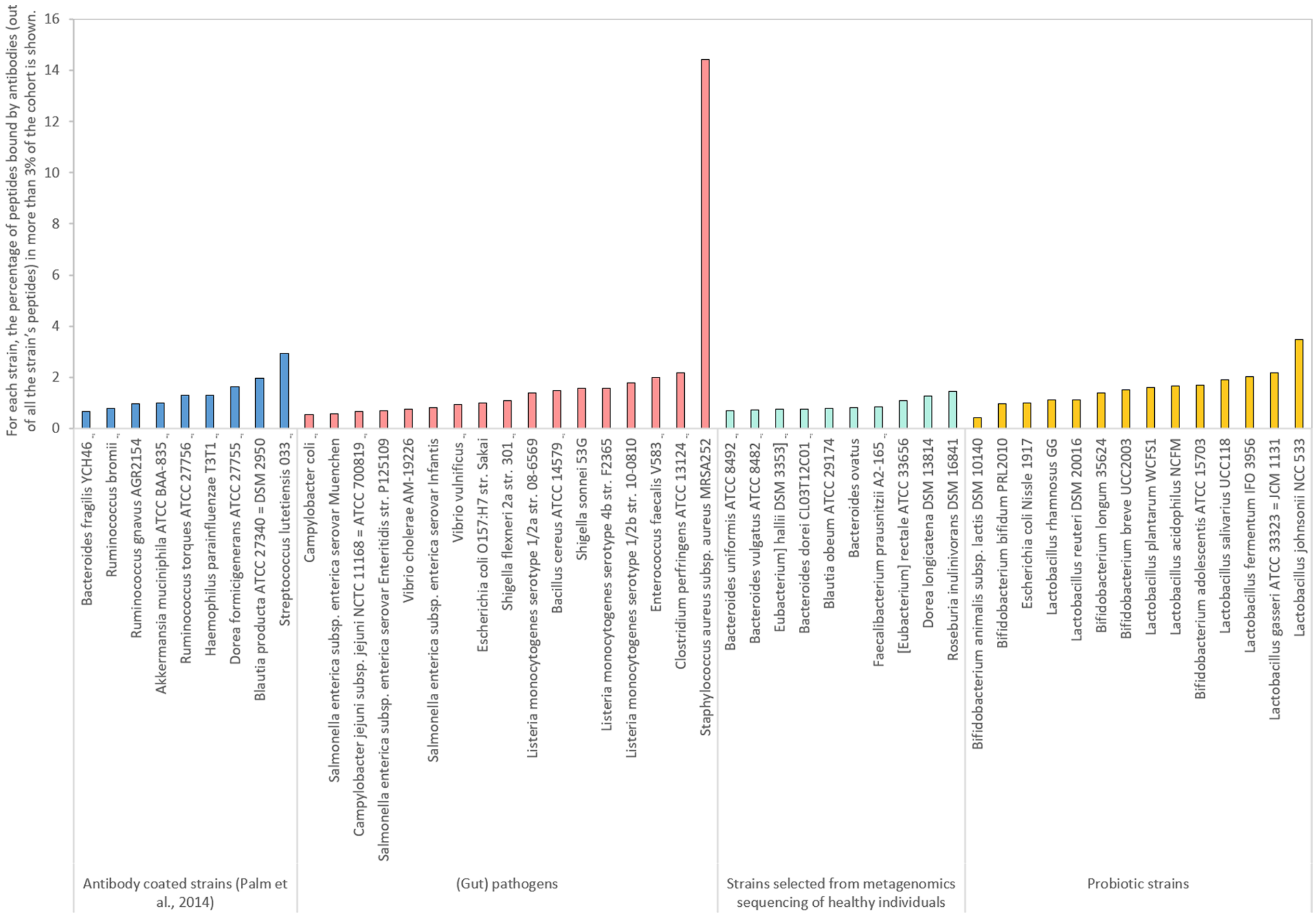
** HER2 is not part of the antigen library, but 9/9 bound peptides are known antibody binding proteins

*** TNFalpha is not part of the antigen library, but 5/7 bound peptides are known antibody binding proteins

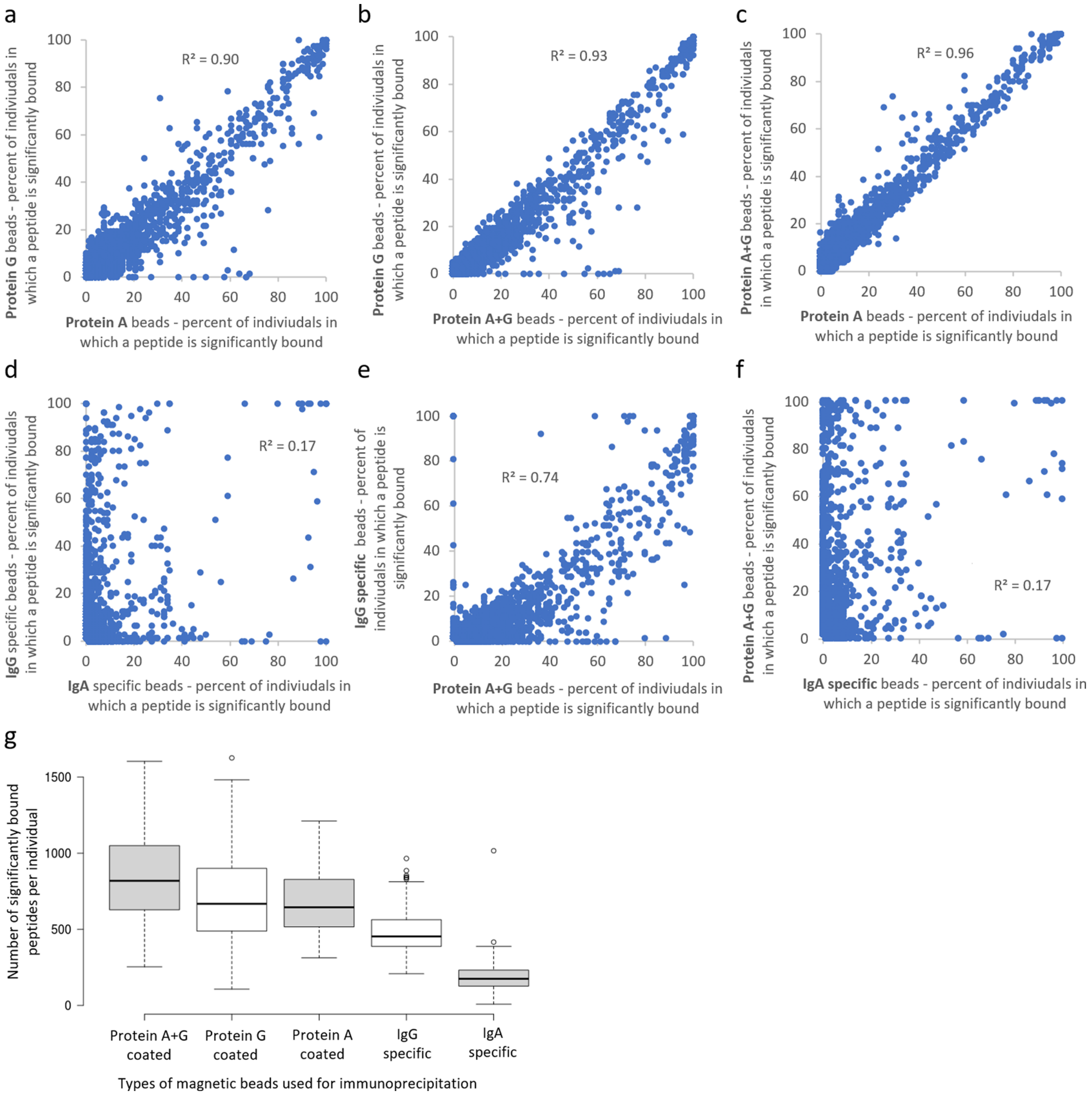
**** The appearance of peptides originating from antibody binding proteins indentified by Anti-HER2 or Anti-TNFalpha is counted.

Extended Data Fig. 3 | See next page for caption.

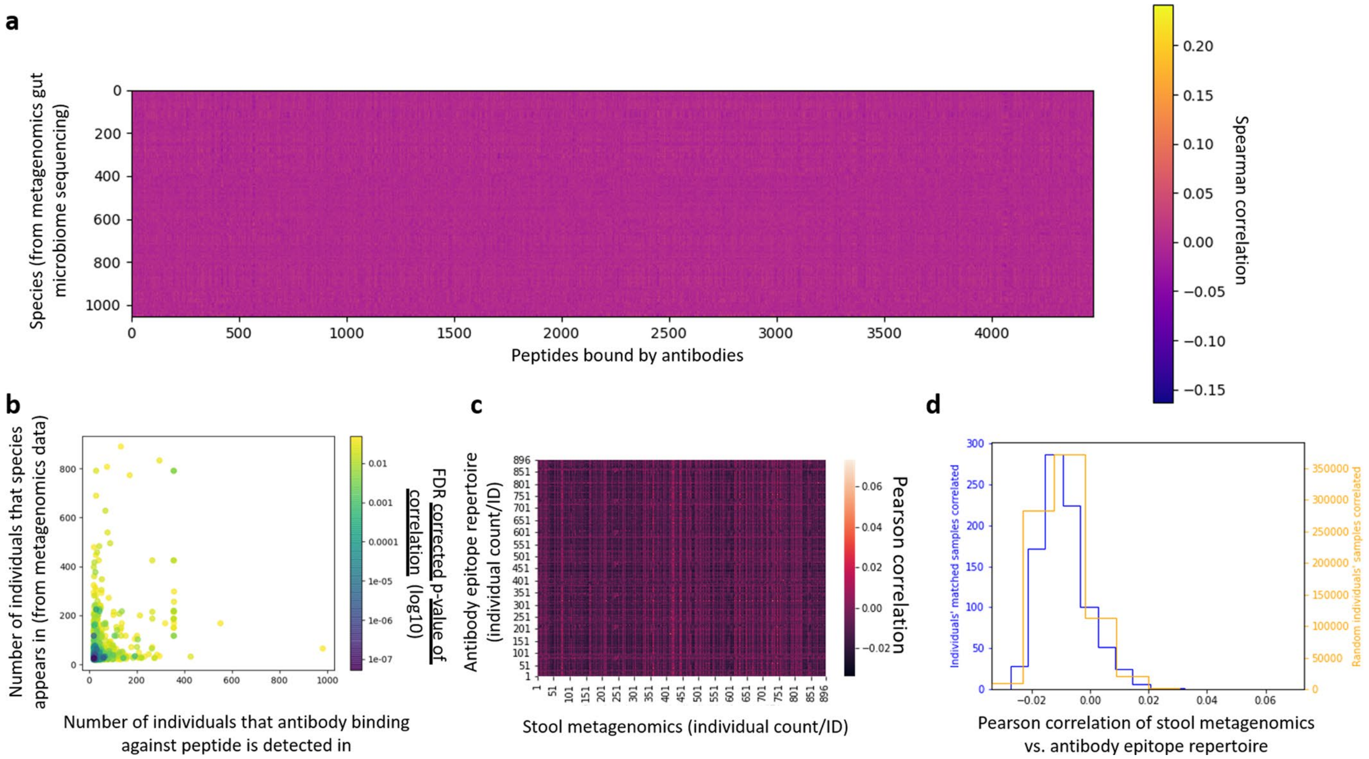
Extended Data Fig. 3 | The PhIP-Seq workflow robustly identifies peptide targets of antibodies generated against immunogens of full-length proteins and bacterial cells. a-i, Commercially available antibodies were measured with our PhIP-Seq microbiota library following the same standard approach applied to human serum samples. Antibody amounts were normalized by the concentrations specified by the manufacturers. Panels a-g represent antibody preparations targeting microbial antigens, panels h and i represent negative controls of monoclonal antibodies targeting human proteins. See the first sheet of supporting file.xlsx file Supplementary table 2 for details on immunogens and properties related to each antibody. Measurements of each antibody were performed in triplicates and peptides appearing in all replicates were used in the analysis. The correlation of Fold change values for two random replicates [Rep. 1, Rep. 2] are shown (Pearson R^2). 'Fold change' refers to the ratio between reads in the IP reaction with antibodies vs. input sequencing of the phage library (a proxy for binding strength). Note the different scales on the axes and note the use of a logarithmic scale for the axes of panel f (to adequately represent weakly enriched peptides). '*' in panel c denotes an antibody preparation, that was protein A purified according to the manufacturer. The exact bound peptides are listed in the second sheet of supporting file.xlsx file Supplementary table 2. Only peptides related to the bound antigens are listed (background reactivity of the whole sera from rabbit/goat omitted). **j**, Assessment of potential cross-reactivity or background reactivity of the antibody preparations. The list of bound peptides by each antibody preparation (marked in panels a-i) was searched among the bound peptides by every other antibody preparation (and is marked in all plots, only nearly identical E. coli and Shigella peptides show up in the other sample as well). The numbers of bound peptides are listed. Thereby we have verified that the marked peptides in panels a-i are not appearing due to background/ cross-reactivity of the whole animal sera, as they only appear in reactions of the respective antibodies.



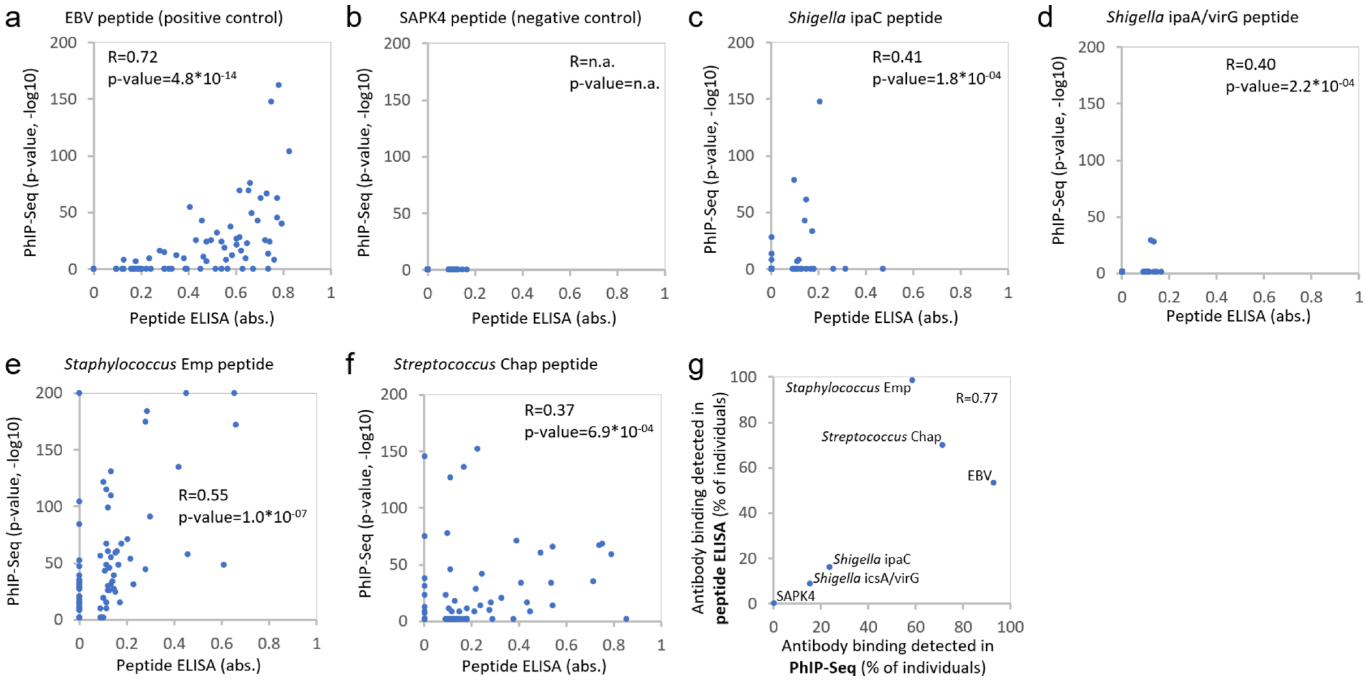
Extended Data Fig. 4 | Bacterial strains of different functional groups within the library (Fig. 1b, methods section and Supplementary table 1) all elicit substantial population wide antibody responses. The fraction of peptides per strain (out of all the strain's peptides) bound in >3% of the cohort (n = 997) is shown. See Supplementary table 1 for details on the bacterial strains listed. Antibody responses are not limited to pathogenic strains, but extend to strains selected from healthy individuals' gut microbiota (from metagenomics sequencing, see the methods section), probiotic strains, and strains previously reported to be coated by antibodies[19]. A large fraction of Staphylococcus aureus peptides were bound, possibly owing to its ubiquitous role in the upper respiratory tract and human skin microbiome along its large number of virulence factors potentially eliciting antibody responses[29,61].



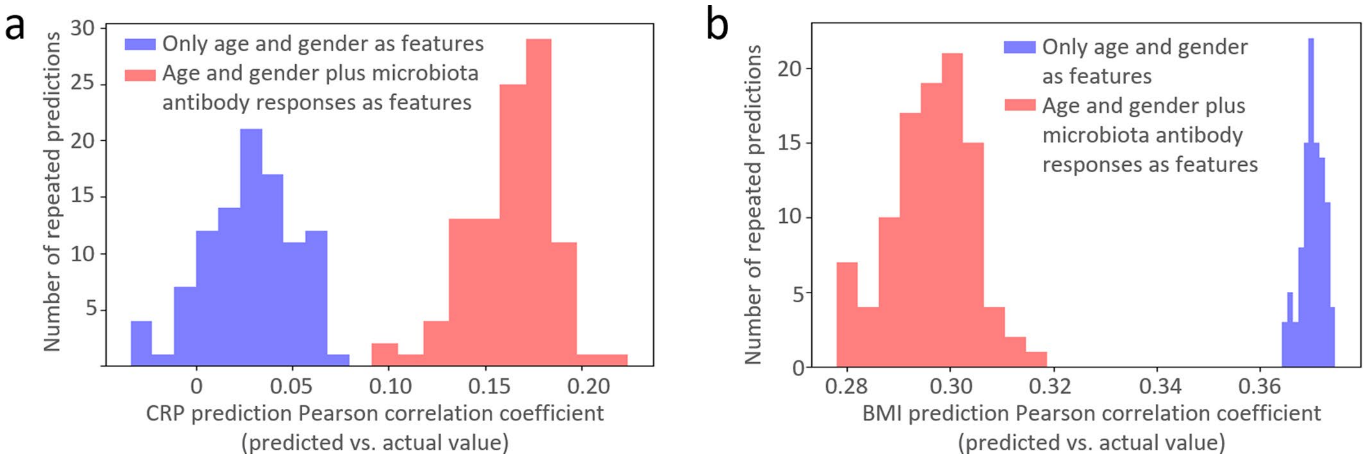
Extended Data Fig. 5 | Antibody binding with protein A and protein G coated beads separately (a-c) and antibody coated beads capturing IgA and IgG separately (d). Supplementary table 4 provides detailed lists on the respective peptides bound. **a-c** Relying on different binding affinities of protein A and G for antibody classes, we processed 80 serum samples each with 1.) a mixture of protein A and G, 2.) protein A alone, and 3.) protein G alone. **a**, Comparison of peptides bound by protein A vs. protein G. **b**, Comparison of peptides bound by a mixture of protein A and G vs. protein G. **c**, Comparison of peptides bound by a mixture of protein A and G vs. protein A. In panels a-c data of 78/80 samples are shown, as samples with <200 significantly enriched peptides per sample were excluded (same cutoff as for the other human sera measured). **d** Experimental workflow to detect IgA and IgG subclasses separately (following procedures reported in the literature[27] and Methods). In panel d, a comparison of peptides bound by IgA vs. IgG specific beads is shown. Samples with IgG specific beads were sequences with 0.8 million reads, however we do not expect a strong impact thereof, as the number of detected peptides typically saturates[22]. **e** Comparison of peptides bound by a mixture of protein A and G vs. IgG specific beads. **f** Comparison of peptides bound by a mixture of protein A and G vs. IgA specific beads. In panels d-f data of 80 samples is shown (as for IgA many samples would not have passed the threshold of >200 peptides applied in other figures, see panel g). For the IgA vs. IgG experiments a different batch of phages was used. In panels a-f Pearson R² is shown. **g** Number of bound peptides per sample with each set of magnetic beads used. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. n = 80 sample points.



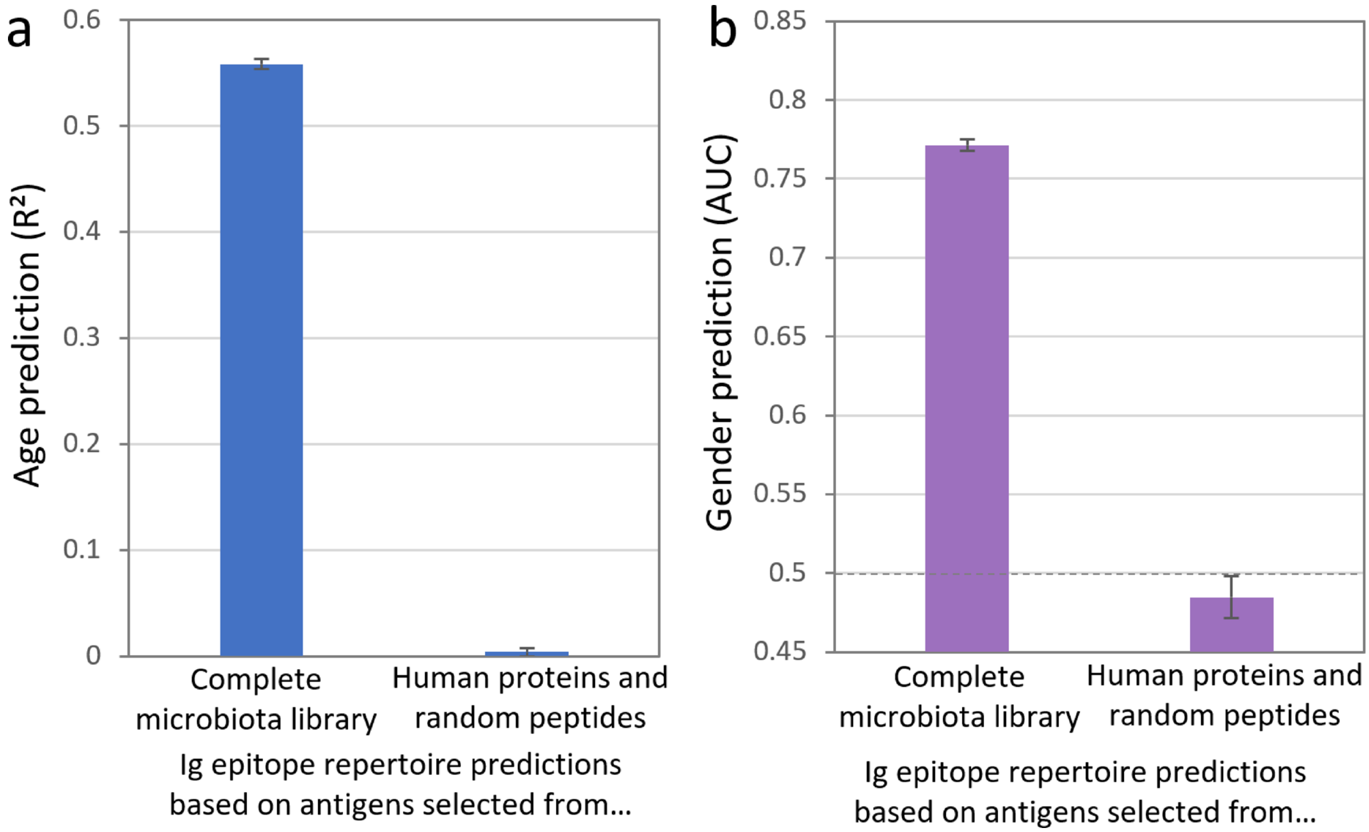
Extended Data Fig. 6 | Associations between serum antibody responses and abundances inferred from metagenomics sequencing. a, Testing antibody bound peptides which appeared in >2% of individuals (4469 peptides) vs. relative abundances of species (SGBs[50]) which appear in >2% of individuals (1056 SGBs). Of them, 1706 pairs (listed in Supplementary table 5) passed FDR correction (p-value <0.05) for multiple hypothesis testing (approximately 4.7 million tests). Most of these associations were from peptides and species that appeared in a small percentage (2-5%) of individuals. We also performed the same test of peptides and species which appear in >5% of individual (745 species and 1566 peptides) with 12 pairs passing FDR correction. Some of the species abundances are correlated with the fold change of up to 23 peptides per species (histogram in Fig. 3b). This analysis includes also associations of multiple SGBs with the same peptide. For example antibody binding of the Shigella IpaC protein (antibody binding against which we had found to be associated with age [Fig. 4c]) was associated with abundances of various SGBs, suggesting multiple factors contributing to its biological effects (for example potentially increased translocation as well as effects mediated by the adaptive immune system). These results are affected by detection thresholds of PhIP-Seq and metagenomics sequencing and we cannot rule out that small amounts of bacteria undetectable in microbiome sequencing eliciting weak antibody responses would associate in a larger fraction of individuals. Another technical consideration beyond the detection threshold is the library content size, with the option of creating PhIP-Seq antigen libraries specific to individuals potentially allowing to capture links between metagenomics data and antibody epitope repertoires at greater depth. **b**, Representation of the 1,706 significant population scale associations between antibody binding against peptides (x-axis) and detection in metagenomics sequencing (y-axis). Every dot represents one of the significant associations listed in (Supplementary table 5). Each dot is colored by the FDR-corrected p-value of the Spearman correlation (also listed in Supplementary table 5). **c,d** Correlation of every person's metagenomics gut microbiome sequencing data with the antibody repertoire data (similar to Fig. 6a,b) on gut metagenomics antigens/genes (methods section).



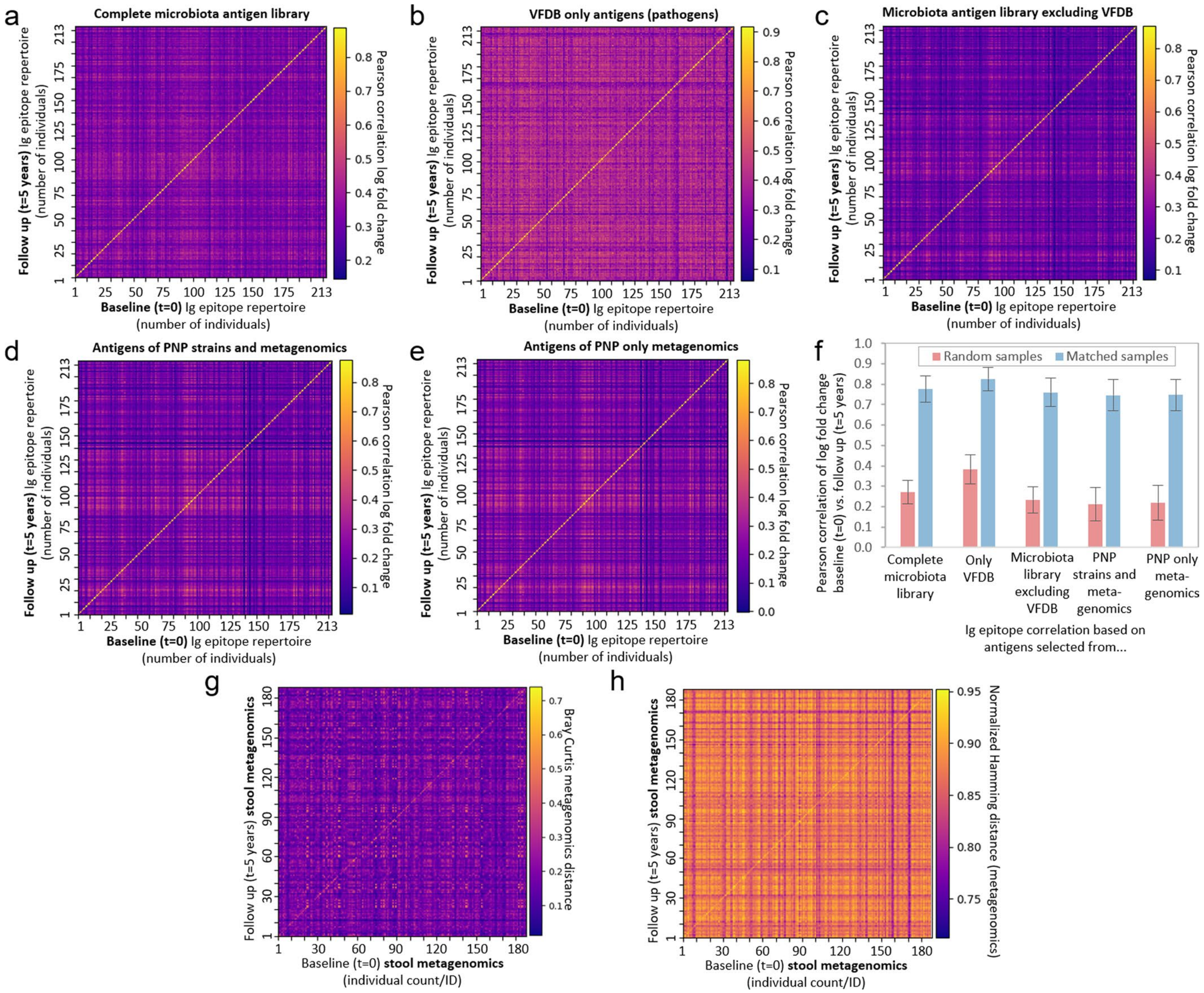
Extended Data Fig. 7 | Peptide ELISAs validations of PhIP-Seq results. Sections of 20 amino acids (aa) of six peptides included within our PhIP-Seq library were chemically synthesized and tested in a peptide ELISA against sera of 80 individuals, for whom also PhIP-Seq data was available (see M&Ms section "Peptide ELISAs" for the selection criteria and sequences of the peptides). **a-f**, Comparison of peptide ELISA and PhIP-Seq data for each peptide (as indicated by the title above each panel). Each dot represents data of one individual. Absorption values of ELISA data and p-values of significance of enrichment of binding in PhIP-Seq (for one peptide) are shown on the x and y axes respectively. Absorption values below the average of the negative control were normalized to 0. Spearman correlation (R) with associated p-value (Spearman rank-order correlation coefficient, nonparametric measure) was computed for each pair of PhIP-Seq and ELISA data (shown in each panel). The negative control peptides were not bound in PhIP-Seq, hence Spearman R/p-val are not applicable (n.a.). **g**, Summary of the results shown in panels a-f. The percentage of ELISA or PhIP-Seq binding in the 80 individuals was calculated for each peptide with the standard Generalized Poisson cutoffs for PhIP-Seq (with binding of multiple peptides summarized if applicable, see text below) and the ELISA data was counted as positive when the absorption value was greater than the average of the negative control. The calculated Spearman correlation (R) between the frequency of antibody responses in PhIP-Seq and ELISA is shown in the panel.



Extended Data Fig. 8 | Antibody epitope repertoires against the microbiota antigen library significantly predict C-reactive protein (CRP) levels (measured with a wide CRP range test) by machine learning, albeit with lower predictive power than age (Fig. 5a) or gender (Fig. 5b). a, Age and gender alone are confounding factors of machine learning based predictions of serum CRP levels (measured with a wide range test for ca. 400 individuals). As antibody epitope repertoires also carry a wealth of age/gender related information (Fig. 5) the contribution of age and gender alone vs. antibody epitope repertoires was assessed here. CRP levels were predicted using age and gender alone as features and with a combination of age, gender, and microbiota antibody epitope repertoires as features, using Ridge Regression 10-fold cross validation. Each model was repeated 100 times (different cross validation sets) and a histogram of the resulting 100 Pearson correlation coefficients (correlation of actual vs. predicted CRP values) are shown in panel a. The analysis has been corrected for multiple hypothesis testing: Pearson correlation of predicted (on antibody bound peptides + age + gender) to actual value is 0.12, with p-value of 0.011, which after FDR correction becomes 0.018 (<0.05, i.r. passes FDR correction). Pearson correlation of machine learning based prediction on age & gender alone to actual value is 0, so that all predictive power comes directly from antibody bound peptides, and not from their prediction of age and gender. Thereby a significant added predictive value of microbiota antibody epitope repertoires is demonstrated. **b**, For both 1.) other blood tests beyond CRP or 2.) anthropometrics such as body mass index (BMI), adding microbiota antibody epitope repertoire information rather worsens machine learning based predictions compared to age and gender alone (as additional meaningless features increase noise, see methods section) or did not pass FDR correction. This notion is exemplified with machine learning based prediction of BMI in panel b.



Extended Data Fig. 9 | In contrast to antibody responses targeting microbiota antigens, reactivities against human self-proteins and random peptides carry virtually no predictive power by machine learning for age (a) and gender (b). This finding precludes that self-reactivity or potential cross-reactivity against random peptides underlie the strong associations observed. The machine learning based predictions based on human proteins and random peptides encompassed ca. 6,300 peptides included in the antigen library as part of the IEDB (autoantigens) or as controls (covering abundant proteins such histones, glycolytic enzymes etc., see the methods section and Supplementary table 1 for details). Average and standard deviation derived by 10 repeats of XGBoost with 10-fold cross validation (as in Fig. 5). Ideally, the same number of controls as microbial peptides should have been used. However, that would have doubled the cost of the PhIP-Seq library as well as doubling the cost of every assay performed (as we would have had to sequence deeper and use more beads to retain the same signal strength). Given these cost considerations, we could not afford to include a set of nearly 250,000 controls. However, we believe that also the set of only 6,300 peptides serves as an important control: We detected very little binding against these controls (discussed in more detail in S2/S3), and they do not carry any predictive power by machine learning, demonstrating that there is no exceedingly large cross-reactivity or background signal with our PhIP-Seq system.



Extended Data Fig. 10 | Additional analyses of longitudinal stability. Complete correlation diagrams for the five-year longitudinal antibody stability results of 213 individuals shown in Fig. 6 (panels a, c, e) and additional subgroups of the antigen library (c,d) as well as two different approaches to assess gut microbiome stability from metagenomics sequencing data (g,h). Pearson correlations of log fold changes of all baseline (t = 0) and follow up (t = 5 years) samples compared with each other are shown. Correlations based on antigens of the entire microbiota library (**a**) [also shown in Fig. 6a], only the VFDB (**b**), and the microbiota library excluding VDFB (**c**) are shown. In addition to these antigens obtained from databases, two analyses with antigens from microbiome sequencing of this cohort[28] (Methods) were performed (**d,e**). **f**, Summary figure on the correlation coefficients of the stability of antibody epitope repertoires from antigen subgroups shown in panel a to e of this figure, comparing correlation of random pairs of samples and pairs of matched individual's samples collected five years apart. Mean values and standard deviations of n = 213. Sample sizes: random pairs of samples: 213^2-213 comparisons; individuals' matched samples: 213 comparisons (see Fig. 6b,c for details). Antigen groups sizes for panels a-f: All microbiota – 231,975 peptides, VFDB – 24,164 peptides, Library excluding VFDB – 207,811 peptides, Metagenomics antigens - 147,061 peptides. **g,h** Gut microbiome stability inferred from metagenomics sequencing of stool samples collected five years apart of 188 individuals. In panel g, stability is calculated from gene abundances. The Bray Curtis distances for all baseline (t = 0) and follow up (t = 5 years) samples compared with each other are shown (the higher the value, the closer the samples resemble). In panel h, stability is calculated based on presence/absence (existence) of genes appearing in individuals (by applying a cutoff threshold to the gene abundances). The Normalized Hamming distances for all baseline (t = 0) and follow up (t = 5 years) samples compared with each other are shown (the higher the value, the closer the samples resemble).

nature research

Corresponding author(s): Eran Segal

Last updated by author(s): 5/22/2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

All antigen/peptide creation 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).

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), and 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).
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).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	strict FDR/Bonferroni correction for differences between groups and cross validation for predictions (see below), the sample size was chosen as the largest number of samples available for each phenotype
Data exclusions	Samples not passing the threshold of >200 peptides significantly bound and were excluded from analyses (see materials and methods). For every phenotype, outliers >3 standard deviations from mean were removed (mean and std estimated from central 90%). These cutoffs apply to all the experiments shown in the manuscript. The exclusion criteria were not pre-established, as we had used the PhIP-seq assay for the first time on such a large cohort.
Replication	Gradient boosting decision trees (XGBoost classifier) with 10-fold cross validation were used for predictions. Details on replications of experiments are provided in the respective figure captions (e.g. Extended Figure 1b).
Randomization	All analyses are performed on one group of healthy individuals. For machine learning cross validation the order of particpatns (divison to 10 folds) was done by using the Python random library (no limitations on the order).
Blinding	All analyses are performed on one group of healthy individuals, there was no blinding necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Details on the antibodies are provided in Supplementary Table 2 Sheet 1.
Validation	Validation of the antibodies is shown in Extended Data Fig. 3 and Supplementary Table 2 Sheet 2.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Full population characteristics are provided in Fig. 1c and the materials and methods section.
Recruitment	n.a. (samples had been previsouly collected and recruitment described in detail, Zeevi et al., 2015)
Ethics oversight	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).

Note that full information on the approval of the study protocol must also be provided in the manuscript.