

ORIGINAL ARTICLE

Allergens

The immunome of soy bean allergy: Comprehensive identification and characterization of epitopes

Karolin Kern¹ | Heide Havenith² | Nicolas Delaroque¹ | Paul Rautenberger³ |
 Jörg Lehmann³ | Markus Fischer⁴ | Holger Spiegel² | Stefan Schillberg² |
 Eva Ehrentreich-Foerster⁵ | Stefanie Aurich⁶ | Regina Treudler⁶ |
 Michael Szardenings¹ 

¹Ligand Development Unit, Fraunhofer IZI, Leipzig, Germany

²Molecular Biotechnology Division, Fraunhofer IME, Aachen, Germany

³Department Therapy Validation, Fraunhofer IZI, Leipzig, Germany

⁴PolyQuant GmbH, Bad Abbach, Germany

⁵Department of Molecular and Cellular Bioanalytics, Fraunhofer IZI, Potsdam-Golm, Germany

⁶Department of Dermatology, Venereology and Allergology, Comprehensive Allergy Centre, Universitätsklinikum Leipzig, Leipzig, Germany

Correspondence

Michael Szardenings, Ligand Development Unit, Fraunhofer Institute for Cell Therapy and Immunology (IZI), Leipzig, Germany.
 Email: michael.szardenings@izi.fraunhofer.de

Funding information

This study was funded by the Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V., MAVO project "LowAllergen"; and the German Federal Ministry for Economic Affairs and Energy, grant KF2302706. Parts of this work were additionally funded by the Fraunhofer Zukunftsstiftung, project "FoodAllergen".

Summary

Background: The precise mapping of multiple antibody epitopes recognized by patients' sera allows a more detailed and differentiated understanding of immunological diseases. It may lead to the development of novel therapies and diagnostic tools.

Objective: Mapping soy bean specific epitopes relevant for soy bean allergy patients and persons sensitized to soy bean, and analysis of their IgE/IgG binding spectrum.

Methods: Identification of epitopes using sera, applying an optimized peptide phage display library followed by next-generation sequencing, specially designed in silico data analysis and subsequent peptide microarray analysis.

Results: We were able to identify more than 400 potential epitope motifs in soy bean proteins. More than 60% of them have not yet been described as potential epitopes. Eighty-three peptides, representing the 42 most frequently found epitope candidates, were validated by microarray analysis using 50 sera from people who have been tested positive in skin prick test (SPT). Of these peptides, 56 were bound by antibodies, 55 by serum IgE, 43 by serum IgG and 30 by both. Person-specific epitope patterns were found for each individual and protein.

Conclusions: For individuals with clinical symptoms, epitope resolved analyses reveal a high prevalence of IgE binding to a few soy bean specific epitopes. Evaluation of individual immune profiles of patients with soy bean sensitization allows the identification of peptides that do facilitate studying individual IgE/IgG epitope binding patterns. This enables discrimination of sensitization from disease, such assay test has the potential to replace SPT assays

1 | INTRODUCTION

Food allergies are an increasing health concern in Western societies and may have relevant impact on quality of life.^{1,2} Soy bean is considered to be an important allergenic food and requires labelling in

the European Union when used as an ingredient in compound foods.³⁻⁵ The major allergens of soy bean are the major seed storage proteins 11S glycinin (Gly m 6), 7S β -conglycinin (Gly m 5), the oil body-associated Gly m Bd 30K as well as the proteins Gly m Bd 28K and Gly m 3.⁶ Gly m 5 and Gly m 6 are associated with severe

allergic reactions.^{7,8} Other soy bean proteins, such as Gly m 2, Gly m 3 and Gly m 4 are highly cross-reactive to other species.⁹ Over the last years, much progress has been made in the identification and characterization of soy bean allergens.^{10–14} Nevertheless, only a few allergenic protein structures responsible for provoking an allergic reaction are currently known beyond the protein level.¹⁵ Several approaches for the identification of epitopes include ELISPOT, Western blot, microarray, and/or bioinformatics methods. Sun et al¹⁶ carried out a study to identify epitopes of the α -subunit of β -conglycinin (Gly m 5.01). Using bioinformatics, they investigated the secondary structure and the immunogenicity of the protein, and 15 potential epitopes with a length of 20–30 amino acids were identified. Eleven out of these bound to IgE antibodies from soy bean-sensitive patients in dot blot experiments. Additionally, several software tools with distinct algorithms for epitope prediction have been used.^{17,18}

Another approach was taken by Schulten et al.¹⁹ 2D-PAGE, Western blot and MALDI-TOF was used for the identification of serum IgE antibody-binding proteins. Fu et al²⁰ combined MALDI-TOF analysis with a peptide array technique identifying 47 IgE-binding epitopes in Gly m 5 with sera from soy bean-sensitive pigs.

Microarrays allow parallel screening of a large variety of allergen molecules. They can provide detailed information regarding the individual sensitization profile of allergic patients. This technology allows the simultaneous detection of sensitized patients' antibody reactivity profiles towards each of the allergens requiring only minute amounts of serum. Lupinek et al²¹ developed the mechanisms for the development of allergies (MeDALL) allergen-chip. The chip covers more than 170 allergen proteins for the detection of allergen-specific antibody patterns.²¹

Efforts have been made to identify epitopes applying peptide phage display. Bøgh et al²² have investigated three serum samples from peanut-sensitive patients for their individual IgE and IgG binding patterns, using the Ph.D.^{TM-723} -phage display library from New England Biolabs (Ipswich, MA, USA). They report that "distinct IgE as well as IgG4 epitope recognition profiles and some IgE epitopes were common to all patients." The same library was used by Christiansen et al²⁴ for the investigation of four sera from peanut-sensitive individuals. This group identified patient-specific peptide clusters. To investigate the authenticity of the epitopes, these authors deconvoluted serum reactivity towards Ara h 1 on peptide microarrays. Thereby, just one 12mer peptide recognized by three out of four sera could be characterized.

Although antibodies usually recognize patterns defined by just five to seven amino acids and/or three to four sugar moieties,²⁵ the above methods could only restrict the epitopes to peptides of 10–30 amino acid in length.

Here, we describe the patient-specific and simultaneous identification of soy bean-related epitopes recognized by antibodies from human serum samples by applying a novel method combining phage display with downstream next-generation sequencing (NGS) analysis. As recently published by our group, we were able to generate variants of Gly m 4 with reduced binding by patient sera on the basis of data achieved using this experimental approach.²⁶

2 | MATERIALS AND METHODS

2.1 | Human sera

Sera were collected from 50 subjects (median age 36 years, range 19–66, 26 males) presenting at the Department of Dermatology, Venereology and Allergy, Comprehensive Allergy Centre, Universitätsklinikum Leipzig, Germany (Table S1). All subjects had a positive skin prick test (SPT) to soy bean (wheal ≥ 3 mm). SPT was performed with commercially available, supplement free soy bean drink (ALPRO, Uelzena, Germany). Specific IgE was positive ≥ 0.35 kU/L in all subjects for at least one of the following soy bean allergens: Gly m 4, Gly m 5, Gly m 6 (according to the results from ImmunoCAP assay, Thermo Scientific Phadia, Freiburg, Germany). Total IgE levels were also determined. All patients had given their written informed consent according to the vote from the local ethical committee of the University of Leipzig.

Clinical studies: Ethikkommission an der Medizinischen Fakultät der Universität Leipzig, Käthe-Kollwitz-Str. 82, 04109 Leipzig, Germany. Approval number 428-12-17122012.

2.2 | Peptide phage display

A trinucleotide-based, 16mer peptide phagemid library covering $>10^9$ clones ("ENTE-1")²⁷ was used for two rounds of panning against individual serum from patients sensitized to soy bean. We coated 24 μ g total serum protein in 2.5 mL PBS to Immuno Tubes (MaxiSorpTM; Thermo Fisher Scientific, Waltham, MA, USA) for 1 hour at 4°C under constant agitation. Under these conditions, all Ig molecules should be bound on the tube surface. A negative control was incubated with PBS only. The coated tubes were washed with 2.5 mL blocking buffer once (5% [w/v] non-fat dry milk powder in PBS) and incubated with 4 mL blocking buffer at 4°C for at least 15 minutes under constant agitation. Before adding the ENTE-1 peptide phage display library, the tubes were washed three times with 4 mL PBS. To preferably select patient-specific antibodies, and in particular to prevent selection of peptides binding serum proteins, a competition with a mixture of ten sera from healthy donors was conducted. The ENTE-1 library was added to each tube in 1 mL blocking buffer containing 30 μ g of non-allergic serum protein. In the first selection round, an amount of phage corresponding to 4×10^{11} colony forming units (cfu) of ENTE-1 library and in the second selection only 10^9 – 10^{10} cfu of recovered phage were used. Tubes were incubated for 2 hour at 4°C with constant agitation, washed five times with 1 mL 0.1% (v/v) Tween 20 in PBS in the first selection round, and with 0.5% (v/v) Tween 20 in PBS in the second selection round. Bound phage was eluted in 1 mL of elution buffer (0.1 mol/L glycine-HCl, pH 2.2), immediately neutralized with 200 μ L neutralization buffer (1 mol/L Tris-HCl, pH 8.2) and allowed to infect 10 mL of actively growing *E. coli* TG1(λ)²⁷ for 30 minutes at 37°C. Infected cells were spread on large dYT-agar plates (~ 10 clones per mm²) and incubated overnight. Colonies were washed off the plate with a

small amount of dYT and an aliquot was cultivated with M13K07 helper phage²⁸ in dYT for 7 hours for phagemid production. On average 10^5 - 10^6 phagemid clones were recovered.

2.3 | DNA isolation

Phagemid DNA was isolated from phage-infected bacteria recovered from dYT-agar plates after each panning round. After washing the plates, 1.5 mL bacteria corresponding to an optical density at 600 nm (OD_{600}) of 1.5-2.0 were centrifuged. DNA was purified from bacterial cells using the Nucleobond miniprep kit (Macherey-Nagel, Düren, Germany). DNA concentration and purity were determined using a ND-1000 Nanodrop spectrophotometer (Thermo Fisher Scientific).

2.4 | Sample preparation for NGS

DNA isolated from bacteria or purified phage²⁹ was used as PCR template to amplify the hypervariable 16mer peptide coding sequence. Oligonucleotide primers were from Eurofins Genomics (Ebersberg, Germany). The forward primer P5 N501 is specific for the ompA signal peptide leader sequence gene in the pPepPr3A phagemid vector.²⁷ The reverse primer mix P7 A7xx (xx denotes one distinct primer out of 24 used) is complementary to the gene-3 coding sequence downstream of the variable-peptide coding sequence. All primers include at their 5' ends sequences

compatible with the sequencing primers and Illumina (San Diego, CA, USA) adapter sequences.³⁰ The different reverse primers contain the variable index sequence. After amplification, the PCR products were purified using the QIAquick PCR purification kit (Qiagen, Venlo, the Netherlands). DNA concentrations were determined using the Qubit broad range kit (Life Technologies, Wilmington, MA, USA). An equimolar library of all samples was prepared. Mixed PCR products were gel-purified using the E.Z.N.A.[®] Gel Extraction Kit (Omega, Norcross, GA, USA) and their concentration measured again using Qubit system (Invitrogen, Carlsbad, CA, USA). 4 nmol of the DNA were denatured with 1 mol/L NaOH, from which 7 pmol were used for NGS using Illumina MiSeq with MiSeq Kit V2 performing 300 cycles (Illumina).

2.5 | Data analysis

The Illumina data sets of the sequence reads in both directions of the DNA template in each spot were combined by applying routines of the EMBOSS software package to generate a highly reliable data set. Further processing was carried out with a specially developed interactive JAVA™ based software (LibDB, EPITOPIC, Leipzig, Germany). As the ENTE library is using a restricted codon set, at first the sequences were compared to the expected codons incorporated by the trinucleotide-based library synthesis. Sequences with one or more nucleotide deviations were excluded from further analysis. LibDB then used the remaining sequences to generate a table

TABLE 1 Alignment of phage display sequences with the soy bean protein sequence. Listed are only sequences with identical and continuous 5 amino acids

Sequence	Serum	Frequency
LLLLVFASQV VQTEG RVCSQSHGF	<i>Gly m 2 (naïve sequence)</i>	
---GVFGSIT VQTEG VYSC-----	13003	1
----GWKNCA VQTEG CNCYF-----	13003	2
----GHQACE VQTEG FLLHY-----	13006	2
-GSGPSSATQ VQTEG QD-----	14011	1
-GTWSCRYYQQ VQTEG QD-----	13003	1
-----GVNV VQTEG AWHAQAY---	13003	1
-GVHMCKLHLL VQTEG QC-----	13003	1
-----GSIWSQ VQTEG YRHNHA----	13003	2
----GWIVSFA VQTEG HGVAD-----	13003	70
---GKDSCVSR VQTEG LAVD-----	13006	1
----GQLVCPH VQTEG ECVSC-----	13006	2
---GIDGCPIT VQTEG QGPD-----	14011	1
-GGEASIPDIF VQTEG FI-----	14011	36
-GGVHSQDYEN VQTEG IF-----	14011	2
-GQPTCYNHNKQ QTEGR C-----	13003	1
-----GEELC QTEGR GPEVQPY--	13003	2
-----GYIVS QTEGR HPQHQA--	13003	4
----GRINSGSR QTEGR YGTY-----	13006	1
---GWWFSVLLT QTEGR KQC-----	13006	1
-----GAHTCIF QTEGR IHPQF----	13006	1
-----GHRPCVP QTEGR QAHD----	13006	1
-----GTWNSQ QTEGR GSDNCA---	13006	1
-----GSVPSKRF QTEGR EQII-----	14011	4

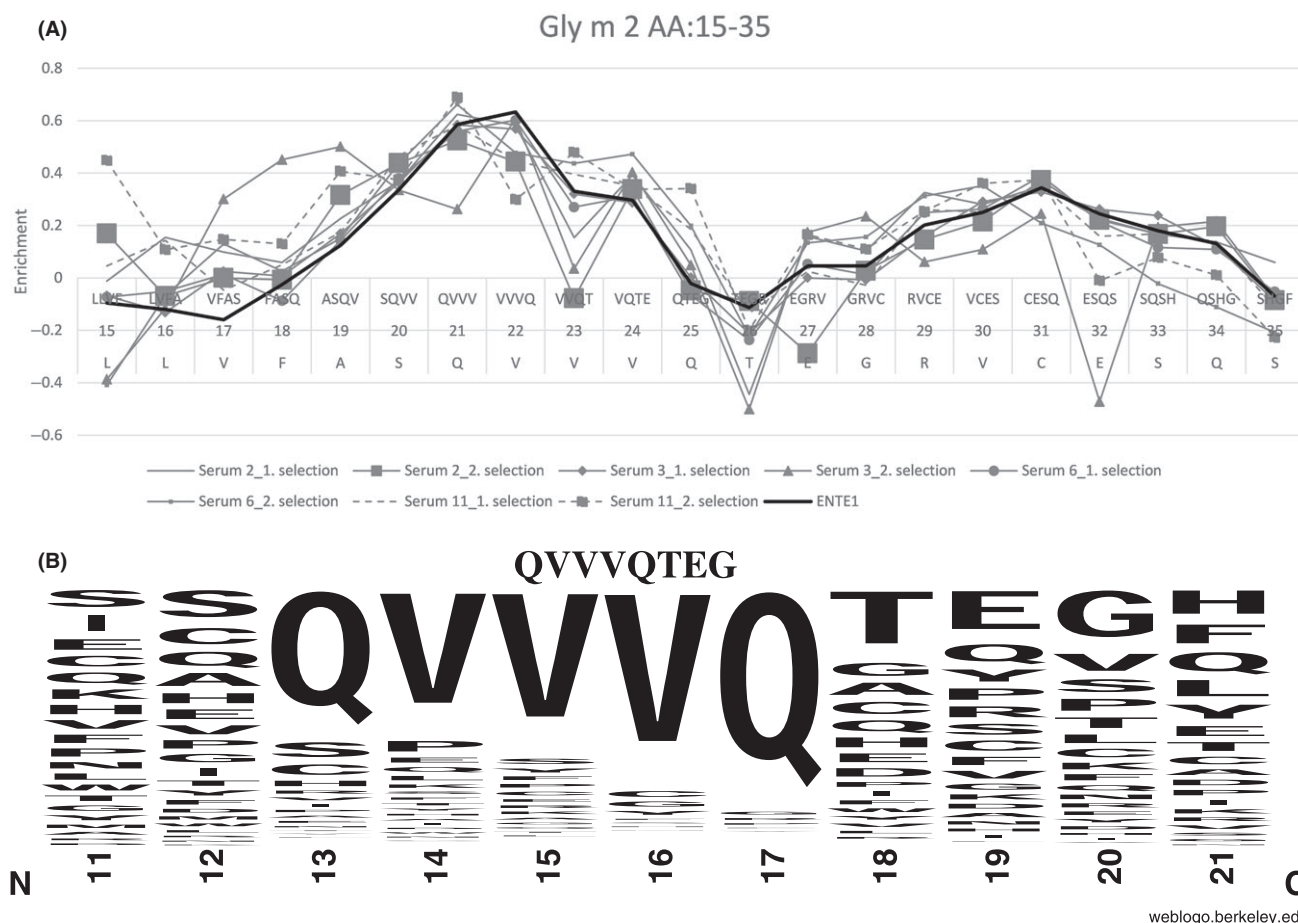


FIGURE 1 A, B show the steps of epitope identification and validation. The first step is the comparison enrichment of 4mer motifs' within sequence data sets identified by peptide phage display with 4mer motifs' enrichment of sequences from the naïve library plotted against the 4mers of the potential antigen (A). If a motif is enriched in comparison to the naïve library full-length sequences are retrieved from a database and aligned to generate a longer consensus sequence. Epitope fingerprint indicates most important amino acids. A weblogo⁵⁴ is obtained by aligning 5267 sequences from experiments with serum 13003, 13006 and 13011 containing the 4mer motifs QVVV, QxVVQ, QVxVQ, QVVxQ, VVVQ, VxVQT, VVxQT, VVVxT, VVQT, VQTE, VxTEG, VQxEG, VQTxG and QTEG (B)

of sequence frequencies and a distribution of amino acids at each position of the library. Based on these sequences, an indexed database of 3mer and 4mer amino acid motifs was generated. The frequencies of each motif in the data set were calculated, as well as their theoretically expected frequencies, as expected based on the number of sequences and the design of the starting library, and finally by comparison the yield the enrichment of each motif. For practical reasons, all these data are stored for every motif in the database. Listing or graphically displaying these values for all 4mer motifs present in an antigen's sequence identifies regions of enriched motifs.

Enriched motifs may be potential parts of epitopes or of mimotopes, peptides mimicking of the essential features of an epitope.³¹ The sequences that contain an enriched motif can be retrieved from the database. These sequences were aligned (Table 1) and checked for extensions homologous to the protein's sequence corresponding to 5 or 6 amino acid superposition. An epitope fingerprint derived from the alignment of all motif sequences indicated the amino acids

most relevant for binding to the serum antibodies (Figure 1B). Epitope/mimotope candidate peptides were validated in a microarray approach (Table 2).

2.6 | PEPperCHIP® peptide microarray

The PEPperCHIP® peptide microarray has previously been described.²⁶ In brief, peptides coated to the microarray covered 10 soy bean proteins (Gly m 1, Gly m 2, Gly m 3, Gly m 4, Gly m 5.01, Gly m 5.02, Gly m 5.03, Gly m 6.01, Gly m 6.02, Gly m 6.03, Gly m 6.04, Gly m 6.05, Gly m agglutinin, Gly m P34 und Gly m TI). All soy bean protein sequences were divided into 15mer fragments with four-residue overlap and the chip was designed bearing duplicates of every peptide. Microarrays were purchased from PEPperPRINT (Heidelberg, Germany), performing peptide synthesis one residue at a time directly onto a glass slide using solid-phase Fmoc chemistry.

Each slide was washed in staining buffer (PBS containing 0.05% [v/v] Tween 20 and 0.1% [w/v] BSA, pH 7.4) for 10 minutes at room

temperature (RT) and blocked for 30 minutes at RT in blocking buffer (PBS containing 0.05% [v/v] Tween 20 and 1% [w/v] BSA, pH 7.4). The slides were then incubated in 16/50 sera from soy bean-sensitive patients diluted 1:50 in staining buffer and placed overnight at 4°C. After three washes in standard buffer (PBS containing 0.05% [v/v] Tween 20, pH 7.4), slides were agitated at 200 rpm for 30 minutes at RT in a solution of goat anti-human IgE antibody (Thermo Fisher Scientific) diluted to 200 ng/mL in staining buffer. After three further washes, binding was detected by incubating the array for 30 minutes at RT with a secondary Cy5-labelled donkey anti-goat antibody (Abcam, Cambridge, UK) diluted to 200 ng/mL in staining buffer. Finally, the slides were washed again three times and detected with microarray reader (Genepix 4300; Molecular Devices, San Jose, CA, USA).

2.7 | Soy bean mimotope microarray preparation

Epitope candidates were supplied externally (Ines Neundorff, University of Cologne). The peptides had the following structure: Acetyl (Ac) – N-epitope sequence-C – Ebes – Lys(ε-N₃) – amide (Ebes = amino-3,6-dioxaoctyl-succinamic acid). All peptide sequences were based on either the naïve soy bean protein sequence or the mimotope from phage display experiments. The peptide length varied between eight and nineteen amino acids. Peptides were coupled to amino-DBCO and immobilized with different concentration (81 ng/μL, 27 ng/μL, 9 ng/μL) in two different clusters on GLYMO ((3-glycidyloxypropyl)trimethoxysilane)-modified slides. In total, each slide contained 2160 spots.

2.8 | Soy bean mimotopes microarray

All 50 sera were applied to the soy bean mimotopes microarray. Each slide was blocked for 30 minutes in blocking buffer (PBS containing 0.1% [v/v] Tween 20 and 1% [w/v] BSA, pH 7.4). For the incubation with sera, cover slides (HybriWell™ 60 mm × 21 mm × 0.15 mm) were used. The slides were incubated with 100 μL sera from soy bean-sensitive individuals diluted 1:50 in blocking buffer overnight at 4°C. After two washes in washing buffer (PBS containing 0.1% [v/v] Tween 20, pH 7.4), the slides were incubated for 60 minutes with a mouse anti-human IgE antibody (Abcam) solution, diluted to 100 ng/mL, or goat anti-human IgG antibody Cy5 (Abcam), diluted to 100 ng/mL in blocking buffer. All sera were investigated with both antibodies. After two further washing steps, binding was detected by incubating the array for 60 minutes at RT with the secondary Cy5-labelled goat anti-mouse antibody solution (Thermo Fisher Scientific), diluted to 400 ng/mL in blocking buffer, respectively. Fluorescence signals of the array spots were detected using a microarray reader (Agilent G2565BA, Santa Clara, CA, USA) at 635 nm, and the data were evaluated using GenePix® Pro 6 software (Molecular Devices) to determine the signal intensity of each peptide spot. The software calculated the signal-to-noise ratio (SNR) for every spot. An overall threshold of 5 SNR units was set, and signals above this threshold were assumed to indicate positive recognition by patient sera.

3 | RESULTS

3.1 | Data processing

Sequence data from phage bound by serum were obtained by NGS. The data from NGS were read into the LibDB software, resulting in 0.2–1.0 million sequences from each NGS run after removal of all ambiguous sequences. According to the manufacturer (Illumina), an error rate of 0.1% per base can be expected, which was in good accordance with our observation that about 20%–35% of all sequences contained potential read errors, since they did not fit to the library's trinucleotide sets. Only data sets from panning experiments giving more than 170 000 sequences that were included in downstream analyses (see section Discussion). Typically, sequences obtained from the first selection round data include approximately 138 000 different 4mer motifs of 160 000 theoretically possible. This fits well to the expected numbers, given the almost complete absence of methionine and low frequency of tryptophan and cysteine in the library is taken into account.

3.2 | Epitope identification

From the theoretically expected frequency and the observed frequency in the experimental data set, the enrichment can be calculated for each motif within all sequences in the database. This is plotted against the 4mer sequences of a potential antigen together with the same data showing the observed variations of the naïve ENTE1 library (Figure 1A), revealing epitopes when the motif is found only 2× more frequently in the data after selection than expected by the data of the starting library. Only their alignment (Table 1) reveal, whether a potential epitope has been identified, by showing multiple sequences with identical or similar amino acids to the soy bean protein sequence. Additionally, motif statistics from different sera data sets were compared (see also patient's individual fingerprints in the Appendix S1).

Table S2 (appendix) lists potential soy bean epitopes meeting our selection criteria:

- the central motif is enriched compared to the naïve library
- the motifs derived from different sequences are at least individually five amino acids long, or overlapping shorter motifs suggest this size to be reached
- data from at least two different patients support the presence of antibodies recognizing such a motif.

A total of 405 potential motifs for immunogenic epitopes from fourteen soy bean proteins have been identified with the phage display approach data obtained from 22/50 sera. The number of potential epitope motifs in each soy bean protein in total is listed in Table S3. In accordance with the array results, 60% of them were characterized from Gly m 6, followed by Gly m 5. Only three antibody-binding motifs were identified from Gly m T1, see Tables S2 and S3 (appendix). Each sequence in Table S2 is listed only once, even if the sequence can be found in several related protein



FIGURE 2 Epitope map of Gly m 2. This is typical for all observed allergen maps. The x-axis displays the protein sequence of Gly m 2 (GenBank: Z13956). The serum donors are listed on the y-axis. Each line represents a single individual, the last two persons are not sensitized to soy. The spectrum of observed epitope regions is highly heterogeneous even amongst the individuals with soy allergy, illustrating the complexity of the immune response as well as the value of the knowledge of individual epitopes for diagnosis and prognosis. Positions are marked, when the sequence was identified by: ■ phage display, □ PEPperPRINT array²⁴, ■ soy bean mimotope array. Not all sera were used for each investigation

subunits, eg, in Gly m 5.01 and Gly m 5.02. It is noted in Table S2 when an identical epitope is found in several subunits.

Individual sera recognize varying amino acids of each epitope. Based on the alignment of enriched motifs, the binding stretch comprises five to fourteen amino acids, with an average of approximately seven amino acids (Table S2). These may either represent long linear epitopes or parts of structural epitopes, eg, loops, and probably also reflect the individual genetic and B-cell maturation effects.

A complete list of all identified epitopes can be found in the appendix (Table S2). As examples maps of identified epitopes of Gly m 5 and Gly m 4 are shown as supplementary material. Gly m 5 consists of three subunits, which are very similar in their amino acid sequence. The α' and β subunits share 90.14% and 76.2% sequence identity, respectively, with the α subunit.³² Homologous sequences in other soy bean proteins, eg, Gly m 5.02 and/or Gly m 5.03, are listed in a separate row of Table S2. Twenty epitopes recognized by IgE for Gly m 5.01 are currently listed in the Immune Epitope Database (IEDB). Thirty-six new potential epitopes have been identified in our study. Many published epitopes claim a length of more than 20 amino acids. Our data suggest that some of these longer sequences may contain more than one antibody-binding site. Sixteen epitopes are listed more than once with different IEDB IDs, indicating that these epitopes have been described several times from different experiments. Even though soy bean allergens have been studied in great detail, nine new epitopes of Gly m 5.01 were identified using the phage display approach.

3.3 | Variations in antigenic proteins recognized by individual sera

For all soy bean-sensitized patient sera investigated in this study several potential epitopes in soy bean proteins have been identified,

but the antibody-binding recognition pattern of each serum is indeed specific for each patient. All sera recognized various epitopes of Gly m 5 and Gly m 6, while epitopes of the highly cross-reactive protein Gly m 4 (homologue of the major birch allergen Bet v 1) and many other PR10 proteins³³ were found for only a few. In all patient sera, epitopes from two or more soy bean proteins were found.

As an example, an overview of all potential epitopes for Gly m 2 is presented in Figure 2.

3.4 | Results from peptide arrays

Two types of microarrays were used for the validation of epitopes. To initially confirm the results above, assays using parallel overlapping peptide chips were performed on a subset of 10 proteins and 16 sera. Serum samples were tested with PEPperCHIP[®] peptide microarrays. A total of 241 IgE-binding peptides could be identified in 10 soy bean proteins. Different epitope recognition patterns were found for each serum. Peptides from Gly m 5 and Gly m 6 were the most frequent IgE-binders, while only a few peptides from Gly m 1 and Gly m 3 were recognized by IgE antibodies (Table S3).

The second approach was to use a microarray of the peptide epitope candidates identified by the phage display approach (Table 2). Short peptide sequences comprising mimotopes or native peptide fragments were immobilized on specially coated glass slides using “click chemistry”. The peptide spots represented epitopes of 10 different soy bean proteins.

It is well known that Gly m 5 and Gly m 6 are the major causes of soy bean allergy.^{10,12,34} However, we have chosen to expand our initial investigations to other soy bean proteins, for example, Gly m 1, Gly m 2 and Gly m BD30K to validate phage display data showing that more than a few soy bean proteins are

involved in soy bean allergy. Eighty-three peptides, representing 42 different epitopes out of more than 400 epitope candidates, were chosen and tested by microarray. These 42 epitopes represent the soy bean protein regions recognized by the majority of human sera. The selected peptides were based on the native soy bean protein sequence (Pr) and/or on a phage display sequence (Ph). The latter should be referred to as "mimotopes" of identified epitopes. The chosen peptide was based on phage display data incorporating also C- and N-terminal amino acids according to sequence alignment data.

We selected 36 of the 83 peptides based on the native soy bean protein sequence and 47 of the 83 mimotopes were mimotopes.

3.5 | Array data evaluation

The results of the microarray experiments are listed in Table S4. Only peptides with a signal five times over background are listed in this table. Under these criteria antibody binding to 56 out of 83 peptides could be detected. Fifty-five of them were recognized by IgE and 43 by IgG antibodies. Fifty-six of 83 peptides, corresponding to 34 out of 42 epitopes, bound serum antibodies, 22 were recognized by at least five different sera.

Several examples indicate the accurate selection of epitope sequences. For example, peptide pairs EN3 (Pr)/DO14 (Ph), EN17 (Pr)/EN18 (Ph) and DN11 (Pr)/EN8 (Ph) contain amino acid combinations required for antibody binding. However, in several examples the mimotope sequences are better suited for serum binding, eg, the corresponding peptide pairs EN5 (Ph)/DN6 (Pr), EM20 (Ph)/EM19 (Pr) and DO19 (Ph)/EN4 (Pr) (Table S4). In contrast, for the Gly m 2 5mer motif LCNRD the natural sequence is better, see DO2(Pr)/EL14(Ph) in Table S4. Epitopes identified with phage display, overlapping microarray and soy bean mimotope microarray are summarized in Figure 2.

3.6 | Patient-specific recognition patterns

The majority of antibody-binding peptides are from Gly m BD 30K, Gly m 5, Gly m 6, Gly m BD28K and Gly m TI. Gly m BD 30K, Gly m 5, Gly m 6 epitopes were frequently recognized by IgE and IgG antibodies to the same extent (see Appendix, Table S4). The IgE/IgG antibody recognition pattern of each serum is completely individual-specific. Sera can be subclassified into three clusters of similar size by their epitope recognition patterns:

- IgE/IgG recognition
- solely IgG recognition
- solely IgE recognition

Clinical reactions with mild to severe symptoms after consuming soy bean had been reported by 19 individuals (see Appendix, Table S1). In the array, their IgG pattern shows no or only a few positive peptides. As can be seen from Figure 3A,B, some peptides are recognized by sera from this cluster as well as by sera from

silent sensitized persons to the same extent. A few peptides are recognized by a larger number of clinical sensitized individuals. Evaluation of the IgE-binding data alone shows that with this approach it may be possible to identify 81%-94% of all clinical sensitized individuals utilizing only three peptides (Figure 3). For a number of peptides, a difference between the two patient clusters in the IgE/IgG ratio can be seen. Without further optimization, the IgE pattern of four peptides would be sufficient to identify 80% of the apparently clinical relevant patients vs only 50% of the sensitized patients.

4 | DISCUSSION

The initial intention of this work was to establish reliable and sensitive identification of a few peptide epitopes directly from sera. The phage display approach can be tuned to be much more powerful than recent^{24,35,36} publications in the field suggest. The large number of observed epitopes allowed a broad screening of patient sera. We have shown here that the epitope-based immunological analysis can provide a much more detailed picture of the individual's personal sensitization status than expected.

4.1 | Patient characterization

The persons' soy bean sensitization was determined by SPT. All individuals have completed a standardized medical interview. For 19 of these individuals, clinical reactions were confirmed by their convincing medical history. With respect to the scope of this study and the available funding, we could not use sera from patients with provocation studies. The results actually confirm that these patients' epitope profiles are different to the majority of the other persons.

4.2 | Peptide phage display for serum epitope mapping

Peptide phage display has been recently applied with limited success to the analysis of serum antibody epitopes. Usually, high antibody titres are required to successfully identify at least a few epitopes.^{24,35,36} Casey et al³⁵ have used affinity-purified antibodies from serum of an Epstein-Barr virus (EBV)-infected patient to enrich peptides for the detection of patient antibodies. After six selection rounds, only 1 out of 30 clones showed specificity for IgG antibodies to EBV. Folgori et al³⁶ have used sera from individuals immunized with hepatitis B virus envelope protein (HBsAg) identifying just two different disease-specific phagotopes. Christiansen et al²⁴ have used the Ph.D.[™] 7 peptide phage display library for the investigation of four patient sera from peanut-sensitive individuals. They aligned the patient-specific peptide clusters to the primary sequence of Ara h 1. For the final epitope identification, these authors investigated serum reactivity towards Ara h 1 on a 12mer peptide microarray. Thereby, just one 12mer peptide from Ara h 1 recognized by IgE antibodies from three sera could be characterized, being consistent with the patient-specific

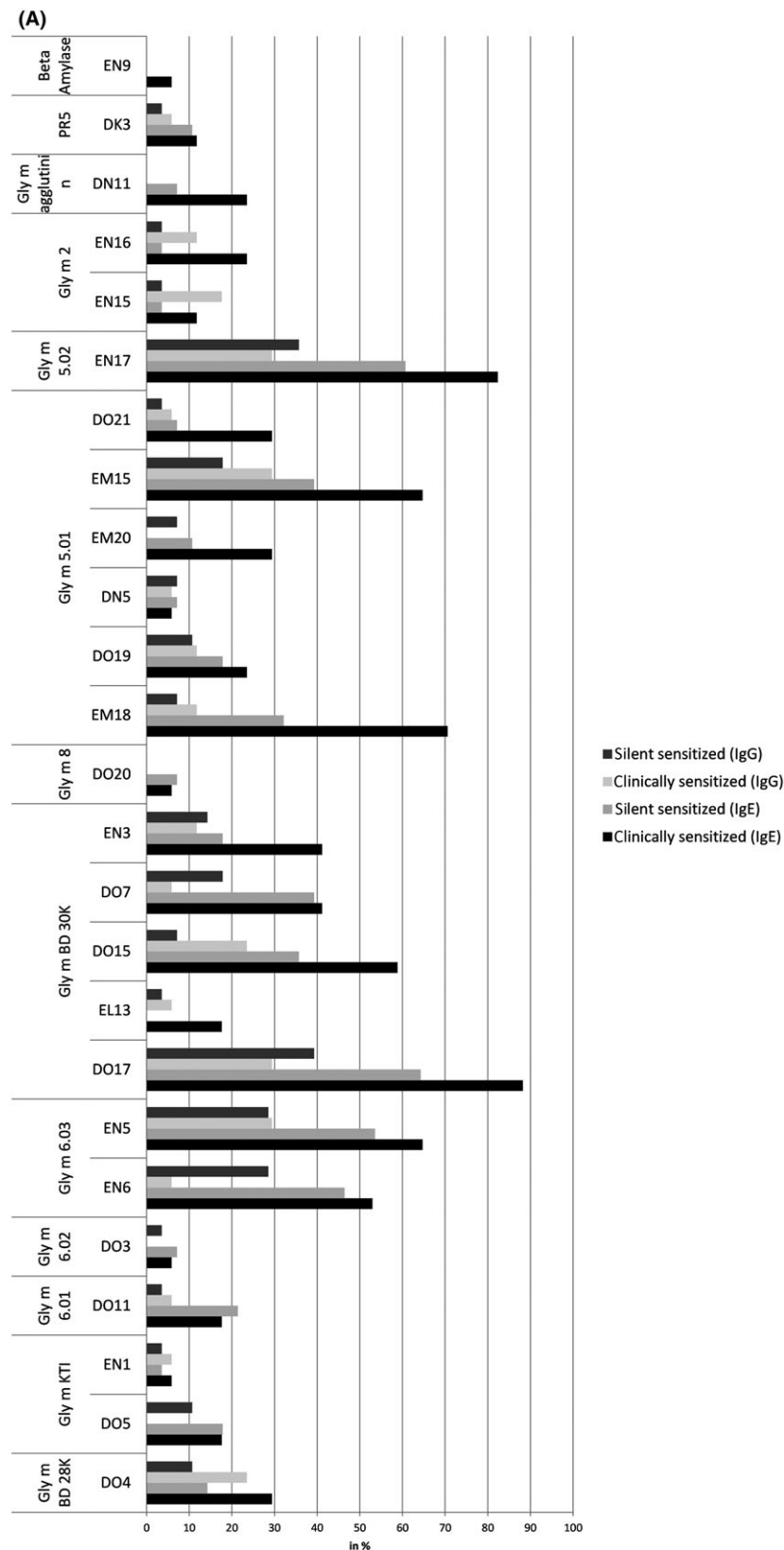
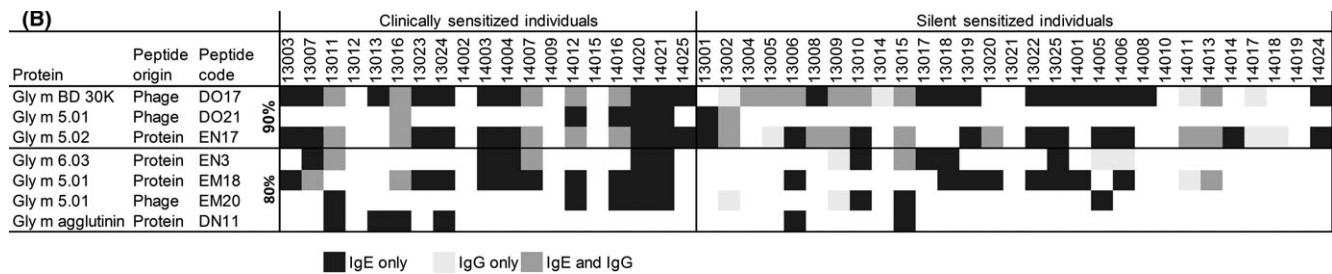


FIGURE 3 A, B, Selection of immunoglobulin binding patterns from persons with clinically relevant sensitization and silent sensitization in the microarray based on soy bean mimotopes. Serum IgE from people with clinically relevant sensitization shows more only IgE-positive peptides than immunoglobulin from other people (A). The combination of three peptides (D017, D021, EN17) allows the identification of 90% of clinically relevant sensitized people vs only 50% of the silently sensitized patients; the combination of four peptides (EN3, EM18, EM20 and DN11) allows the identification of 80% of all soy bean-sensitized people (B)



peptide cluster. These are only three examples of the labour-intensive procedures to identify just a few epitopes/mimotopes. In all these cases, the affinity selection was preceded by enrichment of IgE or IgG antibodies from serum. Experimental data gave less complex results (less enriched motifs) probably because of loss of antibodies in the enrichment steps with anti-IgE or -IgG antibodies (data not shown).

We used full serum protein bound to immuno tubes. Here, we describe for the first time the selection of multiple epitopes by a procedure with only two panning rounds. We assume that in the selection 20-30 immobilized Ig molecules may be sufficient to enrich binding motifs representing protein epitopes. The presented data are based on phage display experiments with 22 individual's sera resulting in individual epitope patterns. Recently, a similar approach using phage display in combination with NGS required complex algorithms to filter only epitopes shared between different serum samples.³⁷ By choosing a similar approach, we would have missed many important features in the allergen epitope maps (Table 2).

4.3 | Epitope identification

In recent procedures, identification of multiple epitopes (or mimotopes) was solely based on the enrichment of individual peptide sequences. The most frequent sequences obtained after several selection rounds on multiple targets typically are not the best binders, if specific at all. The analysis of motif enrichment is working on a statistical basis, thereby allowing identification of multiple sequences in a data set with defined homology. This is enabled by the application of a library with almost statistical even, and therefore, predictable distribution of amino acids in each position, even when looking at 4mer motifs after two rounds of selection (see Figure 1). Since it can be expected that the observed enrichment is caused not by a single, but by multiple B-cell clones producing slightly different antibodies, the motif-based approach allows identifying minimalistic mimotopes, probably allowing the binding of different antibody variants to the same structure (compare also epitope walking; Table 3).

In our workflow, epitope identification is based on the analysis of 4mer motifs' statistics. The ENTE1 library, containing 5×10^9 clones, provides sufficient complexity beyond all the 4mer motifs and for almost all possible 7mer motifs. This was confirmed by the identification and enrichment of sequence motifs of this length in different clones.

The selection protocol and NGS sequencing procedure provide sufficiently large datasets for the initial statistical analysis, especially when considering each sequence containing multiple 4mer motifs. For a statistical epitope analysis, at least 1.2 million motifs (average ~6-8 sequences per motif) are required. This corresponds to about 1.2×10^5 peptide sequences of a 16mer library, but only if the statistical distribution of amino acids in the library is normal, as it is the case in our library.

A sound statistical analysis based on 5mer motifs would require least 3.4×10^7 phage sequences. Such a number is corresponding to almost 1% of all phage particle variants in the selection, eg, 5×10^9 , and would require the unrealistic recovery of a corresponding number of binding clones, respectively, a much larger number of target binding sites.

In most cases, the enriched 4mer motifs identified for one antibody-binding region constitute only a part of a larger epitope sequence or structure. They comprise at least some amino acids essential for binding. The motif patterns “walk” within the immunogenic sequence (Table 3). An alignment analysis of all sequences containing these motifs reveals overlaps, potential gaps and additional matching residues outside the 4mer context, and enables both: the definition of the entire epitope structure and the identification of the most critical amino acids required for antibody recognition. Epitope maps of additional soy bean proteins are provided in the supplementary material.

4.4 | Comparison to published data

Highly allergenic soy bean proteins identified here have been previously described in other studies. Soy bean allergy-related proteins have been investigated thoroughly on the protein level. Ogawa et al³⁸ have used 69 sera from soy bean-sensitive patients to identify IgE-binding epitopes in several soy bean proteins. The three major allergens are Gly m BD 28K, Gly m BD 30K and the α -subunit of β -conglycinin (Gly m 5.01). Other major allergens are storage proteins glycinin (Gly m 6) and profilin (Gly m 3).³⁹ The acidic subunit of glycinin G1 and the basic subunit of glycinin G2¹¹ are regarded as most important allergens in patients with food allergy to soy beans. In these studies, antibody-binding sites have been identified by means of immunoblot or ELISpot.^{10,11} Potential epitopes of 15 amino acids and more resulted from these studies. Characterization of recognition patterns on the amino acid level is only possible by using time-consuming and cost-intensive ala-scan⁴⁰ The potential epitopes described here have been characterized to a much higher resolution. The

epitope fingerprint of potential epitopes, being derived from aligned all-motif sequences, highlights the most important amino acids (Figure 1B).

4.5 | Peptide arrays

For alternative initial epitope identification, a microarray of overlapping 15mer peptides covering the sequence of 10 soy bean proteins was used. Two hundred forty-one epitopes recognized by IgE antibodies could be identified with the peptide microarray approach based on overlapping naïve sequences comparing to 405 epitopes identified by peptide phage display method (Table S3). Several overlapping peptides consisting of 10 or more amino acids and representing one epitope in the array did turn out to probably consist out of two shorter epitopes. Certain epitopes are not physically or chemically representable in a microarray format, reflecting our finding that sometimes only mimotopes from the phage library bound serum antibodies in the array format. Thus, phage display is capable of providing mimotopes displaying

structural determinants that allow recognition by antibodies that otherwise may not recognize the native sequence of the epitope.

Peptide microarrays with overlapping fragments with a typical length of 10-20 amino acids have been applied as a standard method for epitope mapping.⁴¹⁻⁴⁴ Our phage display approach in contrast results in details of all Ig-binding epitopes, with sizes ranging from five to fourteen amino acids (Table S2). A peptide array reveals only the approximate location of an epitope, restricted by the peptide overlap length. The phage display method based on motif enrichment analysis renders more detailed information about the potential epitopes. Every antibody (clone) may enrich multiple-related peptide sequences. If the number of different sequences is sufficiently large, this results in an epitope fingerprint revealing the accepted amino acid variations within the antibody-binding site (Figure S4a-d). Such detailed epitope information does allow the design of selective peptide-based diagnostics for the specific identification of cross-reactivity and/or multi-sensitization.

TABLE 2 Influence of peptide sequence on the array results

Peptide	Sequence origin	IgE/IgG positive sera
LLVFASQVVVQTEG-----	Protein sequence	1/1
---FASQVVVQTEGRVCES	protein sequence	5/1
---SANQVVVQTHV-----	phage sequence	3/4

TABLE 3 Epitope walking. List of sequences with consecutive 5mer identities to Gly m 5.01. The complete epitope would probably be best defined as "SEQDIFVIPAGYPV", although none of the individual antibodies may be able to recognize the entire sequence

Sequence	Serum	Frequency
-----RAELSEQDIFVIPAGYPVVNA---	Gly m 5.01 (naïve Sequence)	
-----GVQRSEQDIFTEPEAHD-----	13006	count=2
-----GEPQSEQDIFQKCEKQF-----	14008	count=1
--GEAHCKHSEQDICAIA-----	13006	count=4
--GNLACWYNEQDIFVAHD-----	13003	count=1
GAGFCYQQYLEQDIFVY-----	13003	count=2
GEEFSQLHAFEQDIFSC-----	13006	count=46
GTKRCFNVYPCQDIFVI-----	14018	count=1
GYPSVGHQTCQDIFVA-----	13003	count=8
-----GIARSNQDIFVYDAHKC-----	13003	count=9
-----GLNRCDIFVIPECELHN-----	13009	count=2
-----GSENCDFIVIQEIQSCF-----	13009	count=3
--GTFHSIQQQRDIFVIVI-----	14008	count=1
-----GSWPSLEIFVIPIFVQI-----	13007	count=112
-----GYQWSTEHIIFVIPQORA-----	14008	count=4
-----GLNRCDIFVIPECELHN-----	13009	count=2
-----GIRACWQAPFVIPAGIC-----	13016	count=1
-----GSAVCSYGFVIPAGCII-----	14012	count=5
-----GLVSSQLTQFVIPAHCD-----	13002	count=2
-----GGTQCLQFKVIPAGHFC-----	13003	count=40
-----GEKHSEDPFLVIPAGGD-----	13021	count=8
-----GTSQSEAVIPAGHHEHY-----	13021	count=15
-----GTVVCFIWWPAGYPVDC-----	14011	count=6
-----GLKVCHGPAGYPVCPCD----	14018	count=3
-----GIWGSAGYPEECVVRD-----	14012	count=3
-----GIQRSAGYPVFKDVIDID-----	13001	count=4

Working with serum samples in allergy research makes the definition of a control very difficult. Although each individual serum sample, from healthy individuals as well as sensitized individuals may contain antibodies against food proteins,⁸ the immunological response to food proteins seems to be less pronounced in serum from healthy volunteers (own unpublished observation with other sera). Although specific immunoreactivity varies amongst individuals, it seems to be extremely difficult to obtain a universal IgE-containing serum pool as a standard comprising antibodies for all important epitopes.²⁵ Also, many individuals with allergic symptoms are multi-sensitized, and it is often difficult to assess whether this multi-sensitization is due to true species-specific co-sensitization or to cross-reactivity.⁴⁵

4.6 | Patient-specific allergen recognition patterns

State-of-the-art personalized allergy stratification is presently determined protein by protein. This allows to differentiate the individual pattern of antibodies at a much higher resolution than using allergen extracts.⁴⁶ In the microarray format, 50 serum samples from sensitized people have been tested. In general, we observed more IgE than IgG binding events. In seven sera, IgE and IgG antibodies bound the same peptide. Our results show that specific peptide reactivity is highly variable between individual patients. A similar observation was reported by Deak et al.⁴⁷ They developed “nanoallergens”, a nanoparticle-based platform for multivalent testing of potential epitopes of Ara h 2 in a cell-based assay. They concluded that some epitopes are more “immunogenic” than others. In our case, eight peptides representing different soy bean proteins have been recognized by at least 20 patient sera. These are peptides from Gly m BD 30K, Gly m BD 28K, Gly m 5 and Gly m 6. These proteins are classified as “major allergenic proteins” of soy bean.^{10–12} Holzhauser et al.⁷ have proposed that Gly m 5 and Gly m 6 are diagnostic markers for allergic reaction to soy bean.

Figure 3 shows the immune response of clinical sensitized people and people with silent sensitization of most frequently detected peptides. The mimotopes seem to be more suitable for allergy diagnosis than protein-derived sequences. Serum antibodies derived from clinically sensitized as well as from silently sensitized individuals bound to these mimotopes, demonstrating potential functionality of the peptides for future diagnostic test.

Most importantly, the combination of three peptides from Gly m 5, Gly m 6 and Gly m BD30K allow the identification of 80% of the clinically relevant sensitized people. This value is higher than standard test systems, for example, SPT (44%)⁴⁸ and ISAC (70%),⁴⁹ respectively. Even beyond, these sera from clinically relevant sensitized individuals demonstrated stronger and more frequent IgE responses to selected peptide epitopes than other sera. For example, Peptides EN17, EM18 and D017 were bound more frequently by IgE serum antibodies from clinically relevant sensitized people. These three peptides are sufficient for differentiation between clinically sensitized and silently sensitized persons.

Several researchers have found strong cross-reactivity between PR 10 proteins, for instance Bet v 1 and Gly m 4.^{51–53} However, Blankestijn et al.⁵¹ have demonstrated that storage proteins, for

example, Gly m 6, Ara h 1, Ara h 3 as well as Ana o 2, showed less cross-reactivity. The microarray used in our study shows further that Gly m 5 and Gly m 6 are suitable to discriminate between sensitized and clinically apparent or unapparent individuals.

4.7 | Detection of species cross-reactivity

A further advantage of using peptides instead of full-length proteins is in the reduction in the cross-reactivity with other proteins, eg, peanut or birch. For example, Gly m 5, Ara h 1, and Cor a 11 are members of the Fabales family. Patient sera showed cross-reactivity in *in vitro* tests.^{54,55} It is possible that a patient has cross-reactive antibodies recognizing Gly m 4 and Bet v 1, but also specific antibodies against Gly m 5. Peptides can be chosen from sequences highly conserved (eg, EM18, DN3; Table S3) or highly divergent between cross-reactive proteins.

4.8 | Further application of peptide epitopes/mimotopes

A peptide array to be applied in clinical diagnostics may of course require further optimization of peptide sequences and selection of appropriate surfaces and has to be clinically validated with a larger number of patient sera. Nevertheless, the data derived from the peptide array with a limited number of peptides used in the present study could demonstrate the high potential of peptide epitope-based diagnostics, matching the precision of existing methods without prior optimization. The selected peptides are useful for the detection of individual patient's epitope as well as sensitization patterns for particular geographic areas.

A further application is antigen-specific immunotherapy (SIT). Allergen extracts used for SIT are prepared from natural allergen sources and usually show high variation in their protein composition.⁵⁸ The relative importance of several allergens varies from patient to patient.⁵⁹ Allergen extracts also contain other proteins that are not relevant for the treatment but might even sensitize patients. Personalized patient recognition patterns can enable a selection of proteins including cross-reactive proteins for SIT.

This report demonstrates the first broad application of peptide phage display for the identification of epitopes of several soy bean proteins recognized by serum antibodies. By analysing and comparing NGS data obtained with the sera from 22 soy bean-sensitized patients, we could identify up to now more than 400 potential epitopes at the amino acid resolution level. This enormous amount of information formed the basis for an initial investigation of the patient-related IgE vs IgG binding to single epitopes on larger scale.

ACKNOWLEDGEMENTS

This study was funded by the Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V., MAVO project “LowAllergen”; and the German Federal Ministry for Economic Affairs and Energy, grant KF2302706. Parts of this work were additionally funded by the Fraunhofer Zukunftsstiftung, project “FoodAllergen”. Technical

assistance by Lena Larsson, Pia Klante, Annika Anders and Tobias Jumel is gratefully acknowledged. Holger Fehr is acknowledged for reviewing the manuscript.

CONFLICT OF INTEREST

K.K., H.H., H.S. and M.S. are named inventors on a patent application describing soy epitopes; N.D. and M.S. are named inventors on a patent covering the phage display method; M.S. is CEO and minority shareholder of EPITOPIC GmbH, a spin-off having in licensed the above technology for epitope mapping. Separate declarations by the authors are in preparation.

AUTHORS CONTRIBUTIONS

K.K. carried out most of the phage display and array work described in this manuscript and was responsible for compiling the manuscript. P.R. supported parts of this work as part of his master thesis. H.H. performed the design and the analysis of overlapping peptide arrays. H.S. contributed to the design and the evaluation of overlapping peptide arrays. S.A. and R.T. recruited patients, performed clinical investigations including SPT and collected sera. D.N., M.F., and M.S. designed and supported phage display experiments, evaluation methods and software. E. E.-F. supervised the generation of peptide arrays. S.S., R.T., H.S., J.L., and M.S. contributed to the manuscript. M.S. initiated and supervised the project. All authors have provided support throughout the project, analysed data, discussed the results and provided input for the manuscript.

ORCID

Michael Szardenings  <http://orcid.org/0000-0003-1553-9765>

REFERENCES

- Muraro A, Halken S, Arshad SH, et al. EAACI food allergy and anaphylaxis guidelines. *Pri Prev Food Allergy*. 2014;69(5):590-601.
- Beyer S, Franke A, Simon JC, Treudler R. Measurement of health-related quality of life in adult patients with birch pollen-associated food allergy. *J Dtsch Dermatol Ges*. 2016;14(4):397-404.
- Commission directive 2006/142/EC; 2006 [cited 2018 May 9].
- Holzhauser T, Franke A, Treudler R, et al. The BASALIT multicenter trial: Gly m 4 quantification for consistency control of challenge meal batches and toward Gly m 4 threshold data. *Mol Nutr Food Res* 2017; 61(3).
- Treudler R, Werner M, Thiery J, et al. High risk of immediate-type reactions to soy drinks in 50 patients with birch pollinosis. *J Invest Allergol Clin Immunol*. 2008;18(6):483-484.
- Yang WW, Mejia EGd, Zheng H, Lee Y. *Soybean Allergens: Presence, Detection and Methods for Mitigation*: INTECH Open Access Publisher; 2011.
- Holzhauser T, Wackermann O, Ballmer-Weber BK, et al. Soybean (Glycine max) allergy in Europe: Gly m 5 (beta-conglycinin) and Gly m 6 (glycinin) are potential diagnostic markers for severe allergic reactions to soy. *J Allergy Clin Immunol*. 2009;123(2):452-458.
- Ito K, Sjölander S, Sato S, et al. IgE to Gly m 5 and Gly m 6 is associated with severe allergic reactions to soybean in Japanese children. *J Allergy Clin Immunol*. 2011;128(3):673-675.
- Csáky I, Fekete S. Soybean: feed quality and safety. Part 1: biologically active components. *Acta Vet Hung* 2004; 52(3):299-313.
- Beardslee TA, Zeece MG, Sarath G, Markwell JP. Soybean glycinin G1 acidic chain shares IgE epitopes with peanut allergen Ara h 3. *Int Arch Allergy Immunol*. 2000;123(4):299-307.
- Helm RM, Cockrell G, Connaughton C, et al. A soybean G2 glycinin allergen. 1. Identification and characterization. *Int Arch Allergy Immunol* 2000; 123(3):205-212.
- Hiemori M, Bando N, Ogawa T, et al. Occurrence of IgE antibody-recognizing N-linked glycan moiety of a soybean allergen, Gly m Bd 28K. *Int Arch Allergy Immunol*. 2000;122(4):238-245.
- Ogawa T, Samoto M, Takahashi K. Soybean Allergens and Hypoallergenic Soybean Products. *J Nutr Sci Vitaminol* 2000; 46(6):271-279.
- Ogawa T, Tsuji H, Bando N, et al. Identification of the Soybean Allergenic Protein, Gly m Bd 30K, with the Soybean Seed 34-kDa Oil-body-associated Protein. *Biosci Biotechnol Biochem*. 1993;57(6):1030-1033.
- Ogawa T, Bando N, Tsuji H, Nishikawa K, Kitamura K. Alpha-subunit of beta-conglycinin, an allergenic protein recognized by IgE antibodies of soybean-sensitive patients with atopic dermatitis. *Biosci Biotechnol Biochem*. 1995;59(5):831-833.
- Sun X, Shan X, Yan Z, Zhang Y, Guan L. Prediction and characterization of the linear IgE epitopes for the major soybean allergen β -conglycinin using immunoinformatics tools. *Food Chem Toxicol*. 2013; 56:254-260.
- Dall'Antonia F, Gieras A, Devanaboyina SC, Valenta R, Keller W. Prediction of IgE-binding epitopes by means of allergen surface comparison and correlation to cross-reactivity. *J Allergy Clin Immunol* 2011; 128(4):872-879.e8
- Niemi M, Jänis J, Jylhä S, et al. Characterization and crystallization of a recombinant IgE Fab fragment in complex with the bovine beta-lactoglobulin allergen. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2008;64(Pt 1):25-28.
- Schulten V, Greenbaum JA, Hauser M, et al. Previously undescribed grass pollen antigens are the major inducers of T helper 2 cytokine-producing T cells in allergic individuals. *Proc Natl Acad Sci U S A*. 2013;110(9):3459-3464.
- Fu CJ, Jez JM, Kerley MS, Allee GL, Krishnan HB. Identification, characterization, epitope mapping, and three-dimensional modeling of the alpha-subunit of beta-conglycinin of soybean, a potential allergen for young pigs. *J Agric Food Chem*. 2007;55(10):4014-4020.
- Lupinek C, Wollmann E, Baar A, et al. Advances in allergen-microarray technology for diagnosis and monitoring of allergy: the MeDALL allergen-chip. *Methods*. 2014;66(1):106-119.
- Bøgh KL, Nielsen H, Eiwegger T, et al. IgE versus IgG4 epitopes of the peanut allergen Ara h 1 in patients with severe allergy. *Mol Immunol*. 2014;58(2):169-176.
- Devlin J, Panganiban L, Devlin P. Random peptide libraries: a source of specific protein binding molecules. *Science*. 1990;249(4967):404-406.
- Christiansen A, Kringelum JV, Hansen CS, et al. High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci Rep*. 2015;5:12913.
- Cordle CT. Soy protein allergy: incidence and relative severity. *J Nutr*. 2004;134(5):1213S-1219S.
- Havenith H, Kern K, Rautenberger P, et al. Combination of two epitope identification techniques enables the rational design of soy allergen Gly m 4 mutants. *Biotechnol J* 2017; 12(2).
- Szardenings M, Delaroque N, inventor. Phage display library. US20170058276A1.
- Vieira J, Messing J. Production of single-stranded plasmid DNA. In: Wu R, editor. *Recombinant DNA*. San Diego, Calif.: Academic Press; 1987. p. 3-11 (Methods in Enzymology; vol. 153).

29. Szardenings M, Törnroth S, Mutulis F, et al. Phage Display Selection on Whole Cells Yields a Peptide Specific for Melanocortin Receptor 1. *J Biol Chem*. 1997;272(44):27943-27948.
30. Illumina. MiSeq System Guide; 2015 [cited 2018 Jul 30].
31. Geysen HM, Rodda SJ, Mason TJ. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Mol Immunol*. 1986;23(7):709-715.
32. Maruyama N, Katsube T, Wada Y, et al. The roles of the N-linked glycans and extension regions of soybean beta-conglycinin in folding, assembly and structural features. *Eur J Biochem* 1998; 258(2):854-862.
33. Scala E, Alessandri C, Palazzo P, et al. IgE recognition patterns of profilin, PR-10, and tropomyosin panallergens tested in 3,113 allergic patients by allergen microarray-based technology. *PLoS ONE*. 2011;6(9):e24912.
34. Xiang P, Haas EJ, Zeece MG, Markwell J, Sarath G. C-Terminal 23 kDa polypeptide of soybean Gly m Bd 28 K is a potential allergen. *Planta*. 2004;220(1):56-63.
35. Casey JL, Coley AM, Parisi K, Foley M. Peptide mimics selected from immune sera using phage display technology can replace native antigens in the diagnosis of Epstein-Barr virus infection. *Protein Eng Des Sel*. 2009;22(2):85-91.
36. Folgori A, Tafi R, Meola A, et al. A general strategy to identify mimotopes of pathological antigens using only random peptide libraries and human sera. *EMBO J*. 1994;13(9):2236-2243.
37. Gerasimov E, Zelikovsky A, Mändou I, Ionov Y. Identification of cancer-specific motifs in mimotope profiles of serum antibody repertoire. *BMC Bioinformatics*. 2017;18(Suppl 8):244.
38. Ogawa T, Bando N, Tsuji H, OKAJIMA H, Nishikawa K, SASAOKA K. Investigation of the IgE-Binding Proteins in Soybeans by Immunoblotting with the Sera of the Soybean-Sensitive Patients with Atopic Dermatitis. *J Nutr Sci Vitaminol* 1991; 37(6):555-565.
39. Kleine-Tebbe J, Wangorsch A, Vogel L, Crowell DN, Haustein U-F, Vieths S. Severe oral allergy syndrome and anaphylactic reactions caused by a Bet v 1-related PR-10 protein in soybean, SAM22. *Journal of Allergy and Clinical Immunology*. 2002;110(5):797-804.
40. Morrison KL, Weiss GA. Combinatorial alanine-scanning. *Curr Opin Chem Biol*. 2001;5(3):302-307.
41. Lin J, Bardina L, Shreffler WG, et al. Development of a novel peptide microarray for large-scale epitope mapping of food allergens. *J Allergy Clin Immunol* 2009; 124(2):315-22-322.e1-3.
42. Taliencio E, Kim SW. Epitopes from two soybean glycinin subunits are antigenic in pigs. *J Sci Food Agric*. 2013;93(12):2927-2932.
43. Martínez-Botas J, Cerecedo I, Zamora J, et al. Mapping of the IgE and IgG4 sequential epitopes of ovomucoid with a peptide microarray immunoassay. *Int Arch Allergy Immunol*. 2013;161(1):11-20.
44. Vereda A, Andrae DA, Lin J, et al. Identification of IgE sequential epitopes of lentil (Len c 1) by means of peptide microarray immunoassay. *J Allergy Clin Immunol*. 2010;126(3):596-601.e1.
45. Weber RW. Patterns of pollen cross-allergenicity. *Journal of Allergy and Clinical Immunology*. 2003;112(2):229-239.
46. Valenta Lidholm, Niederberger Hayek, Kraft Gronlund. The recombinant allergen-based concept of component-resolved diagnostics and immunotherapy (CRD and CRIT). *Clin Exp Allergy*. 1999;29(7):896-904.
47. Deak PE, Vrabel MR, Kiziltepe T, Bilgic B. Determination of Crucial Immunogenic Epitopes in Major Peanut Allergy Protein, Ara h2, via Novel Nanoallergen Platform. *Sci Rep*. 2017;7(1):3981.
48. Sampson HA. Utility of food-specific IgE concentrations in predicting symptomatic food allergy. *J Allergy Clin Immunol*. 2001;107(5):891-896.
49. Berneder M, Bublin M, Hoffmann-Sommergruber K, Hawranek T, Lang R. Allergen chip diagnosis for soy-allergic patients: Gly m 4 as a marker for severe food-allergic reactions to soy. *Int Arch Allergy Immunol*. 2013;161(3):229-233.
50. Griffiths LLM, El-Shanawany T, Jolles SRA, et al. Comparison of the Performance of Skin Prick, ImmunoCAP, and ISAC Tests in the Diagnosis of Patients with Allergy. *Int Arch Allergy Immunol*. 2017;172(4):215-223.
51. Blankestijn MA, Knulst AC, Knol EF, et al. Sensitization to PR-10 proteins is indicative of distinctive sensitization patterns in adults with a suspected food allergy. *Clin Transl Allergy*. 2017;7:42.
52. Ferreira F, Hauser M, Egger M, Wallner M, Wopfner N, Schmidt G. Molecular Properties of Plant Food Allergens: A Current Classification into Protein Families. *TOIJ*. 2008;1(1):1-12.
53. Hoffmann-Sommergruber K, Mills ENC. Food allergen protein families and their structural characteristics and application in component-resolved diagnosis: new data from the EuroPrevall project. *Anal Bioanal Chem*. 2009;395(1):25-35.
54. Bernhiselbroadbent J, Sampson H. Cross-allergenicity in the legume botanical family in children with food hypersensitivity. *J Allergy Clin Immunol*. 1989;83(2):435-440.
55. Eigenmann PA, Burks AW, Bannon GA, Sampson HA. Identification of unique peanut and soy allergens in sera adsorbed with cross-reacting antibodies. *J Allergy Clin Immunol*. 1996;98(5):969-978.
56. Cox L, Williams B, Sicherer S, et al. Pearls and pitfalls of allergy diagnostic testing: report from the American College of Allergy, Asthma and Immunology/American Academy of Allergy, Asthma and Immunology Specific IgE Test Task Force. *Ann Allergy Asthma Immunol*. 2008;101(6):580-592.
57. Verstege A, Mehl A, Rolinck-Werninghaus C, et al. The predictive value of the skin prick test wear size for the outcome of oral food challenges. *Clin Exp Allergy*. 2005;35(9):1220-1226.
58. de Weck AL, Sanz ML. Cellular allergen stimulation test (CAST) 2003, a review. *J Investig Allergol Clin Immunol*. 2004;14(4):253-273.
59. Jutel M, Jaeger L, Suck R, Meyer H, Fiebig H, Cromwell O. Allergen-specific immunotherapy with recombinant grass pollen allergens. *J Allergy Clin Immunol*. 2005;116(3):608-613.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kern K, Havenith H, Delaroque N, et al. The immunome of soy bean allergy: Comprehensive identification and characterization of epitopes. *Clin Exp Allergy*. 2019;49:239–251. <https://doi.org/10.1111/cea.13285>

List of Tables

Table S1: Clinical characteristics of patient sera.....	2
Table S2: Table of identified, potential epitope candidates in all investigated soybean proteins.....	4
Table S3: Comparing PEPperPRINT array and phage display experiments	21
Table S4: Summary of the microarray experiments.....	22

List of Figures

Figure S1: Atlas of potential epitopes in Gly m 2.....	25
Figure S2: Atlas of potential epitopes in Gly m 4.....	26
Figure S3: Atlas of potential epitopes in Gly m 5.....	27
Figure S4a-d: Epitope fingerprint indicating most important amino acids.	28
Figure S5: IgE and IgG binding patterns.	30

Table S1: Clinical characteristics of patient sera

Patient code	SPT Soydrink in mm	SPT Birch in mm	Total IgE in kU/l	Symptoms	ASIT
13001	5	6	86.4	not known	no
13002	5	7	124	not known	no
13003	7	4	88	moderate	birch, hazel, alder
13004	5	5	26.9	not known	birch, hazel, alder
13005	5	3	-	tolerance	no
13006	3	8	>5000	not known	not known
13007	6	7	964	severe	not known
13008	8	7	302	tolerance	not known
13009	8	0	>5000	tolerance	not known
13010	10	11	248	tolerance	birch
13011	9	8	127	mild	birch
13012	10,5	10	75.5	mild	birch
13013	-	12	-	mild	not known
13014	9	4	456	tolerance	birch
13015	3	5	1174	not known	not known
13016		8	336	moderate	birch, hazel, alder, grasses
13017	4	7	272	tolerance	no
13018	5	6	260	not known	no
13019	5	5	-	tolerance	not known
13020	10	13	78.2	tolerance	grasses, rye
13021	5	7	261	tolerance	birch, hazel, alder
13022	5	7	>100	not known	no
13023		14,5	872	mild	birch
13024	11	11,5	220	moderate	birch
13025	9	7,5	859	tolerance	birch
14001	5	5	72.3	tolerance	not known
14002	9	13	157	mild	birch
14003	5	5	56	mild	no
14004	5	8	650	moderate	birch, hazel, alder, grasses, rye
14005	10	11	68.1	tolerance	birch
14006	5	7	60.7	tolerance	birch, hazel, alder
14007	5	-	-	moderate	not known
14008	-	-	>5000	not known	not known
14009	4	6	49.4	severe	not known
14010	6	-	-	not known	not known
14011	3	10	51	tolerance	birch
14012	5	6	117	moderate	birch
14013	5	-	2427	not known	no
14014	4	9	31.9	not known	no
14015	-	-	>5000	mild	birch, hazel, alder
14016	2,5	8	309	mild	birch

14017	-	-	206	tolerance	not known
14018	-	-	>5000	tolerance	not known
14019	8,5	-	-	tolerance	not known
14020	3	-	-	mild	not known
14021	11	-	138	mild	no
14022	-	-	>5000	tolerance	not known
14023	-	-	-	not known	not known
14024	-	-	>5000	tolerance	not known
14025	-	-	134	mild	birch, hazel, alder, house dust mite

Table S2: Table of identified, potential epitope candidates in all investigated soybean proteins. Epitopes occurring in several subunits are listed only once.

Nr.	Epitope	Position	Homology in other proteins	IEDB ID published epitopes
GenBank ID / Protein: A29385/ Gly m 1				
1	RPSCPD	05-11		
2	LGGSLGTV	18-25		
3	VDDCC	25-29		
4	CCALIG	28-33		
5	IEAIVCLC	38-45		
6	LCIQLRA	44-50		
7	CGRSY	67-71		
8	RSYPSNA	69-75		
9	NATCPR	74-79		
GenBank ID / Protein: Z13956/ Gly m 2				
10	RKSCGF	04-09		
11	FASQVV	18-23		
12	SQVVVQTEGRVCESQS	20-35		
13	SQSHGFH	33-39		
14	HGLCNRDH	39-46		
15	RDHNCALVC	44-52		
16	LVCERNE	50-55		
17	EGFSGG	55-60		
18	SRRCF	65-69		
GenBank ID / Protein: AJ223981/ Gly m 3.01				
19	LLCGI	11-15		
20	AAIIGQD	23-29	AAIIGQD/ Gly m 3.02	
21	DGSVWLQS	29-36	DGSVW / Gly m 3.02	
22	TPGQCN	111-116	TPGQCN / Gly m 3.02	
23	GDYLID	123-128	GDYLID / Gly m 3.02	

24	IDQGY	127-131	IDQGY / Gly m 3.02	
GenBank ID / Protein:AJ223982/ Gly m 3.02				
25	DGSVWAQS	29-36	DGSVW / Gly m 3.01	78563
26	VERPGDYL	119-126		
GenBank ID / Protein: x60043/ Gly m 4				
27	EINSP	09-13		
28	NSPVA	11-15		
29	PVAPA	13-17		
30	KSVENV	39-44		
31	ENVEG	42-46		
32	FLEDGET	58-64		
33	GETKFV	62-67		
34	LGYSY	79-83		
35	SYSVV	82-86		
36	VVGGA	85-90		
37	NGGSAGK	109-115		
38	YETKG	120-124		
39	GDAEP	124-128		
40	AEPNQDE	126-132		
41	QDELK	130-134		
GenBank ID / Protein: X17698/ Gly m 5.01				
42	FLASVSVSFGI	15-25	FLASVSVSFGI / Gly m 5.02, Gly m 5.03	
43	YWEKE	27-31	YWEKE / Gly m 5.03	
44	CLQSCN	38-43	CLQSCN / Gly m 5.03m, CLQSC / sucrose binding protein	
45	SERDSY	44-49	SERDSY / Gly m 5.02	
46	SYRNQACHC	48-56	SYRNQACHC / Gly m 5.02, Gly m 5.03	
47	ARCNLL	56-61	ARCNL / Gly m 5.02, Gly m 5.03	
48	EKEECEEG	64-71	EKEECEEG / Gly m 5.03; EECEE / Gly m 5.02	181299

49	EIPRPR	72-77	EIPRPR / Gly m 5.03, IPRPR / Gly m 5.02	181299
50	RPRPQHPE	77-84	RPRPQHPE / Gly m 5.02, Gly m 5.03	181299
51	HPEREPQQPG	82-91	HPEREPQQPG / Gly m 5.03; HPERE / Gly m 5.02; EPQQP / Gly m 6.01	181299
52	QQPGEKE	88-94	QQPGEKE / Gly m 5.03	181299
53	EEDEDE	94-99	EEDEDE / Gly m 5.03, EDEDE / Gly m 6.04, Gly m 6.05	181432
54	EDEQPRP	97-103	EDEQPRP / Gly m 5.03	
55	PRPIPFPR	101-108		181447
56	PFPRPQ	105-110	PFPRPQ / Gly m 5.03; PFPRP / Gly m 5.02	181432, 181447
57	PRQEEEH	111- 117	PRQEEEH/ Gly m 5.02, Gly m 5.03; PRQEE / Gly m 6.04	181447
58	EEEHEQ	114-119	EEEHEQ / Gly m 5.02, Gly m 5.03	181447
59	EHEQREEQE	116-124		181334, 181447
60	SEEEDEDE	137-144	EEEDE/ Gly m 6.01, sucrose binding protein; EEEDEDE / Gly m 5.03; EDEDE / Gly m 6.04, Gly m 6.05	181334, 181432
61	EDDEQDERQFP	144-154	EDDEQDERQFP / Gly m 5.03,EDDEQDER / Gly m 5.02, EEQDE / sucrose binding protein	
62	RPPHQKEER	157-165	RPPHQKEER / Gly m 5.03; PHQKEE / Gly m 5.02	181432
63	ERNEEE	164-169		181334, 181432, 11732
64	EEEDEDEE	167-174	EEEDEDE / Gly m 5.03; EEEDE / Gly m 6.01, sucrose binding protein; EDEDE / Gly m 6.04, Gly m 6.05	181288, 181334, 181432, 11305, 11732
65	EDDEQQ	171-176	EDDEQ / Gly m 5.02, Gly m 5.03	181288, 181432, 11305
66	EQQRESE	174-180		181288, 181432, 11305
67	ESEESEDS	178-185	ESEESEDS / Gly m 5.03	181288, 11305
68	DSELRR	184-189	DSELRR / Gly m 5.03	181288, 57535
69	LFGSNR	197-202	LFGSNR / Gly m 5.03	32585, 181406
70	RFETLF	202-207	RFETLF / Gly m 5.03	32585, 45674, 181406
71	KNQYGRI	208-214	KNQYG / Gly m 5.02; KNQYGRI / Gly m 5.03	45674, 181406, 45646

72	QRFNQ	218-222	QRFNQ / Gly m 5.03	38973, 45646
73	SPQLQN	224-229	SPQLQN / Gly m 5.03	38973
74	YRILEFN	233-239	YRILEFN / Gly m 5.02, Gly m 5.03	181386
75	NSKPN	239-243	NSKPN / Gly m 5.02, Gly m 5.03	181386
76	SYRLQSGDA	276-284	OSGDA / Gly m 5.02; SYRLQSGDA / Gly m 5.03	181474, 9963
77	VPSGTT	287-292	VPSGTT / Gly m 5.03	181474
78	TTYVV	291-296	TTYVV / Gly m 5.02, Gly m 5.03	
79	FLSSTEAAQSYLQGFSRNIL	323-342	FLSSTEAAQSYLQGFSRNIL / Gly m 5.03; FLSST, AQQSYLQGFS / Gly m 5.02	181316, 17831
80	LEASYDT	342-348	LEASYDT / Gly m 5.02, Gly m 5.03	17831
81	FEEIN	350-354	FEEIN / Gly m 5.02, Gly m 5.03	27677, 227839
82	VLFSREEGQQQGEQR	356-370	VLFSREEGQQQGEQR / Gly m 5.03; GQQQGE / Gly m 5.02	11655, 27677, 227839,
83	QRLQESVIVE	369-378	RLQESVIVE / Gly m 5.02, Gly m 5.03; SVIVE / β -Amylase	11655
84	ISKEQIRA	379-386	ISKEQIRA / Gly m 5.03	227835, 227858, 181307
85	AKSSSRKT	391-398	AKSSSRKT / Gly m 5.02; Gly m 5.03	181307; 227829, 227834
86	NLRSRDP	407-413	NLRSRDP / Gly m 5.02, Gly m 5.03	181455, 227814, 227847
87	PQLRD	431-435	PQLRD / Gly m 5.02, Gly m 5.03	181292, 227769, 227787, 227812
88	IFLSIVD	438-444	IFLSIVD / Gly m 5.03	
89	HFNSK	454-458	HFNSK / Gly m 5.02, Gly m 5.03	2716, 227784
90	LKEQQQE	477-483	LKEQQQE / Gly m 5.03, KEQQQ / Gly m 5.02	13307
91	QEEQPL	486-491	QEEQPL / Gly m 5.02, Gly m 5.03	
92	EVRKYRAELSE	491-502	EVRKYRAELSE / Gly m 5.02, Gly m 5.03	
93	QDIFVIPAGYPVVVN	503-517	QDIFVIPAGYPVVVN / Gly m 5.02, Gly m 5.03	26180
94	ATSNLN	518-523	ATSNLN / Gly m 5.03	227740
95	RNFLA	536-540	RNFLA / Gly m 5.02, Gly m 5.03	181403, 27590
96	GSQDNVISQIPS	541-552	GSQDNVISQIPS / Gly m 5.03; ISQIPS / Gly m 5.02	
97	SQVQELAF	552-559	SQVQELA / Gly m 5.02, Gly m 5.03	227813, 227819

98	FPGSAQAVE	559-567	FPGSAQVE/ Gly m 5.03, FPGSA/ Gly m 5.02,	227819, 22286
99	NQRESYFVDAQPKK	572-585	NQRESYFVDAQPKK / Gly m 5.03; ESYFVDAQP / Gly m 5.02	181434, 227756, 227797
100	PLSSILR	596-602	PLSSILR / Gly m 5.02, Gly m 5.03	181434
GenBank ID / Protein: M13759/ Gly m 5.02				
101	YWEKQ	27-31		
102	QNPSHNK	31-37		
103	RSCNSE	40-45	SCNSE / Gly m 5.01, Gly m 5.03	
104	KVEEEEECEEGQ	62-73	VEEEE / sucrose binding protein; EEECEEGQ / Gly m 5.0101; EECEEGQ / Gly m 5.03	11655, 27677
105	RPRPQHPERERQQHGEK	76-92	RPRPQHPERE / Gly m 5.01, Gly m 5.03	181299
106	EDEGEQPR	94-101		
107	PRPRQ	106-110	PRPRQ / Gly m 5.0101, Gly m 5.03, Gly m 6.04	
108	QPHQEE	110-115		
109	EEEHEQ	114-119	EEEHEQ / Gly m 5.0101, Gly m 5.03	181447
110	EQKEEHE	118-124	EQKEE / Gly m 5.0101, Gly m 5.03	181432, 227821
111	KEEKHG	128-133		
112	SEEEQDERE	137-145	SEEEQDER / Gly m 5.0101, Gly m 5.03; EEEQDE / sucrose binding protein	
113	EHPRPH	145-150	EHPRPH / Gly m 5.0101	
114	QPHQKEEE	151-158	PHQKE / Gly m 5.01, Gly m 5.03	181334, 181432, 11732, 17285
115	HEWQHKQEKH	160-169		
116	ESEEE	173-177	ESEEE / Gly m 5.0101, Gly m 5.03	
117	EEEDQDEDEEQK	177-188	DQDEDE / Gly m 6.04; DEDEE / Gly m 8, Gly m 5.0101, Gly m 5.03, Gly m 6.05	181432
118	ESQESEGSESQREP	190-203	ESEGS / Gly m BD 30K; SESQRE / Gly m 5.0101, ESQRE / Seed biotinylated protein	
119	HFNSKR	213-218	HFNSK / Gly m 5.01, Gly m 5.03	227784
120	RFQTLF	218-223	FQTLF / Gly m 5.03	
121	NKRSQQLQ	237-244		

122	LPHHAD	263-268	LPHHAD / Gly m 5.03	227782
123	SYNLQ	292-296		
124	LRVPAG	301-306		
125	FLSSTQAQQSYLQGFS	358-373	FLSST / Gly m 5.01; FLSSTQAQQSYLQGF / Gly m 5.03; AQQSYLQGF / Gly m 5.01	181316
126	GREEGQ	394-399	REEGQ / Gly m 5.01 , Gly m 5.03	
127	EERLQESVIVE	403-413	RLQESVIVE / Gly m 5.01, Gly m 5.03	11655
128	KQIREL	417-422		
129	TQRNPQLRD	461-469	NPQLRD / Gly m 5.01, Gly m 5.03	181292, 227769, 227787, 227812
130	VFLSVVD	472-478		
131	EGALF	481-485		
132	INEGEAN	499-505		
133	IKEQQQ	511-516	IKEQQQ / Gly m 5.03; KEQQQ / Gly m 5.01	
134	QEEQPL	520-525	QEEQPL / Gly m 5.0101, Gly m 5.03	
135	NATSD	551-555		
136	FPRSAK	593-598		
137	KSQSES	605-610		
138	AQPQQ	615-619	AQPQQ / Gly m 5.03	227797
GenBank ID / Protein: S44893/ Gly m 5.03				
139	FLASVCVS	16-23	FLASV / Gly m 5.01	
140	YFRSSNSFQTLFEN	35-48	FQTLF / Gly m 5.02	227751, 227798
141	QNVRIPLLQ	49-57		227822
142	SPQLEN	63-68		227764, 227832
143	ENLRD	67-71		227764
144	YRIVQFQSK	72-80		
145	FVLSGRAIL	97-105		227775
146	SYNLHP	115-120		227783
147	DAQRIP	122-127		227776

148	YLVNPHDHQN	133-142		227772, 227805
149	KPGRYD	154-159		227744, 227790
150	FLSSTQAQQSYLQGFS	162-177	FLSST, AQQSYLQGFS / Gly m 5.01; FLSSTQAQQSYLQGFS / Gly m 5.02	181316
151	LETSFHSEFEEINR	181-194	FEEIN / Gly m 5.0101, Gly m 5.02	227820, 227839
152	NRVLFG	193-198		227839
153	LFGESEE	196-202		
154	QRQQEGVI	203-210		227835
155	VELSKEQI	211-218	SKEQI / Gly m 5.01	227835, 227858
156	IFLSSVDIN	272-280		227812
157	VILVINE	295-301	VILVINE / Gly m 5.0101; LVINE / Gly m 5.02	227765
158	AYPFVV	345-350		227740
GenBank ID / Protein: AB113349/ Gly m 6.01				
159	KPFQCAGVA	60-68	KPFQCAGVA / Gly m 6.02, Gly m 6.03; QCAGV / Gly m 6.05	187829, 32708
160	RPSYT	81-85	RPSYT / Gly m 6.02, Gly m 6.03	54971
161	NGPQE	86-90	NGPQE / Gly m 6.02	44051, 54971, 76131
162	EIYIQQG	90-96	EIYIQQG / Gly m 6.02, Gly m 6.03	44051, 76131, 74358
163	IYPGCPSTFEE	103-113	PGCPSTFE / Gly m 6.02, Gly m 6.03; PGCP / Gly m 6.04, Gly m 6.05, Gly m TI; PSTF / Lectin; TFEE / Gly m 6.04	16009, 52031
164	EPQQPQQRG	113-121	EPQQP / Gly m 5.0101, Gly m 5.03	
165	GQSSRPQD	121-128	GQSSRPQD / Gly m 6.02, Gly m 6.03	
166	PQDRHQ	126-131	PQDRHQ / Gly m 6.02, Gly m 6.03	
167	IAVPTG	142-147	IAVPTG / Gly m 6.02, Gly m 6.03	
168	ENQLDQ	172-177	ENQLDQ / Gly m 6.02; NQLDQ / Gly m 6.03, Gly m 6.04, Gly m 6.05	45574
169	PRRFY	179-183	PRRFY / Gly m 6.02, Gly m 6.03	45574
170	GNQEQ	186-190	GNQEQ / Gly m 6.02, Gly m 6.03	53180
171	FLKYQQE	192-198	FLKYQQ / Gly m 6.02	13815, 37068

172	QQEQGGHQ	196-203		
173	HQSQKG	202-207	QSQKG / Gly m 6.02, Gly m 6.03	52128
174	KHQQEE	208-213	HQQEE / Gly m 6.04	
175	GFTLEF	223-228		19632, 20021, 58037, 187850
176	FLEHA	228-232	FLEHA / Gly m 6.02	19632, 58037, 187850
177	EHAFSVD	230-236		58037
178	NLQGE	242-246	NLQGE / Gly m 6.02	
179	QGENE	244-248	QGENE / Gly m 6.02, Gly m 6.03	
180	LSVIKP	262-267		
181	PTDEQQQ	268-274		
182	QRPQEE	274-279		
183	EEEEDEK	281-287	EEEEED / Gly m Gly m 5.02, sucrose binding protein, EEEEDE / Gly m 5.0101, Gly m 5.03, sucrose binding protein	
184	EKPQC	286-290		
185	KDKHC	293-297	KDKHC / Gly m 6.02, Gly m 6.03	
186	ETICT	314-318	ETICT / Gly m 6.02, Gly m 6.03	
187	FVPHYNL	370-374	FVPHYNL / Gly m 6.02, Gly m 6.03	41548, 59438
188	FLVPPQESQK	482-491	FLVPP / Gly m 6.02, Gly m 6.03, VPPQESQ / Gly m 6.02	
GenBank ID / Protein: D00216/ Gly m 6.02				
189	QQEQGGHQ	12-18	LFSGC / Gly m 6.01, Gly m 6.03	6238, 32233
190	HQSQKG	19-31	QQNECQIQ / Gly m 6.01, Gly m 6.03	6238
191	KHQQEE	38-43	PDNRIE / Gly m 6.01, Gly m 6.03	
192	GFTLEF	41-46	RIESEG / Gly m 6.01, Gly m 6.03	187845
193	FLEHA	44-49	SEGGFI / Gly m 6.03	
194	EHAFSVD	100-108	IFPGCPS / Gly m 6.03; FPGCP / Gly m 6.05; PGCPST / Gly m 6.01; GCPST / Gly m 6.03	16009, 52031
195	NLQGE	110-116	PQESQ / Gly m 6.01	

196	QGENE	112-118	QQRG / Gly m 6.01	
197	LSVIKP	118-128	RPQDRHQ /Gly m 6.01, Gly m 6.03	
198	PTDEQQQ	128-133		
199	QRPQEE	142-147	PTGVAV / Gly m 6.01	
200	EEEEDEK	189-195	FLKYQQ / Gly m 6.01	13815, 37068
201	EKPQC	195-201		13815, 37068, 52128
202	KDKHC	199-205	QSQKG / Gly m 6.01, Gly m 6.03	37068, 52128
203	ETICT	209-214	EEENEG / Gly m 6.01, Gly m 6.03	60470
204	FVPHYNL	219-223	SGFAP / Gly m 6.03	58026
205	FLVPPQESQK	238-244	NLQGEN / Gly m 6.01; LQGEN / Gly m 6.03	
206	LFSGCFA	257-262		19919, 57973
207	LREQAQQNECQIQ	275-282		
208	PDNRIE	287-292		538733
209	RIESEG	313-317		
210	SEGGFI	317-322		
211	IFPGCPSTY	323-328	IYNPQA / Gly m 6.01	
212	EPQESQQ	345-352	LKLSAQY / Gly m 6.03; LSAQY / Gly m 6.04, Gly m 6.05	36948, 35160, 187962
213	QESQQRG	350-357	GSLRKN / Gly m 6.01; SLRKN / Gly m 6.03	36948, 59438
214	GRSQRPDHRHQ	369-374	NSIIY / Gly m 6.01, Gly m 6.03	41548, 43262, 45850, 70332,
215	QKVHRF	455-460	NLKSQQ / Gly m 6.01	51031, 69084
216	PTGVAV	460-464	QARQV / Gly m 6.03	33488, 51031
217	FLKYQQQ	384-390	VNCNGER / Gly m 6.01, Gly m 6.03	
218	QQQGGSQ	387-392	NGERVF / Gly m 6.01, Gly m 6.03	
219	GSQSQKG	394-400	GELQEG / Gly m 6.01, Gly m 6.03	
GenBank ID / Protein: X15123/ Gly m 6.03				
220	SGCCFAF	14-20	SGCCFAF / Gly m 6.01, Gly m 6.02	
221	CFAFSF	17-22	CFAFS / Gly m 6.01; CFAFSF, Gly m 6.02	
222	FSFREQPQ	20-27	FSFREQPQ / Gly m 6.02; REQPQ / Gly m 6.01	

223	QPQQNECQIQ	25-34	QPQQNECQIQ / Gly m 6.01, Gly m 6.02	
224	NRIESEG	43-49	NRIESEG / Gly m 6.01, Gly m 6.02	
225	YIQQGSIG	92-99	YIQQG / Gly m 6.01; YIQQGSIG / Gly m 6.02	44051, 74358, 76131
226	EPQQKG	113-118	EPQQK / Gly m 6.02	
227	QQKGQSSR	115-122	QQKGQSSR / Gly m 6.02; QQSSR / Gly m 6.01	
228	IYHFRE	130-135	IYHFRE / Gly m 6.02	
229	IAVPTGF	139-145	IAVPVG / Gly m 6.01; IAVPTGF / Gly m 6.02	
230	FQNQLDQ	168-174	FQNQLDQ / Gly m 6.02; NQLDQ / Gly m 6,0101, Gly m 6,0301, Gly m 6,0401, Gly m 6,0501	45574
231	EQEFL	186-190	EQEFL / Gly m 6.01, Gly m 6.02	45544, 13815, 53810
232	FLQYQPQ	189-195	FLQYQPQ / Gly m 6.02	
233	GLSVISP	260-266	GLSVISP / , Gly m 6.02, Gly m 6.04, Gly m 6.05; LSVIS / Gly m 6.01	18633, 187789, 187855
234	PTTEEQQ	266-272	PTTEEQQ / Gly m 6.02	
235	RPEEEKPP	274-281	RFEEKPP / Gly m 6.02	
236	EEKPDCDE	278-285	EEKPD / Gly m 6.02, Lea Protein, EKPDCDE / Gly m 6.02	
237	QSQSRN	291-296	QSQSRN / Gly m 6.02	
238	TSSPDIF	314-320	TSSPDIF / Gly m 6.01, Gly m 6.02, TSSPDI / Gly m 6.01	
GenBank ID / Protein: X02626/ Gly m 6.04				
239	LEPDH	40-44	LEPDH / Gly m 6.05	
240	SQHPEL	58-63	SQHPEL / Gly m 6.05	187903
241	GVTVSK	67-72	GVTVSK / Gly m 6.05	187903
242	NGLHSP	78-83		
243	KQQLQD	128-133	QQLQD / Gly m 6.05	
244	NEGDVLV	142-148	NEGDVLV / Gly m 6.05	
245	QGQHQQ	215-220	QGQHQQ / Gly m 6.05	
246	QDEDE	279-284	QDEDE / Gly m 5.02	

247	GQDEDE	335-340	QDEDE / Gly m 5.02	
248	PRQEEPR	363-369	PRQEE / Gly m 5.0101, Gly m 5.02, Gly m 5.50301	
249	ARPSRA	394-399	ARPSRA / Gly m 6.05	187752
250	QFQLSAQY	422-429	LSAQY / Gly m 6.02, Gly m 6.05	187962, 35160, 36948
251	AVTSY	507-511		
252	GPLVN	545-549	GPLVN / Gly m 6.05	
GenBank ID / Protein: M10962/ Gly m 6.05				
253	ECQLNN	32-37	ECQLNN / Gly m 6.04	
254	VESEGG	47-52	VESEGG / Gly m 6.04; ESEGG / Gly m 6.01, Gly m 6.02, Gly m 6.03, Seed biotinylated protein	
255	NRNGSH	77-82		
256	NGSHLPS	79-85		
257	LPSYLPY	83-89		
258	YLPYPQ	86-91		
259	AFPGCP	105-110	FPGCP / Gly m 6.02, Gly m 6.03	16009, 52031
260	GCPETFE	108-114	GCPETFE / Gly m 6.04	
261	HFNEG	140-144	HFNEG / Gly m 6.04	
262	LVIPLG	147-152		
263	EPVVAI	163-168	EPVVAI / Gly m 6.04	
264	DIEHPET	194-200		
265	QEEEGG	221-226	EEEGG / Gly m 6.04	
266	EGGSV	224-228	EGGSV / Gly m 6.04	
267	KLRSPDD	250-256		
268	GGLSVISP	266-273	GGLSVI / Gly m 6.0101; GGLSVISP / Gly m 6.02, Gly m 6.03, Gly m 6.04	18633, 187789, 187855
269	SPKWQ	272-276	SPKWQ / Gly m 6.04	187855
270	WQEQEDED	275-282		
271	DEDEEYG	284-290	DEDEE / Gly m 8, Gly m 5.0101, Gly m 5.02, Gly m 5.03	

272	TPSYPP	292-297	TPSYP / β - Amylase	
273	RPSHGKH	299-305	RPSHGK / Gly m 6.04	187775
274	EDEDE	309-313	EDEDE / Gly m 5.0101, Gly m 5.03, Gly m 6.04	
275	EDEEEDQPRP	311-320	EDEEE / Gly m 8, Gly m BD28K; EEEDQ / Gly m 5.02; EDQPR / Gly m 6.04	
276	QPRPDH	317-322		
277	PDHPPQ	320-325		
278	QRPSRPEQQE	325-334	RPSRP / Gly m 6.04	
279	QQEPRG	332-337		
280	RGRGCQ	336-341		
281	GCQTRN	339-344		
282	AVFDGE	433-438	AVFDGE / Gly m 6.04; VFDGE / Gly m 6.01, Gly m 6.02; Gly m 6.03	
283	THHNA	469-473	THHNA / Gly m 6.04	
284	VSSYIK	474-479		
285	LSNSYN	490-495		
286	NSYNLG	492-497		
287	NSGPLVN	509-515	GPLVN / Gly m 6.04	
GenBank ID / Protein: U59626/ Gly m 7				
288	KEIHVEKHR	16-24		
289	KHRVPK	22-27		
290	HAGKA	67-71		
291	KESQREA	110-116	KESQR / Gly m 6.02, Gly m 6.03; ESQRE / Gly m 5.01, Gly m 5.02	
292	GGQVVAEK	163-170		
293	HVVEGA	388-393		
294	EEAQREL	432-438		
295	QPQEAEERP	445-453		
296	EERPSEGIG	450-458		

297	GGESE	508-512		
298	VTGRA	630-634		
299	AKHEEG	634-639		
GenBank ID / Protein: AAB71140/ Gly m 8				
300	LFCIAHTCS	12-20		
301	SASKWQH	20-26		
302	QQDSCR	27-32		
303	SCRKQL	30-35		
304	KQLQGVN	33-39		
305	NLTPCEK	39-45		
306	QGRGD	52-56		
307	EDEEEEG	86-92	EDEEE /Gly m BD 28K, Gly m 6.05 ; EEEEG / Gly m 6.04	
308	QKCCT	95-99		
309	TEMSEL	99-104		
310	CKALQK	111-116		
311	NQSEEEEEK	120-128		
312	MCRFGP	143-148		
313	IQCDLS	150-155		
GenBank ID / Protein: S45092/ Gly m TI				
314	RAAPTG	54-59		
315	CVGIPT	110-115		
316	TEWSVV	115-120		
GenBank ID / Protein: K00821/ Gly m agglutinin				
317	QTHAGY	134-139		
318	FRNSWD	160-165		
319	SWDPP	163-167		
320	FASNLPH	260-266		

321	HASSNID	266-272		
322	DLTSF	275-279		
GenBank ID / Protein: J05560/ Gly m BD30K				
323	SSSSIST	16-22		37300
324	HRSILD	23-28		55861
325	QKQVSSLFQL	37-46	SLFQL / β - Amylase	39985, 61155
326	HEEEA	58-62		74125
327	ITPQEFSK	98-105		8803, 48987, 48988
328	KYLQAP	106-111		8803, 74873
329	DVSQQIK	113-119		50328, 60532
330	QYSCDH	129-134		
331	GVITQV	145-150		31559
332	EESEGCYNG	192-200	ESEGS / Gly m 5.02	14141
333	GWQYQS	200-205		
334	HYQSFWEV	202-209		14141, 44118
335	KYQGCGG	151-157		78253
336	QGGCGR	153-158		78253
337	LEHGGI	210-215		12334
338	DDYPYR	219-224		12334, 63127
339	EGRCKAN	227-233		75579, 53282
340	IDGYE	241-245		78118
341	IMSDES	248-253		78118
342	SVSIDAK	272-278		28777
343	GYGSAD	307-312		23875
344	GEDGYI	328-333		11472
345	NLLGVC	341-346		
GenBank ID / Protein: AAA33984/ 18kD Seed Maturation Protein				
346	IGASA	18-22	IGASA/ Lea protein	

347	ATVQEKAER	31-39	ATVQEKAER/ Lea protein	
348	ARDPVQ	42-47	ARDPVQ/ Lea protein	
349	ELATQK	49-54	ELATQK/ Lea protein	
350	QHNTA	71-75	QHNTA/ Lea protein	
351	GHGHHT	87-92	GHGHHT/ Lea protein	
352	GEYGQP	106-111	GEYGQP/ Lea protein	
353	HQTSA	115-119	HQTSA/ Lea protein	
354	HGTGQPTGH	123-131	HGTGQPTGH/Lea protein	
355	HVTEGV	131-136	HVTEGV / Lea protein	
356	VGSHPIG	137-143	VGSHPIG / Lea protein	
GenBank ID / Protein: AJ011935/ Albumin				
357	NGACSPFEV	23-31		
358	PPCRSRD	32-38		
359	RDCRC	37-41		
360	RCVPIGLF	70-47		
361	GLSSVA	56-61		
362	VDEHP	64-68		
363	HPNLCQSD	67-74		
364	QSDDEC	72-77		
365	GNFCARYP	83-90		
366	SDSEA	102-106		
GenBank ID / Protein: AB084260/ Basic 7S globulin				
367	ASILHY	02-06		
368	LSFSFL	13-18		
369	SDSVPIP	22-28		
370	VPIPQH	25-30		
371	VLPVQN	42-47		
372	QVPVLVD	66-72		

373	HLWVNCE	77-83		
374	NCEQHYS	81-87		
375	YSSKTYQA	86-93		
376	TYQAPFCH	90-97		
377	PFCHSTQCSR	94-103		
378	QCLSCPAASRPGCHKN	108-123		
379	NPITQ	131-135		
380	GELGQDVL	140-147		
381	IHATQGSTQQL	149-159		
382	TQQLGPL	156-162		
383	VTVPQF	163-168		
384	RNIQGVA	184-190		
385	LASHFGLQ	203-210		
386	GLQHQFT	208-214		
387	TSKGAL	222-227		
388	QQFHNQ	237-242		
389	VTPQGEYNVR	255-264		
390	VRVSSIR	263-269		
391	NQHSV	271-275		
392	ISTSTPH	294-300		
393	SLYQAFTQ	306-313		
394	QVFAQQLEKQ	312-322		
395	LEKQAQVKSV	319-328		
396	KSVAPF	326-331		
397	GLCFNSN	332-338		
398	NAYPSVDL	341-348		
399	VQAQPG	367-372		
400	VTCLGV	373-378		

401	QPRAEV	384-389		
402	GTRQLE	392-397		
403	ARSRVG	406-411		
404	FSTSSLHSH	412-420		

Table S3: Comparing PEPperPRINT array and phage display: Total numbers of identified epitopes

Protein	Peptide phage display	PEPperPRINT array
Gly m 1	9	2
Gly m 2	9	11
Gly m 3.01	6	4
Gly m 3.02	2	2
Gly m 4	15	2
Gly m 5.01	59	28
Gly m 5.02	38	27
Gly m 5.03	19	9
Gly m 6.01	30	26
Gly m 6.02	31	20
Gly m 6.03	21	20
Gly m 6.04	14	22
Gly m 6.05	35	39
Gly m 7	12	10
Gly m 8	14	*
Gly m TI	3	2
Gly m Agglutinin	6	5
Gly m BD30K	23	3
18kD SMP	11	*
Albumin	10	*
Basic 7S globulin	38	*
	405	232

*. = not on the PEPperPRINT array

Table S4: Summary of the microarray experiments. Spots were regarded positive, if the signal intensity was at least five times over background; b.t. means signal intensity below threshold for all sera. Peptides are based on the native soybean protein sequence (Pr) and/or phage displays sequence (Ph). Underlined amino acids in Ph-peptides are the identical to the native protein sequence. Underlined peptides bear “clinical relevant” epitopes.

Protein	Peptide origin	Peptide code	Peptide sequence	Position	Only IgE-positive	Only IgG-positive	IgE/IgG positive	In total	IEDB code
Beta amylase	Pr	<u>EN9</u>	PGIGDFQSYD	202-211	2	b.t.	b.t.	2	
	Ph	DL2	GVFPSPTNFHPY <u>GDFQI</u>	205-208	1	b.t.	b.t.	1	
Gly m 2	Pr	EN14	LLVFASQVVVQTEG	15-28	1	1	b.t.	2	
	Ph	<u>EN16</u>	EVSFASQVLIY	18-21	4	2	1	7	
	Pr	DO1	FASQVVVQTEGRVCES	18-33	2	b.t.	1	3	
Gly m 2	Ph	<u>EN15</u>	SANQVVVQTHV	21-25	1	2	2	5	
	Pr	EN11	VQTEGRVCESQSH	24-36	b.t.	b.t.	b.t.	0	
	Pr	DO23	QTEGRVCESQSHGFH	25-39	3	b.t.	b.t.	3	
	Ph	EN13	TEHTEGRVLVN	26-30	b.t.	b.t.	b.t.	0	
Gly m 2	Ph	EN12	QLQRVCESLHY	29-33	2	b.t.	b.t.	2	
	Pr	DO2	SQSHGFHGLCNRDHNCA	33-49	4	1	b.t.	5	
	Ph	EN10	VCHQSHGFGHH	34-38	2	2	b.t.	4	
Gly m 2	Ph	EL14	GFNL <u>CNRDR</u> PAP	42-45	3	3	1	7	
Gly m 5.01	Pr	EM19	KCLQSCNSERDSYRN	37-51	2	1	b.t.	3	
	Ph	<u>EM20</u>	AINQSCNSHPI	40-44	7	3	b.t.	10	
	Pr	<u>DN7</u>	QSCNSERDSYRNQ	40-52	2	b.t.	b.t.	2	
Gly m 5.01	Ph	EL11	GNANSERDSIIQ	43-48	b.t.	b.t.	b.t.	0	
Gly m 5.01	Pr	<u>EM18</u>	PRPRPQHPERE	76-86	18	1	3	22	181299
	Ph	EL10	GSPPRPQHPIDV	78-83	b.t.	b.t.	b.t.	0	181299, 181447
Gly m 5.01	Pr	DN1	QHPEREPQQPGEKEE	81-95	b.t.	b.t.	b.t.	0	181299
Gly m 5.01	Pr	EN4	<u>PRPQPRQ</u> EEE	107-116	b.t.	b.t.	b.t.	0	181447
	Ph	<u>DO19</u>	SHHDQPRQHA	110-113	9	4	1	14	181447
Gly m 5.01	Pr	DN2	PRQEEEEHEQREEQEW	111-125	b.t.	b.t.	b.t.	0	181447
Gly m 5.01	Pr	EM16	KRGEKGS <u>EEE</u>	131-140	b.t.	1	b.t.	1	181334
	Ph	EM17	VYAEKGS <u>EAY</u>	134-137	b.t.	b.t.	1	1	181334
Gly m 5.01	Ph	<u>DO21</u>	FFLS <u>STEAQ</u>	324-328	6	b.t.	2	8	181316

Gly m 5.01	Pr	<u>EM14</u>	KVLF <u>S</u> REEGQQQ	355-366	16	5	5	26	
	Ph	<u>EM15</u>	HIT <u>F</u> SREEGSI	358-363	10	2	b.t.	12	
Gly m 5.01	Pr	EM12	PFNL <u>R</u> SRDPIY	405-415	2	1	b.t.	3	181455
	Ph	EM13	ATY <u>L</u> RSRDQAV	408-412	3	1	b.t.	4	181455
Gly m 5.01	Pr	DN3	DANIELV <u>G</u> LKEQQQE	469-483	1	b.t.	b.t.	1	
	Ph	<u>DN5</u>	NIELV <u>G</u> LKEQQQRQQQ	471-482	b.t.	b.t.	3	3	
Gly m 5.01	Pr	EL15	GLKEQQQEQQQ	476-486	b.t.	b.t.	b.t.	0	
	Ph	EM11	QEGKEQQQPNV	478-482	2	1	2	5	
Gly m 5.02	Pr	EN19	YRNQACHARSN	49-59	1	1	b.t.	2	
	Ph	EN21	CDTYN <u>Q</u> ACHLAL	51-55	b.t.	b.t.	b.t.	0	
Gly m 5.02	Ph	EN20	CFNVIQACHANPC	52-56	b.t.	b.t.	b.t.	0	
	Pr	<u>EN17</u>	VEEEEECEEGQI	63-74	18	3	14	35	
	Ph	EN18	HIQEEEECEGDL	66-70	1	2	b.t.	3	
Gly m 5.02	Pr	DN4	DEGEQPRPFPPRPR	95-109	2	2	b.t.	4	
	Ph	EO13	TQHEGEQPFY	96-99	1	b.t.	b.t.	1	
Gly m 5.02	Pr	DN8	<u>N</u> KVLFSGREEGQQQG	354-367	1	2	b.t.	3	
	Ph	EO14	SDQNKVLQPH	389-392	b.t.	b.t.	b.t.	0	
Gly m 6.01	Ph	<u>EN5</u>	ISQQDRHRI	127-130	19	5	9	33	538745
Gly m 6.01	Pr	<u>EN2</u>	EFLKYQQEQG	191-200	3	b.t.	1	4	13815, 37068
	Ph	DO13	EFLKYQQQQQ	192-197	b.t.	1	b.t.	1	13815, 37068
	Ph	<u>DO11</u>	SDKYQEEFQPR	194-196	7	b.t.	2	9	13815, 37068
	Ph	DO12	TEKYQQNSSG	194-197	1	3	b.t.	4	13815, 37068
Gly m 6.02	Ph	DL3	MIFPGCPS	101-104	b.t.	b.t.	b.t.	0	16009, 52031,
Gly m 6.02	Pr	EM21	EDDD <u>D</u> EEEEQPQS	273-284	b.t.	b.t.	b.t.	0	
	Ph	<u>DO3</u>	GVYNSQVD <u>D</u> EEEQNQRD	276-281	3	b.t.	1	4	
Gly m 6.03	Pr	DN6	EPQQKGQSSRPQDRH	113-127	2	b.t.	1	3	538745
	Ph	DO22	DRHQKIYHF	128-131	2	b.t.	1	3	
Gly m 6.03	Pr	<u>DN9</u>	RRFYLAGNQE <u>Q</u> EFL	177-190	3	1	1	5	13815, 53810
	Ph	<u>EN6</u>	HVVE <u>Q</u> EFLD	186-190	17	3	8	28	13815, 53810
Gly m 6.05	Pr	DN10	NPDIHPETMQQ	192-203	b.t.	b.t.	b.t.	0	
	Ph	EN7	NQDDI <u>H</u> PEPI	194-199	1	b.t.	b.t.	1	

Gly m 8	Pr	<u>DO20</u>	KCQCKALQKI	108-117	3	b.t.	b.t.	3	561750
Gly m agglutinin	Pr	<u>DN11</u>	HAAANVVGNKESQREAR	101-117	6	b.t.	b.t.	6	
	Ph	EN8	FDQKESQHFS	110-113	1	1	b.t.	2	
Gly m BD 28K	Ph	<u>DO24</u>	QDQEEDEEDE	237-244	3	1	1	5	
	Ph	<u>DO4</u>	GYNPCRQEEDEELHHKC	239-244	4	4	3	11	
	Ph	DL7	GYVPCRQEEDEELHHKC	239-244	b.t.	b.t.	b.t.	0	
Gly m BD 30K	Pr	<u>EN3</u>	VYHNHEEEAK	54-63	11	1	3	15	74125
	Pr	<u>DO15</u>	YHNHEEEAKR	55-64	18	3	3	24	74125
	Ph	DO14	GENHEEQHVI	57-60	4	2	1	7	74125
Gly m BD 30K	Ph	<u>DO7</u>	GEFSKKYLQAP	105-108	13	1	1	15	8803, 48988
	Ph	<u>DO6</u>	GPVQKKYLFDS	105-108	9	5	b.t.	14	8803, 48988
Gly m BD 30K	Ph	<u>DO17</u>	GEYSCDITYT	130-133	20	4	12	36	
	Ph	DO18	KEQYSCDHPP	132-135	1	b.t.	b.t.	1	
Gly m BD 30K	Ph	DO16	TCDHPVKGQP	132-135	4	3	b.t.	7	
Gly m BD 30K	Pr	<u>EL13</u>	GYGSADGVVDYW	307-317	3	2	1	6	
	Ph	DL6	GVSSADGVGQINIDQN	310-314	b.t.	b.t.	b.t.	0	
Gly m KTI	Pr	<u>EN1</u>	IRFIAEGHPLSL	88-89	1	2	1	4	
	Ph	DO9	GSVQVIAEGHPHS	91-96	5	3	b.t.	8	
	Ph	DO8	GCVQVIAEGHPHC	91-96	3	1	b.t.	4	
	Pr	D010	AEGHPLSLKF	92-101	2	b.t.	b.t.	2	
Gly m KTI	Pr	EO12	QKLDKESLAK	198-207	b.t.	b.t.	b.t.	0	
	Ph	DL9	GATSSDRQLDKESPGYVI	200-204	b.t.	b.t.	b.t.	0	
Gly m KTI	Pr	<u>EM22</u>	SLAKKNHGLSR	204-214	2	1	b.t.	3	
	Ph	<u>DO5</u>	GTHFSKAVLGKKNHGDEF	207-211	7	2	1	10	
PR5	Pr	EL12	GRTNSTFDNAG	69-79	b.t.	1	b.t.	1	
	Ph	<u>DK3</u>	GVINCETFDQYQPIEIGC	72-76	5	1	1	7	

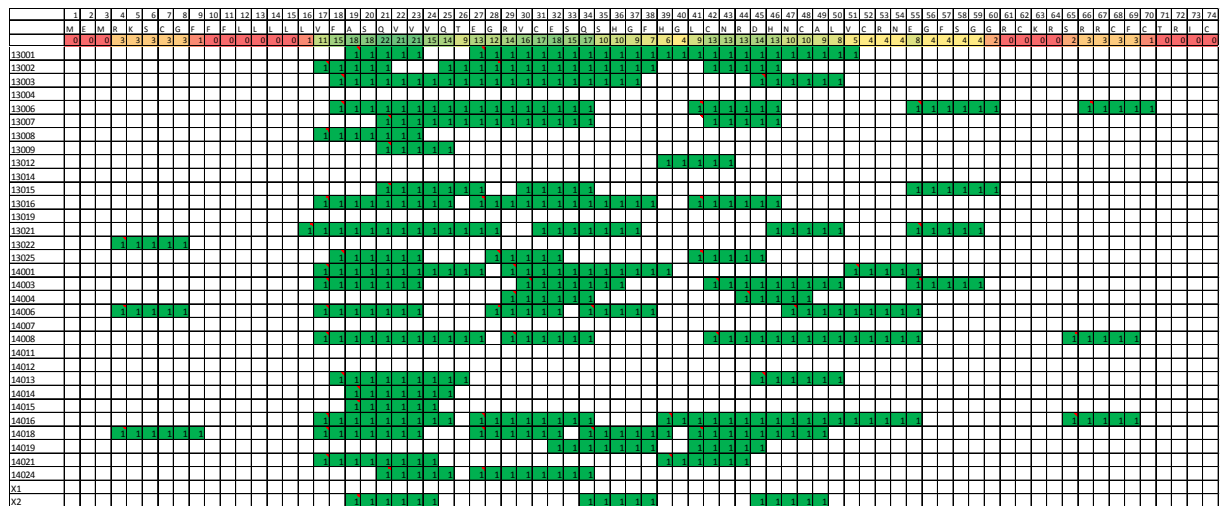
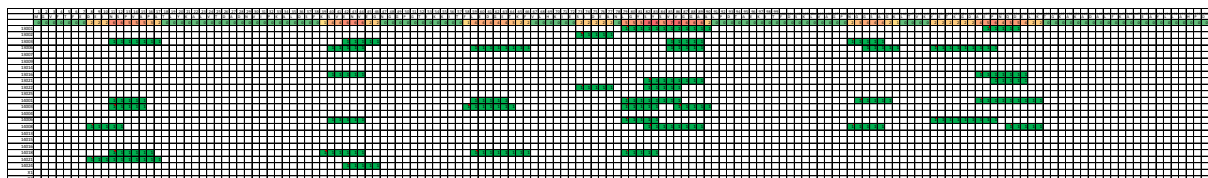


Figure S1: Atlas of potential epitopes in Gly m 2 identified by phage display. This is typical for all observed allergen maps. The x-axis displays the protein sequence of Gly m 2 (GenBank: Z13956). The serum donors are listed on the y-axis. Each line represents a single patient, the last two persons are not sensitized to soy. The spectrum of observed epitope regions is highly heterogeneous even amongst the people sensitized to soy, illustrating the complexity of the immune response as well as the value of the knowledge of individual epitopes for diagnosis and prognosis. First row: Sequence position; second row: amino acid; third row: frequency of the position in potential epitopes; green: epitope by peptide phage display.



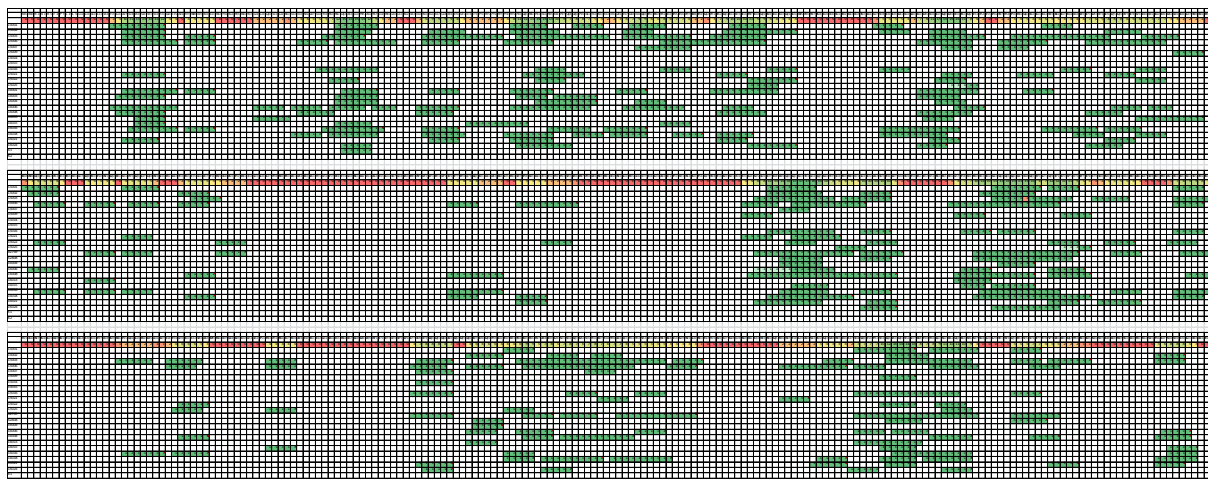


Figure S3: Atlas of potential epitopes in Gly m 5 identified by phage display. The second line displays the protein sequence of Gly m 5 (GenBank: X17698). Each line represents a single patient, X1/X2 are not sensitized to soy. The spectrum of epitope regions is highly heterogeneous, illustrating the complexity of the immune response as well as the value of the knowledge of individual epitopes for diagnosis and prognosis. First row: Sequence position; second row: amino acid; third row: frequency of the position in potential epitopes; green: epitope by peptide phage display.

