Supplemental Material

Supplemental Methods: Bioinformatics

Alignment algorithm selection: The optimal alignment parameters were determined by benchmarking the performances of several alignment algorithms (BWA (1), Bowtie2 (2), and Novoalign (Novocraft, Petaling Jaya Selangor, Malaysia), using a procedure with 1,000 iterations to identify differences between the sequences (l = 50, n = 10,000) and their reference. Perfectly-matched sequences were modified using a random number ($\leq$ 5) of all three types of mismatch events (substitution, deletion, and insertion) and were aligned relative to their reference. Different parameter values were tested, and the global performance was determined by counting the number of mapped sequences and the portion of differences that were appropriately identified. The parameter values that yielded the best performance were conserved. Results were then compared among tools, with the optimal parameters applied to each. The BWA was determined to be the best alignment tool evaluated.

Arachis hypogaea and Dermatophagoides pteronyssinus sequencing: Because of the lack of curated genome annotation, a seed transcriptome reference was created using RNA sequencing from seed tissue samples (NCBI Sequence Read Archive [SRA] project #PRJNA291488) for *Arachis hypogaea* and mites (#PRJNA339131). The SOAPdenovo (3) and Trinity (4) tools were used to assemble reads. Datasets were then mapped using Gmap (5) against the genomes of *Arachis duranensis* and *Arachis ipaensis* (two wild peanut species from which the *Arachis hypogaea* hybrid was derived) and *Dermatophagoides pteronyssinus*. Only full-length mapped sequences with a greater than 90% identity with one or both genomes were conserved for *Arachis hypogaea*, and only full-length mapped sequences with a greater than 90% identity with *Dermatophagoides pteronyssinus* genome were conserved. Sequence redundancies were removed using the Cuffcompare tool (6). Finally, 305,194 and 92,783 transcripts from *Arachis hypogaea* and *Dermatophagoides pteronyssinus*, respectively, were conserved and considered to be the reference sequences for the alignments. Transcript models were then matched against the NCBI non-redundant nucleic acids (NR) database (BlastX). Most inferred coding sequences were partial, but 18,088 of the predicted transcripts were mapped completely with their full-length homologs in the NR database for *Arachis hypogaea*; 16,514 sequences similarly were mapped for *Dermatophagoides pteronyssinus*. The TI analysis focused on this full-length coding sequence dataset.

Glycine max and Phaseolus vulgaris sequencing: RNA reference sequences were created from exon sequences of the reference genomes (GCF_000004515.4 and GCF_000499845.1, respectively). The exons were joined to create transcript reference sequences for each isoform of all known and predicted genes. Sequencing from seed- and whole-tissue RNA, respectively, was downloaded from the NCBI SRA database (projects #PRJNA291488 and PRJNA24236, respectively).

Preprocessing: FASTQ files were trimmed, and low-quality extremities were eliminated according to a quality threshold (Q = 30). Reads with remaining lengths lower than 20 bases were excluded.

Alignment and data filtering: The trimmed reads were aligned against reference sequences using BWA 0.5.8c. Only alignments with a greater than 90% identity and the best scores were assigned. Subsequently, reads with alignments that involved less than70% of their lengths were

excluded. The first and last six bases of each remaining alignment were excluded to avoid misalignments created by the algorithm. Only paired-end reads appropriately mapped to the same reference were conserved. The alignments were then parsed to quantify the number of reads that were identical to the reference sequence. The number of reads that deviated at any position from these reference sequences was also identified. Read positions with Illumina quality score < 30 (Q30) were not counted. The data were then used to calculate the rates of RDD (Fig. 1A) and identify single variations encoding TI proteins. Positions with more than 20% sequence variation were not used in the RDD rate calculation, as they might correspond to natural single-nucleotide polymorphisms.

The RDD gap rate was defined as the ratio of the number of differences to the number of reads to normalize for variations resulting from coverage differences at given positions. The RDD gap rate was calculated for each sequencing of peanut, soybean, and green bean. To improve data quality, only positions at which the error rate exceeded the machine error (1 error per 1,000 reads) were retained.

References for the Supplemental Material:

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–1760.

2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357–359.

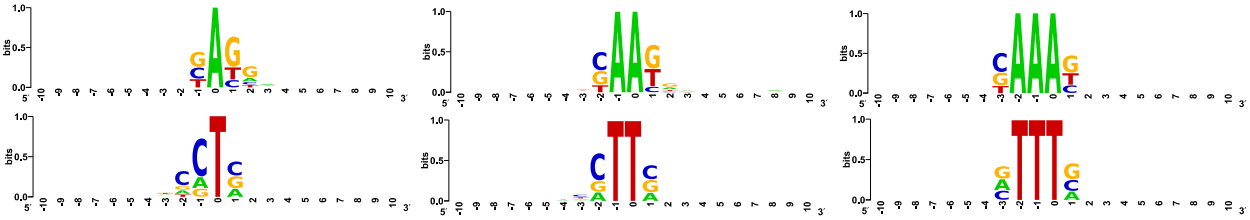
3. Luo R et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience.* 2012;1(1):18.

4. Grabherr MG et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol.* 2011;29(7):644–652.

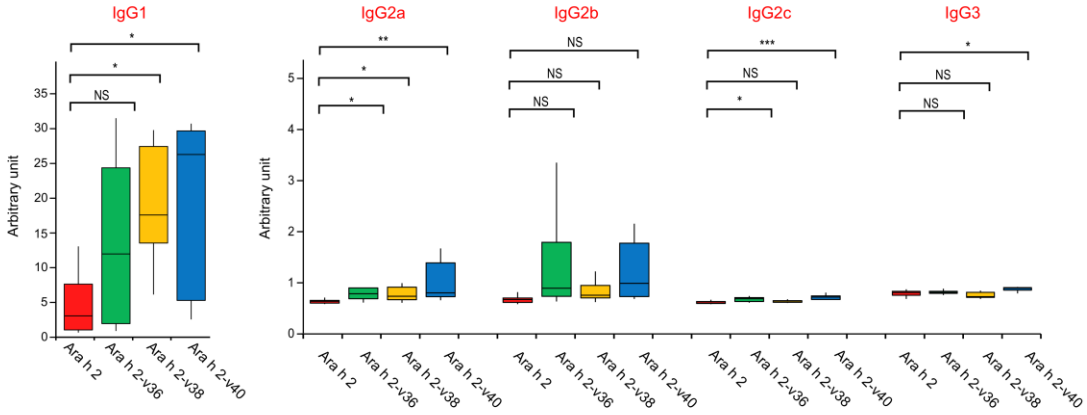
5. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics.* 2005;21(9):1859–1875.

6. Trapnell C et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012;7(3):562–578.

Supplemental Figure 1. A MEME (Multiple EM for Motif Elicitation) search showed that the ten bases surrounding (-10 to +10) the A and T repeats are not random (RDD gap position = b0). The height of the letter corresponds to the base frequency in the analyzed sequences.



Supplemental Figure 2. The observed effects of injections (at 1-week intervals) of the canonical Ara h 2 and three of its transcription infidelity variant on IgG were almost entirely due to the increase in IgG1 in mice. T-test, Welch T-test and Mann-Whitney tests with Holm-Bonferroni correction: *p < 0.05, **p < 0.01, ***p < 0.001. Box-and-whisker plot: box = interquartile range, bar = median, whiskers = 95% CI.

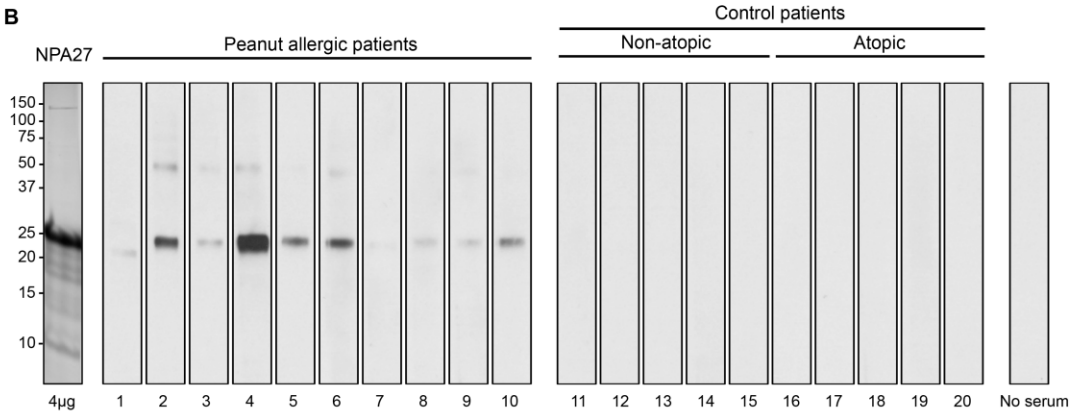


Supplemental Figure 3. Discovery of a peanut allergen (NPA) using their transcription infidelity profiles. (**A**) Genomic and protein sequence characteristics of NPA27. The table lists the known allergens with sequence similarity to NPA27 in the Allergome database. **(B)** First lanes on the left: Coomassie blue staining showing purified NPA27 protein (4 µg). The upper bands are believed to be NPA27 multimers. Center and rightmost lanes, IgE western immunoblotting with sera of peanut-allergic and -tolerant children (atopic and non-atopic controls). (**C)** Specific IgE directed towards recombinant NPA27 using sera of peanut-allergic (n = 52) and -tolerant patients (atopic and non-atopic subjects; n = 44) measured by ELISA.

A

Uniprot ID	Species	Allergen name	Identity %
B6CQT9	Prunus persica	Pru p 2	40
B6CQT8	Prunus dulcis	Pru du 2	40
Q5DWG1	Cryptomeria japonica	Cry j 3.5	42
C1K3P2	Pyrus pyrifolia	Pyr py 2	40

B



C