Article

Not peer-reviewed version

Linking Phenotypes to Protein Characteristics in 3D Structures Predicted by Alphafold

Atit Parajuli * , Robert Brueggeman , Steven Wagner , Marilyn Warburton , Michael Peel , Longxi Yu , Deven See , Zhiwu Zhang

Posted Date: 16 May 2023

doi: 10.20944/preprints202305.1102.v1

Keywords: Alfalfa; Plant Growth Vigor; Alphafold; Protein 3D Structure; Association Study

Article

Linking Phenotypes to Protein Characteristics in 3D Structures Predicted by Alphafold

Atit Parajuli [1,*], Robert Brueggeman [1], Steve Wagner [2], Marilyn L. Warburton [3], Michael Peel [4], Long-Xi Yu [3], Deven See [5] and Zhiwu Zhang [1]

[1] Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99164, USA
[2] Corteva Agriscience. West Salem. WI 54669, USA
[3] Plant Germplasm Introduction and Testing Research, USDA-ARS and Washington State University, Prosser, WA 99350, USA
[4] Forage and Range Research Lab, USDA-ARS and Utah State University, Logan, UT 84322, USA
[5] Wheat Health, Genetics, and Quality Research Unit, USDA-ARS and Washington State University, Pullman, WA 99164, USA
* Correspondence: atit.parajuli@wsu.edu

Abstract: Plant breeding aims to develop elite crop varieties appropriate for various environments with higher quality and quantity of production. Researchers use quantitative trait loci (QTL) mapping and association studies to identify regions in the genome responsible for the variation of the quantitative traits of interest. However, mapped regions do not always translate to functional proteins, which makes it challenging to identify genes associated with traits of interest. The biological functions of proteins are strongly dependent on their 3D structure. Alternatively, if proteins can be directly linked with the phenotypes, the effect of mutations on phenotypic changes can be assessed. Innovation of deep learning models in biology opens new avenues of exploration. AlphaFold is an AI system that predicts the 3D structure of a protein from its amino acid sequence with near experimental accuracy and was used in this study. Point mutations with a significant influence on the 3D structure of a protein can capture the effect on phenotypes through association study, and this provides insights into the regions that are of functional importance. In the current study, 534 plants were selected based on plant vigor, and 154 missense variants that change amino acid sequences, including 5 significant hits from previous study, were included. The changes in protein 3D structure were assessed by association with the phenotype. The analysis identified five significant associations, four of which were also identified in previous study of SNPs GWAS, however, a new fifth association was also identified which was annotated as disease resistance gene in *Medicago truncatula*. This study helps to associate SNPs that could be missed by GWAS due to stringent Bonferroni corrected p-values by providing a more robust filter for SNPs using features from predicted protein 3D structures.

Keywords: Alfalfa; Plant Growth Vigor; Alphafold; Protein 3D Structure; Association Study

Introduction

The ultimate goal of plant breeding is to develop elite crop varieties with improved production and quantity suitable for various environments. Traditional crop improvement methods rely on crossing chosen plants to incorporate desirable traits of interest into a new crop variety. Crop improvement programs have been using quantitative trait loci (QTL) mapping and association studies to identify regions of the genome associated with desirable traits of interest. These QTLs are responsible for variation in quantitative traits such as yield, quality, and biotic and abiotic stress tolerance. Identified QTLs become useful tools when they can be used in breeding programs to reduce the breeding cycle through a marker-assisted selection of genotypes of interest.

Mapped QTLs do not always translate to actual functional proteins as they are not always linked to genes or genetic variants associated with the traits of interest. This makes it challenging to identify genes that produce functional proteins associated with these traits. Additional studies such as transcriptomic or proteomic analysis, directed mutagenesis, or gene editing are generally required to confirm and validate the functional role of the genes or genetic variants identified by QTL mapping and association study. Additionally, rare genetic variants with minor allele frequency below 5% can also contribute to changes in the phenotypes. However, they are generally filtered to increase statistical power and reduce the false positive rate. To effectively identify mutations that affect phenotypes of interest, a more robust approach that includes the genetic variants prevalent in the population in lower frequency is crucial.

An alternative approach to associate proteins directly to the phenotype could help to assess the effect of mutation on phenotype if the mutation has a destabilizing effect on the protein. The biological function of the protein is strongly dependent on its 3D structure (Sotomayor-Vivas, Hernández-Lemus, and Dorantes-Gilardi 2022). Experimental processes involving the prediction of protein 3D structures have been painstaking and time-consuming, and past computational methods had very low prediction accuracy Thus, neither method has been reliably useful in the field of structural biology. With the integration of deep learning models, the AI system AlphaFold (Jumper et al. 2021) is able to predict the 3D structure of proteins with near experimental accuracy. This system is able to predict not only the domain structures but also the side chains and amino acid residues of protein models with high accuracy.

Mutations, including point mutations, can have significant consequences on the structure of the protein, ultimately affecting its function (Tóth-Petróczy and Tawfik 2014). Not all mutations lead to changes in the amino acid sequences. Synonymous mutations are benign as they do not change the sequence of amino acids in the polypeptide chain. Non-synonymous mutations lead to changes in the amino acid sequence and thus may change the 3D structure and the function of the protein (Tokuriki and Tawfik 2009). Proteins must fold to their active native state, with the lowest free energy, to assume a sTable 3D structure capable of functional activity ((Bai et al. 1995; Bai and Englander 1996; Levinthal 1968; Luheshi, Crowther, and Dobson 2008). Stability and folding are consequences of interactions among amino acid residues (Anfinsen 1973). Changes in one site of a protein are determined by interactions of the site with residues in nearby positions. Effects of mutations on protein stability are thus constrained, and protein structures are highly conserved during the evolution of protein homologs with similar structures and functions (Ashenberg, Gong, and Bloom 2013). Thus, changes in amino acids from mutations are non-random, and with complete information on protein 3D structure, it can be predictable.

With the atomic level information available from AlphaFold predicted 3D structure of tested proteins, features such as the local distance difference test (LDDT) and co-ordinates information between the mutated and the non-mutated protein can be used in an association study. We hypothesize that point mutations with a destabilizing effect on protein structure will cause significant differences in the backbone structure of the protein and in their neighboring residues. This is the first report of a study associating features from the predicted 3D structure of proteins and plant phenotype.

Material and Methods

Plant Materials

The project was funded by USDA, Alfalfa and Forage Research Program, to develop 200 inbred lines in collaboration with Corteva. The lines consisted of diverse alfalfa association panels (Prosser, Washington), breeding lines for drought resistance (Logan, Utah), and alfalfa breeding lines from Corteva (West Salem, Wisconsin). The plant materials were grown in a USDA, ARS greenhouse located on the Campus of Utah State University and Irrigated Agriculture Research and Extension Center at Prosser, Washington of Washington State University. For plant materials from Corteva, after germinating the seeds in flats, the most vigorous plants were transplanted to the field. All the transplanted plants were potted and brought into the greenhouse for self-pollination before flowering.

The plant materials from Utah were clone materials from three populations originating from distinct parental germplasm sources. These consisted of a diploid falcata population, a tetraploid falcata population, and a tetraploid sativa population. The falcata populations were used to develop drought-hardy materials for use on Western US grazing lands, while the sativa population was used for developing lines tolerant to saline conditions. The plant materials from Washington consisted of diverse alfalfa association panels selected for developing the drought-tolerant lines. Finally, the materials from Corteva were elite breeding lines selected for winter hardiness. During growth, the plants were not intentionally subjected to any stresses. These plants were irrigated at regular intervals and supplied with fertilizers when necessary.

The program started with 100 lines each from Washington and Utah, while 15 lines from Corteva. The lines were self-pollinated for three generations (S_0, $S_{1,}$ and S_2). In each generation of self-pollination, plants were paired with vigor traits (strong and weak) within lines. The information on plant samples, generations, and their vigor trait is presented in Table 1. The S_1 generation from Washington, S_1 generation and S_2 generation from Utah, and S_2 generation from Wisconsin were pooled into 42 pools. The pooling was based on health and the number of seeds produced. The strong individuals and weak individuals within the same generation were pooled together, similarly, for Wisconsin samples, individuals from the S2 generation were pooled based on the number of seeds and produced as top, low, and no seeded pools. Finally, from the 534 plants, 121 samples were generated that included individual plants as well as pooled ones.

Table 1. Information of Plants used in the study.

Location	Generation	Lines	Individual	Phenotyping	DNA Bulk
Washington	S1	21	42	Strong vs. weak Pair	3 strong and 3 weak pooled samples each with 7 individuals
Washington	S2	9	18	Strong vs. weak pair	No pooling, 18 individual samples
Utah	S1	19	38	Strong vs. weak pair	2 strong and 2 weak pools each with 9-10 individuals
Utah	S2	137	274	Strong vs. weak pair	11 strong and 12 weak pools each with 11-13 individuals
Wisconsin	S0	5	5	Seed count	No pooling, 5 individual samples
Wisconsin	S1	15	56	Seed count	No pooling, 56 individual samples
Wisconsin	S2	15	91	Seed count	3 top, 3 low and 3 no seeded pools each with 5-14 individuals
Total	3	534		NA	121 DNA samples (42 pools and 79 individuals)

Exome capture sequencing

The probe set for exome capture sequencing was developed from 112,626 contigs generated by *de novo* transcriptome assembly from two alfalfa subspecies, *M. sativa ssp. sativa* (B47) and *M. sativa* *ssp. Falcata* (F56) by (O'Rourke et al. 2015) using Illumina RNA-seq technology. The transcripts for this assembly were taken from roots, root nodules, leaves, flowers, elongated stem internodes, and post-elongating stem internodes. At the time of this library preparation, the complete genome annotation was not available and so the sequence contigs were used for probe development.

For DNA isolation, leaf tissues were collected and lyophilized in 96 deep-well microplates. DNA was extracted utilizing the oKtopure automated DNA extraction system (LGC, Biosearch Technologies, Teddington, Middlesex, UK). Libraries were prepared using the SeqCap EZ HyperCap Workflow and KAPA HyperPlus Library Preparation kit (ver. 2.3; Roche Sequencing Solutions, Inc. Pleasanton, CA) with the following modifications: starting DNA was normalized at 100 ng per sample; the pre-capture PCR amplification used seven cycles, and one µg per samples were hybridized for the exome capture. Sample libraries were purified using AMPure XP Beads (Beckman Coulter, Indianapolis, IN). Sequencing was outsourced to John Hopkins University, where 150-bp paired-end reads were generated with a total of around 4 billion reads for all 121 samples, which averages approximately 25 million reads per sample.

Sequence Alignment and Variant Calling

High-performance computing cluster Steptoe was used for data analysis. First, quality control of the reads was carried out using the FastQC program (Babrahman n.d.), and all the reads with a quality score between 20 and 40 were considered as high-quality reads. The next step involved removing adapter sequences as well as trimming low-quality sequences for which the BBMap function in BBduk (Bushnell 2014) was used. The high-quality trimmed reads were used for mapping to the alfalfa reference genome (Shen et al. 2020) using short read aligner bwa-mem (Li and Durbin 2009), which accurately maps paired-end reads to the reference genome passing more quality filters (Koboldt et al. 2012; Li and Durbin 2009; Yao et al. 2020)

The output was a sequence alignment map (SAM), which was converted, sorted, and indexed to Binary alignment mapping (BAM) using samtools (Li and Durbin 2009). Further processing of the bam files was carried out to remove duplicate reads resulting from PCR amplification during library preparation using Picard Markduplicate tools (http://broadinstitute.github.io/picard). Finally, the Samtools mpileup command along with Varscan (Koboldt et al. 2012) was used with the following parameters (minimum coverage 20, minimum variant frequency 0.05, p-value 0.005, and other parameters kept as default) for variant calling. Altogether 13 million SNPs were generated which were filtered based on the following parameters of no missing rate, minor allele frequency (0.05), genotype call quality (30), minimum mean depth (20), no indels, and only biallelic SNPs. The resulting variant call file contains 8,207 markers.

Variant Annotation

The variant file generated after filter markers based on above mentioned parameters was then annotated using snpEff (Cingolani et al. 2012). Alfalfa genomic data (fasta), gene annotation file (gff), coding sequence, and protein sequences (fasta) files were used to annotate the variant file. The SNPs were annotated as different functional effects based on their location in the Chromosome. The detailed distribution of the annotated variant effects is presented in Figure 1. Among the effects, only missense mutation that changes the amino acid sequences of the protein was retained.

doi:10.20944/preprints202305.1102.v1

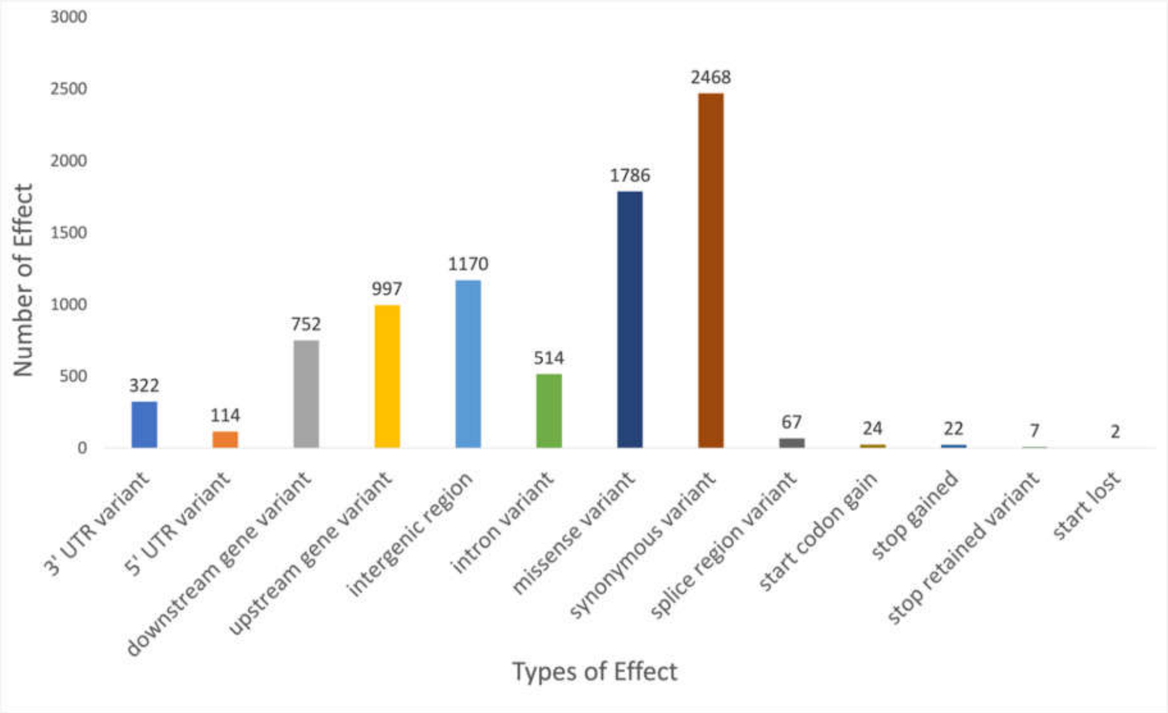

Figure 1. Types and distribution of effects as annotated by snpEff. The different color indicates different effect types as annotated by snpEff with the effect name provided in the x-axis and their total number on top of the bar plot. Among the annotated effects, the missense variant is one that we used for association study in this study as they are responsible for changing the amino acid sequence of the protein, ultimately affecting the final 3D structure of proteins.

Mutation effect prediction using MutPred 2

The pathogenicity of amino acid substitution was predicted using Mutpred 2 (Pejaver et al. 2020). Mutpred 2 is a machine learning-based tool for predicting the pathogenicity of missense variants which is mostly used in the case of humans. It uses a combination of sequence-based features, structure-based features, and functional annotation to predict the probability of a given missense variant being deleterious or benign. It provides an effect value ranging between 0–1. Any prediction value of 0.5 or above for missense mutation substitution is considered deleterious.

Prediction of the 3D structure of protein

Alphafold2 (Jumper et al. 2021) was used to predict the 3D structure of the protein. The high-performance computing GPU cluster Steptoe was used for running the alphafold. For the protein 3D structure prediction, template data from 2020-05-14 and the reduced dataset were used. The program was run using all the default parameters and the monomer models. At a time, 10 predictions were carried out for protein sequences with amino acid lengths less than 200 while the bulkier proteins with more than 1000 amino acid sequences were used one at a time. All the mutated and non-mutated protein structures were predicted through the alphafold.

Filtering of Markers based on LDDT values and their assignment to SNP genotypes

The Local distance difference test (LDDT) values are the measure to compare the accuracy of the predicted structure and experimental structure of the protein. Since any LDDT value below 50 is considered the lowest prediction accuracy, we used this value to further filter the SNPs to have protein features with the highest prediction accuracy structures. For association analysis, LDDT values of the amino acid residue between mutated and non-mutated proteins were used. Since the SNPs are coded as 0, 1, and 2 for the reference homozygote, heterozygote, and alternate homozygote, a similar coding method was used. The LDDT value for the reference protein was used as 0, the LDDT

doi:10.20944/preprints202305.1102.v1

value for the mutated amino acid was used as 2 and the heterozygote was the average value between the two. Mathematically, the assignment of LDDT values this way is similar to using the 0, 1, and 2 coding patterns of the SNP genotype. So, for convenience, we used the SNP coding way to perform association analysis.

Association Study and Annotation of significant gene

Altogether 154 genes with missense mutation that included the 5 significant genes from (Parajuli et al., 2023) was included in the final analysis. The genome-wide association study was performed using GAPIT (Wang 2021) package in R version 1.3.959. The number of principal components was selected based on the scree plot. Principal components along with location, pools, and generation were fixed as covariates in the model. The p-value threshold was set to the usual Bonferroni correction method which was 3.2×10^{-04}. For the significant hits, the protein sequence of significant genes is extracted from the Alfalfa annotation file and blasted against nr database of the National Center for Biotechnological Information for functional annotation of the genes.

Results

Variant calling and filtering SNPs

The variant calling on 121 alfalfa samples generated approximately 13 million SNPs. The raw SNPs were filtered based on the following parameters: no missing genotype, a minor allele frequency of 0.05, a minimum mean depth of 20, and a genotype call quality of 30 (probability of calling 1 false genotype in 1000 calls). In addition to this, insertions and deletion were also removed and only biallelic SNPs were kept for the final analysis, removing multi-allelic variant. After this, we acquire a total of 8,207 as final SNPs for further processing. For the annotation of these genetic variants, snpEff tools were used. The snpEff uses information from the Alfalfa genome assembly and annotation file for the annotation of the genetic variants. The variants were annotated with information from Alfalfa genomic data, gene annotation files, coding sequences, and protein sequences. Detailed information for the effect annotation is provided in Figure 1.

Most of the genetic variants were annotated as missense variants, synonymous variants, variants in upstream and downstream regions, intergenic regions, and intronic regions of the gene. Similarly, a smaller number of genetic variants (altogether 122) were annotated as splice region variant, start codon gain, stop gained, stop retained variant, and start lost variants. Synonymous mutations were the highest annotated genetic variant with 2,468 in number followed by the missense variant of 1,786. Other genetic variants that are outside the coding region of the genes were variants within the intronic region, intergenic region, and upstream and downstream of the genes. These variants do not directly alter the structure of the proteins.

While Synonymous mutations are silent mutations that do not change in the amino acid sequence of a protein ultimately not altering the 3D structure of the protein, missense mutations are the non-synonymous mutations that change the amino acid sequences leading to changes in the final 3d structure of the protein. Since missense variants are the ones that are actually impacting the final 3D structure of the protein, these were selected for further processing in the association study analysis. Although genetic variants leading to changes in the start and end codon are also vital in changing the protein structure, they are very few in number (only 55) and so they were not retained for the final association study analysis.

Mutation Prediction using Mutpred 2

An average gene is estimated to have several non-synonymous missense SNPs that result in substituting an amino acid residue in protein (Cargill et al. 1999). In humans, it has been estimated that there are around 13, 000 Exonic variants of which around 58% are non-synonymous (Tennessen et al. 2012). Even though non-synonymous mutations lead to changes in the amino acid sequences, in most cases these non-synonymous variants are not discernible and do not result in changes in protein stability and functions (Yates and Sternberg 2013). However, there are non-synonymous variants that can be destabilizing leading to changes in protein structure and function ultimately changing the phenotypic effect (Tennessen et al. 2012). Therefore, we further filtered out missense variants to select only the variants that result in a destabilizing effect on the phenotypes (Figure 2).

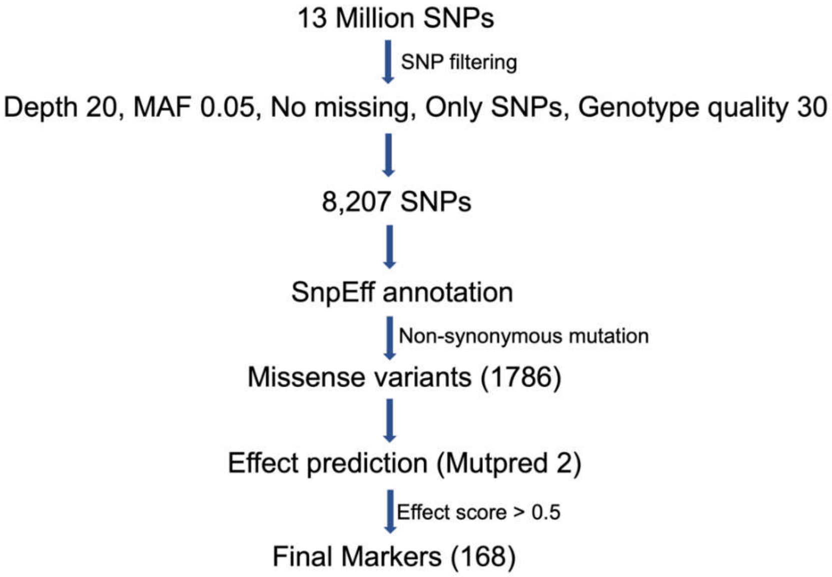

Figure 2. Selection for the 168 markers used in the study. The initial 13 million SNPs called from 121 Alfalfa samples were filtered and annotated using bcftools and SnpEff. From the non-synonymous SNPs, only missense variants were selected. The mutation effect prediction tool Mutpred 2 was used to select the final 168 markers based on an effect score of 50 or above.

For this, we used Mutpred 2 which is a tool to predict whether a missense variant due to amino acid substitution has any pathogenic effect/ deleterious effect. MutPred 2 returns a pathogenicity score between 0 and 1. Any predicted mutation effect with a score of 0.5 or above is considered as having deleterious effects due to mutation. It uses probabilistic models to estimate the likelihood of a given variant affecting the function of the protein. During the prediction, it also considers other factors such as sequence conservation, physiochemical properties of amino acids, and location of the variants within the protein structure. In humans, it helps in the identification of structural and functional mutational signatures related to Mendelian and complex neurodevelopmental disorders. The distribution of the mutation effect score for the missense variants is presented in Figure S4.

Alphafold prediction of protein 3D structure

The output from alphafold provides atomic level information of every residue of the protein 3D structure. The sample output from the alphafold is presented in Table 2. The first column header is called atom that represents information about residues in the protein information for amino acids and their elements. The Second and third column provides information about the atom number and the name of the atom that includes Nitrogen, Carbon, Oxygen, and Sulphur, which either forms the backbone of the amino acids or are connected to the side chain. The following columns provide the name of the amino acid as the residue name, the chain identifier as A for single chain and B and C for more than one chain, the residue number which is the number for the amino acid in the protein sequences, the X, Y, and Z coordinate of protein 3D structure in space. The final three columns contain the information for occupancy, the local distance difference test values, and the element symbol. Occupancy is the fraction of time that a given residue resides in the specific conformation.

Table 2. Detailed output from the alphafold protein 3D structure prediction.

A	ANB	AN	RN	CI	RNB	X	Y	Z	O	LDDT	ES
ATOM	1	N	MET	A	1	-10.427	13.062	-7.962	1	60.05	N
ATOM	2	CA	MET	A	1	-10.032	11.915	-8.775	1	60.05	C
ATOM	3	C	MET	A	1	-9.71	10.711	-7.896	1	60.05	C
ATOM	4	CB	MET	A	1	-8.824	12.264	-9.646	1	60.05	C
ATOM	5	O	MET	A	1	-8.88	10.802	-6.99	1	60.05	O
ATOM	6	CG	MET	A	1	-9.175	12.537	-11.1	1	60.05	C
ATOM	7	SD	MET	A	1	-7.722	12.399	-12.212	1	60.05	S
ATOM	8	CE	MET	A	1	-8.557	12.332	-13.822	1	60.05	C
ATOM	9	N	PRO	A	2	-10.577	9.55	-7.943	1	81.7	N
ATOM	10	CA	PRO	A	2	-10.288	8.391	-7.095	1	81.7	C
ATOM	11	C	PRO	A	2	-8.922	7.773	-7.386	1	81.7	C
ATOM	12	CB	PRO	A	2	-11.414	7.412	-7.44	1	81.7	C
ATOM	13	O	PRO	A	2	-8.545	7.619	-8.55	1	81.7	O
ATOM	14	CG	PRO	A	2	-11.943	7.887	-8.754	1	81.7	C
ATOM	15	CD	PRO	A	2	-11.509	9.31	-8.958	1	81.7	C

* The sample output from alphafold 3D Protein structure prediction of gene MsG0480023677.01.T01. The column represents Atoms(A), Atomic Number (ANB), Atomic Name (AN), Residue Name (RN), Chain ID (CI), Residue Number (RNB), X, Y, and Z coordinates of the protein 3D structure, Occupancy (O), Local distance difference test (LDDT) and Element Symbol (ES).

The most important part of this output that is used in this study is the local distance difference test values. This information from LDDT values for the mutated and non-mutated amino acids was used for filtering the SNPs in the final association study. These values could be converted to genotype values of 0, 1, and 2, however, for convenience of use we used the SNP coding values as it is mathematically the same for the association study. In general, the LDDT values range between 0–100, with structures >90 considered as very high confidence, between 70–90 as confident, between 50–70 as low, and anything below 50 considered very low (Figure S3). The very low score prediction could be due to the complexity of the structure, less information on the database about the region, and the destabilizing effect of amino acid substitution. Any LDDT values below 50 is considered very low prediction and so these values were further filtered. The distribution of LDDT values for the 168 SNPs is presented in Figure 3. Even though a single gene could have multiple locations of mutation within it, at a time only one mutational effect is used for the prediction of protein structure. Most of the values in our study are above 50 with few predictions below 50 which were further filtered.

Association study

The LDDT values of the mutated and non-mutated proteins were used for further filtering the SNPs while the SNP itself was used for performing GWAS. The Farmcpu model (Liu et al. 2016) which used fixed and random models, is used in GAPIT for the association analysis. Along with SNPs, 13 principal components as determined by scree plot along with location, generation, and pool were fixed as covariates. For the final analysis, 5 markers that were within the region of the gene from previous study (Parajuli et al., 2023) were also included in the final study. There were 5 significant SNPs identified by the association study. Out of the 5 SNPs, 4 were the same SNPs from the previous study (Chapter 3), while the analysis also identified a new locus in Chromosome 3 as a significant association.

The protein sequence of the new significant marker was extracted from Alfalfa genome assembly data and blasted against the NCBI data against the non-redundant (nr) protein database. The top results from the blast search identified the gene as a disease-resistance gene in *Medicago truncatula*. The gene was disease resistance protein RPV1 isoform X2 and disease resistance protein RPV1 isoform X1 in *Medicago truncatula*. The other genes were annotated as MDIS1-interacting receptorlike kinase 2, serine/threonine-protein kinase(d6pk1), T-complex protein 1 subunit theta, and putative RNA-directed DNA polymerase.

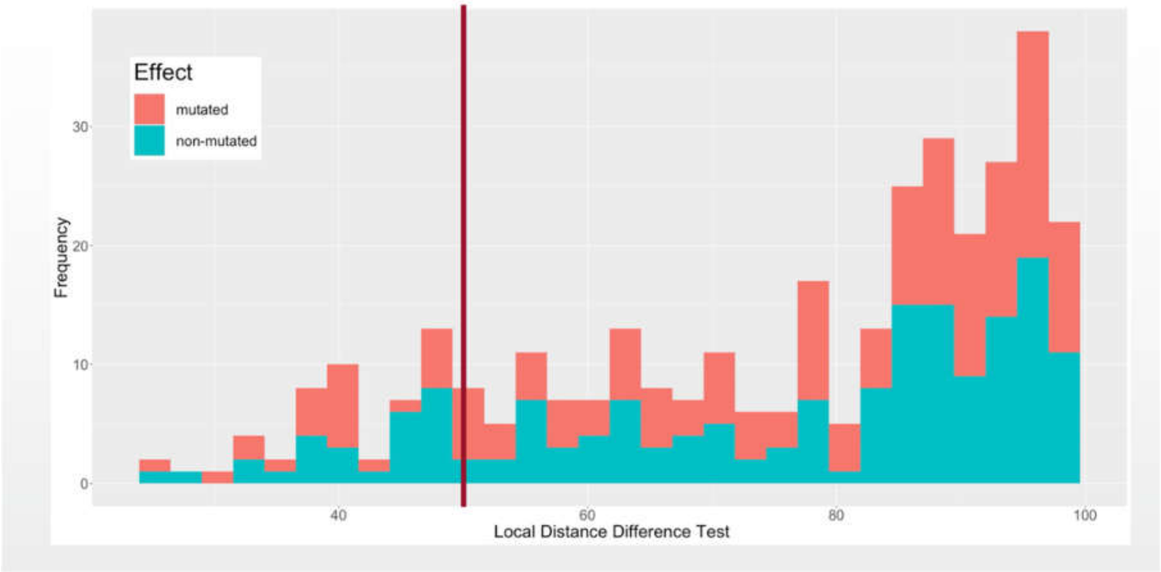

Figure 3. Distribution of LDDT values for the mutated and non-mutated amino acids for the 168 markers. The red line marks the point of 50 below which all the markers were filtered, and the final marker was 149, with further 5 significant markers from Chapter 3 making the total to 154 final markers.

Discussion

From GWAS to PWAS

In the present study, we used 154 markers to link phenotypes after filtering SNPs using features from Alphafold to predict the 3D structure of a protein with alfalfa plant growth vigor. Although QTL analysis and genome-wide association studies exist to connect phenotype with genotype, these mapping techniques suffer from specific limitations. QTL mapping suffers from allelic richness (Borevitz and Nordborg 2003) and a smaller number of genetic recombination leading to low mapping resolution (Balasubramanian et al. 2009). On the other hand, the statistical power of genome-wide association study is limited by rare variants with small effect sizes (Manolio et al. 2009). Moreover, due to linkage disequilibrium and population structure, pinning exact causal variants implicating a phenotype is still a complex task (Manolio et al. 2009). A significant mutation in GWAS analysis could be determined as non-significant due to the use of a stringent Bonferroni correction p-value threshold. With the availability of atomic-level protein 3D structures from the alphafold, we wanted to test whether we can use features from protein 3D structures (Figure S1) to gain important information from the SNP data. As a protein's biological function strongly depends upon its 3D structure, features from high-confidence protein 3D structure could be an important parameter to filter SNPs for final association analysis.

Marker Filtering and Computational Efficiency

We only used 154 markers because alphafold utilizes deep learning models that are computationally expensive (2) and the time required for predicting protein 3D structures takes a lot with the increased size of protein sequences. Filtering markers help reduce the number of genes, consequently reducing the expensive computational part, however, the genome-wide coverage of markers will always be less once the stringent threshold for filtering is applied. Besides, not all mutations leading to changes in the amino acid sequence are deleterious. Filtering based on the predicted effect of a missense mutation (non-synonymous) can capture only the effective mutation actually responsible for the phenotypic variation. Additionally, using genotype call quality as a parameter to filter SNPs reduces the number of false positive genotypes, further improving the quality of mutation (Figure 2).

General Structure of Protein

Protein is a major component of living cells that perform their respective biological function after folding into the most sTable 3D structure. Dysfunction in the structure of protein due to mutation is known to cause the development of deleterious conditions (Uversky and Dunker 2010). To understand specific feature changes in protein 3D structure due to point mutation, we need to understand as a whole about protein structure. A protein is a long chain of amino acids joined together by polypeptide bonds (Figure S2). The backbone of protein consists of Amide (NH_2) group, Alpha carbon, and carboxylic acid (COOH) group. On one side of the alpha carbon, hydrogen is attached, and the other side contains an alkyl group (R-group). Protein folding involves the formation of a complex network consisting of hydrogen bonds, Van der Waals force, and electrostatic interaction. These interactions stabilize protein into alpha helices and beta sheets, leading to the final three-dimensional structure.

3D Structure and Bond Angles of Protein

The final 3D structure of the protein is influenced by the torsion angle between alpha carbon and its surrounding residues. These angles describe peptide bonds connecting alpha carbon to carbonyl carbon and nitrogen atoms of the adjacent amide group. Since the amino acids are connected through peptide bonds, the torsion angle between the alpha carbon and its neighbor residues is not independent, meaning, a change in the position of alpha carbon leads to a change in the position of other nearby residues and vice versa. This change due to mutation affects the local environment around the alpha carbon, changing its position as well as its neighboring residues. The atomic level protein 3D structure prediction by Alphafold includes a metric called the local distance difference test, which is the difference in distance of residues between the experimental structure and predicted structure when they are superimposed to each other. As mentioned above, a change in the amino acid sequence leading to point mutation alters the torsion angle between backbone alpha carbon and its residues leading to either lower confidence prediction or higher confidence in that region. Using this information on the difference between the non-mutated protein structure and mutated protein structure we carry out the association analysis between protein 3D structure characteristics and phenotype.

Local Distance Difference Test (LDDT)

The local distance difference test is a statistical method used in protein alignment and comparison to identify structural differences between two proteins (Mariani et al. 2013). The conventional method of measuring the accuracy of predicted protein 3D structures has been modified from time to time (Mariani et al. 2013). The root-mean-square-deviation (RMSD), used in the first installment of Critical Assessment of Protein Structure Prediction (CASP) has been modified to overcome deficiencies in comparison of predicted and the experimental structures. Global Distance Test (GDT) (A. Zemla 2003; Adam Zemla et al. 2001), replacing the RMSD, also had limitations when applied to flexible proteins composed of several domains (Mariani et al. 2013). To overcome these failings, local superposition-free measures were developed. It was introduced to assess how well local atomic interactions in the reference structure are reproduced in the prediction (Mariani et al. 2011). This metric has an important role as it helps to determine the accuracy of protein 3D structure with high confidence as a result we used this information in our study for filtering SNPs for final association analysis.

Principal Component Analysis and Association Study

We used the LDDT value to capture the destabilizing effect of point mutation on the 3D structure of protein and their association with the phenotypes. As the biological function of protein is strongly dependent on its native 3D structure, any deviation in its structure due to point mutation will lead to protein misfolding which ultimately will affect the structure of the protein. Once its structure is affected, the protein is unable to perform its biological function. Using this confidence metric, genetic variants can be easily filtered from a bulk of SNP information. And we followed the same and used 154 SNPs for the final association analysis.

Principal component analysis (PCA) was performed in R version 1.3.959 to reduce the dimension of the SNP data and to extract the amount of the variance explained by it. Using the scree plot from the eigenvalues of the PCA, we selected 13 principal components to be fixed as covariates along with location, generation, and pools in the final analysis. The matrix plot for 6 principal components is presented in Figure 4. The first two principal components (PC) separate the samples into three clusters separating samples from Utah, Washington, and Wisconsin. Since the samples used in the analysis are not only from three different locations but are also from different lines themselves. The PC1 is able to separate that variation between the samples from Utah and other locations. Since the lines from Washington are diversity lines, which could have their roots in breeding lines as well, we can see that these lines from Washington are separated from Wisconsin, however, there are still lines that share the same variation with Wisconsin.

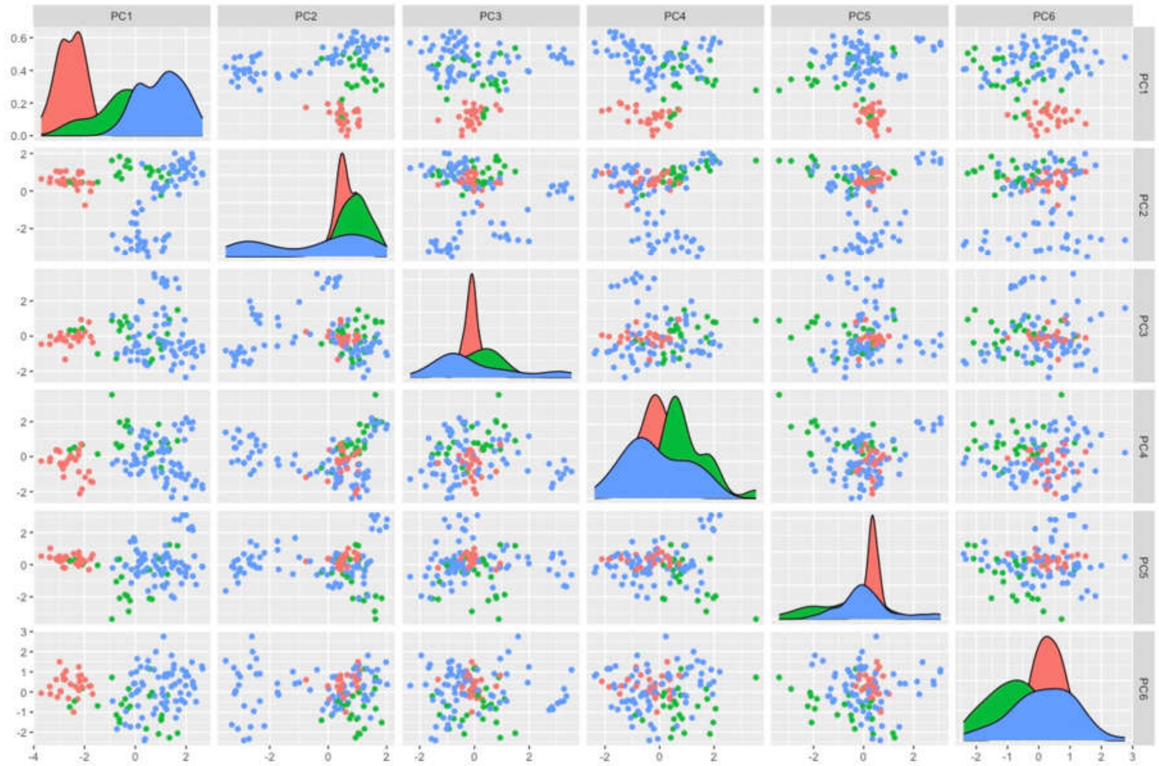

Figure 4. Scatterplot matrix of six principal components generated using the Prcomp function in R. The three different colors indicate the locations where the plants have been collected (orange-Utah, green-Washington, and blue-Wisconsin). These six PCs were among the first 6 PCs out of fixed as covariates in the FarmCPU model. These principal components were generated using all 5 features for 840 markers.

The association study identified 5 significant loci out of which four were the loci already identified as significant in the SNP GWAS analysis (Figure 5). However, in the current analysis, we identified a new association on Chromosome 3 which was annotated as the MsG0380017260.01.T01 gene. The association analysis identified the previously identified 4 SNPs using GWAS with just 154 markers, it can be inferred that this method of filtering SNPs from the large number of SNPs will be effective. It can help to identify the actual association that causes the changes in the protein structure and finally into the phenotype through a reduced significance threshold for the identification of genetic variants associated with the trait of interest. Also, since these SNPs are selected based on their effect on the protein structure, the identified variant can be directly translated into functional proteins as they will be annotated within the region of the gene. However, the computationally expensive part of the alphafold prediction is still a stumbling block to including more markers for association in the final study. For this, improvement in the mutation prediction tools that are solely developed for plant protein prediction will be required in the future.

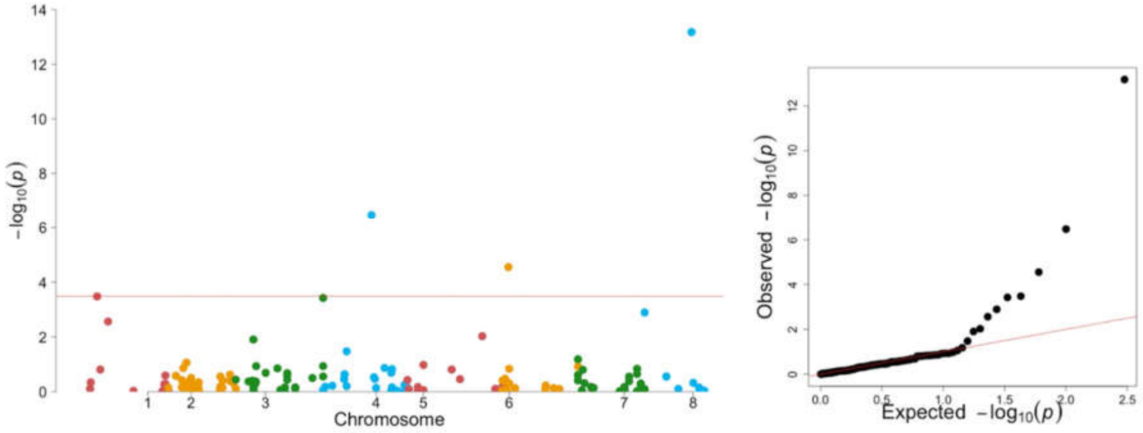

Figure 5. Manhattan and QQ plot of association study between alfalfa plant vigor phenotypes and features from protein 3D structures. The Manhattan plot presents the association study conducted using SNPs from the selected markers after they were filtered using the MutPred 2 prediction followed by the protein feature of LDDT values. The 4 significant SNPs are the ones identified in previous study while the SNP at Chromosome 3 is a new one.

Conclusions

We present the first association study that uses the features from protein 3D structure to filter SNPs and then uses the filtered SNPs for association study. The association study performed this way can help to identify genetic variants that can be easily translated to functional proteins. It can be used to link the effect of point mutation on the phenotypes. Effective genetic variants identified as non-significant due to stringent p-value threshold can also be filtered out and identified using this method. In this study, we identified a new association along with the previously identified association. The newly identified association was annotated as isoforms of a disease-resistance gene in *Medicago truncatula*. Association study involving proteins provides a more direct way of understanding the molecular basis of trait development. It can readily link the effect of mutation with the phenotypes which is not easily possible using SNPs only. However, the association study is limited only to the features predicted from the alphafold and it cannot be used to understand the effect of post-translational modification of proteins and protein-protein interactions in its current form.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization: A.P. and Z.Z. Funding: D.S., L.-X.Y., M.P., and Z.Z. Data analysis: A.P. Phenotype data collection: S.W., M.P., and L.-X.Y. Plant sample collection: S.W., M.P., and L.-X.Y. Manuscript draft: A.P. Manuscript revision: M.L.W. and R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This project was partially supported by the USDA National Institute of Food and Agriculture (Hatch project 1014919 and Award # 2018-70005-28792) and the Washington Grain Commission (Endowment and Award #s 126593 and 134574).

Data, Script, Code and Supplementary Information Availability: Data are available online: https://figshare.com/s/4958ff01c55a954d7585; R codes for GWAS analysis is available online at https://github.com/atitparajuli2020/Alfalfa_BSA_Final.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anfinsen, Christian B. 1973. "Principles That Govern the Folding of Protein Chains." *Science* 181(4096): 223–30. https://www.science.org/doi/10.1126/science.181.4096.223.
2. Ashenberg, Orr, L. Ian Gong, and Jesse D. Bloom. 2013. "Mutational Effects on Stability Are Largely Conserved during Protein Evolution." *Proceedings of the National Academy of Sciences* 110(52): 21071–76. https://pnas.org/doi/full/10.1073/pnas.1314781111.
3. Babrahman. "Babrahman Bioinformatics." https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
4. Bai, Yawen et al. 1995. "[15] Thermodynamic Parameters from Hydrogen Exchange Measurements." In, 344–56. https://linkinghub.elsevier.com/retrieve/pii/007668799559051X.
5. Bai, Yawen, and S. Walter Englander. 1996. "Future Directions in Folding: The Multi-State Nature of Protein Structure." *Proteins: Structure, Function, and Genetics* 24(2): 145–51. https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0134(199602)24:2%3C145::AID-PROT1%3E3.0.CO;2-I.
6. Balasubramanian, Sureshkumar et al. 2009. "QTL Mapping in New Arabidopsis Thaliana Advanced Intercross-Recombinant Inbred Lines" ed. Brian P. Dilkes. *PLoS ONE* 4(2): e4318. https://dx.plos.org/10.1371/journal.pone.0004318.
7. Borevitz, Justin O., and Magnus Nordborg. 2003. "The Impact of Genomics on the Study of Natural Variation in Arabidopsis." *Plant Physiology* 132(2): 718–25. https://academic.oup.com/plphys/article/132/2/718/6111721.
8. Bushnell, Brian. 2014. "BBMap: A Fast, Accurate, Splice-Aware Aligner." In *9th Annual Genomics of Energy & Environment Meeting*, Walnut Creek.
9. Cargill, Michele et al. 1999. "Characterization of Single-Nucleotide Polymorphisms in Coding Regions of Human Genes." *Nature Genetics* 22(3): 231–38. https://www.nature.com/articles/ng0799_231.
10. Cingolani, Pablo et al. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6(2): 80–92. http://www.tandfonline.com/doi/abs/10.4161/fly.19695.
11. Jumper, John et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596(7873): 583–89. https://www.nature.com/articles/s41586-021-03819-2.
12. Koboldt, Daniel C. et al. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." *Genome Research* 22(3): 568–76. http://genome.cshlp.org/lookup/doi/10.1101/gr.129684.111.
13. Levinthal, Cyrus. 1968. "Are There Pathways for Protein Folding?" *Journal de Chimie Physique* 65: 44–45. http://jcp.edpsciences.org/10.1051/jcp/1968650044.
14. Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25(14): 1754–60. https://academic.oup.com/bioinformatics/article/25/14/1754/225615.
15. Liu, Xiaolei et al. 2016. "Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies" ed. Jennifer Listgarten. *PLOS Genetics* 12(2): e1005767. https://dx.plos.org/10.1371/journal.pgen.1005767.
16. Luheshi, Leila M, Damian C Crowther, and Christopher M Dobson. 2008. "Protein Misfolding and Disease: From the Test Tube to the Organism." *Current Opinion in Chemical Biology* 12(1): 25–31. https://linkinghub.elsevier.com/retrieve/pii/S1367593108000392.
17. Manolio, Teri A. et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461(7265): 747–53. http://www.nature.com/articles/nature08494.
18. Mariani, Valerio et al. 2011. "Assessment of Template Based Protein Structure Predictions in CASP9." *Proteins: Structure, Function, and Bioinformatics* 79(S10): 37–58. https://onlinelibrary.wiley.com/doi/10.1002/prot.23177.
19. Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. "LDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests." *Bioinformatics* 29(21): 2722–28. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt473.
20. O'Rourke, Jamie A. et al. 2015. "The Medicago Sativa Gene Index 1.2: A Web-Accessible Gene Expression Atlas for Investigating Expression Differences between Medicago Sativa Subspecies." *BMC Genomics* 16(1): 502. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1718-7.

doi:10.20944/preprints202305.1102.v1

21. Parajuli, A.; Brueggeman, R.; Wagner, S.; Warburton, M.; Peel, M.; Yu, L.; See, D.; Zhang, Z. Bulked Target Capture Sequencing Identified Numerous Genetic Loci Associated with Alfalfa Growth Vigor During Inbreeding. *Preprints.org* 2023, 2023050898. https://doi.org/10.20944/preprints202305.0898.v1.
22. Pejaver, Vikas et al. 2020. "Inferring the Molecular and Phenotypic Impact of Amino Acid Variants with MutPred2." *Nature Communications* 11(1): 5918. https://www.nature.com/articles/s41467-020-19669-x.
23. Shen, Chen et al. 2020. "The Chromosome-Level Genome Sequence of the Autotetraploid Alfalfa and Resequencing of Core Germplasms Provide Genomic Resources for Alfalfa Research." *Molecular Plant* 13(9): 1250–61. https://linkinghub.elsevier.com/retrieve/pii/S1674205220302161.
24. Sotomayor-Vivas, Cristina, Enrique Hernández-Lemus, and Rodrigo Dorantes-Gilardi. 2022. "Linking Protein Structural and Functional Change to Mutation Using Amino Acid Networks" ed. Sriparna Saha. *PLOS ONE* 17(1): e0261829. https://dx.plos.org/10.1371/journal.pone.0261829.
25. Tennessen, Jacob A. et al. 2012. "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes." *Science* 337(6090): 64–69. https://www.science.org/doi/10.1126/science.1219240.
26. Tokuriki, Nobuhiko, and Dan S Tawfik. 2009. "Stability Effects of Mutations and Protein Evolvability." *Current Opinion in Structural Biology* 19(5): 596–604. https://linkinghub.elsevier.com/retrieve/pii/S0959440X09001249.
27. Tóth-Petróczy, Ágnes, and Dan S Tawfik. 2014. "The Robustness and Innovability of Protein Folds." *Current Opinion in Structural Biology* 26: 131–38. https://linkinghub.elsevier.com/retrieve/pii/S0959440X14000724.
28. Uversky, Vladimir N., and A. Keith Dunker. 2010. "Understanding Protein Non-Folding." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1804(6): 1231–64. https://linkinghub.elsevier.com/retrieve/pii/S1570963910000324.
29. Yao, Zhen et al. 2020. "Evaluation of Variant Calling Tools for Large Plant Genome Re-Sequencing." *BMC Bioinformatics* 21(1): 360. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03704-1.
30. Yates, Christopher M., and Michael J.E. Sternberg. 2013. "The Effects of Non-Synonymous Single Nucleotide Polymorphisms (NsSNPs) on Protein–Protein Interactions." *Journal of Molecular Biology* 425(21): 3949–63. https://linkinghub.elsevier.com/retrieve/pii/S0022283613004440.
31. Zemla, A. 2003. "LGA: A Method for Finding 3D Similarities in Protein Structures." *Nucleic Acids Research* 31(13): 3370–74. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg571.
32. Zemla, Adam, ?eslovas Venclovas, John Moult, and Krzysztof Fidelis. 2001. "Processing and Evaluation of Predictions in CASP4." *Proteins: Structure, Function, and Genetics* 45(S5): 13–21. https://onlinelibrary.wiley.com/doi/10.1002/prot.10052.