**Employing homology modelling software to identify genomic variation implicated in enabling the colonisation of alpine environments by *Arabidopsis arenosa***

Name: Matthew Gaskins

Names of other project members: Bailey Francis, Lewis Wood, Halima Abdulsalam

**Abbreviations:**

AFD: Allele Frequency Difference

BLOSUM62: Blocks Substitution Matrix 62

CASP: Critical Assessment of Protein Structure Prediction

CoA: Coenzyme A

DREPP: Developmentally Regulated Plasma Membrane Polypeptide

FAR2: Fatty Acyl-CoA Reductase 2

FAR5: Fatty Acyl-CoA Reductase 5

FST: Fixation Index

MAP18: Microtubule-Associated Protein 18

MSA: Multiple Sequence Alignment

NAD(P)H: Reduced Nicotinamide Adenine Dinucleotide (Phosphate)

PGM: Population Genomic Metric

PSI-BLAST: Position-Specific Iterative Basic Local Alignment Search Tool

PtdInsP: Phosphatidylinositol Phosphate

SNP: Single-Nucleotide Polymorphism

UPGMA: Unweighted Pair Group Method with Arithmetic mean

**Project Role:**

I employed AlphaFold to generate all three homology models for MAP18. I then conducted secondary-structure prediction for FAR5 using JPred4, before extracting PGM values from sample data for our two candidate-genes to generate AFD-plots. Lastly, I employed ChimeraX to visualise frequently-observed variants in three-dimensional space within homology models.

**Aims and Experimental Question:**

Ancestral *Arabidopsis arenosa* thrives at low-altitude, however the species has sporadically colonised multiple geographically-distinct alpine environments [1]. In 2020, Knotek *et al.* demonstrated this high-altitude ecotype arises through independent, parallel adaptation, rather than a shared divergence event [2]. Well-documented examples of plant parallel-evolution are rare, therefore this study aims to provide further insight into genomic mechanisms facilitating environmental adaptation [2].

Text

Description automatically generated with medium confidenceAs displayed in Figure 1, alpine populations exhibit marked phenotypic variation compared to foothill relatives, with shorter stems, darker flowers, and more-compact morphology. This report aims to explain phenotypic differences by interrogating two geographically-isolated foothill-alpine pairs, SUB/ZEP and HRA/TKO, again displayed in Figure 1. Our research-group identified two candidate-proteins - MAP18 and FAR5 - displaying significant genomic variation within five distinct foothill-alpine pairs. These proteins represent the focus of our attempts to explain phenotypic disparities between alpine and foothill *A.arenosa.*

We attempt to interpret functional effects of frequently-observed MAP18 and FAR5-mutants on high-altitude acclimatisation by examining variation at each level of protein-structure. Expected significance of missense mutations is quantified using a Grantham matrix, before predicted secondary-structures of alpine and non-alpine proteins are compared to determine whether environmental adaptation derives from variation at this structure level.

We then employ two homology modelling programs to predict tertiary-structures for low- and high-altitude candidate-proteins, with the highest-quality model selected to visualise variants previously implicated in environmental adaptation. A secondary aim of our report is therefore to identify an optimised method of modelling candidate-proteins.

Localising structural variation within three-dimensional models incites speculation regarding physiological consequences of particular mutations. However, in absence of functional characterisation for individual residues through experimental targeted-mutagenesis, our findings remain conjecture. Nevertheless, our project aims to provide further insight into high-altitude adaptation in *Arabidopsis*, and extend our understanding of parallel evolution in plants.

**Background and Motivation:**

Diagram

Description automatically generated*Arabidopsis arenosa* is a diploid-autotetraploid species of flowering plant located primarily within foothill regions of Central and Eastern Europe, with scattered occurrences identified at high-elevation [1]. *A.arenosa* colonises at least four geographically-distinct mountainous regions - as displayed in Figure 2, phylogenetic studies indicate our two experimental populations colonised alpine habitats through independent, parallel evolution, rather than a shared divergence event [2]. This parallel alpine adaptation therefore represents a suitable paradigm for understanding convergent-evolution.

Genome-wide SNP variation within foothill-alpine *A.arenosa* pairs can be interrogated to elucidate genomic mechanisms underlying high-altitude adaptation - here we examine two geographically-isolated SUB/ZEP and HRA/TKO populations. SNP-frequency is quantified using population genomic metrics, including AFD and FST. Allele-frequency divides the total number of allele-sequences by the incidence of a particular variant. Large AFD-values therefore indicate significant variation between populations [3]. FST quantifies the difference in allelic variant frequency in one population compared to its mean frequency amongst all populations combined. Large FST-values therefore also indicate important allelic variation, as FST peaks when variants become fixed in one population and not another [3]. Both metrics are employed to identify selection of particular alleles in specific environments.

By analysing PGMs across genomes of five foothill-alpine *A.arenosa* pairs, our research-group identified two candidate-genes which display significant variation - *FAR5 and MAP18,* located on chromosomes 3 and 5 respectively [4,5]. To locate variants within candidate-proteins facilitating environmental adaptation, we align primary-sequences of MAP18 and FAR5 for one foothill-alpine(SUB/ZEP) pair using ClustalOmega [6]. Figure 3 illustrates the main stages of this alignment algorithm.

Diagram

Description automatically generatedTable

Description automatically generatedSNPs within characterised protein domains are likely to have functional consequences. This can be further evidenced using the Grantham matrix shown in Figure 4, which quantifies physiochemical differences between substituted amino-acids [7]. Larger Grantham-scores indicate decreased residue similarity; these substitutions are more likely to elicit physiological consequences. Functionally significant residues are also predisposed to conservation amongst kingdoms, as evolutionary pressures select against variation at this location; mutations within conserved residues suggest important functional change.

SNPs can manifest variation at higher levels of protein-structure. Evolutionary repetition of canonical secondary-structures is exploited by JPred to predict query-protein organisation based on homologous sequence [8]. Homology modelling relates known protein structures to query-sequence at the tertiary-level to produce three-dimensional predictions. Homology model quality can be assessed using various metrics, including the number of Ramachandran outliers. These residues possess energetically-unfavourable dihedral-angles which consequently lie outside permitted regions of Ramachandran plots; numerous outliers indicate poor model-quality [9]. Clashscore represents another widely-used quality metric, defined as the number of serious steric overlaps (>0.4Å) per 1000-atoms - larger clashscores depict less-realistic models [9].

**Methods:**

Nucleotide and primary-sequence data for MAP18 and FAR5 from reference *A.lyrata* and all four *A.arenosa* populations were provided by Dr. Sian Bray. ClustalOmega(version1.2.4) was then run between foothill SUB and alpine ZEP strains using default parameters [6]. Subsequent MSAs were visualised with JalView(version2.11.2.0), with alignment-quality assessed using a BLOSUM62 scoring-matrix [10].

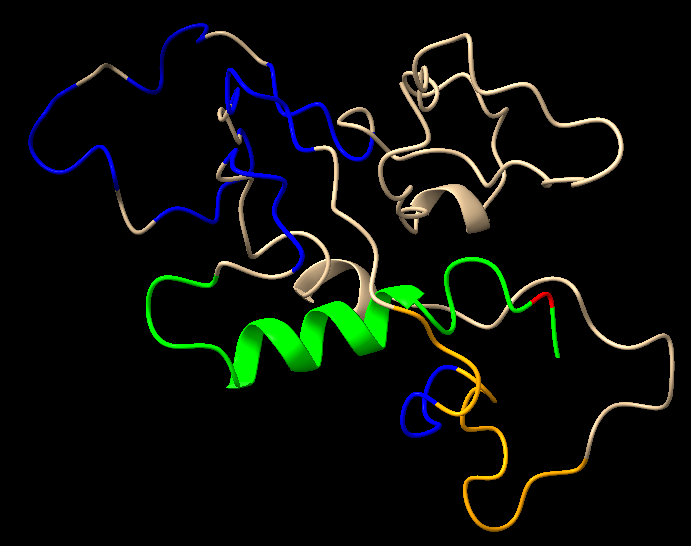
Missense variants with AFD-values exceeding 0.15 for SUB/ZEP and 0.1 for HRA/TKO were extracted from provided population genomic metric tables (see appendices). These high-frequency missense mutations were then further interrogated through analysis of Weir & Cockerham FST values and Grantham distance scores to estimate functional significance.

PSI-BLAST was employed to align *A.arenosa* candidate-proteins with homologues from fourteen model organisms [11]. The program was run under default parameters using the Model Organisms(landmark) database and a BLOSUM62 scoring-matrix. Expect-thresholds of 0.05 and 1 were used for FAR5 and MAP18 respectively. Secondary-structures for both candidate-proteins were then predicted using JPred4 under default parameters, and visualised using the 2dSS webserver [8,12].

Lastly, the three-dimensional configuration of each candidate protein was modelled using two different software - RaptorX and AlphaFold(version2.1.0), with both webservers run under default parameters [13,14]. Four from five chains produced by RaptorX were deleted to allow quality evaluation.

The quality of each model was assessed using MolProbity to identify preferred homology modelling software [9]. Hydrogen-atoms were added using nuclear x-H bond lengths and automatically repositioning asparagine, glutamine and histidine side-chains. Favourable criteria include infrequent Ramachandran outliers, a low clashscore and few energetically unfeasible rotamer-angles; RaptorX and AlphaFold were identified as optimum programs for modelling MAP18 and FAR5 respectively. Three-dimensional structures of models built using preferred software were then visualised using ChimeraX(version1.1) to identify residues implicated in high-altitude adaptation [15].

**The Structure of MAP18 and FAR5:**

In *Arabidopsis*, *MAP18* encodes the eponymous 168-amino acid plasma membrane-associated protein displayed in Figure 5. MAP18 regulates intracellular signalling including response to abscisic-acid, cold and salt-stress - all conditions plants experience at high-altitude [16]. A helical-wheel model of the N23-domain is presented in Figure 6, which reveals basic residues are predominantly aligned along the same helical face. This polybasic face allows the N23-domain to mediate electrostatic forces-of-attraction between positively-charged residue R-groups and negatively-charged phosphate moieties from inositol-rings of PtdInsP ligands, however calcium-calmodulin complexes also competitively bind this domain [17].

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceChart, radar chart

Description automatically generated

The remainder of *MAP18* is characterised by seven VEEKK repeat motifs, theorised to bind microtubules in other organisms, including analogous mouse-MAP1B [19]. Considering MAP18 is predominantly expressed in developing tissues, particularly root-hair cells, this represents a likely function of this protein - accordingly in 2017, Kang demonstrated MAP18 inhibits microtubule-polymerisation in *Arabidopsis thaliana* [19,20]. VEEKK motifs are therefore strongly implicated in MAP18-mediated microtubule destabilisation, essential for appropriate cell morphogenesis.

*FAR5* encodes the eponymous 496-amino acid helix-rich enzyme displayed in Figure 7. FAR5 catalyses fatty acyl-CoA reduction into alcohols required for synthesis of essential biomolecules, including suberin and cuticular root-wax [21]. This gene is consequently expressed in vascular leaf-tissue, where thick waxy cuticles are required to prevent water-loss [21].

A picture containing shape

Description automatically generatedFAR5 has two well-characterised domains; a 303-residue N-terminal NAD(P)H-binding domain, and a shorter 103-amino acid C-terminal Sterile domain, shown in green and red respectively in Figure 7 [21]. The NAD(P)H-binding domain facilitates donation of two electrons from NADPH-cofactor to fatty-acyl CoA substrate [21]. The function of the Sterile domain is less clear, however in *A.thaliana* FAR2, a similar domain mediates pollen-exine formation during male gametogenesis [22]. Interestingly, multiple mutations between the NAD(P)H-binding and Sterile domains are characterised to alter synthesised alcohol chain-lengths, despite their significant distance from the active-site [23].

**Results and Discussion:**

A screenshot of a computer

Description automatically generated with medium confidenceTimeline

Description automatically generatedTo achieve our aims of identifying variation between low- and high-elevation *A.arenosa* candidate-proteins, we aligned MAP18 and FAR5 primary-sequences from foothill(SUB) and alpine(ZEP) populations using ClustalOmega [6]. MSAs for MAP18 and FAR5 are displayed alongside reference *A.lyrata* in Figures 8 and 9 respectively. PSI-BLAST demonstrates *A.lyrata* homologues share over 95% identity with candidate *A.arenosa* proteins, thus this species represents a suitable reference [11]. We selected ClustalOmega as this program produces more accurate MSAs than alternative software, with comparably low computing times to competitors including Mafft [24].

Three SNPs identified within the N23-domain of alpine *A.arenosa* warrant further investigation, as these twenty-three N-terminal residues mediate MAP18-ligand contact with both PtdInsPs and calcium-calmodulin. However, the seven VEEKK-motifs, coloured blue in Figure 8, remain highly conserved between *Arabidopsis* populations - we therefore conclude high-altitude adaptation is unlikely to affect MAP18-mediated microtubule polymerisation.

The majority of SNPs reside within the MAP18 C-terminus, however this region is predicted to contain significant intrinsic disorder, and with no characterised function, C-terminal variation is presumably inconsequential [18]. Moreover, 27% of C-terminal substitutions possess at least weak similarity with the original residue, reducing their likelihood of functional importance. As later described, the insignificance of C-terminal variation is supported by PGMs; AFD-values for all C-terminal SNPs (with one exception) feasibly result from chance, demonstrating these mutations are unlikely to facilitate high-altitude adaptation across multiple geographically-distinct populations.

Timeline

Description automatically generatedTimeline

Description automatically generated with medium confidence

A close-up of a document

Description automatically generated with low confidenceTimeline

Description automatically generated

The NAD(P)H-binding domain of FAR5 exhibits significant conservation across all three populations - only 3% of residues contain SNPs. As displayed in Figure 9, only one variant within this domain exists in alpine *arenosa* but neither SUB nor the reference; this D216E mutation warrants further investigation, although both residues display strong physiochemical similarity.

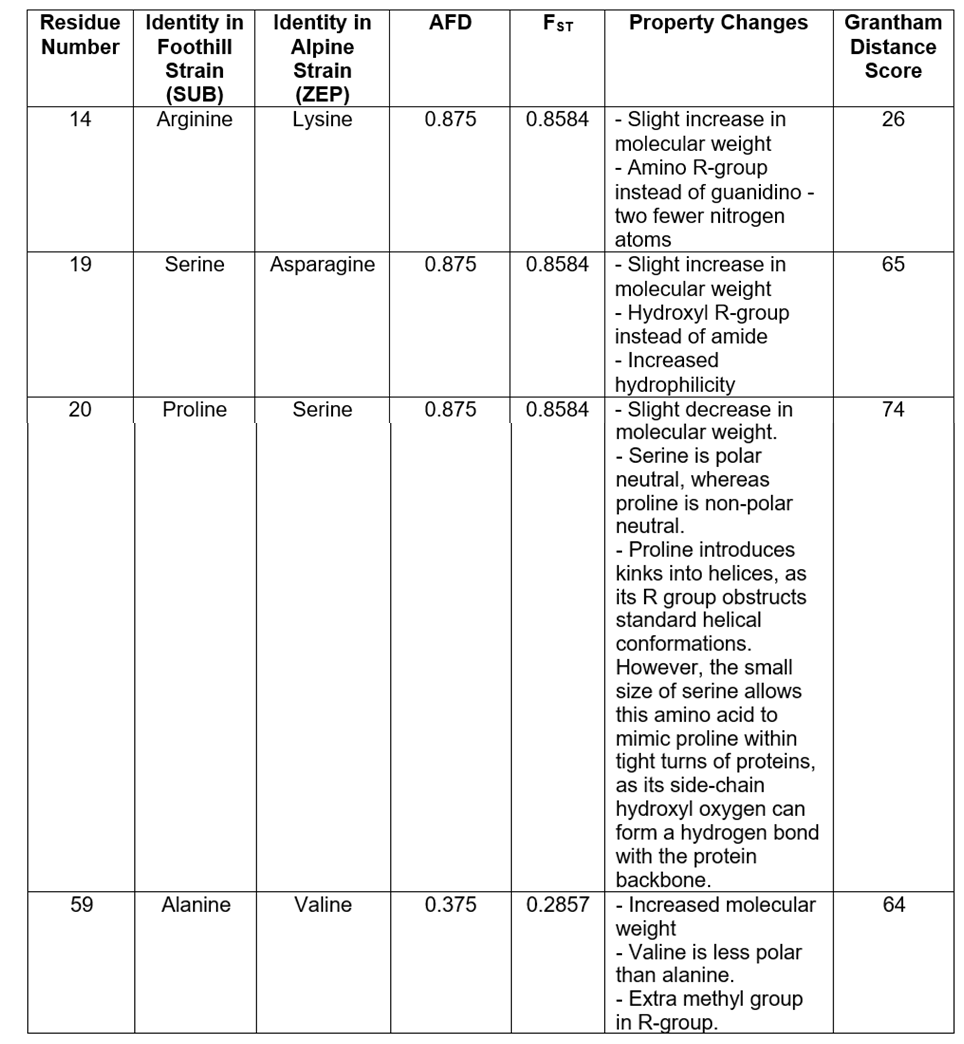
Six SNPs are located between the end of the NAD(P)H-binding domain and the start of the Sterile domain. Three of these substitutions are between ZEP and at least one of the foothill and reference populations - two of which closely proximate the Sterile domain. Point-mutations directly proximal to the Sterile domain are characterised to alter chain-length of FAR5-synthesised alcohols in *Arabidopsis thaliana*, implicating these variants in high-altitude adaptation - longer, less-polar fatty-alcohols comprising root-waxes will increase cuticular permeability to non-polar carbon-dioxide [23,27]. This is selectively advantageous for alpine plants to mitigate decreased carbon-dioxide availability for photosynthesis at high-altitude [28]. We therefore closely interrogate variants proximal to the Sterile domain.

However, the Sterile domain itself contains the most significant proportion of variation (48%) within our MSA. This domain is tentatively predicted to support pollen-exine formation, therefore mutations within this region may facilitate environmental adaptation - for example, tougher exines could enable pollen to avoid desiccation at high-altitude, increasing fertilisation success probabilities. As later described, two Sterile mutants are especially frequent amongst ZEP samples, and demand further explanation.

To isolate SNPs facilitating high-altitude adaptation, we extracted all variants with AFD-values greater than a specified threshold between both foothill-alpine pairs under investigation. All plant samples are diploids to maintain allele-number consistency. The total number of allele-sequences for SUB/ZEB and HRA/TKO is 14 and 32 respectively, hence minimum AFD-thresholds of 0.15 and 0.1 were employed for the two respective populations - this only extracts AFD-values with magnitudes larger than expected by chance [3].

Tables 1 and 2 display significantly frequent SNPs within MAP18 between SUB/ZEP and HRA/TKO respectively. The leading observation is repeated incidence of three missense-mutations within the N23-domain previously identified in our MSA. When we consider the N23-helical-wheel model displayed in Figure 6, lysine-14 is implicated at the protein-ligand interface, with serine-19 and proline-20 at 45° and 90° respectively from this residue, away from the binding-site. We speculate replacement of proline-20 for serine modifies helical orientation such that the N23-domain shifts to involve alternative residues at the protein-ligand interface [29]. As less-basic residues consequently mediate weaker electrostatic forces-of-attraction between the N23-domain and PtdInsPs, this conformational change decreases MAP18-ligand affinity.

P20S is repeatedly observed alongside two other high-frequency SNPs, but possesses the largest Grantham-score, therefore we hypothesise this substitution arose first, with the remaining two variants arising secondarily in-response. Secondary A14K and S19N mutations both reduce electrostatic forces-of-attraction with PtdInsPs - arginine forms two additional electrostatic bonds with ligands than lysine, as its guanidinium side-chain contains two more nitrogen-atoms than the lysine amino R-group, whilst three lone electron-pairs of asparagine repel negatively-charged moieties more strongly than one pair from serine [30,31]. Reducing the extent of PtdInsP-binding frees these molecules to modulate downstream target-effectors that improve plant-stress response [32]. This biochemical modification may therefore enable *A.arenosa* to tolerate high-altitude, and requires further investigation through homology modelling.

Unlike our MSA, an A59V substitution within the coiled-coil is recognised as significant. AFD-values are beneath 0.5 for A59V, therefore the substitute residue is absent from *A.arenosa* consensus sequence and is not observed in MSAs. Contrastingly, numerous C-terminal missense mutations are observed in our MSA, however their PGM values are beneath thresholds for inclusion in our tables - suggesting they have little functional consequence. However, one C-terminal variant worth investigation is P142L, despite its barely significant AFD-value. This substitution possesses a notably large Grantham-score, and may therefore invoke significant MAP18 structural variation.

Table

Description automatically generatedTable

Description automatically generated

Chart

Description automatically generated with medium confidenceTable

Description automatically generated

Text

Description automatically generated

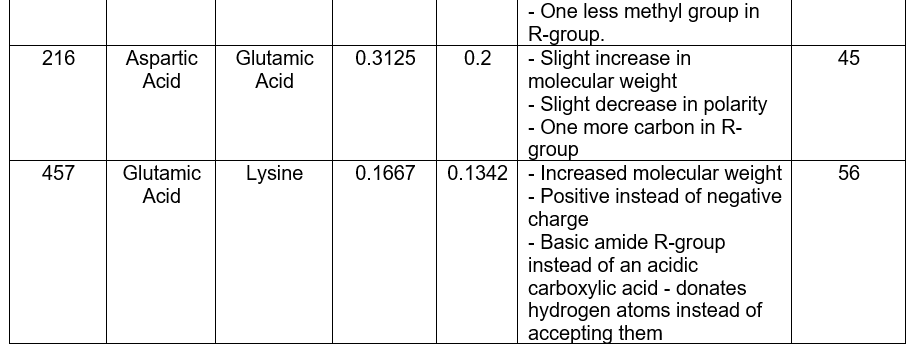
It is also worth mention significant AFD-values for MAP18 variants exist outside coding-regions. Figure 10 displays a SUB/ZEP absolute-AFD plot relative to the start and end-position of MAP18; whilst most high-frequency SNPs are located in the N23-domain, scattered intronic-variants are also regularly observed, alongside one upstream-mutation. Our project is limited by excluding analysis of these variants.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

Tables 3-4 exhibit high-frequency FAR5-SNPs between SUB/ZEP and HRA/TKO populations respectively.

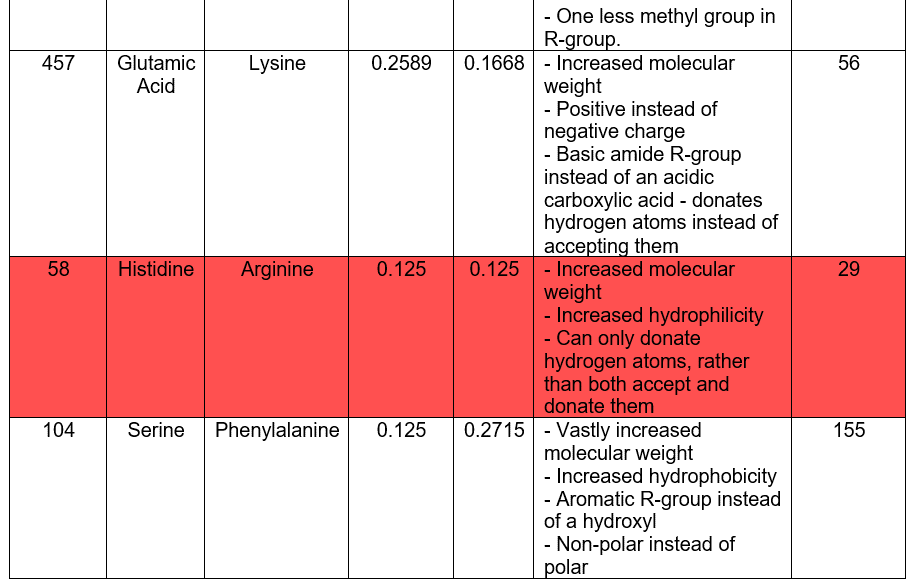
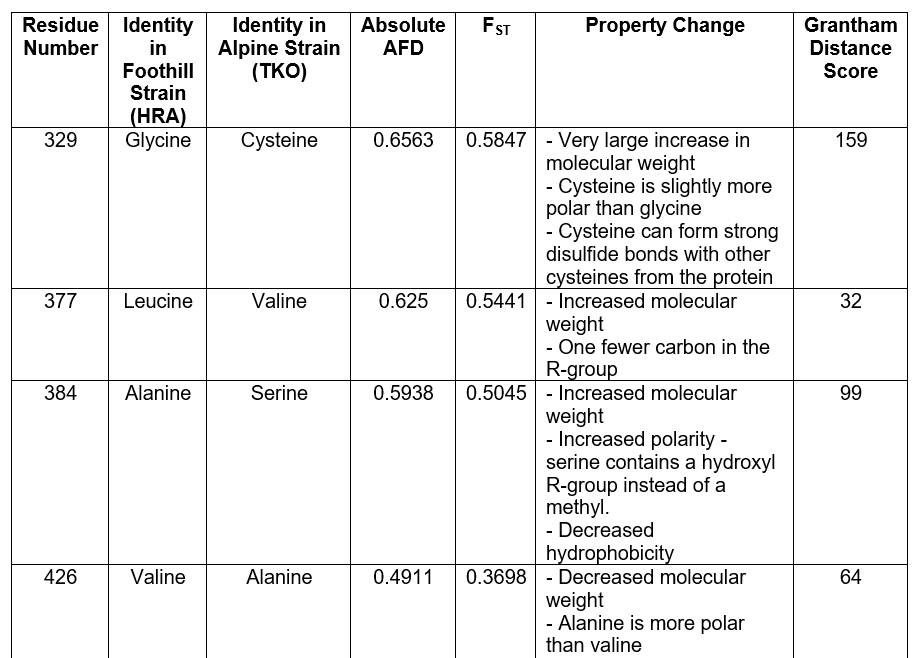


-De

-De

A screenshot of a computer

Description automatically generated with low confidence



-De

Table, calendar

Description automatically generated



Four FAR5 SNPs are completely distinct between SUB and ZEP populations. However, only I490V is located within a functionally-characterised (Sterile) domain - this substitution has a particularly low Grantham-score, and is therefore unlikely to confer extensive structural variation. Correspondingly, the G329C SNP with a larger Grantham-score is significantly more likely to elicit conformational change - cysteine residues can introduce strong disulfide-bonds that define tertiary-structures, which can be investigated further through homology modelling.

However, two substitutions directly proximal to the Sterile domain are particularly noteworthy; mutations in this location are characterised to alter FAR5-synthesised alcohol chain-length in *Arabidopsis thaliana*. In *thaliana*, a V377M substitution catalyses C16:0 fatty-alcohol formation instead of C18.0 [23]. Reintroduction of valine through L377V in *arenosa* may therefore produce reciprocal effect; leucine and methionine are highly-similar residues with low Grantham-substitution scores, thus this mutation may facilitate C18.0 fatty-alcohol synthesis. In *thaliana*, chain-length modification requires a second substitution - alanine is replaced by leucine [23]. Here in *arenosa*, we also observe alanine substituted for a dissimilar residue, in this case serine. These two SNPs could therefore combine to decrease cuticular polarity and increase carbon-dioxide uptake at altitudes where gaseous exchange is more limited [28]. We recommend targeted point mutagenesis of these two residues, with subsequent mass-spectrometry identifying whether FAR5-synthesised alcohol chain-length increases post-mutation.

Whilst the majority of variation identified in our MSA occurs in the Sterile domain, this does not translate to large PGM values - residue identity is likely unconserved and mutations vary in nature. However, two Sterile substitutions are frequently observed for both foothill-alpine pairs. Due to analogy between *Arabidopsis-*FAR5 and FAR2 Sterile domains, it could be speculated these SNPs affect pollen-exine synthesis [22]. However, the role of this FAR5 Sterile domain remains unclear - predictions regarding mutations affecting probabilities of fertilisation success represent conjecture.

Chart, scatter chart

Description automatically generatedFigure 11 displays Grantham-scores for high-frequency mutations identified within FAR5 for either foothill-alpine pair. Outside aforementioned high-frequency variants, we observe a particularly large value for S104F in the NAD(P)H-binding domain. Another L106F variant is also observed between HRA and TKO; introduction of phenylalanine within this region may alter helical orientation and consequently promote NAD(P)H-binding. Further investigation through homology modelling is required.

Returning to MAP18, we note the helical-wheel model of *A.arenosa* N23-domains in Figure 6 only contains one residue inconsistent with *thaliana*, and even substituted amino-acids are similar. A high-level of N23-domain conservation across species and potentially kingdoms indicates important functional consequence. Therefore, to determine the extent of interkingdom conservation within other identified SNPs, we employ PSI-BLAST to align MAP18 and FAR5 homologues from fourteen model-organisms. We select this program ahead of standard BLASTp as PSI-BLAST detects additional distantly-related proteins - this is essential as MAP18 homologues are scarce within the BLAST database[11].

Even when the expect-threshold is raised to 1, we are unable to identify MAP18 homologues from model organisms, except for two leguminous soybeans: *Glycine max* and *Arabidopsis thaliana* (as displayed in Figure 12). Only the first sixteen MAP18 residues display conservation between these three species, indicating this N-terminal is functionally significant. Contrastingly, C-terminal residues are not conserved, indicating variants identified in our MSA likely represent intrinsic disorder rather than biologically consequential change. Our analysis also reveals the previously-described R14K substitution returns *A.arenosa* to sequence observed in other leguminous soybeans, highlighting the potential relevance of this substitution.

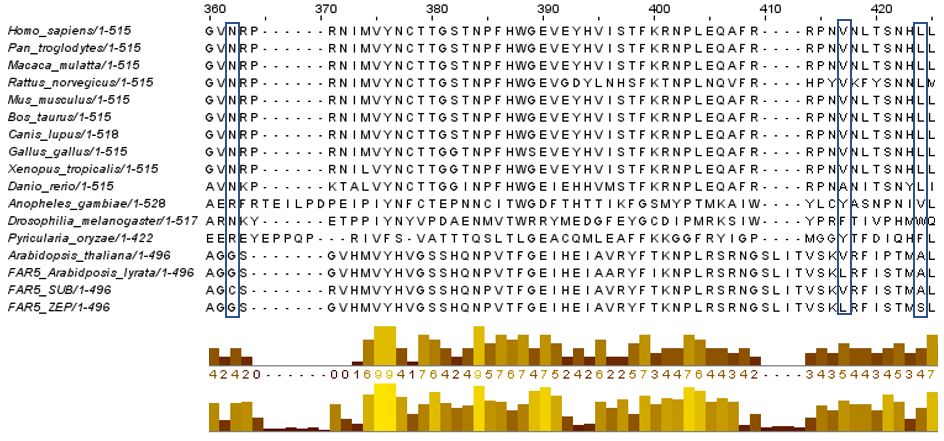
A picture containing chart

Description automatically generated

The BLAST database contains significantly more FAR5 homologues, therefore Figures 13a-c display FAR5 MSAs against fourteen model organisms. Figure 13a exhibits little conservation of previously identified SNPs within the NAD(P)-H binding domain, however characterised variants proximal to the Sterile domain are moderately conserved in 13b. Residue-426 exhibits a high degree of conservation within the Sterile domain - combined with its large Grantham-score, this residue is expected to have significant functional character, with speculated involvement in male-gametogenesis. Contrastingly, Figure 13c demonstrates E457 is absent from other kingdoms, suggesting this residue is biologically unimportant.

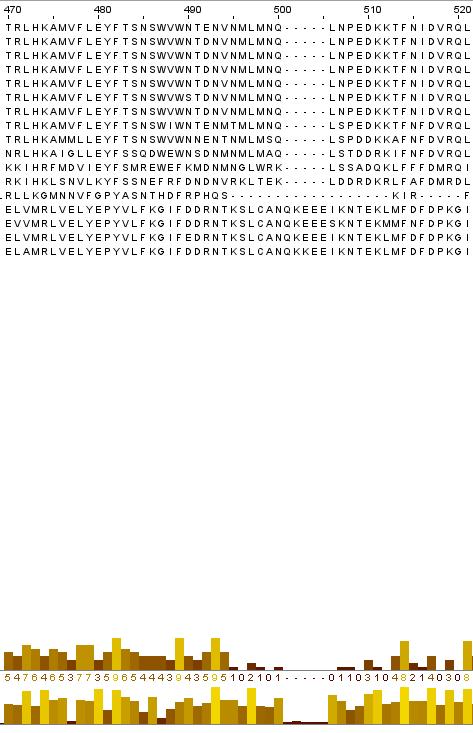
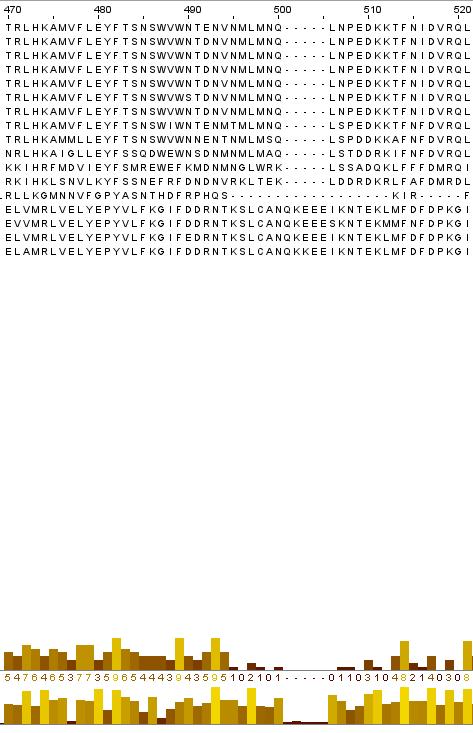
A picture containing table

Description automatically generated



A close-up of a document

Description automatically generated with low confidence



Text

Description automatically generated

We then employed JPred4 to determine whether secondary-structure variation exists between SUB and ZEP populations. This program provides more accurate secondary-structure predictions than alternative software including PSI-PRED [8]. As displayed in Figure 14, no secondary-structure differences exist between SUB and ZEP except a small region of β-sheet in the MAP18 C-terminus - this domain is characterised to contain intrinsic disorder, therefore variation is unlikely to prompt functional consequences.

Application, timeline

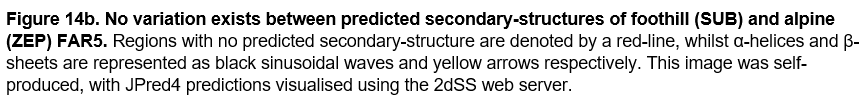
Description automatically generated

Text, letter

Description automatically generated

Table

Description automatically generated



We therefore investigate variation at the tertiary level, selecting two programs to generate homology models for candidate-proteins - RaptorX and AlphaFold. We selected RaptorX because this software successfully models proteins without many sequence homologs, such as MAP18 [13]. AlphaFold produces significantly more accurate backbone and side-chain models than competing programs, prompting selection of this software [14].

Table

Description automatically generatedTable

Description automatically generatedText

Description automatically generatedModel quality is evaluated using MolProbity, to ultimately determine an optimised approach for modelling candidate-proteins [9]. RaptorX generates five near-identical three-dimensional chains per model, therefore chains 2-5 are deleted to allow quality assessment. Tables 5-6 display a range of metrics comparing model quality across software.

Text

Description automatically generated

A picture containing diagram

Description automatically generatedFigure 15 demonstrates alpine-MAP18 adopts more compact structure than foothill protein, with fewer regions of α-helix - particularly in the C-terminus. Whilst this C-terminal is functionally uncharacterised, we speculate such intrinsic disorder actually contains internal structure related to MAP18 activity, and this C-terminal may require reclassification following additional functional assays [18]. Once the C-terminus is hidden in Figure 16, both models appear more similar, although the coiled-coil of foothill-MAP18 appears surprisingly helical, and the region between the first VEEKK-motif and the coiled-coil extends more vertically than its alpine equivalent - eliciting a narrower structure.

Text

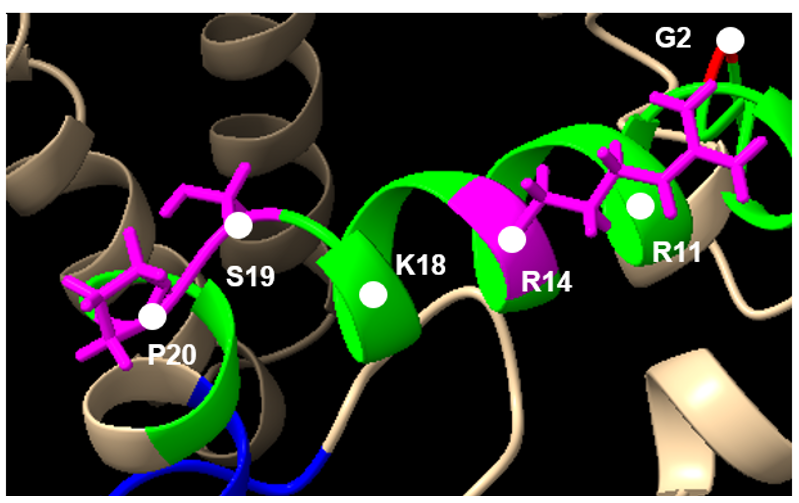
Description automatically generated

Text

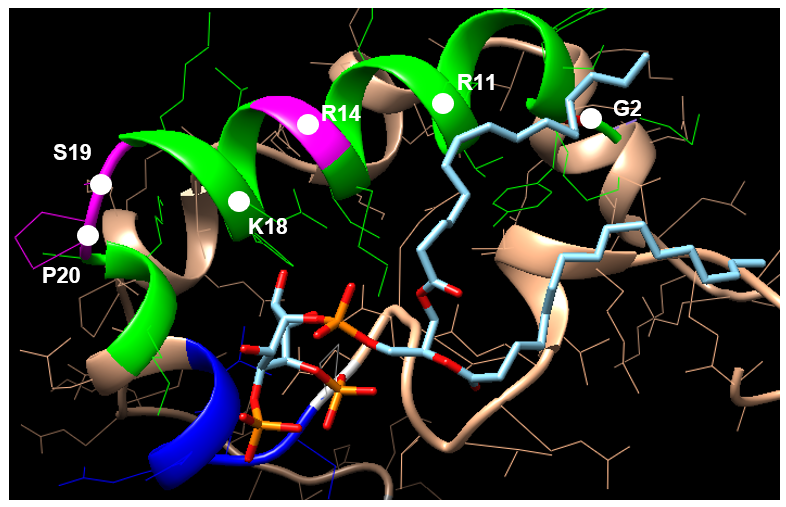
Description automatically generatedA picture containing diagram

Description automatically generated

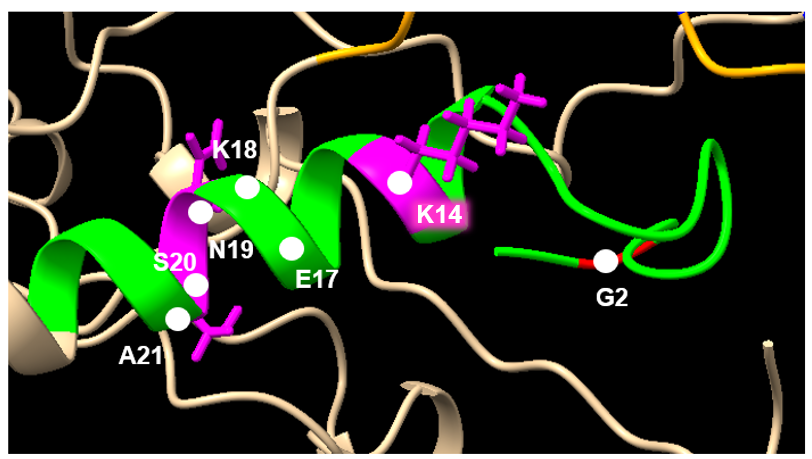
When we analyse the three N-terminal residues implicated in high-altitude adaptation in Figure 17, we observe these amino-acids reside along the same helical face for SUB-MAP18, but along opposite faces for the ZEP-strain. The alpine protein-ligand interface instead contains an acidic E17-residue rather than basic K18, whilst residues-19 and 20 are now separate from the binding-site. Negatively-charged moieties of PtdInsP-ligands will experience less attraction to the N23-domain, and are therefore more frequently available to initiate signalling cascades that mediate stress responses at high-altitude [21].



**SUB MAP18**



***lyrata* MAP18**



**ZEP MAP18**

Text

Description automatically generated

Absolute AFD-values for A59V are beneath 0.5, therefore functional effects of this SNP cannot be interrogated here, as the substituted-residue is not included within alpine consensus-sequence. However, A59 can still be visualised, located within the MAP18 coiled-coil as shown in Figure 18. Alanine residues confer high helix-formation propensity due to their short, non-polar side-chain - it is possible A59V constrains alpine-MAP18 to adopt the conserved coiled-coil motif that may improve protein efficiency.

A picture containing diagram

Description automatically generated

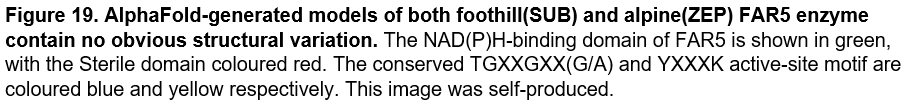
A picture containing text, light, vector graphics

Description automatically generatedA picture containing text, light

Description automatically generatedFigure 19 displays AlphaFold-generated FAR5 models - unlike MAP18, no obviously observable differences exist between populations. As predicted from previous evidence, we identify structurally similar NAD(P)H-binding domains between foothill and alpine models. Figure 20 also demonstrates implicated adaptive S104 and L106 residues are orientated comparably between models - although neither SNP is present.

**SUB FAR5**

**ZEP FAR5**



Background pattern

Description automatically generatedA picture containing text, chain

Description automatically generated

Text

Description automatically generated

We then investigate residues between the NAD(P)H-binding and Sterile domains. As displayed in Figure 21, G329C elicits no conformational change within FAR5; no disulfide-bonds are formed, as no adjacent residues contain sulphur-atoms. FAR5 tertiary-structure is also unaffected by L377V and A384S SNPs. However, in Figure 22 we observe these residues are proximal to the NAD(P)H-binding domain catalytic site, providing additional evidence these amino-acids modify fatty-alcohol synthesis. We speculate the predicted effect of these mutations - to alter synthesised alcohol chain-length - results from modifying the precise binding-site of fatty acyl-CoA substrate within FAR5, changing its subsequent catalysis [23].

A picture containing chain, metalware

Description automatically generatedA picture containing chain, metalware, outdoor object

Description automatically generatedDiagram

Description automatically generated with low confidence

**A384**

**S384**

**L377**

**V377**

**SUB MAP18**

**ZEP MAP18**

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, vector graphics

Description automatically generatedDiagram

Description automatically generated with medium confidenceWe finally analyse three Sterile domain SNPs previously implicated in high-altitude adaptation - V426A, E457K and I490V, displayed in Figures 23a-c respectively. These variants do not alter FAR5 conformation, therefore targeted point-mutageneses of these residues is required to understand biological consequences of their variation. Functional characterisation of the Sterile domain more generally is also required to elucidate the genomic basis for high-altitude adaptation in *A.arenosa*.

**ZEP MAP18**

**K457**

**SUB MAP18**

**ZEP MAP18**

Company name

Description automatically generated

The objective to identify genomic mechanisms underlying parallel high-altitude adaptation of *A.arenosa* has mainly been achieved. We also successfully determined RaptorX and AlphaFold represent optimal software for modelling MAP18 and FAR5 respectively. However, our conclusions could be strengthened through improved functional characterisation of key domains from candidate-proteins. Moreover, interrogation of non-protein coding SNPs could provide further insight into the genomic-basis of functional variation. Nevertheless, we identify three SNPs in the MAP18 N23-domain alongside two interdomain L377V and A384S variants from FAR5 that are strongly implicated in high-altitude adaptation.

**Word Count:** 3584

**Appendices:**

See Submitted Spreadsheet Files

**References:**

1. Wos, G., Mořkovská, J., Bohutínská, M., Šrámková, G., Knotek, A., Lučanová, M., Španiel, S., Marhold, K., & Kolář, F. (2019). Role of ploidy in colonization of alpine habitats in natural populations of *Arabidopsis arenosa.* *Annals of botany,* **124(2):** 255–268.

2. Knotek, A., Konečná, V., Wos, G., Požárová, D., Šrámková, G., Bohutínská, M., Zeisek, V., Marhold, K., & Kolář, F. (2020). Parallel Alpine Differentiation in *Arabidopsis arenosa.* *Frontiers in plant science,* **11:** 561526.

3. Berner D. (2019). Allele Frequency Difference AFD - An Intuitive Alternative to FST for Quantifying Genetic Population Differentiation. *Genes,* **10(4):** 308.

4. Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M…. Kazusa DNA Research Institute (2000). Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana. Nature,* **408(6814):** 820–822.

5. Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E.… European Union Arabidopsis Genome Sequencing Consortium (2000). Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana.* *Nature,* **408(6814):** 823–826.

6. Sievers, F. & Higgins, D. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein science: a publication of the Protein Society,* **27(1):** 135–145.

7. Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.),* **185(4154):** 862–864.

8. Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. (2015). JPred4: a protein secondary structure prediction server. *Nucleic acids research,* **43(1):** 389–394.

9. Davis, I., Leaver-Fay, A., Chen, V.., Block, J., Richardson, J. & Richardson, D.(2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research,* **35:** 375–383.

10. Waterhouse A., Procter J., Martin D., Clamp M. & Barton G. (2009). Jalview Version 2- a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* **25:** 1189-1191.

11. Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25:** 3389-3402.

12. Diksha, P., Cochard, C., Vieira, F. & Bernardes, J. (2019). 2dSS: a web server for protein secondary structure visualization. *BioRxiv,* **649426.**

13. Wang, S., Li, W., Liu, S. & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic acids research,* **44(1):** 430–435.

14. Jumper, J., Evans, R., Pritzel, A. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature,* **596:** 583–589.

15. Pettersen, E., Goddard, T., Huang, C., Meng, E., Couch, G., Croll, T., Morris, J., & Ferrin, T. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein science: a publication of the Protein Society,* **30(1):** 70–82.

16. Kato, M., Nagasaki-Takeuchi, N., Ide, Y., & Maeshima, M. (2010). An *Arabidopsis* hydrophilic Ca2+ -binding protein with a PEVK-rich domain, PCaP2, is associated with the plasma membrane and interacts with calmodulin and phosphatidylinositol phosphates. *Plant & cell physiology,* **51(3):** 366–379.

17. Kato, M., Aoyama, T., & Maeshima, M. (2013). The Ca2+-binding protein PCaP2 located on the plasma membrane is involved in root hair development as a possible signal transducer. *The Plant journal: for cell and molecular biology,* **74(4):** 690–700.

18. Di Domenico, T.; Walsh, I., Martin, A., Tosatto, S. (2012). MobiDB - a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics,* **28(15):** 2080–2081. Gene: *PCAP2.* Organism: *Arabidopsis thaliana (Mouse-ear cress).* Taxon ID: 3702. Reference proteome: UP00000654.

19. Kang, E., Zheng, M., Zhang, Y., Yuan, M., Yalovsky, S., Zhu, L., & Fu, Y. (2017). The Microtubule-Associated Protein MAP18 Affects ROP2 GTPase Activity during Root Hair Growth. *Plant physiology,* **174(1):** 202–222.

20. Wang, X., Zhu, L., Liu, B., Wang, C., Jin, L., Zhao, Q., & Yuan, M. (2007). *Arabidopsis* MAP18 functions in directional cell growth by destabilizing cortical microtubules. *The Plant cell*, **19(3):** 877–889.

21. Wang, Y., Wang, M., Sun, Y., Wang, Y., Li, T., Chai, G., Jiang, W., Shan, L., Li, C., Xiao, E., & Wang, Z. (2015). FAR5, a fatty acyl-coenzyme A reductase, is involved in primary alcohol biosynthesis of the leaf blade cuticular wax in wheat. *Journal of experimental botany,* **66(5):** 1165–1178.

22. Doan, T. Carlsson, A., Stymne, S. & Hofvander, P. (2016). Biochemical characteristics of AtFAR2, a fatty acid reductase from *Arabidopsis thaliana* that reduces fatty acyl-CoA and -ACP substrates into fatty alcohols. *Acta biochimica Polonica,* **63(3):** 565–570.

23. Chacón, M, Fournier, A., Tran, F., Dittrich-Domergue, F., Pulsifer, I., Domergue, F. & Rowland, O. (2013). Identification of amino acids conferring chain length substrate specificities on fatty alcohol-forming reductases FAR5 and FAR8 from *Arabidopsis thaliana.* *The Journal of biological chemistry,* **288(42):** 30345–30355.

24. Pais, F., Ruy, P., Oliveira, G. & Coimbra, R. S. (2014). Assessing the efficiency of multiple sequence alignment programs. *Algorithms for molecular biology,* **9(1):** 4.

25. The UniProt Consortium. (2012). Reorganizing the protein space at the Universal Protein Resource. *Nucleic Acids Research,* **40:** 71–75. Entry name: PCAP2\_ARATH. Primary accession number: Q9LU05.

26. The UniProt Consortium. (2012). Reorganizing the protein space at the Universal Protein Resource. *Nucleic Acids Research,* **40:** 71–75. Entry Name: FAR5C\_ARATH. Primary accession number: Q0WRB0.

27. Missner, A., Kügler, P., Saparov, S. M., Sommer, K., Mathai, J. C., Zeidel, M. L., & Pohl, P. (2008). Carbon dioxide transport through membranes. *The Journal of biological chemistry*, **283(37):** 25340–25347.

28. von Caemmerer, S., & Farquhar, G. (1981). Some relationships between the biochemistry of photosynthesis and the gas exchange of leaves. *Planta,* **153(4):** 376–387.

29. Jiao, P., Tian, G., Li, Y., Deng, G., Jiang, Y.,., Kawaoka, Y., & Chen, H. (2008). A single-amino-acid substitution in the NS1 protein changes the pathogenicity of H5N1 avian influenza viruses in mice. *Journal of virology,* **82(3):** 1146–1154.

30. Grosdidier, A., Zoete, V., & Michielin, O. (2011). SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic acids research*, **39:** 270–277.