Accelerating structure–function mapping using the ViVa webtool to mine natural variation

2018-11-15

Thousands of sequenced genomes are now publically available and represent a large swath of plant evolution; yet, much of this data remains inaccessible to researchers without significant bioinformatics experience. Here, we present a webtool called ViVa (Visualizing Variation) which aims to empower any researcher to take advantage of the amazing genetic resource collected in the Arabidopsis 1001 Genomes Project. ViVa facilates data mining on the gene, gene family or gene network level. To test the utility and accessibility of ViVa, we assembled a team with a range of expertise with biology and bioinformatics to analyze the natural variation in the well-studied nuclear auxin signaling pathway. Our analysis has provided further confirmation of existing knowledge and has also helped generate new hypotheses regarding this well studied pathway. These results highlight how natural variation can be used to generate and test hypotheses about less studied gene families and networks, especially when paired with biochemical and genetic characterization. ViVa is also readily extensible to newly emerging databases of genetic variation, such as the 3,000 Rice Genomes Project.

## List of working titles

* Predicting structure–function relationships using natural genetic variation
* Accelerating structure–function mapping using natural genetic variation
* A comprehensive analysis of natural sequence variation within the Arabidopsis thaliana nuclear auxin signaling pathway
* Visualizing natural variation with ViVa, a new webtool
* Accelerating structure–function mapping using a new webtool to mine natural variation

# Introduction

The first genome sequence of *Arabidopsis thaliana* facilitated rapid advancement of plant biology (The Arabidopsis Genome Initiative [2000](#ref-thearabidopsisgenomeinitiative_analysis_2000)). Since this initial genome, massive scaling of sequencing technology has allowed the survey the genomic variation in natural *Arabidopsis thaliana* populations (Nordborg et al. [2005](#ref-nordborg_pattern_2005); Borevitz et al. [2007](#ref-borevitz_genomewide_2007); Weigel and Mott [2009](#ref-weigel_1001_2009)). This valuable population genetics resource has led to several associations of genetic loci with phenotypic traits and provided insights into how selective pressure has influenced the evolution of plant genomes (Long et al. [2013](#ref-long_massive_2013); Atwell et al. [2010](#ref-atwell_genomewide_2010); Clark et al. [2007](#ref-clark_common_2007)). Despite the high value of this compendium of genetic variation, its use has been largely limited to investigations of population genetics and for genome-wide association studies.

Natural genetic variation provides a catalog of permissible polymorphisms which may facilitate exploration and prediction of genotype/function/phenotype relationships at the gene, gene family and network scales (Joly-Lopez, Flowers, and Purugganan [2016](#ref-joly-lopez_developing_2016)). Outside of large effect variants, these relationships are difficult to map. Such studies in plants have the added obstacles of large gene families and the still quite time and resource-intensive process of allele replacement. Massively parallel assays of variant effects have revolutionized genetic diagnostics and personalized medicine (L. M. Starita et al. [2017](#ref-starita_variant_2017); Gasperini, Starita, and Shendure [2016](#ref-gasperini_power_2016); Matreyek, Stephany, and Fowler [2017](#ref-matreyek_platform_2017)). We envision the use of plant natural variation datasets as a tool to similarly revolutionize breeding and genetic engineering of crop plants by rapidly advancing our understanding of genotype/function/phenotype relationships. Biochemical or synthetic assays of gene function and quantitative phenotyping of the germplasm of collections of natural variation can facilitate functional and phenotypic quantification (R. C. Wright et al. [2017](#ref-wright_insights_2017)). However, the high dimensionality of natural variation datasets often limits access to those with bioinformatics expertise.

Here, we present ViVa: a web application and R-package for Visualizing Variation, which allows plant molecular biologists of any level access to the 1001 genomes database. ViVa facilitates formulation of hypotheses as to the genotype/function/phenotype relationships within any gene, family, or network. Using ViVa researchers may: 1. Identify polymorphisms to facilitate biochemical assays of variant effects (L. M. Starita et al. [2017](#ref-starita_variant_2017); R. C. Wright et al. [2017](#ref-wright_insights_2017)); 2. Produce family-wise alignments of variants to facilitate de novo functional domain identification (Melamed et al. [2015](#ref-melamed_combining_2015)); 3. Generate lists of accessions containing polymorphisms to facilitate phenotypic analysis of gene variant effects (Park et al. [2017](#ref-park_distributions_2017)); and 4. Quantify metrics of genetic diversity to facilitate the study of gene, gene family and network evolution (Delker et al. [2010](#ref-delker_natural_2010); Kliebenstein [2008](#ref-kliebenstein_role_2008)).

# Results and Discussion

## An introduction to ViVa

The current version of ViVa is specifically focused on visualizes natural variation in the coding sequences of genes. Future versions will expand upon this to examine noncoding sequences; however, indels are much more common in noncoding sequences, which limits the certainty of this resequencing data.

**Morgan will provide screenshots** ### Data input: Select the desired genes

### SNPstats: Summary of gene information, structure, and diversity

### Diversity plot: Visualize allelic diversity across the coding sequence

### Accession mapping

### Alignment

### Gene tree

### SNP browser

## Visualizing Variation within the auxin signaling pathway

To test the useability and accessibility of this tool we assembled a group of alpha testers comprising postdoctoral, graduate, and undergraduate researchers. This team of researchers analyzed natural variation in the nuclear auxin signaling pathway. This signaling pathway has been well studied, revealing abundant functional knowledge and crystal structures of several protein domains, providing several points of ground truth reference for comparing structure-function variation and natural genetic variation.

Assuming there is selective pressure against mutations which alter a protein’s function, nonsynonymous genetic variation in critical functional domains of a protein should be low. Therefore scanning gene coding sequences for regions of low nonsynonymous diversity, may predict functional domains. Within the auxin signaling pathway several domains have been identified as critical to transmitting auxin signals.

Perhaps the most critical domain is the degron of the Aux/IAA transcriptional repressor proteins. Mutations which prevent this domain from being recognized and marked for proteasomal degradation by SCFTIR1/AFB are dominant and often result in obvious phenotypes (Berleth, Krogan, and Scarpella [2004](#ref-berleth_auxin_2004); Yang et al. [2004](#ref-yang_iaa1_2004)). As expected this critical functional region is highly conserved at the amino acid level (1). Similarly, the EAR motif which facilitates Aux/IAA interaction with TPL/TPR transcriptional repressors, are also conserved.

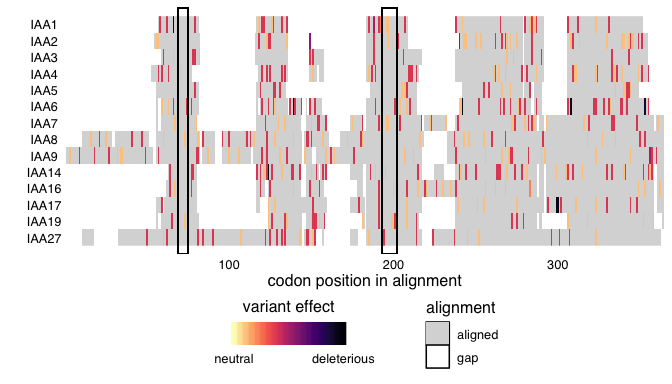


Figure 1 **Canonical IAAs have conserved Degron and EAR motifs.** Protein sequences were aligned (E. S. Wright [2015](#ref-wright_decipher_2015)) and variants were mapped to this alignment and colored according to the predicted functional effect of the most deleterious variant at that position, with light colors being less deleterious and darker colors being more deleterious. Red indicates missense variants. Color scale is explained in [Methods](#methods). EAR motif is highlighted by a black box including alignment positions 70-74. Degron domain is highlighted by a gray box including alignment positions 194-201.

The Aux/IAA proteins bind to ARF transcription factors allowing the Aux/IAA’s to effect transcriptional repression at specific loci. Thus in the ARF transcription factor proteins we expect the DNA-binding function to be conserved. Indead, the DNA-binding domains of the ARFs had very few, low-diversity missense polymorphisms, as did the PB1 domain which facilitates ARF-Aux/IAA interaction (Figure 2).

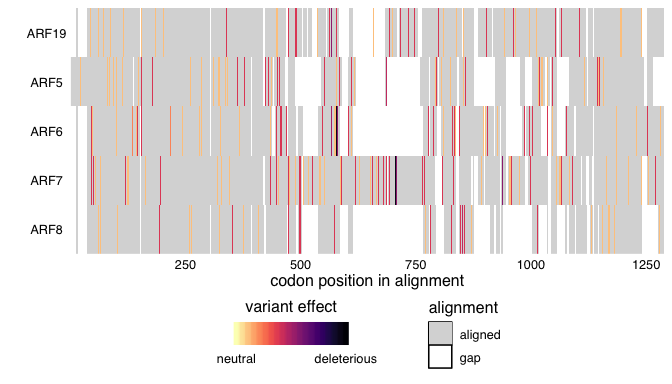


Figure 2 **Canonical ARFs have conserved DNA binding and PB1 domains.** Protein sequences were aligned (E. S. Wright [2015](#ref-wright_decipher_2015)) and variants were mapped to this alignment and colored according to the predicted functional effect of the most deleterious variant at that position, with light colors being less deleterious and darker colors being more deleterious. Red indicates missense variants. Color scale is explained in [Methods](#methods).

Natural variation also provides a means to study how gene families are evolving. Mapping diversity at nonsynonymous variant sites relative to synonymous sites to gene trees allows comparison of recent functional divergence to the sequence divergence throughout the growth of the gene family. This visualization facilitates hypothesis generation regarding the functional redundancy or novelty of recent gene duplications.

Previous research has found evidence of both broad genetic redundancy and specificity within closely related pairs or groups of IAA proteins (Overvoorde et al. [2005](#ref-overvoorde_functional_2005)). For example, the *iaa8-1 iaa9-1* double mutant and the *iaa5-1 iaa6-1 iaa19-1* triple mutant have wild-type phenotypes (Overvoorde et al. [2005](#ref-overvoorde_functional_2005)), yet the *IAA6/19* sister pair has significant differences in expression patterns, protein abundances and functions suggesting they have undergone functional specialization since their divergence (Winkler et al. [2017](#ref-winkler_variation_2017)).

In our analysis of the *IAA* gene family, we frequently observed that one member of most *IAA* sister pairs had high nonsynonymous diversity while the other sister was more conserved 3. In an interspecific comparison between *A. thaliana* and *A. lyrata*, *IAA19* is more conserved, while *IAA6* has regions of increased sequence divergence upstream of the degron and in the N-terminal half of the PB1 domain (Winkler et al. [2017](#ref-winkler_variation_2017)). Consistent with these results, our analysis of intraspecific natural variation revealed higher conservation for *IAA19* ( = 0.55) compared to *IAA6* ( = 2.3) 3, and also detected high diversity within the same regions of *IAA6* as seen in Winkler *et al*.

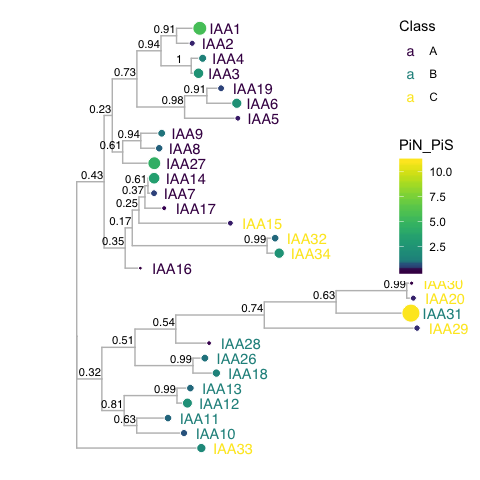


Figure 3 **IAA protein sequence tree mapped with .** Protein sequences were aligned (E. S. Wright [2015](#ref-wright_decipher_2015)) and low information content regions were masked (Kück et al. [2010](#ref-kuck_parametric_2010)) prior to infering a phylogeny (Ronquist and Huelsenbeck [2003](#ref-ronquist_mrbayes_2003)). Tips of the tree are mapped with circles of diameter proportional to and also are colored according to . Nodes are labeled with the poster probability of monophyly.

Within the Group B *IAA* genes, there are four sister pairs—*IAA10*/*IAA11*, *IAA12*/*IAA13*, *IAA18*/*IAA26*, and *IAA20*/*IAA30* 3 (Remington et al. [2004](#ref-remington_contrasting_2004)). Interestingly, among these sister pairs only *IAA12* and *IAA13* have a noticeable difference in nonsynonymous diversity. The *Arabidopsis thaliana* EFP browser shows that *IAA10* and *IAA11* have almost identical expression patterns, with the exception of *IAA10* being more highly expressed in the developing seed, and *IAA11* being more highly expressed in root tissue during lateral root development **need to cite original data source here**. In our analysis, *IAA10* and *IAA11* both showed functional conservation ( of 0.80 and 0.67 respectively) in support of *IAA10* and *IAA11* playing redundant roles. Accessions and associated variants identified here provide a potentially expedited route to elucidating the extent of redundancy between family members such as *IAA10* and *IAA11*.

# Conclusion

# Methods

## Data Sources

### Variant Data

Variant data was queried from the 1001 genomes project (<http://1001genomes.org>) via URL requests to their API service (<http://tools.1001genomes.org/api/index.html>). These queries provided subsets of the whole-genome VCF. The whole-genome VCF file can be found on the project’s website at <http://1001genomes.org/data/GMI-MPI/releases/v3.1/>

### Accession Information

Details on each of the 1135 accessions including CS stock numbers, and geographic location where the samples were collected were taken from the 1001 genomes website at <http://1001genomes.org/accessions.html>, via the download link at the bottom of the page. This data file has been embedded in the R package.

### Gene and transcript information

Information on the genes and transcripts including chromosomal coordinates, start and end location, and transcript length were taken from either the TAIR10 or Araport11 databases. The TAIR10 database, found at <http://arabidopsis.org>, was accessed via the biomart protocol, using the R package biomaRt. The araport11 database, which can also be found on the arabidopsis.org website (<https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FAraport11_genome_release>), has been embedded in the R package.

## Nucleotide Diversity Calculation

Nei defined the nucleotide diversity statistic in his original paper as: “…the average number of nucleotide differences per site between two randomly chosen DNA sequences.” (Nei, 1979) And provided the equation:

Where Where is the frequency of the ith sequence in the population and is the number of sites that are different between the ith and jth sequence divided by sequence length.

a more general form, that treats each sequence in the population as unique can be written as:

where is the nucleotide (A, T, C or G) at position k on the *i*th sequence of the population. is the length of the sequence (the same for all sequences in the population, assumes only SNPs exist, no indels). n is the total number of sequences in the population.

from this form we can re-arrange summations to the form below:

here can be thought of as the site-wise nucleotide diversity at position k, and is equal to the nucleotide diversity of a sequence of length 1 at location k. we can calculate for each site, then average those over the sequence length to calculate , the nucleotide diversity of the sequence.

The function Nucleotide\_diversity in the r1001genomes package calculates for each position in the gene or region that contains a variant. Note, is equal to 0 at all locations without variants. This is also what is displayed in the Diversity Plot tab of the web app.

## Detailed calculation simplification.

The formula for above requires comparing every sequence to every other sequence at location k, however, we know there are only a few variant forms at each individual location.

so we can revert back to using Nei’s original formula (1), modifying it slightly, replacing with , being the number of sequences in the population of the form at location k.

Note that in equation (1) subscripts i and j are summed over all sequences in the population, however in equation (4) i and j are only summed over unique variants at a particular location k.

we will define as the number of sequences different from i at k.

we can see the summed term will be zero if , and if so:

next we substitute our definition of

then distribute and split summation:

finally, summing is equal to n:

this simplified form for is used by the app, because the counts of unique variants at a single nucleotide location can easily be summarized in R.

## Software

The r1001genomes package has many software dependencies on other R packages, a few of the key bioinformatics packages used are listed below.

**biomaRt:** The biomaRt package was used for accessing the TAIR10 database on arabidopsis.org

**vcfR:** vcfR was used to read in the VCF files in a flat “tidy” format for easy manipulaiton

**BSgenome:** The BSgenome package BSgenome.Athaliana.TAIR.TAIR9 was used as the source for the complete DNA string of the reference genome (Col-0).

**DECIPHER:** DECIPHER was used to align nucleotide and amino acid sequences of homologous genes

**GenomicFeatures:** The GenomicFeatures package was used for handling sequence annotations.

**Biostrings:** Biostrings provides the underlying framework for the sequence manipulations used for generating and aligning sequences with BSgenome, Decipher, and GenomicFeatures

# References

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. 2010. “Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines.” *Nature* 465 (7298): 627–31. doi:[10.1038/nature08800](https://doi.org/10.1038/nature08800).

Berleth, Thomas, Naden T. Krogan, and Enrico Scarpella. 2004. “Auxin Signals–Turning Genes on and Turning Cells Around.” *Current Opinion in Plant Biology* 7 (5): 553–63. doi:[10.1016/j.pbi.2004.07.016](https://doi.org/10.1016/j.pbi.2004.07.016).

Borevitz, Justin O., Samuel P. Hazen, Todd P. Michael, Geoffrey P. Morris, Ivan R. Baxter, Tina T. Hu, Huaming Chen, et al. 2007. “Genome-Wide Patterns of Single-Feature Polymorphism in Arabidopsis Thaliana.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (29): 12057–62. doi:[10.1073/pnas.0705323104](https://doi.org/10.1073/pnas.0705323104).

Clark, Richard M., Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, et al. 2007. “Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis Thaliana.” *Science* 317 (5836): 338–42. doi:[10.1126/science.1138632](https://doi.org/10.1126/science.1138632).

Delker, Carolin, Yvonne Pöschl, Anja Raschke, Kristian Ullrich, Stefan Ettingshausen, Valeska Hauptmann, Ivo Grosse, and Marcel Quint. 2010. “Natural Variation of Transcriptional Auxin Response Networks in Arabidopsis Thaliana.” *The Plant Cell* 22 (7): 2184–2200. doi:[10.1105/tpc.110.073957](https://doi.org/10.1105/tpc.110.073957).

Gasperini, Molly, Lea Starita, and Jay Shendure. 2016. “The Power of Multiplexed Functional Analysis of Genetic Variants.” *Nature Protocols* 11 (10): 1782–7. doi:[10.1038/nprot.2016.135](https://doi.org/10.1038/nprot.2016.135).

Joly-Lopez, Zoé, Jonathan M Flowers, and Michael D Purugganan. 2016. “Developing Maps of Fitness Consequences for Plant Genomes.” *Current Opinion in Plant Biology*, SI: 30: Genome studies and molecular genetics, 30 (April): 101–7. doi:[10.1016/j.pbi.2016.02.008](https://doi.org/10.1016/j.pbi.2016.02.008).

Kliebenstein, Daniel J. 2008. “A Role for Gene Duplication and Natural Variation of Gene Expression in the Evolution of Metabolism.” *PLOS ONE* 3 (3): e1838. doi:[10.1371/journal.pone.0001838](https://doi.org/10.1371/journal.pone.0001838).

Kück, Patrick, Karen Meusemann, Johannes Dambach, Birthe Thormann, Björn M. von Reumont, Johann W. Wägele, and Bernhard Misof. 2010. “Parametric and Non-Parametric Masking of Randomness in Sequence Alignments Can Be Improved and Leads to Better Resolved Trees.” *Frontiers in Zoology* 7 (March): 10. doi:[10.1186/1742-9994-7-10](https://doi.org/10.1186/1742-9994-7-10).

Long, Quan, Fernando A Rabanal, Dazhe Meng, Christian D Huber, Ashley Farlow, Alexander Platzer, Qingrun Zhang, et al. 2013. “Massive Genomic Variation and Strong Selection in Arabidopsis Thaliana Lines from Sweden.” *Nature Genetics* 45 (8): 884–90. doi:[10.1038/ng.2678](https://doi.org/10.1038/ng.2678).

Matreyek, Kenneth A., Jason J. Stephany, and Douglas M. Fowler. 2017. “A Platform for Functional Assessment of Large Variant Libraries in Mammalian Cells.” *Nucleic Acids Research* 45 (11): e102. doi:[10.1093/nar/gkx183](https://doi.org/10.1093/nar/gkx183).

Melamed, Daniel, David L. Young, Christina R. Miller, and Stanley Fields. 2015. “Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites.” *PLoS Genet* 11 (2): e1004918. doi:[10.1371/journal.pgen.1004918](https://doi.org/10.1371/journal.pgen.1004918).

Nordborg, Magnus, Tina T Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, et al. 2005. “The Pattern of Polymorphism in Arabidopsis Thaliana.” *PLoS Biol* 3 (7): e196. doi:[10.1371/journal.pbio.0030196](https://doi.org/10.1371/journal.pbio.0030196).

Overvoorde, Paul J., Yoko Okushima, José M. Alonso, April Chan, Charlie Chang, Joseph R. Ecker, Beth Hughes, et al. 2005. “Functional Genomic Analysis of the AUXIN/INDOLE-3-ACETIC ACID Gene Family Members in Arabidopsis Thaliana.” *The Plant Cell* 17 (12): 3282–3300. doi:[10.1105/tpc.105.036723](https://doi.org/10.1105/tpc.105.036723).

Park, Briton, Matthew T. Rutter, Charles B. Fenster, V. Vaughan Symonds, Mark C. Ungerer, and Jeffrey P. Townsend. 2017. “Distributions of Mutational Effects and the Estimation of Directional Selection in Divergent Lineages of Arabidopsis Thaliana.” *Genetics* 206 (4): 2105–17. doi:[10.1534/genetics.116.199190](https://doi.org/10.1534/genetics.116.199190).

Remington, David L., Todd J. Vision, Thomas J. Guilfoyle, and Jason W. Reed. 2004. “Contrasting Modes of Diversification in the Aux/IAA and ARF Gene Families.” *Plant Physiology* 135 (3): 1738–52. doi:[10.1104/pp.104.039669](https://doi.org/10.1104/pp.104.039669).

Ronquist, F., and J. P. Huelsenbeck. 2003. “MrBayes 3: Bayesian Phylogenetic Inference Under Mixed Models.” *Bioinformatics* 19 (12): 1572–4. doi:[10.1093/bioinformatics/btg180](https://doi.org/10.1093/bioinformatics/btg180).

Starita, Lea M., Nadav Ahituv, Maitreya J. Dunham, Jacob O. Kitzman, Frederick P. Roth, Georg Seelig, Jay Shendure, and Douglas M. Fowler. 2017. “Variant Interpretation: Functional Assays to the Rescue.” *The American Journal of Human Genetics* 101 (3): 315–25. doi:[10.1016/j.ajhg.2017.07.014](https://doi.org/10.1016/j.ajhg.2017.07.014).

The Arabidopsis Genome Initiative. 2000. “Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis* *Thaliana*.” *Nature* 408 (6814): 796–815. doi:[10.1038/35048692](https://doi.org/10.1038/35048692).

Weigel, Detlef, and Richard Mott. 2009. “The 1001 Genomes Project for Arabidopsis Thaliana.” *Genome Biology* 10 (5): 107. doi:[10.1186/gb-2009-10-5-107](https://doi.org/10.1186/gb-2009-10-5-107).

Winkler, Martin, Michael Niemeyer, Antje Hellmuth, Philipp Janitza, Gideon Christ, Sophia L. Samodelov, Verona Wilde, et al. 2017. “Variation in Auxin Sensing Guides AUX/IAA Transcriptional Repressor Ubiquitylation and Destruction.” *Nature Communications* 8 (June): 15706. doi:[10.1038/ncomms15706](https://doi.org/10.1038/ncomms15706).

Wright, Erik S. 2015. “DECIPHER: Harnessing Local Sequence Context to Improve Protein Multiple Sequence Alignment.” *BMC Bioinformatics* 16 (October): 322. doi:[10.1186/s12859-015-0749-z](https://doi.org/10.1186/s12859-015-0749-z).

Wright, R. Clay, Mollye L. Zahler, Stacey R. Gerben, and Jennifer L. Nemhauser. 2017. “Insights into the Evolution and Function of Auxin Signaling F-Box Proteins in Arabidopsis Thaliana Through Synthetic Analysis of Natural Variants.” *Genetics* 207 (2): 583–91. doi:[10.1534/genetics.117.300092](https://doi.org/10.1534/genetics.117.300092).

Yang, Xiaoqing, Sungsu Lee, Jai-Hyun So, Suni Dharmasiri, Nihal Dharmasiri, Lei Ge, Carolyn Jensen, Roger Hangarter, Lawrence Hobbie, and Mark Estelle. 2004. “The Iaa1 Protein Is Encoded by Axr5 and Is a Substrate of SCF(TIR1).” *The Plant Journal: For Cell and Molecular Biology* 40 (5): 772–82. doi:[10.1111/j.1365-313X.2004.02254.x](https://doi.org/10.1111/j.1365-313X.2004.02254.x).