A comprehensive analysis of natural sequence variation within the Arabidopsis thaliana nuclear auxin signaling pathway using an accessible web application

2018-07-02

Table of Contents

knitr::opts\_chunk$set(warning = FALSE, collapse = TRUE, comment = "#>", tidy = TRUE, cache = TRUE, cache.lazy = FALSE, out.width = "90%", fig.height=4, fig.width=7, out.extra='style="margin: auto; display: block; padding-top: 15px;"')  
  
#devtools::install\_github("wrightrc/r1001genomes", ref = "auxin-natural-variation")  
library(ape)

## Warning: package 'ape' was built under R version 3.4.4

library(r1001genomes)

## Loading required package: plyr

## Loading required package: DECIPHER

## Loading required package: Biostrings

## Loading required package: BiocGenerics

## Loading required package: parallel

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':  
##   
## clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
## clusterExport, clusterMap, parApply, parCapply, parLapply,  
## parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, append, as.data.frame, cbind, colMeans,  
## colnames, colSums, do.call, duplicated, eval, evalq, Filter,  
## Find, get, grep, grepl, intersect, is.unsorted, lapply,  
## lengths, Map, mapply, match, mget, order, paste, pmax,  
## pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,  
## rowMeans, rownames, rowSums, sapply, setdiff, sort, table,  
## tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##   
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:plyr':  
##   
## rename

## The following object is masked from 'package:base':  
##   
## expand.grid

## Loading required package: IRanges

##   
## Attaching package: 'IRanges'

## The following object is masked from 'package:plyr':  
##   
## desc

## Loading required package: XVector

##   
## Attaching package: 'XVector'

## The following object is masked from 'package:plyr':  
##   
## compact

##   
## Attaching package: 'Biostrings'

## The following object is masked from 'package:ape':  
##   
## complement

## The following object is masked from 'package:base':  
##   
## strsplit

## Loading required package: RSQLite

## Warning: replacing previous import 'Biostrings::collapse' by  
## 'dplyr::collapse' when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::intersect' by  
## 'dplyr::intersect' when loading 'r1001genomes'

## Warning: replacing previous import 'biomaRt::select' by 'dplyr::select'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::setdiff' by  
## 'dplyr::setdiff' when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::setequal' by  
## 'dplyr::setequal' when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::union' by 'dplyr::union'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::failwith' by 'plyr::failwith'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::id' by 'plyr::id' when loading  
## 'r1001genomes'

## Warning: replacing previous import 'dplyr::summarize' by 'plyr::summarize'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::count' by 'plyr::count' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::desc' by 'plyr::desc' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::mutate' by 'plyr::mutate' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::arrange' by 'plyr::arrange' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::rename' by 'plyr::rename' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::summarise' by 'plyr::summarise'  
## when loading 'r1001genomes'

library(DT)  
library(tidyverse)

## ── Attaching packages ──────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 2.2.1 ✔ purrr 0.2.4  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.4  
## ✔ tidyr 0.8.0 ✔ stringr 1.3.0  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ─────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::arrange() masks plyr::arrange()  
## ✖ dplyr::collapse() masks Biostrings::collapse(), IRanges::collapse()  
## ✖ dplyr::combine() masks BiocGenerics::combine()  
## ✖ purrr::compact() masks XVector::compact(), plyr::compact()  
## ✖ dplyr::count() masks plyr::count()  
## ✖ dplyr::desc() masks IRanges::desc(), plyr::desc()  
## ✖ tidyr::expand() masks S4Vectors::expand()  
## ✖ dplyr::failwith() masks plyr::failwith()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::first() masks S4Vectors::first()  
## ✖ dplyr::id() masks plyr::id()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::mutate() masks plyr::mutate()  
## ✖ ggplot2::Position() masks BiocGenerics::Position(), base::Position()  
## ✖ purrr::reduce() masks IRanges::reduce()  
## ✖ dplyr::rename() masks S4Vectors::rename(), plyr::rename()  
## ✖ dplyr::slice() masks XVector::slice(), IRanges::slice()  
## ✖ dplyr::summarise() masks plyr::summarise()  
## ✖ dplyr::summarize() masks plyr::summarize()

library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(RColorBrewer)  
library(ggpmisc)

## For news about 'ggpmisc', please, see http://www.r4photobiology.info/

## For on-line documentation see http://docs.r4photobiology.info/ggpmisc/

library(ggthemes)  
library(ggseqlogo)  
library(ggtree)

## Loading required package: treeio

##   
## Attaching package: 'treeio'

## The following object is masked from 'package:Biostrings':  
##   
## mask

## The following objects are masked from 'package:ape':  
##   
## drop.tip, Nnode, Ntip

## ggtree v1.10.5 For help: https://guangchuangyu.github.io/ggtree  
##   
## If you use ggtree in published research, please cite:  
## Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017, 8(1):28-36, doi:10.1111/2041-210X.12628

##   
## Attaching package: 'ggtree'

## The following object is masked from 'package:magrittr':  
##   
## inset

## The following object is masked from 'package:dplyr':  
##   
## collapse

## The following object is masked from 'package:tidyr':  
##   
## expand

## The following object is masked from 'package:Biostrings':  
##   
## collapse

## The following object is masked from 'package:IRanges':  
##   
## collapse

## The following object is masked from 'package:S4Vectors':  
##   
## expand

## The following object is masked from 'package:plyr':  
##   
## .

## The following object is masked from 'package:ape':  
##   
## rotate

library(reticulate)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(viridis)

## Warning: package 'viridis' was built under R version 3.4.4

## Loading required package: viridisLite

##   
## Attaching package: 'viridis'

## The following object is masked from 'package:scales':  
##   
## viridis\_pal

library(ips)

## Loading required package: colorspace

## Loading required package: XML

## Warning: package 'XML' was built under R version 3.4.4

##   
## Attaching package: 'ips'

## The following object is masked from 'package:treeio':  
##   
## read.beast

library(knitr)  
#use\_python(python = "/usr/local/bin/python")

# List of working titles

* A comprehensive analysis of natural sequence variation within the Arabidopsis thaliana nuclear auxin signaling pathway
* **Add your suggestions**

# Abstract

* Motivating, high level statement
* Statement defining the cutting edge of your research area
* A statement of the form: “However, it remains unclear …” that defines what problem you are  
  addressing.
* A statement of the form:“Here, we …” that concisely states your contribution.
* A series of 3?5 statement about your results.
* A concluding statement.

# Intro

The first genome sequence of *Arabidopsis thaliana* facilitated rapid advancement of plant biology through molecular genetics. Since this initial genome, massive scaling of sequencing technology has allowed a pioneering group to survey the global genomic variation in natural *Arabidopsis thaliana* populations. This valuable population genetics resource has led to several associations of genetic loci with phenotypic traits and provided insights into how selective pressure has and is influencing the evolution of plant genomes. Outside of genome-wide studies however, this valuable dataset has seen little use. Natural genomic variation also provides a set of feasible functional variation at the gene, gene family and gene network level that may provide insight into the function and evolution of genes, families and networks. Here, we present a web application and R-package through which plant molecular biologists with little-to-no bioinformatics experience can make use of this rich dataset of genetic variation to formulate hypotheses as to the sequence/function/phenotype relationships determined by the gene, family, or network of their interest. We demonstrate the utility of this tool through comprehensive analysis and identification of potential functional variation in the nuclear auxin signaling pathway.

This application allows easy access to 1. lists of missense polymorphisms to facilitate biochemical assays of variant effects (Starita et al., [2017](#ref-starita_variant_2017)), 2. family-wise alignments of variants to facilitate de novo functional domain identification (Melamed, Young, Miller, & Fields, [2015](#ref-melamed_combining_2015)), 3. lists of accessions containing missense (or any type of) polymorphisms to facilitate segregation analysis and measurment of the distribution of variant effects on phenotype (Park et al., [2017](#ref-park_distributions_2017)), 4. as well as the study of gene and network evolution (**???**; Delker et al., [2010](#ref-delker_natural_2010)).

Here, we have utilized this tool to analyze the genetic variation in the nuclear auxin signaling pathway and formulate hypotheses regarding the functional implications of this variation and the evolution of the genes and gene families in this pathway. Our analysis also provided further confirmation of much of the existing knowledge of these genes demonstrating the validity of this approach.

# Methods

## 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 ith 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 calculaiton 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, modifying it slightly, replacing with , being the number of sequences in the population of the form at location k: subscripts i and k are summed only over unique forms [need to clarify this]

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.

# Results

## *TIR1/AFB* genes

Auxin acts by binding to receptors that in turn target co-repressors for degradation. Auxin receptors (six in the model plant *Arabidopsis thaliana*) evolved through gene duplication and diversification early in the history of vascular plants (**???**). The rate of co-repressor degradation is determined by the identity of both the receptor and co-repressor (**???**), and this rate sets the pace of lateral root development (**???**).

Although it is unclear what unique roles each receptor plays in growth and development, a number of studies have pointed out differences in the ways the six different receptors in A. thaliana differ in biochemical function and expression domain(Dharmasiri et al., [2005](#ref-dharmasiri_plant_2005); Parry et al., [2009](#ref-parry_complex_2009); Prigge et al., [2016](#ref-prigge_arabidopsis_2016)). Although TIR1/AFBs are expressed ubiquitously in A. thaliana tissues, TIR1, AFB2, and AFB3 have been shown to accumulate in the shoot and root meristems and leaf tissues, with slightly different expression patterns for TIR1 (**???**). Additionally, the expression of AFB5 is stronly circadian-regulated (**???**) and AFB3 is more highly expressed in the roots in the presence of nitrate, allowing increased lateral root formation (**???**), suggesting more broad environmental regulation of this gene family may exist. Functionally, all members of this family have been shown to bind auxin and Aux/IAA proteins. However, AFB1 has drastically reduced ability to assemble into an SCF complex, due to the substitution E8K in its F-box domain, preventing it from inducing degradation of Aux/IAAs(**???**). This lack of complexation may allow observed high ubiquitous AFB1 accumulation (**???**). Higher order mutants in the family containing *afb1* mutants suggest that *AFB1* has a moderate positive effect on auxin signaling. Additionally, AFB4 and AFB5 have been shown to preferentially and functionally bind the synthetic auxin picloram, while other family members preferentially bind indole-3-acetic acid (**???**). Interestingly, the strength and rate with which TIR1/AFBs are able to bind and mark Aux/IAAs for degradation are variable (**???**; Calderón Villalobos et al., n.d.). AFB2 induced the degradation of certain Aux/IAA proteins at a faster rate than TIR1, suggesting some functional specificity has arisen since the initial duplication between the *TIR1/AFB1* and *AFB2/AFB3* clades.

gene\_info <- geneInfoFromFile(fname = system.file("shiny-app", "gene\_family\_data",   
 "AFB\_gene\_ids.csv", package = "r1001genomes"))  
#> [1] "new genes:"  
#> character(0)  
annotation <- readAnnotationFile(filename = system.file("shiny-app", "gene\_family\_data",   
 "AFB\_annotations.csv", package = "r1001genomes"), gene\_info = gene\_info)  
alignment <- alignCDS(gene\_info$transcript\_ID)  
#> Determining distance matrix based on shared 5-mers:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Clustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Aligning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |========== | 16%  
 |   
 |===================== | 33%  
 |   
 |================================ | 50%  
 |   
 |=========================================== | 66%  
 |   
 |====================================================== | 83%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.18 secs  
#>   
#> Iteration 1 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.08 secs  
#>   
#> Alignment converged - skipping remaining iteration.  
# make an alignment data frame  
aln\_df <- makeAlnDF(alignment[[2]])  
  
# make a collapsed alignment for translating positions  
collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)  
  
VCF <- readRDS(system.file("shiny-app", "gene\_family\_data", "AFBs\_VCF.rds",   
 package = "r1001genomes"))  
div\_stats <- ldply(VCF, diversityStats, geneInfo = gene\_info, .id = "transcript\_ID")  
formatRound(datatable(div\_stats, colnames = c("transcript", "symbol", "&pi;<sub>N</sub>",   
 "&pi;<sub>S</sub>", "&pi;<sub>N</sub>/&pi;<sub>S</sub>", "&pi; coding",   
 "&pi; transcript"), rownames = FALSE, escape = FALSE, options = list(paging = FALSE,   
 searching = FALSE)), columns = 2:7, digits = 6)

Examining the natural sequence variation across the *AFB* family revealed that *TIR1* and *AFB1* both had very low nonsynonymous diversity, hinting at their likely functional importance and bringing in to question the inconclusive role of *AFB1* in auxin signaling. *AFB3* and *AFB4* had higher nonsynonymous diversity, while their sister genes, *AFB2* and *AFB5* were more conserved. This matches our current understanding of *AFB3* as playing a minor role in the auxin signaling pathway and *AFB4* perhaps undergoing pseudogenization.

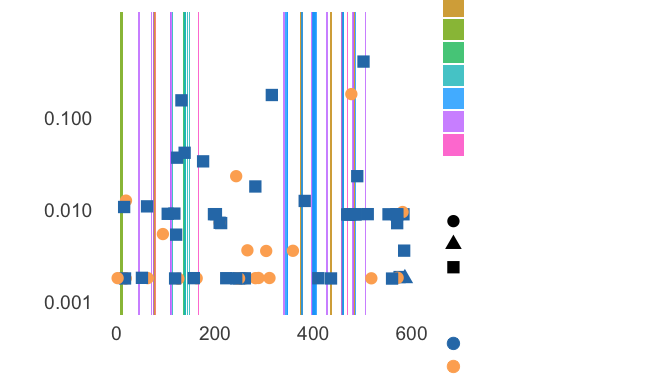
Although most known functional regions are highly conserved in *AFB1*, there is a nonsynonymous polymorphism in the oligomerization domain, only found in the Can-0 accession. Mutations in this domain of *TIR1* frequently have a semidominant effect on root phenotypes (Dezfulian et al., [2016](#ref-dezfulian_oligomerization_2016); Wright, Zahler, Gerben, & Nemhauser, [2017](#ref-wright_insights_2017a)). Characterization of this accession may identify potential *AFB1* phenotypes and may also test the hypothesis that AFB1 functions through oligomerization with other TIR1/AFBs.

coding\_vcf <- ldply(.data = VCF, .fun = subset, !is.na(Transcript\_ID) & gt\_GT !=   
 "0|0")  
aln\_df <- addSNPsToAlnDF(aln\_df, coding\_vcf, seq\_name = Transcript\_ID, seq\_pos = Codon\_Number)  
collapsed\_aln\_df %<>% left\_join(y = subset(aln\_df, seq\_name == gene\_info[gene\_info$tair\_symbol ==   
 "TIR1", "transcript\_ID"]))  
#> Joining, by = "aln\_pos"  
  
#' Select variants for importing into PyMOL  
#'  
#' @param gene\_info a gene info data frame  
#' @param aln\_df an alignment data frame with attached SNPs  
#' @param structure\_ref string, which tair\_symbol or transcript\_ID (depending on   
#' \code{name\_by\_symbols}) corresponds to the protein structure to which   
#' selections should be mapped  
#' @param effect\_type string of the effect type to select from the alignment   
#' data frame  
#' @param name\_by\_symbols logical, if true (default) tair\_symbols will be used   
#' to name the selections, otherwise transcript\_IDs will be used  
#'  
#' @return  
#' @export  
#'  
#' @examples  
select\_variants <- function(gene\_info, aln\_df, structure\_ref, effect\_type = "missense",   
 name\_by\_symbols = TRUE) {  
 collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)  
 if (name\_by\_symbols)   
 column <- "tair\_symbol" else column <- "transcript\_ID"  
   
 selections <- ldply(.data = gene\_info[, column], .fun = function(gene) {  
 collapsed\_aln\_df %<>% dplyr::left\_join(y = subset(aln\_df, seq\_name ==   
 gene\_info[gene\_info[, column] == gene, "transcript\_ID"]))  
 variants <- collapsed\_aln\_df[grepl(effect\_type, collapsed\_aln\_df$effects),   
 gene\_info[gene\_info[, column] == structure\_ref, "transcript\_ID"]]  
 c(gene, paste0(na.omit(as.numeric(variants[[1]])), collapse = "+"))  
 })  
 return(selections)  
}  
  
selections <- select\_variants(gene\_info, aln\_df, structure\_ref = "TIR1")  
#> Joining, by = "aln\_pos"  
#> Joining, by = "aln\_pos"  
#> Joining, by = "aln\_pos"  
#> Joining, by = "aln\_pos"  
#> Joining, by = "aln\_pos"  
#> Joining, by = "aln\_pos"  
#> Joining, by = "aln\_pos"  
selections$V2 <- paste0("chain B and resi ", selections$V2)  
write.table(selections, file = "AFB\_variant\_selections", row.names = FALSE,   
 col.names = FALSE, sep = ",", quote = FALSE)

#  
# -- readSelection.py  
#  
from pymol import cmd, stored  
def readSelection (fileName):  
 '''  
 DESCRIPTION  
 reads a csv file containing a column of selection names and a column of  
 selection strings and creates selection objects for each  
 '''  
 #read in the selection file  
 with open(fileName) as f:  
 lines = f.readlines()  
 for line in lines:  
 sel = line.split(",")  
 cmd.select(sel[0], sel[1])  
 f.close()  
cmd.extend('readSelection',readSelection);

#from within pymol  
cmd.load("Dropbox/180705\_NIH\_renewal/Aim1-figures/2P1Q.pdb")  
hide everything  
  
run ~/Desktop/natural-variation-webtool/manuscript/readSelection.py  
  
readSelection ~/Desktop/natural-variation-webtool/manuscript/AFB\_variant\_selections  
  
# Define selections  
# select TIR1variants, (chain B and resi 154+239+546)  
select fbox, (chain B and resi 1-50)  
select LRR, (chain B and resi 50-576)  
select TIR1, (chain B and resi 1-576)  
# select AFB2variants, (chain B and resi 38+174+181+184+209+259+401+496)  
select tir1\_1, (chain B and resi 147)  
select TIR1dm, (chain B and resi 170+473)  
select insPcontacts, (chain B and resi 74+78+113+114+344+401+403+436+460+484+485+509)  
select dimerNULL, (chain B and resi 142+143+151+117+147)  
select auxin, resn IAC  
select AUXcontacts, (chain B and resi 78+79+82+378+380+403+404+405+438+439+462+463+464+489)  
select insP, resn IHP  
select IAA7, chain C  
select IAA7contacts, (chain B and resi 81+82+84+139+347+349+350+351+352+380+405+406+407+409+464+465+489)  
select IAA7\_N, chain C and resi 1  
select IAA7\_C, chain C and resi 13  
select ASK1, chain A  
select Y, chain C and resi 11  
select Ycontacts, chain B and resi 165+166+195  
  
#Define color palette  
#set\_color name, [ red-float, green-float, blue-float ]  
set\_color blue1, [0.00000000, 0.20392157, 0.6980392]  
set\_color blue2, [0.09803922, 0.36078431, 1.0000000]  
set\_color blue3, [0.00000000, 0.29019608, 1.0000000]  
set\_color teal, [0.09803922, 1.00000000, 0.6980392]  
set\_color LRR, [1.00000000, 0.71764706, 0.0000000]  
set\_color Fbox, [0.69803922, 0.49411765, 0.0000000]  
set\_color TIR1l, [0.67058824, 0.07843137, 0.8000000]  
set\_color TIR1f, [0.63529412, 0.34117647, 0.6980392]  
set\_color AFB2l, [0.37647059, 0.70196078, 1.0000000]  
set\_color AFB2f, [0.07843137, 0.64313725, 0.8000000]  
  
show cartoon, 2P1Q  
  
color lrr, LRR  
  
color TIR1l, TIR1  
show sticks, TIR1  
color TIR1f, AFB1  
show sticks, AFB1  
color AFB2l, AFB2  
show sticks, AFB2  
color AFB2f, AFB3  
show sticks, AFB3  
  
set ray\_trace\_mode, 3  
  
# used get\_view() to return the current view (position) of the molecule   
# copy the code to be able to set the identical viewpoint later  
# to return the current view from Pymol  
# get\_view  
  
set\_view (\  
 0.044671144, 0.083668359, -0.995409012,\  
 0.774909019, 0.625905454, 0.087390237,\  
 0.630409002, -0.775298953, -0.036878210,\  
 -0.001916260, 0.005584979, -202.498916626,\  
 -5.538011074, -8.507736206, 39.046234131,\  
 -219.237945557, 624.123413086, -20.000000000 )  
   
#to render a high definition image use ray  
ray

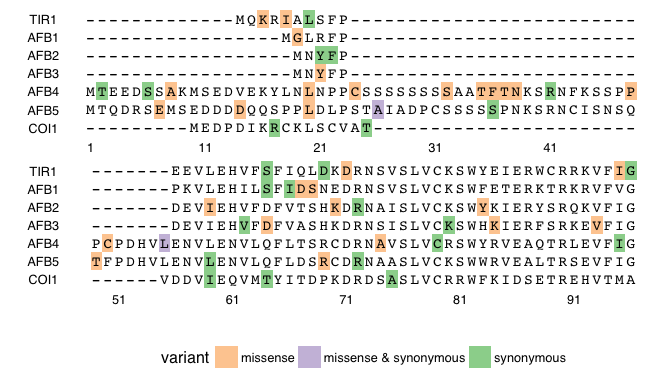
coding\_variants <- getCodingDiv(VCF[[gene\_info[gene\_info$tair\_symbol == "AFB1",   
 "transcript\_ID"]]])  
p <- plotCodingDiv(uniqueCodingVars = coding\_variants)  
annotation <- addAlnPosToAnno(anno\_df = annotation, aln\_df)  
  
  
# translate positions from TIR1 to AFB1  
AFB1\_positions <- annotation$positions  
AFB1\_positions %<>% left\_join(y = select(collapsed\_aln\_df, aln\_pos, gene\_info[gene\_info$tair\_symbol ==   
 "AFB1", "transcript\_ID"]), by = "aln\_pos")  
AFB1\_positions$position <- AFB1\_positions$AT4G03190.1  
  
p <- append\_layers(p, list(geom\_rect(data = AFB1\_positions, mapping = aes(xmin = as.integer(position) -   
 1, xmax = as.integer(position) + 2, fill = annotation), ymin = -Inf, ymax = Inf,   
 inherit.aes = FALSE, alpha = 0.8)), position = "bottom")  
p



The AFB4 and AFB5 receptors have an N-terminal extension prior to the F-box domains. This extension had very high nonsynonymous diversity(Fig. 4), suggesting that this extension does not play an important functional role in these proteins. Additionally, two frameshift variants and one stop-gained variant were observed in *AFB4* supporting its pseudogenization.

Figure 4: Alignment of A. thaliana TIR1/AFB F-box domains showing variants. Cyan marks the F-box binding domain. Yellow, green, and magenta, mark missense and synonymous, synonymous only, and missense only variant locations respectively. Yellow squared E’s show Cul1-interacting residues, differing in AFB1.

aln\_df <- makeAlnDF(alignment[[2]])  
aln\_df <- left\_join(aln\_df, dplyr::select(gene\_info, "tair\_locus", "tair\_symbol",   
 "transcript\_ID"), by = c(transcript\_ID = "transcript\_ID"))  
aln\_df$tair\_symbol <- fct\_rev(fct\_relevel(aln\_df$tair\_symbol, "TIR1"))  
aln\_df <- chunkAlnDF(aln\_df = aln\_df, chunk\_width = 55)  
VCF <- ldply(.data = VCF, .fun = subset, gt\_GT != "0|0")  
VCF <- getCodingDiv(VCF)  
aln\_df <- addSNPsToAlnDF(aln\_df, VCF)  
F\_box <- subset(aln\_df, aln\_pos <= 96)  
F\_box <- chunkAlnDF(aln\_df = F\_box, chunk\_width = 80)  
p <- ggplot(F\_box, aes(x = aln\_pos, y = tair\_symbol, group = seq\_pos, text = variants))  
p <- p + geom\_tile(data = na.omit(F\_box), mapping = aes(fill = effects), width = 1,   
 height = 1, alpha = 0.8) + geom\_text(aes(label = letter), alpha = 1, family = "Courier") +   
 scale\_fill\_brewer(type = "qual", palette = 1, direction = -1) + scale\_x\_continuous(breaks = seq(1,   
 max(F\_box$aln\_pos), by = 10)) + scale\_y\_discrete() + # expand increases distance from axis  
labs(x = "", y = "", fill = "variant") + theme\_logo(base\_family = "Helvetica") +   
 theme(panel.grid = element\_blank(), panel.grid.minor = element\_blank()) +   
 facet\_wrap(facets = ~chunk, ncol = 1, scales = "free") + theme(strip.background = element\_blank(),   
 strip.text.x = element\_blank(), legend.box = "vertical")  
p



## *Aux/IAA* genes

The Aux/IAA proteins (IAAs) mediate both auxin perception and transcriptional co-repression of auxin responsive genes. Auxin perception is carried out along with TIR1/AFB proteins, which together with IAAs act as co-receptors to bind and perceive nuclear auxin. Interaction of IAAs with auxin and TIR1/AFBs is conferred by the highly conserved 13 amino acid degron motif (Tan et al., n.d.). Formation of this auxin co-receptor complex triggers ubiquitination and subsequent degradation of IAAs. This degradation in turn relieves their repression upon ARF transcription factors. Stabilizing mutations within the degron can lead to dominant phenotypes (Berleth, Krogan, & Scarpella, [2004](#ref-berleth_auxin_2004); Yang et al., [2004](#ref-yang_iaa1_2004)). The *Arabidopsis thaliana* Aux/IAA family has 29 members, most of which possess a canonical auxin degron. Notable exceptions are the auxin-insensitive IAA20 and IAA30 that lack a degron sequence entirely, while the closely related IAA31 retains a semi-functional degron sequence with a very slow degradation rate (Dreher, Brown, Saw, & Callis, [2006](#ref-dreher_arabidopsis_2006); Havens et al., n.d.). Additionally, IAA degradation rate is influenced by rate motifs found outside the degron, such as the KR motif (Dreher et al., [2006](#ref-dreher_arabidopsis_2006); Moss et al., [2015](#ref-moss_rate_2015)). IAA-mediated repression of ARFs depends upon both an EAR domain, N-terminal of the degron, that recruits co-repressors of the TOPLESS (TPL) family, and a Phox and Bem1 (PB1) domain that enables interaction via hetero-oligomerization with ARF PB1 domains (Kagale & Rozwadowski, [2011](#ref-kagale_ear_2011); Nanao et al., [2014](#ref-nanao_structural_2014)).

gene\_info <- geneInfoFromFile(fname = system.file("shiny-app", "gene\_family\_data",   
 "IAA\_gene\_ids.csv", package = "r1001genomes"))  
#> [1] "new genes:"  
#> character(0)  
# annotation <- readAnnotationFile(filename = system.file('shiny-app',  
# 'gene\_family\_data', 'IAA\_annotations.csv', package = 'r1001genomes'),  
# gene\_info = gene\_info)  
  
VCF <- readRDS(system.file("shiny-app", "gene\_family\_data", "IAAs\_VCF.rds",   
 package = "r1001genomes"))  
div\_stats <- ldply(VCF, diversityStats, geneInfo = gene\_info, .id = "transcript\_ID")  
div\_stats$tair\_symbol %<>% fct\_reorder(.x = , div\_stats$tair\_symbol, .fun = function(x) {  
 as.integer(str\_remove(x, pattern = "IAA"))  
})  
div\_stats <- formatRound(datatable(div\_stats[order(div\_stats$tair\_symbol), ],   
 colnames = c("transcript", "symbol", "&pi;<sub>N</sub>", "&pi;<sub>S</sub>",   
 "&pi;<sub>N</sub>/&pi;<sub>S</sub>", "&pi; coding", "&pi; transcript"),   
 rownames = FALSE, escape = FALSE, options = list(paging = FALSE, searching = FALSE)),   
 columns = 2:7, digits = 6)  
  
kable(div\_stats$x$data)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | transcript\_ID | tair\_symbol | Pi\_non\_syn | Pi\_syn | Pi\_NS\_Ratio | Pi\_coding | Pi\_transcript |
| 22 | AT4G14560.1 | IAA1 | 0.0009780 | 0.0001767 | 5.5362351 | 0.0011547 | 0.0046573 |
| 18 | AT3G23030.1 | IAA2 | 0.0014169 | 0.0044385 | 0.3192218 | 0.0058554 | 0.0099218 |
| 2 | AT1G04240.1 | IAA3 | 0.0010998 | 0.0004480 | 2.4548590 | 0.0015478 | 0.0029756 |
| 27 | AT5G43700.1 | IAA4 | 0.0032561 | 0.0031912 | 1.0203537 | 0.0064473 | 0.0162985 |
| 6 | AT1G15580.1 | IAA5 | 0.0005073 | 0.0020145 | 0.2518279 | 0.0025218 | 0.0038507 |
| 8 | AT1G52830.1 | IAA6 | 0.0026281 | 0.0011744 | 2.2379097 | 0.0037731 | 0.0060295 |
| 19 | AT3G23050.1 | IAA7 | 0.0013970 | 0.0026000 | 0.5373104 | 0.0039970 | 0.0058084 |
| 11 | AT2G22670.1 | IAA8 | 0.0011169 | 0.0014635 | 0.7631611 | 0.0025805 | 0.0066460 |
| 29 | AT5G65670.1 | IAA9 | 0.0008058 | 0.0008839 | 0.9116481 | 0.0016896 | 0.0059209 |
| 1 | AT1G04100.1 | IAA10 | 0.0004919 | 0.0007381 | 0.6664307 | 0.0012277 | 0.0015522 |
| 23 | AT4G28640.1 | IAA11 | 0.0005417 | 0.0006757 | 0.8017605 | 0.0012174 | 0.0021788 |
| 4 | AT1G04550.1 | IAA12 | 0.0012202 | 0.0005767 | 2.1157879 | 0.0017969 | 0.0034247 |
| 12 | AT2G33310.1 | IAA13 | 0.0001873 | 0.0006675 | 0.2806617 | 0.0008548 | 0.0032696 |
| 21 | AT4G14550.1 | IAA14 | 0.0006636 | 0.0002034 | 3.2624668 | 0.0008670 | 0.0062600 |
| 9 | AT1G80390.1 | IAA15 | 0.0004124 | 0.0015116 | 0.2728366 | 0.0019207 | 0.0055105 |
| 14 | AT3G04730.1 | IAA16 | 0.0000622 | 0.0017632 | 0.0353022 | 0.0018254 | 0.0087302 |
| 3 | AT1G04250.1 | IAA17 | 0.0000502 | 0.0005422 | 0.0926491 | 0.0005924 | 0.0029045 |
| 7 | AT1G51950.1 | IAA18 | 0.0014673 | 0.0012853 | 1.1416256 | 0.0027460 | 0.0083006 |
| 15 | AT3G15540.1 | IAA19 | 0.0023656 | 0.0042675 | 0.5543224 | 0.0066330 | 0.0141514 |
| 13 | AT2G46990.1 | IAA20 | 0.0003983 | 0.0013118 | 0.3036532 | 0.0017102 | 0.0068619 |
| 16 | AT3G16500.1 | IAA26 | 0.0004433 | 0.0004886 | 0.9071937 | 0.0009319 | 0.0038490 |
| 24 | AT4G29080.1 | IAA27 | 0.0004636 | 0.0001038 | 4.4668822 | 0.0005413 | 0.0018219 |
| 26 | AT5G25890.1 | IAA28 | 0.0000654 | 0.0006997 | 0.0935064 | 0.0007651 | 0.0051677 |
| 25 | AT4G32280.1 | IAA29 | 0.0003527 | 0.0008899 | 0.3963790 | 0.0012426 | 0.0025075 |
| 20 | AT3G62100.1 | IAA30 | 0.0001038 | 0.0017895 | 0.0580112 | 0.0018933 | 0.0058015 |
| 17 | AT3G17600.1 | IAA31 | 0.0007073 | 0.0000641 | 11.0405664 | 0.0007714 | 0.0026228 |
| 10 | AT2G01200.1 | IAA32 | 0.0003362 | 0.0004071 | 0.8256960 | 0.0007433 | 0.0044000 |
| 28 | AT5G57420.1 | IAA33 | 0.0005714 | 0.0003498 | 1.6335067 | 0.0009212 | 0.0026045 |
| 5 | AT1G15050.1 | IAA34 | 0.0025070 | 0.0010654 | 2.3530711 | 0.0035659 | 0.0041099 |

# saveWidget(div\_stats, file = 'IAA\_div\_stats.html', )

Within the available natural sequence variation, we found several missense polymorphisms affecting the EAR domain or degron of several canonical IAAs. However, no missense polymorphisms were present in the KR motif or residues critical for PB1-PB1 domain interactions (Supp. Fig. Alignment of canonical IAAs). The majority of missense mutations observed were in the variable regions between the conserved domains.

###Sister Pair Diversification

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 (Figure IAA phylogeny heat map). 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) (Figure IAA phylogeny heat map), and also detected high diversity within the same regions of *IAA6* as seen in Winkler et al. (Supp Fig IAA6 diversity plot).

alignment <- alignCDS(gene\_info$transcript\_ID)  
#> Determining distance matrix based on shared 4-mers:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Clustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Aligning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |== | 3%  
 |   
 |===== | 7%  
 |   
 |====== | 10%  
 |   
 |========= | 14%  
 |   
 |=========== | 17%  
 |   
 |============== | 21%  
 |   
 |================ | 25%  
 |   
 |================== | 28%  
 |   
 |===================== | 32%  
 |   
 |======================= | 35%  
 |   
 |========================= | 39%  
 |   
 |=========================== | 42%  
 |   
 |============================== | 46%  
 |   
 |================================ | 50%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================= | 60%  
 |   
 |========================================== | 64%  
 |   
 |============================================ | 67%  
 |   
 |============================================== | 71%  
 |   
 |================================================= | 75%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================== | 89%  
 |   
 |============================================================ | 92%  
 |   
 |============================================================== | 96%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.37 secs  
#>   
#> Iteration 1 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=== | 5%  
 |   
 |====== | 10%  
 |   
 |========== | 15%  
 |   
 |============= | 20%  
 |   
 |================ | 25%  
 |   
 |==================== | 30%  
 |   
 |======================= | 35%  
 |   
 |========================== | 40%  
 |   
 |============================= | 45%  
 |   
 |================================ | 50%  
 |   
 |==================================== | 55%  
 |   
 |======================================= | 60%  
 |   
 |========================================== | 65%  
 |   
 |============================================== | 70%  
 |   
 |================================================= | 75%  
 |   
 |==================================================== | 80%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================== | 90%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.4 secs  
#>   
#> Iteration 2 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Refining the alignment:  
#>   
 |   
 | | 0%  
 |   
 |======= | 11%  
 |   
 |============================= | 44%  
 |   
 |==================================== | 56%  
 |   
 |=========================================== | 67%  
 |   
 |=================================================== | 78%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.16 secs  
# make an alignment data frame  
aln\_df <- makeAlnDF(alignment[[2]])  
  
# make a collapsed alignment for translating positions  
collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)

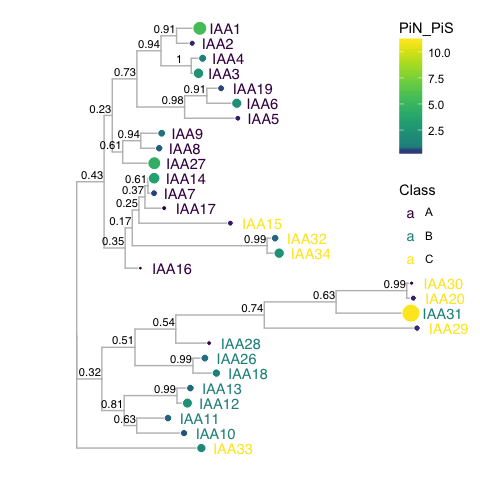
aliscore(as.AAbin(alignment[[2]]), path = "/Volumes/SSD/Applications/ALISCORE\_v2.0/Aliscore\_v.2.0/")  
id <- scan(file = "input.fas\_List\_random.txt", sep = " ", quiet = TRUE)  
AAAlign\_no\_gaps <- as.matrix(alignment[[2]])[, id]  
write.nexus.data(AAAlign\_no\_gaps, file = "manuscript/IAAs.nex", format = "protein")

~/Desktop/natural-variation-webtool/manuscript  
  
/Applications/MrBayes/mb  
  
execute IAAs.nex  
# set IAA33 as the outgroup  
outgroup AT5G57420.1  
  
lset nucmodel=protein Rates=invgamma  
# or include Nst=6 ngammacat=4 omegavar=M3   
prset aamodelpr=fixed(jones) statefreqpr=fixed(empirical)   
# initial run with aamodelpr=mixed yielded posterior prob of 1 for jones model  
propset ExtTBR$prob=0  
  
mcmcp ngen=250000 nchains=4 printfreq=1000 samplefreq=100 burnin=200  
  
mcmc  
sump  
sumt

tree <- treeio::read.beast(file = "IAAs.nex.con.tre")  
  
  
IAA\_data <- read.csv("IAAs\_table.csv")  
# tip\_labels need to be in the first column  
IAA\_data %<>% select(name, everything())  
IAA\_data$name %<>% as.character()  
IAA\_data$AGI %<>% as.character()  
  
IAA\_data <- IAA\_data[match(gsub(x = tree@phylo$tip.label, pattern = "\\.1",   
 replacement = ""), IAA\_data$AGI), ]  
tree@phylo$tip.label <- IAA\_data$name

*I think IAA20, 30, and 15 should also be in class C. So there are really two large clades A and B and then there are truncations/loss of function/dominant variants scattered throughout, Class C.*

p <- ggtree(tree, layout = "rectangular", color = "grey") %<+% IAA\_data + theme\_tree() +   
 scale\_x\_continuous(expand = c(0, 1.2)) + geom\_tippoint(mapping = aes(size = 4 \*   
 PiN\_PiS, fill = PiN\_PiS), shape = 21, color = "white", show.legend = TRUE) +   
 scale\_fill\_viridis(values = rescale(c(min(IAA\_data$PiN\_PiS), 1, max(IAA\_data$PiN\_PiS)),   
 to = c(0, 1)), guide = "colourbar") + scale\_size(guide = "none") + theme(legend.position = c(0.9,   
 0.7)) + geom\_nodelab(mapping = aes(label = round(prob, digits = 2)), nudge\_x = -0.3,   
 nudge\_y = 0.5, size = 3) + geom\_tiplab(mapping = aes(color = Class), hjust = -0.25) +   
 scale\_color\_viridis(discrete = TRUE)  
p



(ref:IAA tree)

(ref:IAA tree) IAA protein sequence tree mapped with

Within the Group B *IAA* genes, there are four sister pairs—*IAA10*/*IAA11*, *IAA12*/*IAA13*, *IAA18*/*IAA26*, and *IAA20*/*IAA30* @ref(IAA tree) (Remington, Vision, Guilfoyle, & Reed, [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*.

### New Loss-of-Function/Gain-of-Function alleles

Accessions with functional variants can be potentially used to investigate the extent of redundancy/specialization within sister pairs. For example, three accessions have a missense variant affecting the EAR domain of IAA10 (HR-5, UKSE06-118 and UKSW06-240, all from southern UK). This variant (L49S) is in the second leucine of the LxLxL motif and likely reduces the ability of IAA10 to recruit the TOPLESS co-repressor (Lee, An, & Cho, [2016](#ref-lee_biological_2016)). The nonsense variant (IAA10-Y219\*) resulting in truncated PB1 domain may be more indicative of *IAA10* null phenotype, as it likely prevents IAA10 from interacting with an ARF. This variant is only found in one accession from northeastern Spain (IP-Svi-0). These accessions and potential loss-of-function alleles may facilitate exploration of the function of *IAA10*.

Several examples of possible gain-of-function missense variants were found in the degron of *IAA2*, *IAA3*, *IAA6*, *IAA7*, *IAA8*, *IAA13*, *IAA15*, *IAA18*, *IAA19*, *IAA27*, *IAA28* and *IAA29*. Most degron variants were very minor alleles in the population, only represented in one or two accessions. Interestingly, degron missense variants occured more frequently in *IAA* genes that have a closely related sister, which may indicate redundant function within the sister-pairs. *I think this is an interesting finding, but I’m not sure about this hypothesis of redundant function. If pairs are acting redundantly and we assume the degron variants are dominant, we would expect the auxin signal they transmit (their function) to be lower, in sum, as a pair. So I would actually hypothesize that there has been some divergence in function as a pair.*

*IAA6* has two potentially dominant missense mutations found within the degron (Supp Fig. alignment), one previously believed to lead to the *shy1-1* phenotype (Kim, Soh, Kang, Furuya, & Nam, [1996](#ref-kim_two_1996)). Interestingly, this variant is widely represented in 31 accessions across Europe (Supplemental Map Fig). This finding supports the proposition that iaa6-C78R may not be the mutation causing the *shy1-1D* phenotype but rather represents natural variation in the degron sequence ([note from Mike Prigge in TAIR](https://www.arabidopsis.org/servlets/TairObject?type=notepad&id=11247)). These accessions could be used to confirm whether *iaa6-C78R* is responsible for the *shy1-1D* phenotype. Additionally, an accession was identified with the IAA6-V77M variant, also in the degron, that may be useful to further explore *iaa6* gain-of-function phenotypes.

### Identifying novel functional sequences in IAAs

An exciting possible use for the Natural Variation Webtool is to search for conserved, low-diversity sites that may represent possible sites of functional import, such as post-translational modification or protein-protein interaction. For example, recent detailed biochemical analysis identified putative lysine ubiquitination sites in members of the *IAA6/19* sister pairs (Winkler et al., [2017](#ref-winkler_variation_2017)); Supp Fig IAA6/19 alignment). These sites are conserved across natural sequence variation, while several other lysine residues vary (Supp Figs alignment and map). In IAA6, there are 17 lysines of which ~6 are putative ubiquitination sites~ *I think this should be 5, K3, K32, K33, K91, and K97. Missense variants occured in K34Q, K140N* . One is proximal to the KR motif, while two others show natural variation. In IAA19, there are 15 total lysines, 8 of which are putative ubiquitination sites (K3, K25, K68, L87, K93, K100, K111, and K141). Two of these ubiquitination sites are proximal to the KR motif and degron. **This sentence is not clear. Only 1 of the ubiquitination sites varies? or one lysine, I think you mean this only 1 lysine that varies, but it is not ubiquitinated. Ubiquitylated Missense K129T** Natural sequence diversity may be used to narrow down a subset of lysines to assess for functional ubiquitination.

coding\_vcf <- ldply(.data = VCF, .fun = subset, !is.na(Transcript\_ID) & gt\_GT !=   
 "0|0")  
aln\_df <- addSNPsToAlnDF(aln\_df, coding\_vcf, seq\_name = Transcript\_ID, seq\_pos = Codon\_Number)  
  
selections <- select\_variants(gene\_info, aln\_df, structure\_ref = "IAA17")  
selections$V2 <- paste0("chain X and resi ", selections$V2)  
write.table(selections, file = "IAA\_variant\_selections", row.names = FALSE,   
 col.names = FALSE, sep = ",", quote = FALSE)

The members of the *IAA8/9/27* clade have a unique N-terminal extension with a highly conserved region common to orthologs in several other species (Dreher et al., [2006](#ref-dreher_arabidopsis_2006)). *IAA8* and *IAA9* play redundant roles in leaf serration (Koenig, Bayer, Kang, Kuhlemeier, & Sinha, [2009](#ref-koenig_auxin_2009)), while *IAA8* also plays an important role in lateral root development (Arase et al., [2012](#ref-arase_iaa8_2012)). IAA8 has been reported to uniquely localize in the cytosol (Arase et al., [2012](#ref-arase_iaa8_2012)) where it may interact with LSD1 protein (Coll, Epple, & Dangl, [2011](#ref-coll_programmed_2011)). The N-terminal extension has low natural sequence diversity in *IAA8* and *IAA9* (Supp Fig IAA8/9 alignment) that may represent a site of novel functionality, possibly mediating interaction with LSD1. The few accessions with missense alleles in this region in both *IAA8* and *IAA9* may be valuable in probing the functionality of this region (Supplemental Fig IAA8/9 map).

Members of the *IAA17/7/14/16* clade have a second EAR motif (LxLxL) with no observed missense variants in *IAA7*, *IAA16*, or *IAA17* (Supp Fig IAA7/14/16/17 alignment). This observation supports the experimental evidence for IAA7 that the second EAR motif is functional, *i.e.* it can interact with TPR1 co-repressor and has a minor repressive role in auxin-related developmental phenotypes [lee\_biological\_2016]. For *IAA14*, there were two missense variants that may be of interest to probe functionality of this second EAR motif. One variant, L41F, in the central lysine is represented in only one accession, and the other in the second “x” (K40N) is found in twelve accessions around Copenhagen (Supp Fig. IAA14 map).

## *TPL/TPR* genes

The Auxin signaling pathway utilizes the TOPLESS (TPL) and TOPLESS-related (TPR) family of Gro/TLE/TUP1 type co-repressor proteins to maintain auxin responsive genes in a transcriptionally-repressed state in the absence of auxin (Szemenyei, Hannon, & Long, [2008](#ref-szemenyei_topless_2008)). In *Arabidopsis thaliana* the five member *TPL/TPR* family includes *TPL* and *TPR1-4*. The resulting proteins are comprised of three structural domains: an N-terminal TPL domain and two WD-40 domains (**???**). TPL/TPR proteins are recruited to the AUX/IAA proteins through interaction with the conserved Ethylene-responsive element binding factor-associated amphiphilic repression (EAR) domain (Szemenyei et al., [2008](#ref-szemenyei_topless_2008)). Canonical EAR domains have the amino acid sequence LxLxL, as found in most AUX/IAAs (figure xA) (Overvoorde et al., [2005](#ref-overvoorde_functional_2005)). TPL/TPR co-repressors bind EAR domains via their C-terminal to LisH (CTLH) domains found near their N-termini (citations of pre-structure founding papers/reviews). Recent structural analyses of the TPL N-terminal domain have highlighted the precise interaction interface between TPL and AUX/IAA EAR domains, as well as the TPL-TPL dimerization and tetramerization motifs (**???**; Martin-Arevalillo et al., [2017](#ref-martin-arevalillo_structure_2017)). The residues required for higher-order multimers of TPL tetramers have also been identified (**???**). Additional interactions with transcriptional regulation and chromatin modifying machinery are likely mediated by two tandem beta propeller domains of TPL/TPRs. These domain have not yet been crystallized, however numerous high resolution crystal structures of beta propeller domains allow for structural prediction of the TPL C-terminal beta propeller domains and the locations of likely interaction faces (Figure, X).

The TOPLESS co-repressor family generally exhibits a high level of sequence conservation at the amino acid sequence level across resequenced *Arabidopsis thaliana* accessions, with all values below 1 (Table X). The closely related *TPL* and *TPR1* have the highest values (TPL-0.425, TPR1-0.739), suggesting that these these two related genes tolerate a higher degree of sequence and potentially functional diversity compared to *TPR2/3/4*. The N-terminal TPL domain of the TPL/TPR family is particularly conserved (Figure X). All nonsynonymous polymorphisms observed in this region are either in the coils between helices or are highly conservative mutations within helices (i.e. Valine to Isoleucine), which would be predicted to exhibit little effect on folding and function.

gene\_info <- geneInfoFromFile(fname = system.file("shiny-app", "gene\_family\_data",   
 "TPL\_gene\_ids.csv", package = "r1001genomes"))  
#> [1] "new genes:"  
#> character(0)  
# annotation <- readAnnotationFile(filename = system.file('shiny-app',  
# 'gene\_family\_data', 'IAA\_annotations.csv', package = 'r1001genomes'),  
# gene\_info = gene\_info)  
  
VCF <- readRDS(system.file("shiny-app", "gene\_family\_data", "TPL\_VCF.rds", package = "r1001genomes"))  
div\_stats <- ldply(VCF, diversityStats, geneInfo = gene\_info, .id = "transcript\_ID")  
div\_stats$tair\_symbol %<>% fct\_reorder(.x = , div\_stats$tair\_symbol, .fun = function(x) {  
 as.integer(str\_remove(x, pattern = "IAA"))  
})  
div\_stats <- formatRound(datatable(div\_stats[order(div\_stats$tair\_symbol), ],   
 colnames = c("transcript", "symbol", "&pi;<sub>N</sub>", "&pi;<sub>S</sub>",   
 "&pi;<sub>N</sub>/&pi;<sub>S</sub>", "&pi; coding", "&pi; transcript"),   
 rownames = FALSE, escape = FALSE, options = list(paging = FALSE, searching = FALSE)),   
 columns = 2:7, digits = 6)  
  
kable(div\_stats$x$data)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | transcript\_ID | tair\_symbol | Pi\_non\_syn | Pi\_syn | Pi\_NS\_Ratio | Pi\_coding | Pi\_transcript |
| 1 | AT1G15750.1 | TPL | 0.0001777 | 0.0004179 | 0.4252898 | 0.0005956 | 0.0010113 |
| 2 | AT1G80490.1 | TPR1 | 0.0003242 | 0.0004382 | 0.7397442 | 0.0007602 | 0.0013711 |
| 4 | AT3G16830.1 | TPR2 | 0.0000776 | 0.0004481 | 0.1732263 | 0.0005258 | 0.0010098 |
| 5 | AT5G27030.1 | TPR3 | 0.0003393 | 0.0017365 | 0.1953753 | 0.0020758 | 0.0035165 |
| 3 | AT3G15880.1 | TPR4 | 0.0001044 | 0.0009106 | 0.1146741 | 0.0010145 | 0.0018425 |

The high degree of conservation in the entire N-terminal domain underscores its importance in TPL/TPR function. For example, the initial *tpl-1* mutation (N176H) in the ninth helix is a dominant gain-of-function allele (**???**), which is capable of binding wild-type TPL protein and inducing protein aggregation (**???**). It is therefore understandable that this helix had very low diversity as nonsynonymous variants in this domain could act in a dominant negative fashion. Addtionally, the crystal structure of TPL has revealed that the The LisH domain and CRA domain coordinate TPL dimerization and tetramerization respectively (26601214, 28698367), the CTLH domain is critical for interaction with DNA-binding transcription factors (26601214), and helix nine within the CRA domain may be critical for the formation of higher-order TPL complexes (multimers of TPL tetramers, 28630893).

Compared to the TPL domain, both the proline rich and WD-40 domains exhibited a higher frequency of missense variants. We utilized Phyre2.0 modeling software to predict the structure of the TPL/TPR WD-40 repeats in order to map non-synonymous mutations to protein structure (Figure X). Here we can see both faces of each predicted beta-propeller structure of the WD-40 domains, which are connected by a short peptide linker. This model reveals amino acid conservation in the core and central residues of the WD-40 domains. ~~These observations could be used as the basis for forward mutagenesis strategies to test whether this region of TPL functions similarly to its homolog TLE, which binds its target genes EH1 and the C-terminal WRPW/Y motif (Hairy/Hes/Runx) in the center of the WD-40 domain (pmid 16762837).~~ **Can we say something more general here? There is not enough background to support this. What is TLE from? What is EH1? Is TLE really binding to EH1 DNA?**

Despite the presence of five *TPL/TPR* genes in the genome, which are considered to be redundant (pmid 16763149), very few observed SNPs are predicted to result in loss of function for any *TPL/TPR* gene. Only one frameshift variant was observed, *TPR4-665delG*, in a single accession (Aiell-4 from Italy). This deletion occurs in codon 222, just after the TPL N-terminal domain, and generates an early stop eleven codons downstream from the deletion. It is unclear whether this N-terminal fragment would be expressed at any appreciable level, as it is possible that the generation of an early stop codon would trigger nonsense-mediated decay. However, if expressed, it would be fascinating to examine whether this naturally occurring N-terminal truncation retains activity as a co-repressor. Given the observation that the TPL N-terminal domain is sufficient to bind histones (28630893), and repress transcription via synthetic assays in yeast (24979769), it is possible that this represents a semi-functional TPL variant. There has been no evidence for prioritization of individual *TPL/TPR* family members in the auxin response pathway, but experiments modulating the gene copy number of *TPR1* revealed the importance of maintaining the proper dosage of *TPL/TPR* in order to accurately trigger the salicylic acid innate immune response (20647385). These results, taken together with the absence of *TPL/TPR* loss-of-function alleles in the *Arabidopsis thaliana* accessions suggest that perhaps the entire family is being conserved at the nucleotide level in the population to meet a gene dosage requirement, with little opportunity for neo- or sub-functionalization.

## ARF genes

Auxin response is ultimately mediated by the auxin responsive transcription factors (ARFs). There are 23 ARFs in *Arabidopsis thaliana* that are divided into three phylogenetic classes. Class A ARFs (ARF5, ARF6, ARF7, ARF8 and ARF19) activate transcription. These ARFs have a glutamine-rich region in the middle of the protein that may mediate activation (Guilefoyle, 2007). It has recently been shown that the middle region of ARF5 interacts with the SWI/SNF chromatin remodeling ATPases BRAMA and SPLAYED, possibly to reduce nucleosome occupancy and allow for the recruitment of transcription machinery (Wu, 2016). Additionally, ARF7 interacts with Mediator subunits, directly tethering transcriptional activation machinery to its binding sites in the chromosome (Ito, 2016). Class B and C ARFs are historically categorized as repressor ARFs, though the mechanism through which they confer repression has not been identified. Their middle regions tend to be proline- and serine-rich (citation, Ulmasov?).

Canonical ARFs are comprised of three major domains. Recent crystallization of these domains have informed structure-function analysis of the ARFs (Boer, 2014; Nanao, 2014; Korasick, 2014). These domains are conserved throughout land plants (Dolf’s new paper). ARFs share an N-terminal B3 DNA binding domain. Flanking this DNA-binding domain is a dimerization domain, which folds up into a single “taco-shaped” domain to allow for dimerization between ARFs. There is an auxiliary domain that immediately follows and interacts with the dimerization domain. The middle region is the most variable between ARFs, as mentioned above, but is characterized by repetitive units of glutamine (class A), serine, or proline residues (classes B and C).

The C-terminal domain of the ARF is a protein-protein interaction domain mediating interactions among ARFs, between ARFs and other transcription factors, and between ARFs and the Aux/IAA repressors. This interaction domain was recently characterized as a Phox and Bem1 (PB1) domain, which is comprised of a positive and negative face with conserved basic and acidic residues, respectively (Nanao, 2014; Korasick, 2014). The dipolar nature of the PB1 domain may mediate multimerization by the pairwise interaction of these faces on different proteins as the ARF7 PB1 domain was crystallized as a multimer (**???**). However, it is unclear whether ARF multimerization occurs or plays a developmental significant role in vivo. Interfering with ARF dimerization in either the DNA-binding proximal dimerization domain or the PB1 domain decreases the ability of class A ARFs to activate transcription in a heterologous yeast system (Pierre-Jerome, 2016).

While domain architecture is broadly conserved among the ARFs, there are exceptional cases. Three ARFs do not contain a PB1 domain at all, ARF3, ARF13, and ARF17, and several more have lost the conserved acidic or basic residues in the PB1 domain, suggesting they may be reduced to a single interaction domain. Several ARFs additionally have an expanded conserved region within the DNA-binding domain, of unknown function. The majority of domain variation among ARFs occurs in the large B-class subfamily. The liverwort *Marchantia polymorpha* has a single representative ARF of each class (Flores-Sandoval, 2015). The expansion of these classes in flowering plants is the result of both whole genome and tandem duplication events (Remington, 2004). The growth of the ARF family may have allowed for the expansion of the quantity and complexity of loci regulated by the ARFs and subsequent expansion in their regulation of developmental processes.

Class A ARFs are the most well-studied ARF subfamily—the five family members all act as transcriptional activators and have well-characterized, distinct developmental targets. Overall the diversity of class A ARFs was generally low, especially compared to the class B and C ARFs, suggesting that class A ARFs are central to auxin signal transduction and plant development. Analysis of class A ARF nonsynonymous diversity suggests that the majority of these ARFs are highly functionally conserved, with values much lower than 1 with the exception of ARF19, with value of 1.8. Comparing diversity within sister pairs, there is a similar trade-off as seen in most *IAA* sister pairs, with one sister being highly conserved and the other more divergent. ARF19 and ARF8 are the more divergent class A ARFs, with values at least three time those of their sisters, ARF7 and ARF6 respectively. This may suggest that ARF6 and ARF7 serve more essential purposes in plant development.

gene\_info <- geneInfoFromFile(fname = system.file("shiny-app", "gene\_family\_data",   
 "ARF\_gene\_ids.csv", package = "r1001genomes"))  
#> [1] "new genes:"  
#> character(0)  
# annotation <- readAnnotationFile(filename = system.file('shiny-app',  
#'gene\_family\_data', 'ARF\_annotations.csv',   
# package = 'r1001genomes'), gene\_info = gene\_info)  
  
VCF <- readRDS(system.file("shiny-app", "gene\_family\_data", "ARF\_VCF.rds", package = "r1001genomes"))  
div\_stats <- ldply(VCF, diversityStats, geneInfo = gene\_info, .id = "transcript\_ID")  
div\_stats$tair\_symbol %<>% fct\_reorder(.x = div\_stats$tair\_symbol, .fun = function(x) {  
 as.integer(str\_remove(x, pattern = "ARF"))  
})  
div\_stats <- formatRound(datatable(div\_stats[order(div\_stats$tair\_symbol), ],   
 colnames = c("transcript", "symbol", "&pi;<sub>N</sub>", "&pi;<sub>S</sub>",   
 "&pi;<sub>N</sub>/&pi;<sub>S</sub>", "&pi; coding", "&pi; transcript"),   
 rownames = FALSE, escape = FALSE, options = list(paging = TRUE, searching = FALSE)),   
 columns = 2:7, digits = 6)  
  
kable(div\_stats$x$data)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | transcript\_ID | tair\_symbol | Pi\_non\_syn | Pi\_syn | Pi\_NS\_Ratio | Pi\_coding | Pi\_transcript |
| 12 | AT1G59750.1 | ARF1 | 0.0010829 | 0.0018316 | 0.5912201 | 0.0029145 | 0.0057536 |
| 23 | AT5G62000.1 | ARF2 | 0.0003347 | 0.0005510 | 0.6074751 | 0.0008837 | 0.0011934 |
| 15 | AT2G33860.1 | ARF3 | 0.0006598 | 0.0004778 | 1.3808762 | 0.0011298 | 0.0011285 |
| 22 | AT5G60450.1 | ARF4 | 0.0007088 | 0.0002019 | 3.5104349 | 0.0009107 | 0.0012847 |
| 2 | AT1G19850.1 | ARF5 | 0.0001846 | 0.0003683 | 0.5011695 | 0.0004925 | 0.0007762 |
| 3 | AT1G30330.1 | ARF6 | 0.0001502 | 0.0007548 | 0.1989807 | 0.0009031 | 0.0011078 |
| 20 | AT5G20730.1 | ARF7 | 0.0006147 | 0.0014901 | 0.4125288 | 0.0021018 | 0.0032876 |
| 21 | AT5G37020.1 | ARF8 | 0.0003166 | 0.0004747 | 0.6668362 | 0.0007913 | 0.0016994 |
| 18 | AT4G23980.1 | ARF9 | 0.0007284 | 0.0022270 | 0.3270608 | 0.0029554 | 0.0037027 |
| 14 | AT2G28350.1 | ARF10 | 0.0008466 | 0.0021845 | 0.3875547 | 0.0030295 | 0.0038111 |
| 16 | AT2G46530.1 | ARF11 | 0.0006729 | 0.0005824 | 1.1554266 | 0.0012282 | 0.0014574 |
| 5 | AT1G34310.1 | ARF12 | 0.0033899 | 0.0014727 | 2.3018584 | 0.0047328 | 0.0053961 |
| 4 | AT1G34170.1 | ARF13 | 0.0090789 | 0.0021896 | 4.1463840 | 0.0100616 | 0.0107053 |
| 10 | AT1G35540.1 | ARF14 | 0.0044175 | 0.0015942 | 2.7708916 | 0.0054870 | 0.0066620 |
| 9 | AT1G35520.1 | ARF15 | 0.0023227 | 0.0006413 | 3.6218117 | 0.0029243 | 0.0035744 |
| 19 | AT4G30080.1 | ARF16 | 0.0003146 | 0.0004215 | 0.7462421 | 0.0007343 | 0.0007074 |
| 13 | AT1G77850.1 | ARF17 | 0.0014579 | 0.0017239 | 0.8456947 | 0.0031818 | 0.0032026 |
| 17 | AT3G61830.1 | ARF18 | 0.0002835 | 0.0002117 | 1.3387989 | 0.0004344 | 0.0009262 |
| 1 | AT1G19220.1 | ARF19 | 0.0007552 | 0.0004125 | 1.8309428 | 0.0011677 | 0.0024942 |
| 8 | AT1G35240.1 | ARF20 | 0.0031381 | 0.0021624 | 1.4512521 | 0.0052198 | 0.0055010 |
| 7 | AT1G34410.1 | ARF21 | 0.0026459 | 0.0013376 | 1.9780448 | 0.0037495 | 0.0040307 |
| 6 | AT1G34390.1 | ARF22 | 0.0025578 | 0.0021914 | 1.1671816 | 0.0044653 | 0.0052844 |
| 11 | AT1G43950.1 | ARF23 | 0.0084679 | 0.0020734 | 4.0841576 | 0.0084942 | 0.0103227 |

# saveWidget(div\_stats, file = 'IAA\_div\_stats.html', )

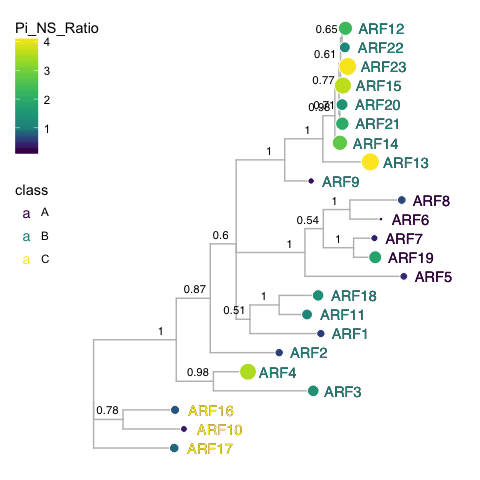
alignment <- alignCDS(gene\_info$transcript\_ID)  
#> Determining distance matrix based on shared 5-mers:  
#>   
 |   
 | | 0%  
 |   
 |===== | 8%  
 |   
 |========== | 16%  
 |   
 |================ | 24%  
 |   
 |===================== | 32%  
 |   
 |========================= | 39%  
 |   
 |============================== | 46%  
 |   
 |================================== | 52%  
 |   
 |====================================== | 58%  
 |   
 |========================================== | 64%  
 |   
 |============================================= | 69%  
 |   
 |=============================================== | 73%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.03 secs  
#>   
#> Clustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |===== | 8%  
 |   
 |========== | 16%  
 |   
 |================ | 24%  
 |   
 |===================== | 32%  
 |   
 |========================= | 39%  
 |   
 |============================== | 46%  
 |   
 |================================== | 52%  
 |   
 |====================================== | 58%  
 |   
 |========================================== | 64%  
 |   
 |============================================= | 69%  
 |   
 |=============================================== | 73%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Aligning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=== | 4%  
 |   
 |====== | 9%  
 |   
 |======== | 13%  
 |   
 |============ | 18%  
 |   
 |============== | 22%  
 |   
 |================== | 27%  
 |   
 |==================== | 31%  
 |   
 |======================= | 36%  
 |   
 |========================== | 40%  
 |   
 |============================= | 45%  
 |   
 |================================ | 50%  
 |   
 |=================================== | 54%  
 |   
 |====================================== | 59%  
 |   
 |========================================= | 63%  
 |   
 |============================================ | 68%  
 |   
 |=============================================== | 72%  
 |   
 |================================================== | 77%  
 |   
 |===================================================== | 81%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================== | 90%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.86 secs  
#>   
#> Iteration 1 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |===== | 8%  
 |   
 |========== | 16%  
 |   
 |================ | 24%  
 |   
 |===================== | 32%  
 |   
 |========================= | 39%  
 |   
 |============================== | 46%  
 |   
 |================================== | 52%  
 |   
 |====================================== | 58%  
 |   
 |========================================== | 64%  
 |   
 |============================================= | 69%  
 |   
 |=============================================== | 73%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |===== | 8%  
 |   
 |========== | 16%  
 |   
 |================ | 24%  
 |   
 |===================== | 32%  
 |   
 |========================= | 39%  
 |   
 |============================== | 46%  
 |   
 |================================== | 52%  
 |   
 |====================================== | 58%  
 |   
 |========================================== | 64%  
 |   
 |============================================= | 69%  
 |   
 |=============================================== | 73%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============= | 20%  
 |   
 |================= | 26%  
 |   
 |===================== | 33%  
 |   
 |========================== | 40%  
 |   
 |============================== | 46%  
 |   
 |================================== | 53%  
 |   
 |======================================= | 60%  
 |   
 |=========================================== | 66%  
 |   
 |=============================================== | 73%  
 |   
 |==================================================== | 80%  
 |   
 |======================================================== | 86%  
 |   
 |============================================================ | 93%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.57 secs  
#>   
#> Iteration 2 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |===== | 8%  
 |   
 |========== | 16%  
 |   
 |================ | 24%  
 |   
 |===================== | 32%  
 |   
 |========================= | 39%  
 |   
 |============================== | 46%  
 |   
 |================================== | 52%  
 |   
 |====================================== | 58%  
 |   
 |========================================== | 64%  
 |   
 |============================================= | 69%  
 |   
 |=============================================== | 73%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |===== | 8%  
 |   
 |========== | 16%  
 |   
 |================ | 24%  
 |   
 |===================== | 32%  
 |   
 |========================= | 39%  
 |   
 |============================== | 46%  
 |   
 |================================== | 52%  
 |   
 |====================================== | 58%  
 |   
 |========================================== | 64%  
 |   
 |============================================= | 69%  
 |   
 |=============================================== | 73%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Refining the alignment:  
#>   
 |   
 | | 0%  
 |   
 |============= | 20%  
 |   
 |========================== | 40%  
 |   
 |======================================= | 60%  
 |   
 |==================================================== | 80%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.59 secs  
# make an alignment data frame  
aln\_df <- makeAlnDF(alignment[[2]])  
  
# make a collapsed alignment for translating positions  
collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)

aliscore(as.AAbin(alignment[[2]]), path = "/Volumes/SSD/Applications/ALISCORE\_v2.0/Aliscore\_v.2.0/")  
id <- scan(file = "/Volumes/SSD/Applications/ALISCORE\_v2.0/Aliscore\_v.2.0/input.fas\_List\_random.txt",   
 sep = " ", quiet = TRUE)  
AAAlign\_no\_gaps <- as.matrix(alignment[[2]])[, id]  
write.nexus.data(AAAlign\_no\_gaps, file = "ARFs.nex", format = "protein")

/Applications/MrBayes/mb  
  
execute ARFs.nex  
# set ARF17 as the outgroup  
outgroup AT1G77850.1  
  
lset nucmodel=protein Rates=invgamma  
# or include Nst=6 ngammacat=4 omegavar=M3   
prset aamodelpr=fixed(jones) statefreqpr=fixed(empirical)   
# initial run with aamodelpr=mixed yielded posterior prob of 1 for jones model  
propset ExtTBR$prob=0  
  
mcmcp ngen=250000 nchains=4 printfreq=1000 samplefreq=100 burnin=200  
  
mcmc  
sump  
sumt

tree <- treeio::read.beast(file = "ARFs.nex.con.tre")  
  
  
ARF\_data <- read.csv(system.file("shiny-app", "gene\_family\_data", "ARF\_gene\_ids.csv",   
 package = "r1001genomes"))  
ARF\_data %<>% left\_join(y = div\_stats$x$data, by = c(name = "tair\_symbol"))  
# tip\_labels need to be in the first column  
ARF\_data %<>% select(name, everything())  
ARF\_data$tair\_locus %<>% as.character()  
  
ARF\_data <- ARF\_data[match(gsub(x = tree@phylo$tip.label, pattern = "\\.1",   
 replacement = ""), ARF\_data$tair\_locus), ]  
  
tree@phylo$tip.label <- ARF\_data$name

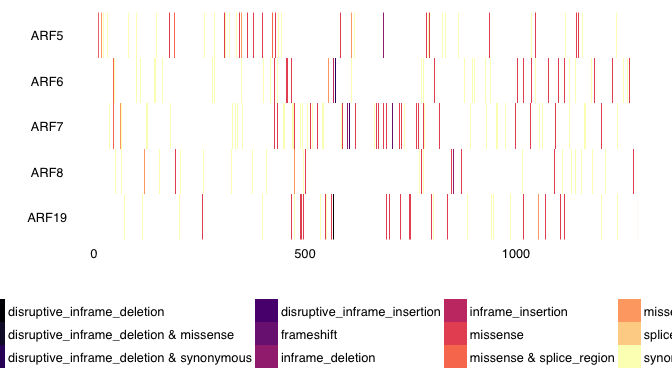
p <- ggtree(tree, layout = "rectangular", color = "grey") %<+% ARF\_data + theme\_tree() +   
 scale\_x\_continuous(expand = c(0, 1.2)) + geom\_tiplab(hjust = -0.25) + geom\_tippoint(mapping = aes(size = 4 \*   
 Pi\_NS\_Ratio, fill = Pi\_NS\_Ratio), shape = 21, color = "white", show.legend = TRUE) +   
 scale\_fill\_viridis(values = rescale(c(min(ARF\_data$Pi\_NS\_Ratio), 1, max(ARF\_data$Pi\_NS\_Ratio)),   
 to = c(0, 1)), guide = "colourbar") + scale\_size(guide = "none") + theme(legend.position = c(0.1,   
 0.7)) + geom\_nodelab(mapping = aes(label = round(prob, digits = 2)), nudge\_x = -0.3,   
 nudge\_y = 0.5, size = 3) + geom\_tiplab(mapping = aes(color = class), hjust = -0.25) +   
 scale\_color\_viridis(discrete = TRUE)  
p



### Class A ARFs

For all class A ARFs, the middle region of the protein was the predominant high diversity region. The DNA-binding domain had very few, low-diversity missense mutations, as did the PB1 domain. Considering the necessity of their conserved functions, the low level of variation in these key DNA and protein-protein interaction domains is expected.

gene\_info <- subset(gene\_info, tair\_symbol %in% c("ARF5", "ARF6", "ARF7", "ARF8",   
 "ARF19"))  
alignment <- alignCDS(gene\_info$transcript\_ID)  
#> Determining distance matrix based on shared 5-mers:  
#>   
 |   
 | | 0%  
 |   
 |========================== | 40%  
 |   
 |============================================== | 70%  
 |   
 |========================================================== | 90%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Clustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |========================== | 40%  
 |   
 |============================================== | 70%  
 |   
 |========================================================== | 90%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Aligning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |================ | 25%  
 |   
 |================================ | 50%  
 |   
 |================================================= | 75%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.23 secs  
#>   
#> Iteration 1 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |========================== | 40%  
 |   
 |============================================== | 70%  
 |   
 |========================================================== | 90%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |========================== | 40%  
 |   
 |============================================== | 70%  
 |   
 |========================================================== | 90%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Alignment converged - skipping remaining iteration.  
# make an alignment data frame  
aln\_df <- makeAlnDF(alignment[[2]])  
# BrowseSeqs(alignment[[2]])  
aln\_df <- makeAlnDF(alignment[[2]])  
aln\_df <- left\_join(aln\_df, dplyr::select(gene\_info, "tair\_locus", "tair\_symbol",   
 "transcript\_ID"), by = c(transcript\_ID = "transcript\_ID"))  
aln\_df$tair\_symbol %<>% as.factor()  
aln\_df$tair\_symbol %<>% fct\_reorder({  
 as.integer(str\_remove(aln\_df$tair\_symbol, pattern = "ARF"))  
}, .desc = TRUE)  
  
  
aln\_df <- chunkAlnDF(aln\_df = aln\_df, chunk\_width = 100)  
VCF <- ldply(.data = VCF, .fun = subset, gt\_GT != "0|0")  
VCF <- getCodingDiv(VCF)  
aln\_df <- addSNPsToAlnDF(aln\_df, VCF)  
A\_ARFs <- subset(aln\_df, tair\_symbol %in% c("ARF5", "ARF6", "ARF7", "ARF8",   
 "ARF19"))  
p <- ggplot(A\_ARFs, aes(x = aln\_pos, y = tair\_symbol, group = seq\_pos, text = variants))  
p <- p + geom\_raster(data = na.omit(A\_ARFs), mapping = aes(fill = effects)) +   
 scale\_fill\_viridis(option = "A", discrete = TRUE) + scale\_y\_discrete() +   
 # expand increases distance from axis  
labs(x = "", y = "", fill = "variant") + theme\_logo(base\_family = "Helvetica") +   
 theme(panel.grid = element\_blank(), panel.grid.minor = element\_blank()) +   
 theme(strip.background = element\_blank(), strip.text.x = element\_blank(),   
 legend.box = "vertical")  
p



# ggsave(p,filename = 'A\_ARFs\_alignment\_variants.pdf', width = 12, height =  
# 4)

Class A ARFs all contain the canonical B3 DNA-binding domain, PB1 interaction domain and glutamine-rich middle region (Guilefoyle, 2007). In the analyzed natural variation, ARF7 had several expansions of polyglutamine sequences in the middle region. Polyglutamine regions are known to readily expand and contract throughout evolutionary time due to replication error, and variation in polyglutamine length can be acted on by natural selection and have phenotypic consequences (Press, 2014).

A few missense variants in ARF5 highlighted potential functional variation in the class A ARFs. A high frequency allele (*What are the variants?*), resulting in substitutions of small amino acids on the positive face of the PB1 domain for bulky phenylalanines, may disrupt interactions with other ARF or Aux/IAA PB1 domains. Interestingly, these variants were always found together in accessions Tu-B2-3, Ru-2, Haes-1, Rd-0, 627ME-13Y1, Le-0, Bu-0 and Bsch-0 *Where are these?*. Another highly represented polymorphism in ARF5 results in change in the 17th amino acid from a leucine to a methionine. This position is in the N-terminal extension of ARF5 which is of unknown function. It is possible that this methionine acts as an alternative start site in the accessions.

### Class B ARFs

*ARF23*, which has a truncated DNA-binding domain, has a high value of 4.1, and has many high-diversity nonsense variants, as may be expected of a gene that is pseudogenizing.

ARF3, ARF13 and ARF17 all lack a C-terminal PB1 domain, which suggests they may be unable to interact with Aux/IAA repressors, possibly acting as auxin-insensitive competitive inhibitors of ARF-mediated transcription (Fig ARF3\_13\_17 alignment). These ARFs span the range of ratios relative to the other ARFs, ranging from 0.85 to 4.1. ARF17 is more conserved than the activator ARF, ARF19. ARF13 has many nonsense variants of high diversity and may be pseudogenizing.

ARF14, ARF15, ARF16, ARF20, and ARF21 all do not have one or multiple of the conserved acidic residues in the PB1 domain, and themselves form a subclade of the B ARFs. Interestingly, no non-synonymous variation is seen in these amino acid residues of these ARFs, suggesting that the amino acids that are in the regions may still be important for the function of these proteins.

### Class C ARFs

Class C ARFs show low nucleotide diversity scores, all values substantially lower than 1. ARF16 was the most conserved, whereas its clade members (ARF10, ARF17) had scores at least four times higher (Remington et al., [2004](#ref-remington_contrasting_2004)).

Structurally, all three members of Class C ARFs contain a canonical B3 DNA-binding domain, but only ARF10 and ARF16 contain a PB1 domain. The DNA binding domains exhibit overall low diversity and presence of largely synonymous changes for all the ARFs. Only the ARF10 DNA binding domain contains missense variants (S147T and A157S) (Fig. ARF10\_DBD Variants distribution). **Any insights into whether these may affect function? They are quite conservative.** Of the PB1 domain containing class C ARFs, ARF16 exhibits several missense variants which are sporadically distributed, in contrast to the conserved PB1 domain of ARF10 (Fig. Arf16\_PB\_variants distribution).

Several members of class B (ARF2, 3 and 4) and C ARFs (ARF10, 16 and 17) are distinguished from other ARFs on the basis of their post transcriptional regulation by trans-acting small interfering RNAs (tasiRNAs) and microRNAs (miRNAs) (Allen, Xie, Gustafson, & Carrington, [2005](#ref-allen_micrornadirected_2005); Rhoades et al., [2002](#ref-rhoades_prediction_2002)). **((???) Arf 6 and 8 targeted by miRNA 167 - check for conservation).** The miRNA target of class C ARFs showed a high degree of conservation (Fig. ARF10\_16\_17\_Alignment). This is consistent with perturbation studies that have shown drastic developmental phenotype for class C ARFs (**???**; Mallory, Bartel, & Bartel, [2005](#ref-mallory_micrornadirected_2005); Wang et al., [2005](#ref-wang_control_2005)). For the class B ARFs (2, 3 and 4) the tasiRNA binding motif is also highly conserved suggesting strong regulatory constraints on these genes (Fig ARF2\_3\_4\_Alignment). The DNA binding domain of ARF4 exhibits non-synonymous variation in 4 accession and would be interesting to carry out phenotypic characterization of these accessions for developmental growth defects (Fig ARF4\_DBD Variants map).

#Discussion

Allen, E., Xie, Z., Gustafson, A. M., & Carrington, J. C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, *121*(2), 207–221. <https://doi.org/10.1016/j.cell.2005.04.004>

Arase, F., Nishitani, H., Egusa, M., Nishimoto, N., Sakurai, S., Sakamoto, N., & Kaminaka, H. (2012). IAA8 involved in lateral root formation interacts with the TIR1 auxin receptor and ARF transcription factors in Arabidopsis. *PloS One*, *7*(8), e43414. <https://doi.org/10.1371/journal.pone.0043414>

Berleth, T., Krogan, N. T., & Scarpella, E. (2004). Auxin signals–turning genes on and turning cells around. *Current Opinion in Plant Biology*, *7*(5), 553–563. <https://doi.org/10.1016/j.pbi.2004.07.016>

Calderón Villalobos, L. I. A., Lee, S., De Oliveira, C., Ivetac, A., Brandt, W., Armitage, L., … Estelle, M. (n.d.). A combinatorial TIR1/AFB-Aux/IAA co-receptor system for differential sensing of auxin. *Nat Chem Biol*, *8*(5), 477–85. <https://doi.org/10.1038/nchembio.926>

Coll, N. S., Epple, P., & Dangl, J. L. (2011). Programmed cell death in the plant immune system. *Cell Death and Differentiation*, *18*(8), 1247–1256. <https://doi.org/10.1038/cdd.2011.37>

Delker, C., Pöschl, Y., Raschke, A., Ullrich, K., Ettingshausen, S., Hauptmann, V., … Quint, M. (2010). Natural Variation of Transcriptional Auxin Response Networks in Arabidopsis thaliana. *The Plant Cell*, *22*(7), 2184–2200. <https://doi.org/10.1105/tpc.110.073957>

Dezfulian, M. H., Jalili, E., Roberto, D. K. A., Moss, B. L., Khoo, K., Nemhauser, J. L., & Crosby, W. L. (2016). Oligomerization of SCF TIR1 Is Essential for Aux/IAA Degradation and Auxin Signaling in Arabidopsis. *PLOS Genet*, *12*(9), e1006301. <https://doi.org/10.1371/journal.pgen.1006301>

Dharmasiri, N., Dharmasiri, S., Weijers, D., Lechner, E., Yamada, M., Hobbie, L., … Estelle, M. (2005). Plant Development Is Regulated by a Family of Auxin Receptor F Box Proteins. *Developmental Cell*, *9*(1), 109–119. <https://doi.org/10.1016/j.devcel.2005.05.014>

Dreher, K. A., Brown, J., Saw, R. E., & Callis, J. (2006). The Arabidopsis Aux/IAA Protein Family Has Diversified in Degradation and Auxin Responsiveness. *The Plant Cell*, *18*(3), 699–714. <https://doi.org/10.1105/tpc.105.039172>

Havens, K. A., Guseman, J. M., Jang, S. S., Pierre-Jerome, E., Bolten, N., Klavins, E., & Nemhauser, J. L. (n.d.). A synthetic approach reveals extensive tunability of auxin signaling. *Plant Physiol*, *160*(1), 135–42. <https://doi.org/10.1104/pp.112.202184>

Kagale, S., & Rozwadowski, K. (2011). EAR motif-mediated transcriptional repression in plants: An underlying mechanism for epigenetic regulation of gene expression. *Epigenetics*, *6*(2), 141–146.

Kim, B. C., Soh, M. S., Kang, B. J., Furuya, M., & Nam, H. G. (1996). Two dominant photomorphogenic mutations of Arabidopsis thaliana identified as suppressor mutations of hy2. *The Plant Journal*, *9*(4), 441–456. <https://doi.org/10.1046/j.1365-313X.1996.09040441.x>

Koenig, D., Bayer, E., Kang, J., Kuhlemeier, C., & Sinha, N. (2009). Auxin patterns Solanum lycopersicum leaf morphogenesis. *Development (Cambridge, England)*, *136*(17), 2997–3006. <https://doi.org/10.1242/dev.033811>

Lee, M.-S., An, J.-H., & Cho, H.-T. (2016). Biological and molecular functions of two EAR motifs of Arabidopsis IAA7. *Journal of Plant Biology*, *59*(1), 24–32. <https://doi.org/10.1007/s12374-016-0453-1>

Mallory, A. C., Bartel, D. P., & Bartel, B. (2005). MicroRNA-directed regulation of Arabidopsis AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. *The Plant Cell*, *17*(5), 1360–1375. <https://doi.org/10.1105/tpc.105.031716>

Martin-Arevalillo, R., Nanao, M. H., Larrieu, A., Vinos-Poyo, T., Mast, D., Galvan-Ampudia, C., … Parcy, F. (2017). Structure of the Arabidopsis TOPLESS corepressor provides insight into the evolution of transcriptional repression. *Proceedings of the National Academy of Sciences*, 201703054. <https://doi.org/10.1073/pnas.1703054114>

Melamed, D., Young, D. L., Miller, C. R., & Fields, S. (2015). Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites. *PLoS Genet*, *11*(2), e1004918. <https://doi.org/10.1371/journal.pgen.1004918>

Moss, B. L., Mao, H., Guseman, J. M., Hinds, T. R., Hellmuth, A., Kovenock, M., … Nemhauser, J. L. (2015). Rate Motifs Tune Auxin/Indole-3-Acetic Acid Degradation Dynamics. *Plant Physiology*, *169*(1), 803–813. <https://doi.org/10.1104/pp.15.00587>

Nanao, M. H., Vinos-Poyo, T., Brunoud, G., Thévenon, E., Mazzoleni, M., Mast, D., … Dumas, R. (2014). Structural basis for oligomerization of auxin transcriptional regulators. *Nature Communications*, *5*, 3617. <https://doi.org/10.1038/ncomms4617>

Overvoorde, P. J., Okushima, Y., Alonso, J. M., Chan, A., Chang, C., Ecker, J. R., … Theologis, A. (2005). Functional Genomic Analysis of the AUXIN/INDOLE-3-ACETIC ACID Gene Family Members in Arabidopsis thaliana. *The Plant Cell*, *17*(12), 3282–3300. <https://doi.org/10.1105/tpc.105.036723>

Park, B., Rutter, M. T., Fenster, C. B., Symonds, V. V., Ungerer, M. C., & Townsend, J. P. (2017). Distributions of Mutational Effects and the Estimation of Directional Selection in Divergent Lineages of Arabidopsis thaliana. *Genetics*, *206*(4), 2105–2117. <https://doi.org/10.1534/genetics.116.199190>

Parry, G., Calderon-Villalobos, L. I., Prigge, M., Peret, B., Dharmasiri, S., Itoh, H., … Estelle, M. (2009). Complex regulation of the TIR1/AFB family of auxin receptors. *Proceedings of the National Academy of Sciences*, *106*(52), 22540–22545. <https://doi.org/10.1073/pnas.0911967106>

Prigge, M. J., Greenham, K., Zhang, Y., Santner, A., Castillejo, C., Mutka, A. M., … Estelle, M. (2016). The Arabidopsis Auxin Receptor F-Box Proteins AFB4 and AFB5 Are Required for Response to the Synthetic Auxin Picloram. *G3: Genes|Genomes|Genetics*, *6*(5), 1383–1390. <https://doi.org/10.1534/g3.115.025585>

Remington, D. L., Vision, T. J., Guilfoyle, T. J., & Reed, J. W. (2004). Contrasting modes of diversification in the Aux/IAA and ARF gene families. *Plant Physiology*, *135*(3), 1738–1752. <https://doi.org/10.1104/pp.104.039669>

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., & Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, *110*(4), 513–520.

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., … Fowler, D. M. (2017). Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics*, *101*(3), 315–325. <https://doi.org/10.1016/j.ajhg.2017.07.014>

Szemenyei, H., Hannon, M., & Long, J. A. (2008). TOPLESS mediates auxin-dependent transcriptional repression during Arabidopsis embryogenesis. *Science (New York, N.Y.)*, *319*(5868), 1384–1386. <https://doi.org/10.1126/science.1151461>

Tan, X., Calderon-Villalobos, L. I. A., Sharon, M., Zheng, C., Robinson, C. V., Estelle, M., & Zheng, N. (n.d.). Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature*, *446*(7136), 640–5. <https://doi.org/10.1038/nature05731>

Wang, J.-W., Wang, L.-J., Mao, Y.-B., Cai, W.-J., Xue, H.-W., & Chen, X.-Y. (2005). Control of root cap formation by MicroRNA-targeted auxin response factors in Arabidopsis. *The Plant Cell*, *17*(8), 2204–2216. <https://doi.org/10.1105/tpc.105.033076>

Winkler, M., Niemeyer, M., Hellmuth, A., Janitza, P., Christ, G., Samodelov, S. L., … Calderón Villalobos, L. I. A. (2017). Variation in auxin sensing guides AUX/IAA transcriptional repressor ubiquitylation and destruction. *Nature Communications*, *8*, 15706. <https://doi.org/10.1038/ncomms15706>

Wright, R. C., Zahler, M. L., Gerben, S. R., & Nemhauser, J. L. (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), 583–591. <https://doi.org/10.1534/genetics.117.300092>

Yang, X., Lee, S., So, J.-H., Dharmasiri, S., Dharmasiri, N., Ge, L., … Estelle, M. (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), 772–782. <https://doi.org/10.1111/j.1365-313X.2004.02254.x>