Data annotation of *Epichloe amarillans* and *bromicola* genes

Imogen Dumville

27/02/2020

**Document purpose** This R markdown document describes production of datasets for *Epichloe amarillans* and *bromicola* species.

1. Lineage specific genes (LSG) are isolated as potenital sources of lineage-specific evolutoon, formation of new genes or increased evoltuionary rate.
2. The length of proteins produced is isolated. If LSG are much shorter than non-lineage speciifc genes they may be ‘new’ genes.
3. The distnace of LSG to the nearest ATrich repeat region is measured to determine potential of gene production.
4. Quantification of genes is carried out in datasets with RNAseq data ( *amarillans* ) to determine expression of LSG.
5. Gene ID, lineage specificity, length, TPM, and distance to AT regions is tabulated.
6. GO functional annotation of LSG is carried out and displayed in a seperate table to confer potential functions of new or altered genes.
7. Data is statsitically and graphically analysed.
8. This project has huge opportunity for further study.

**Making lineage specific gene datasets.** *amarillans* shown.

orthos <- read.table("ortho\_long.tsv", stringsAsFactors=FALSE, col.names=c("ortho\_id", "gene\_id", "strain")) #load dataset  
# convert to named list, with strains represented by each ortho  
by\_ortho <- tapply(orthos$strain, orthos$ortho\_id, FUN=unique)   
  
strains <- unique(orthos$strain)  
  
#fxn to find orthos with only one set of IDs  
lineage\_specific <- function(ortho\_set, lineage\_ids){  
 all(ortho\_set %in% lineage\_ids) & length(ortho\_set) == length(lineage\_ids)  
}  
  
# which indices in by\_ortho are amarillans speific.  
all\_amarillans <- sapply(by\_ortho, lineage\_specific, c("EamaE4668", "Eam702", "EamaE57")) #all amarillans clade  
  
long\_amarillans <- sapply(by\_ortho, lineage\_specific, c("Eam702")) #just long read  
trueamaog <- data.frame(which(long\_amarillans == TRUE))  
colnames(trueamaog) <- "num"  
amaorthonum <- by\_ortho[trueamaog$num]  
amaorthos <- data.frame(names(amaorthonum))  
  
amageneID <- orthos[c(which(orthos$ortho\_id %in% amaorthos$names.amaorthonum.)),]  
write.table(amageneID$gene\_id, file = "amarillansuniquegeneid.tsv", row.names = FALSE, col.names = FALSE, quote = FALSE)

There were 142 genes shared by all *amarillans* and no other clades There were 302 genes in Eam702 and no other strains

There were 32 genes shared by all *bromicola* and no other clades There are 394 genes in long read *bromicola* genome not found in any other strains

Dataset is produced from Eam702 and EbroNfe1 as these are long read genomes. Head of unqiue amarillans tsv is shown

## Eam702\_000028.T1  
## 1 Eam702\_000039-T1  
## 2 Eam702\_000046-T1  
## 3 Eam702\_000048-T1  
## 4 Eam702\_000086-T1  
## 5 Eam702\_000091-T1  
## 6 Eam702\_000104-T1

**Quantification of *amarillans* genes is carried out using RNAseq data.** This was carried out in shell, using Salmon. File; Eamquant.sf; is produced. For more information see <https://github.com/Imogen-D/Cyclic-Peptide-Expression>

salmon index -t <transcripts.fa> -i transcripts\_index -k 31 salmon quant -i transcripts\_index -l A -r reads.fq –validateMappings -p 8 -o transcripts\_quant scp transcripts\_quant/quant.sf ./quant.sf

**Calculation of distance to AT rich repeat regions** This is carried out in shell, using bedtools. gff files are supplied with locations of AT regions.

bedtools closest -d -a Epichloe\_amarillans.gff3 -b Eam\_AT\_rich.gff | grep “gene” | cut -f9,18 > closest\_AT.tsv

**Functional Annotation** To gain functional annotation, protein files from Eam702 and EbroNfe1 were uploaded to <http://ekhidna2.biocenter.helsinki.fi/sanspanz/>. This returned a dataset of GO annotations, titled AmarillansAnnoout.txt and BromicolaAnnoout.txt respectively. They were read into R to undergo GO enrichment, following code from <https://github.com/dwinter/genome_factory/wiki/Taking-a-gene-list-and-PANNZER-output-through-to-GO-term-enrichment> *amarillans* code shown, with *amarillans* and *bromicola* datasets displayed

AmarillansAnnoout <- read.delim("~/CoxExtension/GO output/AmarillansAnnoout.txt", stringsAsFactors=FALSE, na.strings = "n.d.")  
  
subsetammaGO <- AmarillansAnnoout[which(AmarillansAnnoout$PPV >= 0.5), c(1, 5)]  
  
go\_filt <- AmarillansAnnoout[which(AmarillansAnnoout$PPV >= 0.5),]   
go\_filt$id <- paste0('GO:', go\_filt$id)  
panzer\_to\_golist <- function(panzer\_df){  
 go\_df <- aggregate( id ~ qpid, data=panzer\_df, FUN=c)  
 structure(go\_df$id, .Names=go\_df$qpid)  
}  
all\_golist <- panzer\_to\_golist(go\_filt)  
  
  
make\_topGO\_DO <- function(gene\_list, ontology, gene2GO\_list){  
 topGO\_data <- new("topGOdata", ontology = ontology, allGenes = gene\_list,  
 annot = annFUN.gene2GO, gene2GO = gene2GO\_list)  
 fishers\_result <- runTest(topGO\_data, algorithm = "elim", statistic = "fisher")  
 fishers\_table <- GenTable(topGO\_data, Fishers = fishers\_result, useLevels = TRUE)  
 fishers\_table$Ontology <- ontology  
 fishers\_table  
}  
  
amarillansuniquegeneid <- read.table("~/CoxExtension/LineageSpecificity/amarillansuniquegeneid.tsv", quote="\"", comment.char="", stringsAsFactors=FALSE)  
ortho\_long <- read.delim("~/CoxExtension/ortho\_long.tsv", header=FALSE, stringsAsFactors=FALSE)  
Eam702 <- ortho\_long[which(ortho\_long$V3 == "Eam702"),]  
uniquegenes <- amarillansuniquegeneid$V1  
geneList <- factor(as.integer(Eam702$V2 %in% uniquegenes))  
names(geneList) <- Eam702$V2  
  
  
topGO\_BP\_table <- make\_topGO\_DO(geneList, "BP", all\_golist)  
topGO\_MF\_table <- make\_topGO\_DO(geneList, "MF", all\_golist)  
topGO\_CC\_table <- make\_topGO\_DO(geneList, "CC", all\_golist)  
  
topGO\_all\_table <- (rbind(topGO\_BP\_table, topGO\_MF\_table, topGO\_CC\_table))  
topGO\_all\_table <- topGO\_all\_table[order(topGO\_all\_table$Fishers),]  
  
write.csv(topGO\_all\_table, file = "topGOfull.csv", quote = FALSE)  
head(topGO\_all\_table)

## GO.ID Term Level Annotated  
## 11 GO:0030619 U1 snRNA binding 7 2  
## 1 GO:0006122 mitochondrial electron transport, ubiqui... 15 4  
## 2 GO:0006030 chitin metabolic process 7 6  
## 21 GO:0005685 U1 snRNP 12 6  
## 22 GO:0043189 H4/H2A histone acetyltransferase complex 15 8  
## 12 GO:0008061 chitin binding 4 11  
## Significant Expected Fishers Ontology  
## 11 2 0.03 0.00019 MF  
## 1 2 0.05 0.0010 BP  
## 2 2 0.08 0.0025 BP  
## 21 2 0.10 0.0040 CC  
## 22 2 0.14 0.0073 CC  
## 12 2 0.15 0.00942 MF

Genes included relate to chitin porduction, a key influencer in the interaction of *Epichloe* and grass. They also include key componenets in RNA splicing and may relate to alternative splicing sites.

## GO.ID Term Level Annotated  
## 2 GO:0006334 nucleosome assembly 9 12  
## 3 GO:0048846 axon extension involved in axon guidance 15 1  
## 4 GO:0033615 mitochondrial proton-transporting ATP sy... 9 1  
## 5 GO:0034723 DNA replication-dependent nucleosome org... 8 1  
## 6 GO:0071711 basement membrane organization 6 1  
## 7 GO:0000395 mRNA 5'-splice site recognition 15 1  
## Significant Expected Fishers Ontology  
## 2 2 0.12 0.0060 BP  
## 3 1 0.01 0.0099 BP  
## 4 1 0.01 0.0099 BP  
## 5 1 0.01 0.0099 BP  
## 6 1 0.01 0.0099 BP  
## 7 1 0.01 0.0099 BP

*bromicola* genes are also related to key cellular functions. There is one related to RNA-directed DNA Polymerase activity: EbroNfe1\_002229-T1, MF\_ARGOT, 7.7824356421, 0.7097647, <GO:0003964>, RNA-directed DNA polymerase activity

**Code to produce dataset of *amarillans* Information**

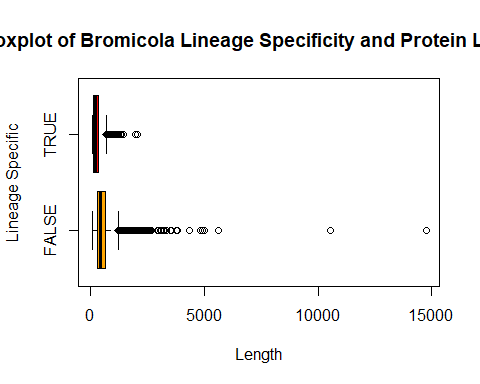
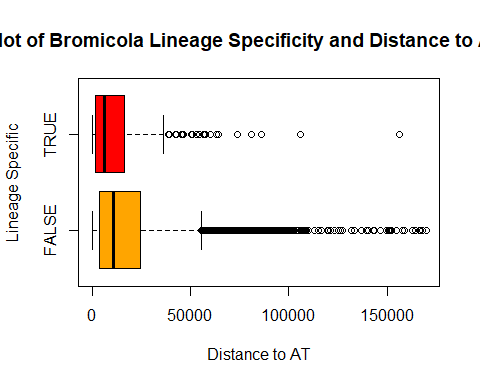
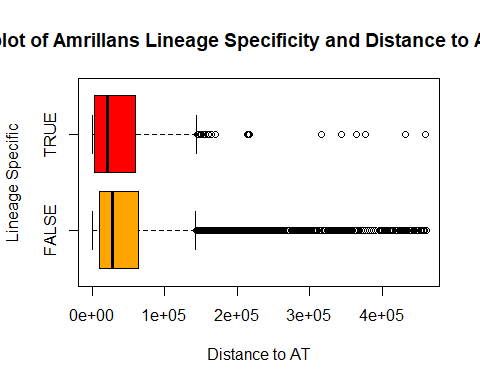
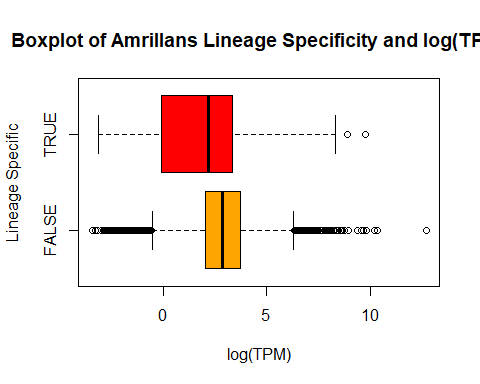
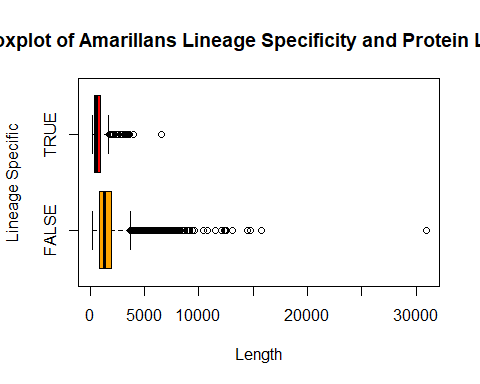
Create csv containing information on length, distance to AT, TPM and lineage specificity

ortho\_long <- read.delim("~/CoxExtension/ortho\_long.tsv", header=FALSE, stringsAsFactors=FALSE) #load ortholog file  
Eamquant <- read.delim("~/CoxExtension/RNA/quantfiles/Eamquant.sf", stringsAsFactors=FALSE) #load quant file  
amarillansuniquegeneid <- read.table("~/CoxExtension/LineageSpecificity/amarillansuniquegeneid.tsv", quote="\"", comment.char="", stringsAsFactors=FALSE) #load list of lineage specific genes  
  
Eam702 <- ortho\_long[which(ortho\_long$V3 == "Eam702"),] #subsetting  
colnames(Eam702) <- c("ortholog", "Name", "Strain") #renaming for merging  
mergedTPMandid <- merge(Eam702, Eamquant) #first merge  
  
uniquegenes <- amarillansuniquegeneid$V1 #subsetting  
geneList <- factor(as.integer(Eam702$Name %in% uniquegenes))  
names(geneList) <- Eam702$Name  
TFgenes <- data.frame(TF = geneList, Name = Eam702$Name, row.names = NULL)  
mergedTPMandTF <- merge(mergedTPMandid, TFgenes)  
  
Amaclosest\_AT <- read.delim("~/CoxExtension/Amaclosest\_AT.tsv", header=FALSE, stringsAsFactors=FALSE)  
Amaclosest\_AT$V1 <- str\_remove(Amaclosest\_AT$V1, "ID=")  
Amaclosest\_AT$V1 <- str\_replace(Amaclosest\_AT$V1, ";", "-T1")  
colnames(Amaclosest\_AT) <- c("Name", "DistanceAT")  
  
Amarillansfinalframe <- merge(Amaclosest\_AT, mergedTPMandTF)  
  
Amarillansfinalframe$TF <- str\_replace(Amarillansfinalframe$TF, "0", "FALSE")  
Amarillansfinalframe$TF <- str\_replace(Amarillansfinalframe$TF, "1", "TRUE")  
  
write.csv(Amarillansfinalframe, file = "AmarillansData.csv")  
  
head(Amarillansfinalframe)

## Name DistanceAT ortholog Strain Length EffectiveLength TPM  
## 1 Eam702\_000001-T1 63298 og\_6853 Eam702 1971 1722 0.519836  
## 2 Eam702\_000002-T1 0 og\_0644 Eam702 960 711 62.836155  
## 3 Eam702\_000003-T1 308 og\_0157 Eam702 894 645 468.333379  
## 4 Eam702\_000004-T1 61144 og\_7959 Eam702 1566 1317 34.479049  
## 5 Eam702\_000005-T1 1732 og\_7542 Eam702 543 294 88.851458  
## 6 Eam702\_000006-T1 595 og\_3582 Eam702 1257 1008 75.403795  
## NumReads TF  
## 1 11 FALSE  
## 2 549 FALSE  
## 3 3712 FALSE  
## 4 558 FALSE  
## 5 321 FALSE  
## 6 934 FALSE

*bromicola* final frame code not shown, although some edits made from *amarillans* code. Due to a lack of RNAseq (and therefore quantification) data, no TPM is available. Length was caulcated manually utilising protein sequences and Bioawk, before being read into R and appended to dataset. Head of *bromicola* frame is shown.

## Name DistanceAT ortholog Strain TF Length  
## 1 EbroNfe1\_000001-T1 1403 og\_4422 EbroNfe1 FALSE 259  
## 2 EbroNfe1\_000002-T1 798 og\_5274 EbroNfe1 FALSE 330  
## 3 EbroNfe1\_000003-T1 3767 og\_1293 EbroNfe1 FALSE 163  
## 4 EbroNfe1\_000004-T1 5046 og\_0409 EbroNfe1 FALSE 341  
## 5 EbroNfe1\_000005-T1 1690 og\_29412 EbroNfe1 TRUE 213  
## 6 EbroNfe1\_000006-T1 113 og\_7009 EbroNfe1 FALSE 148

**Graphical representation** Boxplots were produced for variables in comparison to lineage specificity 

**Statistical Analysis** It can be observed that lineage specific genes appear to be shorter, closer to ATrich repeat regions and less expressed (the latter only shown in *amarillans*). Significance of these results was tested and all significant (below).

##   
## Call:  
## glm(formula = TPM < 2 ~ TF, family = "binomial", data = Amarillansfinalframe)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8468 -0.4399 -0.4399 -0.4399 2.1835   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.28699 0.04088 -55.94 <2e-16 \*\*\*  
## TFTRUE 1.44599 0.13191 10.96 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4865.5 on 7448 degrees of freedom  
## Residual deviance: 4766.7 on 7447 degrees of freedom  
## AIC: 4770.7  
##   
## Number of Fisher Scoring iterations: 5

##   
## Call:  
## glm(formula = DistanceAT < 2000 ~ TF, family = "binomial", data = Amarillansfinalframe)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7023 -0.4963 -0.4963 -0.4963 2.0763   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.03230 0.03696 -54.988 < 2e-16 \*\*\*  
## TFTRUE 0.75813 0.14407 5.262 1.42e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5466.9 on 7448 degrees of freedom  
## Residual deviance: 5442.7 on 7447 degrees of freedom  
## AIC: 5446.7  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = Length < 500 ~ TF, family = "binomial", data = Amarillansfinalframe)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0556 -0.4578 -0.4578 -0.4578 2.1484   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.20299 0.03952 -55.75 <2e-16 \*\*\*  
## TFTRUE 1.90951 0.12286 15.54 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5248.0 on 7448 degrees of freedom  
## Residual deviance: 5042.7 on 7447 degrees of freedom  
## AIC: 5046.7  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = Length < 500 ~ TF, family = "binomial", data = BromicolaFinalFrame)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.984 -1.316 1.044 1.044 1.044   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.32097 0.02384 13.46 <2e-16 \*\*\*  
## TFTRUE 1.49770 0.14731 10.17 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10285 on 7614 degrees of freedom  
## Residual deviance: 10145 on 7613 degrees of freedom  
## AIC: 10149  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = DistanceAT < 2000 ~ TF, family = "binomial", data = BromicolaFinalFrame)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8265 -0.6106 -0.6106 -0.6106 1.8823   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.58521 0.03132 -50.607 < 2e-16 \*\*\*  
## TFTRUE 0.68661 0.11543 5.948 2.71e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 7091.5 on 7614 degrees of freedom  
## Residual deviance: 7059.3 on 7613 degrees of freedom  
## AIC: 7063.3  
##   
## Number of Fisher Scoring iterations: 4