Investigating the presence of genes in the Extrachromosomal Circular DNA of plants

Aman Sidhant

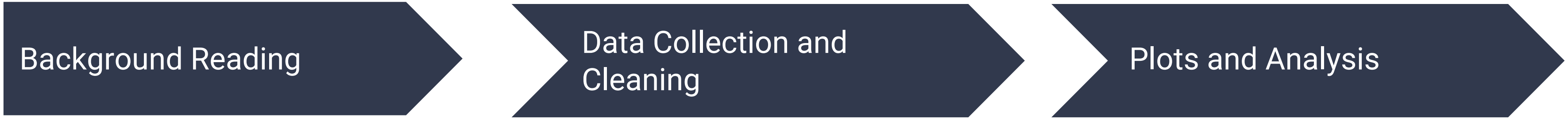
Hypothesis

Many genes on eccDNA might be responsible for plant adaptation to the environment.

Questions

1. Are genes present on Extrachromosomal Circular DNA (eccDNA)?
2. If genes are present, are they induced in mutated genes and/or across different samples as well?
    a. If the genes are induced, where are they located, and how dense are they?

Process



Background Reading

Primarily covered the paper “Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants” by Lanciano et al.

Data Collection and Cleaning

Used plants.ensembl and phytozome.jgi.doe.gov to download sequence plant genome data

Cleaned and prepared the data according to specifications for creating suitable graphs

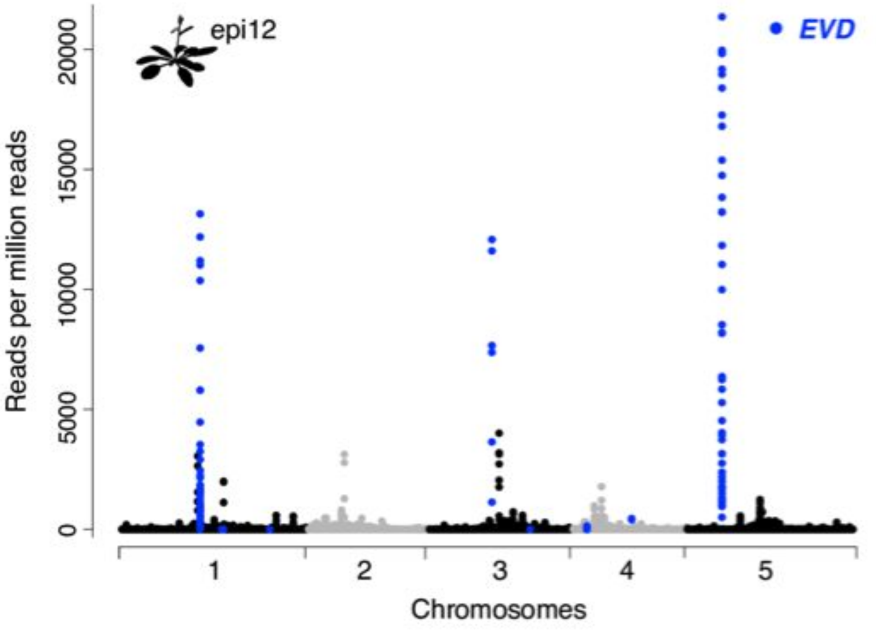
Plots and Analysis

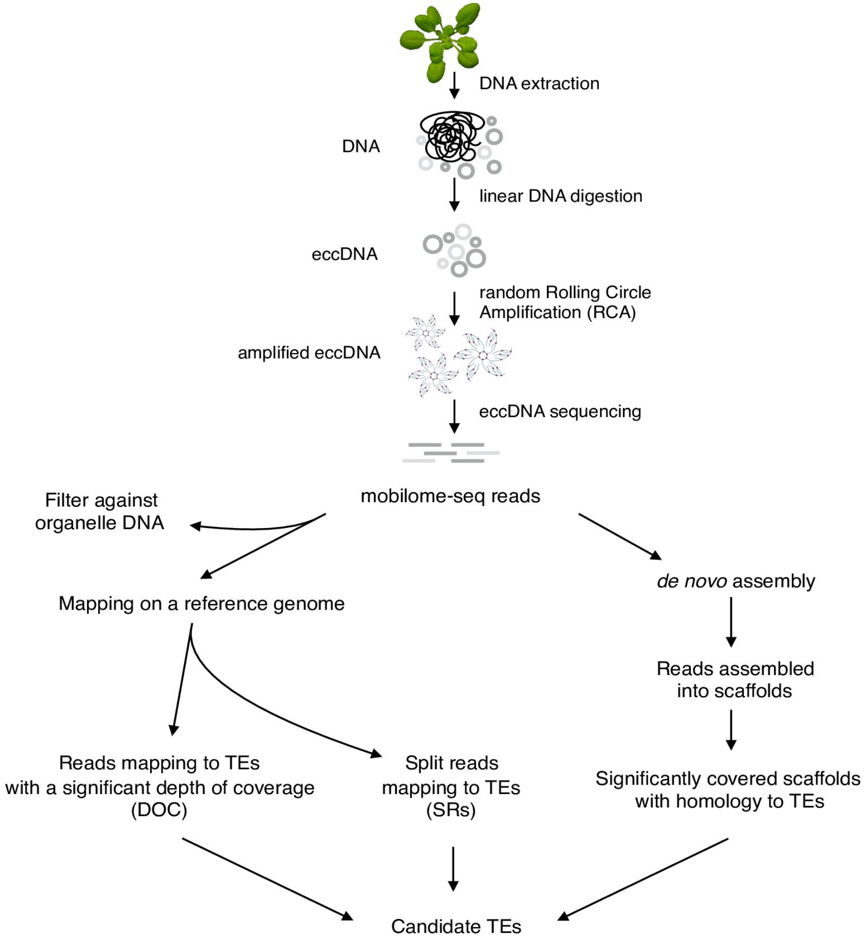
Used statistical techniques and created density graphs and Manhattan plots using R to verify hypothesis

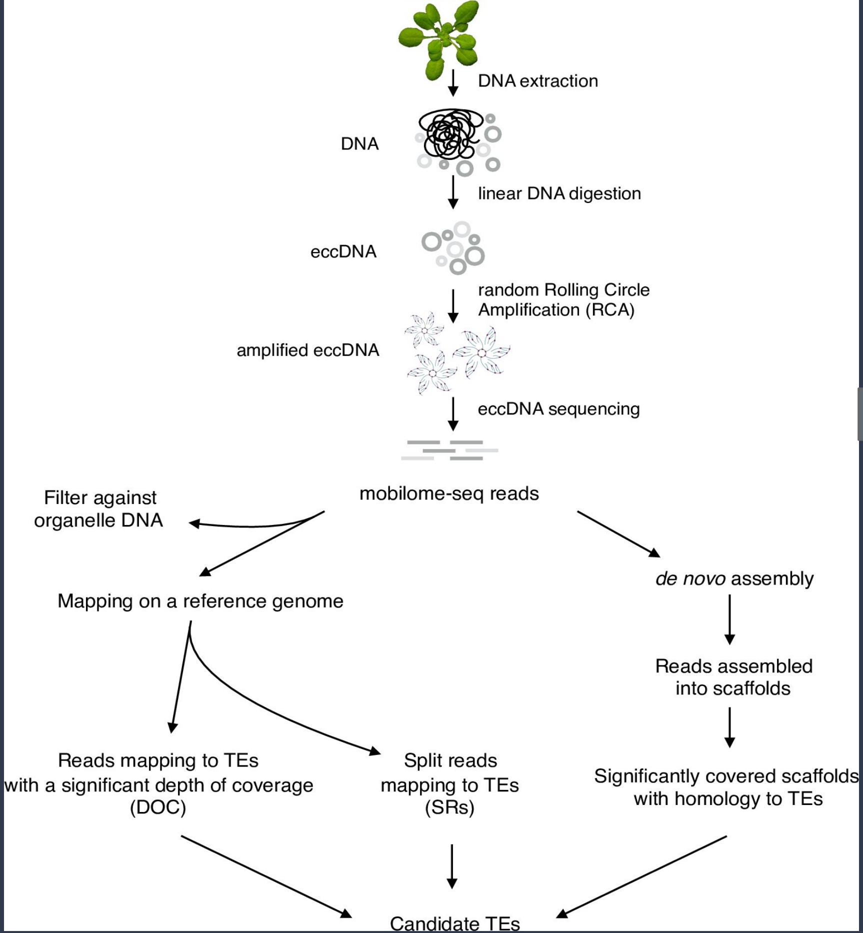
Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants, Lanciano et al. - https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006630#abstract0

Background Reading

Prior information from Lanciano et al.




DNA extraction
DNA
linear DNA digestion
eccDNA
random Rolling Circle Amplification (RCA)
amplified eccDNA
eccDNA sequencing
mobilome-seq reads
Filter against organelle DNA
Mapping on a reference genome
de novo assembly
Reads assembled into scaffolds
Reads mapping to TEs with a significant depth of coverage (DOC)
Split reads mapping to TEs (SRs)
Significantly covered scaffolds with homology to TEs
Candidate TEs



DNA extraction

DNA

linear DNA digestion

eccDNA

random Rolling Circle Amplification (RCA)

amplified eccDNA

eccDNA sequencing

mobilome-seq reads

Filter against organelle DNA

Mapping on a reference genome

de novo assembly

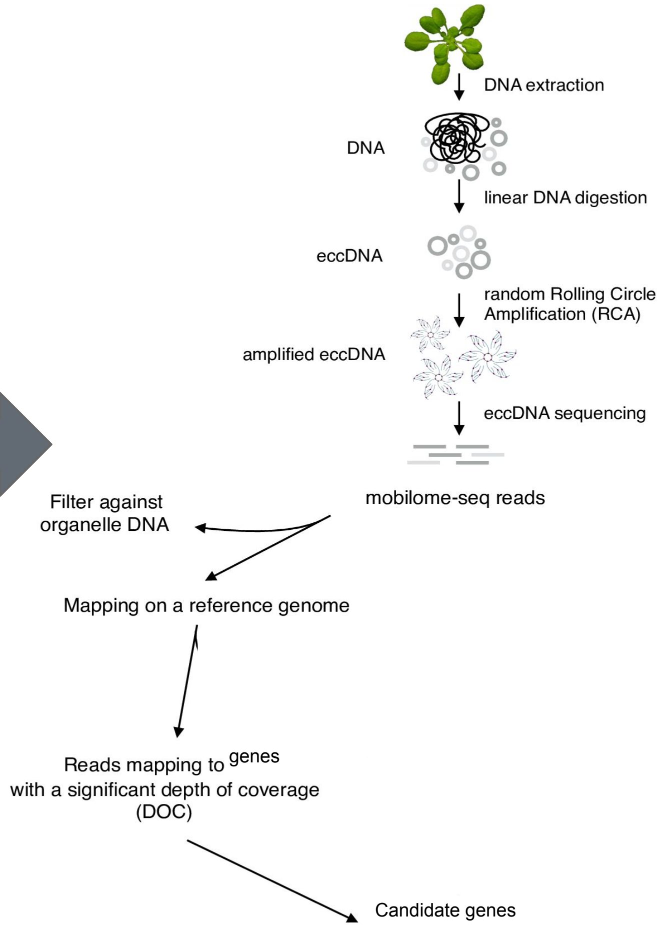
Reads assembled into scaffolds

Reads mapping to TEs with a significant depth of coverage (DOC)

Split reads mapping to TEs (SRs)

Significantly covered scaffolds with homology to TEs

Candidate TEs



DNA extraction

DNA

linear DNA digestion

eccDNA

random Rolling Circle Amplification (RCA)

amplified eccDNA

eccDNA sequencing

mobilome-seq reads

Filter against organelle DNA

Mapping on a reference genome

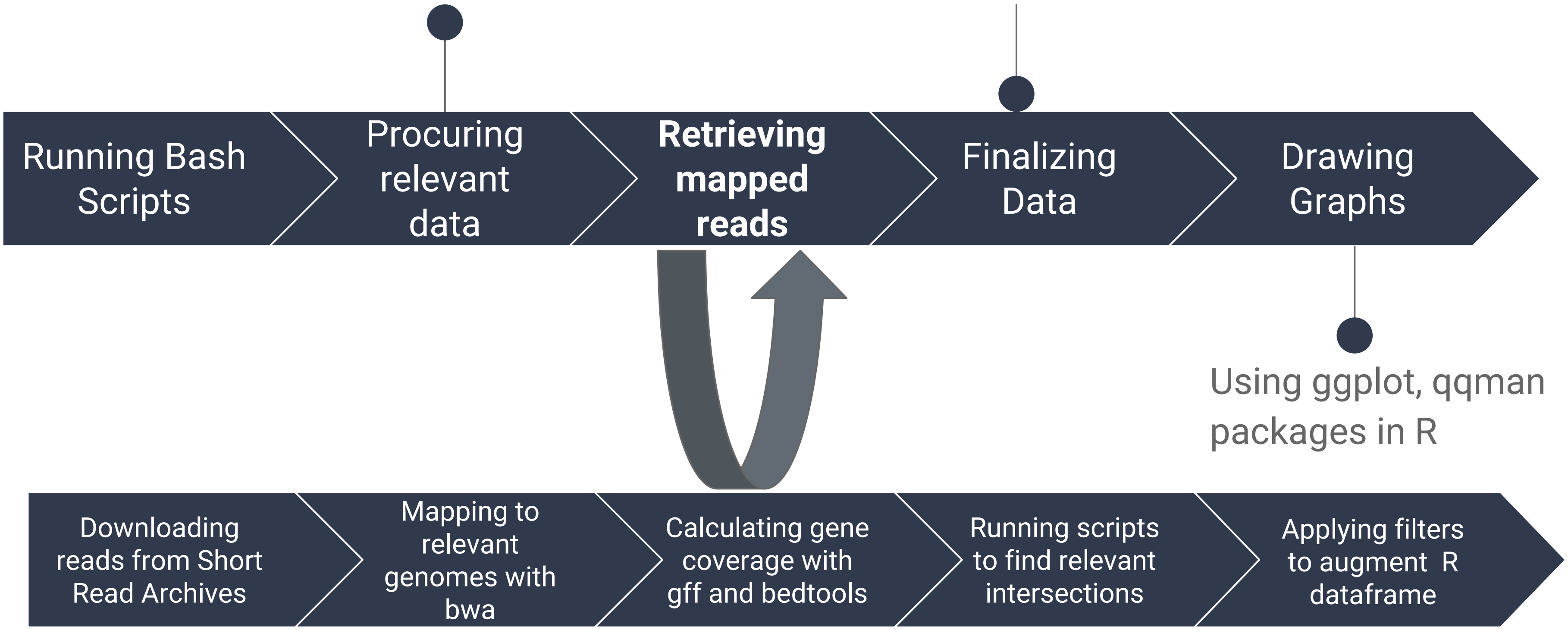
Reads mapping to genes with a significant depth of coverage (DOC)

Candidate genes

Data Collection

Gff, fastq and bam files from Phytozome, plantsensembl

Data Cleaning and Analysis in R



Running Bash Scripts

Procuring relevant data

Retrieving mapped reads

Finalizing Data

Drawing Graphs

Using ggplot, qqman packages in R

Downloading reads from Short Read Archives

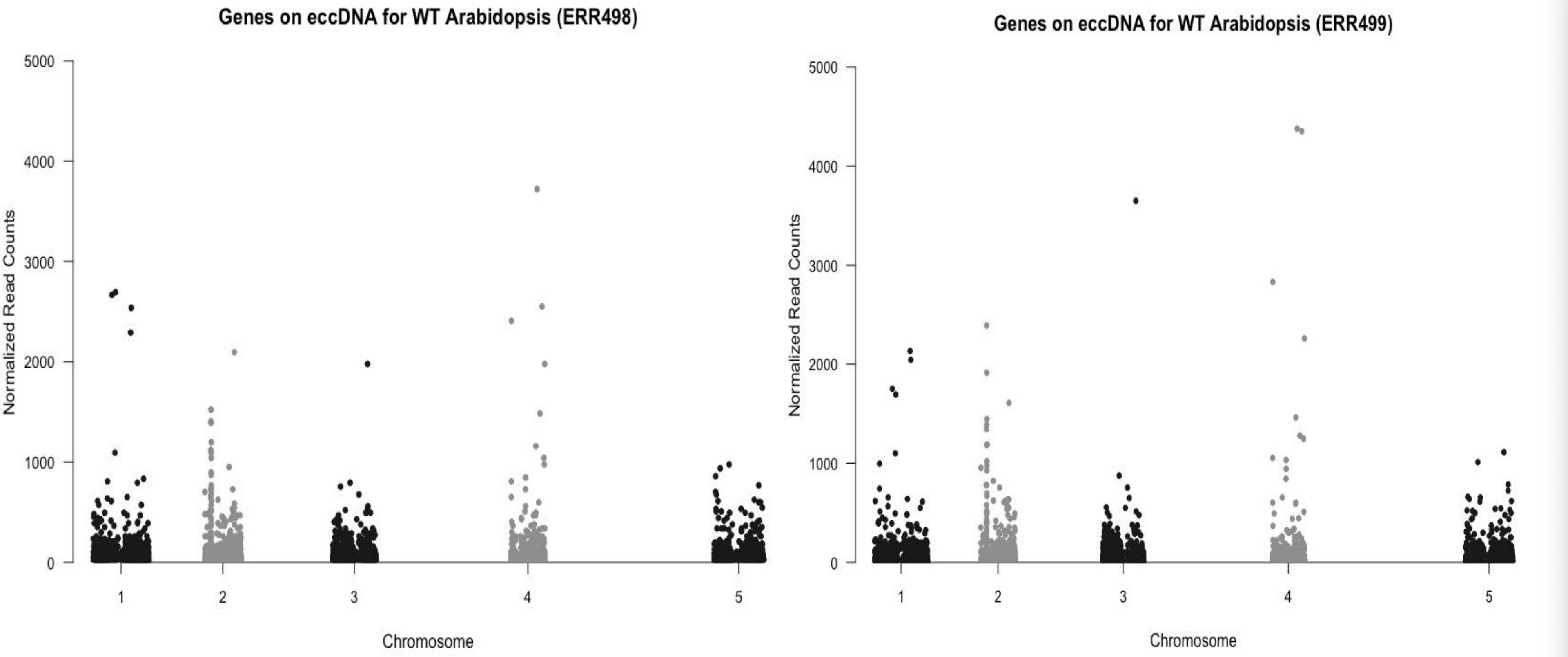
Mapping to relevant genomes with bwa

Calculating gene coverage with gff and bedtools

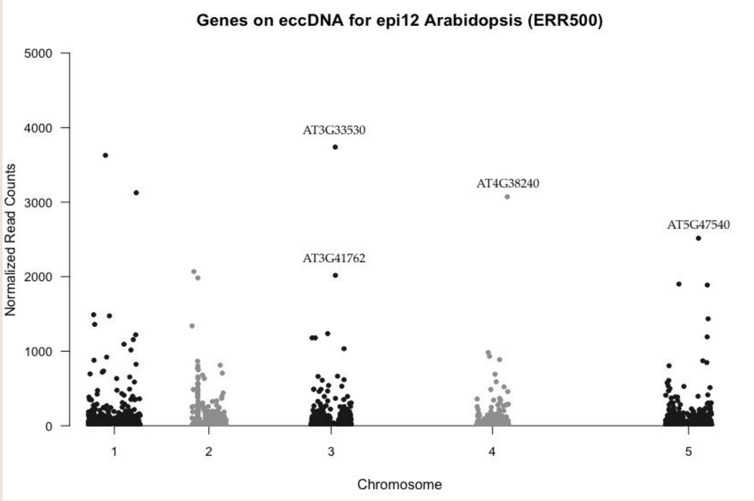
Running scripts to find relevant intersections

Applying filters to augment R dataframe

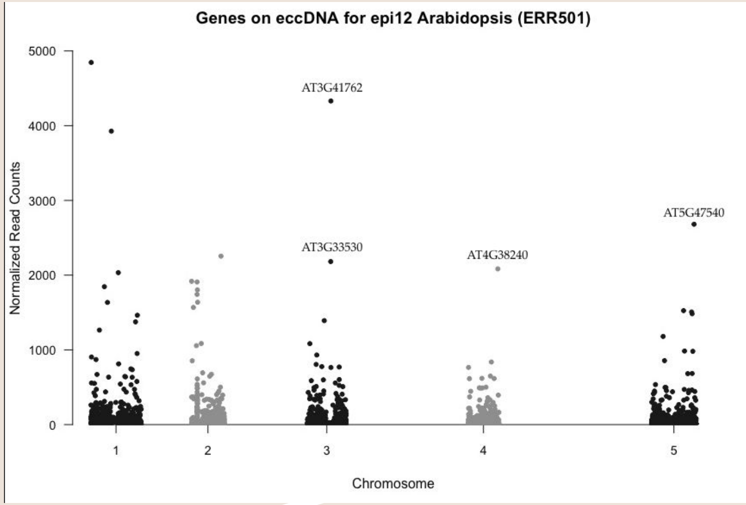
Plots


Genes on eccDNA for WT Arabidopsis (ERR498)
Normalized Read Counts
5000
4000
3000
2000
1000
0
1
2
3
4
5
Chromosome
Genes on eccDNA for WT Arabidopsis (ERR499)
Normalized Read Counts
5000
4000
3000
2000
1000
0
1
2
3
4
5
Chromosome

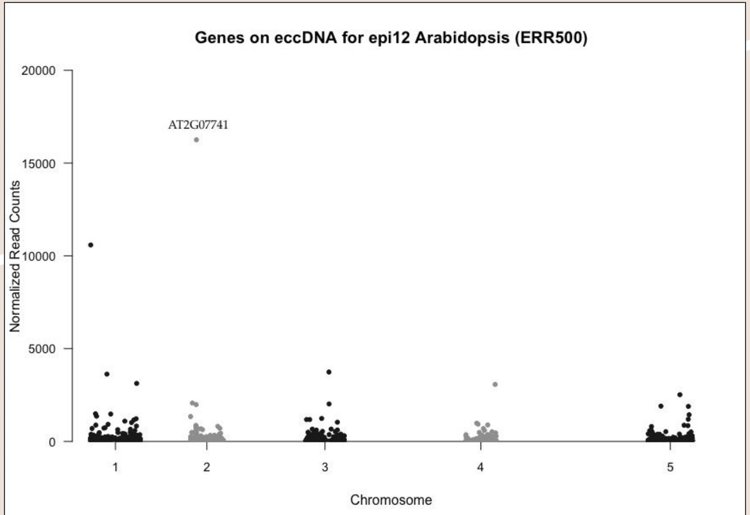
Plots



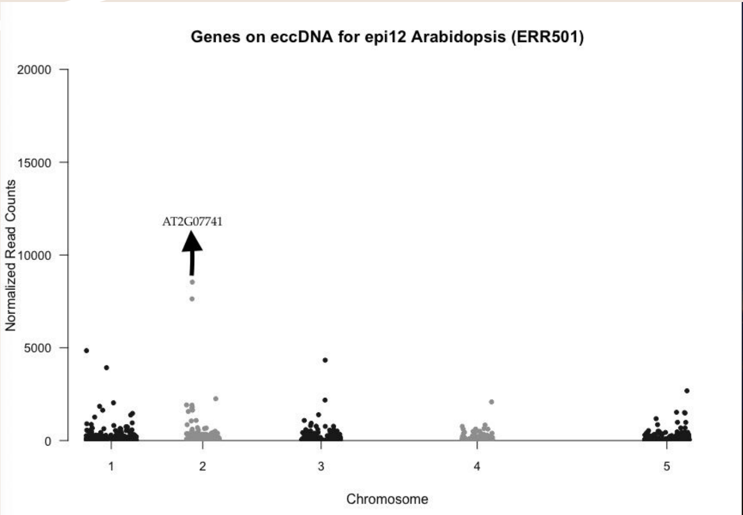
Genes on eccDNA for epi12 Arabidopsis (ERR500)



Genes on eccDNA for epi12 Arabidopsis (ERR501)

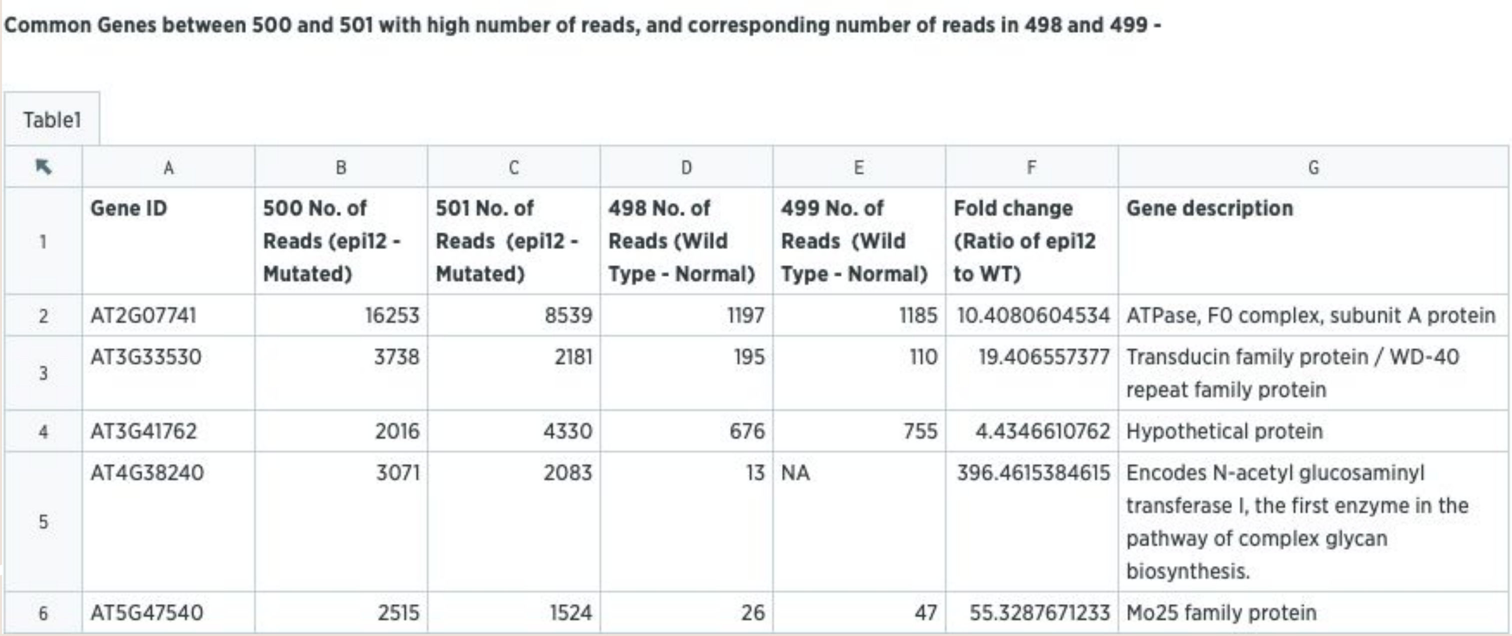


Genes on eccDNA for epi12 Arabidopsis (ERR500)



Genes on eccDNA for epi12 Arabidopsis (ERR501)

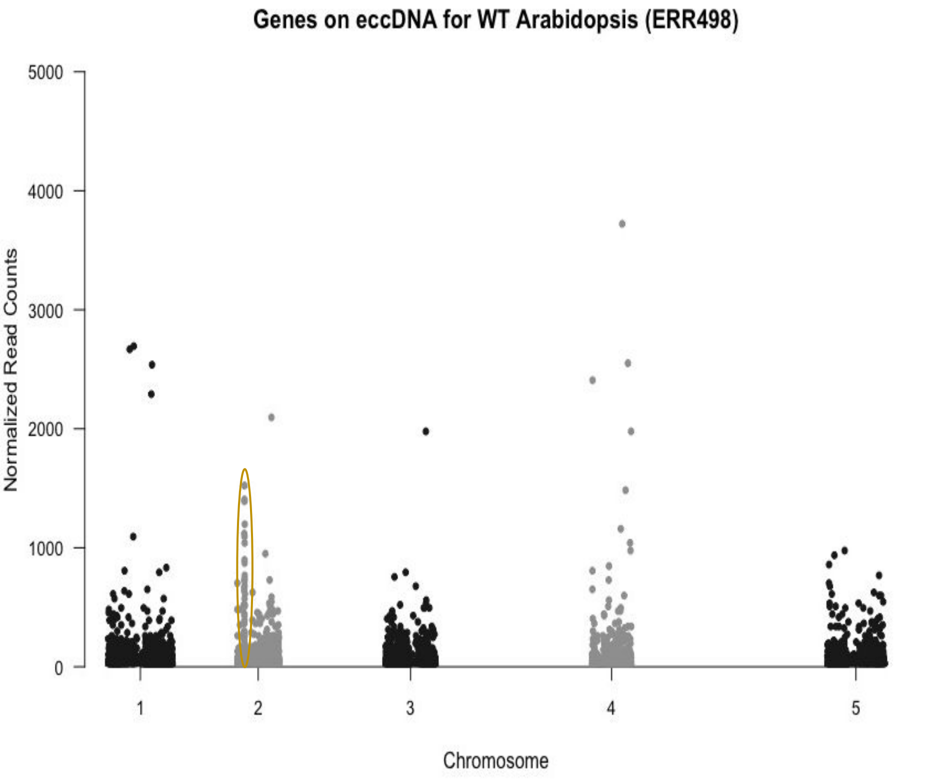
Results

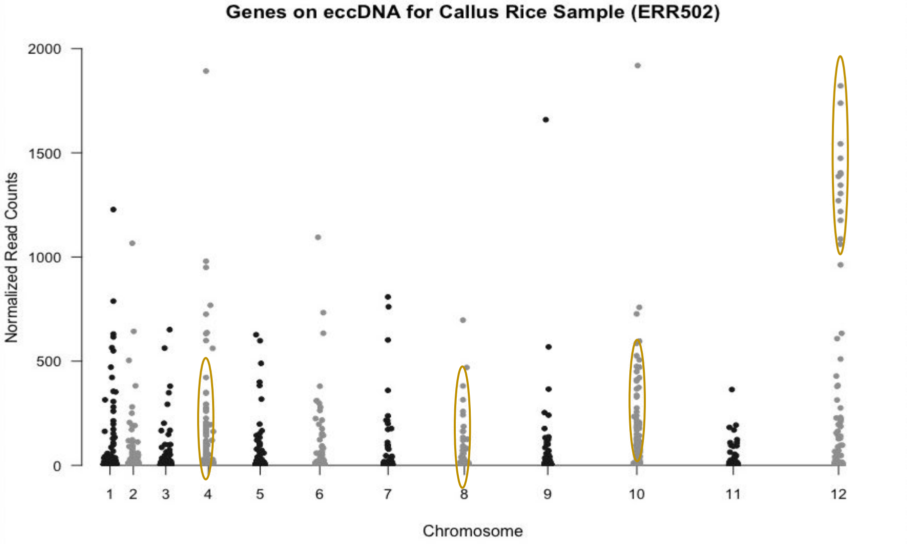


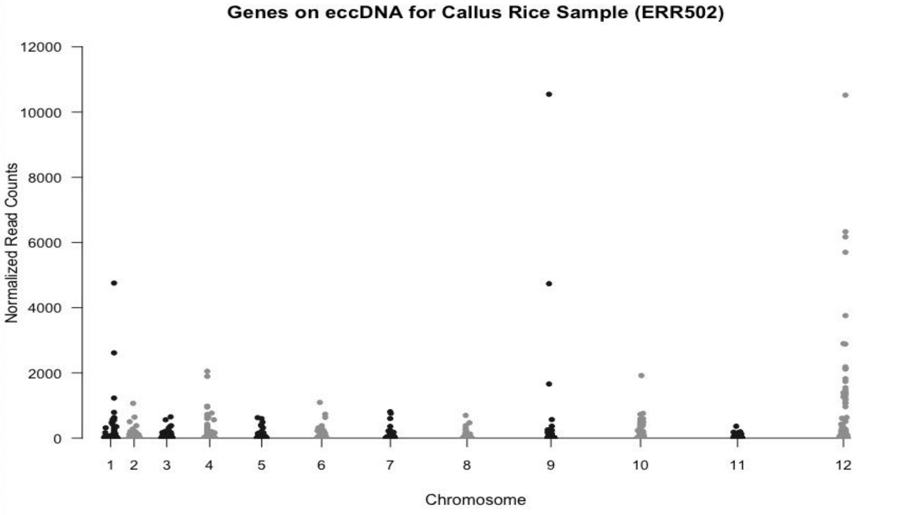
Common Genes between 500 and 501 with high number of reads, and corresponding number of reads in 498 and 499 -

	A	B	C	D	E	F	G
1	**Gene ID**	**500 No. of Reads (epi12 - Mutated)**	**501 No. of Reads (epi12 - Mutated)**	**498 No. of Reads (Wild Type - Normal)**	**499 No. of Reads (Wild Type - Normal)**	**Fold change (Ratio of epi12 to WT)**	**Gene description**
2	AT2G07741	16253	8539	1197	1185	10.4080604534	ATPase, F0 complex, subunit A protein
3	AT3G33530	3738	2181	195	110	19.406557377	Transducin family protein / WD-40 repeat family protein
4	AT3G41762	2016	4330	676	755	4.4346610762	Hypothetical protein
5	AT4G38240	3071	2083	13	NA	396.4615384615	Encodes N-acetyl glucosaminyl transferase I, the first enzyme in the pathway of complex glycan biosynthesis.
6	AT5G47540	2515	1524	26	47	55.3287671233	Mo25 family protein

Next Steps


Genes on eccDNA for WT Arabidopsis (ERR498)
Normalized Read Counts
0
1000
2000
3000
4000
5000
1
2
3
4
5
Chromosome


Genes on eccDNA for Callus Rice Sample (ERR502)
Normalized Read Counts
0
500
1000
1500
2000
1
2
3
4
5
6
7
8
9
10
11
12
Chromosome


Genes on eccDNA for Callus Rice Sample (ERR502)
Normalized Read Counts
0
2000
4000
6000
8000
10000
12000
1
2
3
4
5
6
7
8
9
10
11
12
Chromosome

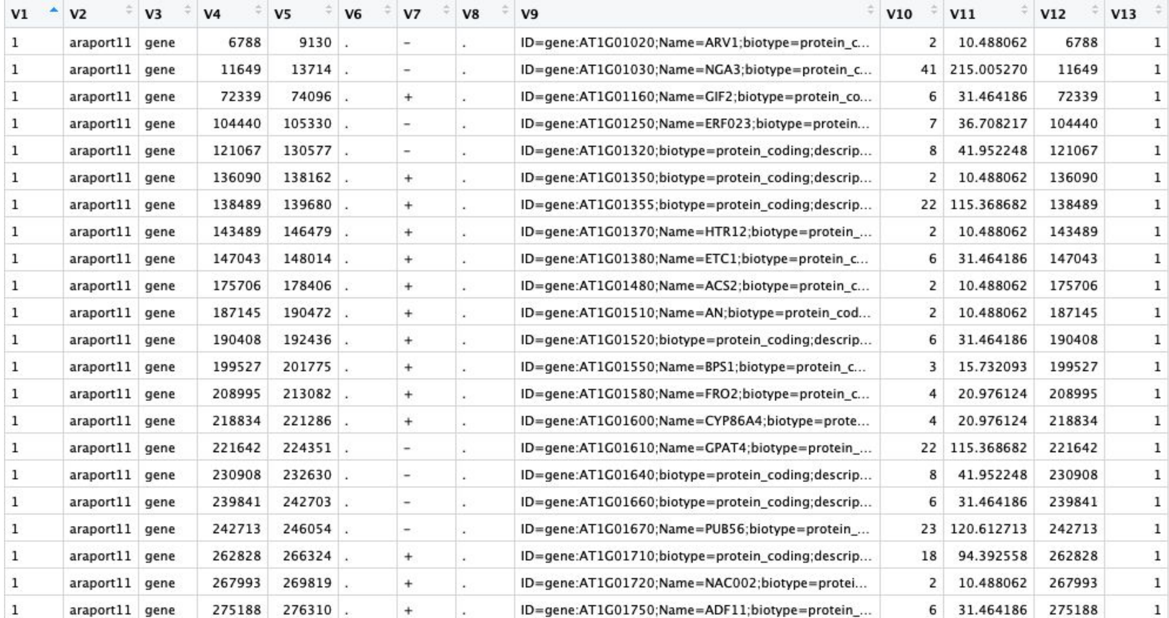
Conclusions:

- Genes are indeed present on eccDNA. These gene families have been shown to increase disease resistance and are highly polymorphic.★
- Genes present on eccDNA are induced across both wild type and mutated samples. Density is much higher in epigenetically deficient species.
- A small portion of such genes is induced by stress or methylation changes to the DNA (difference between epi12 and Wild Type Arabidopsis)

★ “Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri” http://www.pnas.org/content/115/13/3332

Lessons

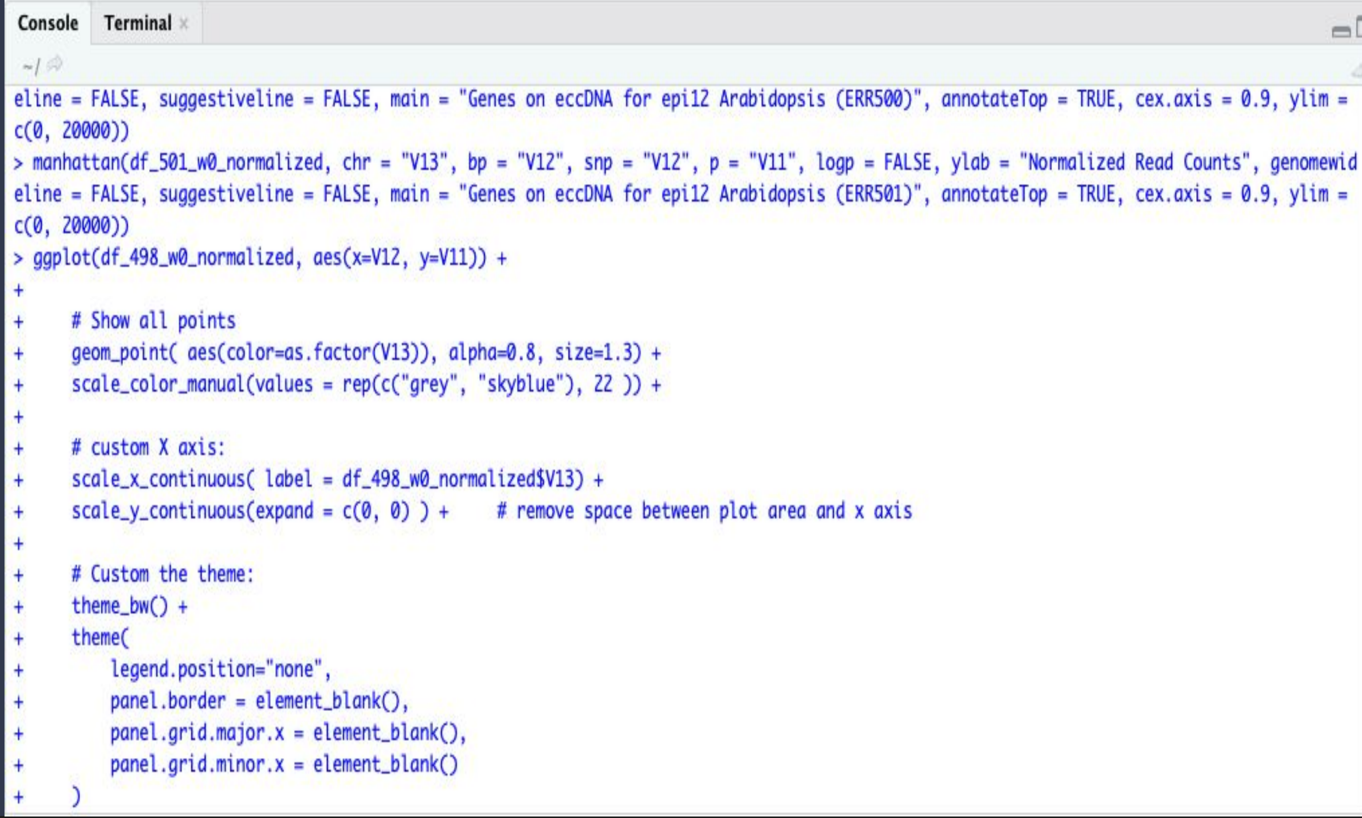
- Learning biology again
- Operating in a research environment



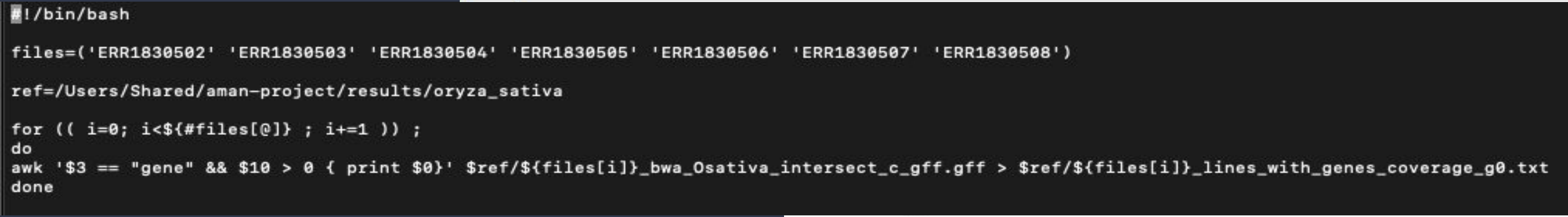
V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	araport11	gene	6788	9130	.	-	.	ID=gene:AT1G01020;Name=ARV1;biotype=protein_c...	2	10.488062	6788	1
1	araport11	gene	11649	13714	.	-	.	ID=gene:AT1G01030;Name=NGA3;biotype=protein_c...	41	215.005270	11649	1
1	araport11	gene	72339	74096	.	+	.	ID=gene:AT1G01160;Name=GIF2;biotype=protein_co...	6	31.464186	72339	1
1	araport11	gene	104440	105330	.	-	.	ID=gene:AT1G01250;Name=ERF023;biotype=protein...	7	36.708217	104440	1
1	araport11	gene	121067	130577	.	-	.	ID=gene:AT1G01320;biotype=protein_coding;descrip...	8	41.952248	121067	1
1	araport11	gene	136090	138162	.	+	.	ID=gene:AT1G01350;biotype=protein_coding;descrip...	2	10.488062	136090	1
1	araport11	gene	138489	139680	.	+	.	ID=gene:AT1G01355;biotype=protein_coding;descrip...	22	115.368682	138489	1
1	araport11	gene	143489	146479	.	+	.	ID=gene:AT1G01370;Name=HTR12;biotype=protein_...	2	10.488062	143489	1
1	araport11	gene	147043	148014	.	+	.	ID=gene:AT1G01380;Name=ETC1;biotype=protein_c...	6	31.464186	147043	1
1	araport11	gene	175706	178406	.	+	.	ID=gene:AT1G01480;Name=ACS2;biotype=protein_c...	2	10.488062	175706	1
1	araport11	gene	187145	190472	.	+	.	ID=gene:AT1G01510;Name=AN;biotype=protein_cod...	2	10.488062	187145	1
1	araport11	gene	190408	192436	.	+	.	ID=gene:AT1G01520;biotype=protein_coding;descrip...	6	31.464186	190408	1
1	araport11	gene	199527	201775	.	+	.	ID=gene:AT1G01550;Name=BPS1;biotype=protein_c...	3	15.732093	199527	1
1	araport11	gene	208995	213082	.	+	.	ID=gene:AT1G01580;Name=FRO2;biotype=protein_c...	4	20.976124	208995	1
1	araport11	gene	218834	221286	.	+	.	ID=gene:AT1G01600;Name=CYP86A4;biotype=prote...	4	20.976124	218834	1
1	araport11	gene	221642	224351	.	-	.	ID=gene:AT1G01610;Name=GPAT4;biotype=protein_...	22	115.368682	221642	1
1	araport11	gene	230908	232630	.	-	.	ID=gene:AT1G01640;biotype=protein_coding;descrip...	8	41.952248	230908	1
1	araport11	gene	239841	242703	.	-	.	ID=gene:AT1G01660;biotype=protein_coding;descrip...	6	31.464186	239841	1
1	araport11	gene	242713	246054	.	-	.	ID=gene:AT1G01670;Name=PUB56;biotype=protein_...	23	120.612713	242713	1
1	araport11	gene	262828	266324	.	+	.	ID=gene:AT1G01710;biotype=protein_coding;descrip...	18	94.392558	262828	1
1	araport11	gene	267993	269819	.	+	.	ID=gene:AT1G01720;Name=NAC002;biotype=protei...	2	10.488062	267993	1
1	araport11	gene	275188	276310	.	+	.	ID=gene:AT1G01750;Name=ADF11;biotype=protein_...	6	31.464186	275188	1

Skills Learnt

Statistical Modeling in R
Data Processing
Data Cleaning
Plotting Graphs using dplyr, ggplot2
Bash Scripting



```
eline = FALSE, suggestiveline = FALSE, main = "Genes on eccDNA for epi12 Arabidopsis (ERR500)", annotateTop = TRUE, cex.axis = 0.9, ylim =
c(0, 20000))
> manhattan(df_501_w0_normalized, chr = "V13", bp = "V12", snp = "V12", p = "V11", logp = FALSE, ylab = "Normalized Read Counts", genomewid
eline = FALSE, suggestiveline = FALSE, main = "Genes on eccDNA for epi12 Arabidopsis (ERR501)", annotateTop = TRUE, cex.axis = 0.9, ylim =
c(0, 20000))
> ggplot(df_498_w0_normalized, aes(x=V12, y=V11)) +
+
+     # Show all points
+     geom_point( aes(color=as.factor(V13)), alpha=0.8, size=1.3) +
+     scale_color_manual(values = rep(c("grey", "skyblue"), 22 )) +
+
+     # custom X axis:
+     scale_x_continuous( label = df_498_w0_normalized$V13) +
+     scale_y_continuous(expand = c(0, 0) ) +     # remove space between plot area and x axis
+
+     # Custom the theme:
+     theme_bw() +
+     theme(
+         legend.position="none",
+         panel.border = element_blank(),
+         panel.grid.major.x = element_blank(),
+         panel.grid.minor.x = element_blank()
+     )
```



```
#!/bin/bash

files=('ERR1830502' 'ERR1830503' 'ERR1830504' 'ERR1830505' 'ERR1830506' 'ERR1830507' 'ERR1830508')

ref=/Users/Shared/aman-project/results/oryza_sativa

for (( i=0; i<${#files[@]} ; i+=1 )) ;
do
awk '$3 == "gene" && $10 > 0 { print $0}' $ref/${files[i]}_bwa_Osativa_intersect_c_gff.gff > $ref/${files[i]}_lines_with_genes_coverage_g0.txt
done
```