RNA-seq Report

Nathan Hughes (JIC)

May 30, 2019

Contents

1 General helper functions **3**
1.1 Making nice tables 3
1.2 Excel reader and loading count data 3
1.3 Gprofiler function 4

2 Recursive Feature Elimination **4**
2.1 25 vars 4
2.1.1 Forest on this RFE set 5
2.2 25 vars 7
2.2.1 Forest on this RFE set 9

1 General helper functions

1.1 Making nice tables

```python
import tabulate
import IPython

class OrgFormatter(IPython.core.formatters.BaseFormatter):
    format_type = IPython.core.formatters.Unicode('text/org')
    print_method = IPython.core.formatters.ObjectName('_repr_org_')

def pd_dataframe_to_org(df):
    return tabulate.tabulate(df, headers='keys', tablefmt='orgtbl', showindex='always')

ip = get_ipython()
ip.display_formatter.formatters['text/org'] = OrgFormatter()

f = ip.display_formatter.formatters['text/org']
f.for_type_by_name('pandas.core.frame', 'DataFrame', pd_dataframe_to_org)
```

1.2 Excel reader and loading count data

```python
import pandas as pd
import warnings
warnings.filterwarnings('ignore')


def read_xl(fn="/Users/nathan/PHD/Transcripts/Data/diff_from_col0:False_onlyDiff:False.xlsx"):
    xl = pd.ExcelFile(fn)
    sheet_names = xl.sheet_names
    dfs = []
    for s in sheet_names:
        d = xl.parse(s)
        d['sample'] = s.split("|")[0].replace(" ", "")
        dfs.append(d)

    DE = pd.concat(dfs)
    DE = DE.rename_axis('gene').sort_values(by=['gene', 'log2FoldChange'],
                                            ascending=[False, False])
    return DE

counts = pd.read_csv(
    "/Users/hughesn/PHD/Transcripts/Data/norml_count_data.csv",index_col=0)
counts[[c for c in counts.columns if 'cer_c' in c]].head(5)
```

	cer_c_05h_a37	cer_c_05h_b38	cer_c_05h_c39	cer_c_6h_a85	cer_c_6h_b86	cer_c_6h_c87
AT1G01010	7.65333	7.73449	7.5679	7.63575	7.62055	7.81064
AT1G01020	7.93999	7.79909	7.79347	7.95616	7.924	7.88399
AT1G01030	7.27285	7.09544	7.00389	6.88372	6.72014	6.58998
AT1G01040	9.16837	9.09566	9.13567	9.05724	9.0856	9.21304
AT1G01050	9.825	9.80514	9.76124	9.82781	9.91565	9.77211

1.3 Gprofiler function

```
from gprofiler import GProfiler

def get_gene_names(geneList):

    gp = GProfiler(return_dataframe=True)
    df = gp.convert(organism='athaliana',
             query=geneList)[['incoming', 'name', 'description']]
    df['description'] = df.apply(lambda x: x['description'].split('[')[0].split(';')[0], axis=1)
    return df
get_gene_names(list(counts.head(5).index))
```

	incoming	name	description
0	AT1G01010	NAC001	NAC domain-containing protein 1
1	AT1G01020	ARV1	ARV1 family protein
2	AT1G01030	NGA3	B3 domain-containing transcription factor NGA3
3	AT1G01040	DCL1	Dicer-like 1
4	AT1G01050	PPA1	Soluble inorganic pyrophosphatase 1

2 Recursive Feature Elimination

2.1 25 vars

```
from sklearn.feature_selection import RFE, RFECV
from sklearn.linear_model import LogisticRegression

# load data
DE_pairings_05hr = read_xl('./Data/pairings_05hr.xlsx')
sig = DE_pairings_05hr[DE_pairings_05hr['padj'] < 0.05]
sig = sig['log2FoldChange'].sort_values()
locs = sig.index
df = counts.loc[locs][[c for c in counts.columns if ('_05h' in c and 'col' in c)]].T
df = df.loc[:,~df.columns.duplicated()]
df = df[[c for c in set(df.columns.values)]]
```

```
# Feature Extraction with RFE
X = df.values
y = [y.rsplit('_',1)[0] for y in df.reset_index()['index']]
# feature extraction
model = LogisticRegression()
rfe = RFE(model, n_features_to_select=25)
fit = rfe.fit(X, y)
print("Num Features: {0}".format(fit.n_features_))
print("Selected Features: {0}".format(fit.support_))
print("Feature Ranking: {0}".format(fit.ranking_))
```

```
Num Features: 25 Selected Features: [False False False ... False False False] Feature Ranking: [ 394 3832 2928
... 3537 4774 1915]
```

	incoming	name	description
0	AT5G61590	ERF107	Ethylene-responsive transcription factor ERF107
1	AT1G25440	COL16	Zinc finger protein CONSTANS-LIKE 16
2	AT1G56242	AT1G56242	other RNA
3	AT5G24110	WRKY30	Probable WRKY transcription factor 30
4	AT2G21210	AT2G21210	SAUR-like auxin-responsive protein family
5	AT3G09275	AT3G09275	None
6	AT1G07160	AT1G07160	PP2C-type phosphatase AP2C2
7	AT2G18440	GUT15	GUT15 (GENE WITH UNSTABLE TRANSCRIPT 15)
8	AT1G14540	PER4	Peroxidase
9	AT1G56240	PP2B13	F-box protein PP2-B13
10	AT2G37430	ZAT11	ZAT11
11	AT5G60390	A1	Elongation factor 1-alpha 4
12	AT5G11140	AT5G11140	Phospholipase-like protein (PEARLI 4) family protein
13	AT3G02840	AT3G02840	ARM repeat superfamily protein
14	AT5G59780	MYB59	Transcription factor MYB59
15	AT5G47230	ERF5	ERF5
16	AT4G38840	AT4G38840	At4g38840
17	AT5G37260	RVE2	Protein REVEILLE 2
18	AT5G27420	ATL31	E3 ubiquitin-protein ligase ATL31
19	AT4G19700	BOI	E3 ubiquitin-protein ligase BOI
20	AT1G68520	COL6	Zinc finger protein CONSTANS-LIKE 6
21	AT5G25350	EBF2	EIN3-binding F-box protein 2
22	AT1G66090	AT1G66090	Disease resistance protein (TIR-NBS class)
23	AT1G72520	LOX4	Lipoxygenase 4, chloroplastic
24	AT3G59940	SKIP20	F-box/kelch-repeat protein SKIP20

2.1.1 Forest on this RFE set

gene	importance	name	description
AT1G68520	0.03995	COL6	Zinc finger protein CONSTANS-LIKE 6
AT1G72520	0.03995	LOX4	Lipoxygenase 4, chloroplastic
AT1G56242	0.03875	AT1G56242	other RNA
AT1G25440	0.03825	COL16	Zinc finger protein CONSTANS-LIKE 16
AT3G59940	0.03805	SKIP20	F-box/kelch-repeat protein SKIP20
AT5G59780	0.0379	MYB59	Transcription factor MYB59
AT1G66090	0.0379	AT1G66090	Disease resistance protein (TIR-NBS class)
AT4G38840	0.03775	AT4G38840	At4g38840
AT5G37260	0.0377	RVE2	Protein REVEILLE 2
AT1G07160	0.0377	AT1G07160	PP2C-type phosphatase AP2C2
AT5G27420	0.03745	ATL31	E3 ubiquitin-protein ligase ATL31
AT5G11140	0.03705	AT5G11140	Phospholipase-like protein (PEARLI 4) family protein
AT5G24110	0.0369	WRKY30	Probable WRKY transcription factor 30
AT5G47230	0.0369	ERF5	ERF5
AT3G09275	0.03685	AT3G09275	None
AT2G21210	0.0367	AT2G21210	SAUR-like auxin-responsive protein family
AT5G61590	0.03665	ERF107	Ethylene-responsive transcription factor ERF107
AT4G19700	0.0365	BOI	E3 ubiquitin-protein ligase BOI
AT3G02840	0.03645	AT3G02840	ARM repeat superfamily protein
AT2G18440	0.0362	GUT15	GUT15 (GENE WITH UNSTABLE TRANSCRIPT 15)
AT1G14540	0.0361	PER4	Peroxidase
AT2G37430	0.0354	ZAT11	ZAT11
AT1G56240	0.035	PP2B13	F-box protein PP2-B13
AT5G25350	0.03455	EBF2	EIN3-binding F-box protein 2
AT5G60390	0.02035	A1	Elongation factor 1-alpha 4

```
<Figure size 432x288 with 1 Axes>
```

obipy-resources/featbar.png

Figure 1: Genes importance in determining Chitin and Water tr :tangle recursive_feature.pyeatments in Col0

2.2 25 vars

```
from sklearn.feature_selection import RFE, RFECV
from sklearn.linear_model import LogisticRegression

# load data
DE_pairings_05hr = read_xl('./Data/pairings_05hr.xlsx')
sig = DE_pairings_05hr[DE_pairings_05hr['padj'] < 0.05]
sig = sig['log2FoldChange'].sort_values()
locs = sig.index
df = counts.loc[locs][[c for c in counts.columns if ('_05h' in c and 'col' in c)]].T
df = df.loc[:,~df.columns.duplicated()]
df = df[[c for c in set(df.columns.values)]]
```

```
# Feature Extraction with RFE
X = df.values
y = [y.rsplit('_',1)[0] for y in df.reset_index()['index']]
# feature extraction
model = LogisticRegression()
rfe = RFE(model, n_features_to_select=250)
fit = rfe.fit(X, y)
print("Num Features: {0}".format(fit.n_features_))
print("Selected Features: {0}".format(fit.support_))
print("Feature Ranking: {0}".format(fit.ranking_))
```

Num Features: 250 Selected Features: [False False False ... False False False] Feature Ranking: [169 3607 2703 ... 3312 4549 1690]

	incoming	name	description
0	AT5G51190	ERF105	Ethylene-responsive transcription factor ERF105
1	AT2G47440	AT2G47440	Tetratricopeptide repeat (TPR)-like superfamily protein
2	AT1G15690	AVP1	VHP1
3	AT2G41840	RPS2C	40S ribosomal protein S2-3
4	AT2G43340	AT2G43340	At2g43340
5	AT4G31550	WRKY11	Probable WRKY transcription factor 11
6	AT1G78100	AT1G78100	AUF1
7	AT5G61590	ERF107	Ethylene-responsive transcription factor ERF107
8	AT5G20290	RPS8A	40S ribosomal protein S8
9	AT3G29000	CML45	Probable calcium-binding protein CML45
10	AT5G64905	PEP3	Elicitor peptide 3
11	AT1G25440	COL16	Zinc finger protein CONSTANS-LIKE 16
12	AT3G53870	RPS3B	40S ribosomal protein S3-2
13	AT1G56242	AT1G56242	other RNA
14	AT1G80080	TMM	TMM
15	AT2G25735	AT2G25735	Expressed protein
16	AT1G30370	AT1G30370	DLAH
17	AT2G01180	ATPAP1	phosphatidic acid phosphatase 1
18	AT5G15870	AT5G15870	Glycosyl hydrolase family 81 protein
19	AT5G61570	AT5G61570	Protein kinase superfamily protein
20	AT1G13930	AT1G13930	At1g13930/F16A14.27
21	AT3G22121	AT3G22121	other RNA
22	AT5G24110	WRKY30	Probable WRKY transcription factor 30
23	AT2G21210	AT2G21210	SAUR-like auxin-responsive protein family
24	AT5G61600	ERF104	Ethylene-responsive transcription factor ERF104
25	AT3G49010	RPL13B	60S ribosomal protein L13-1
26	AT4G13930	SHM4	Serine hydroxymethyltransferase 4
27	AT3G63200	PLP9	Probable inactive patatin-like protein 9
28	AT4G16720	RPL15A	Ribosomal protein L15
29	AT2G10940	AT2G10940	At2g10940/F15K19.1
30	AT2G40140	CZF1	Zinc finger CCCH domain-containing protein 29
31	AT1G50040	AT1G50040	F2J10.8 protein
32	AT4G27280	KRP1	Calcium-binding protein KRP1
33	AT1G67430	RPL17B	60S ribosomal protein L17-2
34	AT3G28340	GATL10	Hexosyltransferase (Fragment)
35	AT5G59730	ATEXO70H7	Exocyst subunit Exo70 family protein
36	AT4G38860	AT4G38860	At4g38860
37	AT1G26380	FOX1	Berberine bridge enzyme-like 3
38	AT4G14450	AT4G14450	Uncharacterized protein At4g14450, chloroplastic
39	AT1G61360	AT1G61360	Serine/threonine-protein kinase
40	AT1G72370	RPSAA	40S ribosomal protein SA
41	AT4G18197	PUP7	Probable purine permease 7
42	AT1G59590	ZCF37	At1g59590
43	AT2G35930	PUB23	E3 ubiquitin-protein ligase PUB23
44	AT1G27730	ZAT10	Zinc finger protein ZAT10
45	AT1G79110	BRG2	Probable BOI-related E3 ubiquitin-protein ligase 2
46	AT3G18710	PUB29	RING-type E3 ubiquitin transferase
47	AT2G45180	AT2G45180	At2g45180/T14P1.1
48	AT5G56030	HSP81-2	Heat shock protein 81-2
49	AT1G78170	AT1G78170	E3 ubiquitin-protein ligase
50	AT1G19020	AT1G19020	CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase
51	AT2G21660	RBG7	Glycine-rich RNA-binding protein 7
52	AT5G41750	AT5G41750	Disease resistance protein (TIR-NBS-LRR class) family
53	AT3G55980	SZF1	Salt-inducible zinc finger 1
54	AT3G23810	SAHH2	Adenosylhomocysteinase
55	AT5G57760	AT5G57760	At5g57760
56	AT1G12090	ELP	ELP
57	AT5G25340	AT5G25340	Ubiquitin-like superfamily protein
58	AT3G23250	MYB15	Transcription factor MYB15
59	AT1G24145	AT1G24145	At1g24145
60	AT5G44680	AT5G44680	DNA glycosylase superfamily protein
61	AT3G44260	CAF1-9	Probable CCR4-associated factor 1 homolog 9
62	AT2G36530	ENO2	LOS2
63	AT2G28000	CPN60A1	SLP
64	AT4G13940	SAHH1	Adenosylhomocysteinase 1

2.2.1 Forest on this RFE set

gene	importance	name	description
AT3G07195	0.00485	AT3G07195	RPM1-interacting protein 4 (RIN4) family protein
AT5G15200	0.0048	RPS9B	40S ribosomal protein S9-1
AT5G56030	0.0047	HSP81-2	Heat shock protein 81-2
AT4G02410	0.0046	LECRK43	L-type lectin-domain containing receptor kinase IV.3
AT4G13930	0.00455	SHM4	Serine hydroxymethyltransferase 4
AT2G18440	0.0045	GUT15	GUT15 (GENE WITH UNSTABLE TRANSCRIPT 15)
AT4G14370	0.0045	AT4G14370	Disease resistance protein (TIR-NBS-LRR class) family
AT2G45180	0.0045	AT2G45180	At2g45180/T14P1.1
AT1G69530	0.00445	ATEXPA1	Expansin
AT1G12090	0.0044	ELP	ELP
AT1G21120	0.0044	AT1G21120	O-methyltransferase family protein
AT3G09275	0.0044	AT3G09275	None
AT3G10930	0.0044	AT3G10930	Uncharacterized protein At3g10930
AT5G25250	0.00435	FLOT1	Flotillin-like protein 1
AT4G39260	0.00435	RBG8	GRP8
AT1G68550	0.00435	ERF118	Ethylene-responsive transcription factor ERF118
AT5G59780	0.00435	MYB59	Transcription factor MYB59
AT4G23220	0.00435	CRK14	Cysteine-rich receptor-like protein kinase 14
AT1G79680	0.0043	WAKL10	Wall-associated receptor kinase-like 10
AT2G27820	0.0043	ADT3	Arogenate dehydratase 3, chloroplastic
AT1G17420	0.0043	LOX3	Lipoxygenase 3, chloroplastic
AT1G14540	0.0043	PER4	Peroxidase
AT3G52400	0.0043	SYP122	Syntaxin-122
AT3G09500	0.0043	RPL35A	60S ribosomal protein L35-1
AT2G35935	0.0043	AT2G35935	None
AT2G40140	0.00425	CZF1	Zinc finger CCCH domain-containing protein 29
AT5G02500	0.00425	MED37E	Probable mediator of RNA polymerase II transcription subunit 37e
AT5G44680	0.00425	AT5G44680	DNA glycosylase superfamily protein
AT2G36530	0.00425	ENO2	LOS2
AT4G14450	0.00425	AT4G14450	Uncharacterized protein At4g14450, chloroplastic
AT5G41740	0.00425	AT5G41740	Disease resistance protein (TIR-NBS-LRR class) family
AT5G47850	0.0042	CCR4	Serine/threonine-protein kinase-like protein CCR4
AT4G11470	0.0042	CRK31	cysteine-rich RLK (RECEPTOR-like protein kinase) 31
AT4G22030	0.0042	AT4G22030	F-box family protein with a domain of unknown function (DUF295)
AT5G22250	0.0042	CAF1-11	Probable CCR4-associated factor 1 homolog 11
AT4G23180	0.00415	CRK10	Cysteine-rich receptor-like protein kinase 10
AT1G80840	0.00415	WRKY40	Probable WRKY transcription factor 40
AT5G04340	0.00415	ZAT6	Zinc finger protein ZAT6
AT5G37260	0.00415	RVE2	Protein REVEILLE 2
AT1G61360	0.00415	AT1G61360	Serine/threonine-protein kinase
AT4G01250	0.00415	WRKY22	WRKY transcription factor 22
AT3G59940	0.0041	SKIP20	F-box/kelch-repeat protein SKIP20
AT2G02630	0.0041	AT2G02630	Cysteine/Histidine-rich C1 domain family protein
AT4G19700	0.0041	BOI	E3 ubiquitin-protein ligase BOI
AT2G22880	0.0041	AT2G22880	At2g22880
AT1G30370	0.0041	AT1G30370	DLAH
AT1G20310	0.0041	AT1G20310	Syringolide-induced protein
AT3G49010	0.0041	RPL13B	60S ribosomal protein L13-1
AT2G23270	0.0041	AT2G23270	At2g23270
AT4G28460	0.0041	PIP1	PAMP-induced secreted peptide 1
AT3G05590	0.0041	RPL18B	RPL18
AT2G44670	0.00405	FLZ3	FCS-Like Zinc finger 3
AT1G69890	0.00405	AT1G69890	Actin cross-linking protein (DUF569)
AT2G30520	0.00405	RPT2	Root phototropism protein 2
AT1G61560	0.00405	MLO6	MLO-like protein 6
AT5G52050	0.004	DTX50	Protein DETOXIFICATION 50
AT1G09780	0.004	PGM1	IPGAM1
AT2G25735	0.004	AT2G25735	Expressed protein
AT4G18100	0.004	RPL32A	60S ribosomal protein L32-1
AT4G37540	0.004	LBD39	LOB domain-containing protein 39
AT4G18197	0.004	PUP7	Probable purine permease 7
AT1G24140	0.004	3MMP	Metalloendoproteinase 3-MMP
AT5G13930	0.004	CHS	Chalcone synthase family protein
AT1G69760	0.004	AT1G69760	At1g69760
AT2G01180	0.004	ATPAP1	phosphatidic acid phosphatase 1

```
<Figure size 1440x2160 with 2 Axes>
```

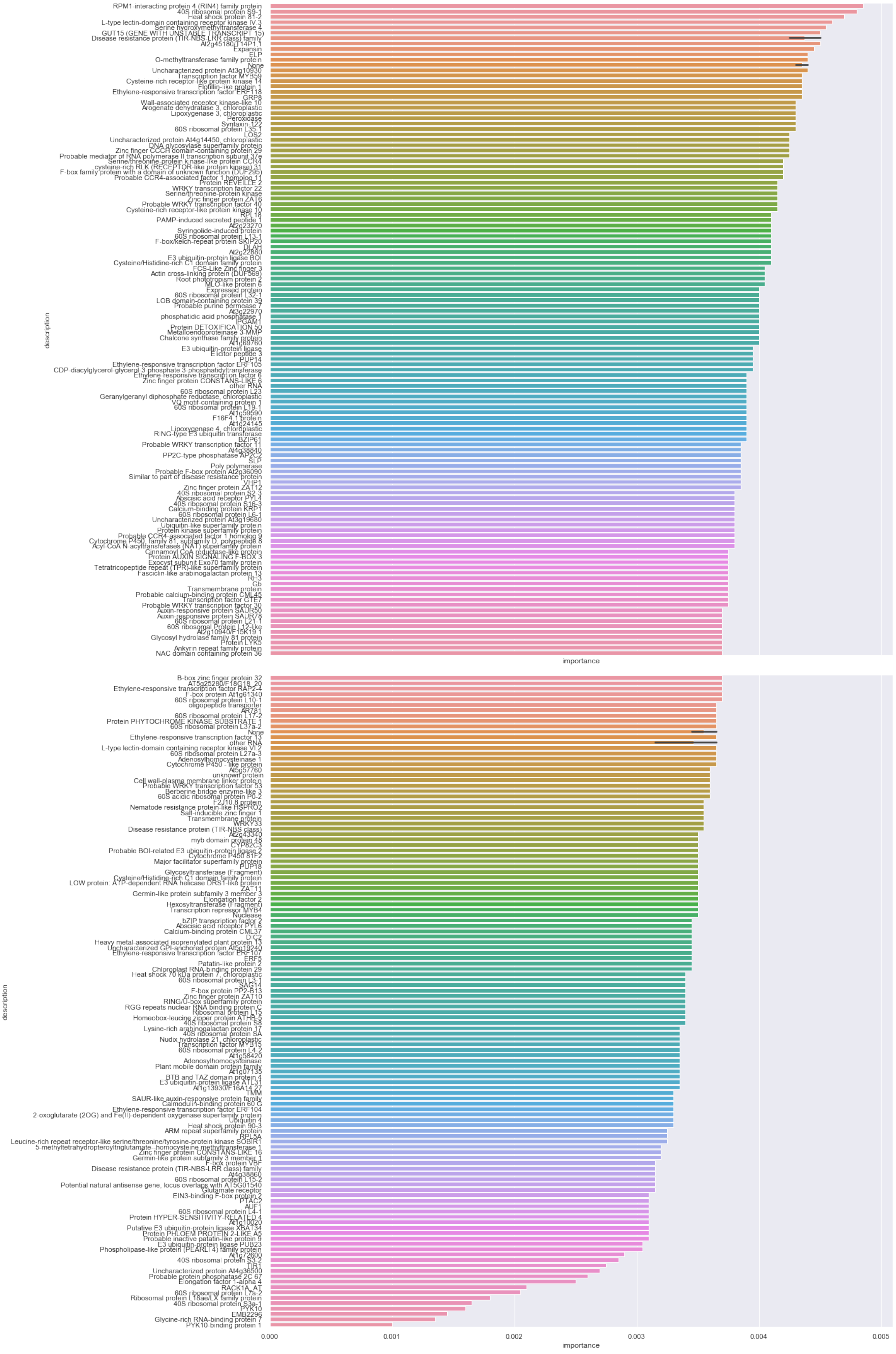

Figure 2: Genes importance in determining Chitin and Water tr :tangle recursive_feature.pyeatments in Col0