Analysis of Gene Imprinting - Preliminary Findings

Gitanshu Munjal

March 19th, 2015

Dr. Comai,

Presented herein are some preliminary findings (with sufficiently commented R code) from exploratory analyses of the data you shared with us. Overall, the findings look promising and we saw evidence for both paternally expressed genes (PEGs) and maternally expessed genes (MEGs). As you will find in reviewing this document, the Harada data set was of great value in identifying MEGs and also helped confirm some of the PEG findings.

Kind regards,
Gitanshu

Setting-up shop

```
#########################
# For reproducibility, make a project directory similar to mine.
# You only need change the working directory variable path for the rest of the analysis.

# My project directory contains three folders as follows:
# (1) data   - this folder contains the raw data my hypothetical PI and you shared with me
# (2) output - this folder is where the tables and figures I generate here will go
# (3) code   - this folder contains all the R code presented here combined into one script

#########################

#Set working directory for this session. This should be the only thing needing changes
#assuming the name of the data file is still the same.
setwd("C:\\Users\\uglysweaters\\Desktop\\Comai\\")

#read in the table you shared with me.
#please modify this if you're using a different file name now.
imprint <- read.table("data\\f_ggg_lc.dat", header = T, sep = "\t")
```

Identifying Paternally Expressed Genes (PEGs)

In order to identify PEGs from the complete dataset, I naively defined PEGs as instances where paternal contribution in both (reciprocal) hybrids were greater than maternal contributions. PEGs identified in this manner were also compared with the Harada set to determine how many of these were present/expressed in the endosperm at all during the globular stage.

```
########################
# Begin Exploratory Analyses
########################

########################
# PEGs
########################

#retreieve naively defined raw PEGs and write to table
#Write a separate table containing GeneIDs for complete set
peg <- imprint[imprint$X.Col_in_ColxPur < imprint$X.Pur_in_ColxPur &
                 imprint$X.Col.PurxCol > imprint$X.Pur.PurxCol ,]
peg <- na.omit(peg)

write.table(peg, "output\\rawpegs.txt", sep = "\t", quote = F, row.names = F)
write.table(peg[,1], "output\\PEGs-id.txt",
            sep = "\t", quote = F, row.names = F, col.names = "GeneID")
nrow(peg)
```

```
## [1] 45
```

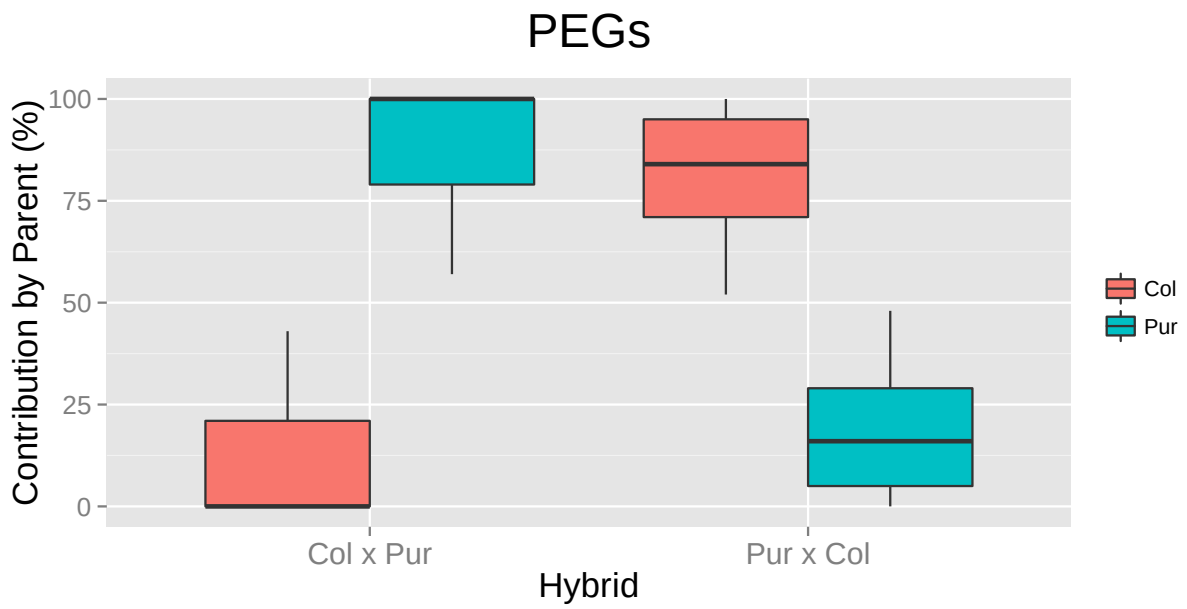
```
#filter PEGs by using Harada data
#Write filtered PEGs to table
filterpeg <- peg[peg$micropylar_endosperm != 0 |
                   peg$peripheral_endosperm != 0 |
                   peg$chalazal_endosperm != 0, ]

write.table(filterpeg, "output\\PEGs-Haradaconfirmed.txt",
            sep = "\t", quote = F, row.names = F)
nrow(filterpeg)
```

```
## [1] 32
```

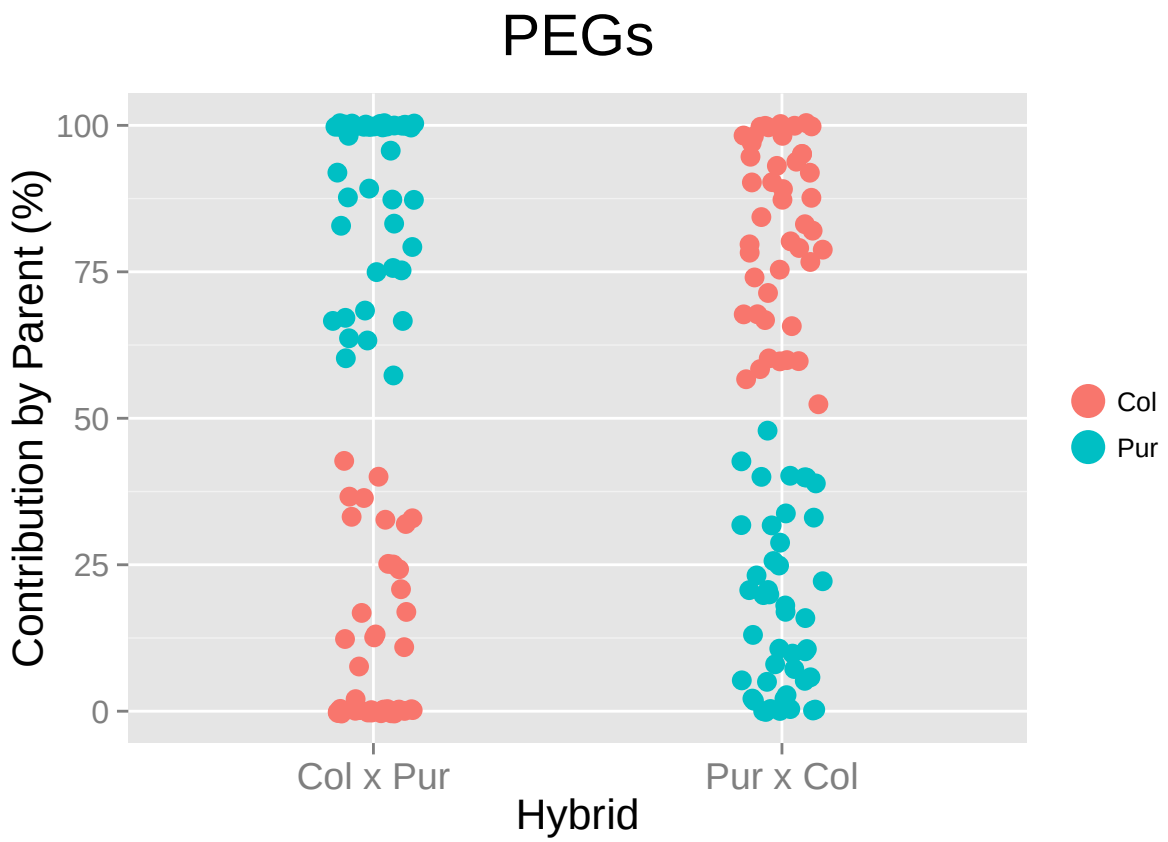
So overall, it would appear that there were 45 putative PEGs in our dataset and 32 of these are confirmed as expressed in the endosperm by the Harada dataset.

Next, I sought to visualize these data by making some box and jitter plots as follows:

```
##################
# Visualize PEGs
##################

#make an empty data frame with columns for Parent, Contribution, and Hybrid
pegplot <- matrix( nrow = nrow(peg), ncol = 3)
colnames(pegplot) <- c("Parent","Contribution","Hybrid")

#use rounded-up percent contributions from PEGs
pegplot[, 1:3] <- cbind("Col", round(peg[,3],0), "Col x Pur")
pegplot <- rbind(pegplot, cbind("Pur",round(peg[,4],0),"Col x Pur"))
pegplot <- rbind(pegplot, cbind("Col",round(peg[,6],0),"Pur x Col"))
pegplot <- rbind(pegplot, cbind("Pur",round(peg[,7],0),"Pur x Col"))

pegplot <- as.data.frame(pegplot)
pegplot$Contribution <- as.double(as.character(pegplot$Contribution))

library(ggplot2)
p2 <- ggplot(pegplot, aes(factor(Hybrid), Contribution))

#boxplot
pegbox <- p2 + geom_boxplot(aes(fill = factor(Parent)),) +
  labs(y = "Contribution by Parent (%)", x = "Hybrid", title = "PEGs") +
  theme(legend.title= element_blank(),legend.key = element_rect(fill='NA')) +
  guides(colour = guide_legend(override.aes = list(size=6))) +
  theme(plot.title = element_text(size=22,lineheight=.2, vjust=2),
        axis.title.x = element_text(size=16, lineheight=4),
        axis.title.y = element_text(size=16, vjust=0.7),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=12))

pegbox
```



```
#jitterplot
pegjit <- p2 + geom_jitter(aes(color = factor(Parent)),
                           size=3.5,position = position_jitter(width = .1)) +
  labs(y = "Contribution by Parent (%)", x ="Hybrid", title = "PEGs") +
  theme(legend.title = element_blank(), legend.key = element_rect(fill='NA')) +
  guides(colour = guide_legend(override.aes = list(size=6))) +
  theme(plot.title = element_text(size=22,lineheight=.2, vjust=2),
        axis.title.x = element_text(size=16, lineheight=4),
        axis.title.y = element_text(size=16, vjust=0.7),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=12))

pegjit
```



As you can probably tell from both above presented figures, paternal contributions from the dataset naively (as previously defined) reduced from the complete dataset indeed represent the expectation of significantly higher expression from PEGs. Thus, I concluded that **there is evidence for PEGs** in our dataset.

Identifying Maternally Expressed Genes (MEGs)

In order to identify MEGs from the complete dataset, I removed SNPs that were significantly, at the 95% confidence level, different from our 2:1 (Maternal:Paternal) expectation for the endosperm.

```
########################
# Retrieve Non-Significant (for 2:1 endosperm expectation) SNPs
########################

#filter for SNPs that are not ignificantly different at the 95% confidence level
sigimprint <- imprint[imprint$adj_P > 0.05, ]
sigimprint <- sigimprint[sigimprint$SNP.used > 0, ]
sigimprint <- na.omit(sigimprint)

nrow(sigimprint)
```

```
## [1] 7363
```

Only these 7363 SNPs were retained for further MEGs analyses. MEGs were naively defined as instances where maternal contribution in both (reciprocal) hybrids were greater than paternal contributions.

```
meg <- sigimprint[sigimprint$X.Col_in_ColxPur > sigimprint$X.Pur_in_ColxPur &
                    sigimprint$X.Col.PurxCol < sigimprint$X.Pur.PurxCol,]
meg <- na.omit(meg)
write.table(meg, "output\\rawmegs.txt", sep = "\t", quote = F, row.names = F)
nrow(meg)
```

```
## [1] 6936
```

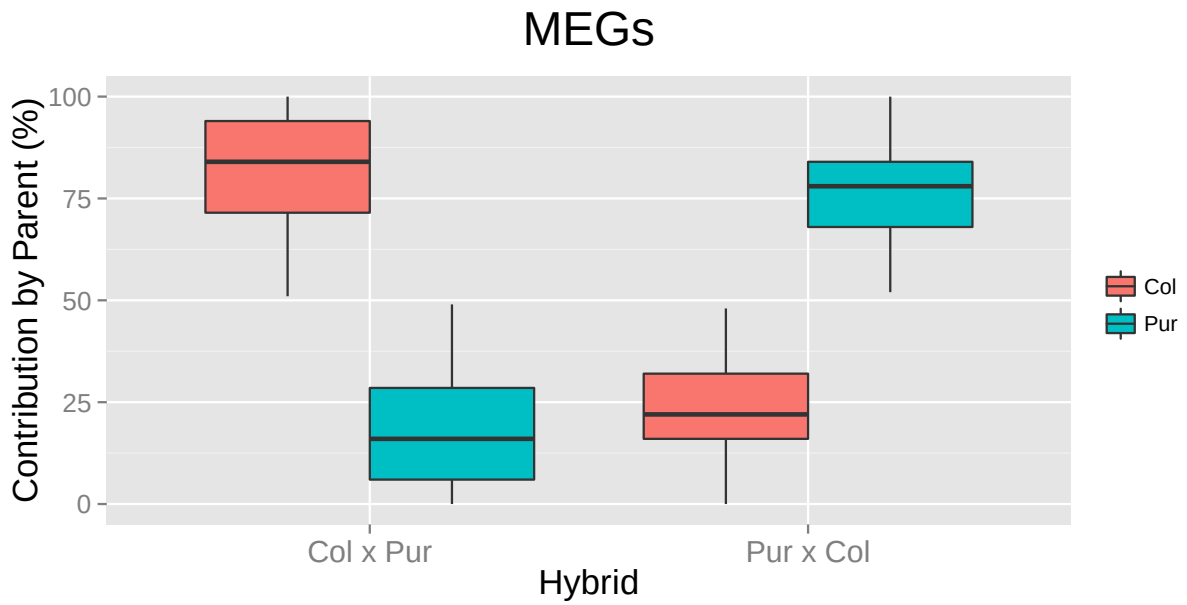
Similar to the plots for PEGs, but with reversed effects, the box and jitter plots (not shown here, included in supplement at the end) made using these 6936 naive SNPs showed **evidence for MEGs**. However, the signal was much noisier and I suspected that even post-filtering for adjusted p-values, the data were still contaminated by non-endosperm maternal tissue. In order to deal with this, I further filtered the naive MEGs by using the Harada data and imposing the condition that expression values for all tissues except endosperm were less than 7.5 and that expression values for at lease one of the endosperm related tissues (chalazal,micropylar, and peripheral) was greater than 9. I arrived at this criteria by trial and error with the PEG data that I was more confident with. The filtering scheme was implemented as follows:

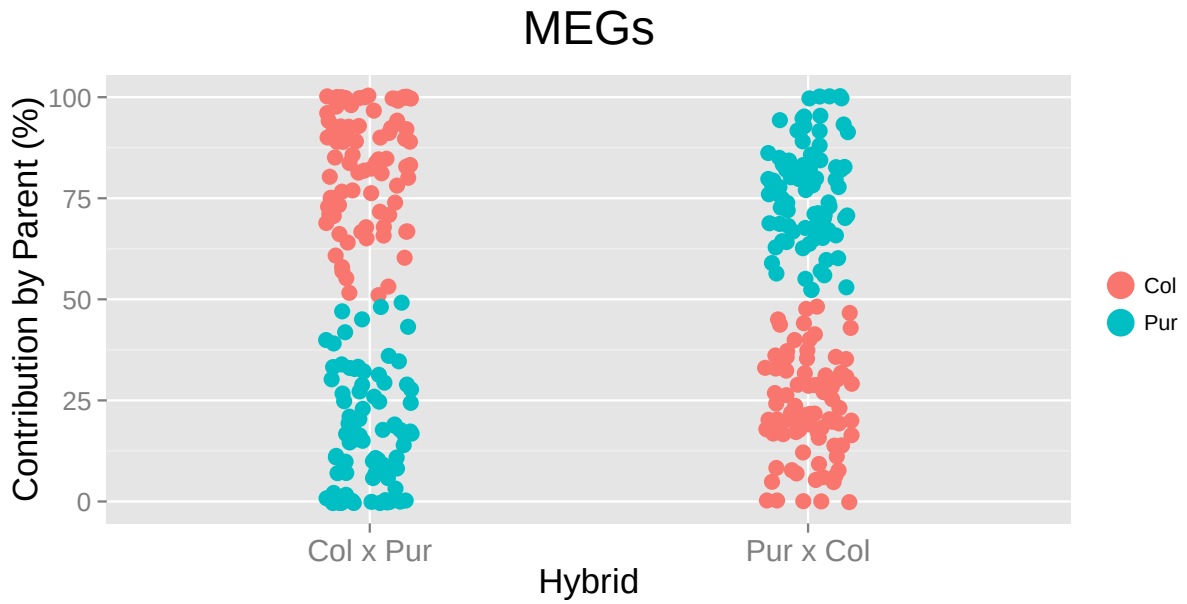
```
#filter MEGs by using Harada data
#Write filtered MEGs to table
#Write a separate table containing GeneIDs for filtered set
filtermeg <- meg[meg$general_seed_coat<7.5 &
                   meg$chalazal_seed_coat<7.5 &
                   meg$embryo<7.5 &
                   meg$suspensor<7.5,]

filtermeg <- filtermeg[filtermeg$micropylar_endosperm>9 |
                         filtermeg$peripheral_endosperm>9 |
                         filtermeg$chalazal_endosperm>9 ,]

write.table(filtermeg, "output\\MEGs-filteredset.txt",
            sep = "\t", quote = F, row.names = F)
```

```
write.table(filtermeg[,1], "output\\MEGs-id.txt",
            sep = "\t", quote = F, row.names = F, col.names = "GeneID")

nrow(filtermeg)
```

```
## [1] 83
```

Thus, using my method described above, I found these 83 putative MEGs. Further, similar to the PEGs, I visualized the MEGs as follows:





Just to summarize what I've presented here already

```
##                            SNPs
## Non-Significant            7363
## potential PEGs               45
## PEGs confirmed by Harada     32
## potential MEGs             6936
## MEGs confirmed by Harada     83
```

Using the Gene ID's for the MEGs confirmed by the Harada dataset and the complete set of PEGs (NOTE: Gene-IDs were written to separate files (see "PEGs-id.txt" and "MEGs-id.txt" in the preceeding code)), I tried to establish the identity of these simply by feeding my Gene-ID files into https://www.arabidopsis.org/tools/bulk/genes/. The results look pretty promising and some insights from a brief scan of these tells me that:

1. The naive methods I used are somewhat reliable, at least for PEGs, for preliminary analyses as I was able to detect 4 genes previously described as "paternally expressed imprinted gene" (namely: AT1G48910, AT1G57800, AT4G11940, AT5G63740). For MEGs, I reliably detected one gene previously known to be specific fo the globular stage of development (namely AT5G59810).
2. A large number of the detected PEGs are described as being involved in the regulation of transcription indicating possibly profound downstream effects for these.
3. Two the MEGs (AT1G05190 and AT4G13380) are described as being involved in "translation, embryo development ending in seed dormancy." A large number of the detectedd MEGs are also involved in the regulation of transcription.

My current efforts will likely be devoted to fine tuning my filtering methods for the MEGs as I feel the parameters I used here (by trial and error in the PEG data) are likely not optimal.

Supplement