Figures for Famdenovo Output

Contents

1. ROC

1.1 Structure of the data sets

The data sets in section 1 have the following structure.

```
#  (mda.family, nci.family, chop.family, dfci.family)
# structure for the data set using mda as an example
knitr::kable(data.frame(
  Variables = colnames(comb4),
  Meaning = c("Family ID", "Truth", "De Novo Probability")
), align = 'c')
```

Variables	Meaning
fam.id	Family ID
truth	Truth
prob	De Novo Probability

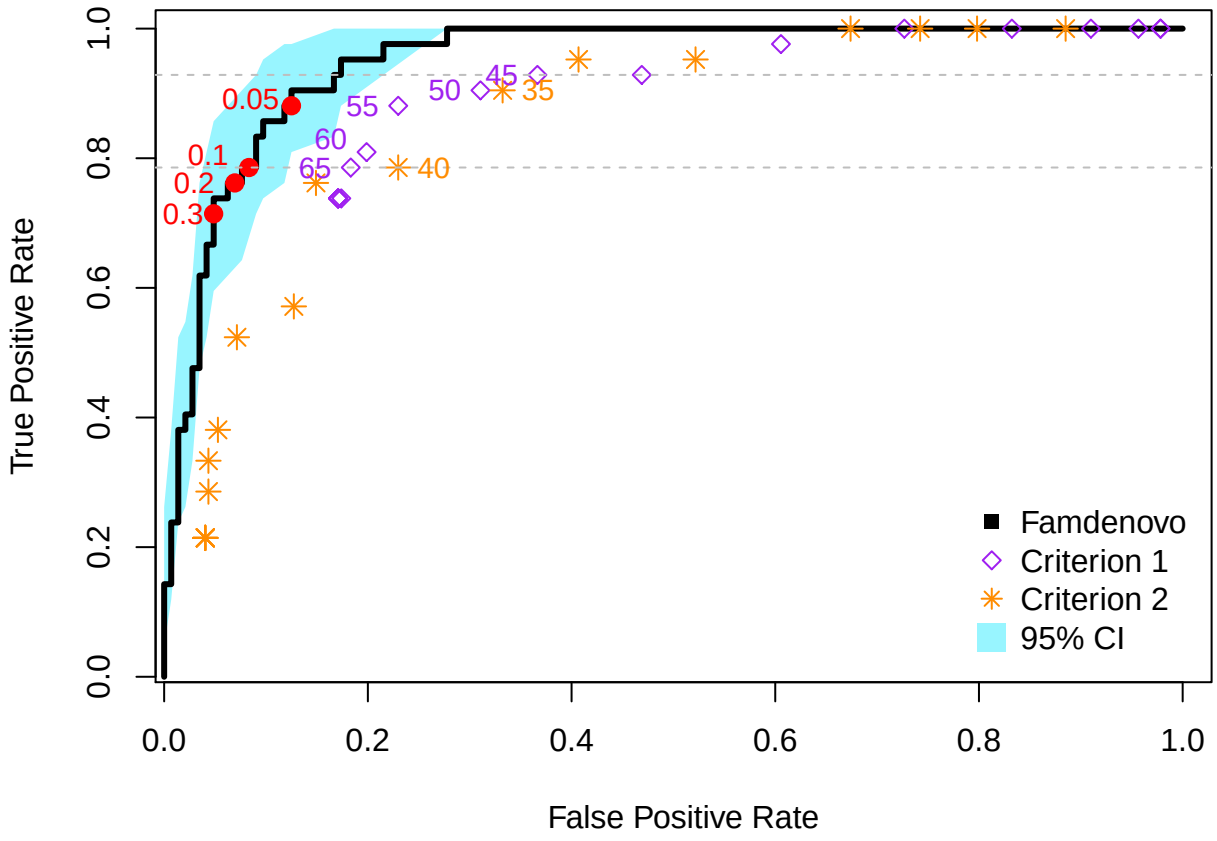
```
# these are the first 3 rows
knitr::kable(head(comb4, n=3), align = 'c')
```

fam.id	truth	prob
79	0	0.0011513
80	0	0.0000202
81	0	0.8014998

1.2 ROC curve

```
# get roc object
rocobj = roc(comb4$truth, comb4$prob)
rocobj$fpr <- 1-rocobj$specificities
rocobj$tpr <- rocobj$sensitivities

# get ci
ci_rocobj = ci.thresholds(rocobj, boot.n=1000)
rocobj_ci_specificity <- as.vector(ci_rocobj$specificity[,2])
ci_rocobj_fpr <- 1- rocobj_ci_specificity
ci_rocobj_upper_bound <- as.vector(ci_rocobj$sensitivity[,3])
ci_rocobj_lower_bound <- as.vector(ci_rocobj$sensitivity[,1])

# plot the ROC
par(mar=c(4,4,1,1))
plot(y=rocobj$sensitivities, x=1-rocobj$specificities,ylab="True Positive Rate",
     xlab="False Positive Rate", xlim = c(0.03,0.99), ylim = c(0.03,0.99), type='l')
polygon(c(ci_rocobj_fpr , rev(ci_rocobj_fpr )), c(ci_rocobj_upper_bound, rev(ci_rocobj_lower_bound)), c
lines(y=rocobj$sensitivities, x=1-rocobj$specificities, lwd=3)
points(c(1 - 0.9514, 1 - 0.9306, 1 - 0.9167, 1 - 0.8750), c(0.7143, 0.7619, 0.7857, 0.8810), col = "red
points(roc.relaxed$fpr, roc.relaxed$tpr, col="purple", pch=23)
```

```
points(roc.stringent$tpr~roc.stringent$fpr, col = "darkorange", pch = 8, cex = 1.2)
legend('bottomright', legend=c('Famdenovo', 'Criterion 1', 'Criterion 2','95% CI'),
       pch=c(15, 23,8,15), pt.cex=c(1,1,1,2), box.lwd = 0,box.col = "white",bg = "white",
       col=c('black',  'purple', 'darkorange', 'cadetblue1'), inset=0.01)
abline(h = roc.relaxed$tpr[9], col = "gray", lty = 2) # age 45
abline(h = roc.relaxed$tpr[13], col = "gray", lty = 2) # age 65
text(y=age_points_stringent$tpr,x=age_points_stringent$fpr+0.035,labels=age_points_stringent$age.cut, c
text(y=age_points_relaxed$tpr,x=age_points_relaxed$fpr-0.035,labels=age_points_relaxed$age.cut, col="pu
text(x=c(1 - 0.9514-0.03, 1 - 0.9306-0.04, 1 - 0.9167-0.04, 1 - 0.8750-0.04), y=c(0.7143, 0.7619, 0.785
```



2. "de novo" probabilities in validation and discovery sets

2.1 Structure of the data sets

The data sets in section 2 have the following structure.

```
#  (mda.valid.extract, nci.valid.extract, chop.valid.extract, dfci.valid.extract)
# structure for the data set using mda.valid as an example
knitr::kable(data.frame(
  Variables = colnames(mda.valid.extract),
  Meaning = c("Family ID","De Novo Probability", "True State", "Prediction", "Is The Prediction Correct
), align = 'c')
```

Variables	Meaning
fam.id	Family ID
prob	De Novo Probability
truth	True State
pred	Prediction
ifcorr	Is The Prediction Correct

```
# these are the first 3 rows
knitr::kable(head(mda.valid.extract, n=3), align = 'c')
```

fam.id	prob	truth	pred	ifcorr
171	0.0026731	FALSE	FALSE	TRUE
172	0.2808838	TRUE	TRUE	TRUE
173	0.2214413	TRUE	TRUE	TRUE

```
# (mda.disc.extract, nci.disc.extract, chop.disc.extract, dfci.disc.extract)
# structure for the data set using mda disc as an example
knitr::kable(data.frame(
  Variables = colnames(mda.disc.extract),
  Meaning = c("Family ID", "De Novo Probability")
), align = 'c')
```

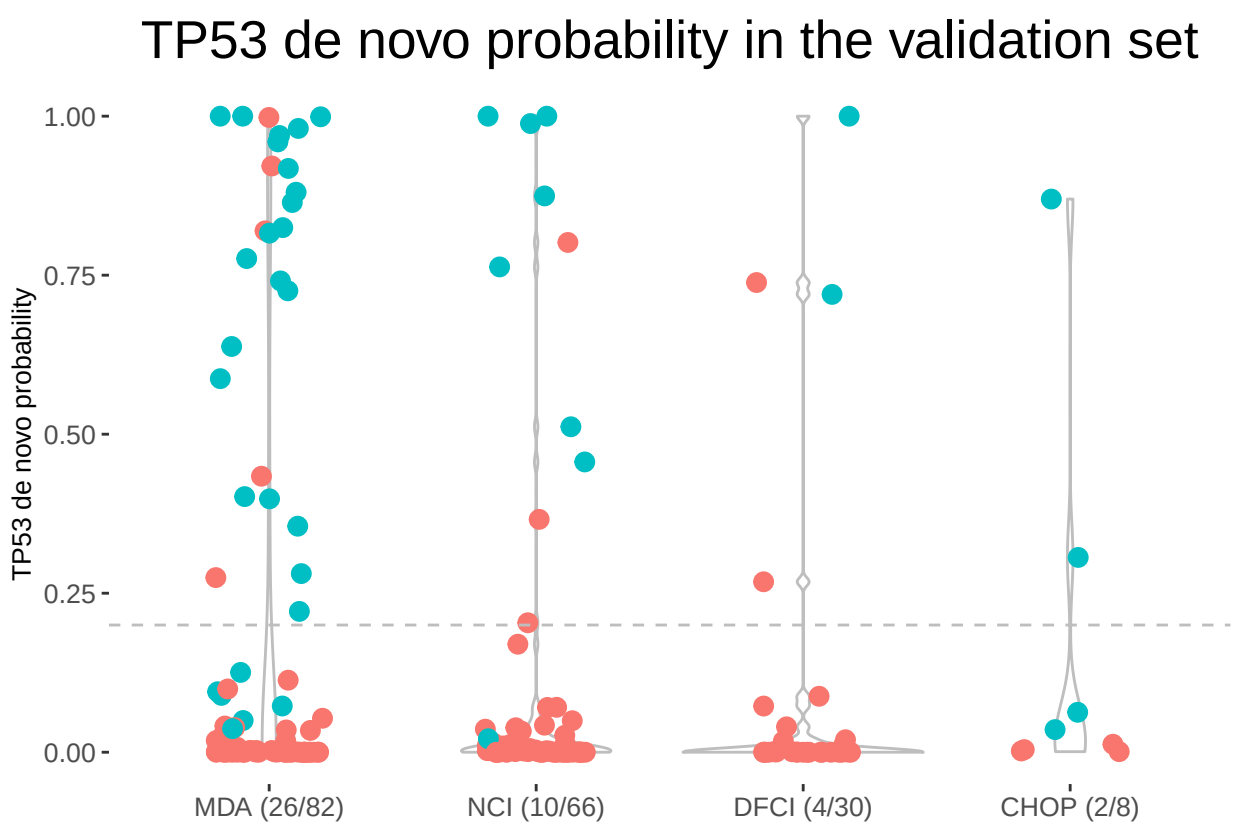
Variables	Meaning
fam.id	Family ID
prob	De Novo Probability

```
# these are the first 3 rows
knitr::kable(head(mda.disc.extract, n=3), align = 'c')
```

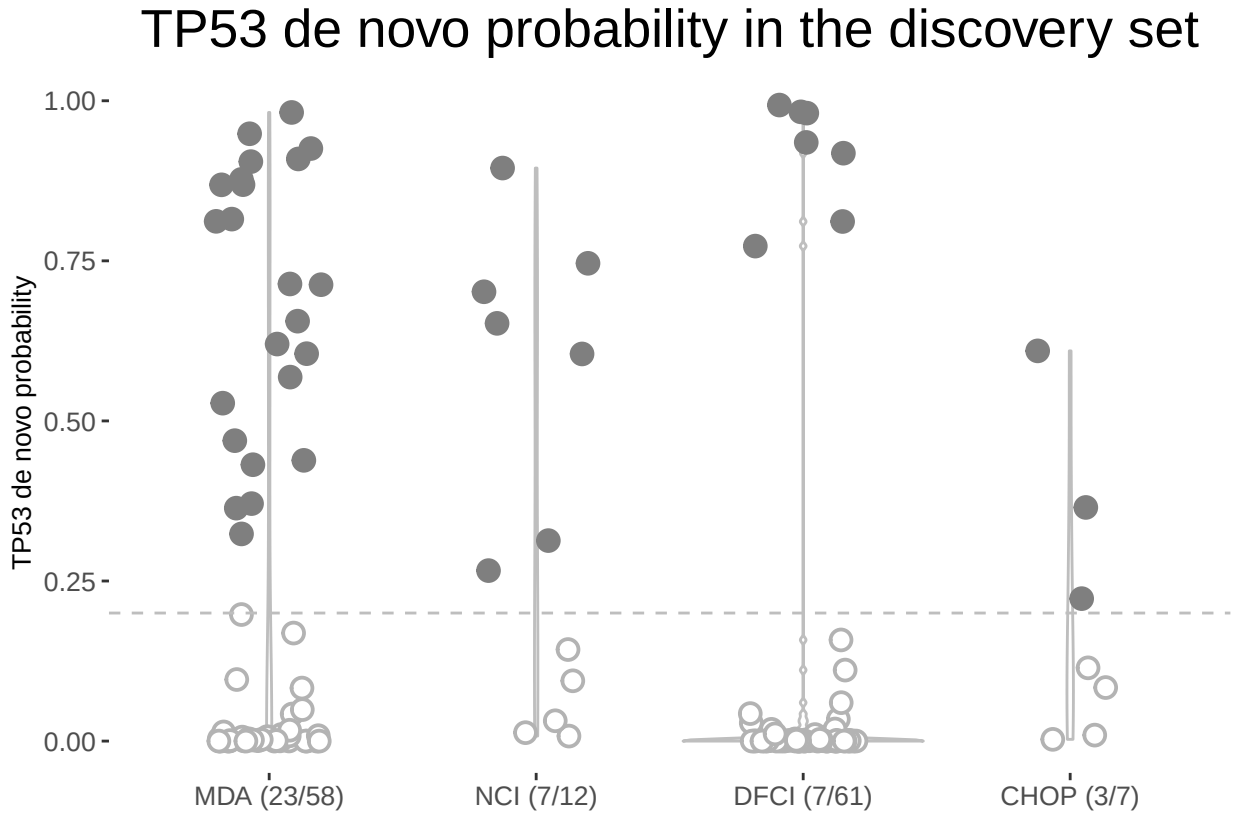
fam.id	prob
170	0.0004403
175	0.0422781
177	0.0000004

```
# ------ De Novo Probability ------ #
# validation set
valid <- rbind(data.frame(name="MDA (26/82)", prob=mda.valid.extract$prob, truth=mda.valid.extract$truth
               data.frame(name="NCI (10/66)", prob=nci.valid.extract$prob, truth=nci.valid.extract$truth
               data.frame(name="DFCI (4/30)", prob=dfci.valid.extract$prob, truth=dfci.valid.extract$tru
               data.frame(name="CHOP (2/8)", prob=chop.valid.extract$prob, truth=chop.valid.extract$tru

# plot the distribution on the validation set
ggplot(valid, aes(factor(name, levels=c("MDA (26/82)", "NCI (10/66)", "DFCI (4/30)", "CHOP (2/8)")), pr
  geom_violin(color="gray") + geom_jitter(height=0, width=0.2, aes(color=truth), size = 3) +
  geom_hline(yintercept=0.2, color="gray", linetype="dashed") +
  labs(title="TP53 de novo probability in the validation set", x="", y="TP53 de novo probability") +
```

```
  theme(panel.background=element_blank(),
        plot.title=element_text(size = 20, hjust=0.5),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10)) +
  guides(color=FALSE)
```




```
# discovery
disc <- rbind(data.frame(name="MDA (23/58)", prob=mda.disc.extract$prob),
              data.frame(name="NCI (7/12)", prob=nci.disc.extract$prob),
              data.frame(name="DFCI (7/61)", prob=dfci.disc.extract$prob),
              data.frame(name="CHOP (3/7)", prob=chop.disc.extract$prob))
disc$pred <- disc$prob >= 0.2

# plot the distribution on the discovery set
ggplot(disc, aes(factor(name, levels=c("MDA (23/58)", "NCI (7/12)", "DFCI (7/61)", "CHOP (3/7)")), prob
  geom_violin(color="gray") +
  geom_point(position=position_jitter(width = 0.2),shape = 21, color=ifelse(disc$pred == TRUE, 'gray50'
  geom_hline(yintercept=0.2, color="gray", linetype="dashed") +
  labs(title="TP53 de novo probability in the discovery set", x="", y="TP53 de novo probability") +
  theme(panel.background=element_blank(),
        plot.title=element_text(size = 20, hjust=0.5),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))+
  guides(color=FALSE)
```



3. Parental and diagnosis age analysis

3.1 Structure of the data set in this section

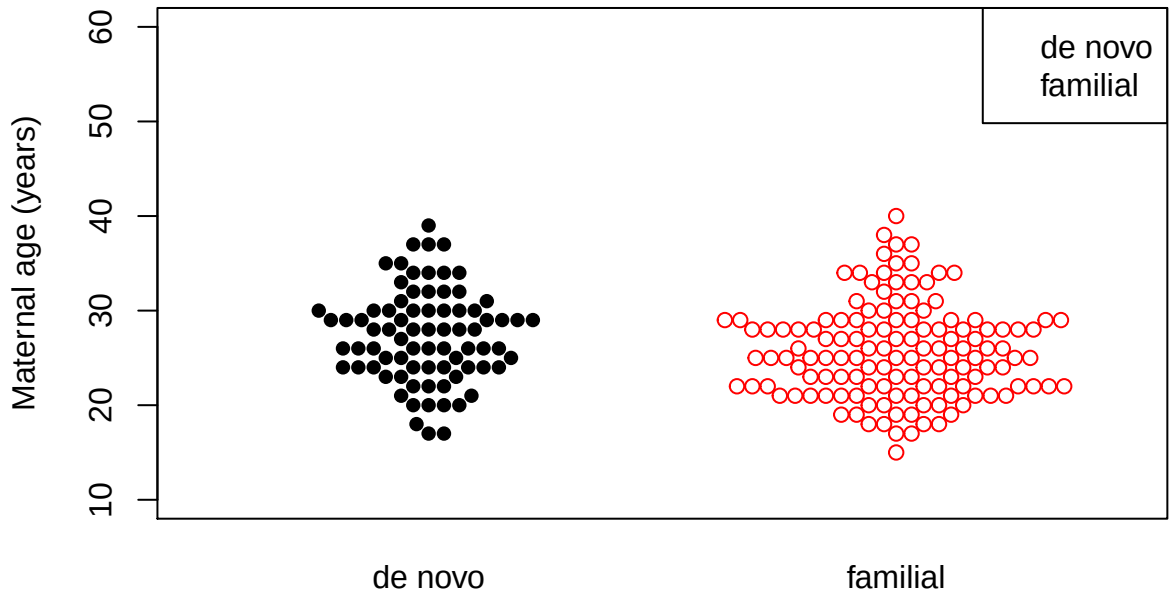
The data set used in this section considers only the probands of each family and has the following structure.

```
# first 3 rows
knitr::kable(head(probands3, n=3), align = 'c')
```

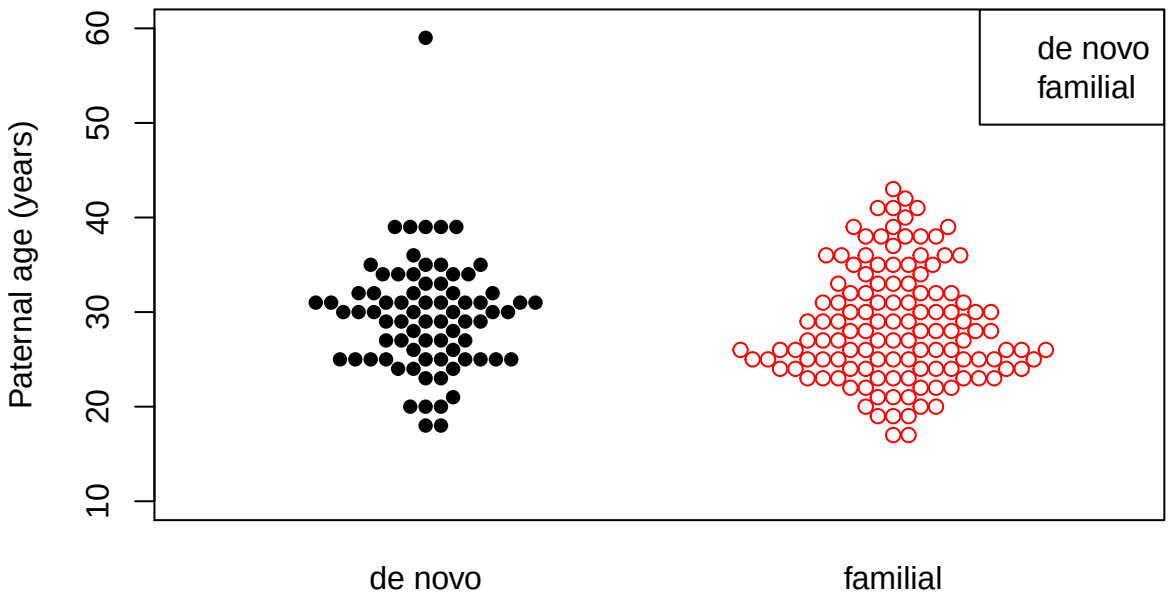
institution.x	Nu.ID	Nu.Fam	prob	truth	paternal.age	maternal.age	proband	Fam.
NCI	93	79	0.0000015	familial	29	24	proband	40
NCI	98	80	0.0000027	familial	21	21	proband	68
NCI	109	81	0.0113485	familial	24	25	proband	18
### 3.2 Several p	lots rega	rding age	of diagnosis	and parental	age at birth of	the proband		

```
#probands 2 mother
beeswarm(as.numeric(probands3$maternal_age) ~ probands3$new_state,
         pwpch = ifelse(probands3$state == "denovo", 16, 1),
         method="swarm", ylim = c(10, 60),
         pwcol = ifelse(probands3$state == "denovo", 1, 2),
```

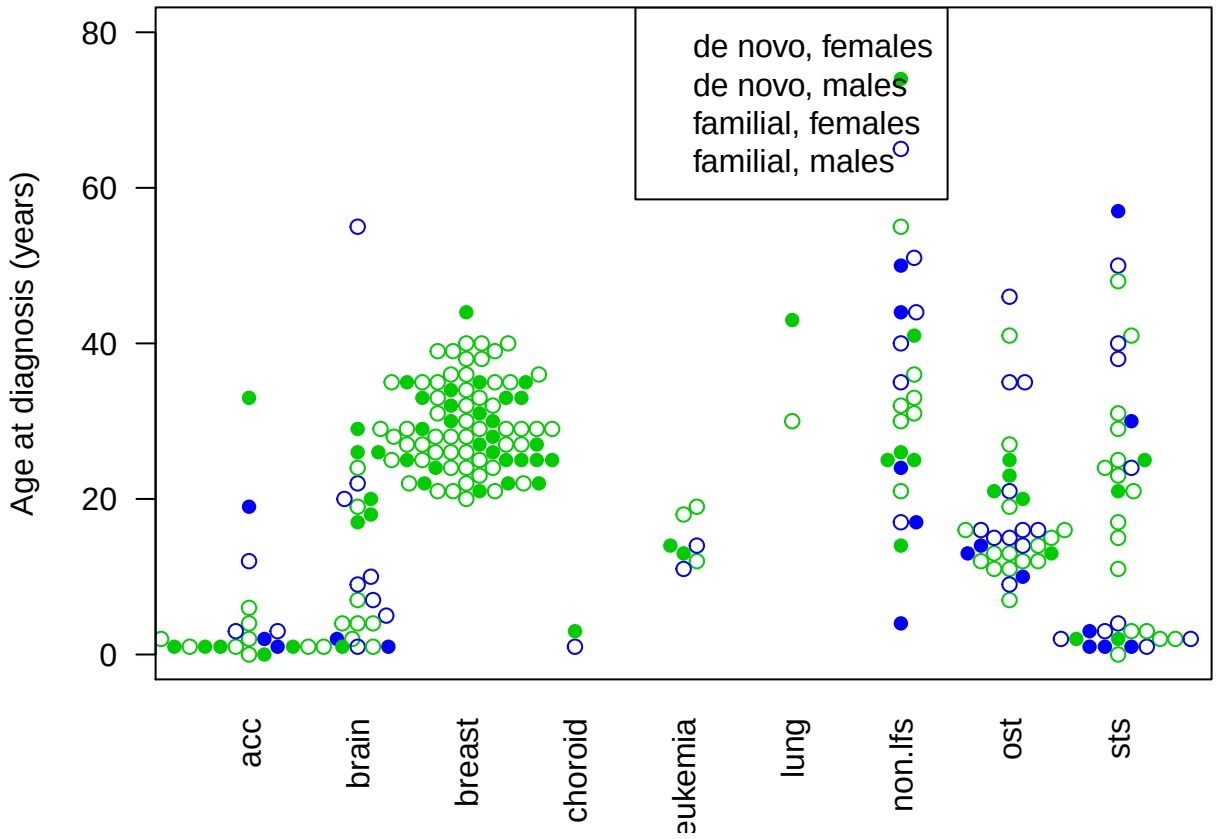
```
        xlab = "", ylab="")
title(ylab="Maternal age (years)")
legend("topright", legend = c("de novo", "familial"), col = c(1, 2))
```




```
# probands 2 father
beeswarm(as.numeric(probands3$paternal_age) ~ probands3$new_state,
         pwpch = ifelse(probands3$state == "denovo", 16, 1),
         method="swarm", ylim = c(10, 60),
         pwcol = ifelse(probands3$state == "denovo", 1, 2),
         xlab = "", ylab="")
title(ylab="Paternal age (years)")
legend("topright", legend = c("de novo", "familial"), col = c(1, 2))
```




```
# all probands
par(mar=c(4,4,1,1))
beeswarm(as.numeric(probands3$diag.age.first) ~ probands3$LFS.Spectrum,#probands3$cancer.first,
         pwpch = ifelse(probands3$state == "denovo", 16, 1),
         pwcol = ifelse(probands3$gender == 0, 3, 4), method="swarm",
         ylim = c(0, 80), xlab="", ylab="", las=2)
title(ylab="Age at diagnosis (years)")
legend("topright", legend = c("de novo, females", "de novo, males", "familial, females", "familial, mal
```



4. Supplementary: Distribution of family size in LFS and BRCA cohorts

5.1 Supplementary: Structure of the data sets in this section

These data sets contain the size of each family belonging to the LFS cohorts used for our Famdenovo.TP53 project and the BRCA families from the CGN cohort used for Famdenovo.BRCA. Lets take a look at the structure of these data sets.

```
# LFS data set structure
knitr::kable(data.frame(Variable = colnames(LFS_Family_Size),
                        Meaning = c("Number of Members Available in Pedigree", "Number of Pedigrees wit
```

Variable	Meaning
lfs_pedigree_size	Number of Members Available in Pedigree
lfs_families	Number of Pedigrees with This Amount of Available Members

```
# first 3 rows
knitr::kable(head(LFS_Family_Size, n =3),align = 'c' )
```

lfs_pedigree_size	lfs_families
3	2

lfs_pedigree_size	lfs_families
4	3
5	1

5.2 Supplementary: Plots of Family Size Distributions

```
# --------- Distribution of Family Size --------- #
# LFS Data plot
plot(x=as.vector(LFS_Family_Size$lfs_pedigree_size), y=as.vector(LFS_Family_Size$lfs_families), type =
      main="LFS Data Family Size Distribution", ylab = 'No. of Families',
     xlim = c(0,150), ylim = c(0,10), xlab = "Family Size")
  minor.tick(nx = 5, ny = 0)
lines(x=as.vector(LFS_Family_Size$lfs_pedigree_size), y=as.vector(LFS_Family_Size$lfs_families),
      type = 'p', pch = 19, cex = 2, col = "Blue")
```