# Discovering mutations

In this module, we’ll use DNA sequencing data from human families to explore the relationship between parental age and *de novo* mutations in their children.

#### Learning objectives

After completing this chapter, you’ll be able to:

1. Create plots to visualize the relationship between two variables.
2. Interpret the results of a linear model.
3. Compare the impact of maternal vs. paternal age on *de novo* mutation counts.
4. Explain what a confidence interval is and why it’s useful.

## *De novo* mutations

Mutation and recombination are two biological processes that generate genetic variation. When these phenomena occur during gametogenesis, the changes that they make to DNA are passed down to the next generation through germline cells (i.e., sperm and oocyte).

***De novo*** **mutations (DNMs)** arise from errors in DNA replication or repair. These mutations can be single-nucleotide polymorphisms (SNPs) or insertions and deletions of DNA. Every individual typically carries around 70 *de novo* SNPs that were not present in either of their parents.

|  |
| --- |
| **Fig. 1.** Sources of DNMs in gametogenesis. |

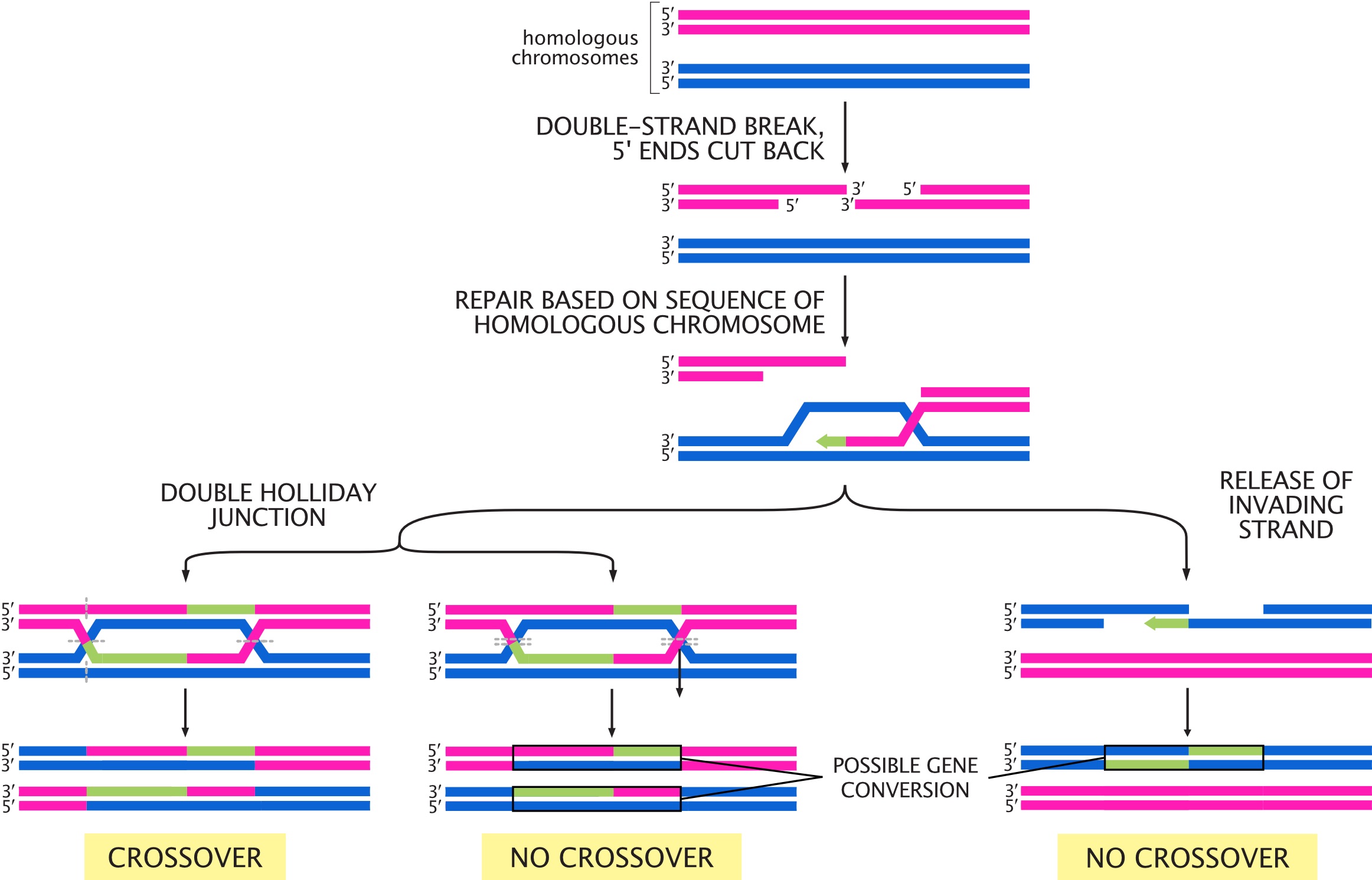
**Fig. 1.** Sources of DNMs in gametogenesis.

## Recombination

**Crossovers**, or meiotic **recombination**, occur during prophase of meiosis I, when homologous chromosomes pair with each other. Double-strand breaks are deliberately generated in the DNA, and are then cut back and repaired based on the sequence of the homologous chromosome. These repairs can sometimes resolve in a crossover event, where sections of DNA are swapped between chromosomes.

Because the sequences of homologous chromosomes differ at sites where they carry different alleles, recombination generates genetic diversity by creating new haplotypes, or combinations of alleles.

Crossovers are required for meiosis in most organisms because they ensure proper homologous chromosome pairing and segregation. Humans experience 1-4 crossover events per chromosome, with longer chromosomes having more crossovers.



**Fig. 2.** Possible outcomes for double-strand breaks generated during meiosis I. Adapted from *Molecular Biology of the Cell, 6th Edition* (Alberts et al.)

## Setup

In this module, we’ll use sequencing data from families to look at the relationship between DNMs, crossovers, and parental age.

### R packages

We’re using R’s tidyverse library to analyze our data. You can load this R package by running:

library(tidyverse)

### Data

Our data comes from the supplementary tables of [this paper by Halldorsson et al.](https://science.sciencemag.org/content/363/6425/eaau1043), which performed whole-genome sequencing on “trios” (two parents and one child) in Iceland. We’ve pre-processed the data to make it easier to work with.

Load the pre-processed data by running the code chunk below.

# read data  
dnm\_by\_age <- read.table("dnm\_by\_age\_tidy\_Halldorsson.tsv",  
 sep = "\t", header = TRUE)  
# preview data  
head(dnm\_by\_age)

## Proband\_id n\_paternal\_dnm n\_maternal\_dnm n\_na\_dnm Father\_age Mother\_age  
## 1 675 51 19 0 31 36  
## 2 1097 26 12 1 19 19  
## 3 1230 42 12 3 30 28  
## 4 1481 53 14 1 32 20  
## 5 1806 61 11 6 38 34  
## 6 2280 63 9 3 38 20

The columns in this table are:

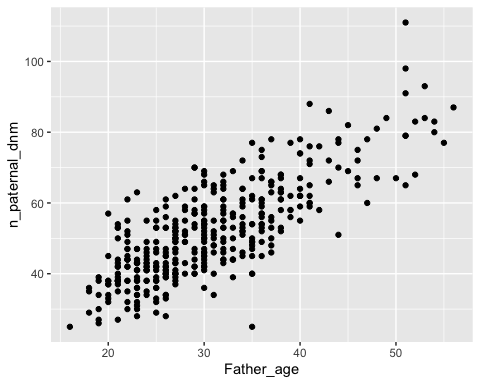
1. Proband\_id: ID of the child (i.e., “proband”)
2. n\_paternal\_dnm: Number of DNMs (carried by the child) that came from the father
3. n\_maternal\_dnm: Number of DNMs that came from the mother
4. n\_na\_dnm: Number of DNMs whose parental origin can’t be determined
5. Father\_age: Father’s age at proband’s birth
6. Mother\_age: Mother’s age at proband’s birth

## Visualizing the data

We can use our tidied data to ask questions about the *de novo* mutation rate in these Icelandic individuals. How does parental age affect the number of DNMs for males and females?

Use the dnm\_by\_age data to plot this relationship for *males*.

ggplot(data = dnm\_by\_age,  
 # specify where ggplot should be getting the x location for each data point  
 aes(x = Father\_age,  
 # specify where ggplot should be getting the y location for each data point  
 y = n\_paternal\_dnm)) +  
 # specify that the data should be plotted as points  
 geom\_point()

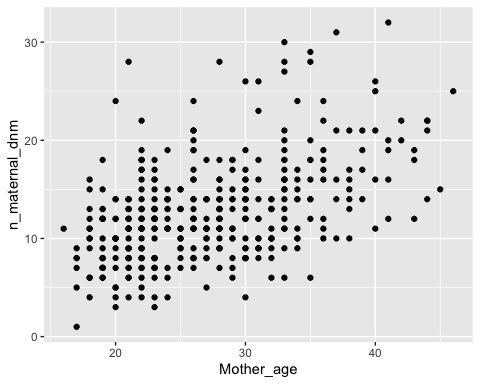


Based on your plot, would you say that there’s an association between paternal age and number of DNMs?

It looks like there’s a pretty strong association between paternal age and number of DNMs, where older males have more DNMs.

Modify your code to plot the relationship between age and number of DNMs for *females*. Does there seem to be an association between maternal age and number of DNMs?

ggplot(data = dnm\_by\_age,  
 aes(x = Mother\_age,  
 y = n\_maternal\_dnm)) +  
 geom\_point()



There’s also a strong positive association between maternal age and number of DNMs, although the slope (i.e., the increase in number of DNMs per year) is shallower.

## Linear models

We can visually observe that age seems associated with number of DNMs in both males and females, but we need a way to ask if that this is a statistically meaningful association.

We can do this with a **linear model**. This model fits a line to the plots that we just made, and asks if the slope is significantly different from 0 (i.e., if there’s a significant increase in DNM count as age increases).

If this is a statistical test, what’s the null hypothesis?

The null hypothesis for this linear model is that the slope is 0 – i.e., that there’s no association between parental age and the number of DNMs from that parent.

If the slope is significantly different from 0, we can reject the null hypothesis.

We’ll fit a linear model using R’s lm function. Run the following code block to open a manual describing the function.

?lm

lm requires two arguments:

* The formula or equation it’s evaluating
* A table of data

The formula must be in the format response variable ~ predictor variable(s), where each variable is the name of a column in our data table.

Is our predictor variable the parental age or the number of DNMs?

The predictor variable is parental age. We expect the number of DNMs to change as a *consequence* of parental age.

## Fitting a linear model for DNMs

Run the following code to fit a model for the effect of age on paternal DNMs.

# fit linear model for paternal DNMs  
fit\_pat <- lm(formula = n\_paternal\_dnm ~ Father\_age,  
 data = dnm\_by\_age)  
  
# print results of model  
summary(fit\_pat)

##   
## Call:  
## lm(formula = n\_paternal\_dnm ~ Father\_age, data = dnm\_by\_age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.785 -5.683 -0.581 5.071 31.639   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.58819 1.70402 6.214 1.34e-09 \*\*\*  
## Father\_age 1.34849 0.05359 25.161 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.426 on 388 degrees of freedom  
## Multiple R-squared: 0.62, Adjusted R-squared: 0.619   
## F-statistic: 633.1 on 1 and 388 DF, p-value: < 2.2e-16

How do you interpret results from a linear model?

For our purposes, the only part of the results you need to look at is the line under (Intercept) in the Coefficients section:

Estimate Std. Error t value Pr(>|t|)  
Father\_age 1.34849 0.05359 25.161 < 2e-16 \*\*\*

* The fourth columm, Pr(>|t|), is the **p-value**.

Because this p-value is < 2e-16, we can reject the null hypothesis and say that there is association between paternal age and the number of paternal DNMs.

* The first column, Estimate, is the **slope**, or **coefficient**.

Linear regression fits a line to our plot of paternal age vs. number of DNMs. The coefficient estimate is the **slope** of that line.

The slope for paternal age given by this linear model is 1.34849. We can interpret this number this way: **For every additional year of paternal age, we expect 1.35 additional paternal DNMs in the child.**

Modify your code to assess the relationship between *maternal* age and number of *maternal* DNMs. Is this relationship significant? How many maternal DNMs do we expect for every additional year of maternal age?

# fit linear model for maternal DNMs  
fit\_mat <- lm(formula = n\_maternal\_dnm ~ Mother\_age,  
 data = dnm\_by\_age)  
  
# print results of model  
summary(fit\_mat)

##   
## Call:  
## lm(formula = n\_maternal\_dnm ~ Mother\_age, data = dnm\_by\_age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.8683 -3.1044 -0.2329 2.2394 17.5379   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.51442 0.98193 2.561 0.0108 \*   
## Mother\_age 0.37846 0.03509 10.785 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.503 on 388 degrees of freedom  
## Multiple R-squared: 0.2307, Adjusted R-squared: 0.2287   
## F-statistic: 116.3 on 1 and 388 DF, p-value: < 2.2e-16

The p-value is <2e-16 and the Mother\_age slope is 0.37846.

This relationship is significant, and we expect 0.38 more maternal DNMs for every additional year of maternal age.

## Confidence intervals

Our models predict that there are 1.35 more DNMs for additional every year of paternal age, and 0.38 more DNMs for every additional year of maternal age. Does this mean that sperm and oocytes accumulate DNMs at different rates?

The maternal and paternal slopes look different, but we need statistical evidence that they actually are. (For example, what if there’s a lot of variability in the maternal DNM data, and the true maternal coefficient could be anywhere between -1 and 10?)

To do this, we compare the **confidence intervals** of our slope estimates.

What is a confidence interval?

We use confidence intervals when estimating a value – in this case, the Mother\_age and Father\_age slope parameters.

A **confidence interval (CI)** is a random interval that has a 95% probability of falling on the parameter we are estimating. So, a 95% CI contains the true value of the slope 95% of the time.

Keep in mind that the definition above (95% of random intervals fall on the true value) is not the same as saying there is a 95% chance that the true value falls within our interval. This latter statement is not accurate.

In R, we get the confidence interval of a parameter from a linear model with the confint function.

?confint

confint requires three arguments:

* A fitted linear model (our fit\_pat variable)
* The parameter we want a CI for (Father\_age)
* The CI’s probability (typically 95%)

## Calculate 95% CIs

Run the following code to calculate the 95% confidence interval for the Father\_age slope parameter.

confint(fit\_pat, 'Father\_age', level = 0.95)

## 2.5 % 97.5 %  
## Father\_age 1.243118 1.45386

So, 95% of the time, the number of additional DNMs per year of paternal age is between 1.24 and 1.45.

Modify your code to get the 95% CI for the *Mother\_age* slope. What’s the interpretation of this confidence interval?

confint(fit\_mat, 'Mother\_age', level = 0.95)

## 2.5 % 97.5 %  
## Mother\_age 0.3094713 0.4474528

95% of the time, the number of additional DNMs per year of maternal age is between 0.31 and 0.45.

Now that we have the confidence intervals for both slope parameters, we can finally compare them.

Our two CI ranges are non-overlapping. The paternal range is [1.24, 1.45] and the maternal range is [0.31, 0.45].

If the 95% CIs for two parameters *don’t* overlap, this strongly supports that the parameters are significantly different from one another. **So, it seems likely that paternal and maternal gametes experience different rates of *de novo* mutation.**

If the CIs for two parameters overlap, are they not significantly different?

Not necessarily. More analysis, like a hypothesis test, is needed to make a final decision.

## Conclusion

In this lab, we explored the relationship between parental age and the number of *de novo* mutations in their gametes.

* We **plotted** the relationship between maternal/paternal age and DNM count. This visualization suggested that DNM count increases with age for both groups.
* We confirmed this hypothesis by using a **linear model**, which tests if additional years of age have a non-zero effect on the number of DNMs.
* The number of paternal DNMs seemed to increase more quickly with age than maternal DNMs. We confirmed this by comparing the **95% confidence intervals** of the slopes of the two models.

One final question – let’s assume that there really is a difference between the effect of age on DNMs in male and female gametes. What biological reasons might be causing this difference?

## Homework

So far, we’ve only looked at the *de novo* mutation data from [the Halldorsson et al. paper](https://science.sciencemag.org/content/363/6425/eaau1043). Now we’ll use their data on the number of maternal and paternal origin crossovers (i.e., how many crossovers occurred across all chromosomes in the maternal and paternal gametes).

#### Learning Objectives

* Practice visualizing data with ggplot2
* Interpret p-values and effect sizes from linear models

## Required homework

The data from the paper has been pre-filtered for you. Run this code block to read it in:

# read data  
crossovers <- read.table("crossovers.tsv", header = TRUE)  
  
# preview data  
head(crossovers)

## Proband\_id n\_pat\_xover n\_mat\_xover Father\_age Mother\_age  
## 1 3 22 51 29 28  
## 2 10 26 50 26 26  
## 3 11 25 38 25 22  
## 4 15 24 50 31 26  
## 5 20 27 35 26 24  
## 6 22 28 40 39 31

The columns in this table are:

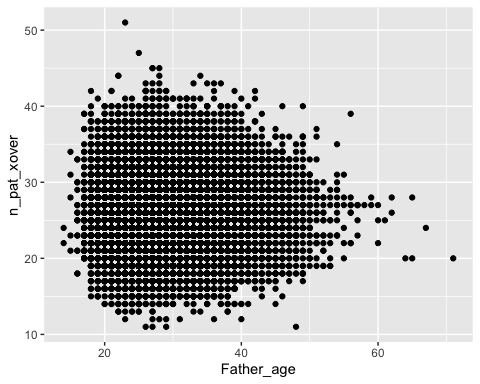
1. Proband\_id: ID of the child
2. n\_pat\_xover: Number of crossovers (carried by the child) that occurred in the paternal gametes
3. n\_mat\_xover: Number of crossovers that occurred in the maternal gametes
4. Father\_age: Father’s age at proband’s birth
5. Mother\_age: Mother’s age at proband’s birth

**Assignment:** Using the ggplot code from this module, plot the relationship between parental age and number of crossovers. As with the DNM data, make one plot for the maternal crossovers and one plot for the paternal. Do you think parental age impacts crossover number?

Solution

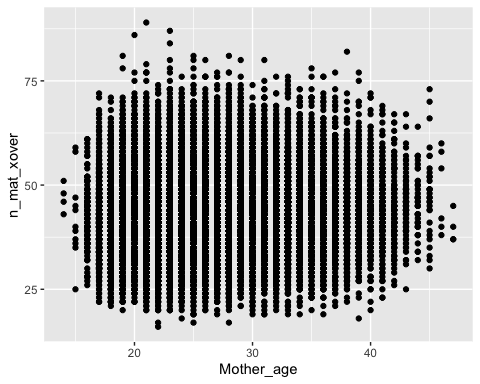
Plot paternal crossovers:

ggplot(data = crossovers,  
 # x axis is paternal age  
 aes(x = Father\_age,  
 # y axis is number of crossovers  
 y = n\_pat\_xover)) +  
 geom\_point()



Plot maternal crossovers:

ggplot(data = crossovers,  
 # x axis is maternal age  
 aes(x = Mother\_age,  
 # y axis is number of crossovers  
 y = n\_mat\_xover)) +  
 geom\_point()



Just by eye, it doesn’t really seem that age affects number of crossovers for either mothers or fathers.

## Optional homework

**Assignment:** Fit *two* linear models (one paternal, one maternal) to ask if there is an association between the number of parental crossovers and parental age. If there is an association, how is the number of crossovers predicted to change with every year of maternal/paternal age?

Solution

# fit the model with paternal age  
fit\_pat <- lm(data = crossovers,  
 formula = n\_pat\_xover ~ Father\_age)  
summary(fit\_pat)

##   
## Call:  
## lm(formula = n\_pat\_xover ~ Father\_age, data = crossovers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.2173 -3.1880 -0.1997 2.8061 24.7652   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 26.369432 0.102736 256.67 <2e-16 \*\*\*  
## Father\_age -0.005852 0.003462 -1.69 0.091 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.388 on 41090 degrees of freedom  
## Multiple R-squared: 6.953e-05, Adjusted R-squared: 4.519e-05   
## F-statistic: 2.857 on 1 and 41090 DF, p-value: 0.09098

There isn’t a significant association between paternal age and the number of paternal crossovers (p = 0.091).

# fit the model with maternal age  
fit\_mat <- lm(data = crossovers,  
 formula = n\_mat\_xover ~ Mother\_age)  
summary(fit\_mat)

##   
## Call:  
## lm(formula = n\_mat\_xover ~ Mother\_age, data = crossovers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.161 -6.095 -0.425 5.641 45.905   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 41.709271 0.206238 202.24 <2e-16 \*\*\*  
## Mother\_age 0.065989 0.007576 8.71 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.685 on 41090 degrees of freedom  
## Multiple R-squared: 0.001843, Adjusted R-squared: 0.001819   
## F-statistic: 75.87 on 1 and 41090 DF, p-value: < 2.2e-16

Surprisingly, there *is* a significant association between maternal age and the number of maternal crossovers (p < 2e-16). For every year of maternal age, we expect the child to carry 0.07 additional maternal origin crossovers.

Although the maternal crossovers plot doesn’t look that impressive, our estimated slope is 0.07, which is probably too small to distinguish visually.