

Exploratory Data Analysis

2023-05-29

```
# Load in packages and data
library(ggplot2)
library(data.table)
library(plyr)
library(dplyr)
library(MASS)
library(RColorBrewer)
library(Hmisc)
library(ggpubr)
library(ks)
library(tidyverse)
load("Rakai_Pangea2_RCCS_pairs_household_202211XX.RData")
```

Inspect structure of the data set

```
# Determine structure of data
str(pairs_tsi)
```

```
## Classes 'data.table' and 'data.frame':  610 obs. of  19 variables:
## $ RECIPIENT      : chr  "AID0011" "AID0023" "AID0102" "AID0110" ...
## $ SOURCE         : chr  "AID1640" "AID0299" "AID7619" "AID1535" ...
## $ SEX.SOURCE     : chr  "M" "M" "M" "M" ...
## $ SEX.RECIPIENT  : chr  "F" "F" "F" "F" ...
## $ CL             : Date, format: "2009-08-09" "2008-12-06" ...
## $ IL             : Date, format: "2010-03-07" "2009-12-09" ...
## $ M              : Date, format: "2011-01-18" "2011-06-09" ...
## $ IU             : Date, format: "2011-10-30" "2012-10-07" ...
## $ CU             : Date, format: "2012-04-07" "2013-07-03" ...
## $ AGE_TRANSMISSION.SOURCE : num  23.1 37.6 29.6 25.1 26.6 23.9 24.2 27 26.8 19.7 ...
## $ AGE_INFECTION.RECIPIENT : num  16.9 18.2 21 28.1 35 17.8 18.9 18.7 31.3 25.6 ...
## $ ROUND.M        : chr  "R014" "R014" "R012" "R012" ...
## $ COMM.SOURCE     : chr  "fishing" "fishing" "fishing" "fishing" ...
## $ COMM.RECIPIENT  : chr  "fishing" "fishing" "fishing" "fishing" ...
## $ COMM_NUM.SOURCE : int   770 771 771 38 23 40 771 602 34 772 ...
## $ COMM_NUM.RECIPIENT : int   770 771 771 38 23 40 771 38 34 772 ...
## $ DATE_COLLECTION.SOURCE : IDate, format: "2012-12-12" "2014-04-28" ...
## $ DATE_COLLECTION.RECIPIENT: IDate, format: "2012-12-12" "2014-04-25" ...
## $ same_hh        : num   1 1 1 0 1 1 1 0 1 1 ...
## - attr(*, ".internal.selfref")=<externalptr>
## - attr(*, "sorted")= chr "RECIPIENT"
```

Description of variables:

- RECIPIENT, SOURCE - IDs for participants involved in case of HIV transmission. The recipients are all unique but sources are not.
- SEX.SOURCE, SEX.RECIPIENT - the sex of source/recipient. M for male and F for female.
- CL, IL, M, IU, CU - *UNKNOWN dates, to be completed*
- AGE_TRANSMISSION.SOURCE - Age of source of transmission.
- AGE_INFECTION.RECIPIENT - Age of recipient of transmission.
- ROUND.M - Round of survey *To be confirmed - this is the first round in which participation was recorded?*
- COMM.SOURCE, COMM.RECIPIENT - Community of source/recipient. Either 'fishing' or 'inland'.
- COMM_NUM.SOURCE, COMM_NUM.RECIPIENT - *UNKNOWN*
- DATE.COLLECTION.SOURCE, DATE.COLLECTION.RECIPIENT - *UNKNOWN*
- same_hh - whether the transmission was within the same household.

Transform data

Bin the data according to age of recipient. Divide into two data sets by household factor.

```
# Divide in 5 years age bands
age_limits <- c(15, 20, 25, 30, 35, 40, 45, 50)
age_labels <- paste0(age_limits, '-', age_limits + 5)
pairs_tsi$AGE_BAND.RECIPIENT <- cut(
  pairs_tsi$AGE_INFECTION.RECIPIENT,
  breaks = age_limits,
  labels = age_labels[1:(length(age_limits) - 1)],
  include.lowest = TRUE)

## HL: Not sure why this is here, such observations do not exist
#Exclude respondents in 'neuro' community
#pairs_tsi <- pairs_tsi[COMM.SOURCE!='neuro' & COMM.RECIPIENT!='neuro',]

# Divide in 2 datasets for transmissions within and out of HH
pairs_tsi_same_hh <- pairs_tsi[same_hh == 1,]
pairs_tsi_diff_hh <- pairs_tsi[same_hh == 0,]
```

Exploratory plots

BY RECIPIENT COMMUNITY AND HOUSEHOLD FACTOR

```
# AB: it's clearer to perform operations here.
# less prone to bugs, more readable, and can focus on graphics only in ggplot snippet
p1_data <- pairs_tsi[, {
  N.in.inland <- sum(COMM.RECIPIENT == 'inland')
  binom.conf <- (.N*binconf(N.in.inland, .N)) |> as.list()
  names(binom.conf) <- c('BC.center', 'BC.min', 'BC.max')
  binom.conf
}, by=same_hh]

# Plot number of counts by comm & hh factor
p1_new <- ggplot(pairs_tsi, aes(x = as.logical(same_hh),
                                fill = COMM.RECIPIENT)) +
  geom_bar() +
```

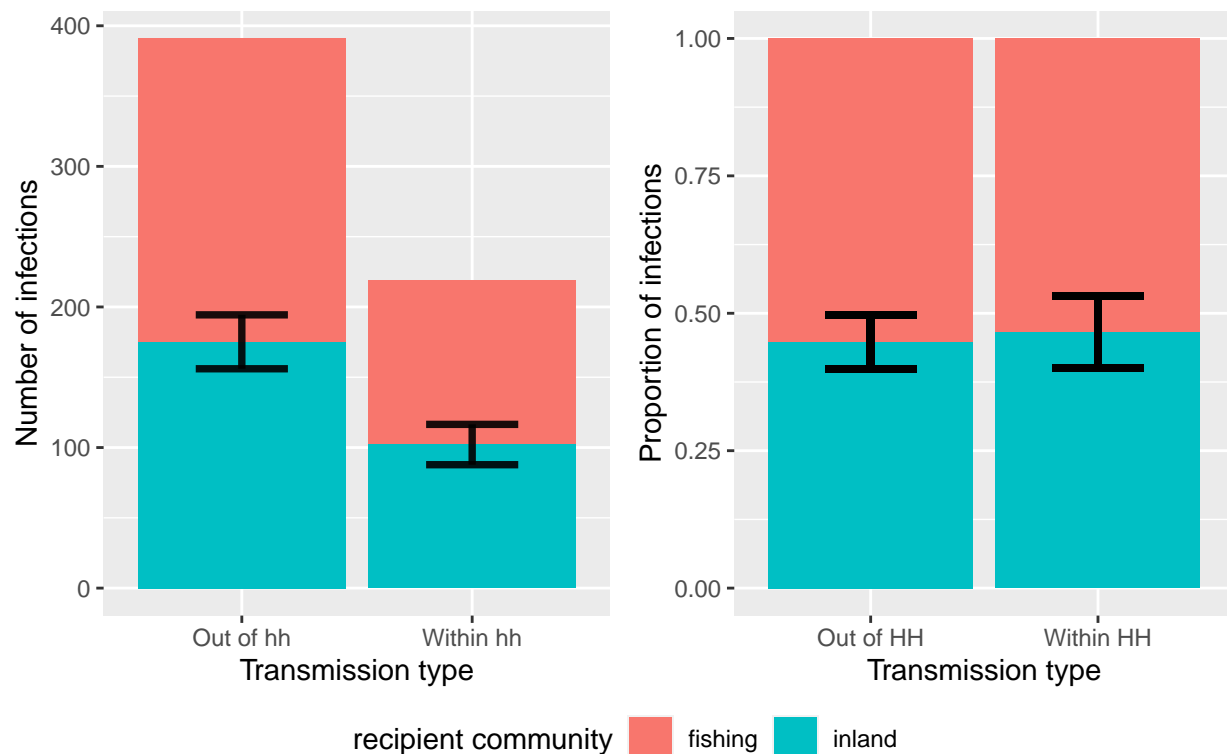
```
geom_errorbar(data = p1_data, aes(ymin = BC.min, ymax = BC.max, fill = NA),
              width = 0.4, colour = "black", alpha = 0.9, linewidth = 1.3) +
labs(x='Transmission type',
     y='Number of infections',
     fill='recipient community') +
scale_x_discrete(labels=c("Out of hh", "Within hh"))
```

```
## Warning in geom_errorbar(data = p1_data, aes(ymin = BC.min, ymax = BC.max, :
## Ignoring unknown aesthetics: fill
```

```
# Plot proportions of infections by household factor
p2<-ggplot(data = pairs_tsi, aes(x = as.logical(same_hh), fill = COMM.RECIPIENT)) +
  geom_bar(position="fill") +
  labs(x='Transmission type',
       y='Proportion of infections',
       fill='Recipient community') +
  scale_x_discrete(labels=c("Out of HH", "Within HH")) +
  geom_errorbar( aes(x='FALSE', ymin=binconf(nrow(pairs_tsi_diff_hh[COMM.RECIPIENT=='inland',]),nrow(pai
  geom_errorbar( aes(x='TRUE', ymin=binconf(nrow(pairs_tsi_same_hh[COMM.RECIPIENT=='inland',]),nrow(pai

# Combine above plots into one
p <- ggarrange(p1_new, p2, ncol=2, nrow=1, common.legend = TRUE,
               legend="bottom")
p2 <- annotate_figure(p, top="Colored by recipient community")
annotate_figure(p2, top="Infections within and out of household")
```

Infections within and out of household
Colored by recipient community



```
# Provide summary tables of counts by comm & hh
table(as.logical(pairs_tsi$same_hh),pairs_tsi$COMM.RECIPIENT)
```

```
##
##      fishing inland
## FALSE      216    175
##  TRUE      117    102
```

```
# 95% confidence intervals for number of infections
cat("Fishing, Outside HH:\n")
```

```
## Fishing, Outside HH:
```

```
binconf(216,610)*610
```

```
## PointEst    Lower    Upper
##      216 193.4724 239.6415
```

```
cat("Inland, Outside HH:\n")
```

```
## Inland, Outside HH:
```

```
binconf(175,610)*610
```

```
## PointEst    Lower    Upper
##      175 153.9719 197.6552
```

```
cat("Fishing, Within HH:\n")
```

```
## Fishing, Within HH:
```

```
binconf(117,610)*610
```

```
## PointEst    Lower    Upper
##      117 99.14088 137.2122
```

```
cat("Inland, Within HH:\n")
```

```
## Inland, Within HH:
```

```
binconf(102,610)*610
```

```
## PointEst    Lower    Upper
##      102 85.21819 121.3226
```

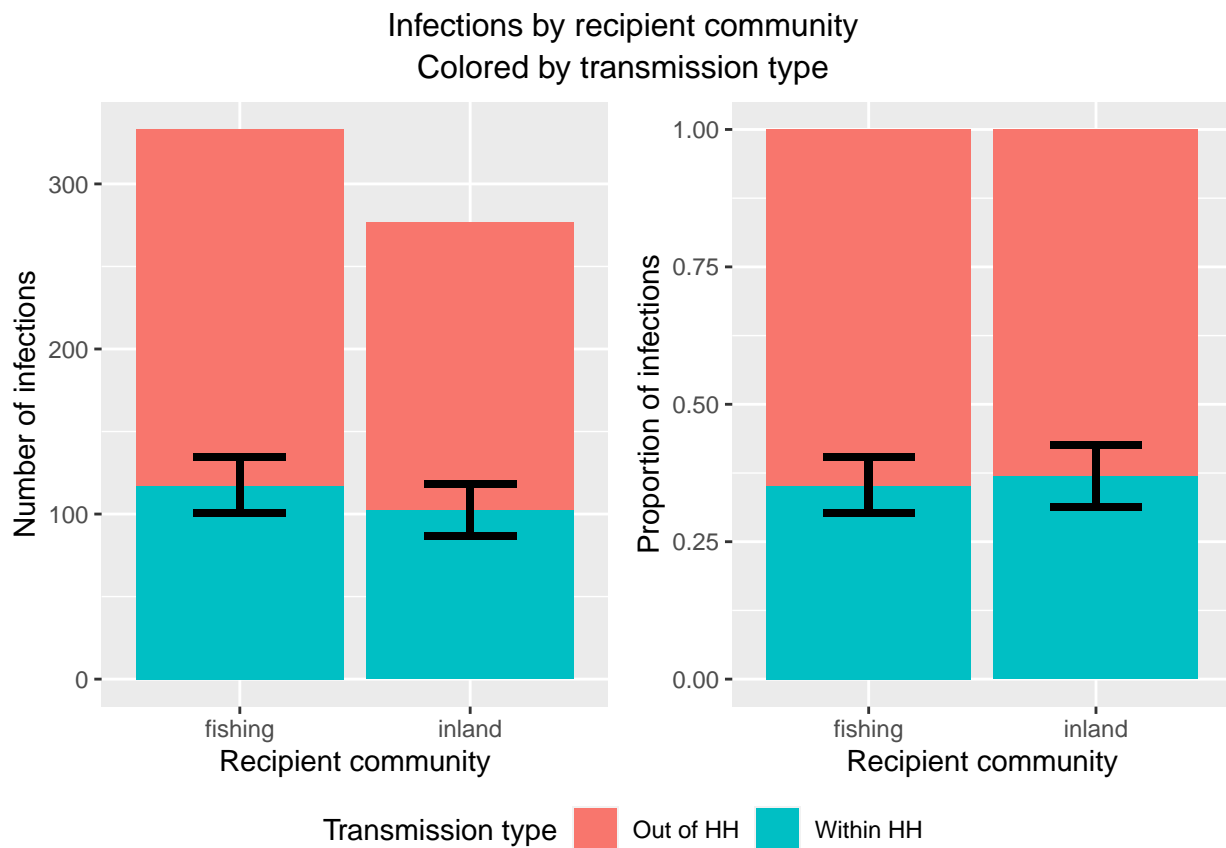
```

# Same as above but flipped hh and comm (x axis vs fill colour)
p3<-ggplot(data = pairs_tsi, aes(x = COMM.RECIPIENT,fill = as.logical(same_hh))) +
  geom_bar()+
  labs(x='Recipient community',
       y='Number of infections',
       fill='Transmission type') +
  scale_fill_discrete(labels=c("Out of HH","Within HH")) +
  geom_errorbar( aes(x='fishing', ymin=binconf(nrow(pairs_tsi_same_hh[COMM.RECIPIENT=='fishing',]),nrow(p
  geom_errorbar( aes(x='inland', ymin=binconf(nrow(pairs_tsi_same_hh[COMM.RECIPIENT=='inland',]),nrow(p

p4<-ggplot(data = pairs_tsi, aes(x = COMM.RECIPIENT,fill = as.logical(same_hh))) +
  geom_bar(position='fill')+
  labs(x='Recipient community',
       y='Proportion of infections',
       fill='Transmission type') +
  scale_fill_discrete(labels=c("Out of HH","Within HH")) +
  geom_errorbar( aes(x='fishing', ymin=binconf(nrow(pairs_tsi_same_hh[COMM.RECIPIENT=='fishing',]),nrow(p
  geom_errorbar( aes(x='inland', ymin=binconf(nrow(pairs_tsi_same_hh[COMM.RECIPIENT=='inland',]),nrow(p

p <- ggarrange(p3, p4, ncol=2, nrow=1, common.legend = TRUE, legend="bottom")
p2<-annotate_figure(p, top="Colored by transmission type")
annotate_figure(p2, top="Infections by recipient community")

```

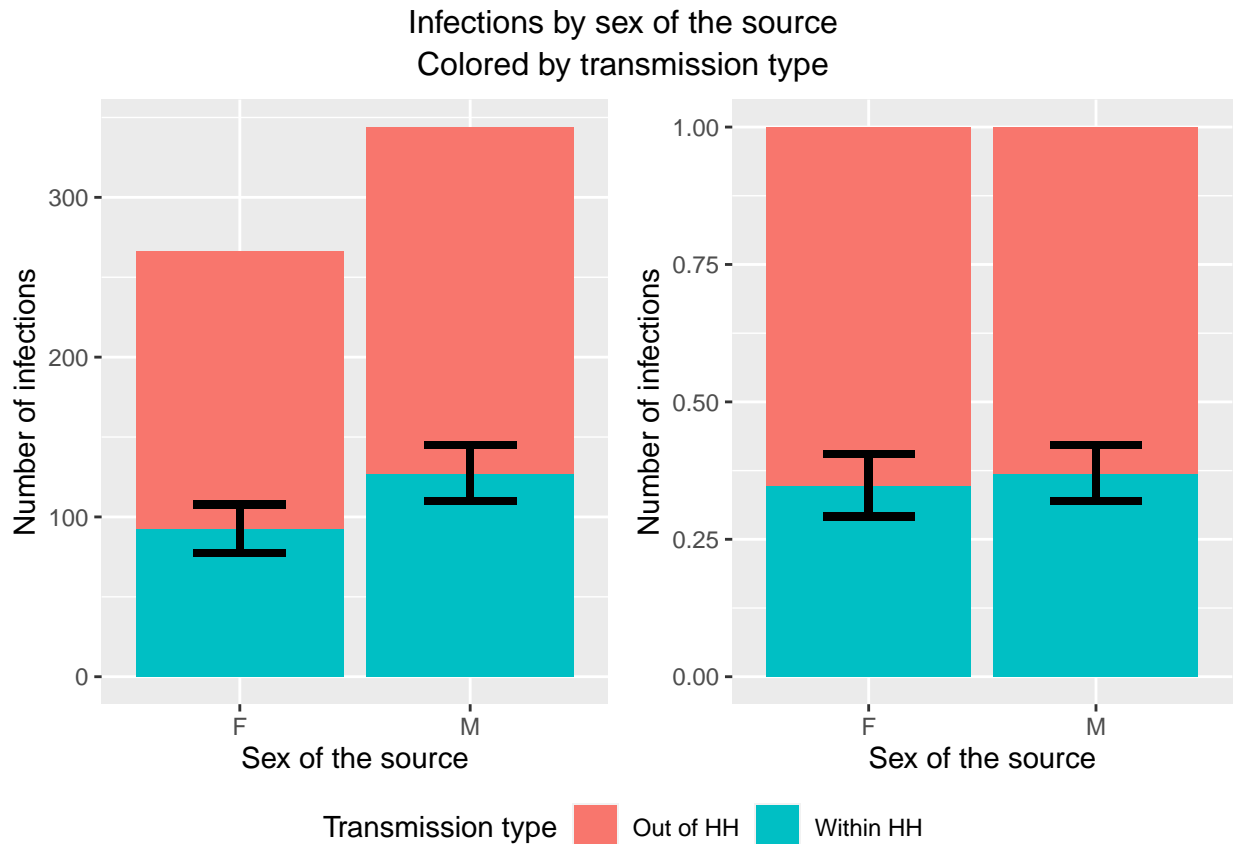


BY SEX AND HOUSEHOLD FACTOR

```
p1<-ggplot(data = pairs_tsi, aes(x = SEX.SOURCE,fill = as.logical(same_hh))) +
  geom_bar()+
  labs(x='Sex of the source',
       y='Number of infections',
       fill='Transmission type') +
  scale_fill_discrete(labels=c("Out of HH","Within HH")) +
  geom_errorbar( aes(x='F', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='F',]),nrow(pairs_tsi[SEX.SOURCE=='F',])),
                    aes(x='M', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='M',]),nrow(pairs_tsi[SEX.SOURCE=='M',]))

p2<-ggplot(data = pairs_tsi, aes(x = SEX.SOURCE,fill = as.logical(same_hh))) +
  geom_bar(position = "fill")+
  labs(x='Sex of the source',
       y='Number of infections',
       fill='Transmission type') +
  scale_fill_discrete(labels=c("Out of HH","Within HH")) +
  geom_errorbar( aes(x='F', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='F',]),nrow(pairs_tsi[SEX.SOURCE=='F',])),
                    aes(x='M', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='M',]),nrow(pairs_tsi[SEX.SOURCE=='M',]))

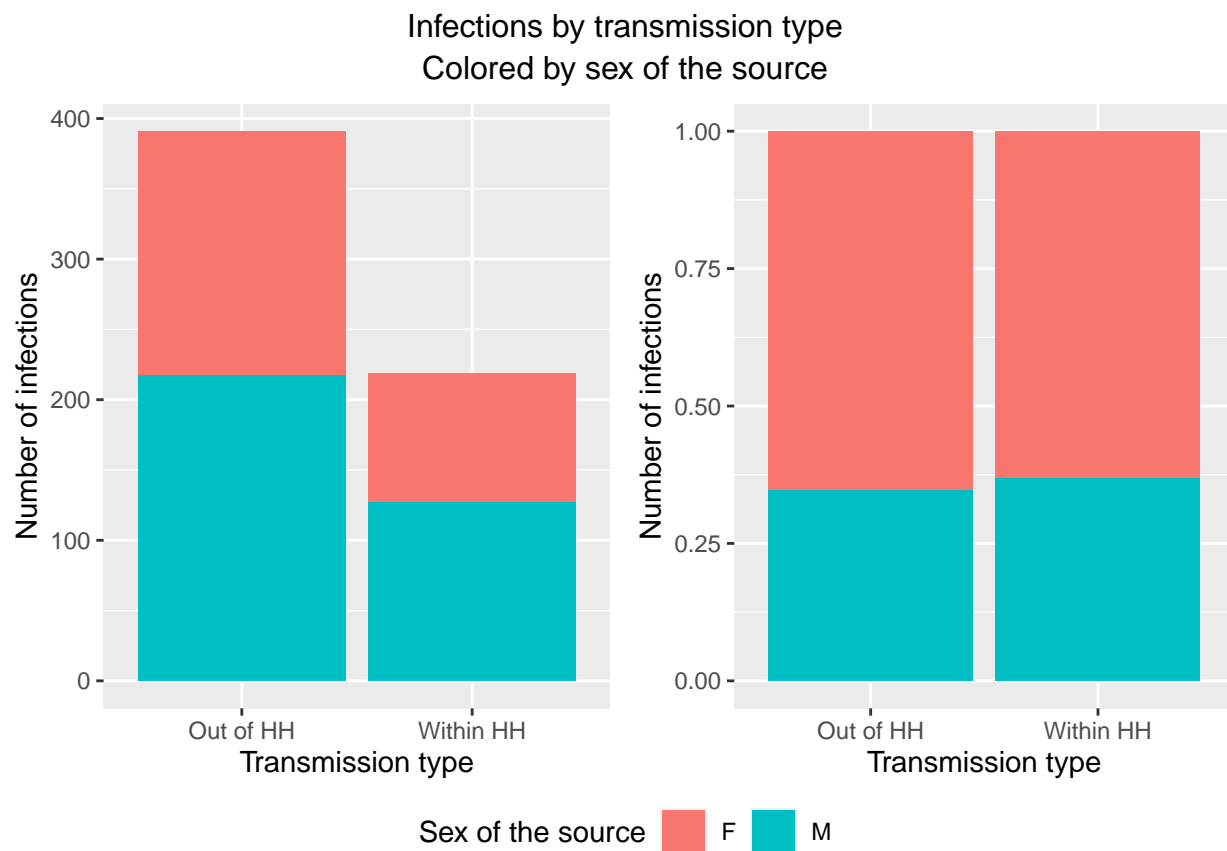
p<-ggarrange(p1, p2, ncol=2, nrow=1, common.legend = TRUE, legend="bottom")
p2<-annotate_figure(p, top="Colored by transmission type")
annotate_figure(p2, top="Infections by sex of the source")
```



```
#### Other way around:
p1<-ggplot(data = pairs_tsi, aes(x = as.logical(same_hh),fill = SEX.SOURCE)) +
  geom_bar()+
  labs(x='Transmission type',
       y='Number of infections',
       fill='Sex of the source') +
  scale_x_discrete(labels=c("Out of HH","Within HH"))
#geom_errorbar( aes(x='F', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='F',])),nrow(pairs_tsi[SEX.SOURCE=='F',])))
#geom_errorbar( aes(x='M', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='M',])),nrow(pairs_tsi[SEX.SOURCE=='M',])))

p2<-ggplot(data = pairs_tsi, aes(x = SEX.SOURCE,fill = as.logical(same_hh))) +
  geom_bar(position="fill") +
  labs(x='Transmission type',
       y='Number of infections',
       fill='Sex of the source') +
  scale_x_discrete(labels=c("Out of HH","Within HH"))
#geom_errorbar( aes(x='F', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='F',])),nrow(pairs_tsi[SEX.SOURCE=='F',])))
#geom_errorbar( aes(x='M', ymin=binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='M',])),nrow(pairs_tsi[SEX.SOURCE=='M',])))

p<-ggarrange(p1, p2, ncol=2, nrow=1, common.legend = TRUE, legend="bottom")
p2<-annotate_figure(p, top="Colored by sex of the source")
annotate_figure(p2, top="Infections by transmission type")
```



BY AGE BAND

```
p1<-ggplot(data = pairs_tsi[is.na(AGE_BAND.RECIPIENT)==FALSE & SEX.RECIPIENT=='F',], aes(fill = as.logi
geom_bar()+
labs(fill='Transmission type',
      y='Number of infections',
      x='Recipient community')+
facet_grid(~ AGE_BAND.RECIPIENT, scales="free_y") +
scale_fill_discrete(labels=c("Out of HH", "Within HH")) +
scale_x_discrete(labels=c('F', 'I')) +
theme(strip.text = element_blank()) +
theme(legend.position='bottom')

p2<-ggplot(data = pairs_tsi[is.na(AGE_BAND.RECIPIENT)==FALSE & SEX.RECIPIENT=='M',], aes(fill = as.logi
geom_bar()+
labs(fill='Transmission type',
      y='Number of infections',
      x='Recipient community')+
facet_grid(~ AGE_BAND.RECIPIENT, scales="free_y") +
scale_x_discrete(labels=c('F', 'I')) +
scale_fill_discrete(labels=c("Out of HH", "Within HH")) +
theme(legend.position='bottom')

p<-ggarrange(p1, p2, ncol=1, nrow=2, common.legend = TRUE, legend="right")
p2<-annotate_figure(p, top="Grouped by transmission type and recipient age at infection")
annotate_figure(p2, top="Infections by recipient community")
```


Figure 2 consists of two stacked bar charts. The top chart shows the number of infections by recipient community (F for female, I for male) across seven age groups (15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50). The bottom chart shows the number of infections by recipient community (F for female, I for male) across seven age groups (15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50). The legend indicates 'Out of HH' (red) and 'Within HH' (teal).

Top Chart: Number of infections by age group and recipient community

Age group	Recipient community	Within HH	Out of HH
15-20	F	21	35
	I	15	25
20-25	F	14	36
	I	23	27
25-30	F	19	27
	I	13	25
30-35	F	10	6
	I	6	21
35-40	F	2	4
	I	4	5
40-45	F	0	3
	I	0	2
45-50	F	0	0
	I	1	0

Bottom Chart: Number of infections by age group and recipient community

Age group	Recipient community	Within HH	Out of HH
15-20	F	3	9
	I	2	5
20-25	F	10	30
	I	3	14
25-30	F	14	26
	I	13	20
30-35	F	15	24
	I	11	20
35-40	F	7	12
	I	6	8
40-45	F	2	4
	I	2	4
45-50	F	0	0
	I	2	0

```
tmp<-pairs_tsi
tmp[is.na(ROUND.M)==TRUE,]$M
```

```
tmp[is.na(ROUND.M)==TRUE & year(M)<=2003,]$ROUND.M<-'before R10'
tmp[is.na(ROUND.M)==TRUE & year(M)>=2018,]$ROUND.M<-'after R18'
pairs_tsi<-tmp
pairs_tsi$ROUND.M<-factor(pairs_tsi$ROUND.M,levels=c('before R10','R010','R011','R012','R013','R014','R015','R016','R017','R018','R019','R020','R021','R022','R023','R024','R025','R026','R027','R028','R029','R030','R031','R032','R033','R034','R035','R036','R037','R038','R039','R040','R041','R042','R043','R044','R045','R046','R047','R048','R049','R050','R051','R052','R053','R054','R055','R056','R057','R058','R059','R060','R061','R062','R063','R064','R065','R066','R067','R068','R069','R070','R071','R072','R073','R074','R075','R076','R077','R078','R079','R080','R081','R082','R083','R084','R085','R086','R087','R088','R089','R090','R091','R092','R093','R094','R095','R096','R097','R098','R099','R100','R101','R102','R103','R104','R105','R106','R107','R108','R109','R110','R111','R112','R113','R114','R115','R116','R117','R118','R119','R120','R121','R122','R123','R124','R125','R126','R127','R128','R129','R130','R131','R132','R133','R134','R135','R136','R137','R138','R139','R140','R141','R142','R143','R144','R145','R146','R147','R148','R149','R150','R151','R152','R153','R154','R155','R156','R157','R158','R159','R160','R161','R162','R163','R164','R165','R166','R167','R168','R169','R170','R171','R172','R173','R174','R175','R176','R177','R178','R179','R180','R181','R182','R183','R184','R185','R186','R187','R188','R189','R190','R191','R192','R193','R194','R195','R196','R197','R198','R199','R200','R201','R202','R203','R204','R205','R206','R207','R208','R209','R210','R211','R212','R213','R214','R215','R216','R217','R218','R219','R220','R221','R222','R223','R224','R225','R226','R227','R228','R229','R230','R231','R232','R233','R234','R235','R236','R237','R238','R239','R240','R241','R242','R243','R244','R245','R246','R247','R248','R249','R250','R251','R252','R253','R254','R255','R256','R257','R258','R259','R260','R261','R262','R263','R264','R265','R266','R267','R268','R269','R270','R271','R272','R273','R274','R275','R276','R277','R278','R279','R280','R281','R282','R283','R284','R285','R286','R287','R288','R289','R290','R291','R292','R293','R294','R295','R296','R297','R298','R299','R300','R301','R302','R303','R304','R305','R306','R307','R308','R309','R310','R311','R312','R313','R314','R315','R316','R317','R318','R319','R320','R321','R322','R323','R324','R325','R326','R327','R328','R329','R330','R331','R332','R333','R334','R335','R336','R337','R338','R339','R340','R341','R342','R343','R344','R345','R346','R347','R348','R349','R350','R351','R352','R353','R354','R355','R356','R357','R358','R359','R360','R361','R362','R363','R364','R365','R366','R367','R368','R369','R370','R371','R372','R373','R374','R375','R376','R377','R378','R379','R380','R381','R382','R383','R384','R385','R386','R387','R388','R389','R390','R391','R392','R393','R394','R395','R396','R397','R398','R399','R400','R401','R402','R403','R404','R405','R406','R407','R408','R409','R410','R411','R412','R413','R414','R415','R416','R417','R418','R419','R420','R421','R422','R423','R424','R425','R426','R427','R428','R429','R430','R431','R432','R433','R434','R435','R436','R437','R438','R439','R440','R441','R442','R443','R444','R445','R446','R447','R448','R449','R450','R451','R452','R453','R454','R455','R456','R457','R458','R459','R460','R461','R462','R463','R464','R465','R466','R467','R468','R469','R470','R471','R472','R473','R474','R475','R476','R477','R478','R479','R480','R481','R482','R483','R484','R485','R486','R487','R488','R489','R490','R491','R492','R493','R494','R495','R496','R497','R498','R499','R500','R501','R502','R503','R504','R505','R506','R507','R508','R509','R510','R511','R512','R513','R514','R515','R516','R517','R518','R519','R520','R521','R522','R523','R524','R525','R526','R527','R528','R529','R530','R531','R532','R533','R534','R535','R536','R537','R538','R539','R540','R541','R542','R543','R544','R545','R546','R547','R548','R549','R550','R551','R552','R553','R554','R555','R556','R557','R558','R559','R560','R561','R562','R563','R564','R565','R566','R567','R568','R569','R570','R571','R572','R573','R574','R575','R576','R577','R578','R579','R580','R581','R582','R583','R584','R585','R586','R587','R588','R589','R590','R591','R592','R593','R594','R595','R596','R597','R598','R599','R600','R601','R602','R603','R604','R605','R606','R607','R608','R609','R610','R611','R612','R613','R614','R615','R616','R617','R618','R619','R620','R621','R622','R623','R624','R625','R626','R627','R628','R629','R630','R631','R632','R633','R634','R635','R636','R637','R638','R639','R640','R641','R642','R643','R644','R645','R646','R647','R648','R649','R650','R651','R652','R653','R654','R655','R656','R657','R658','R659','R660','R661','R662','R663','R664','R665','R666','R667','R668','R669','R670','R671','R672','R673','R674','R675','R676','R677','R678','R679','R680','R681','R682','R683','R684','R685','R686','R687','R688','R689','R690','R691','R692','R693','R694','R695','R696','R697','R698','R699','R700','R701','R702','R703','R704','R705','R706','R707','R708','R709','R710','R711','R712','R713','R714','R715','R716','R717','R718','R719','R720','R721','R722','R723','R724','R725','R726','R727','R728','R729','R730','R731','R732','R733','R734','R735','R736','R737','R738','R739','R740','R741','R742','R743','R744','R745','R746','R747','R748','R749','R750','R751','R752','R753','R754','R755','R756','R757','R758','R759','R760','R761','R762','R763','R764','R765','R766','R767','R768','R769','R770','R771','R772','R773','R774','R775','R776','R777','R778','R779','R780','R781','R782','R783','R784','R785','R786','R787','R788','R789','R790','R791','R792','R793','R794','R795','R796','R797','R798','R799','R800','R801','R802','R803','R804','R805','R806','R807','
```

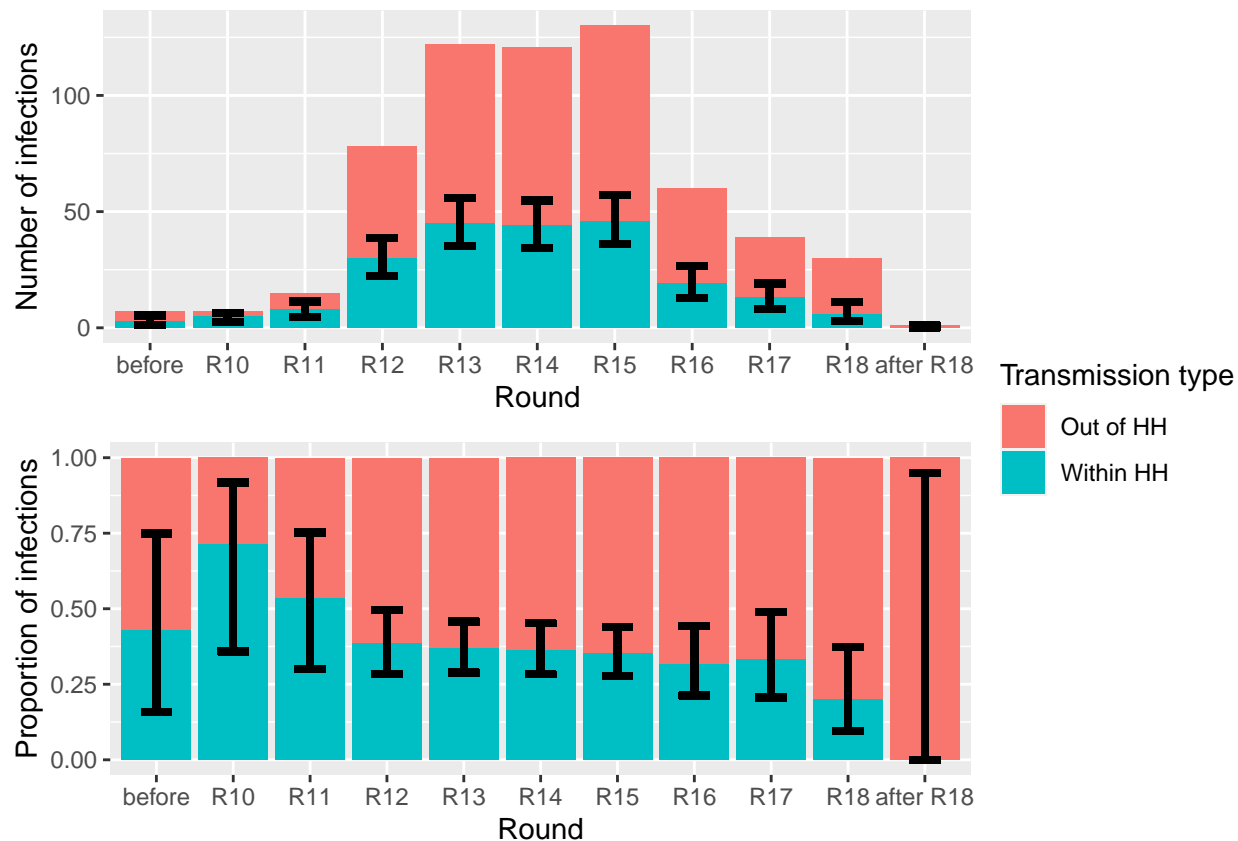
```

geom_errorbar( aes(x='R012', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R012',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='R013', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R013',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='R014', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R014',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='R015', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R015',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='R016', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R016',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='R017', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R017',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='R018', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R018',])),nrow(pairs_tsi[RO
geom_errorbar( aes(x='after R18', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='after R18',])),nrow(pa
scale_fill_discrete(labels=c("Out of HH","Within HH")) +
scale_x_discrete(labels=c("before","R10","R11","R12","R13","R14","R15","R16","R17","R18",'after R18'))

p1<-ggplot(data = pairs_tsi, aes(x = ROUND.M,fill = as.logical(same_hh))) +
  geom_bar()+
  labs(x='Round',
       y='Number of infections',
       fill='Transmission type')+
  geom_errorbar( aes(x='before R10', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='before R10',])),nrow(p
  geom_errorbar( aes(x='R010', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R010',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R011', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R011',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R012', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R012',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R013', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R013',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R014', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R014',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R015', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R015',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R016', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R016',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R017', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R017',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='R018', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='R018',])),nrow(pairs_tsi[RO
  geom_errorbar( aes(x='after R18', ymin=binconf(nrow(pairs_tsi_same_hh[ROUND.M=='after R18',])),nrow(pa
  scale_fill_discrete(labels=c("Out of HH","Within HH")) +
  scale_x_discrete(labels=c("before","R10","R11","R12","R13","R14","R15","R16","R17","R18",'after R18'))

ggarrange(p1, p2, ncol=1, nrow=2, common.legend = TRUE, legend="right")

```



BY AGE OF THE SOURCE AND RECIPIENT

```
p1<-ggplot(data = pairs_tsi[SEX.SOURCE=='M',], aes(x = AGE_INFECTION.RECIPIENT,y=AGE_TRANSMISSION.SOURCE))
  geom_point(size=1)+
  geom_abline(intercept=0,slope=1,color='black')+
  geom_smooth(data=pairs_tsi_diff_hh[SEX.SOURCE=='M',],color='red',size=1,se=FALSE)+
  geom_smooth(data=pairs_tsi_same_hh[SEX.SOURCE=='M',],color='blue',size=1,se=FALSE)+
  xlim(15,50)+
  ylim(15,50)+
  labs(title='M to F',
        x='Age at transmission of the recipient',
        y='Age at transmission of the source',
        color='Transmission type')+
  scale_color_manual(values=c("red", "blue"),labels=c('Out of household','Within household'))
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
p2<-ggplot(data = pairs_tsi[SEX.SOURCE=='F',], aes(x = AGE_INFECTION.RECIPIENT,y=AGE_TRANSMISSION.SOURCE))
  geom_point(size=1)+
```

```

geom_abline(intercept=0,slope=1,color='black')+
geom_smooth(data=pairs_tsi_diff_hh[SEX.SOURCE=='F',],color='red',size=1,se=FALSE)+
geom_smooth(data=pairs_tsi_same_hh[SEX.SOURCE=='F',],color='blue',size=1,se=FALSE)+
xlim(15,50)+
ylim(15,50)+
labs(title='                                F to M',
      x='Age at transmission of the recipient',
      y='Age at transmission of the source',
      color='Transmission type')+
scale_color_manual(values=c("red", "blue"),labels=c('Out of household','Within household'))
p3<-ggarrange(p1, p2, ncol=2, nrow=1, common.legend = TRUE, legend="bottom")

```

```

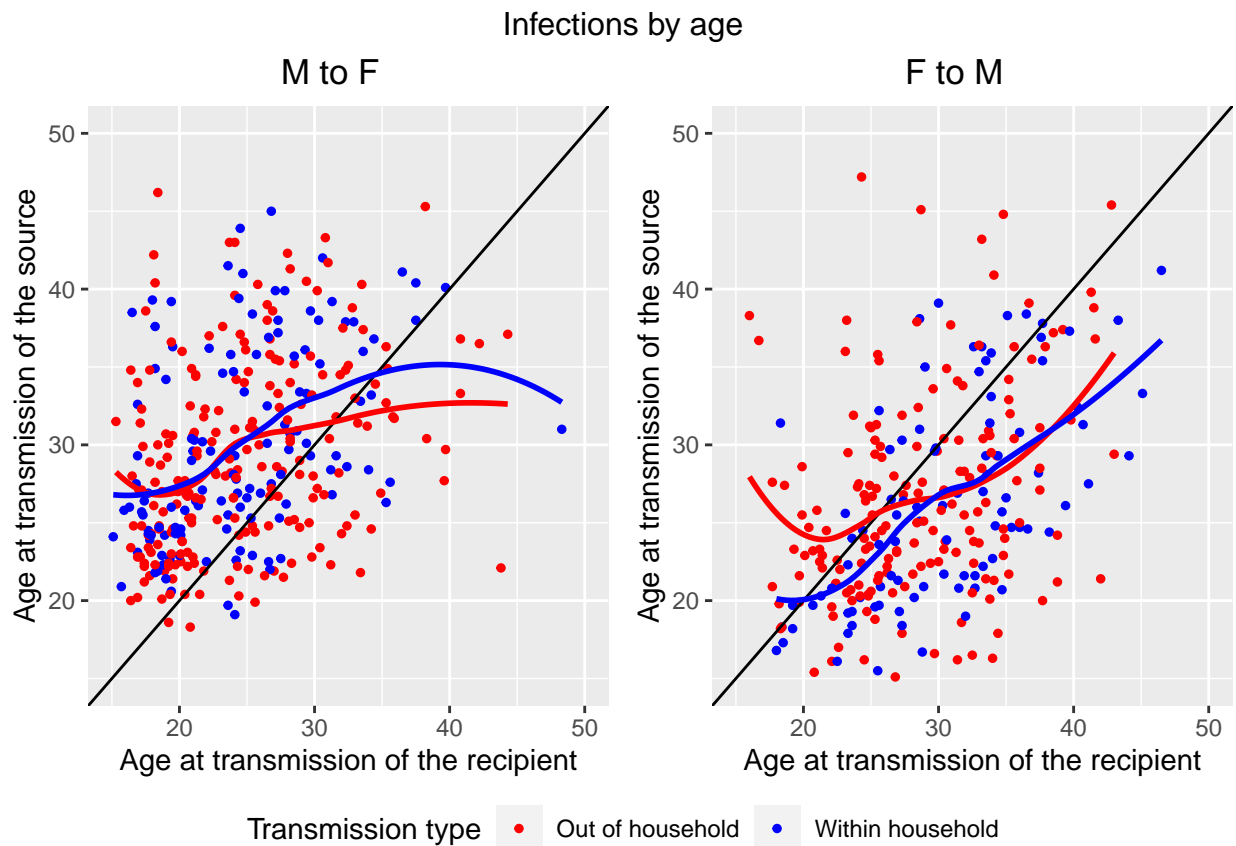
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'

```

```

annotate_figure(p3, top="Infections by age")

```



```

##### SOURCE AGE #####
med.fac = ddply(pairs_tsi, .(SEX.SOURCE, same_hh), function(.d)
data.frame(x=median(.d$AGE_TRANSMISSION.SOURCE)))

```

```
p1<-ggplot(data = pairs_tsi, aes(x = AGE_TRANSMISSION.SOURCE,color=as.logical(same_hh))) +
  geom_density(alpha=.2)+#, fill="#FF6666")+
  labs(title='Histogram of the age at transmission of sources',
        x='Age at transmission',
        color='Transmission type') +
  scale_color_manual(values=c("red", "blue"),labels=c('Out of household','Within household')) +
  facet_grid(~ SEX.SOURCE, scales="free_y") +
  geom_vline(data=med.fac[c(1,3),], aes(xintercept=x),color="red", linetype="dashed", size=.5)+
  geom_vline(data=med.fac[c(2,4),], aes(xintercept=x),color="blue", linetype="dashed", size=.5)+
  theme(legend.position='bottom')
```

#Double check:

```
median(pairs_tsi_diff_hh[SEX.SOURCE=='M',]$AGE_TRANSMISSION.SOURCE)
```

```
## [1] 28.3
```

```
median(pairs_tsi_same_hh[SEX.SOURCE=='M',]$AGE_TRANSMISSION.SOURCE)
```

```
## [1] 28.4
```

```
median(pairs_tsi_diff_hh[SEX.SOURCE=='F',]$AGE_TRANSMISSION.SOURCE)
```

```
## [1] 25.3
```

```
median(pairs_tsi_same_hh[SEX.SOURCE=='F',]$AGE_TRANSMISSION.SOURCE)
```

```
## [1] 25.3
```

RECIPIENT AGE

```
med.fac = ddply(pairs_tsi, .(SEX.RECIPIENT, same_hh), function(.d)
  data.frame(x=median(.d$AGE_INFECTION.RECIPIENT)))
p2<-ggplot(data = pairs_tsi, aes(x = AGE_INFECTION.RECIPIENT,color=as.logical(same_hh))) +
  geom_density(alpha=.2)+
  labs(title='Histogram of the age at transmission of recipients',
        x='Age at transmission ',
        color='Transmission type') +
  scale_color_manual(values=c("red", "blue"),labels=c('Out of household','Within household')) +
  facet_grid(~ SEX.RECIPIENT, scales="free_y") +
  geom_vline(data=med.fac[c(1,3),], aes(xintercept=x),color="red", linetype="dashed", size=.5)+
  geom_vline(data=med.fac[c(2,4),], aes(xintercept=x),color="blue", linetype="dashed", size=.5)+
  theme(legend.position='bottom')
```

#Double check:

```
median(pairs_tsi_diff_hh[SEX.RECIPIENT=='M',]$AGE_INFECTION.RECIPIENT)
```

```
## [1] 28.15
```

```
median(pairs_tsi_same_hh[SEX.RECIPIENT=='M'],)$AGE_INFECTION.RECIPIENT)
```

```
## [1] 30.15
```

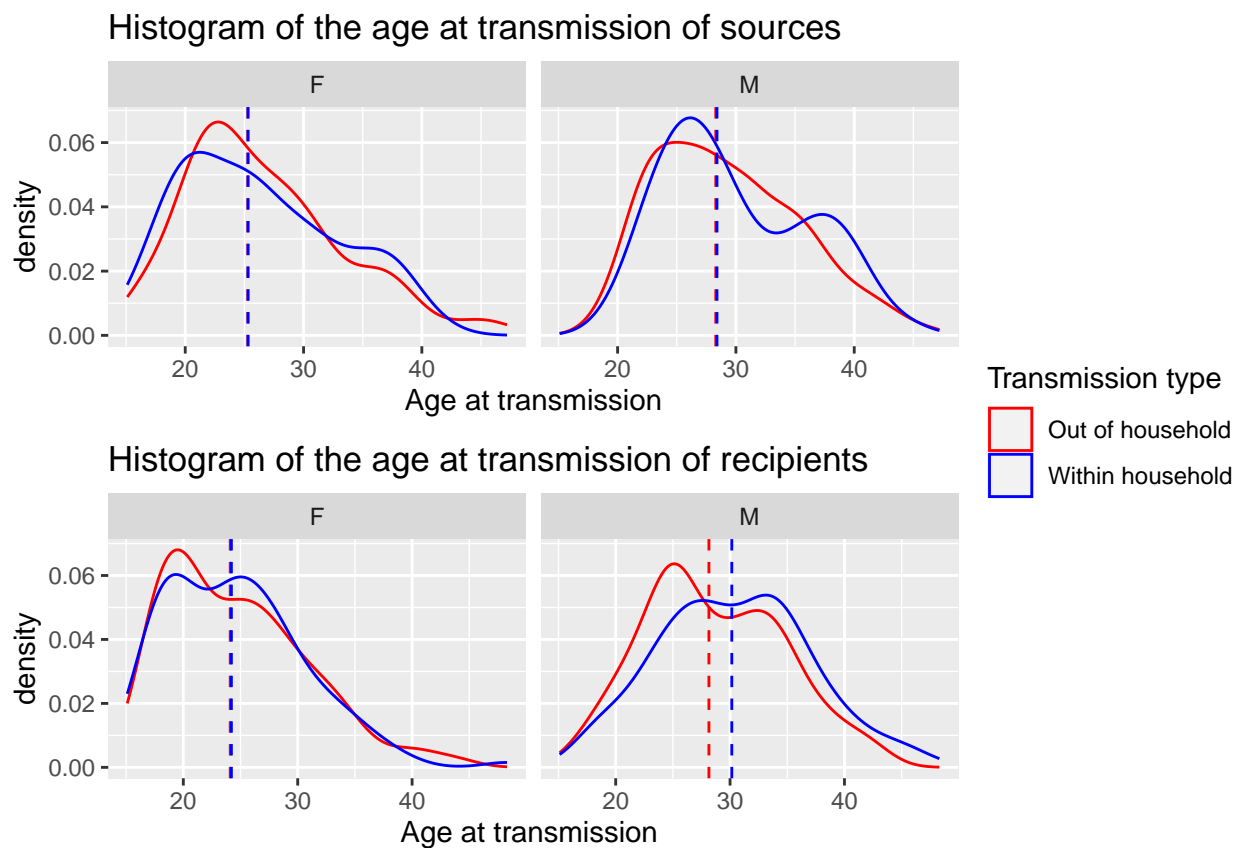
```
median(pairs_tsi_diff_hh[SEX.RECIPIENT=='F'],)$AGE_INFECTION.RECIPIENT)
```

```
## [1] 24.1
```

```
median(pairs_tsi_same_hh[SEX.RECIPIENT=='F'],)$AGE_INFECTION.RECIPIENT)
```

```
## [1] 24.2
```

```
ggarrange(p1, p2, ncol=1, nrow=2, common.legend = TRUE, legend="right")
```



```
##### 2-D KERNEL DENISTY ESTIMATION: #####
```

```
set.seed(1001)
```

```
#DIFFERENT HH, M to F
```

```
d <- matrix(c(pairs_tsi[as.logical(same_hh)==0 & SEX.SOURCE=='M', ]$AGE_INFECTION.RECIPIENT, pairs_tsi[
  magrittr::set_colnames(c("x", "y")) %>%
  as_tibble()
```

```

kd<-ks::kde(d, compute.cont=TRUE, h=0.2)
get_contour <- function(kd_out=kd, prob="5%") {
  contour_95 <- with(kd_out, contourLines(x=eval.points[[1]], y=eval.points[[2]],
                                           z=estimate, levels=cont[prob])[1]))
  as_tibble(contour_95) %>%
    mutate(prob = prob)
}

dat_out <- map_dfr(c("10%", "20%", "50%", "80%", "90%"), ~get_contour(kd, .)) %>%
  group_by(prob) %>%
  mutate(n_val = 1:n()) %>%
  ungroup()

## clean kde output
kd_df <- expand_grid(x=kd$eval.points[[1]], y=kd$eval.points[[2]]) %>%
  mutate(z = c(kd$estimate %>% t))

p1<-ggplot(data=kd_df, aes(x, y)) +
  geom_tile(aes(fill=z)) +
  geom_path(aes(x, y, group = prob),
            data=dat_out, colour = I("black")) +
  geom_text(aes(label = prob), data =
            filter(dat_out, (prob%in% c("10%") & n_val==100 | prob%in% c("20%") & n_val==80) | prob%in%
            colour = I("black"), size =I(3))+
  xlim(15,50)+
  ylim(15,50)+
  geom_abline(intercept=0,slope=1,color='black')+
  scale_fill_gradient(low = "white", high = "red") +
  labs(x='Age of recipient at transmission ',
       y='Age of source at transmission') +
  theme_bw() +
  theme(legend.position = "none")

##SAME HH, M to F
d <- matrix(c(pairs_tsi[as.logical(same_hh)==1 & SEX.SOURCE=='M', ]$AGE_INFECTION.RECIPIENT, pairs_tsi[
  magrittr::set_colnames(c("x", "y")) %>%
  as_tibble()

kd<-ks::kde(d, compute.cont=TRUE, h=0.2)
get_contour <- function(kd_out=kd, prob="5%") {
  contour_95 <- with(kd_out, contourLines(x=eval.points[[1]], y=eval.points[[2]],
                                           z=estimate, levels=cont[prob])[1]))
  as_tibble(contour_95) %>%
    mutate(prob = prob)
}

dat_out <- map_dfr(c("10%", "20%", "50%", "80%", "90%"), ~get_contour(kd, .)) %>%
  group_by(prob) %>%
  mutate(n_val = 1:n()) %>%
  ungroup()

## clean kde output

```

```

kd_df <- expand_grid(x=kd$eval.points[[1]], y=kd$eval.points[[2]]) %>%
  mutate(z = c(kd$estimate %>% t))

p2<-ggplot(data=kd_df, aes(x, y)) +
  geom_tile(aes(fill=z)) +
  geom_path(aes(x, y, group = prob),
            data=dat_out, colour = I("black")) +
  geom_text(aes(label = prob), data =
            filter(dat_out, (prob%in% c("10%") & n_val==100 | prob%in% c("20%") & n_val==80) | prob%in%
            colour = I("black"), size =I(3))+
  xlim(15,50)+
  ylim(15,50)+
  geom_abline(intercept=0,slope=1,color='black')+
  scale_fill_gradient(low = "white", high = "blue") +
  labs(x='Age of recipient at transmission ',
       y='Age of source at transmission') +
  theme_bw() +
  theme(legend.position = "none")
p3<-ggarrange(p1,p2,ncol=2,nrow=1)

```

```
## Warning: Removed 8527 rows containing missing values ('geom_tile()').
```

```
## Warning: Removed 60 rows containing missing values ('geom_path()').
```

```
## Warning: Removed 10153 rows containing missing values ('geom_tile()').
```

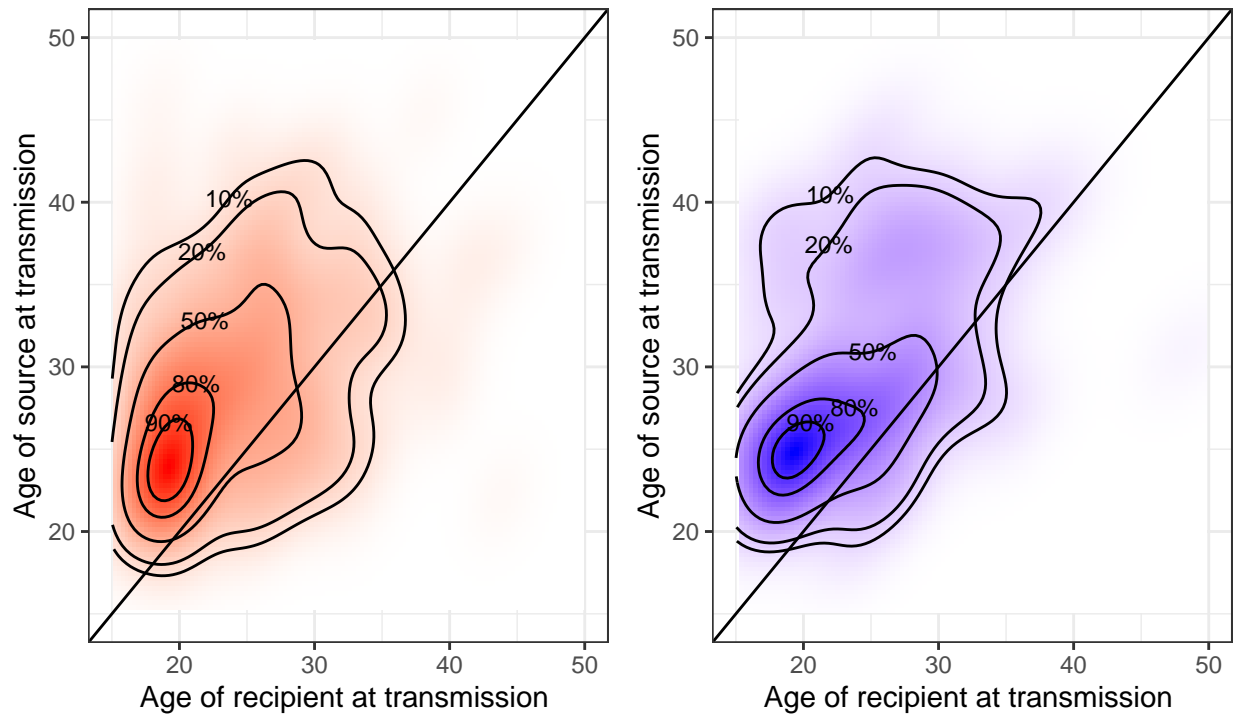
```
## Warning: Removed 76 rows containing missing values ('geom_path()').
```

```

annotate_figure(p3,bottom=('Red: Out of household \nBlue: Within household'), top='Contour plots for in

```


Contour plots for infections M to F



Red: Out of household

Blue: Within household

```
#DIFFERENT HH, F to M
d <- matrix(c(pairs_tsi[as.logical(same_hh)==0 & SEX.SOURCE=='F', ]$AGE_INFECTION.RECIPIENT, pairs_tsi[
  magrittr::set_colnames(c("x", "y")) %>%
  as_tibble()

kd<-ks::kde(d, compute.cont=TRUE, h=0.2)
get_contour <- function(kd_out=kd, prob="5%") {
  contour_95 <- with(kd_out, contourLines(x=eval.points[[1]], y=eval.points[[2]],
                                           z=estimate, levels=cont[prob])[1]))
  as_tibble(contour_95) %>%
  mutate(prob = prob)
}

dat_out <- map_dfr(c("10%", "20%", "50%", "80%", "90%"), ~get_contour(kd, .)) %>%
  group_by(prob) %>%
  mutate(n_val = 1:n()) %>%
  ungroup()

## clean kde output
kd_df <- expand_grid(x=kd$eval.points[[1]], y=kd$eval.points[[2]]) %>%
  mutate(z = c(kd$estimate %>% t))

p1<-ggplot(data=kd_df, aes(x, y)) +
  geom_tile(aes(fill=z)) +
  geom_path(aes(x, y, group = prob),
            data=dat_out, colour = I("black")) +
```

```

geom_text(aes(label = prob), data =
  filter(dat_out, (prob%in% c("10%") & n_val==100 | prob%in% c("20%") & n_val==80) | prob%in%
    colour = I("black"), size =I(3))+
xlim(15,50)+
ylim(15,50)+
geom_abline(intercept=0,slope=1,color='black')+
scale_fill_gradient(low = "white", high = "red") +
labs(x='Age of recipient at transmission ',
  y='Age of source at transmission') +
theme_bw() +
theme(legend.position = "none")

#SAME HH, F to M
d <- matrix(c(pairs_tsi[as.logical(same_hh)==1 & SEX.SOURCE=='F', ]$AGE_INFECTION.RECIPIENT, pairs_tsi[
  magrittr::set_colnames(c("x", "y")) %>%
  as_tibble()

kd<-ks::kde(d, compute.cont=TRUE, h=0.2)
get_contour <- function(kd_out=kd, prob="5%") {
  contour_95 <- with(kd_out, contourLines(x=eval.points[[1]], y=eval.points[[2]],
    z=estimate, levels=cont[prob])[1]))
  as_tibble(contour_95) %>%
  mutate(prob = prob)
}

dat_out <- map_dfr(c("10%", "20%", "50%", "80%", "90%"), ~get_contour(kd, .)) %>%
  group_by(prob) %>%
  mutate(n_val = 1:n()) %>%
  ungroup()

## clean kde output
kd_df <- expand_grid(x=kd$eval.points[[1]], y=kd$eval.points[[2]]) %>%
  mutate(z = c(kd$estimate %>% t))

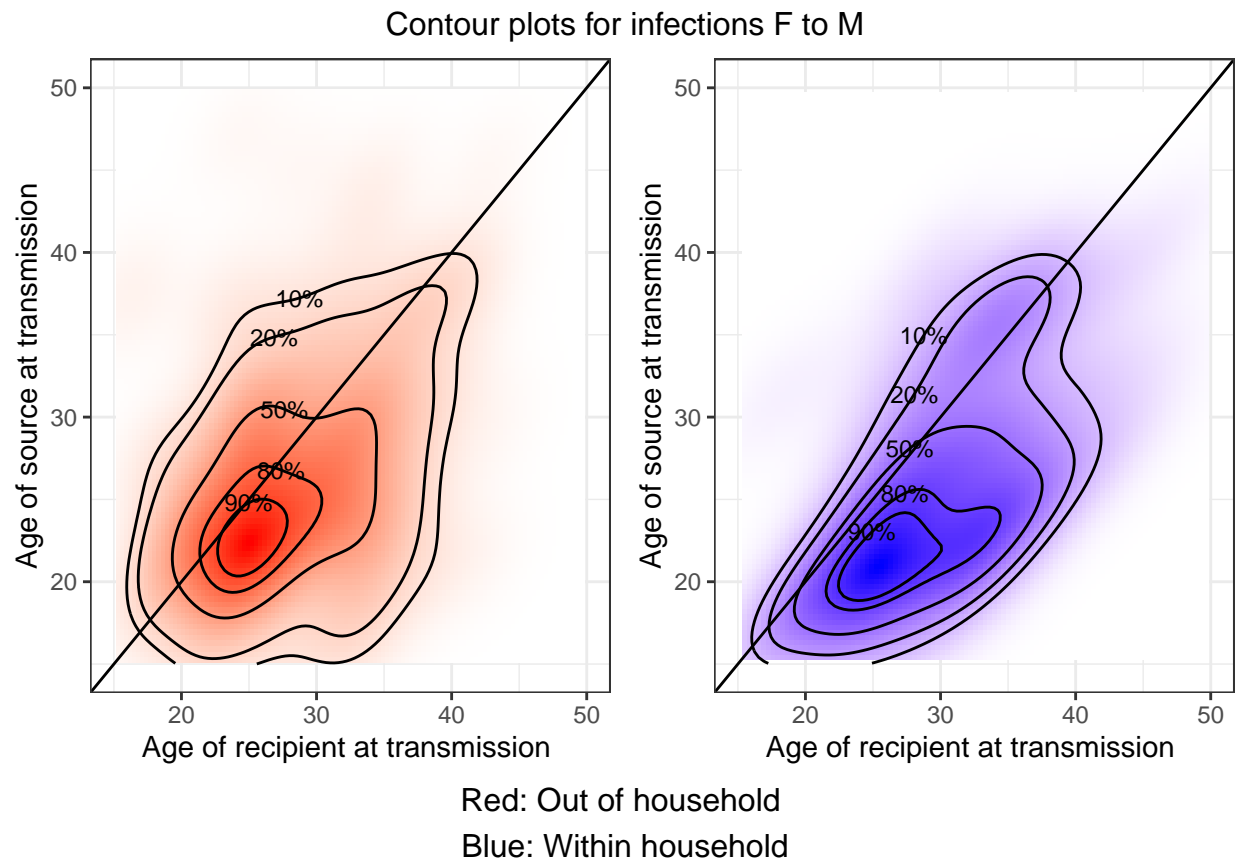
p2<-ggplot(data=kd_df, aes(x, y)) +
  geom_tile(aes(fill=z)) +
  geom_path(aes(x, y, group = prob),
    data=dat_out, colour = I("black")) +
  geom_text(aes(label = prob), data =
    filter(dat_out, (prob%in% c("10%") & n_val==100 | prob%in% c("20%") & n_val==80) | prob%in%
      colour = I("black"), size =I(3))+
xlim(15,50)+
ylim(15,50)+
geom_abline(intercept=0,slope=1,color='black')+
scale_fill_gradient(low = "white", high = "blue") +
labs(x='Age of recipient at transmission ',
  y='Age of source at transmission') +
theme_bw() +
theme(legend.position = "none")
p3<-ggarrange(p1,p2,ncol=2,nrow=1)

```

```
## Warning: Removed 10956 rows containing missing values (‘geom_tile()’).
```

```
## Warning: Removed 11713 rows containing missing values (‘geom_tile()’).
```

```
annotate_figure(p3,bottom=('Red: Out of household \nBlue: Within household'), top='Contour plots for in
```



```
#For the table
```

```
nrow(pairs_tsi_same_hh[SEX.SOURCE=='M',])
```

```
## [1] 127
```

```
nrow(pairs_tsi_same_hh[SEX.SOURCE=='F',])
```

```
## [1] 92
```

```
nrow(pairs_tsi_diff_hh[SEX.SOURCE=='M',])
```

```
## [1] 217
```

```
nrow(pairs_tsi_diff_hh[SEX.SOURCE=='F',])
```

```
## [1] 174
```

```
binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='M',]),nrow(pairs_tsi))*nrow(pairs_tsi)
```

```
## PointEst Lower Upper
## 127 108.4895 147.7383
```

```
binconf(nrow(pairs_tsi_same_hh[SEX.SOURCE=='F',]),nrow(pairs_tsi))*nrow(pairs_tsi)
```

```
## PointEst Lower Upper
## 92 76.01214 110.6538
```

```
binconf(nrow(pairs_tsi_diff_hh[SEX.SOURCE=='M',]),nrow(pairs_tsi))*nrow(pairs_tsi)
```

```
## PointEst Lower Upper
## 217 194.4423 240.6591
```

```
binconf(nrow(pairs_tsi_diff_hh[SEX.SOURCE=='F',]),nrow(pairs_tsi))*nrow(pairs_tsi)
```

```
## PointEst Lower Upper
## 174 153.0154 196.6243
```

Further summary statistics

```
# Convert strings to factors
pairs_tsi[,SEX.SOURCE:=as.factor(SEX.SOURCE)]
pairs_tsi[,SEX.RECIPIENT:=as.factor(SEX.RECIPIENT)]
pairs_tsi[,ROUND.M:=as.factor(ROUND.M)]
pairs_tsi[,COMM.SOURCE:=as.factor(COMM.SOURCE)]
pairs_tsi[,COMM.RECIPIENT:=as.factor(COMM.RECIPIENT)]
pairs_tsi[,same_hh:=as.factor(same_hh)]

# Return summary of data set
summary(pairs_tsi)
```

```
## RECIPIENT SOURCE SEX.SOURCE SEX.RECIPIENT
## Length:610 Length:610 F:266 F:344
## Class :character Class :character M:344 M:266
## Mode :character Mode :character
##
##
##
## CL IL M
## Min. :1998-11-20 Min. :2000-01-17 Min. :2000-04-09
## 1st Qu.:2006-07-21 1st Qu.:2007-08-22 1st Qu.:2009-01-18
## Median :2009-11-18 Median :2010-05-29 Median :2011-03-12
## Mean :2009-08-02 Mean :2010-04-15 Mean :2011-03-18
## 3rd Qu.:2012-05-09 3rd Qu.:2012-07-20 3rd Qu.:2013-01-18
## Max. :2018-08-06 Max. :2018-09-14 Max. :2018-11-17
```

```
##
##          IU                      CU                      AGE_TRANSMISSION.SOURCE
## Min.    :2000-06-28  Min.    :2000-08-13  Min.    :15.10
## 1st Qu.:2010-06-14  1st Qu.:2011-03-07  1st Qu.:23.10
## Median :2011-09-23  Median :2012-02-09  Median :27.10
## Mean   :2012-01-07  Mean   :2012-06-25  Mean   :28.22
## 3rd Qu.:2013-07-17  3rd Qu.:2013-11-21  3rd Qu.:32.60
## Max.   :2019-01-20  Max.   :2019-02-28  Max.   :47.20
##
## AGE_INFECTION.RECIPIENT  ROUND.M  COMM.SOURCE  COMM.RECIPIENT
## Min.    :15.10          R015    :130  fishing:326  fishing:333
## 1st Qu.:21.20          R013    :122  inland :284  inland :277
## Median :25.75          R014    :121
## Mean   :26.66          R012    : 78
## 3rd Qu.:31.40          R016    : 60
## Max.   :48.30          R017    : 39
##                      (Other): 60
## COMM_NUM.SOURCE COMM_NUM.RECIPIENT DATE.COLLECTION.SOURCE
## Min.    : 2.0  Min.    : 0.0  Min.    :2010-01-20
## 1st Qu.: 38.0  1st Qu.: 38.0  1st Qu.:2011-12-02
## Median : 40.0  Median : 39.0  Median :2012-04-29
## Mean   :285.6  Mean   :286.3  Mean   :2012-10-28
## 3rd Qu.:770.0  3rd Qu.:770.0  3rd Qu.:2013-01-20
## Max.   :776.0  Max.   :776.0  Max.   :2019-05-14
##
## DATE.COLLECTION.RECIPIENT same_hh AGE_BAND.RECIPIENT
## Min.    :2010-02-08      0:391  15-20:115
## 1st Qu.:2012-02-07      1:219  20-25:157
## Median :2012-10-14
## Mean   :2013-08-20
## 3rd Qu.:2014-11-21
## Max.   :2019-05-17
##                      40-45: 17
##                      45-50: 3
```

```
# Perform chisq tests of independence
```

```
chisq.test(pairs_tsi$same_hh, pairs_tsi$SEX.SOURCE)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: pairs_tsi$same_hh and pairs_tsi$SEX.SOURCE
## X-squared = 0.26043, df = 1, p-value = 0.6098
```

```
chisq.test(pairs_tsi$same_hh, pairs_tsi$COMM.SOURCE)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: pairs_tsi$same_hh and pairs_tsi$COMM.SOURCE
## X-squared = 0.0083285, df = 1, p-value = 0.9273
```

```

# Plot infections by community (source) and within-household
n_obs <- pairs_tsi[, .N]
data_comm_hh <- pairs_tsi[, .(count = .N), by = .(COMM.SOURCE, same_hh)]

p1_data <- pairs_tsi[, {
  N.in.inland <- sum(COMM.RECIPIENT == 'inland')
  binom.conf <- (.N*binconf(N.in.inland, .N)) |> as.list()
  names(binom.conf) <- c('BC.center', 'BC.min', 'BC.max')
  binom.conf
}, by=same_hh]

p <- ggplot(data_comm_hh, aes(fill = COMM.SOURCE, y = count, x = same_hh)) +
  geom_bar(position = "stack", stat = "identity") +
  geom_errorbar(data = data_comm_hh[, COMM.SOURCE == "inland"],
    aes(ymin = BC.min, ymax = BC.max, fill = NA), width=0.4,
    colour="black", alpha=0.9, linewidth=1.3) +
  labs(x='transmission type',
    y='number of infections',
    fill='recipient community') +
  scale_x_discrete(labels=c("out of hh", "within hh"))

print(p)

```