pure HRT variable creation and summary

Variable summary from Yi:

From Yi and Yu:

Ever	Description
hrt_ref_pm2No	1= no EO and no EP and no any HRT; 0= any EO,EP,any HRT
pure_eo_allNo	1= no EO and no EP and no any HRT; 0= pure EO and no EP and no any HRT
pure_eo_anyNo	1= no EO or no EP or no any HRT; 0 = pure EO and no EP and no any HRT
pure_ep_allNo	1= no EO and no EP and no any HRT; 0 = pure EP and no EO and no any HRT
pure_ep_anyNo	1= no EO or no EP or no any HRT; 0 = pure EP and no EO and no any HRT
eo_allNo	1= no EO and no EP and no any HRT;0 = EO regardless EP/any HRT
eo_anyNo	1= no EO or no EP or no any HRT; 0 = EO regardless EP/any HRT
ep_allNo	1= no EO and no EP and no any HRT; 0 = EP regardless EP/any HRT
ep_anyNo	1= no EO or no EP or no any HRT; 0 = EP regardless EP/any HRT
eo_ref_pm	Post menopausal EO use
ep_ref_pm	Post menopausal EP use
horm_ref_pm	Post menopausal any hormone use
hrt_ref_pm2	0= no EO and no EP and no any HRT; 1= any EO,EP,any HRT
pure_eo_all	0= no EO and no EP and no any HRT; 1 = pure EO and no EP and no any HRT
pure_ep_all	0= no EO and no EP and no any HRT; 1 = pure EO and no EP and no any HRT
eo_all (eo_ref_pm_gxe)	0= no EO and no EP and no any HRT;1 = EO regardless EP/any HRT
ep_all (ep_ref_pm_gxe)	0= no EO and no EP and no any HRT; 1 = EP regardless EP/any HRT

Request from Yu:

> I would like to request for re-analyzing GxE based on pure definition for EO and EP? Because we have found a significant SNP (rs79439591) by EP but neither SNPs in LD with it nor functional annotation, I guess the SNP could be gone if we use stringent EP definition. You could only check for this SNP first, don't have to run analysis for all.

---

recode variables and calculate associations

E+P

ep_ref_pm == "Yes"

```
ep_yes <- filter(figi_gxe, ep_ref_pm == "Yes")
cro(ep_yes$eo_ref_pm, ep_yes$hrt_ref_pm2)
```

ep_yes$hrt_ref_pm2	
No	
Yes	
ep_yes$eo_ref_pm	
No	1955
Yes	493
#Total cases	2448

ep_ref_pm == "No"

```
ep_no <- filter(figi_gxe, ep_ref_pm == "No")
cro(ep_no$eo_ref_pm, ep_no$hrt_ref_pm2)
```

ep_no$hrt_ref_pm2	
No	
Yes	
ep_no$eo_ref_pm	
No	6008
Yes	142
#Total cases	6008
	2938
	3080

`ep_ref_pm_gxe` includes individuals that were treated with ONLY ESTROGEN. Exclude from definition to create `pure_ep_allNo`:

pure_ep_allNo:

```
tmp <- figi_gxe %>%
  mutate(pure_ep_allNo = ifelse(ep_ref_pm == "Yes" & eo_ref_pm == "No", "Yes",
                        ifelse(ep_ref_pm == "No" & hrt_ref_pm2 == "No" & eo_ref_pm == "No", "No", NA)))

cro(tmp$study_gxe, tmp$pure_ep_allNo)
```

tmp$pure_ep_allNo	No	Yes
tmp$study_gxe		
CCFR_1	389	99
CCFR_3	447	69
CCFR_4	336	62
CLUEII	195	2
CPSII_1	318	67
CPSII_2	217	51
Kentucky	334	142
MEC_1	97	
MEC_2	6	1

NFCCR_2

155

NHS_1_2

495

91

NHS_3_AD

320

110

NHS_4

115

NHS_5_AD

94

REACH_AD

54

USC_HRT_CRC

277

149

VITAL

121

61

WHI_1

579

158

WHI_2

1088

416

WHI_3

580

268

#Total cases

6008

1955

Remove studies that only have either 0 or 1 pure_ep_allNo value or very sparse cells e.g. CLUEII. While this isn't necessary for GLM, I do it for GxEScanR so let me keep things consistent just in case:

```r
drops <- data.frame(table(tmp$outcome, tmp[, 'pure_ep_allNo'], tmp$study_gxe)) %>%
  filter(Freq <= 0)
figi_gxe_pure_ep <- tmp %>%
  filter(!study_gxe %in% unique(drops$Var3)) %>%
  dplyr::mutate(study_gxe = fct_drop(study_gxe)) %>%
  dplyr::select(vcfid, outcome, age_ref_imp, sex, study_gxe, pure_ep_allNo, pc1, pc2, pc3) %>%
  filter(complete.cases(.))


cro(figi_gxe_pure_ep$study_gxe, figi_gxe_pure_ep$pure_ep_allNo)
```

figi_gxe_pure_ep$pure_ep_allNo

No

Yes

figi_gxe_pure_ep$study_gxe

CCFR_1

389

99

CCFR_3

447

69

CCFR_4

336

62

CPSII_1

318

67

CPSII_2

217

51

Kentucky

334

142

NHS_1_2

495

91

NHS_3_AD

320

110

USC_HRT_CRC

277

149

VITAL

121

61

WHI_1

579

158

WHI_2

1088

416

WHI_3

580

268

#Total cases

5501

1743

E only

eo_ref_pm == "Yes"

```
eo_yes <- filter(figi_gxe, eo_ref_pm == "Yes")
cro(eo_yes$ep_ref_pm, eo_yes$hrt_ref_pm2)
```

eo_yes$hrt_ref_pm2

No

Yes

eo_yes$ep_ref_pm

No

2938

Yes

493

#Total cases

3431

eo_ref_pm == "No"

```
eo_no <- filter(figi_gxe, eo_ref_pm == "No")
cro(eo_no$ep_ref_pm, eo_no$hrt_ref_pm2)
```

eo_no$hrt_ref_pm2	
No	
Yes	
eo_no$ep_ref_pm	
No	
6008	
142	
Yes	
1955	
#Total cases	
6008	
2097	

similar to estrogen+progesterone variable, `eo_ref_pm_gxe` includes individuals that were treated with E+P. Exclude from definition to create `pure_eo_allNo`:

pure_eo_allNo:

```
tmp <- figi_gxe %>%
  mutate(pure_eo_allNo = ifelse(eo_ref_pm == "Yes" & ep_ref_pm == "No", "Yes",
                        ifelse(eo_ref_pm == "No" & hrt_ref_pm2 == "No" & ep_ref_pm == "No", "No", NA)))

cro(tmp$study_gxe, tmp$pure_eo_allNo)
```

tmp$pure_eo_allNo	
No	
Yes	
tmp$study_gxe	
CCFR_1	
389	
128	
CCFR_3	
447	
122	
CCFR_4	

336

94

CLUEII

195

23

CPSII_1

318

133

CPSII_2

217

81

Kentucky

334

266

MEC_1

97

64

MEC_2

6

17

NFCCR_2

155

NHS_1_2

495

271

NHS_3_AD

320

182

NHS_4

193

NHS_5_AD

163

REACH_AD

54

7

USC_HRT_CRC

277

191

VITAL

121

44

WHI_1

579

232

WHI_2

1088

462

WHI_3

580

265

#Total cases

6008

2938

Again, remove studies that only have either 0 or 1 pure_eo_allNo value or very sparse cells:

```
drops <- data.frame(table(tmp$outcome, tmp[, 'pure_eo_allNo'], tmp$study_gxe)) %>%
  filter(Freq <= 0)
figi_gxe_pure_eo <- tmp %>%
  filter(!study_gxe %in% unique(drops$Var3)) %>%
  dplyr::mutate(study_gxe = fct_drop(study_gxe)) %>%
  dplyr::select(vcfid, outcome, age_ref_imp, sex, study_gxe, pure_eo_allNo, pc1, pc2, pc3) %>%
  filter(complete.cases(.))

cro(figi_gxe_pure_eo$study_gxe, figi_gxe_pure_eo$pure_eo_allNo)
```

figi_gxe_pure_eo$pure_eo_allNo

No

Yes

figi_gxe_pure_eo$study_gxe

CCFR_1

389

128

CCFR_3

447

122

CCFR_4

336

94

CLUEII

195

23

CPSII_1

318

133

CPSII_2

217

81

Kentucky

334

266

MEC_1

97

64

MEC_2

6

17

NHS_1_2

495

271

NHS_3_AD

320

182

USC_HRT_CRC

277

191

VITAL

121

44

WHI_1

579

232

WHI_2

1088

462

WHI_3

580

265

#Total cases

5799

2575

Association with rs79439591 (1:53785007:C:T)

start with pure_ep_allNo

```
exposure <- "pure_ep_allNo"
snp <- "X1.53785007.C.T_dose"

dat <- inner_join(figi_gxe_pure_ep, snps, 'vcfid')

model     <- glm(glue("outcome ~ {exposure}*{snp} + age_ref_imp + pc1 + pc2 + pc3 + study_gxe"), data =
model_ref <- glm(glue("outcome ~ {exposure}+{snp} + age_ref_imp + pc1 + pc2 + pc3 + study_gxe"), data =
summary(model)
```

```
##
## Call:
## glm(formula = glue("outcome ~ {exposure}*{snp} + age_ref_imp + pc1 + pc2 + pc3 + study_gxe"),
##     family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8649  -1.1764   0.7314   1.1343   1.9447
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         0.626358   0.254828   2.458 0.013973
## pure_ep_allNoYes                   -0.287589   0.060940  -4.719 2.37e-06
## X1.53785007.C.T_dose               -0.259573   0.094828  -2.737 0.006194
## age_ref_imp                        -0.012154   0.003553  -3.421 0.000624
## pc1                               -50.100077 135.000998  -0.371 0.710556
## pc2                               -23.166353  69.226211  -0.335 0.737891
## pc3                                35.407356  22.949553   1.543 0.122871
## study_gxeCCFR_3                     0.866529   0.130921   6.619 3.62e-11
## study_gxeCCFR_4                     1.359072   0.148758   9.136  < 2e-16
## study_gxeCPSII_1                    0.544906   0.141263   3.857 0.000115
## study_gxeCPSII_2                    0.377088   0.156023   2.417 0.015655
## study_gxeKentucky                   0.346239   0.132025   2.623 0.008728
```

```
## study_gxeNHS_1_2                       -0.398579   0.126420  -3.153 0.001617
## study_gxeNHS_3_AD                       0.484469   0.134104   3.613 0.000303
## study_gxeUSC_HRT_CRC                    0.171921   0.135870   1.265 0.205752
## study_gxeVITAL                          0.352937   0.177762   1.985 0.047094
## study_gxeWHI_1                          0.272818   0.121594   2.244 0.024854
## study_gxeWHI_2                          0.403223   0.107899   3.737 0.000186
## study_gxeWHI_3                          0.414247   0.116721   3.549 0.000387
## pure_ep_allNoYes:X1.53785007.C.T_dose  -0.627070   0.210217  -2.983 0.002855
##
## (Intercept)                           *
## pure_ep_allNoYes                      ***
## X1.53785007.C.T_dose                  **
## age_ref_imp                           ***
## pc1
## pc2
## pc3
## study_gxeCCFR_3                       ***
## study_gxeCCFR_4                       ***
## study_gxeCPSII_1                      ***
## study_gxeCPSII_2                      *
## study_gxeKentucky                     **
## study_gxeNHS_1_2                      **
## study_gxeNHS_3_AD                     ***
## study_gxeUSC_HRT_CRC
## study_gxeVITAL                        *
## study_gxeWHI_1                        *
## study_gxeWHI_2                        ***
## study_gxeWHI_3                        ***
## pure_ep_allNoYes:X1.53785007.C.T_dose **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10034.8  on 7243  degrees of freedom
## Residual deviance:  9731.2  on 7224  degrees of freedom
## AIC: 9771.2
##
## Number of Fisher Scoring iterations: 4
```

```
lrtest(model, model_ref)
```

```
## Likelihood ratio test
##
## Model 1: outcome ~ pure_ep_allNo * X1.53785007.C.T_dose + age_ref_imp +
##     pc1 + pc2 + pc3 + study_gxe
## Model 2: outcome ~ pure_ep_allNo + X1.53785007.C.T_dose + age_ref_imp +
##     pc1 + pc2 + pc3 + study_gxe
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  20 -4865.6
## 2  19 -4870.3 -1 9.3271   0.002258 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# original ep_ref_pm_gxe variable:
exposure <- "ep_ref_pm_gxe"

model     <- glm(glue("outcome ~ {exposure}*{snp} + age_ref_imp + pc1 + pc2 + pc3 + study_gxe"), data =
model_ref <- glm(glue("outcome ~ {exposure}+{snp} + age_ref_imp + pc1 + pc2 + pc3 + study_gxe"), data =
summary(model)
```

```
##
## Call:
## glm(formula = glue("outcome ~ {exposure}*{snp} + age_ref_imp + pc1 + pc2 + pc3 + study_gxe"),
##     family = "binomial", data = ep_ref_pm_gxe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8642  -1.1560   0.6716   1.1414   1.9270
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        0.657271   0.243725   2.697 0.007001 **
## ep_ref_pm_gxe                     -0.318555   0.055770  -5.712 1.12e-08 ***
## X1.53785007.C.T_dose              -0.259942   0.092330  -2.815 0.004872 **
## age_ref_imp                       -0.012210   0.003386  -3.606 0.000311 ***
## pc1                              -67.022124 128.378276  -0.522 0.601623
## pc2                              -26.687726  64.065311  -0.417 0.676993
## pc3                               39.603054  21.380012   1.852 0.063977 .
## study_gxeCCFR_3                    0.874103   0.128120   6.823 8.94e-12 ***
## study_gxeCCFR_4                    1.343111   0.146960   9.139  < 2e-16 ***
## study_gxeCPSII_1                   0.536419   0.140229   3.825 0.000131 ***
## study_gxeCPSII_2                   0.368415   0.155114   2.375 0.017543 *
## study_gxeKentucky                  0.216880   0.123778   1.752 0.079743 .
## study_gxeMEC_1                     0.278660   0.188738   1.476 0.139826
## study_gxeMEC_2                    -0.176480   0.427227  -0.413 0.679547
## study_gxeNFCCR_2                  -0.568986   0.190609  -2.985 0.002835 **
## study_gxeNHS_1_2                  -0.383315   0.122106  -3.139 0.001694 **
## study_gxeNHS_3_AD                  0.559476   0.129420   4.323 1.54e-05 ***
## study_gxeUSC_HRT_CRC               0.129551   0.129918   0.997 0.318681
## study_gxeVITAL                     0.323038   0.172509   1.873 0.061125 .
## study_gxeWHI_1                     0.266887   0.120383   2.217 0.026625 *
## study_gxeWHI_2                     0.396810   0.106616   3.722 0.000198 ***
## study_gxeWHI_3                     0.408064   0.115546   3.532 0.000413 ***
## ep_ref_pm_gxe:X1.53785007.C.T_dose -0.591457   0.188241  -3.142 0.001678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11123  on 8023  degrees of freedom
## Residual deviance: 10743  on 8001  degrees of freedom
## AIC: 10789
##
## Number of Fisher Scoring iterations: 4
```

```r
lrtest(model, model_ref)
```

```
## Likelihood ratio test
##
## Model 1: outcome ~ ep_ref_pm_gxe * X1.53785007.C.T_dose + age_ref_imp +
##     pc1 + pc2 + pc3 + study_gxe
## Model 2: outcome ~ ep_ref_pm_gxe + X1.53785007.C.T_dose + age_ref_imp +
##     pc1 + pc2 + pc3 + study_gxe
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  23 -5371.6
## 2  22 -5376.8 -1 10.268   0.001353 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```