The genetics package

Utilities for handling genetic data

by Gregory R. Warnes

```
g2 <- genotype( c('AA','AC','CC','CA','',
                  'AA','AC','AC'), sep=1 )
```

Two separate vectors

```
g4 <- genotype(
        c('A','A','C','C','','A','A','A'),
        c('A','C','C','A','','A','C','C')
        )
```

A dataframe or matrix with two columns

```
gm <- cbind(
        c('A','A','C','C','','A','A','A'),
        c('A','C','C','A','','A','C','C') )
g5 <- genotype( gm )
```

For simplicity, the functions makeGenotype and makeHaplotype can be used to convert all of the genetic variables contained in a dataframe in a single pass. (See the help page for details.)

efit: in most contexts factors behave the same as the desired default behavior for genotype objects. Consequently, relatively few additional methods needed to written. Further, in the absence of the genetics package, the information stored in genotype objects

```
+          bp.start=1691,
+          relative.to="intron 1")
[...]
>
> # Look at some of the data
> data[1:5,]
      PID DELTA.BMI c104t a1691g c2249t
1 1127409      0.62   C/C    G/G    T/T
2  246311      1.31   C/C    A/A    T/T
3  295185      0.15   C/C    G/G    T/T
4   34301      0.72   C/T    A/A    T/T
5   96890      0.37   C/C    A/A    T/T
>
> # Get allele information for c104t
> summary(data$c104t)

Marker: MBP2:C-104T (9q35:-104) Type: SNP

Allele Frequency:
  Count Proportion
C   137       0.68
T    63       0.32

Genotype Frequency:
    Count Proportion
C/C    59       0.59
C/T    19       0.19
T/T    22       0.22

>
>
> # Check Hardy-Weinberg Equilibrium
> HWE.test(data$c104t)

        ----------------------------------
        Test for Hardy-Weinberg-Equilibrium
        ----------------------------------

Call:
HWE.test.genotype(x = data$c104t)

Raw Disequlibrium for each allele pair (D)

  C    T
C      0.12
T 0.12

Scaled Disequlibrium for each allele pair (D')

  C    T
C      0.56
T 0.56

Correlation coefficient for each allele pair (r)

  C     T
C  1.00 -0.56
T -0.56  1.00

Overall Values

     Value
  D   0.12
  D'  0.56
  r  -0.56

Confidence intervals computed via bootstrap
using 1000 samples

             Observed 95% CI           NA's
  Overall D   0.121   ( 0.073,  0.159) 0
  Overall D'  0.560   ( 0.373,  0.714) 0
  Overall r  -0.560   (-0.714, -0.373f((-n17(D)-0s.7(D')-1s.560)-
  Overale 0
  Overall [T(test)-520(for)-521(Hardy-Wein)10(berg)-520(Equilibri
```

```
>
> LDtable(ld) # graphical display
```

```
> # fit a model
> summary(lm( DELTA.BMI ~
+                 homozygote(c104t,'C') +
+                 allele.count(a1691g, 'G') +
+                 c2249t, data=data))

Call:
lm(formula = DELTA.BMI ~ homozygote(c104t, "C") +
    allele.count(a1691g, "G") + c2249t,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9818 -0.5917 -0.0303  0.6666  2.7101

Coefficients:
                           Estimate Std. Error
(Intercept)                 -0.1807     0.5996

allele.count(a1691g, "G")   -0.0905     0.1175


                           t value Pr(>|t|)
homozygote(c104t, "C")TRUE    4.46  2.3e-05 ***


---
Signif. codes:  0 `***' 0.001 `**' 0.01
                `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.1 on 95 degrees of
                         freedom
Multiple R-Squared: 0.176,
Adjusted R-squared: 0.141
F-statistic: 5.06 on 4 and 95 DF,
             p-value: 0.000969
```

Conclusion

The current release of the genetics package, 1.0.0, provides a complete set of classes and methods for handling single-locus genetic data as well as functions for computing and testing for departure from Hardy-Weinberg and linkage disequilibrium using a variety of estimators.

As noted earlier, Friedrich Leisch and I collabo-