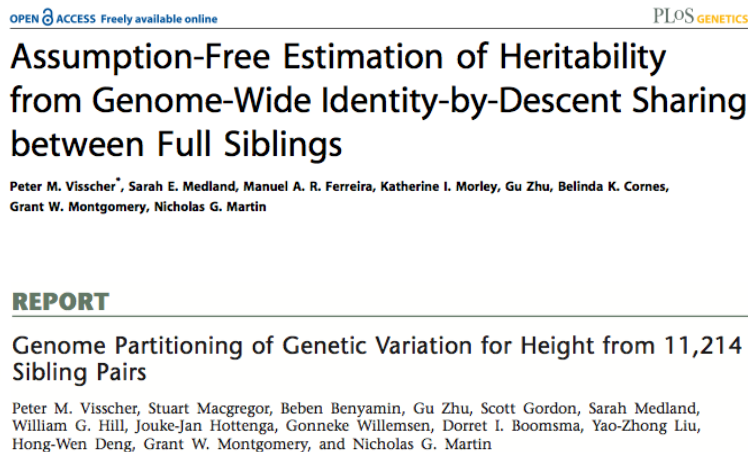


Practical 2 - Estimating genetic variance within families

In this practical we will be using the QTDT software from <http://csg.sph.umich.edu/abecasis/QTDT/> (Abecasis et al., 2000). This software was developed for family based tests of linkage disequilibrium between a marker and a complex trait (i.e., family based association mapping). We will use it to estimate variance components where our marker is the whole genome or a chromosome.

Data

The data used today is real family data and was used in the following studies Visscher et al. (2006) and Visscher et al. (2007). If you are interested take a quick look at these papers and look at the data and conclusions made.



QTDT software

Using QTDT

The QTDT software uses a similar command parsing system to pedstats, with options again called by a -flag system. For example, QTDT allows for customised variance-covariance matrices, with the null and variance structure stipulated with the -w and alternative -v options. Using observed marker genotypes QTDT can implement different association models using the -a option. QTDT can also deal with an identity by descent (IBD) matrix generated from other programs, which we will make use of in this practical.

The two variance component flags -w and -v are coupled with the following letter flags to implement different models. These include

- e - non-shared environment
- g - polygenic components
- a - additive major gene effect
- t - twin environment
- c - common environment

For example, the -w flag, which specifies variances under the null, coupled with the e option specifies that only environmental effects are modelled. The -v flag specifies variances under the alternative, for example, the option combination -veg models environment and polygenic effects.

To estimate heritability, QTDT requires the disabling of the association model via -a- and requires the specification of two models for variance. A typical command line execution of QTDT heritability estimation would look like

```
1 $ qtdt -d example.dat -p example.ped -a- -we -veg
```

Listing 1 Example run of heritability using QTDT

QTDT prints the summary output to the screen and saves the variance component estimates to the `regress.tbl` file.

QTDT estimates linkage via again disabling the association flag `-a-`, specifying the null and alternative as `-weg` and `-vega`, and providing an IBD file via `-i ibd_file.ibd`. A typical execution at the command line looks like

```
1 $ qtdt -d example.dat -p example.ped -i example.ibd -a- -weg -vega
```

Listing 2 Example run of linkage with QTDT

File requirements

The file requirements for this practical include the aptly named `data.txt` file that contains the pre-calculated chromosomal (1-22, X) estimates $\hat{\pi}_a$ (additive coefficient of relationship) and $\hat{\pi}_d$ (coefficient of fraternity). Let's have a look at these data using R.

```
1 > # Read data.txt file
2 > data <- read.table("practical_2/data/data.txt", header = T)
3 > dim(data)
4 11214 59
5 > # Look at the pi_hat_a for the autosomes and X chromosome
6 > data[1:2, 1:24]
7 pairID chr1_pi chr2_pi chr3_pi chr4_pi chr5_pi
8 800301112 0.3025 0.4413 0.4603 0.3322 0.4801
9 801101112 0.5184 0.3173 0.4297 0.6732 0.5803
10 chr6_pi chr7_pi chr8_pi chr9_pi chr10_pi chr11_pi
11 0.2045 0.6479 0.1674 0.3484 0.4224 0.4550
12 0.4266 0.3382 0.6611 0.4706 0.3917 0.4542
13 chr12_pi chr13_pi chr14_pi chr15_pi chr16_pi chr17_pi
14 0.2969 0.5484 0.2646 0.6484 0.7575 0.5294
15 0.4342 0.3589 0.4113 0.4557 0.7145 0.6257
16 chr18_pi chr19_pi chr20_pi chr21_pi chr22_pi chrX_pi
17 0.7745 0.6430 0.6818 0.8810 0.5376 0.4681
18 0.6315 0.8907 0.6510 0.3921 0.4344 0.7597
19 > # Investigate the pi_hat_d values for the autosomes and X chromosome
20 > data[1:2, 25:47]
21 chr1_ibd chr2_ibd chr3_ibd chr4_ibd chr5_ibd chr6_ibd
22 0.0146 0.1576 0.1428 0.0201 0.2047 0.0102
23 0.1987 0.0300 0.1771 0.5068 0.3305 0.1260
24 chr7_ibd chr8_ibd chr9_ibd chr10_ibd chr11_ibd chr12_ibd
25 0.4009 0.0318 0.0458 0.1916 0.1324 0.1090
26 0.0731 0.4637 0.2977 0.1373 0.2333 0.1679
27 chr13_ibd chr14_ibd chr15_ibd chr16_ibd chr17_ibd chr18_ibd
28 0.4508 0.0107 0.3754 0.5247 0.2466 0.5560
29 0.0037 0.0597 0.3768 0.4518 0.3782 0.3355
30 chr19_ibd chr20_ibd chr21_ibd chr22_ibd chrX_ibd
31 0.3028 0.3735 0.7689 0.1238 0.4681
32 0.7852 0.3151 0.0176 0.0309 0.7597
33 > # Take a look at the rest of the data matrix
34 > data[1:2, 48:59]
35 gw_pi gw_ibd sex_sib1 age_sib1 ht_sib1 zht_sib1
36 0.4920500 0.2361227 1 16 178.0 0.47048173
37 0.5118773 0.2498455 1 16 174.5 -0.05328332s
38 sex_sib2 age_sib2 ht_sib2 zht_sib2 sex_pair pop
39 1 16 166 -1.325284 1 2
40 1 16 185 1.518012 1 2
```

Listing 3 Contents of data.txt

In listing 3 the columns 48-59 have the following information, column 48 contains the genome-wide $\hat{\pi}_a$, column 49 contains the genome-wide $\hat{\pi}_d$, column 50 contains the sex (1 = male), column 51 contains the age (years), column 52 contains the height (cm), column 53 contains the z-score of height, columns 54-57 contain the same as 50-53 but for sibling 2, column 58 contains the sex code for the sibling pair, and column 59 contains the country code.

For the $\hat{\pi}_a$ estimates on chromosome 1-22, the estimates are equal to the mean probability of $\pi_a = k_1/2 + k_2$, where k_1 and k_2 are the IBD probabilities for sharing 1 or 2 alleles, across all markers. For the X sex chromosome the expected π_a estimates are 1/2 for male-male pairs, 3/4 for female-female pairs, and 1/4 for male-female pairs. The $\hat{\pi}_d$ estimates for all chromosomes, including the sex chromosome, are equal to the mean probability of $\pi_d = k_2$ across all markers on that chromosome. Further outline of the can be found in the lectures or (Lynch et al., 1998) chapter 7.

Exercise 1

Calculate the sample mean and SD of $\hat{\pi}_a$ and $\hat{\pi}_d$

- for each autosome
- genome-wide
- Do a histogram of the genome-wide $\hat{\pi}_a$ and $\hat{\pi}_d$ (row means) and compare the mean and standard to that in the lectures
- plot genome-wide $\hat{\pi}_a$ against $\hat{\pi}_d$ for each sibling pair
- regress genome-wide $\hat{\pi}_d$ values on genome-wide $\hat{\pi}_a$

```

1 > # Row means over a matrix can be calculated with the below function
2 > rowMeans(data[, 2:23])
3 > # A similar function does not exist for row standard deviations and thus we
4 > # must appeal to the apply function
5 > apply(data[, 2:23], 1, sd)
6 > pihat.mean <- rowMeans(data[, 2:23])
7 > pihat.sd <- apply(data[, 2:23], 1, sd)
8 > ibd.mean <- rowMeans(data[, 25:46])
9 > ibd.sd <- apply(data[, 25:46], 1, sd)
10 > plot(pihat.mean, ibd.mean, pch = 20,
11       col = 4, ylab = "Dominance relationship (mean IBD2 sharing)", xlab = "Additive relationship (mean IBD
       sharing)")
12 > # Regress IBD2 on pi-hat. Use the lm function un R. Try ?lm if you are interested
13 > reg <- lm(ibd.mean ~ pihat.mean)
14 > abline(reg, lwd = 2.5)

```

Listing 4 Hints for exercise 1

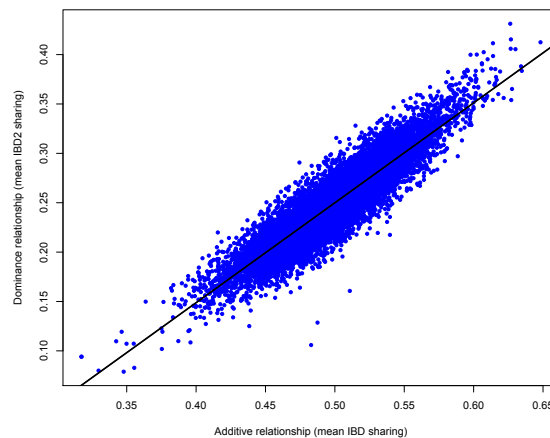


Figure 1 Plot of genome-wide $\hat{\pi}_a$ versus genome-wide $\hat{\pi}_d$. This highlights the dependence between genome wide π_d and π_a .

```

1 > # Regress IBD on pi-hat. Use the lm function in R. Try ?lm if you are interested
2 > summary(lm(ibd.mean ~ pihat.mean))
3 Call:
4 lm(formula = ibd.mean ~ pihat.mean)
5
6 Residuals:
7     Min       1Q   Median       3Q      Max
8 -0.126745 -0.011290  0.000044  0.011251  0.066568
9
10 Coefficients:
11             Estimate Std. Error t value Pr(>|t|)
12 (Intercept) -0.256023   0.002184  -117.2  <2e-16 ***
13 pihat.mean   1.011945   0.004362   232.0  <2e-16 ***
14 ---
15 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
16
17 Residual standard error: 0.01709 on 11212 degrees of freedom
18 Multiple R-squared:  0.8276, Adjusted R-squared:  0.8276
19 F-statistic: 5.382e+04 on 1 and 11212 DF,  p-value: < 2.2e-16

```

Listing 5 Hints for exercise 1 cont.

Quick look at QTDT .ped .dat files

Let's read into R the .ped and .dat files that we will use in the following exercises and see what they consist of.

```

1 > # Read in .ped and .dat files
2 > ped <- read.table("practical_2/data/qttdt.ped")
3 > dat <- read.table("practical_2/data/qttdt.dat")
4 > ibd <- read.table("practical_2/data/qttdt.ibd")
5 > # The head of the .ped file. Two columns have been cut to fit in the listing
6 > head(ped)
7  1  1  0  0  1 X  X  X 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2
8  1  2  0  0  2 X  X  X 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2
9  1  3  1  2  1 0.47 1 16. 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1
10 1  4  1  2  1 -1.33 1 16. 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2 2/2
11 2  1  0  0  1 X  X  X 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2
12 2  2  0  0  2 X  X  X 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2
13 > head(dat)
14   T   Y
15   C SEX
16   C AGE
17 S2  C1
18 S2  C2
19 S2  C3
20 S2  C4
21 S2  C5
22 S2  C6
23 S2  C7
24 S2  C8
25 S2  C9
26 S2 C10
27 S2 C11
28 S2 C12
29 S2 C13
30 S2 C14
31 S2 C15
32 S2 C16
33 S2 C17
34 S2 C18
35 S2 C19
36 S2 C20
37 S2 C21
38 S2 C22
39   M   G
40 > head(ibd)
41  V1 V2 V3 V4      V5      V6      V7
42  1  3  4 C1 0.4096 0.5758 0.0146
43  1  3  4 C2 0.2750 0.5674 0.1576
44  1  3  4 C3 0.2222 0.6350 0.1428
45  1  3  4 C4 0.3557 0.6242 0.0201
46  1  3  4 C5 0.2445 0.5508 0.2047
47  1  3  4 C6 0.6012 0.3886 0.0102

```

Listing 6 QTDT .ped, .dat, and .ibd files

In listing 6 the first five columns of the .ped file contain a pedigree similar to the pedstats format from practical 1. The next three columns of the .ped file contain the information specified in the first three rows of the .dat file. These include the trait - Y, covariate 1 - sex, and covariate 2 - age. The following columns of the .ped contain dummy information for each of the chromosomes 1-22. The s2 command tells QTDT to skip the next 2 columns. The final row specifies that we want to model M the whole genome G.

The columns of the .ibd file contain family, person 1, person 2, marker, and the last three columns are the probabilities that person 1 and person 2 share 0, 1, or 2 alleles IBD at the marker locus. The dummy variables in the .ped file are adequate as we only use these as flags and in combination with the .ibd file.

Exercise 2

- Using R, estimate the sibling correlation for the standardised z-scores (hint: zht_sib1 and zht_sib2)
- Why should we use the z-scores?

IMPORTANT- before executing any binaries at the command line please export the path to your binaries by using the below line

```
1 export PATH=$PATH:~/Desktop/SISG_AQG_2015/bin
```

Listing 7 Hints for exercises 2 and 3

Exercise 3

- Estimate the heritability using QTDT and .dat and .ped files
- Estimate the additive variance from genome wide $\hat{\pi}_a$.

This will require the use of the following QTDT commands

```
1 qtdt -d practical_2/data/qtdt.dat -p practical_2/data/qtdt.ped -a- -we -veg
2 qtdt -d practical_2/data/qtdt.dat -p practical_2/data/qtdt.ped -i practical_2/data/qtdt.ibd
3 -a- -weg -vega
```

Listing 8 Hints for exercises 2 and 3

- Estimate the additive variance for a specific chromosome
- Need to edit the .dat file and change one S2 to an M!

References

- GR Abecasis, LR Cardon, and WOC Cookson. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66(1):279–292, 2000.
- Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.
- Peter M Visscher, Sarah E Medland, MA Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*, 2(3):e41, 2006.
- Peter M Visscher, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G Hill, Jouke-Jan Hottenga, Gonneke Willemsen, Dorret I Boomsma, et al. Genome partitioning of genetic variation for height from 11,214 sibling pairs. *The American Journal of Human Genetics*, 81(5):1104–1110, 2007.