Summary of Decisions for Analysis of Minnesota Clinical GEDI Phenotypes
May 2011

Analysis 1 – GWAS of the 5 clinical phenotypes in the total sample

1. **Phenotypes (missing = -99.0 in file)**
    a. NIC_CON
    b. ALC_CON
    c. ALC_DEP
    d. DRG_FAC
    e. BD_FAC
2. **Analysis 1 Covariates**
    a. Sex (1=Male, 2=Female)
    b. Age (in years)
    c. Generation (1=Offspring, 2=Parent)
    d. Birth Year (BYR) (Actual birth year – 1900)
    e. Generation * Age
    f. Generation * Sex
    g. Generation * BYR
    h. 10 Principal Components
3. **Analysis 1 Sample (Start with the genotyped sample of N=8405)**
    a. Step #1 – eliminate those without clinical phenotype data, N=559
    b. Step #2 – eliminate those who are not 'white', N=611
    c. Step #3 – eliminate those missing 5 or more (out of 17) components used to make the clinical phenotype, N=47
    d. Result – a sample of 7188 individuals. This includes 3336 (46.4%) in offspring generation and 3852 (53.6%) in parent generation; 3328 (46.3%) male and 3860 (53.7%) female.
4. **Analysis #1 - Univariate Analysis**
    a. RFGLS for each of the 5 clinical phenotypes. In the RFGLS analysis there will be two df associated with the SNP effect, the first being the additive main effect and the second being the interaction of the additive effect with generation. We would like p-values both for the combined 2df test of the SNP effect as well as the 1 df interaction test.
    b. VEGAS for each of the 5 clinical phenotypes
    c. Candidate SNPs for each of the 5 clinical phenotypes
5. **Analysis #1 - Multivariate Analysis**
    a. O'Brien test for the 5 clinical phenotypes
    b. VEGAS based on O'Brien
    c. Candidate SNPs based on O'Brien

Analysis #2 – This analysis, low priority, is to determine the effect of covarying out the environmental index on the GWAS analysis. In this case though the sample is markedly reduced – we only have the environmental index on the twins. Also, this analysis requires doing two GWAS for each phenotype – one with covariate and one without.

1. **Phenotypes (missing = -99.0 in file)**
    a. NIC_CON
    b. ALC_CON
    c. ALC_DEP
    d. DRG_FAC
    e. BD_FAC
2. **Analysis 2 Covariates – the same as in Analysis #1, except we have added the ENVIRO covariate**
    a. Sex (1=Male, 2=Female)
    b. Age (in years)
    c. Generation (1=Offspring, 2=Parent)
    d. Birth Year (BYR) (Actual birth year – 1900)
    e. Generation * Age
    f. Generation * Sex
    g. Generation * BYR
    h. 10 Principal Components
    i. ENVIRO composite (this is a composite of the environmental data)
3. **Analysis 2 Sample (There are only up to two members per family and only MZ and DZ FTYPE)**
    a. Step #1 – Starting from 7188 above, eliminate all not having an ENV index (N=4311)
    b. Result – a sample of 2877 individuals, all twins. This includes 1330 (46.2%) male and 1547 (53.8%) female; 1878 (65.3%) in MZ families and 999 (34.7%) in DZ families.
4. **Analysis #2 - Univariate Analysis**
    a. RFGLS for each of the 5 clinical phenotypes both when ENVIRO is used as a covariate and when it is not

Data Sets

I am including two datasets. The first has all the data (N=9882); the second is only for Data Analysis #2 and so contains the N=2877 relevant individuals.

A. **Dataset #1 (file=GEDI_Phen_5_8_2011.csv)**: There are 9882 records in the file, with one record per subject. Of these, 7278 were genotyped and 1127 were MZ cotwins of someone genotyped. Consequently, the sample for which we have genotyped data has N=8405. To pick the sample for Data Analysis #1 use the GSAMPLE =1 indicator (this will give you N=7188 individuals with no missing phenotype or covariate variables). The file is comma delimited with the first record giving the variable names.

1. The file is structured such that for each of the 5 family types (see FTYPE variable) there are four consecutive records – offspring1, offspring2, mother, father (see IND variable). For the first 4 family types, the offspring are exchangeable. For the 5th family type (MIXED), the offspring are ordered such that the biological offspring is first of the four and the adopted offspring is second. The 6th family type (MISC) are single-member families, so all individuals have IND coded as one. The order of the FTYPES in the file is as follows:

FTYPE	# Families in File	# Individuals in File	Comment
MZ	1203	4812	INDIV=1,2 for offspring, 3=mom, 4=dad
DZ	655	2620	INDIV=1,2 for offspring, 3=mom, 4=dad
ADOPT	269	1076	INDIV=1,2 for offspring, 3=mom, 4=dad
BIOL	191	764	INDIV=1,2 for offspring, 3=mom, 4=dad
MIXED	111	444	INDIV=1 for bio off, 2 for adopted, 3=mom,4=dad
Total in Fams	2429	9716	This is the first 9716 records in file
MISC	166	166	All have INDIV = 1
Total		9882	

2. Variables in the file are as follows
    a. ID – 7 digit individual ID, first 5 digits are family number
    b. FTYPE - Family Type (1=MZ, 2=DZ, 3=Adopt, 4=Bio, 5=Mixed, 6=Misc; miss=-9)
    c. INDIV – Individual w/i family code (1=first offspring, 2=second offspring, 3=mother, 4=father; no missing data)
    d. SEX - 1=male, 2=female, missing = -9
    e. GENERATION – Offspring versus parent (1=Offspring generation, 2=Parent generation, -9=missing)
    f. AGE – age in years (missing = -99.0)
    g. BYR – Birthyear-1900 (missing = -9)
    h. WHITE – whether in white sample (1=white, 0 = not, -9 = missing)
    i. PSTAT – whether this person has phenotype data (1=yes, 0=no, -9=missing)
    j. PC1 to PC10 – 10 PCs based on white sample (missing=-99.0)
    k. NIC_FAC - Nicotine Factor (missing = -99.0)

l. CON_FAC – Alcohol consumption factor (missing = -99.0)
m. DEP_FAC – Alcohol dependence factor (missing = -99.0)
n. DRG_FAC – Illicit Drug Consumption factor (missing = -99.0)
o. BD_FAC – Behavioral Disinhibition factor (missing = -99.0)
p. EXT_FAC – Externalizing Factor (Missing=-99.0)
q. ENVIRO – Environmental Factor (missing = -99.0)
r. GSAMPLE – Indicator to identify who qualifies for the GWAS analysis. If = 1 (N=7188), then include; otherwise do not include

B. **Dataset #2 (file=GEDI_Enviro_5_8_2011.csv)**: The format is exactly the same as the first file, except that in this case there are only 2877 cases – those cases that have phenotype and environmental data. There is no missing data in this file.