ENIGMA EEG: EEG Power and peak frequency analysis plan 2015-01-23

**In this document:**

1. Traits of interest

2. Participating studies

3. Table of subjects

4. Genotyping + Imputation

5. Model of association

6. Data exchange

7. Analysis plan

8. Proposed deadline

9. Appendices with analysis details and scripts

**1. Traits of interest**

EEG derived measures of power and peak frequency -- with a focus on EEG traits that are related to brain volumes:

* EEG power in different well-known frequency bands at different locations:
  + Alpha power, occipital (avg of O1O2)
  + Alpha power, Cz
  + Beta power, Cz
  + Theta power, Cz
  + Delta power, Cz
  + Total (broadband) power, Cz
  + Alpha peak frequency, occipital (avg of O1O2)

(Delta .5 – 3.5 Hz, Theta 4 – 7.5 Hz, Alpha 8 – 12.5 Hz, Beta 13 – 30 Hz, EEG alpha peak frequency. Measures derived from eyes-closed resting wakefulness. )

*As a follow-up, the analyses will be held against the brain volume GWAS results. Regarding the relation between EEG power and brain volume: Smit et al., 2012:* [*http://dx.doi.org/10.1017/thg.2012.6*](http://dx.doi.org/10.1017/thg.2012.6)

**2. Participating**

SUNY COGA

Minnesota MTFS

Netherlands Twin Registry NTR

QIMR BATS sample

...

**3. Projected number of subjects:**

*(next page)*



PLUS 130 schizophrenia patients and ca. 300 controls, further specifics unknown.

**4. Genotyping + Imputation**

* Genotyped SNPs (Affymetrix, Illumina, ?)
* Imputation using the 1000 genomes imputation data for all ancestries as outlined in the document on the ENIGMA web site/as sent around.
* Males and females should be analyzed together.
* Use lenient filtering of SNP quality to reduce results-file size, yet allowing further filtering at the meta-analytic stage.
  + Remove NA results
  + Imputation quality (R^2 > 0.3 OR INFO > 0.4 && INFO < 1.02)
  + HWE p>10-5
  + MAF 0.001

**5. Model of association**

**(see appendix for data preparation and scripts)**

* **Additive** model (SNP coded as allele dosage from 0 to 2), which accounts for genotype imputation uncertainty
* **PLINK or Merlin or RFGLS** adjustments for family relatedness/pedigree. The use of Merlin and PLINK’s implementation of Generalized Estimating Equations allows the use of multiple (longitudinal/repeated) observations, complex family designs (including MZ twins).
* **Linear Regression** onto estimated dose adjusting for population structure and covariates.

**6. Data exchange**

* See separate RESULTS\_FORMAT file for details of results file formatting and file naming. Apply coarse filtering to the data to reduce results file size,
* Only summary statistics will be transferred, not individual level genotype or phenotype data. When you are ready to upload the data, please contact Dirk Smit ([d.j.a.smit@vu.nl](mailto:d.j.a.smit@vu.nl)) for the necessary details.

**7. Analysis outline**

*Covariates*

* First 10 PCs to be used as covariates. Use for example EIGENSTRAT.
* Age and Sex covariates NOTE: use your usual best age covariate that can handle nonlinearities that are known to exist for many of the phenotypes. This may be age group for samples homogenous separate groups. This can also be age+age^2 for wide age range adult groups. Preferred is to use log(age) + log(age)^2 for extended ranges and/or when ages under 25 are included. Provide the statistics of a regression analysis of single predictor tests (one df tests for continuous predictors, multiple df tests for age group factors). Include standardized Use unrelated subjects or SPSS GEE for this.
* Covariates coding for chip used (when applicable). Only applicable when multiple platforms were used.
* Disease status covariates (if applicable)

*Subject exclusion criteria*

Subjects should be excluded in the case of: same-day alcohol, marijuana, illicit drug use; same-day medication use likely to affect EEG (e.g., methylphenidate); history of severe head trauma, >24hr unconsciousness, or brain tumor. Data should be scrutinized for sleep status of the subject during recording and subjects excluded if there is evidence for sleep EEG. See SM Malone et al. for details (*Psychophysiology, 51* (2014), 1225–1245.  DOI: 10.1111/psyp.12344 )

*Missing values and data inclusion*

No need to select one from MZ twins. Take dependency into account by specifically modeling the repeated measures in Merlin, or using GEE as in PLINK. Note that it may be necessary to select the most recent of multiple measures for longitudinal measurements, fr example, when using PLINK it may not be feasible to have multiple measurements per person.

* Do not standardize scores. Make sure to have the correct scale of the variables (see Appendix 1).
* Run the analysis as specified above in PLINK/MERLIN/RFGLS on Delta, Theta, Alpha, Beta, Broadband, and Alpha Peak Frequency.

**8. Meta-analysis**

* Meta-analysis will be performed with METAL.

**9. APPENDICES**

**Appendix: Dependent variables**

*Techniques used to extract power:*

*See script in the appendix below for details*

There are many methods to extract power from signal traces, but FFT is probably the most widely used and very fast. The choice of methods seems arbitrary (r>.999 comparing Welch periodogram and FFT, both Hanning windows with 50% overlap). The scripts provided will be based on FFT. Epochs are 2 seconds. Zero padding to 4 s for peak frequency.

*Post-processing power scores*

The algorithm specified in the MATLAB script below has the following features:

* Take the average across power for frequencies that fall within the range (e.g., for alpha power all power values between 8 to 12.5 Hz inclusive). Then it uses the natural log transform (ln[power]) before averaging across epochs. Log transform of the data creates close-to-normal data.
* For peak frequency, no transform is applied. Note that these are calculated on ZERO PADDED DATA. This means that fft is applied to data stretches of 4 seconds, but the last 2 seconds are zeros. This follows the procedure as applied in (Malone et al., 2014).

*Technique used to extract peak frequency:*

Alpha peak frequency determined by taking the full signal’s FFT, and determining the peak frequency using the weighted method. Let P be the power spectrum

Then IAF = , using a fixed-window spanning 7 to 14 Hz.

For details, see:

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, *29*(2), 169-195.

*Reference and channel selection:*

The signal from Cz will be referenced to the average of the two ears (A1 A2), mastoids or nose reference are acceptable if there are no earlobe channels. O1 and O2 are bipolar derivations, to be referenced to P7 and P8, respectively. Because they are likely to be highly correlated (cf. Malone et al., 2014), scores between O1-P7 and O2-P8 leads can be averaged to compute a unitary “occipital channel.”

*Apparatus:*

Although apparatus for EEG are calibrated to measure exactly in uV, (hardware) filtering characteristics of the whole setup may be imperfectly matched. GWA analyses should take apparatus/setup as covariate if multiple setups were used for a combined GWA study set.

*Data cleaning:*

We assume that eyes-closed resting datasets have been cleaned with the following steps

* (Visually) reviewed for suspected artifacts.
* Bad channels removed and NOT imputed. Also see subject exclusion criteria above.
* Eye movements identified (removed as artifact) or removed using ICA/regression

**Appendix: Data file requirements**

Tab-delimited text file, one row per genotyped or imputed SNP, the first row with a header with the labels given below, with the requested information in the following columns:

**OUTPUT FILE FORMAT**

|  |  |  |
| --- | --- | --- |
| Column header | Description | Required format |
| SNP | SNP label for the variant in format CHR:POS beginning with “chr” | CHR:POS |
| rsID | rs number | rs number if available |
| STRAND | Orientation of the site to the human genome strand used | + or - |
| EFFECT\_ALLELE | Allele at this site to which the effect has been estimated | A, C, G, T |
| NON\_EFFECT\_ALLELE | Allele at this site which is not the EFFECT\_ALLELE | A, C, G, T |
| N | Total number of samples analyzed | numeric |
| EAF | Allele frequency of the EFFECT\_ALLELE | Frequency with 3 digits to the right of the decimal |
| BETA | Estimate of the effect size | 3 digits to the right of the decimal |
| SE | Estimated standard error on the estimate of the effect size | 4 digits to the right of the decimal |
| PVAL | Significance of the variant association, uncorrected for genomic control | 3 digits and (please use scientific E notation) |
| IMPUTED | Is the SNP imputed? | 0=genotyped, 1=imputed |
| RSQR | Imputation quality metric; (RSQ for MACH, INFO for PLINK, info |  |
| IF AVAILABLE: |  |  |
| HWE\_P | Exact HWE p-value for the sample analyzed | 4 digits to the right of the decimal |

File name should include:

– the SAMPLE name as indicate in the table

− DATE indicates the date of file generation (DDMMYYYY)

– Variable name: {“deltaCz”,”thetaCz”,“alphaCz”,”betaCz”,”broadCz”, “alphaOcc”,”peakOcc”}

File should be uploaded to the secure FTP server. For increased security on file access use encryption:

tar cz folder\_to\_encrypt | openssl enc -aes-256-cbc -e > out.tar.gz.enc

On a unix prompt. You will be prompted for a passphrase (twice). Pass the passphrase to the recipient meta-analyst.

FOR UPLOAD INSTRUCTION TO THE FTP SERVER, SEE APPENDIX BELOW

**Appendix: Descriptives**

A. Sample characteristics are based on included subjects only. Include

* Total number of subjects
* Breakdown into male/female
* Breakdown by age (for age cohort samples) or a histogram of age for large age range samples without specific age cohorts or a range and mean age for single cohort samples.
* Family composition, including number of families, number of twins (if applicable), number of siblings per family (avg), etc.

B. Measurement parameters should be provided as a short apparatus paragraph in a methods section. Likewise, provide a data cleaning section. Provide number of subjects removed for various reasons: (i) based on data cleaning and measurement issues; (ii) based on exclusion criteria (see exclusion criteria). These section will provide data on:

* Electrode layout
* Reference
* Type of electrodes
* Apparatus

C. Dependent variable descriptives:

* Mean and variance of the Power/Peak frequency scores.
* Covariate effects (including the PCs): effect of gender, effect of each age covariate used (either multiple df age group factor or continuous predictors age, log(age), age^2, and/or log(age)^2. See above.

**Appendix: Computing Principal Components**

This is a basic description of the steps needed to calculate Principal Components. If you have any questions, don’t hesitate to contact us.

Principal Component Analyses (PCA) analyses could be run using the EIGENSOFT package[1](#_ENREF_1) (<http://genepath.med.harvard.edu/~reich/Software.htm>).

We recommend to run two PCAs:

1. *PCA to identify ethnic outliers*: EIGENSOFT has an option to compute PCs in one dataset, and project those PCs onto another (see the *poplistname* parameter from the README file in the POPGEN folder of the EIGENSOFT package). This option can be used to identify subjects with a non-European ancestry, by computing PCs that reflect the global population substructure using the entire 1000 Genomes dataset, and then project those onto your own samples. Individuals with a non-European ancestry will then be able to be identified by checking whether they cluster with the European populations from 1000 Genomes. Since we imputed SNPs using all 1000 Genomes populations, we recommend to use genotyped SNPs for this PCA, as these are a better representation of the population substructure of your dataset.
2. *PCA to capture within population variation*: Once ethnic outliers are excluded, PCs can be computed that reflect the main patterns of variation in your own dataset (the PCs that will be used as covariates in the GWAS). This should be done using unrelated individuals only. If you have family data, we recommend to use GCTA[2](#_ENREF_2) to select the maximum number of unrelated individuals in your dataset by excluding one of each pair of individuals with an estimated genetic relationship of >0.025 (i.e., more related than third or fourth cousin). This can be done with the option --*grm-cutoff   0.025* (see: <http://www.complextraitgenomics.com/software/gcta/manipulation_grm.html>). Once you have a set of unrelated individuals, run a PCA analysis on these individuals, and project the PCs onto the rest of your dataset (i.e., the family members). This can be done on genotyped or imputed SNPs. In case of imputed SNPs, we recommend to use only a subset of ~1 million randomly chosen SNPs, which should result in ~200k SNPs after QC and removing LD (see paragraph below).

Some regions of the genome may be overrepresented in the PCs due to elevated levels of linkage disequilibrium (LD), diluting the genome-wide patterns that reflect ancestry differences. Very strong and/or long-range LD at a particular locus can even result in PCs that only reflect genetic variation at that specific locus[1](#_ENREF_1),[3](#_ENREF_3). We therefore recommend to exclude 24 long-range LD regions[3](#_ENREF_3) from your SNP set and prune the rest of the SNPs for LD (based on a variance inflation factor (VIF) of 2: see <http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml#prune>). These are the coordinates of the 24 long-range LD regions in build 37:

|  |  |  |
| --- | --- | --- |
| CHR | Start BP | End BP |
| 6 | 25392021 | 33392022 |
| 8 | 111930824 | 114930824 |
| 11 | 46043424 | 57243424 |
| 1 | 48287980 | 52287979 |
| 2 | 86088342 | 101041482 |
| 2 | 134666268 | 138166268 |
| 2 | 183174494 | 190174494 |
| 3 | 47524996 | 50024996 |
| 3 | 83417310 | 86917310 |
| 3 | 88917310 | 96017310 |
| 5 | 44464243 | 50464243 |
| 5 | 97972100 | 100472101 |
| 5 | 128972101 | 131972101 |
| 5 | 135472101 | 138472101 |
| 6 | 56892041 | 63942041 |
| 6 | 139958307 | 142458307 |
| 7 | 55225791 | 66555850 |
| 8 | 7962590 | 11962591 |
| 8 | 42880843 | 49837447 |
| 10 | 36959994 | 43679994 |
| 11 | 87860352 | 90860352 |
| 12 | 33108733 | 41713733 |
| 12 | 111037280 | 113537280 |
| 20 | 32536339 | 35066586 |

**References**

1 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909 (2006).

2 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* (2010).

3 Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* **83**, 132 (2008).

**Appendix: MATLAB code**

Below the MATLAB code for getting power in the defined frequency bands (cFreqNames). NOTE set the settings correctly. You must program the loop yourself, for opening the data and iterating through all the subjects (iteration variable ‘subj’). Data are stored in AllPOW (NSubjects,NChannels,NFrequencies) and AllPeakFreq(NSubjects,NChannels) for power and individual alpha peak frequency respectively.

The input data is assumed to have the same shape: columns of doubles in ascii representing uV potential of the signals. Missings should have the value nan in your text file, columnwise.

NOTE Post processing is required in selecting the right channels.

* Extract and Average scores over the occipital channels for O1 and O2 power in the alpha band **(frequency 3)**
* Extract and Average peak frequency scores for channels O1 and O2 in the alpha band **(frequency 6)**
* Extract Cz Power in the delta, theta, alpha, beta bands and broadband power **(frequencies 1, 2, 3, 4, and 5)**

Please contact [d.j.a.smit@vu.nl](mailto:d.j.a.smit@vu.nl) with any questions regarding the script.

% Define constants (use prefix c)

% Frequency settings

cFreqNames = {'Delta','Theta','Alpha','Beta',’Broadband’,’AlphaPeak’};

cFreqs{1} = [ .5 3.5]; % Delta: up to but NOT INCLUDING 4 Hz

cFreqs{2} = [ 4 7.5]; % Theta

cFreqs{3} = [ 8 12.5]; % Alpha

cFreqs{4} = [13 30 ]; % Beta

cFreqs{5} = [ .5 30 ]; % Broadband

cFreqs{6} = [ 7 14 ]; % Alpha for peak frequency

cNFreqs = 6;

% NOTE: Set these correctly!

cWinSec = 2; % use 2 second length windows

cRate = 250; % 250 was our sampling rate

cWinSize = cRate \* cWinSec; % FFT resolution will be .5 Hz

cNSubjects = 1; % set these correctly

cNChans = 12; % set these correctly

% inserted by DBC March 12 2015

% set one more to avoid repeated calculation

cNsampsWin = cWinSize\*cRate;

% get the frequencies from 0 (DC) to Nyqvist frequency

fs = linspace(0,cRate./2,cWinSize./2+1);

% get the frequencies from 0 (DC) to Nyqvist frequency FOR ZERO PADDED DATA

fs2 = linspace(0,cRate./2,cWinSize+1);

%% START LOOP

% This is where the loop (looping through subjects) would be defined. None

% has been defined here.

% Read in data. For example:

subj = 1;

D = dlmread('MyData.txt','\t');

% Data are assumed to be organized in COLUMNS, one channel per column, time

% on the rows. len = the number of samples.

len = size(D,1);

% Windows of length 'cWinSize' equally spaced. Store the starting sample

% indices in 'start' vector. ca. 50% overlapping hanning windows.

start = floor(linspace(1,len-cWinSize+1,(2.\*floor(len/cWinSize))-1));

% inserted by DBC March 12 2015

finish = start + cWinSize - 1;

% Now calculate for power

P = zeros(cWinSize,size(D,2));

for s=1:length(start)

F = fft(repmat(hanning(cWinSize),1,cNChans) .\* D(start(s):finish(s),:));

P = P + ((F.\*conj(F))./cNsampsWin); % sum first, divdide later

end

% save the average power spectrum. Also remove the redundant part with

% frequency of nyqvist and over, and to double power values from 2:end/2.

% For individual differences, it does not matter whether you do or not.

POW = P(1:end/2+1,:)./length(start);

POW(2:end-1,:) = 2\*POW(2:end-1,:);

% Now calculate for peak frequency (with zero padded signal)

P2 = zeros(cWinSize,size(D,2));

for s=1:length(start)

Temp = detrend(D(start(s):finish(s),:), 0); % detrend(,0) removes the DC

F2 = fft([Temp ; zeros(cWinSize,size(D,2))]); % ZERO PADDING

P2 = P2 + ((F2.\*conj(F2))./cNsampsWin); %

end

PW2 = P2(1:end/2+1,:)./length(start);

PW2(2:end-1,:) = 2\*POW2(2:end-1,:);

% Now loop through freqs and extract the right numbers

for freq=1:length(cFreqs)

% which to select and average?

if freq<6

freqndx = fs>=cFreqs{freq}(1) & fs<=cFreqs{freq}(2);

AllPOW(subj,:,freq) = mean(log(POW(freqndx,:)));

% average freq

else % if freq==6

freqndx = fs2>=cFreqs{freq}(1) & fs2<=cFreqs{freq}(2);

AllPeakFreq(subj,:) = sum(PW2(freqndx,:).\*repmat(fs2(freqndx)',[1,size(PW2,2),1]))./sum(PW2(freqndx,:));

end

end

**Appendix: HWE for dosage instructions**

*Step 1: Quicktest*

Download Quicktest from: <http://toby.freeshell.org/software/quicktest.shtml>

Run quicktest for your dosage file

./quicktest095 --geno Genofile.gen.gz --pheno PHENOT.sample --snptest --method-mean --compute-rSqHat --out OutputFile

*Step 2: SPSS*

Then by using the above syntax (SPSS) to calculate HWE\_EXP\_Pval  from the Quicktest output (mean for the expected genotype count for each allele calculated by averaging over the genotype uncertainty =dosage):

COMPUTE FreqA=(meanAA + meanAB\*0.5) / (meanAA+meanAB+meanBB).

EXECUTE.

COMPUTE HWE\_EXP\_AA=FreqA\*FreqA\*(meanAA+meanAB+meanBB).

COMPUTE HWE\_EXP\_AB=2\*(1-FreqA)\*FreqA\*(meanAA+meanAB+meanBB).

COMPUTE HWE\_EXP\_BB=(1-FreqA)\*(1-FreqA)\*(meanAA+meanAB+meanBB).

COMPUTE HWE\_CHI\_AA=((meanAA-HWE\_EXP\_AA)\*(meanAA-HWE\_EXP\_AA))/HWE\_EXP\_AA.

COMPUTE HWE\_CHI\_AB=((meanAB-HWE\_EXP\_AB)\*(meanAB-HWE\_EXP\_AB))/HWE\_EXP\_AB.

COMPUTE HWE\_CHI\_BB=((meanBB-HWE\_EXP\_BB)\*(meanBB-HWE\_EXP\_BB))/HWE\_EXP\_BB.

COMPUTE HWE\_CHI=HWE\_CHI\_AA+HWE\_CHI\_AB+HWE\_CHI\_BB.

EXECUTE.

COMPUTE HWE\_EXP\_Pval=(1-CDF.CHISQ(HWE\_CHI,1)).

EXECUTE.

**Appendix Uploading**

This appendix describes how to use the ftp server at ITM of the faculty FPP of the VU University Amsterdam. The server is named **sftp.psy.vu.nl**.

*Command line (LINUX)*

The protocol used is ftps rather than sftp. This requires using lftp on linux systems.

Start lftp by typing:

lftp

Then within lftp type:

set ftps:initial-prot "";

set ftp:ssl-force true;

set ftp:ssl-protect-data true;

set ftp:passive true;

set ftp:ssl-allow false;

set ssl:verify-certificate false;

open ftps://sftp.psy.vu.nl:1337

user djasmit <password-to-be-communicated-separately>

Then use the ftp commands, such as ls and mput. Use help to list the commands.

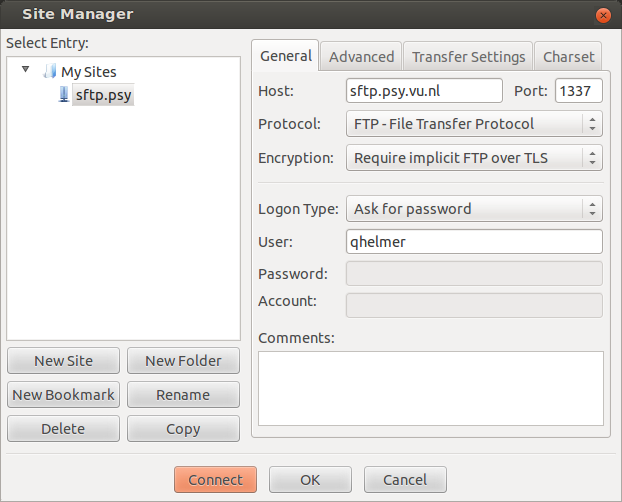
* + 1. *FTP client*

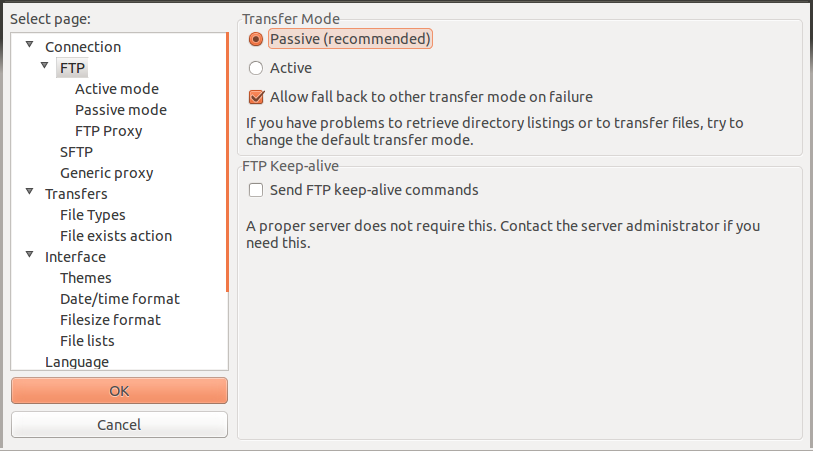
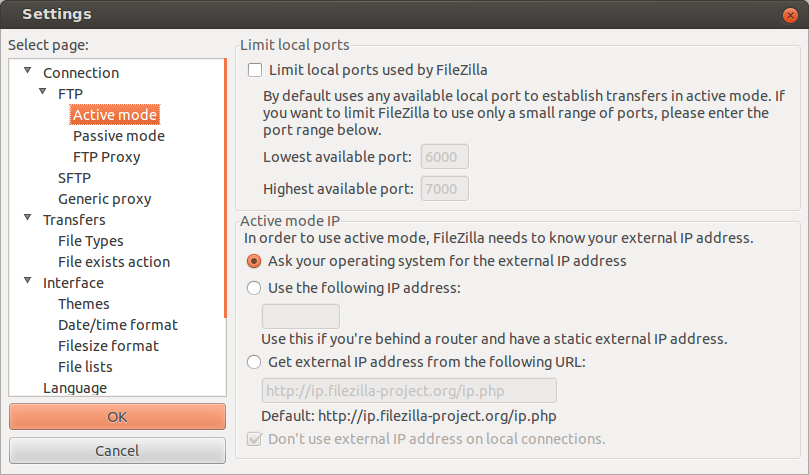
We recommend to use FileZilla (other clients are not supported, but in most clients settings are similar). It can be downloaded from <https://filezilla-project.org/download.php?type=client> for all operating systems.

* + 1. *Settings*

Open FileZilla and follow these steps to set up a connection:

1. Go to File → Site Manager and click on New Site and give the site a name (eg. sftp.psy)
2. The following settings are required on tab General:
   1. Host: sftp.psy.vu.nl
   2. Port: 1337
   3. Protocol: FTP - File Transfer Protocol
   4. Encryption: Require implicit FTP over TLS
   5. Logon Type: Ask for password
   6. User: the user name given to you by ITM or your contact person.



1. Go to tab Transfer Settings and set Transfer mode to Passive
2. Click Connect and type the password given to you by ITM or your contact person.
3. Click OK. A connection is set up and your home directory is listed.
   1. If you get the error message 'Failed to retrieve directory listing'; this can be fixed as follows
      1. Go to Edit → Settings
      2. In sub-menu FTP select Transfer Mode 'Passive'
      3. In sub-menu FTP - Active mode select 'Ask your operating system for the external IP address'
      4. In sub-menu FTP - Passive mode select 'Use the server's external IP address instead'
      5. Click OK to save settings and reconnect.