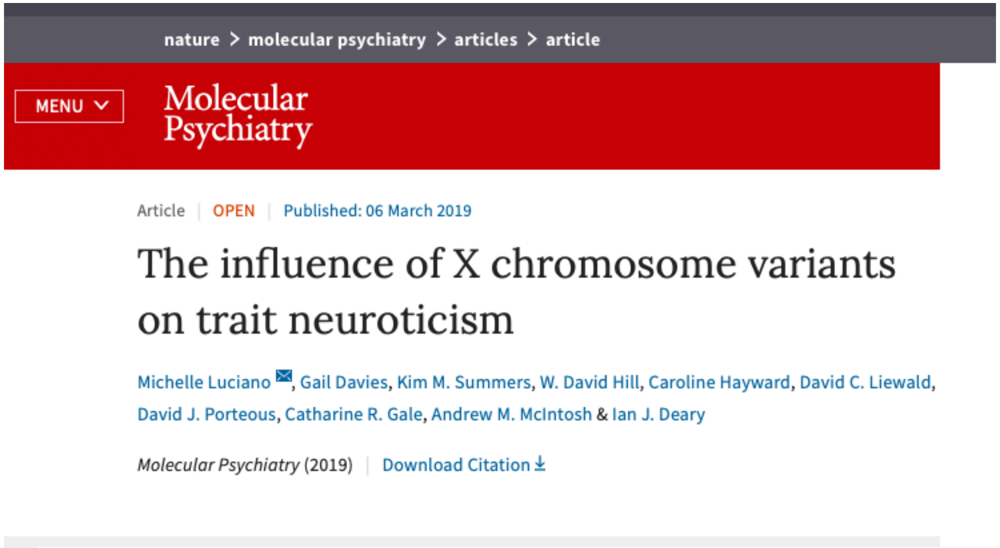
Quality Control
&
Meta-Analysis in METAL

Meike Bartels & Bart Baselmans

Multivariate GWAMA

Michel Nivard & Aysu Okbay



nature > molecular psychiatry > articles > article

MENU

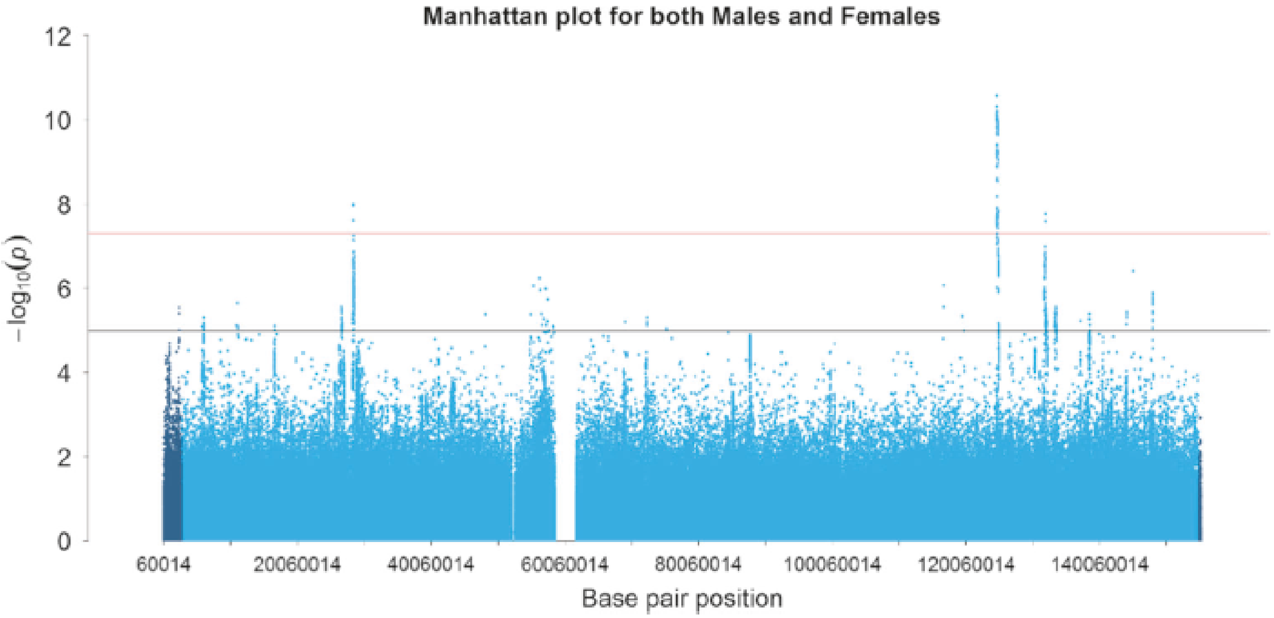
Molecular Psychiatry

Article | OPEN | Published: 06 March 2019

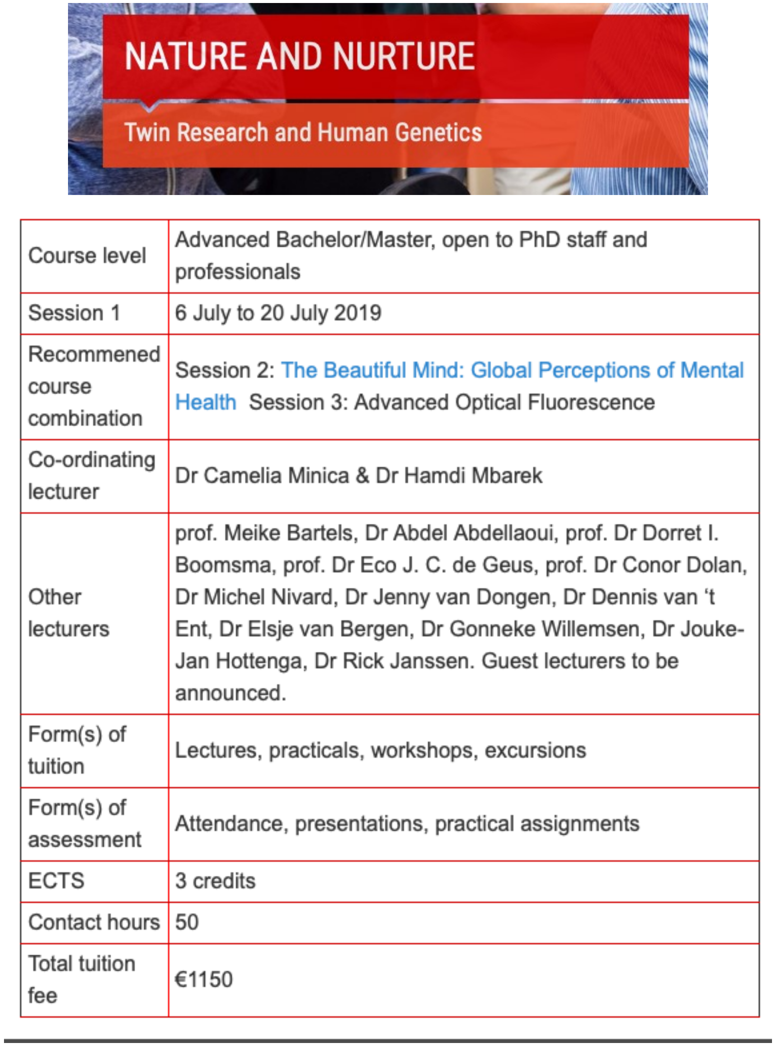
The influence of X chromosome variants on trait neuroticism

Michelle Luciano, Gail Davies, Kim M. Summers, W. David Hill, Caroline Hayward, David C. Liewald, David J. Porteous, Catharine R. Gale, Andrew M. McIntosh & Ian J. Deary

Molecular Psychiatry (2019) | Download Citation



VU Summerschool: 6 July to 20 July 2019



NATURE AND NURTURE

Twin Research and Human Genetics

Course level	Advanced Bachelor/Master, open to PhD staff and professionals
Session 1	6 July to 20 July 2019
Recommened course combination	Session 2: The Beautiful Mind: Global Perceptions of Mental Health Session 3: Advanced Optical Fluorescence
Co-ordinating lecturer	Dr Camelia Minica & Dr Hamdi Mbarek
Other lecturers	prof. Meike Bartels, Dr Abdel Abdellaoui, prof. Dr Dorret I. Boomsma, prof. Dr Eco J. C. de Geus, prof. Dr Conor Dolan, Dr Michel Nivard, Dr Jenny van Dongen, Dr Dennis van 't Ent, Dr Elsje van Bergen, Dr Gonneke Willemsen, Dr Jouke-Jan Hottenga, Dr Rick Janssen. Guest lecturers to be announced.
Form(s) of tuition	Lectures, practicals, workshops, excursions
Form(s) of assessment	Attendance, presentations, practical assignments
ECTS	3 credits
Contact hours	50
Total tuition fee	€1150

GENES IN BEHAVIOUR AND HEALTH (RESEARCH MASTER)

EXPLORING GENE-ENVIRONMENT INTERPLAY ACROSS OUR LIFESPAN

WWW.VU.NL/GENES

Interested in this unique Master programme in the Netherlands?

Visit our website or email rm.gbh.fgb@vu.nl

VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

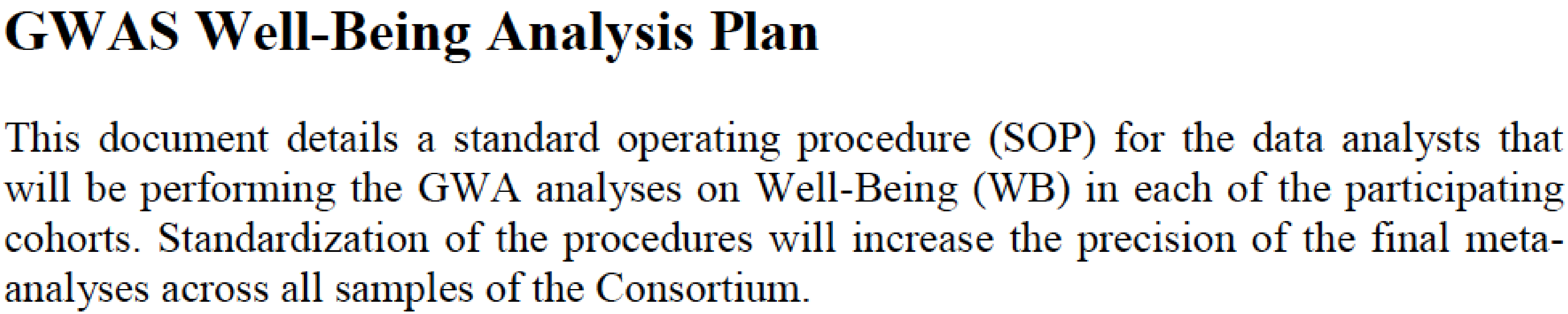
GWAMA

Genome-Wide Association Meta-Analysis

- Large collaborative studies to increase sample size
- Different cohorts run their own GWA and upload summary statistics
- To make these projects successful and trustworthy, we need rigorous organisation AND quality control (QC)

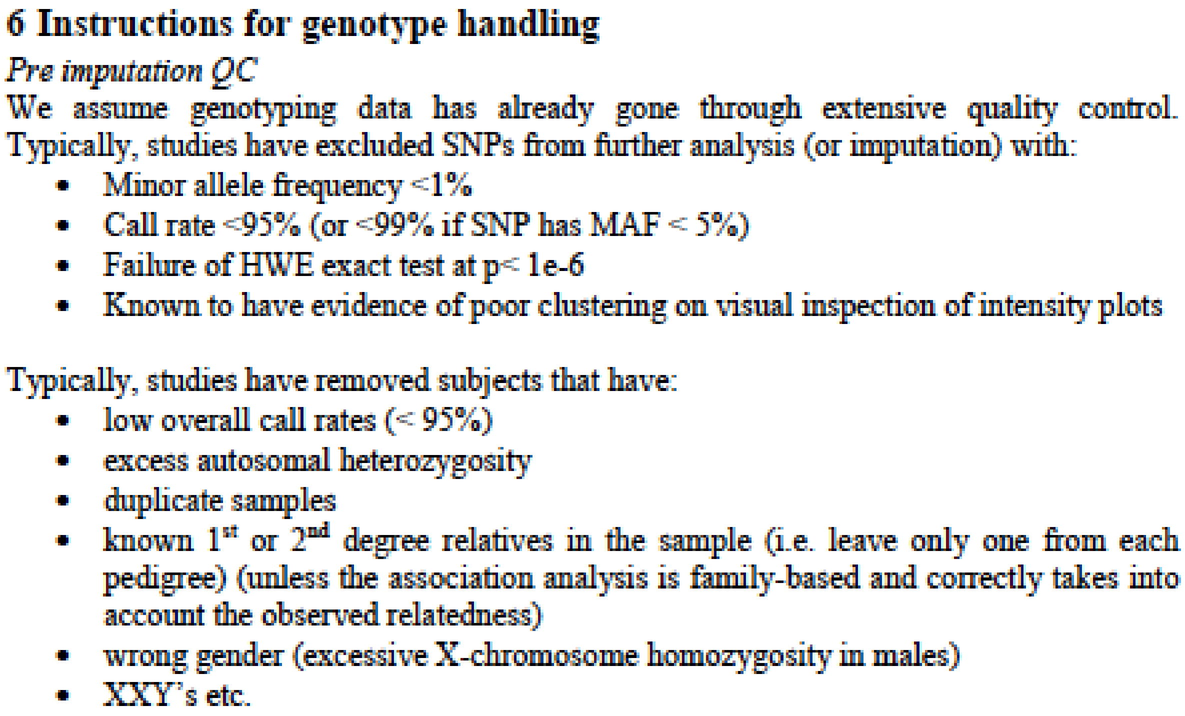
SOP

Standard Operating Procedure



GWAS Well-Being Analysis Plan

This document details a standard operating procedure (SOP) for the data analysts that will be performing the GWA analyses on Well-Being (WB) in each of the participating cohorts. Standardization of the procedures will increase the precision of the final meta-analyses across all samples of the Consortium.



6 Instructions for genotype handling

Pre imputation QC

We assume genotyping data has already gone through extensive quality control. Typically, studies have excluded SNPs from further analysis (or imputation) with:

- Minor allele frequency <1%
- Call rate <95% (or <99% if SNP has MAF < 5%)
- Failure of HWE exact test at p< 1e-6
- Known to have evidence of poor clustering on visual inspection of intensity plots

Typically, studies have removed subjects that have:

- low overall call rates (< 95%)
- excess autosomal heterozygosity
- duplicate samples
- known 1st or 2nd degree relatives in the sample (i.e. leave only one from each pedigree) (unless the association analysis is family-based and correctly takes into account the observed relatedness)
- wrong gender (excessive X-chromosome homozygosity in males)
- XXY's etc.

5 Instructions for WB phenotypes and covariate coding

Inclusion
We propose to limit the analyses to subjects with European ancestry only.

Phenotype handling
The phenotype handling instructions are different depending on whether a cohort's WB measure is based on a single survey question or a set of survey question. However, the instructions regarding covariates, below this section, are the same in both cases.

For cohorts whose WB measure is the response to a single survey question:
Run the GWA using linear regression.

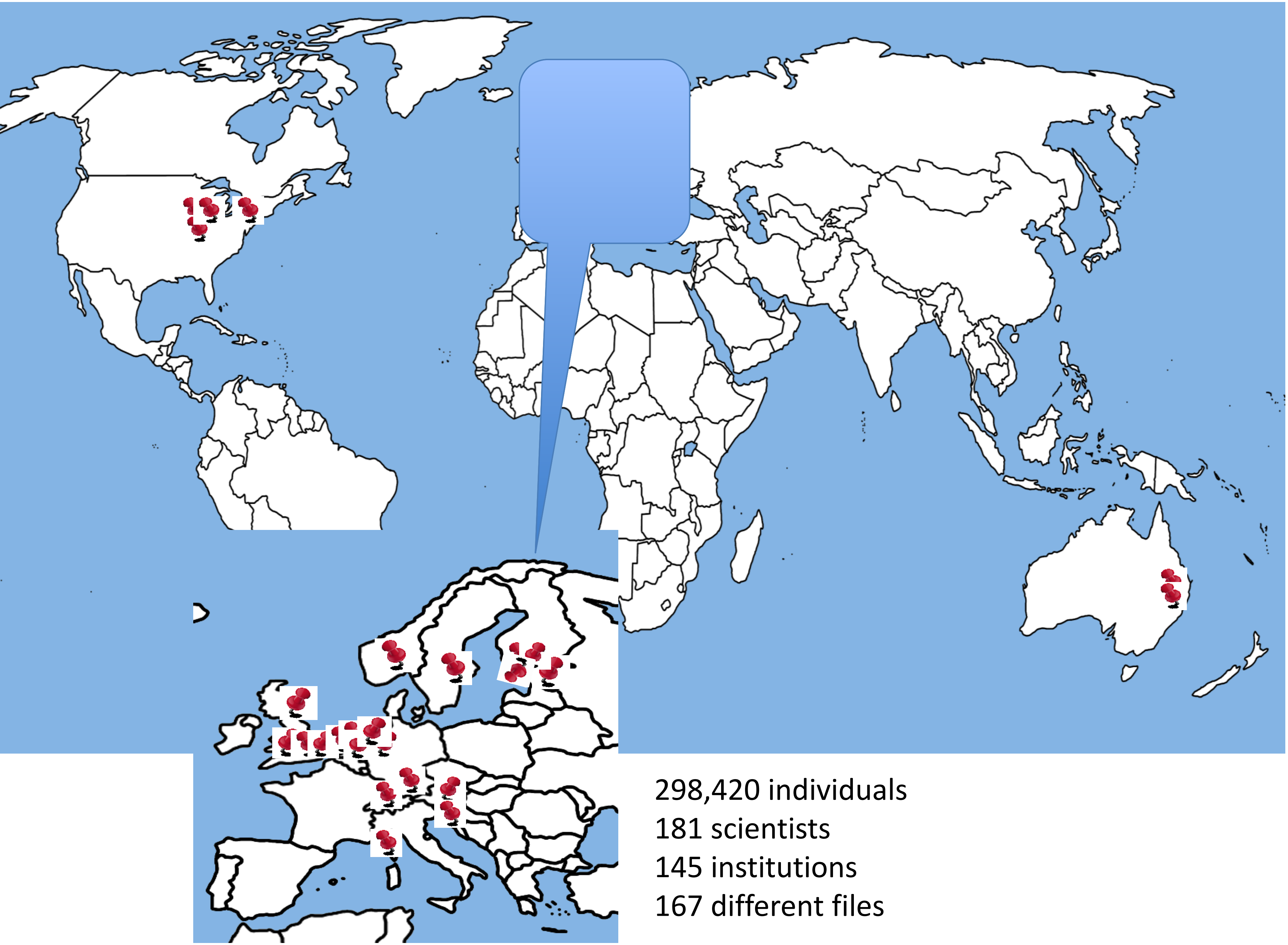
For cohorts whose WB measure is the response to multiple survey questions:
Check with us regarding how to aggregate the responses to these questions into a single WB measure. Once you have this measure, run the GWA using linear regression.

No transformation of the WB measure
Cohorts should **not** transform the WB measure prior to running GWA.[1]
However, make sure that WB is positively measured, i.e. **higher numbers = higher WB**.
Please reverse your measure if you have a scale where higher number = lower WB.

9 Instructions for reporting results from first-pass association analyses

On the joint data from all participating cohorts a meta-analysis will be performed on the study-specific association statistics. This requires each participating study to report the following characteristics for every SNP (*all* imputed and observed SNPs to be reported, i.e. no p-value cut-off, no imputation quality cut-off and no maf cut-off) in plain-text ASCII files for each phenotype separately:

Variable name (case sensitive!!)	**Description**
SNPID	SNP ID as rs number
Chr	Chromosome number (1-22).
position	physical position for the reference sequence (indicate build 35/36 in readme file)
coded_all	Coded allele, also called modelled allele (in example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G)
noncoded_all	The other allele
strand_genome	+ or -, representing either the positive/forward strand or the negative/reverse strand of the human genome reference sequence; to clarify which strand the coded_all and noncoded_all are on
Beta	Beta estimate from genotype-phenotype association, at least 5 decimal places – 'NA' if not available
SE	Standard error of beta estimate, to at least 5 decimal places – 'NA' if not available
Pval	*p*-value of test statistic, here just as a double check – 'NA' if not available
AF_coded_all	Allele frequency for the coded allele – 'NA' if not available
HWE_pval	Exact test Hardy-Weinberg equilibrium *p*-value -- only directly typed SNPs, NA for imputed
callrate	Genotyping call rate after exclusions
n_total	Total sample with phenotype and genotype for SNP



298,420 individuals
181 scientists
145 institutions
167 different files

EasyQC

- For the well-being GWAMA, we used EasyQC for QC-ing the GWAS summary statistics

- Positive experience because it provides guidelines how to perform QC at the:
  - study file level
  - meta-file level
  - meta-analysis OUTPUT level

What needs to be detected

- File name errors -> sounds simple, but with 167 files it is essential that all files can be traced back to a specific cohort
- Incorrect specification of the Phenotype
- Flipped alleles
- Duplicated SNPs
- Bad imputation quality
- Association issues from incorrect analysis models
  - Population stratification
  - Unaccounted relatedness of individuals

These errors

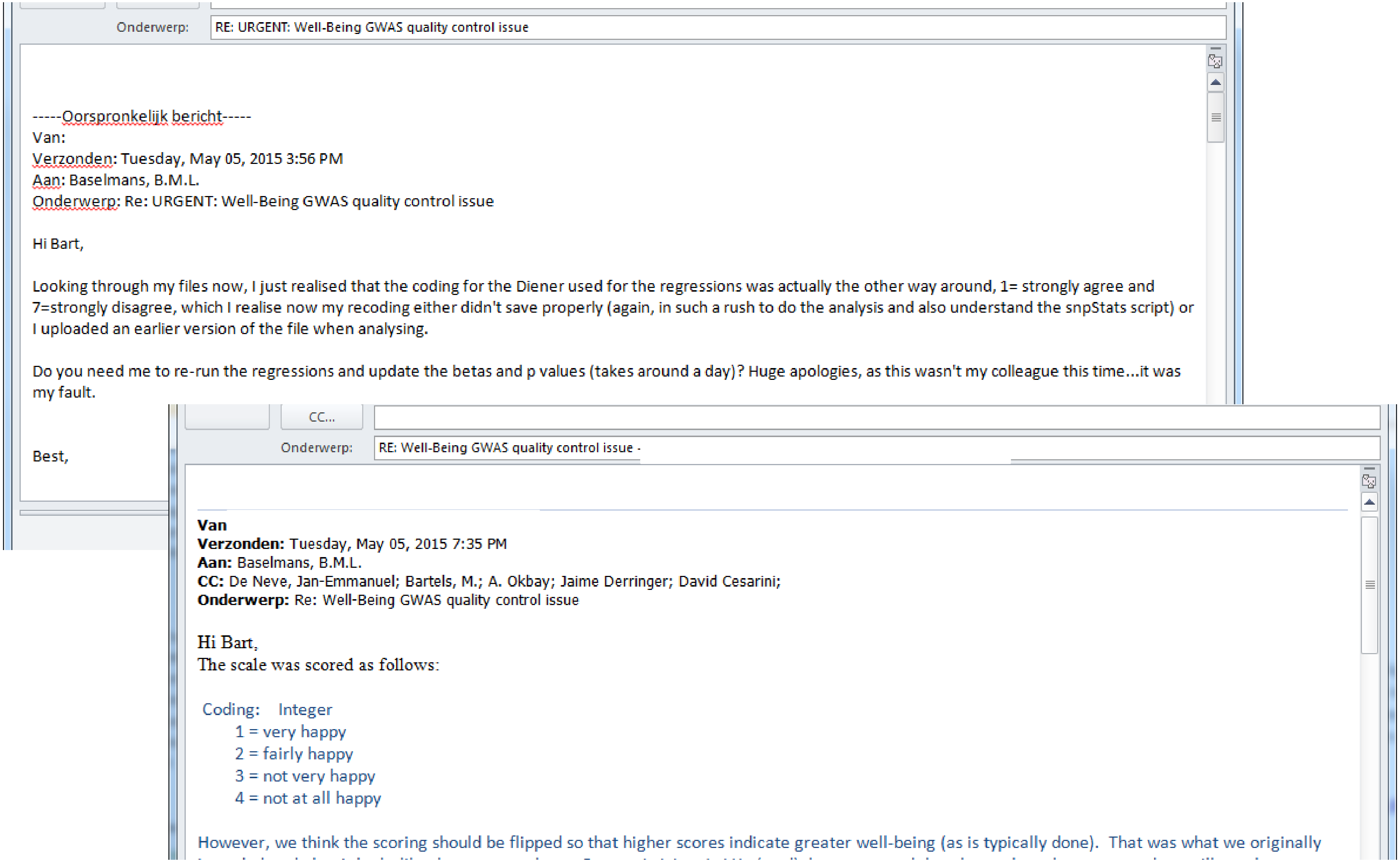
- Limit the contribution of a specific cohort to the meta-analysis

or

- Inflate the number of inflate the number of false positive

Descriptive Summary Statistics

- Participating cohorts were asked to complete a **descriptive statistics summary** file for their sample
  - **8 cohorts** did not specify their question, or did not report the distribution of the question (no categories specified)
  - **3 cohorts** gave **lower values to higher wellbeing** (reversed coding)
  - **8 cohorts** did not map the categories to numeric values, but where the first option was higher wellbeing (suspicious reversed coding)



Onderwerp: RE: URGENT: Well-Being GWAS quality control issue

-----Oorspronkelijk bericht-----
Van:
Verzonden: Tuesday, May 05, 2015 3:56 PM
Aan: Baselmans, B.M.L.
Onderwerp: Re: URGENT: Well-Being GWAS quality control issue

Hi Bart,

Looking through my files now, I just realised that the coding for the Diener used for the regressions was actually the other way around, 1= strongly agree and 7=strongly disagree, which I realise now my recoding either didn't save properly (again, in such a rush to do the analysis and also understand the snpStats script) or I uploaded an earlier version of the file when analysing.

Do you need me to re-run the regressions and update the betas and p values (takes around a day)? Huge apologies, as this wasn't my colleague this time...it was my fault.

Best,

Onderwerp: RE: Well-Being GWAS quality control issue -

Van
Verzonden: Tuesday, May 05, 2015 7:35 PM
Aan: Baselmans, B.M.L.
CC: De Neve, Jan-Emmanuel; Bartels, M.; A. Okbay; Jaime Derringer; David Cesarini;
Onderwerp: Re: Well-Being GWAS quality control issue

Hi Bart,
The scale was scored as follows:

Coding: Integer
1 = very happy
2 = fairly happy
3 = not very happy
4 = not at all happy

However, we think the scoring should be flipped so that higher scores indicate greater well-being (as is typically done). That was what we originally

6 Instructions for genotype handling

Pre imputation QC

We assume genotyping data has already gone through extensive quality control. Typically, studies have excluded SNPs from further analysis (or imputation) with:

- Minor allele frequency <1%
- Call rate <95% (or <99% if SNP has MAF < 5%)
- Failure of HWE exact test at p< 1e-6
- Known to have evidence of poor clustering on visual inspection of intensity plots

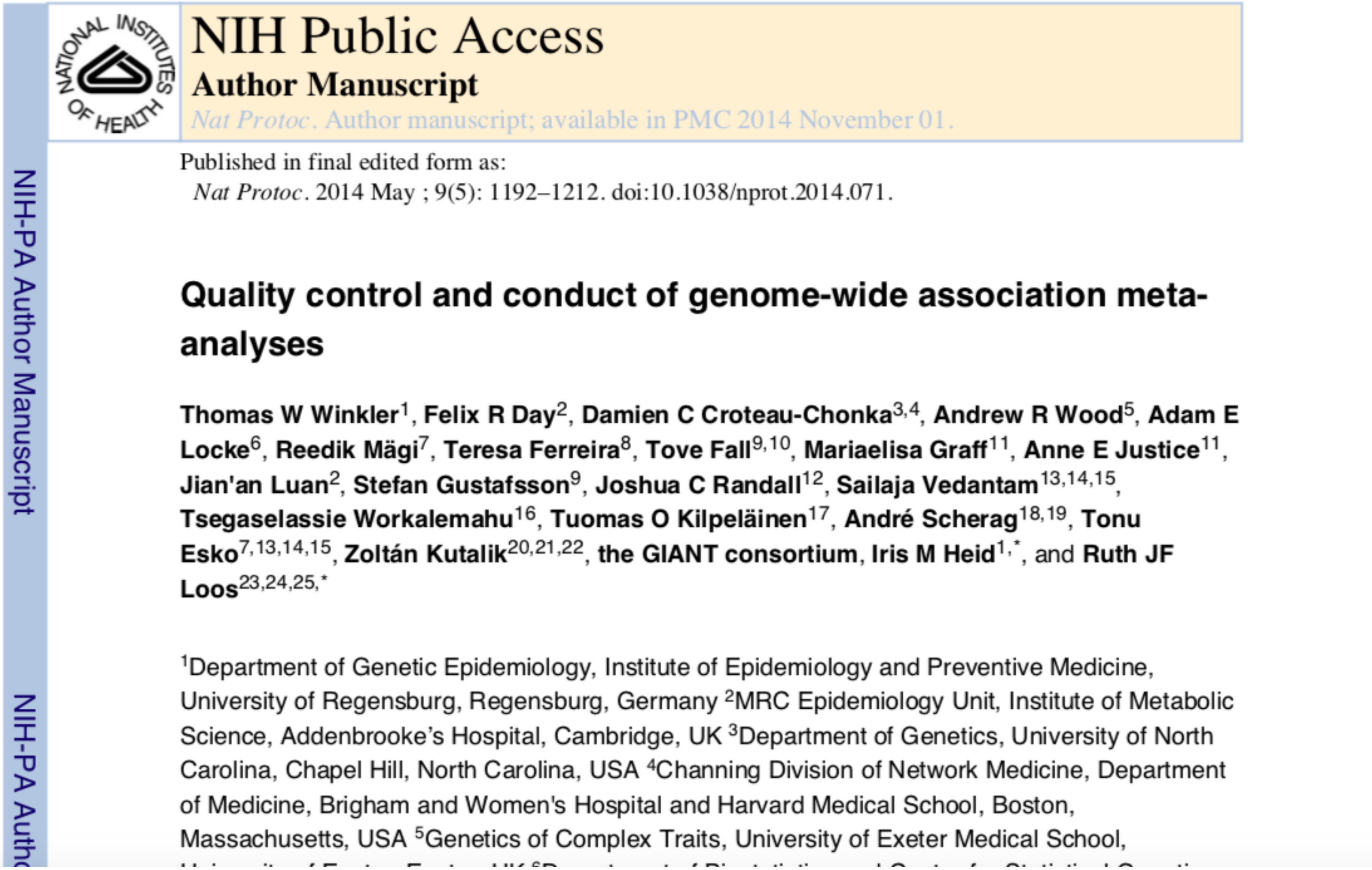
Typically, studies have removed subjects that have:

- low overall call rates (< 95%)
- excess autosomal heterozygosity
- duplicate samples
- known 1st or 2nd degree relatives in the sample (i.e. leave only one from each pedigree) (unless the association analysis is family-based and correctly takes into account the observed relatedness)
- wrong gender (excessive X-chromosome homozygosity in males)
- XXY's etc.

nature protocols

Protocol | Published: 24 April 2014

Quality control and conduct of genome-wide association meta-analyses

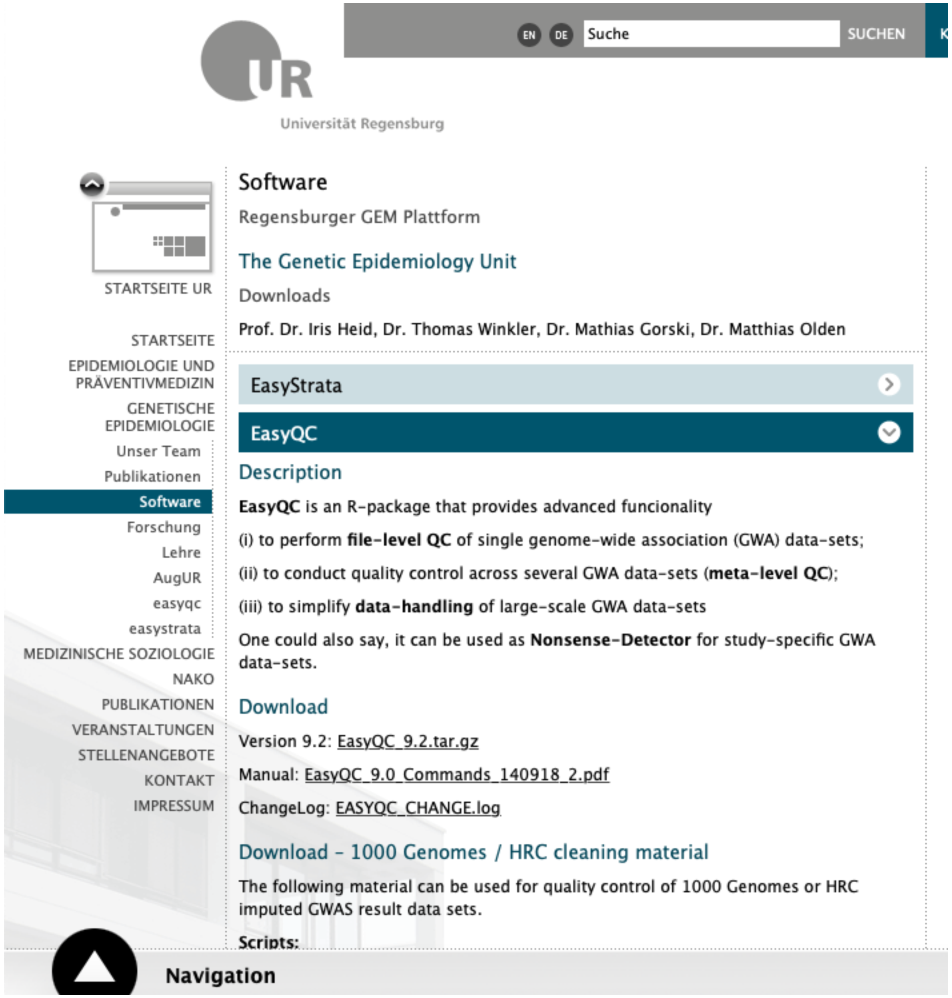


Published in final edited form as:
Nat Protoc. 2014 May ; 9(5): 1192–1212. doi:10.1038/nprot.2014.071.

Quality control and conduct of genome-wide association meta-analyses

Thomas W Winkler[1], **Felix R Day**[2], **Damien C Croteau-Chonka**[3,4], **Andrew R Wood**[5], **Adam E Locke**[6], **Reedik Mägi**[7], **Teresa Ferreira**[8], **Tove Fall**[9,10], **Mariaelisa Graff**[11], **Anne E Justice**[11], **Jian'an Luan**[2], **Stefan Gustafsson**[9], **Joshua C Randall**[12], **Sailaja Vedantam**[13,14,15], **Tsegaselassie Workalemahu**[16], **Tuomas O Kilpeläinen**[17], **André Scherag**[18,19], **Tonu Esko**[7,13,14,15], **Zoltán Kutalik**[20,21,22], **the GIANT consortium**, **Iris M Heid**[1,*], and **Ruth JF Loos**[23,24,25,*]

[1]Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany [2]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK [3]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA [4]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA [5]Genetics of Complex Traits, University of Exeter Medical School,



Software

Regensburger GEM Plattform

The Genetic Epidemiology Unit

Downloads

Prof. Dr. Iris Heid, Dr. Thomas Winkler, Dr. Mathias Gorski, Dr. Matthias Olden

EasyStrata

EasyQC

Description

EasyQC is an R-package that provides advanced funcionality

(i) to perform **file-level QC** of single genome-wide association (GWA) data-sets;

(ii) to conduct quality control across several GWA data-sets (**meta-level QC**);

(iii) to simplify **data-handling** of large-scale GWA data-sets

One could also say, it can be used as **Nonsense-Detector** for study-specific GWA data-sets.

Download

Version 9.2: EasyQC_9.2.tar.gz

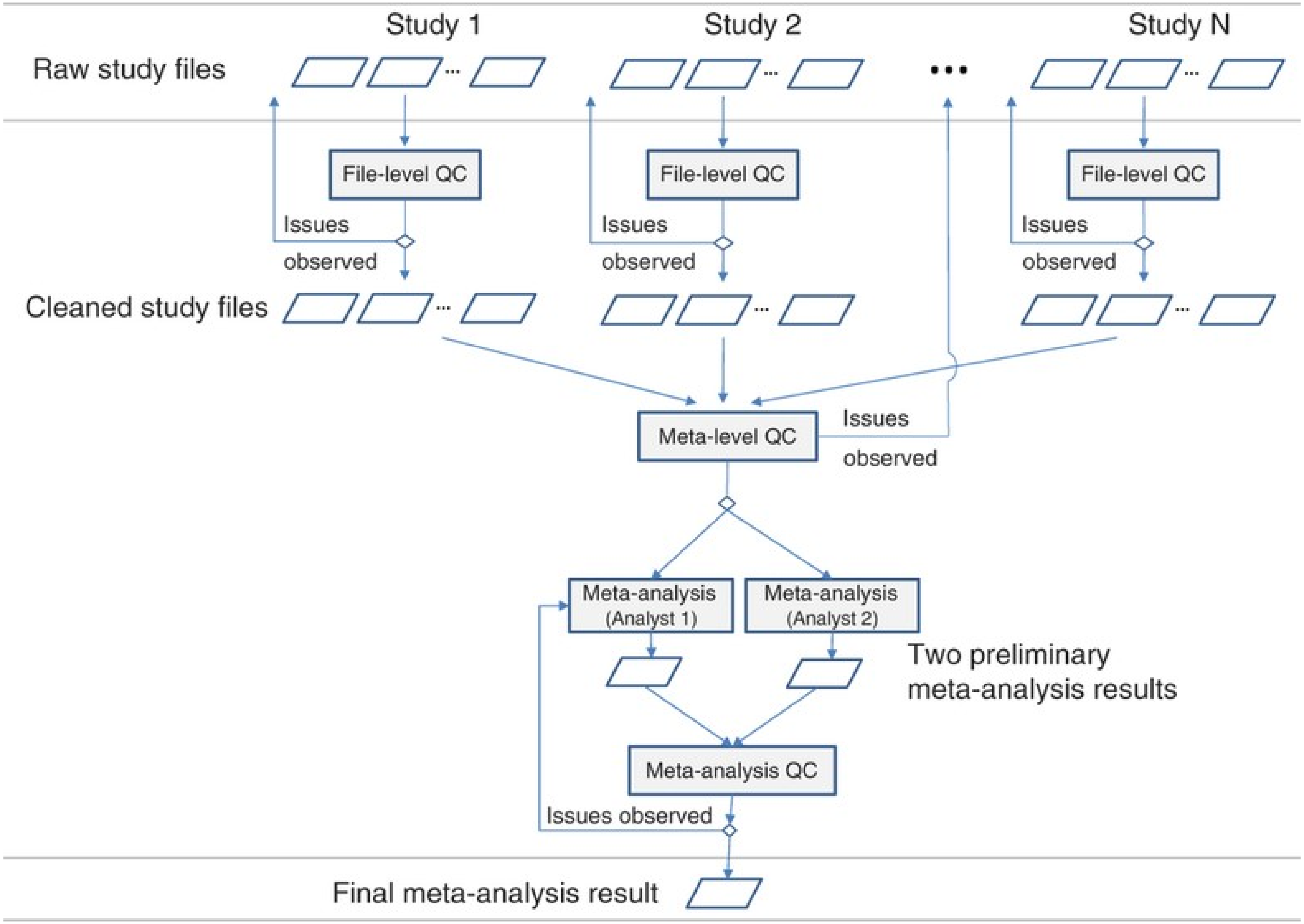
Manual: EasyQC_9.0_Commands_140918_2.pdf

ChangeLog: EASYQC_CHANGE.log

Download - 1000 Genomes / HRC cleaning material

The following material can be used for quality control of 1000 Genomes or HRC imputed GWAS result data sets.

Scripts:


Study 1
Study 2
Study N
Raw study files
• • •
File-level QC
File-level QC
File-level QC
Issues observed
Issues observed
Issues observed
Cleaned study files
Meta-level QC
Issues observed
Meta-analysis (Analyst 1)
Meta-analysis (Analyst 2)
Two preliminary meta-analysis results
Meta-analysis QC
Issues observed
Final meta-analysis result

QC workflow- step 1

Step 1: File level QC

- This stage involves cleaning of the data
  - Deleting poor quality data
  - Provide summaries to judge data quality

Think about

- Monomorphic SNPs
- Missingness (e.g. P-values, Beta's, SE's and more)
- Nonsensical information (e.g Alleles other than A, C, G or T, or p-values larger than 1 or smaller than 0 etc)
- Low number of individuals per SNP (GIANT < 30)
- Harmonization of SNP identifiers using maps with unique SNP identifiers and genomic positions

QC workflow – step 2

Step 2: Meta level QC

- This stage consists of the cross-study comparison of statistics and comparison to reference panels to identify study specific problems.

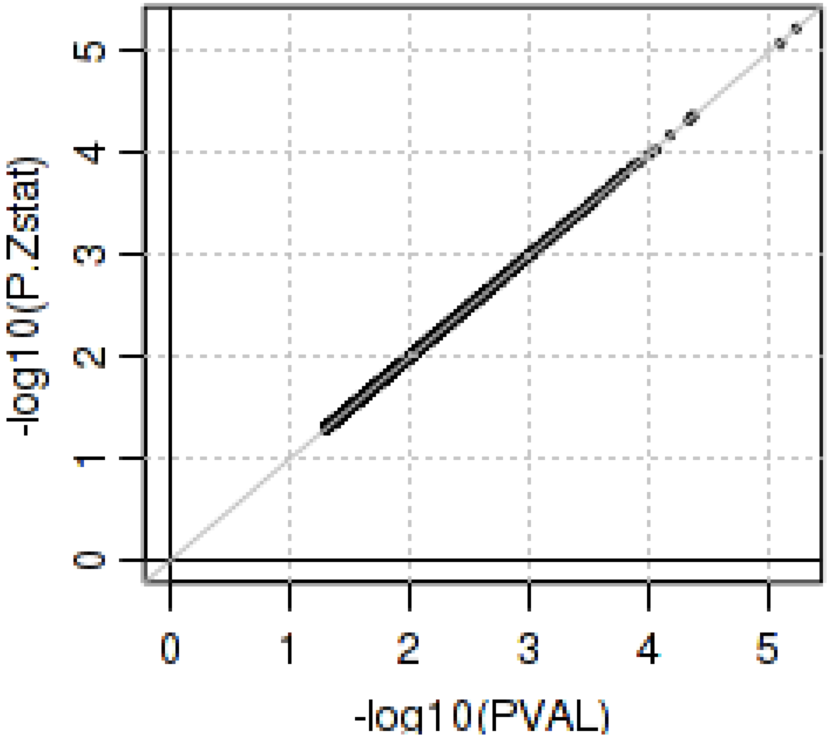
Think about

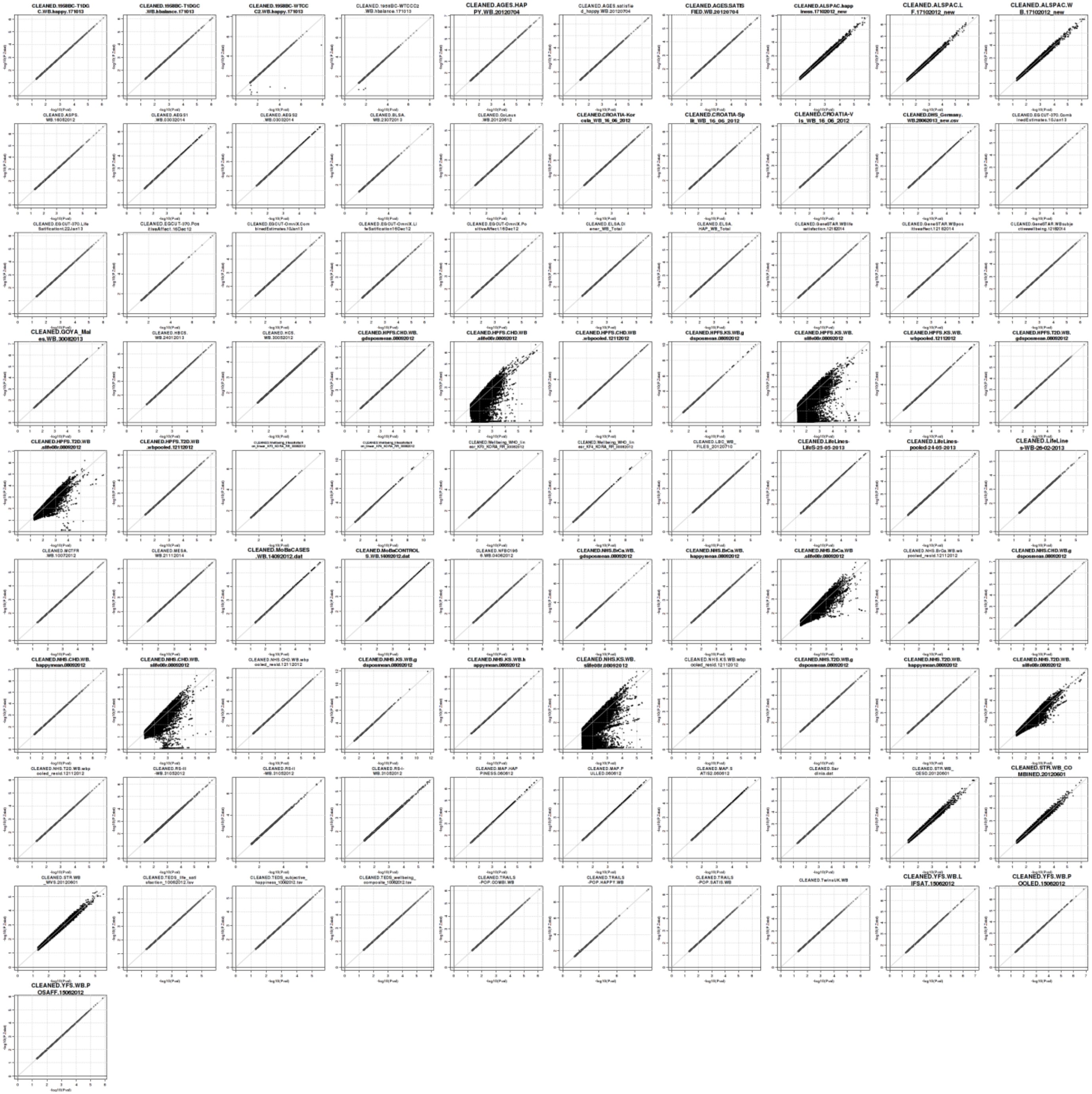
- QQ-plots – to detect early signs of inflation that might be an indication of relatedness problems

- PZ plots -> by B/SE you calculate the corresponding Z statistics. From Z you can obtain P, which can be compared to the actual reported P-value.

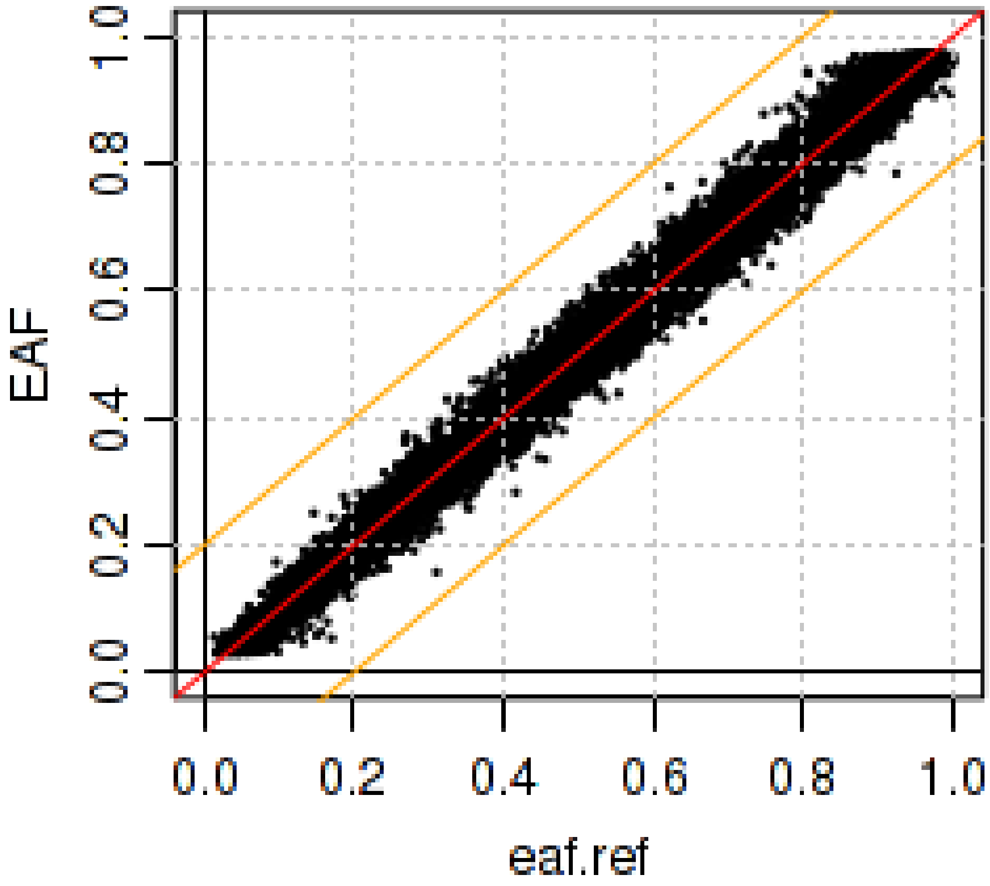
- AF plots -> compare the reported allele frequencies to a reference data set

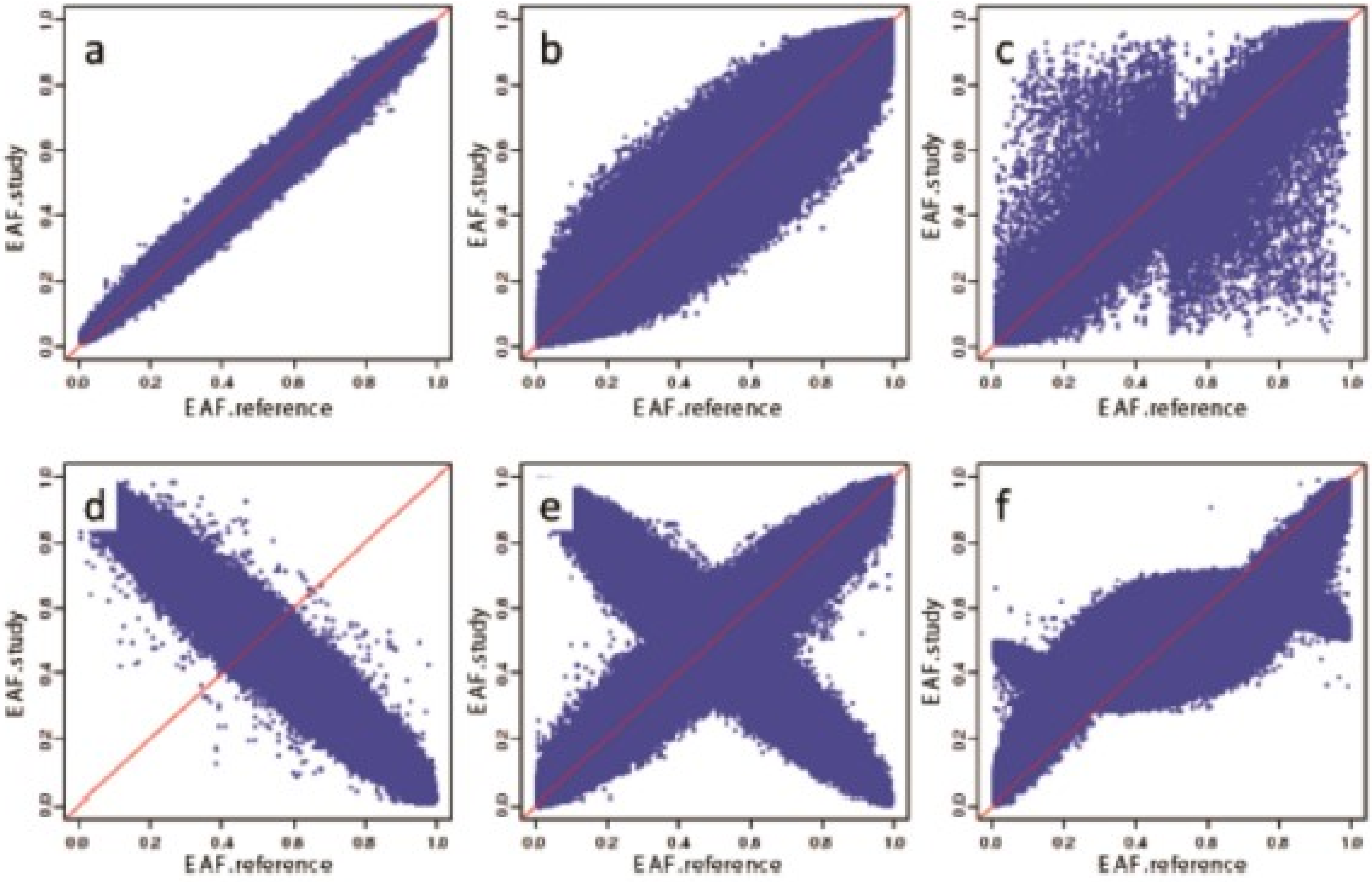
PZ plot



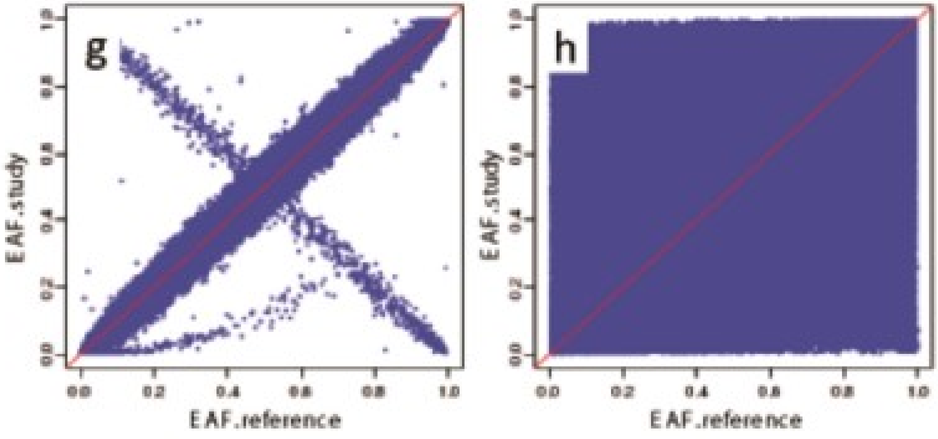

CLEANED.1958BC-T1DG C.WB.happy.171013
CLEANED.1958BC-T1DGC .WB.hbalance.171013
CLEANED.1958BC-WTCC C2.WB.happy.171013
CLEANED.1958BC-WTCCC2 .WB.hbalance.171013
CLEANED.AGES.HAP PY.WB.20120704
CLEANED.AGES.satisfie d_happy.WB.20120704
CLEANED.AGES.SATIS FIED.WB.20120704
CLEANED.ALSPAC.happ iness.17102012_new
CLEANED.ALSPAC.L F.17102012_new
CLEANED.ALSPAC.W B.17102012_new
CLEANED.ASPS. WB.16082012
CLEANED.AEGS1 .WB.03032014
CLEANED.AEGS2 .WB.03032014
CLEANED.BLSA. WB.23072013
CLEANED.CoLaus .WB.20120612
CLEANED.CROATIA-Kor cula_WB_16_06_2012
CLEANED.CROATIA-Sp lit_WB_16_06_2012
CLEANED.CROATIA-V is_WB_16_06_2012
CLEANED.DHS_Germany. WB.28062013_new.csv
CLEANED.EGCUT-370.Comb inedEstimates.10Jan13
CLEANED.EGCUT-370.Life Satificationt.22Jan13
CLEANED.EGCUT-370.Pos itiveAffect.16Dec12
CLEANED.EGCUT-OmniX.Com binedEstimates.10Jan13
CLEANED.EGCUT-OmniX.Li feSatification16Dec12
CLEANED.EGCUT-OmniX.Po sitiveAffect.16Dec12
CLEANED.ELSA.Di ener_WB_Total
CLEANED.ELSA. HAP_WB_Total
CLEANED.GeneSTAR.WBlife satisfaction.12182014
CLEANED.GeneSTAR.WBpos itiveaffect.12182014
CLEANED.GeneSTAR.WBsubje ctivewellbeing.12182014
CLEANED.GOYA_Mal es.WB.30082013
CLEANED.HBCS. WB.24012013
CLEANED.HCS. WB.30052012
CLEANED.HPFS.CHD.WB. gdsposmean.08092012
CLEANED.HPFS.CHD.WB .slife08r.08092012
CLEANED.HPFS.CHD.WB .wbpooled.12112012
CLEANED.HPFS.KS.WB.g dsposmean.08092012
CLEANED.HPFS.KS.WB. slife08r.08092012
CLEANED.HPFS.KS.WB. wbpooled.12112012
CLEANED.HPFS.T2D.WB. gdsposmean.08092012
CLEANED.HPFS.T2D.WB .slife08r.08092012
CLEANED.HPFS.T2D.WB .wbpooled.12112012
CLEANED.Wellbeing_lifesatisfacti on_linear_KF3_KORA_RR_30082012
CLEANED.Wellbeing_lifesatisfacti on_linear_KF4_KORA_RR_30082012
CLEANED.Wellbeing_WHO_lin ear_KF3_KORA_RR_30082012
CLEANED.Wellbeing_WHO_lin ear_KF4_KORA_RR_30082012
CLEANED.LBC_WB_ FILES_20120710
CLEANED.LifeLines- LifeS-25-05-2013
CLEANED.LifeLines- pooled-24-05-2013
CLEANED.LifeLine s-WB-26-02-2013
CLEANED.MCTFR .WB.10072012
CLEANED.MESA. WB.21112014
CLEANED.MoBaCASES .WB.14092012.dat
CLEANED.MoBaCONTROL S.WB.14092012.dat
CLEANED.NFBC196 6.WB.04062012
CLEANED.NHS.BrCa.WB. gdsposmean.08092012
CLEANED.NHS.BrCa.WB. happymean.08092012
CLEANED.NHS.BrCa.WB .slife08r.08092012
CLEANED.NHS.BrCa.WB.wb pooled_resid.12112012
CLEANED.NHS.CHD.WB.g dsposmean.08092012
CLEANED.NHS.CHD.WB. happymean.08092012
CLEANED.NHS.CHD.WB. slife08r.08092012
CLEANED.NHS.CHD.WB.wbp ooled_resid.12112012
CLEANED.NHS.KS.WB.g dsposmean.08092012
CLEANED.NHS.KS.WB.h appymean.08092012
CLEANED.NHS.KS.WB. slife08r.08092012
CLEANED.NHS.KS.WB.wbp ooled_resid.12112012
CLEANED.NHS.T2D.WB.g dsposmean.08092012
CLEANED.NHS.T2D.WB. happymean.08092012
CLEANED.NHS.T2D.WB. slife08r.08092012
CLEANED.NHS.T2D.WB.wbp ooled_resid.12112012
CLEANED.RS-III -WB.31052012
CLEANED.RS-II -WB.31052012
CLEANED.RS-I- WB.31052012
CLEANED.MAP.HAP PINESS.060612
CLEANED.MAP.P ULLED.060612
CLEANED.MAP.S ATIS2.060612
CLEANED.Sar dinia.dat
CLEANED.STR.WB_ CESD.20120601
CLEANED.STR.WB_CO MBINED.20120601
CLEANED.STR.WB _WVS.20120601
CLEANED.TEDS_life_sati sfaction_10062012.tsv
CLEANED.TEDS_subjective_ happiness_10062012.tsv
CLEANED.TEDS_wellbeing_ composite_10062012.tsv
CLEANED.TRAILS -POP.COMBI.WB
CLEANED.TRAILS -POP.HAPPY.WB
CLEANED.TRAILS -POP.SATIS.WB
CLEANED.TwinsUK.WB
CLEANED.YFS.WB.L IFSAT.15062012
CLEANED.YFS.WB.P OOLED.15062012
CLEANED.YFS.WB.P OSAFF.15062012
-log10(Pval)
-log10(P.Zstat)

AF plot



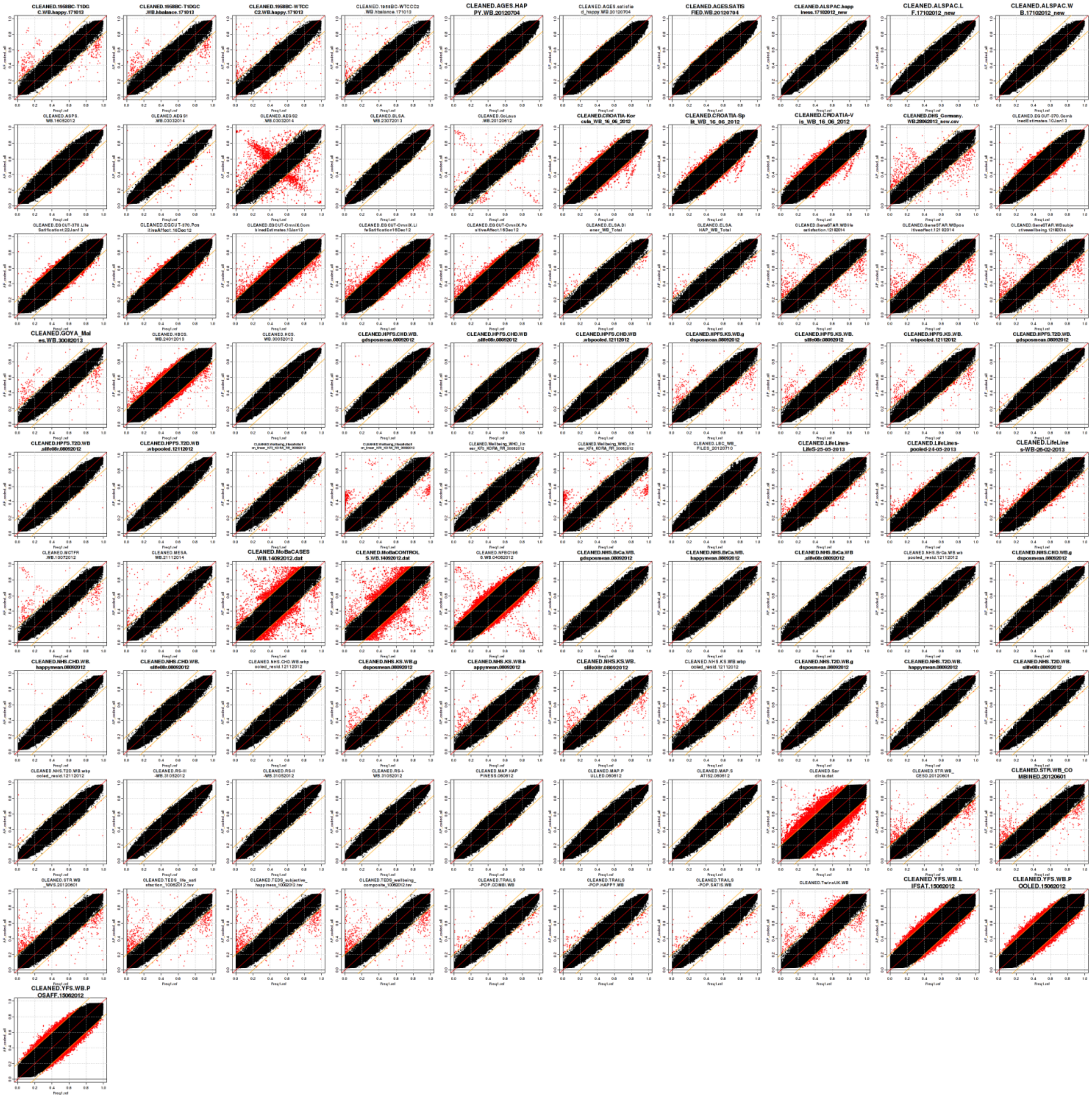


AF correct but different ancestry



(d) wrong allele consistently labeled as effect allele

(e) a fraction of the effect alleles mis-specified, e.g. MAF instead of the effect allele or incorrectly assigning strand


CLEANED.1958BC-T1DG C.WB.happy.171013
CLEANED.1958BC-T1DGC .WB.hbalance.171013
CLEANED.1958BC-WTCC C2.WB.happy.171013
CLEANED.1958BC-WTCCC2 .WB.hbalance.171013
CLEANED.AGES.HAP PY.WB.20120704
CLEANED.AGES.satisfie d_happy.WB.20120704
CLEANED.AGES.SATIS FIED.WB.20120704
CLEANED.ALSPAC.happ iness.17102012_new
CLEANED.ALSPAC.L F.17102012_new
CLEANED.ALSPAC.W B.17102012_new
CLEANED.ASPS. WB.16052012
CLEANED.AEGS1 .WB.03032014
CLEANED.AEGS2 .WB.03032014
CLEANED.BLSA. WB.23072013
CLEANED.CoLaus .WB.20120612
CLEANED.CROATIA-Kor cula_WB_16_06_2012
CLEANED.CROATIA-Sp lit_WB_16_06_2012
CLEANED.CROATIA-V is_WB_16_06_2012
CLEANED.DHS_Germany. WB.28062013_new.csv
CLEANED.EGCUT-370.Comb inedEstimates.10Jan13
CLEANED.EGCUT-370.Life Satificationt.22Jan13
CLEANED.EGCUT-370.Pos itiveAffect.16Dec12
CLEANED.EGCUT-OmniX.Com binedEstimates.10Jan13
CLEANED.EGCUT-OmniX.Li feSatification16Dec12
CLEANED.EGCUT-OmniX.Po sitiveAffect.16Dec12
CLEANED.ELSA.Di ener_WB_Total
CLEANED.ELSA. HAP_WB_Total
CLEANED.GeneSTAR.WBlife satisfaction.12182014
CLEANED.GeneSTAR.WBpos itiveaffect.12182014
CLEANED.GeneSTAR.WBsubje ctivewellbeing.12182014
CLEANED.GOYA_Mal es.WB.30082013
CLEANED.HBCS. WB.24012013
CLEANED.HCS. WB.30052012
CLEANED.HPFS.CHD.WB. gdsposmean.08092012
CLEANED.HPFS.CHD.WB .slife08r.08092012
CLEANED.HPFS.CHD.WB .wbpooled.12112012
CLEANED.HPFS.KS.WB.g dsposmean.08092012
CLEANED.HPFS.KS.WB. slife08r.08092012
CLEANED.HPFS.KS.WB. wbpooled.12112012
CLEANED.HPFS.T2D.WB. gdsposmean.08092012
CLEANED.HPFS.T2D.WB .slife08r.08092012
CLEANED.HPFS.T2D.WB .wbpooled.12112012
CLEANED.Wellbeing_lifesatisfacti on_linear_KF3_KORA_RR_30082012
CLEANED.Wellbeing_lifesatisfacti on_linear_KF4_KORA_RR_30082012
CLEANED.Wellbeing_WHO_lin ear_KF3_KORA_RR_30082012
CLEANED.Wellbeing_WHO_lin ear_KF4_KORA_RR_30082012
CLEANED.LBC_WB_ FILES_20120710
CLEANED.LifeLines- LifeS-25-05-2013
CLEANED.LifeLines- pooled-24-05-2013
CLEANED.LifeLine s-WB-26-02-2013
CLEANED.MCTFR .WB.10072012
CLEANED.MESA. WB.21112014
CLEANED.MoBaCASES .WB.14092012.dat
CLEANED.MoBaCONTROL S.WB.14092012.dat
CLEANED.NFBC196 6.WB.04062012
CLEANED.NHS.BrCa.WB. gdsposmean.08092012
CLEANED.NHS.BrCa.WB. happymean.08092012
CLEANED.NHS.BrCa.WB .slife08r.08092012
CLEANED.NHS.BrCa.WB.wb pooled_resid.12112012
CLEANED.NHS.CHD.WB.g dsposmean.08092012
CLEANED.NHS.CHD.WB. happymean.08092012
CLEANED.NHS.CHD.WB. slife08r.08092012
CLEANED.NHS.CHD.WB.wbp ooled_resid.12112012
CLEANED.NHS.KS.WB.g dsposmean.08092012
CLEANED.NHS.KS.WB.h appymean.08092012
CLEANED.NHS.KS.WB. slife08r.08092012
CLEANED.NHS.KS.WB.wbp ooled_resid.12112012
CLEANED.NHS.T2D.WB.g dsposmean.08092012
CLEANED.NHS.T2D.WB. happymean.08092012
CLEANED.NHS.T2D.WB. slife08r.08092012
CLEANED.NHS.T2D.WB.wbp ooled_resid.12112012
CLEANED.RS-III -WB.31052012
CLEANED.RS-II -WB.31052012
CLEANED.RS-I- WB.31052012
CLEANED.MAP.HAP PINESS.060612
CLEANED.MAP.P ULLED.060612
CLEANED.MAP.S ATIS2.060612
CLEANED.Sar dinia.dat
CLEANED.STR.WB_ CESD.20120601
CLEANED.STR.WB_CO MBINED.20120601
CLEANED.STR.WB _WVS.20120601
CLEANED.TEDS_life_sati sfaction_10062012.tsv
CLEANED.TEDS_subjective_ happiness_10062012.tsv
CLEANED.TEDS_wellbeing_ composite_10062012.tsv
CLEANED.TRAILS -POP.COMBI.WB
CLEANED.TRAILS -POP.HAPPY.WB
CLEANED.TRAILS -POP.SATIS.WB
CLEANED.TwinsUK.WB
CLEANED.YFS.WB.L IFSAT.15062012
CLEANED.YFS.WB.P OOLED.15062012
CLEANED.YFS.WB.P OSAFF.15062012
AF_coded_all
Freq1.ref
0.0
0.2
0.4
0.6
0.8
1.0

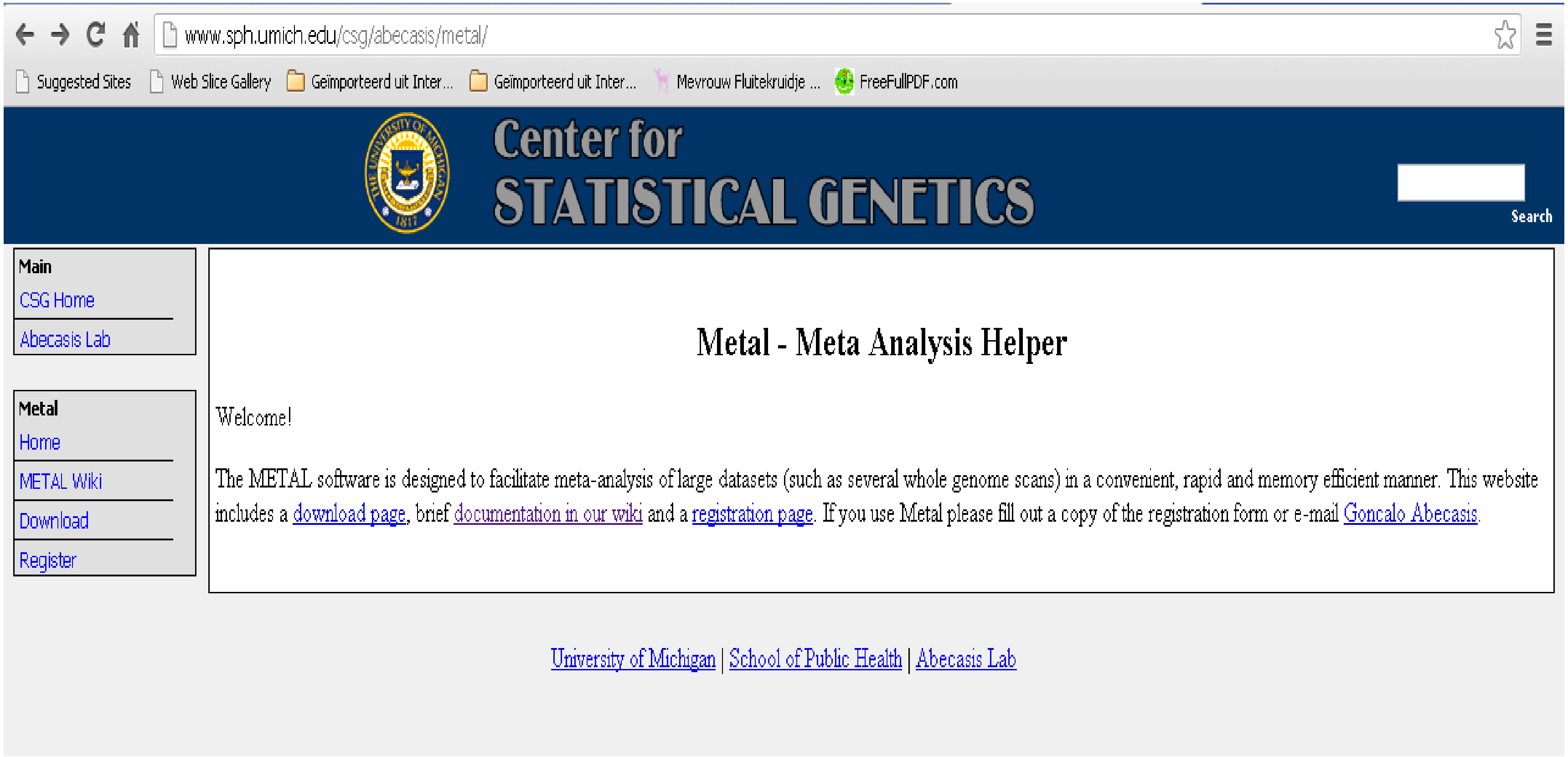
QC workflow – step 3

Compare results with the other data-analyst and resolve any issues remaining

QUALITY CONTROL
APPROVED

METAL

http://www.sph.umich.edu/csg/abecasis/metal/
Documentation can be found at the metal wiki:
http://genome.sph.umich.edu/wiki/Metal_Documentation



Metal - Meta Analysis Helper

Welcome!

The METAL software is designed to facilitate meta-analysis of large datasets (such as several whole genome scans) in a convenient, rapid and memory efficient manner. This website includes a download page, brief documentation in our wiki and a registration page. If you use Metal please fill out a copy of the registration form or e-mail Goncalo Abecasis.

University of Michigan | School of Public Health | Abecasis Lab

METAL

- Metal is flexible
  - By default, METAL combines p-values across studies (sample size, direction of effect)
  - Alternative, standard error based weights (but beta and standard error use same units in all studies)

METAL

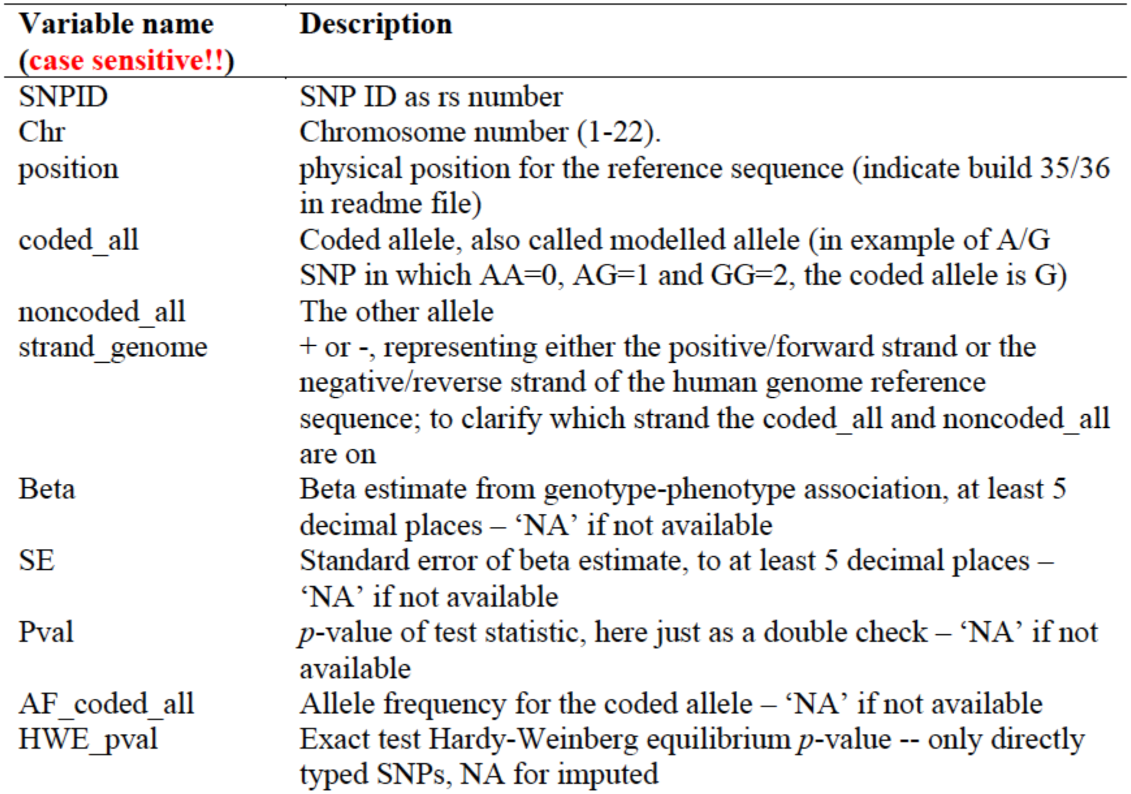
- Requires results files
- ‘Driver’ file
  - Describes the input files
  - Defines meta-analysis strategy
  - Names output file

Steps

1. Check format of results files
    1. Ensure all necessary columns are available
    2. Modify files to include all information
2. Prepare driver file
    1. Ensure headers match description
    2. Crosscheck each results file matches Process name
3. Run metal

Results Files

- Previously asked for standard columns in SOP



Variable name (case sensitive!!)	Description
SNPID	SNP ID as rs number
Chr	Chromosome number (1-22).
position	physical position for the reference sequence (indicate build 35/36 in readme file)
coded_all	Coded allele, also called modelled allele (in example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G)
noncoded_all	The other allele
strand_genome	+ or -, representing either the positive/forward strand or the negative/reverse strand of the human genome reference sequence; to clarify which strand the coded_all and noncoded_all are on
Beta	Beta estimate from genotype-phenotype association, at least 5 decimal places – 'NA' if not available
SE	Standard error of beta estimate, to at least 5 decimal places – 'NA' if not available
Pval	*p*-value of test statistic, here just as a double check – 'NA' if not available
AF_coded_all	Allele frequency for the coded allele – 'NA' if not available
HWE_pval	Exact test Hardy-Weinberg equilibrium *p*-value -- only directly typed SNPs, NA for imputed

- In QC all files are checked (and if necessary corrected)

INPUT FILES

- We will use two GWAs results dataset
  - results1.txt
  - results2.txt

Columns METAL uses

- SNP
- OR
- SE [for standard error meta-analysis]
- P-value [for Z-score meta-analysis]
- If we had two samples of different sizes we would have to add an N/weight column

Meta-analysis running

- We will run meta-analysis based on effect size and on test statistic
- For the weights of test statistic, I’ve assumed that the sample sizes are the same
  - METAL defaults to weight of 1 when no weight column is supplied

Step 2: driver file: meta_run_file

```
# PERFORM META-ANALYSIS based on effect size and on test statistic
# Loading in the input files with results from the  participating samples
# Note: Order of samples is …[sample size, alphabetic order,..]
# Phenotype is ..
# MB March 2019

MARKER  SNP
 ALLELE  A1 A2
 PVALUE  P
 EFFECT  log(OR)
 STDERR  SE                                          specifies column names

PROCESS results1.txt
PROCESS results2.txt                                 processes two results files

OUTFILE meta_res_Z .txt                              Output file naming

ANALYZE                                              Conducts Z-based meta-analysis from test statistic
CLEAR                                                Clears workspace
SCHEME STDERR                                        Changes meta-analysis scheme to beta + SE

PROCESS results1.txt
PROCESS results2.txt                                 processes two results files

OUTFILE meta_res_SE .txt                             Output file naming
ANALYZE                                              Conducts effect size meta-analysis
```

Larger Consortia

```
# Labels
SEPARATOR TAB
MARKER cptid
ALLELE EFFECT_ALLELE OTHER_ALLELE
EFFECT BETA
PVALUE PVAL
FREQ EAF
WEIGHT N

# Options
SCHEME SAMPLESIZE
MINMAXFREQ ON
AVERAGEFREQ ON
GENOMICCONTROL 0.999

# Process files
PROCESS CLEANED.1958T1D.LS.gz
PROCESS CLEANED.BASE.LS.gz
PROCESS CLEANED.HNRSoexpr.LS.gz
PROCESS CLEANED.HRS.LS.gz
PROCESS CLEANED.NHSBRCA.LS.gz
PROCESS CLEANED.RUSHMAP.LS.gz
PROCESS CLEANED.1958WTC.LS.gz
PROCESS CLEANED.EGCUT370.LS.gz
PROCESS CLEANED.HNRSomni1.LS.gz
PROCESS CLEANED.KORAF3.LS.gz
PROCESS CLEANED.NHSCHD.LS.gz
PROCESS CLEANED.TEDS.LS.gz
PROCESS CLEANED.AGES.LS.gz
PROCESS CLEANED.EGCUTOMNI.LS.gz
PROCESS CLEANED.HPFSCHD.LS.gz

Etc

####################################################################

# Analyse and output
MINWEIGHT 50000
ANALYZE HETEROGENEITY

QUIT
```

Running metal

- metal < metal_run_file
- metal is the command
- metal_run_file is the driver file
- This will output information on the running of METAL things to standard out [the terminal]
- It will spawn 4 files:
  - – 2 results files: meta_res_Z1.txt + meta_res_SE1.txt
  - – 2 info files: meta_res_Z1.txt.info + meta_res_SE1.txt.info

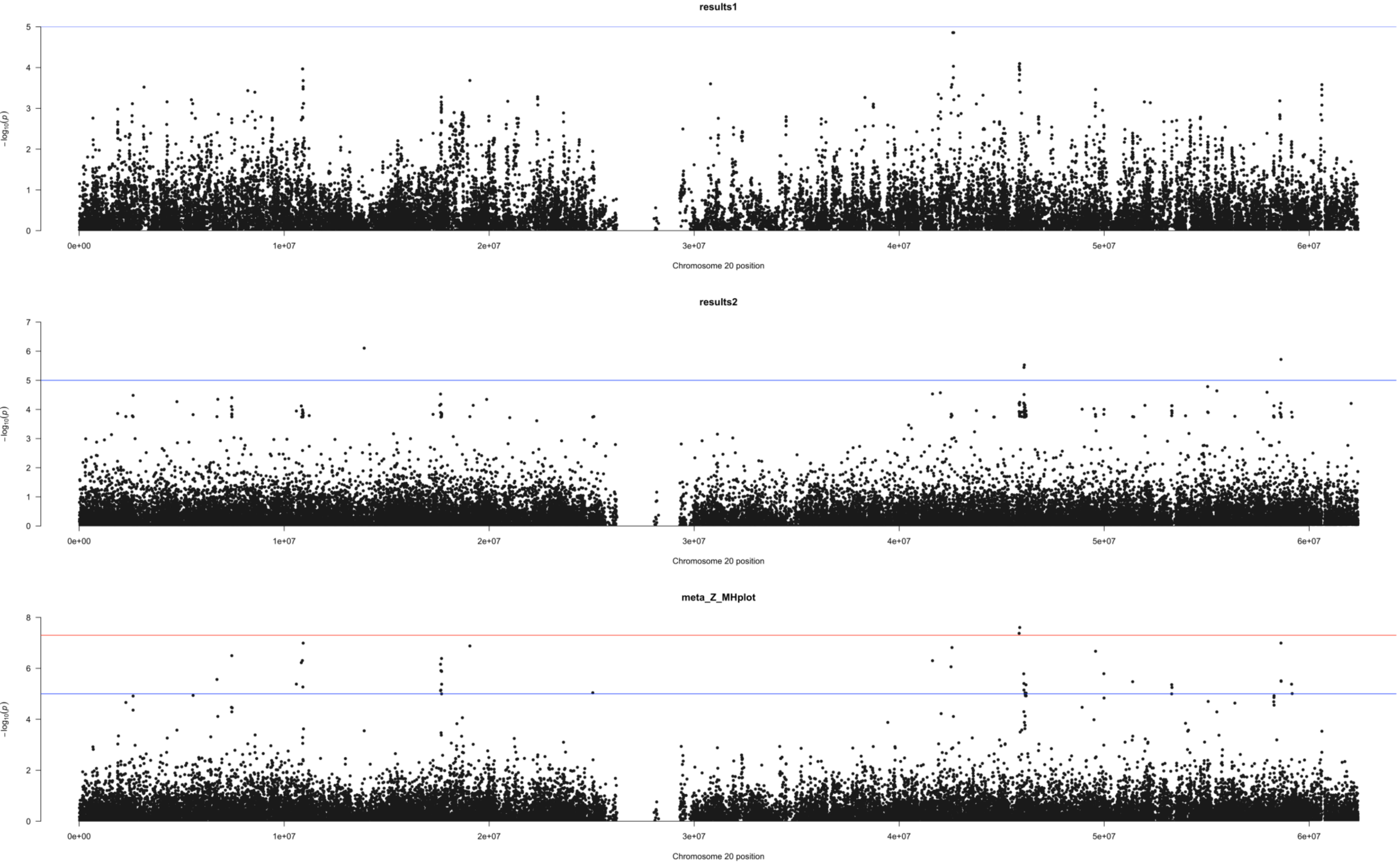
Output

- Overview of METAL commands
- Any errors
- And your best hit from meta-analysis

METAL Practical

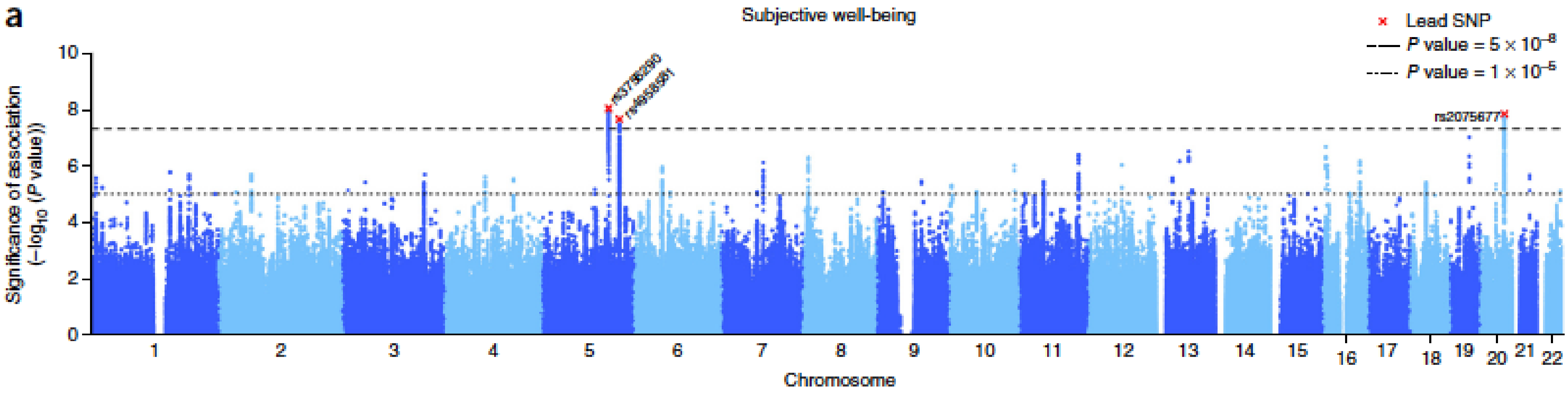
Copy files from **faculty/meike/2019/metal** to your own folder

- Open Metal_prac_Boulder2019.pdf
- follow along
- run the meta-analysis
- create Manhattan plots


results1
results2
meta_Z_MHplot
$-\log_{10}(p)$
0
1
2
3
4
5
6
7
8
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Chromosome 20 position

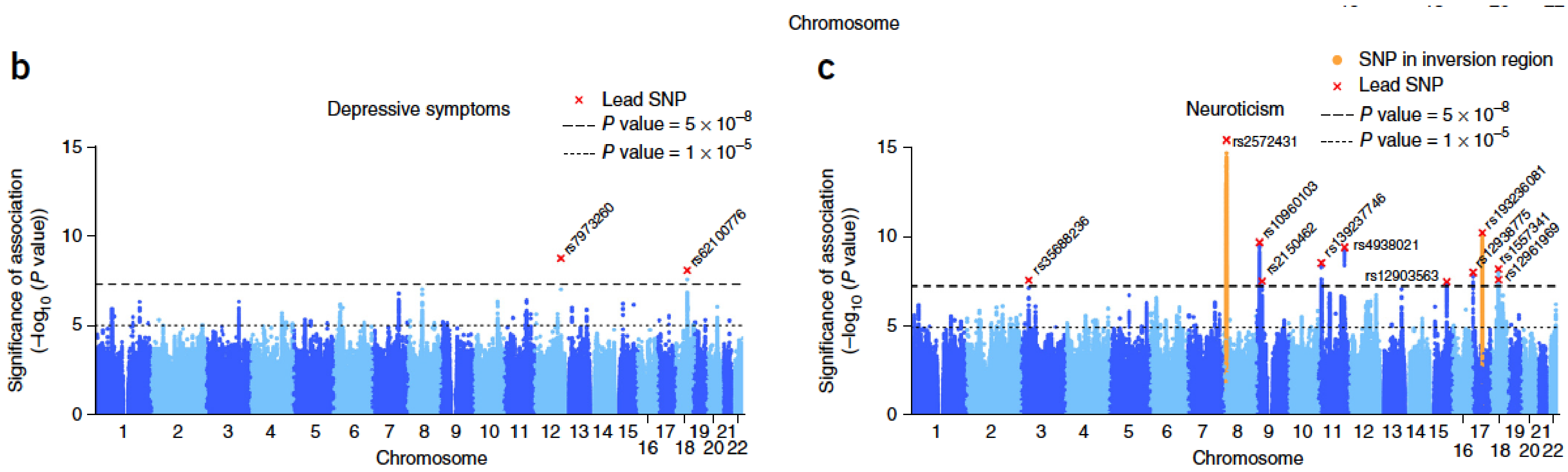
nature genetics

Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses



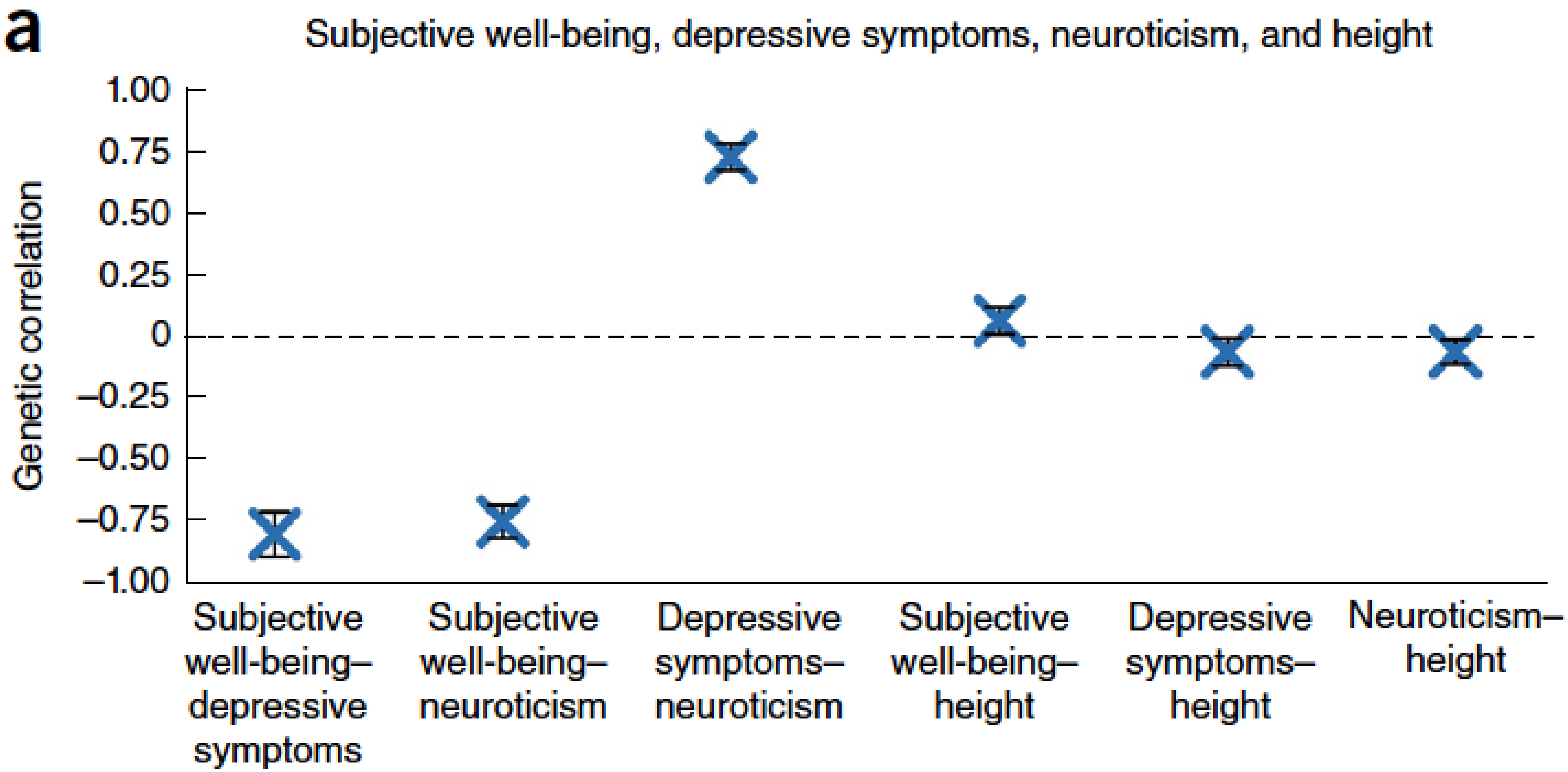
Okbay et al., Nature Genetics, 2016

Depressive Symptoms and Neuroticism

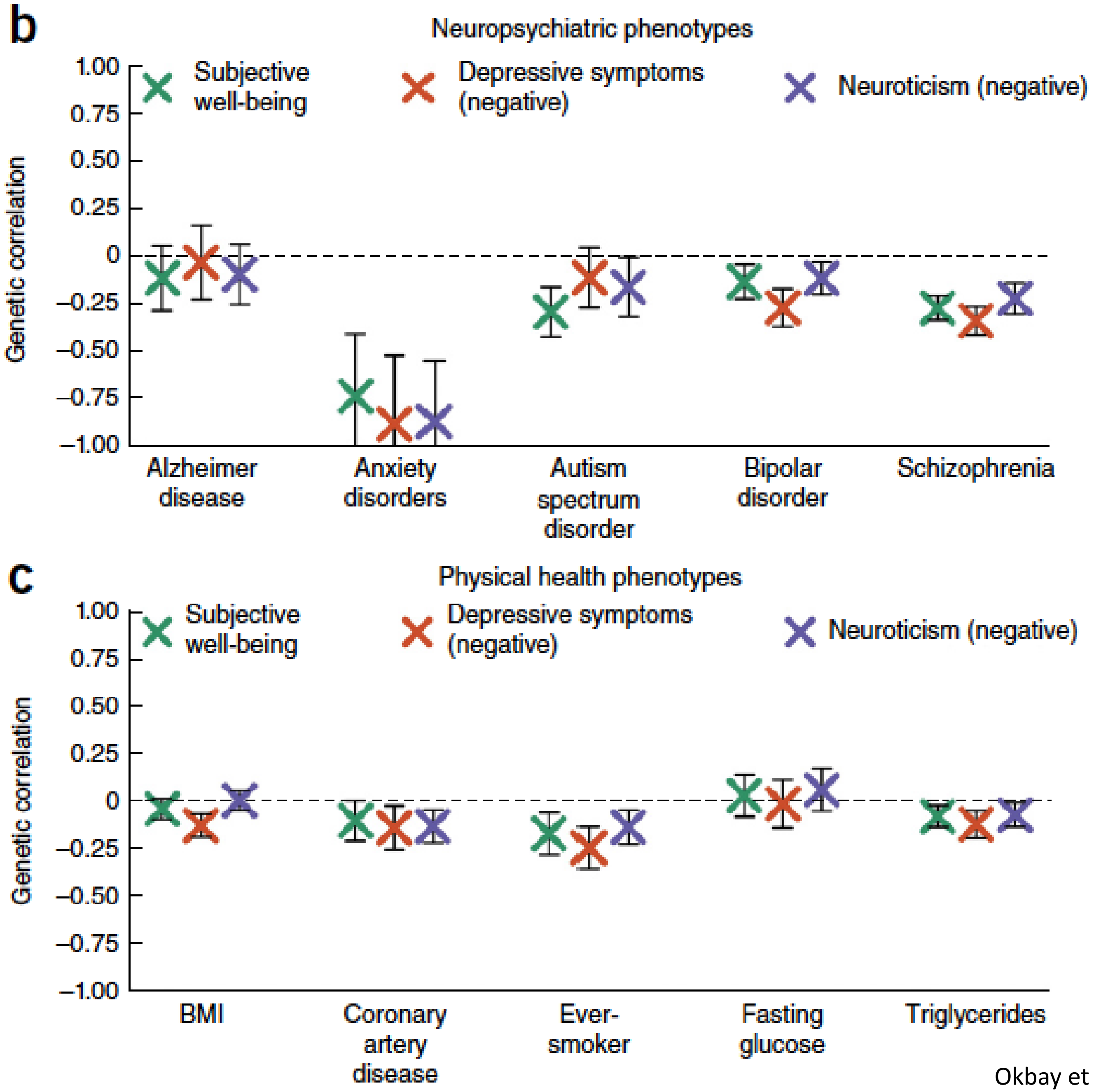


Okbay et al., Nature Genetics, 2016

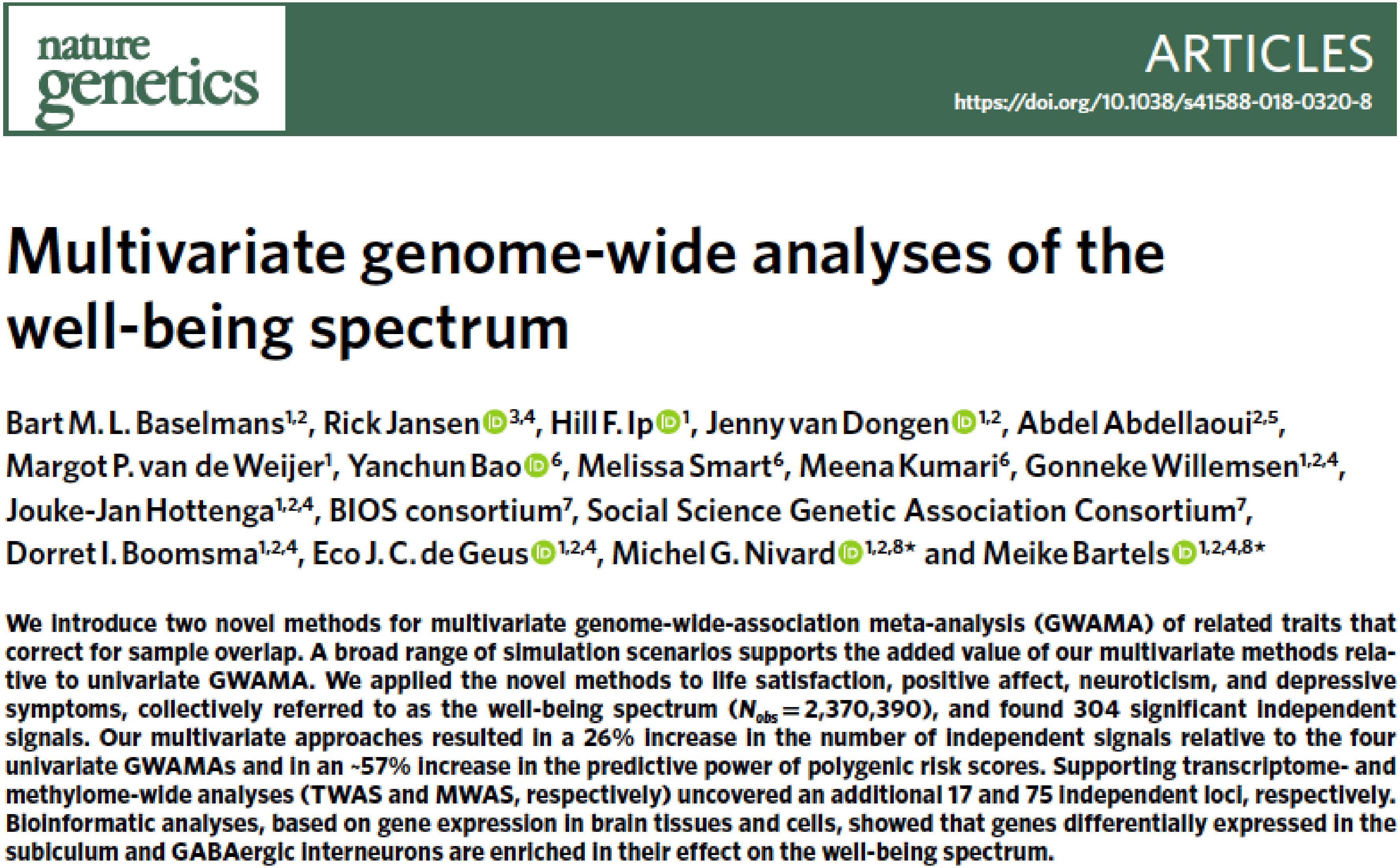
Genetic Correlations



Okbay et al., Nature Genetics, 2016



Okbay et al., Nature Genetics, 2016



nature genetics

ARTICLES

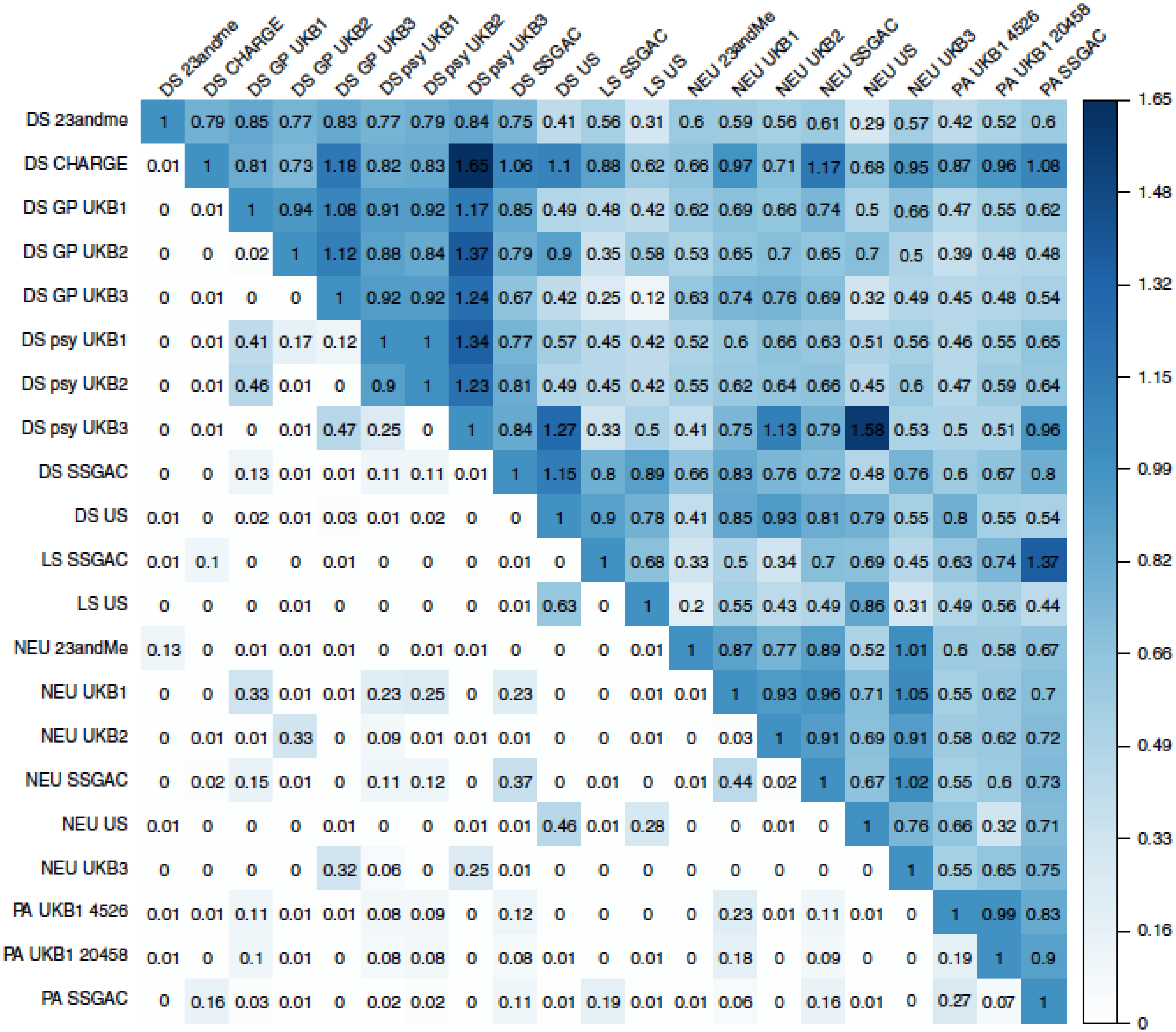
https://doi.org/10.1038/s41588-018-0320-8

Multivariate genome-wide analyses of the well-being spectrum

Bart M. L. Baselmans[1,2], Rick Jansen[3,4], Hill F. Ip[1], Jenny van Dongen[1,2], Abdel Abdellaoui[2,5], Margot P. van de Weijer[1], Yanchun Bao[6], Melissa Smart[6], Meena Kumari[6], Gonneke Willemsen[1,2,4], Jouke-Jan Hottenga[1,2,4], BIOS consortium[7], Social Science Genetic Association Consortium[7], Dorret I. Boomsma[1,2,4], Eco J. C. de Geus[1,2,4], Michel G. Nivard[1,2,8]* and Meike Bartels[1,2,4,8]*

We introduce two novel methods for multivariate genome-wide-association meta-analysis (GWAMA) of related traits that correct for sample overlap. A broad range of simulation scenarios supports the added value of our multivariate methods relative to univariate GWAMA. We applied the novel methods to life satisfaction, positive affect, neuroticism, and depressive symptoms, collectively referred to as the well-being spectrum ($N_{obs} = 2,370,390$), and found 304 significant independent signals. Our multivariate approaches resulted in a 26% increase in the number of independent signals relative to the four univariate GWAMAs and in an ~57% increase in the predictive power of polygenic risk scores. Supporting transcriptome- and methylome-wide analyses (TWAS and MWAS, respectively) uncovered an additional 17 and 75 independent loci, respectively. Bioinformatic analyses, based on gene expression in brain tissues and cells, showed that genes differentially expressed in the subiculum and GABAergic interneurons are enriched in their effect on the well-being spectrum.

Nat Genet. 2019 Mar;51(3):445-451.

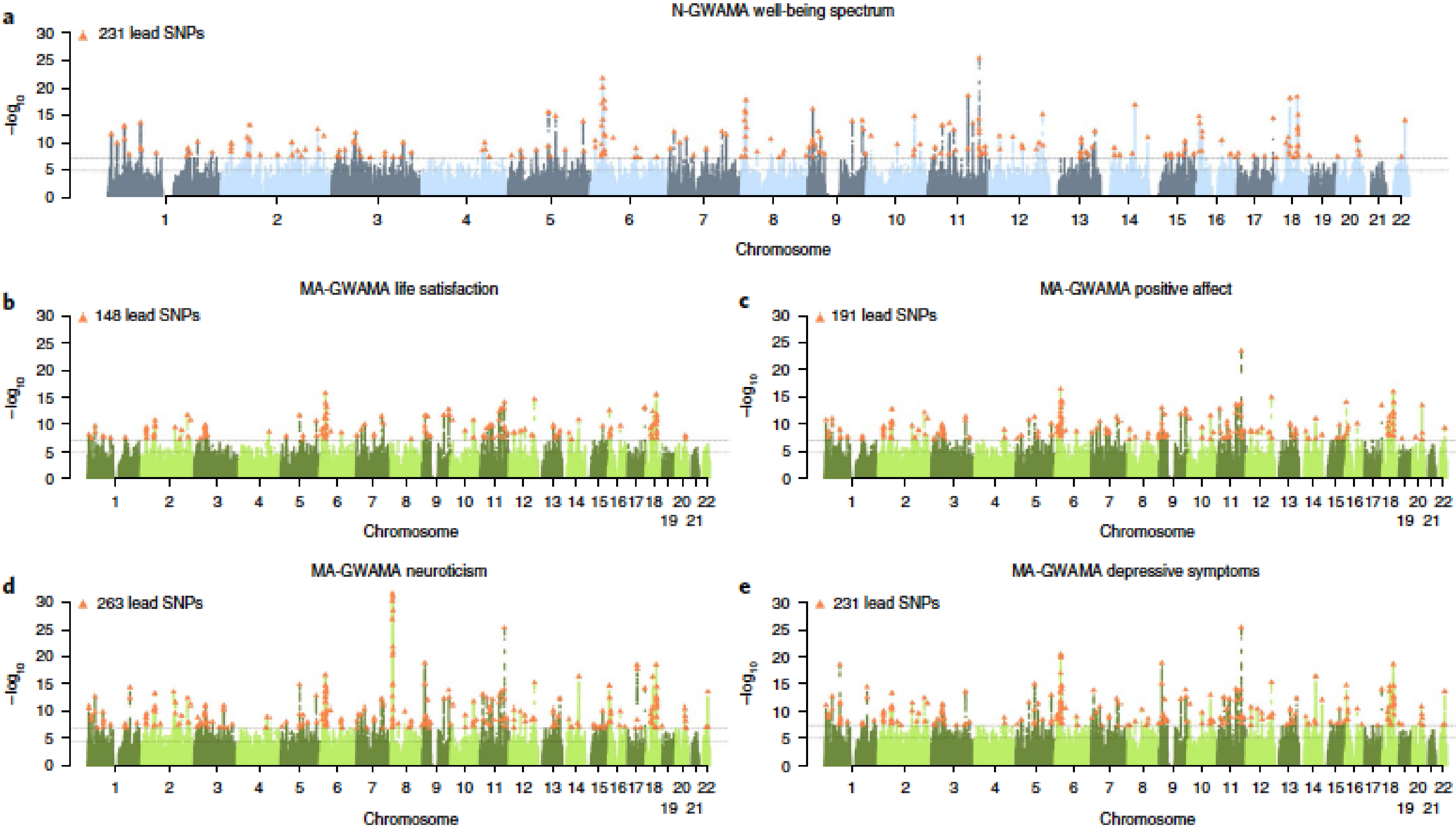


	DS 23andme	DS CHARGE	DS GP UKB1	DS GP UKB2	DS GP UKB3	DS psy UKB1	DS psy UKB2	DS psy UKB3	DS SSGAC	DS US	LS SSGAC	LS US	NEU 23andMe	NEU UKB1	NEU UKB2	NEU SSGAC	NEU US	NEU UKB3	PA UKB1 4526	PA UKB1 20458	PA SSGAC
DS 23andme	1	0.79	0.85	0.77	0.83	0.77	0.79	0.84	0.75	0.41	0.56	0.31	0.6	0.59	0.56	0.61	0.29	0.57	0.42	0.52	0.6
DS CHARGE	0.01	1	0.81	0.73	1.18	0.82	0.83	1.65	1.06	1.1	0.88	0.62	0.66	0.97	0.71	1.17	0.68	0.95	0.87	0.96	1.08
DS GP UKB1	0	0.01	1	0.94	1.08	0.91	0.92	1.17	0.85	0.49	0.48	0.42	0.62	0.69	0.66	0.74	0.5	0.66	0.47	0.55	0.62
DS GP UKB2	0	0	0.02	1	1.12	0.88	0.84	1.37	0.79	0.9	0.35	0.58	0.53	0.65	0.7	0.65	0.7	0.5	0.39	0.48	0.48
DS GP UKB3	0	0.01	0	0	1	0.92	0.92	1.24	0.67	0.42	0.25	0.12	0.63	0.74	0.76	0.69	0.32	0.49	0.45	0.48	0.54
DS psy UKB1	0	0.01	0.41	0.17	0.12	1	1	1.34	0.77	0.57	0.45	0.42	0.52	0.6	0.66	0.63	0.51	0.56	0.46	0.55	0.65
DS psy UKB2	0	0.01	0.46	0.01	0	0.9	1	1.23	0.81	0.49	0.45	0.42	0.55	0.62	0.64	0.66	0.45	0.6	0.47	0.59	0.64
DS psy UKB3	0	0.01	0	0.01	0.47	0.25	0	1	0.84	1.27	0.33	0.5	0.41	0.75	1.13	0.79	1.58	0.53	0.5	0.51	0.96
DS SSGAC	0	0	0.13	0.01	0.01	0.11	0.11	0.01	1	1.15	0.8	0.89	0.66	0.83	0.76	0.72	0.48	0.76	0.6	0.67	0.8
DS US	0.01	0	0.02	0.01	0.03	0.01	0.02	0	0	1	0.9	0.78	0.41	0.85	0.93	0.81	0.79	0.55	0.8	0.55	0.54
LS SSGAC	0.01	0.1	0	0	0.01	0	0	0	0.01	0	1	0.68	0.33	0.5	0.34	0.7	0.69	0.45	0.63	0.74	1.37
LS US	0	0	0	0.01	0	0	0	0	0.01	0.63	0	1	0.2	0.55	0.43	0.49	0.86	0.31	0.49	0.56	0.44
NEU 23andMe	0.13	0	0.01	0.01	0.01	0	0.01	0.01	0	0	0	0.01	1	0.87	0.77	0.89	0.52	1.01	0.6	0.58	0.67
NEU UKB1	0	0	0.33	0.01	0.01	0.23	0.25	0	0.23	0	0	0.01	0.01	1	0.93	0.96	0.71	1.05	0.55	0.62	0.7
NEU UKB2	0	0.01	0.01	0.33	0	0.09	0.01	0.01	0.01	0	0	0.01	0	0.03	1	0.91	0.69	0.91	0.58	0.62	0.72
NEU SSGAC	0	0.02	0.15	0.01	0	0.11	0.12	0	0.37	0	0.01	0	0.01	0.44	0.02	1	0.67	1.02	0.55	0.6	0.73
NEU US	0.01	0	0	0	0.01	0	0	0.01	0.01	0.46	0.01	0.28	0	0	0.01	0	1	0.76	0.66	0.32	0.71
NEU UKB3	0	0	0	0	0.32	0.06	0	0.25	0.01	0	0	0	0	0	0	0	0	1	0.55	0.65	0.75
PA UKB1 4526	0.01	0.01	0.11	0.01	0.01	0.08	0.09	0	0.12	0	0	0	0	0.23	0.01	0.11	0.01	0	1	0.99	0.83
PA UKB1 20458	0.01	0	0.1	0.01	0	0.08	0.08	0	0.08	0.01	0	0.01	0	0.18	0	0.09	0	0	0.19	1	0.9
PA SSGAC	0	0.16	0.03	0.01	0	0.02	0.02	0	0.11	0.01	0.19	0.01	0.01	0.06	0	0.16	0.01	0	0.27	0.07	1

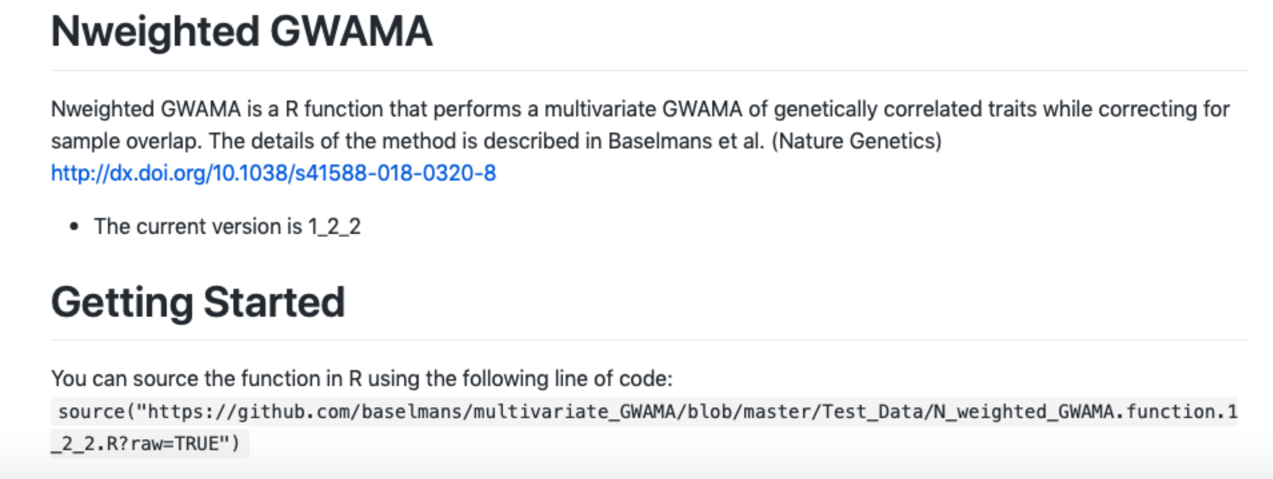
Two Multivariate Approaches

1. N-weighted multivariate GWAMA (N-GWAMA) with a unitary effect of the SNP on all traits

2. Model-averaging GWAMA (MA-GWAMA) in which we relaxed the assumption of a unitary effect of the SNP on all traits.


a
N-GWAMA well-being spectrum
231 lead SNPs
−log10
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
b
MA-GWAMA life satisfaction
148 lead SNPs
−log10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
c
MA-GWAMA positive affect
191 lead SNPs
−log10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
d
MA-GWAMA neuroticism
263 lead SNPs
−log10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
e
MA-GWAMA depressive symptoms
231 lead SNPs
−log10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

https://github.com/baselmans/multivariate_GWAMA



Nweighted GWAMA

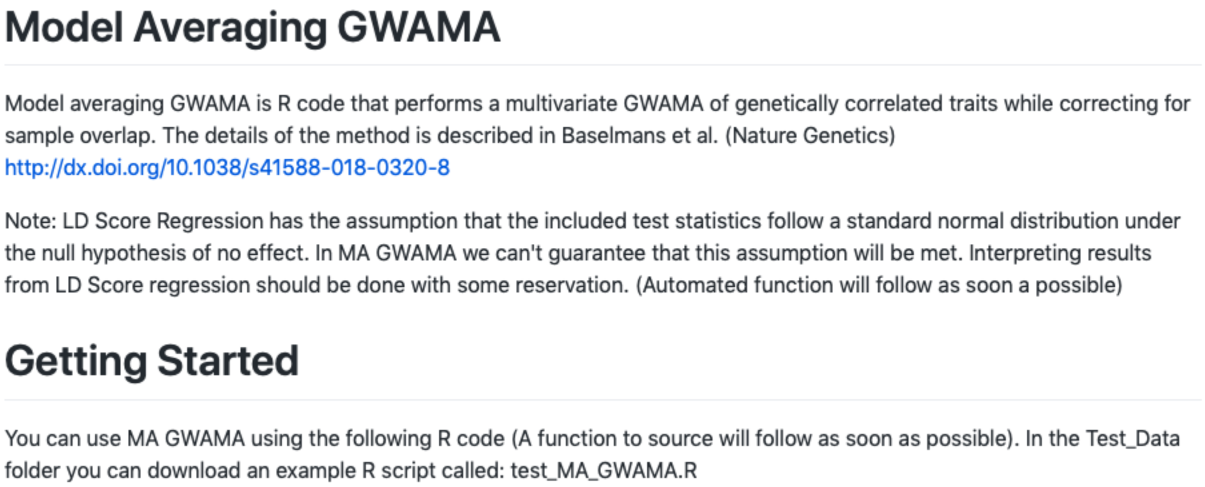
Nweighted GWAMA is a R function that performs a multivariate GWAMA of genetically correlated traits while correcting for sample overlap. The details of the method is described in Baselmans et al. (Nature Genetics) http://dx.doi.org/10.1038/s41588-018-0320-8

- The current version is 1_2_2

Getting Started

You can source the function in R using the following line of code:

```
source("https://github.com/baselmans/multivariate_GWAMA/blob/master/Test_Data/N_weighted_GWAMA.function.1_2_2.R?raw=TRUE")
```



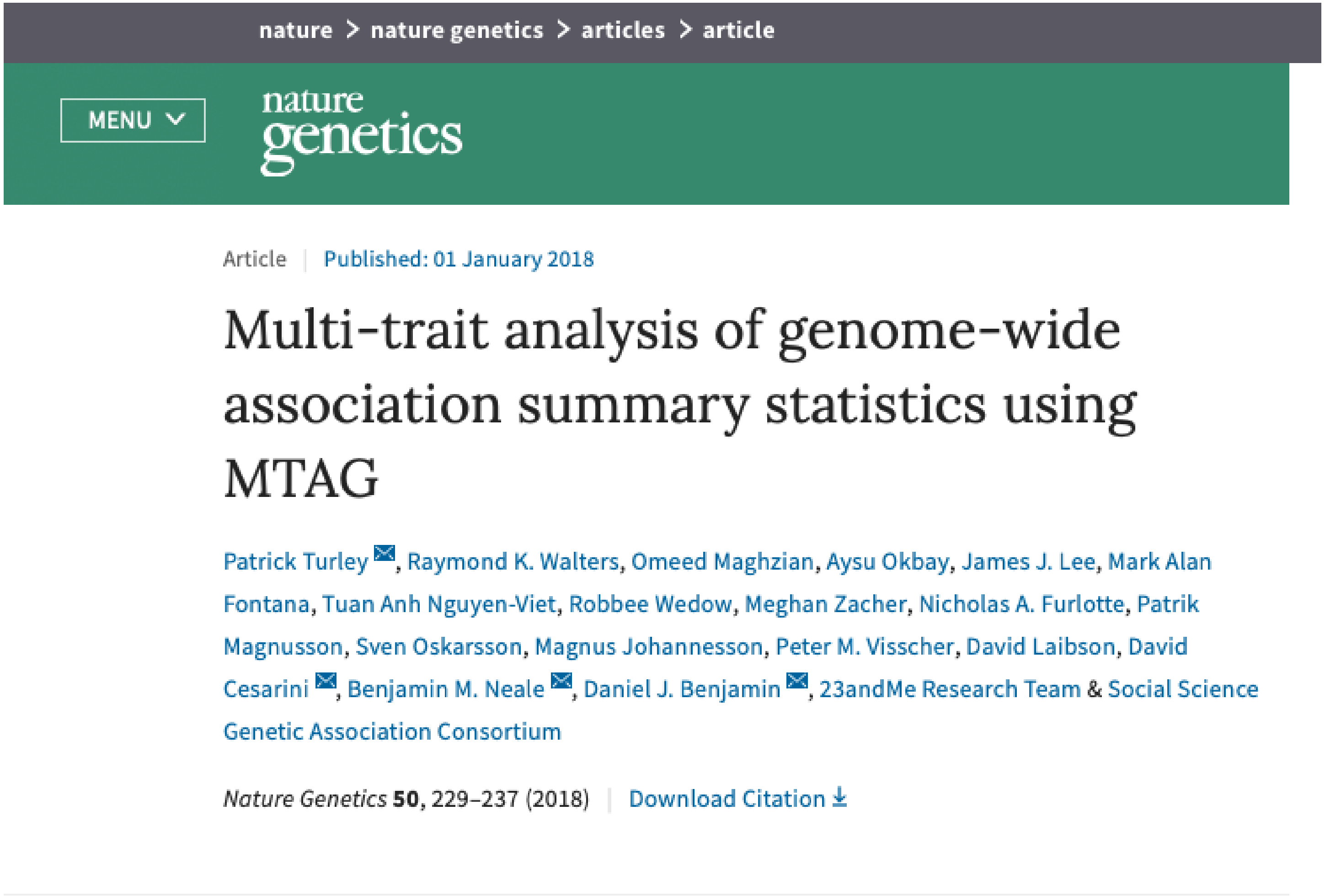
Model Averaging GWAMA

Model averaging GWAMA is R code that performs a multivariate GWAMA of genetically correlated traits while correcting for sample overlap. The details of the method is described in Baselmans et al. (Nature Genetics) http://dx.doi.org/10.1038/s41588-018-0320-8

Note: LD Score Regression has the assumption that the included test statistics follow a standard normal distribution under the null hypothesis of no effect. In MA GWAMA we can't guarantee that this assumption will be met. Interpreting results from LD Score regression should be done with some reservation. (Automated function will follow as soon a possible)

Getting Started

You can use MA GWAMA using the following R code (A function to source will follow as soon as possible). In the Test_Data folder you can download an example R script called: test_MA_GWAMA.R



nature > nature genetics > articles > article

MENU

nature genetics

Article | Published: 01 January 2018

Multi-trait analysis of genome-wide association summary statistics using MTAG

Patrick Turley ✉, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A. Furlotte, Patrik Magnusson, Sven Oskarsson, Magnus Johannesson, Peter M. Visscher, David Laibson, David Cesarini ✉, Benjamin M. Neale ✉, Daniel J. Benjamin ✉, 23andMe Research Team & Social Science Genetic Association Consortium

Nature Genetics **50**, 229–237 (2018) | Download Citation