Introduction to genome-wide association analysis: rvtests

Department of Biomedical Data Sciences

Section of Molecular Epidemiology

LUMC, Leiden

Outline

- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

Outline

- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

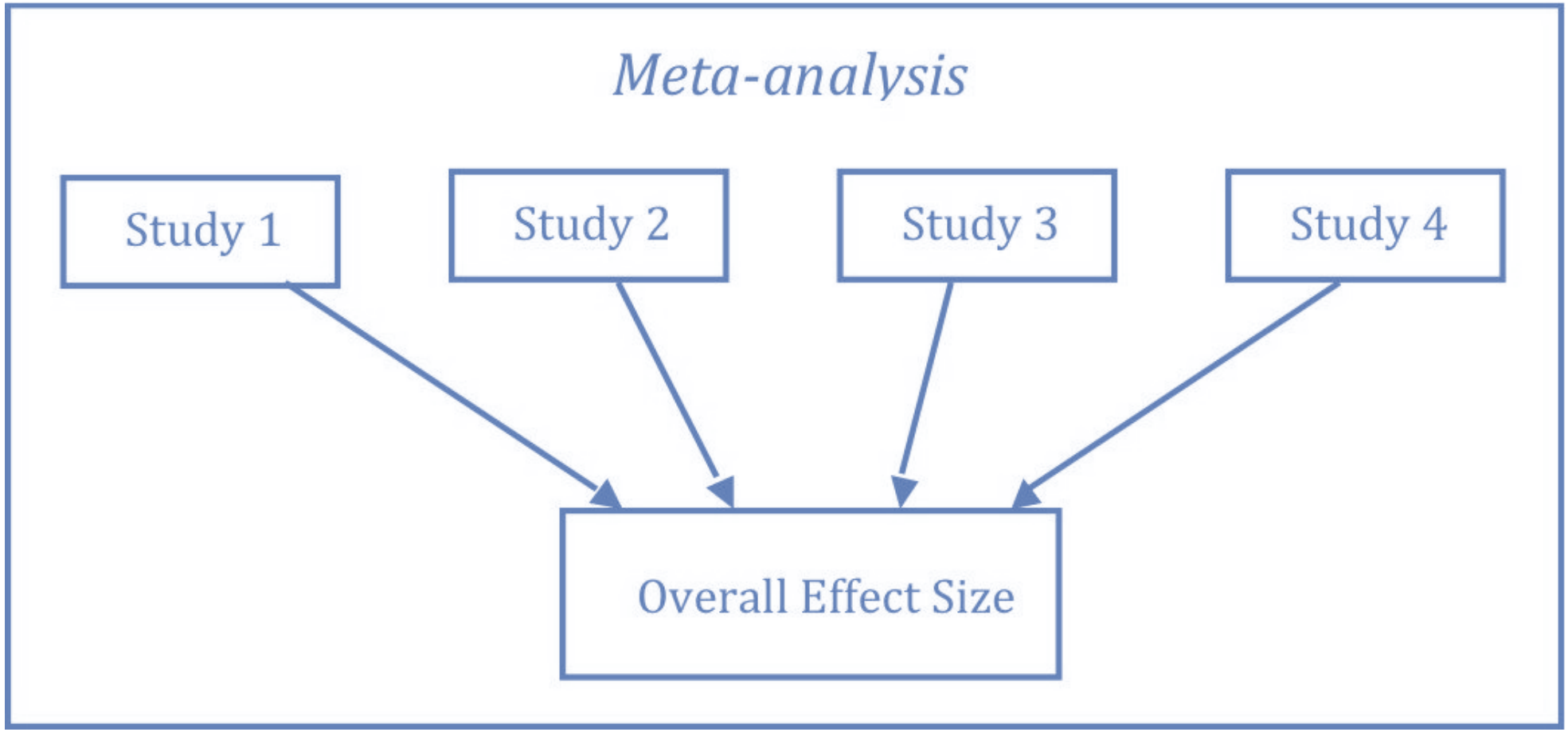
Genome-wide association study (GWAS)

- Millions of genetic variants across the genomes of many individuals tested to identify genotype–phenotype associations
- Various statistical methods and tools available

- Today's practical: Rvtests (Rare Variant tests)
- Available on Linux, MacOS and Windows, developed by Zhan et al.
- Developed to support genetic association analysis for sequence datasets
- Can analyze:
  - unrelated individuals and related (family-based) individuals
  - quantitative and binary outcomes

http://zhanxw.github.io/rvtests/#introduction

Rvtests and meta-analysis



Meta-analysis - combining the results of individual studies with statistical methods

Outline

- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

Input files (1)

Phenotype file

fid	iid	fatid	matid	sex	phenotype1	phenotype2	phenotype3
1	1	0	0	1	5.879	25.888	0
2	2	0	0	2	8.954	19.324	2
3	3	0	0	2	1.909	20.125	1
4	4	0	0	1	NA	28.587	1
5	5	0	0	1	7.888	35.996	1

Covariate file

fid	iid	fatid	matid	sex	covariate1	covariate2
1	1	0	0	1	78.534	1
2	2	0	0	2	67.987	0
3	3	0	0	2	85.123	0
4	4	0	0	1	49.023	1
5	5	0	0	1	55.943	1

Input files (2)

Genotype files – VCF (Variant Call Format) /BGEN/PLINK format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Input files (2)

Genotype files – VCF (Variant Call Format) /BGEN/PLINK format

Header

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Records

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Input files (2)

Genotype files – VCF (Variant Call Format) /BGEN/PLINK format

Header

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20    14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20    17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20    1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20    1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20    1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

INFO meta-information

FILTER meta-information

FORMAT meta-information

Input files (2)

Genotype files – VCF (Variant Call Format) /BGEN/PLINK format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Fixed fields

Optional: FORMAT field specifying data type and per-sample genotype data

Records

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Input files (2)

Genotype files – VCF (Variant Call Format) /BGEN/PLINK format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20    14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20    17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20    1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20    1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20    1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

VCF file --genotype fields

Sample 1 Sample 2 Sample 3

```
FORMAT       NA00001          NA00002          NA00003
GT:GQ:DP:HQ  0|0:48:1:51,51   1|0:48:8:51,51   1/1:43:5:.,.
GT:GQ:DP:HQ  0|0:49:3:58,50   0|1:3:5:65,3     0/0:41:3
GT:GQ:DP:HQ  1|2:21:6:23,27   2|1:2:0:18,2     2/2:35:4
GT:GQ:DP:HQ  0|0:54:7:56,60   0|0:48:4:51,51   0/0:61:2
GT:GQ:DP     0/1:35:4         0/2:17:2         1/1:40:3
```

Haplotype qualities

Read depth at this position for this sample

Conditional genotype quality

Genotype
0 - the reference allele (REF field)
1 - the first allele listed in ALT
2 - the second allele list in ALT

Missing values

// Handle missing genotypes and phenotypes

When genotypes are missing (e.g. genotype = "./.") or genotypes are filtered out, there are three options to handle them: (1) impute to its mean(default option); (2) impute by HWE equilibrium; (3) remove from the model. Use `--impute [mean|hwe|drop]` to specify which option to use.

When quantitative phenotypes are missing, for example, some samples have genotype files, but not phenotypes, rvtests can impute missing phenotype to its mean.

NOTE: Do not use `--imputePheno` for binary trait.

In summary, the following two options can be used:

```
          --impute : Specify either of mean, hwe, and drop
     --imputePheno : Impute phenotype to mean by those have genotypes but no
                     phenotypes
```

Outline

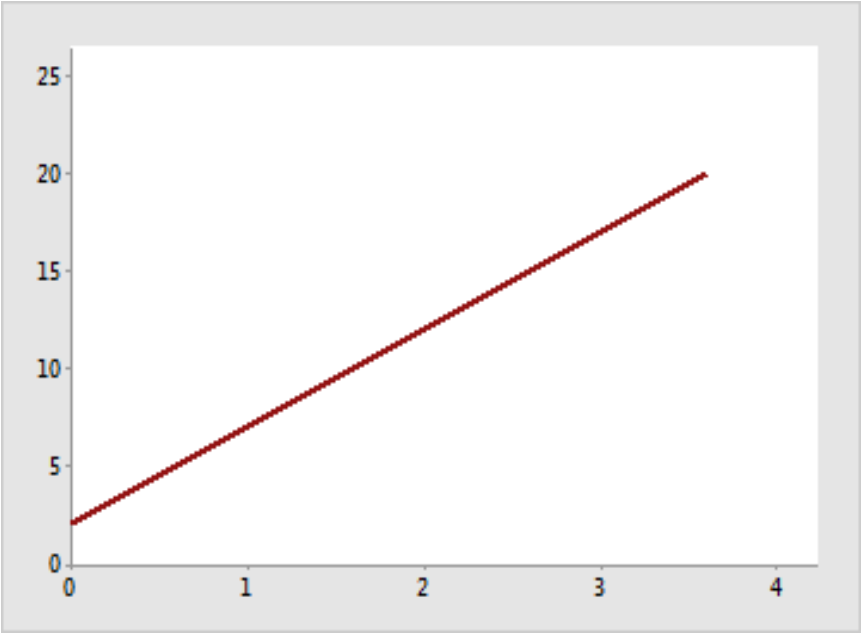
- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

Association analysis

Model: **Trait ~ constant + β_1x SNP + β_2 x covariate 1 + β_3 x covariate 2**

Estimate of regression intercept

Estimate of regression slope



The dependent variable "trait" is:

- Quantitative trait → Linear regression
- Case control status → Logistic regression

test H0: $\beta 1 = 0$, No association

H1: $\beta 1 \neq 0$, Association

Association tests (1)

// Meta-analysis models

Type	Model(#)	Traits(##)	Covariates	Related / unrelated	Description
Score test	score	B,Q	Y	R, U	standard score tests
Dominant model	dominant	B,Q	Y	R, U	score tests and covariance matrix under dominant disease model
Recessive model	recessive	B,Q	Y	R, U	score tests and covariance matrix under recessive disease model
Covariance	cov	B,Q	Y	R, U	covariance matrix
BOLT-LMM score test	bolt	Q	Y	R	BOLT-LMM based score tests (###)
BOLT-LMM covariance	boltCov	Q	Y	R	BOLT-LMM based score tests (###)

(#) Model columns list the recognized names in rvtests. For example, use `--meta score,cov` will generate score statistics and covariance matrix for meta-analysis.

Run GWAS --command

```
rvtest    --inVcf input.vcf              → specify genotype file
          --pheno phenotype.ped          → specify phenotype file
          --pheno-name phenotype1        → specify phenotype
          --covar example.covar          → specify covariate file
          --covar-name age,sex           → specify covariates
          --dosage DS                    → specify dosage tag
          --meta score                   → specify association model
          --out output                   → specify output file
```

Demonstration using HPC (SHARK)

Outline

- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

Output file

CHROM	POS	REF	ALT	N INFORMATIVE	AF	INFORMATIVE ALT AC	CALL RATE	HWE PVALUE	N_REF	N_HET	N_ALT	U_STAT	SQRT_V_STAT	ALT_EFFSIZE	PVALUE
1	13380	C	G	2158	7,22E-01	0.029	1	1	2158	0	0	-0.0098	0.007996	-154.14	0.217756
1	16071	G	A	2158	0.0002	0.796	1	1	2157	1	0	0.03002	0.738133	0.0551133	0.96755
1	16141	C	T	2158	0.0002	0.6	1	1	2158	0	0	0.48214	0.465026	22.296	0.29982
1	16280	T	C	2158	0.0004	1.647	1	1	2158	0	0	0.00312	0.063440	0.777443	0.960663
1	49298	T	C	2158	0.6361	2745.95	1	9.33E-284	6	1777	375	0.72486	526.607	0.02613	0.890518
1	54353	C	A	2158	0.0007	3.171	1	1	2158	0	0	0.09704	0.113511	753.182	0.39258
1	54564	G	T	2158	0.0001	0.518	1	1	2158	0	0	-0.00917	0.033233	-830.354	0.782583
1	54591	A	G	2158	0.0002	0.956	1	1	2158	0	0	0.03387	0.5388	0.1167	0.949864
1	54676	C	T	2158	0.3790	1636.34	1	0	278	1874	6	0.55609	541.083	0.0189943	0.918141

Number of samples analyzed for association

Allele frequency

The number of alternative alleles in the analyzed samples

The fraction of non-missing alleles

P-value Hardy-Weinberg equilibrium

Number of samples carrying homozygous reference/ heterozygous/ homozygous alternative alleles

Outline

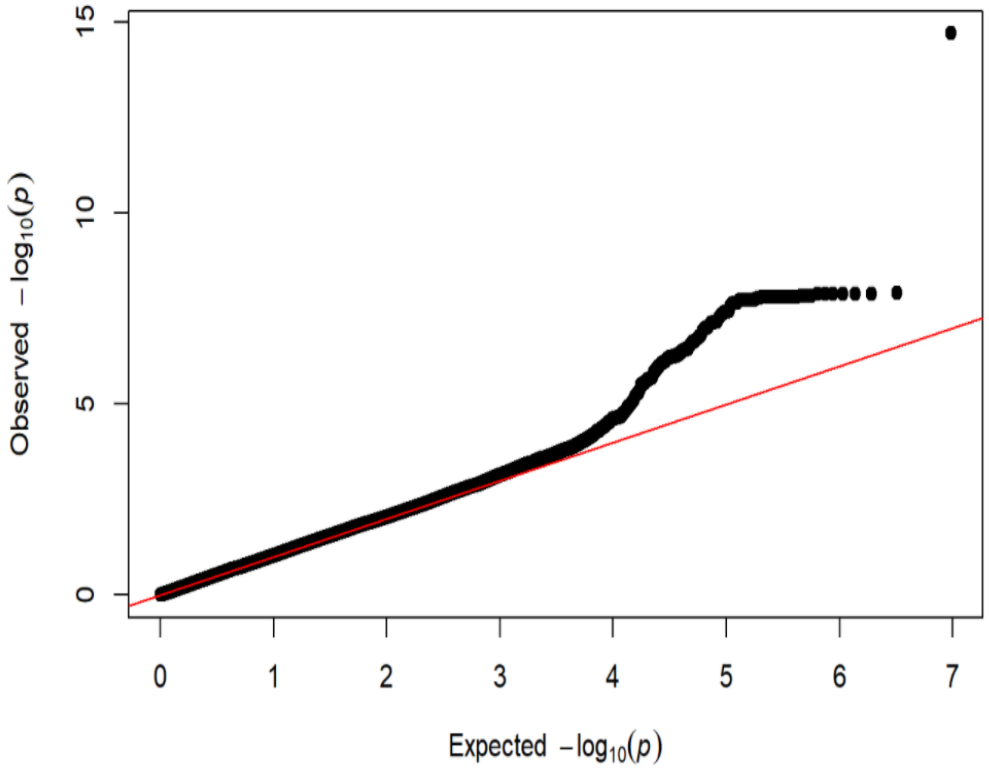
- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

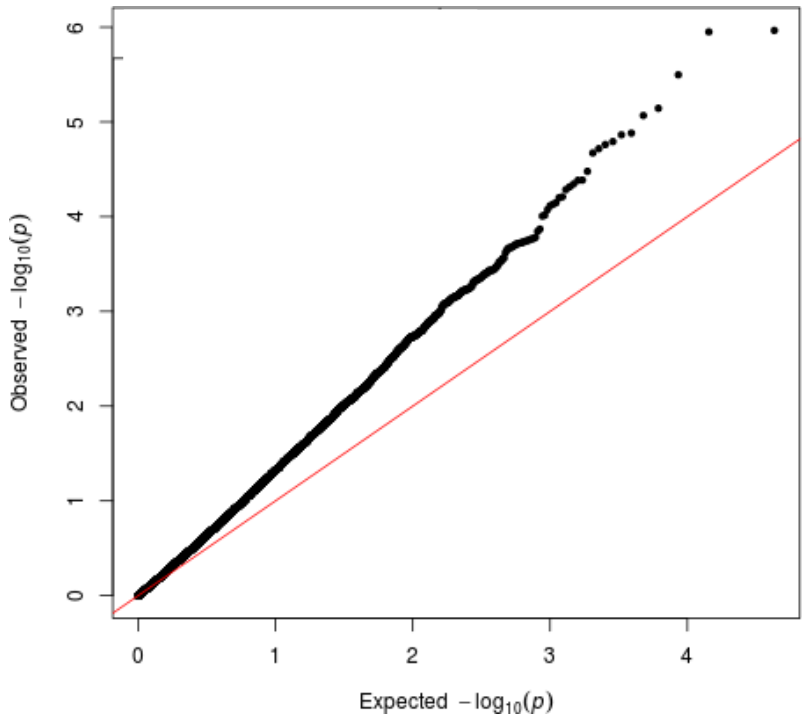
Quality control

1. Cleaning of the data by deleting poor quality data
   - SNPs with low minor allele frequency (MAF)
   - SNP deviating from Hardy-Weinberg equilibrium (HWE)
   - SNPs with low imputation quality ($R^2 < 0.3$)

2. Detect bias – QQ plot

QQ plot – bias detection





Easy QC software

EN DE Suchen nach... SU

Universität Regensburg

STARTSEITE UR

STARTSEITE
EPIDEMIOLOGIE UND PRÄVENTIVMEDIZIN
GENETISCHE EPIDEMIOLOGIE
Unser Team
AugUR
Lehre
Forschung
Publikationen
Software
GWAS summary statistics
MEDIZINISCHE SOZIOLOGIE
NAKO
PUBLIKATIONEN
VERANSTALTUNGEN
STELLENANGEBOTE
KONTAKT

Software

Regensburger GEM Plattform

The Genetic Epidemiology Unit

Downloads

Prof. Dr. Iris Heid, Dr. Thomas Winkler, Dr. Mathias Gorski, Felix Günther

MLA-bilateral (Günther et al. 2020)

EasyStrata (Winkler et al. 2014)

EasyQC (Winkler et al. 2014)

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**);

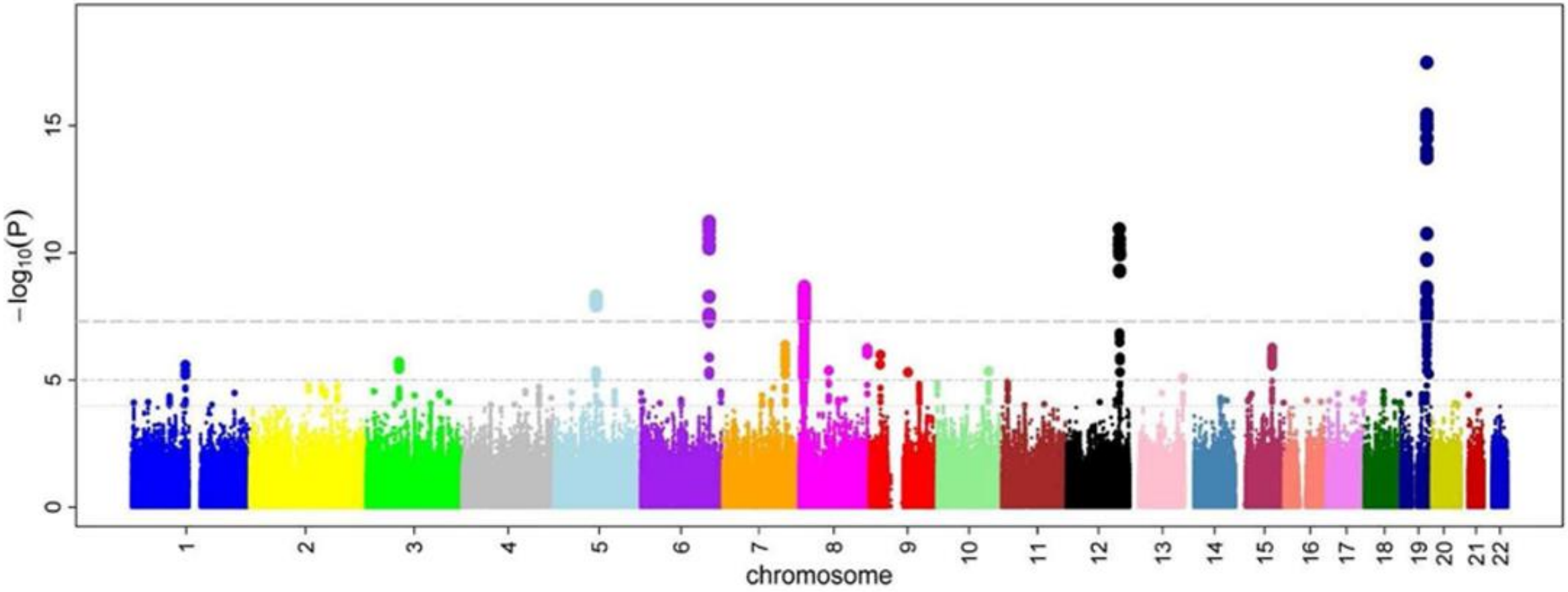
(iii) to simplify **data-handling** of large-scale GWA data-sets

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

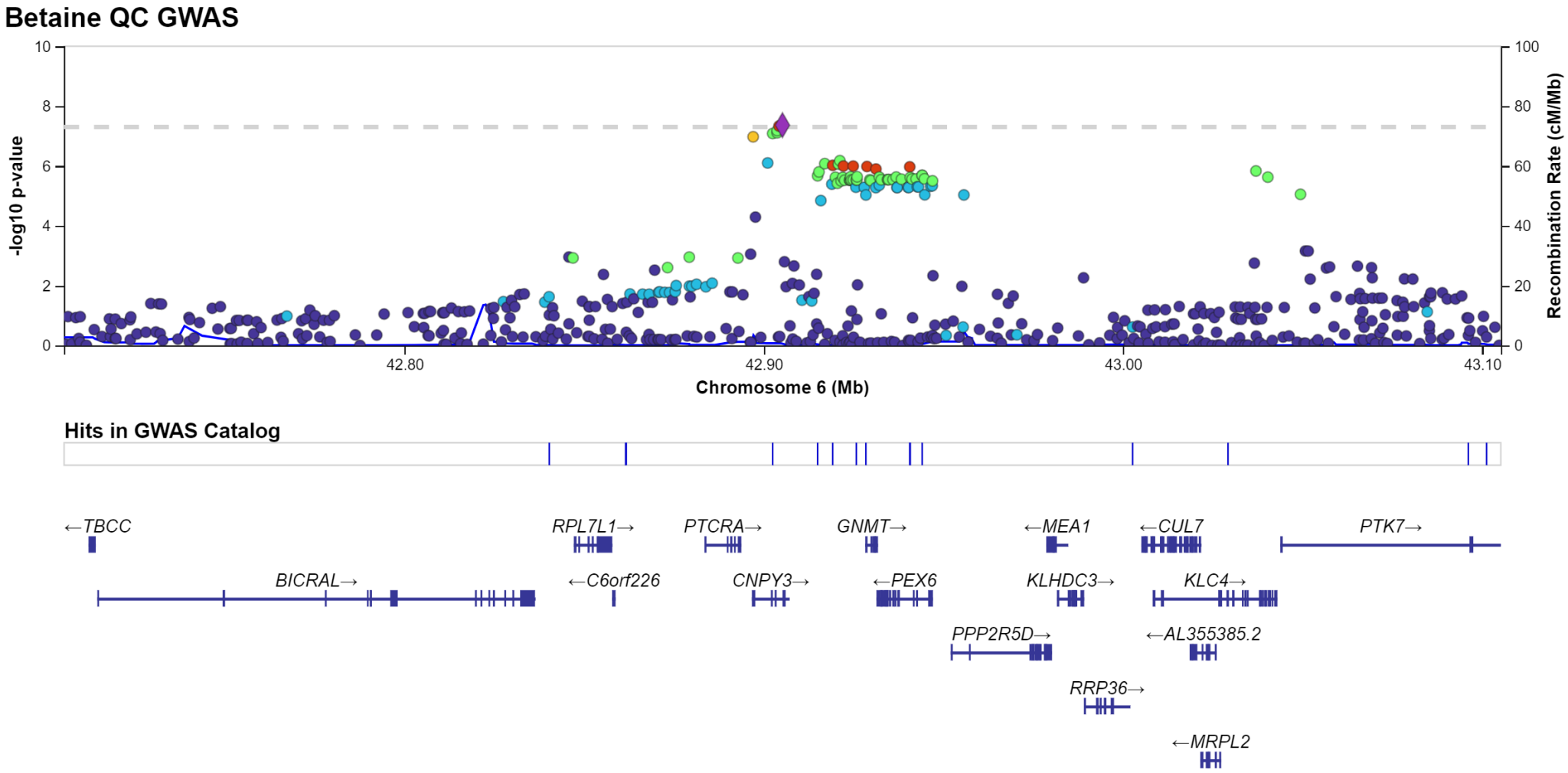
Outline

- Genome-wide association study
  - Introduction to rvtests software
  - Requirements: input files
  - Association analysis
  - Output file
- Post-association analysis steps
  - Quality control
  - Visualization

Manhattan plot



Locus Zoom


Betaine QC GWAS
-log10 p-value
Recombination Rate (cM/Mb)
Chromosome 6 (Mb)
42.80
42.90
43.00
43.10
Hits in GWAS Catalog
←TBCC
RPL7L1→
PTCRA→
GNMT→
←MEA1
←CUL7
PTK7→
BICRAL→
←C6orf226
CNPY3→
←PEX6
KLHDC3→
KLC4→
PPP2R5D→
←AL355385.2
RRP36→
←MRPL2

THANK YOU FOR YOUR ATTENTION!