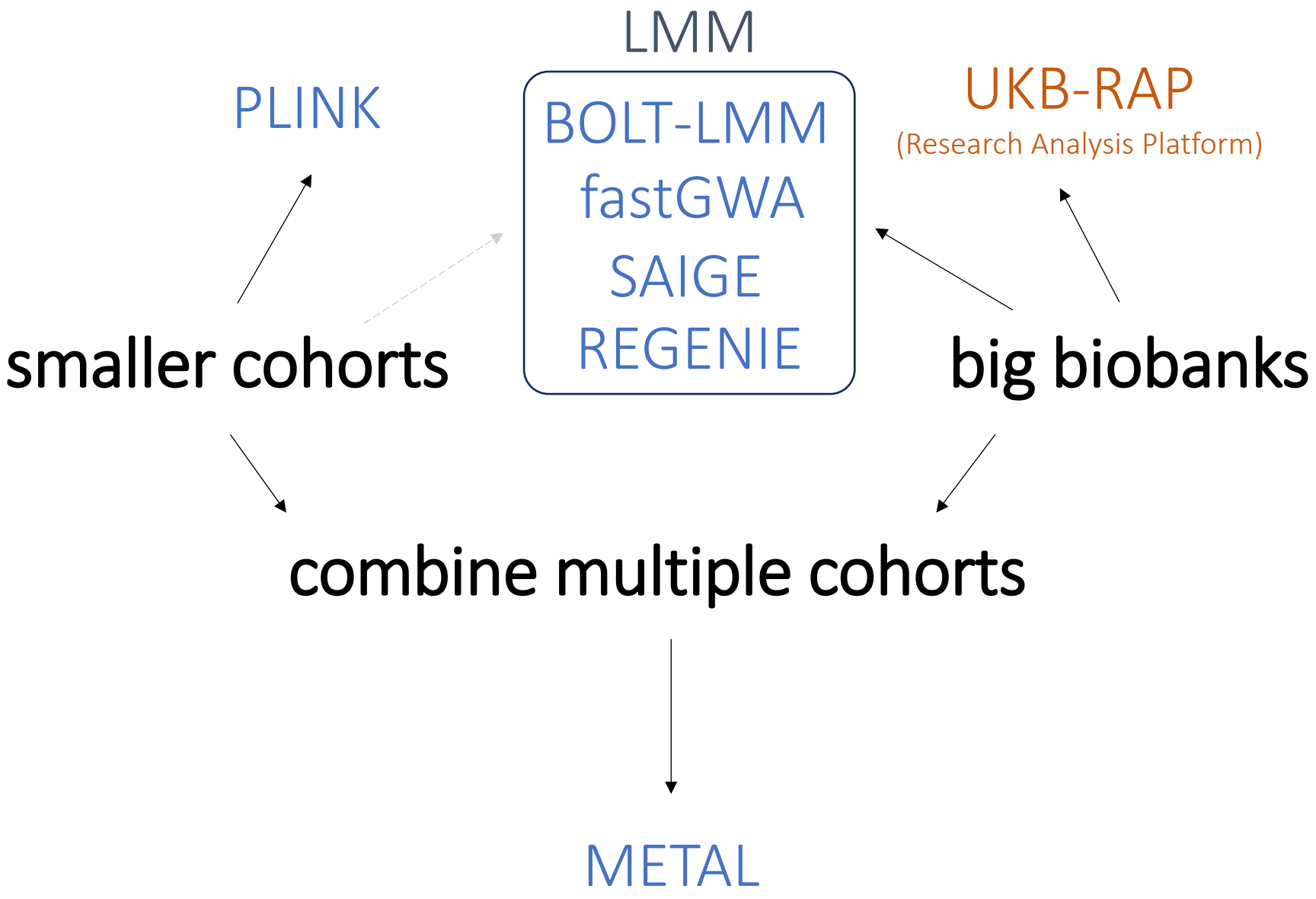
Genome-Wide Association Studies

Part 3: analyzing smaller cohorts, big biobanks, and cohorts combined

International Statistical Genetics Workshop – 2026

Abdel Abdellaoui

a.abdellaoui@amsterdamumc.nl


LMM
PLINK
BOLT-LMM
fastGWA
SAIGE
REGENIE
UKB-RAP
(Research Analysis Platform)
smaller cohorts
big biobanks
combine multiple cohorts
METAL

smaller cohorts (PLINK)

The American Journal of Human Genetics Volume 81 September 2007 REPORT

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham

PLINK – file formats

ped files

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

PLINK – file formats

ped files

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

Family ID
Individual ID
Paternal ID
Maternal ID
sex
phenotype

PLINK – file formats

ped files

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

Family ID

Individual ID

Paternal ID

Maternal ID

sex

phenotype

genotypes

PLINK – file formats

ped files

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

Family ID

Individual ID

Paternal ID

Maternal ID

sex

phenotype

genotypes

chromosome

SNP ID

genetic distance (cM)

base pair position

PLINK – file formats

ped files

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

binary files

*.fam

FID	IID	PID	MID	sex	p
1	1	3	4	2	165
1	2	3	4	2	173
2	5	7	8	1	179

*.bed

Compressed version of the .ped file, but only with information on genotypes. Not in a format readable for humans

*.bim

chr	SNP	GD	BP	A1	A2
1	rs1	0	870000	C	A
1	rs2	0	880000	C	G
1	rs3	0	890000	A	G

PLINK – file formats

ped files

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

binary files

*.fam

FID	IID	PID	MID	sex	p
1	1	3	4	2	165
1	2	3	4	2	173
2	5	7	8	1	179

*.bed

Compressed version of the .ped file, but only with information on genotypes. Not in a format readable for humans

*.bim

chr	SNP	GD	BP	A1	A2
1	rs1	0	870000	C	A
1	rs2	0	880000	C	G
1	rs3	0	890000	A	G

PLINK – file formats

ped files

as input: `--file`

as output: `--recode`

*.ped

FID	IID	PID	MID	sex	p	rs1	rs2	rs3	…
1	1	3	4	2	165	CC	CG	AA	…
1	2	3	4	2	173	CA	GG	AG	…
2	5	7	8	1	179	CA	CG	GG	…

*.map

chr	SNP	GD	BP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

binary files

as input: `--bfile`

as output: `--make-bed`

*.fam

FID	IID	PID	MID	sex	p
1	1	3	4	2	165
1	2	3	4	2	173
2	5	7	8	1	179

*.bed

Compressed version of the .ped file, but only with information on genotypes. Not in a format readable for humans

*.bim

chr	SNP	GD	BP	A1	A2
1	rs1	0	870000	C	A
1	rs2	0	880000	C	G
1	rs3	0	890000	A	G

Many other packages also read binary files
(all LMM approaches discussed later in this video)

binary files

*.fam

FID	IID	PID	MID	sex	p
1	1	3	4	2	165
1	2	3	4	2	173
2	5	7	8	1	179

*.bed

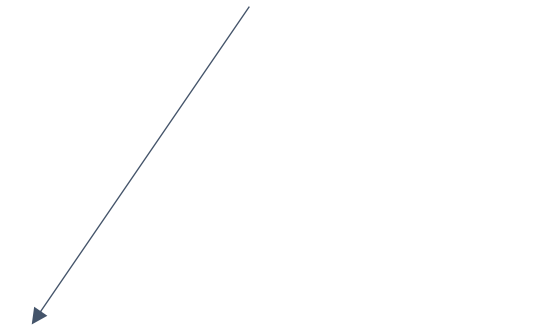
Compressed version of the .ped file, but only with information on genotypes. Not in a format readable for humans

*.bim

chr	SNP	GD	BP	A1	A2
1	rs1	0	870000	C	A
1	rs2	0	880000	C	G
1	rs3	0	890000	A	G

PLINK – regression analysis

```
plink --bfile cohort_name --linear --pheno phenotypes.txt --covar covs.txt --out gwas
```



binary files

*.fam

FID	IID	PID	MID	sex	p
1	1	3	4	2	165
1	2	3	4	2	173
2	5	7	8	1	179

*.bed

Compressed version of the .ped file, but only with information on genotypes. Not in a format readable for humans

*.bim

chr	SNP	GD	BP	A1	A2
1	rs1	0	870000	C	A
1	rs2	0	880000	C	G
1	rs3	0	890000	A	G

PLINK – regression analysis

```
plink --bfile cohort_name --linear --pheno phenotypes.txt --covar covs.txt --out gwas
```

FID	IID	p
1	1	165
1	2	173
2	5	179

Family ID

Individual ID

phenotype

FID	IID	cov1	cov2	cov3	cov4	cov5
1	1	2	31	0.01	0.01	0.03
1	2	2	27	0.03	0.01	0.02
2	5	1	46	0.02	0.02	0.02

Family ID

Individual ID

covariates

PLINK – regression analysis

```
plink --bfile cohort_name --linear --pheno phenotypes.txt --covar covs.txt --out gwas
```

FID	IID	p
1	1	165
1	2	173
2	5	179

Family ID

Individual ID

phenotype

FID	IID	cov1	cov2	cov3	cov4	cov5
1	1	2	31	0.01	0.01	0.03
1	2	2	27	0.03	0.01	0.02
2	5	1	46	0.02	0.02	0.02

Family ID

Individual ID

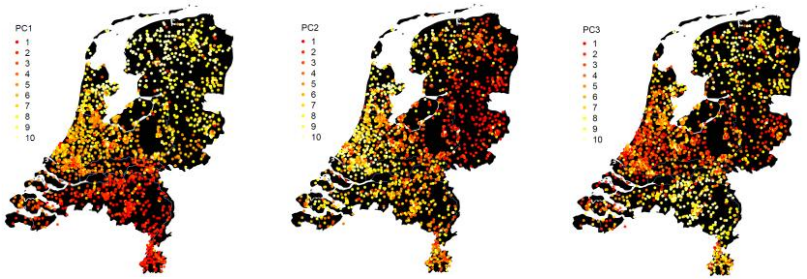
sex

age

PC1

PC2

PC3



PLINK – regression analysis

```
plink --bfile cohort_name --linear --pheno phenotypes.txt --covar covs.txt --out gwas
```

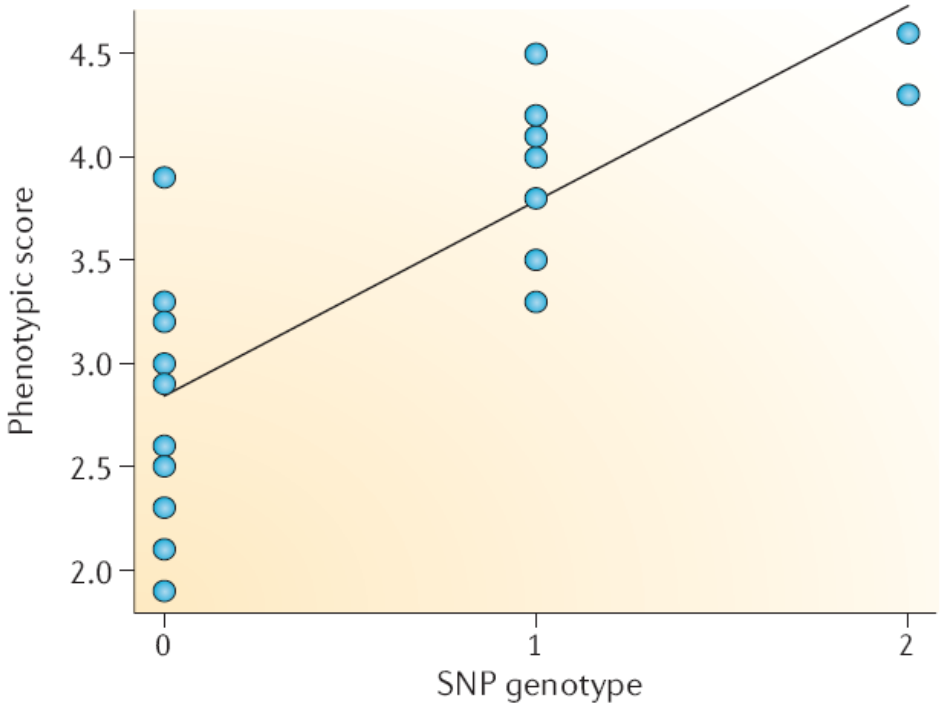
$$y = X\beta + \varepsilon$$

phenotype vector

fixed effects
(SNP + covariates)
β_{SNP} = what we are interested in

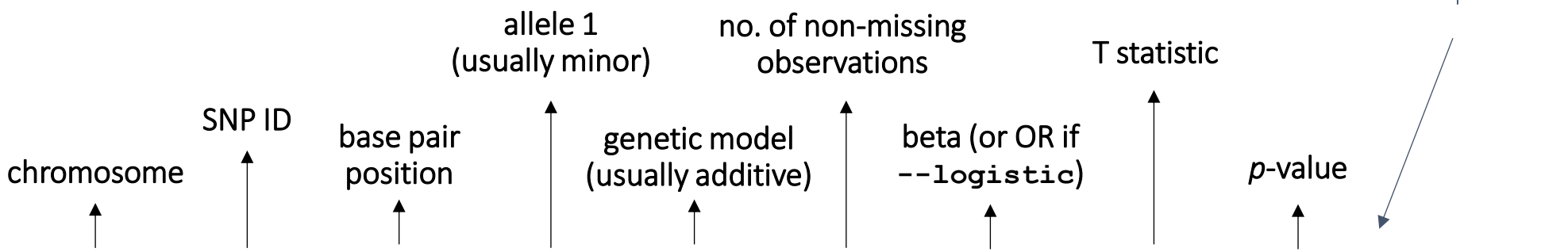
$\sim N(0, \sigma^2 I)$

residual error
(independent)



PLINK – regression analysis

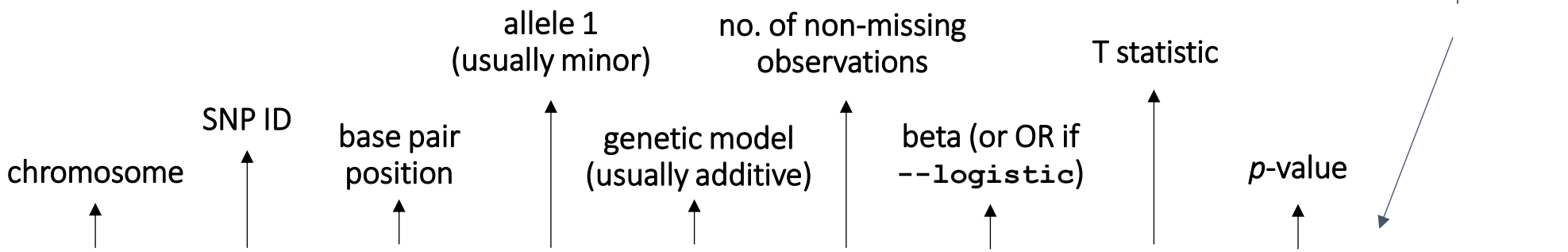
```
plink --bfile cohort_name --linear --pheno phenotypes.txt --covar covs.txt --out gwas
```



CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1	rs1	870000	C	ADD	4987	0.0312	1.847	0.0648
1	rs2	880000	C	ADD	4991	0.1843	6.214	3.2e-09
1	rs3	890000	A	ADD	4983	-0.0071	-0.423	0.6724
... millions more rows ...								

PLINK – regression analysis

```
plink --bfile cohort_name --linear --pheno phenotypes.txt --covar covs.txt --out gwas
```



CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1	rs1	870000	C	ADD	4987	0.0312	1.847	0.0648
1	rs2	880000	C	ADD	4991	0.1843	6.214	3.2e-09
1	rs3	890000	A	ADD	4983	-0.0071	-0.423	0.6724
... millions more rows ...								

Note that PLINK 1.9 doesn't output the standard error (SE) directly (PLINK 2.0 does), but you'll need it for meta-analysis. You can calculate it as: **SE = BETA / STAT**

analyzing big biobanks using linear mixed modelling (LMM)

The linear mixed model (LMM)

$$y = X\beta + g + \varepsilon$$

phenotype vector

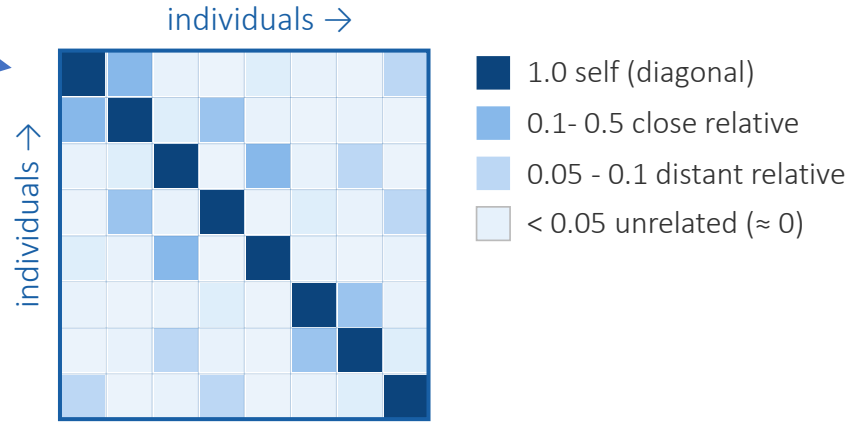
fixed effects
(SNP + covariates)
β_{SNP} = what we are interested in

~ N(0, σ^2 GRM)

genetic random effect, captures relatedness between all individuals

~ N(0, σ^2 I)

residual error
(independent)



Genetic Relationship Matrix (GRM)

Each cell = relatedness between two individuals estimated from genome-wide SNP data

The linear mixed model (LMM)

$$y = X\beta + g + \varepsilon$$

phenotype vector

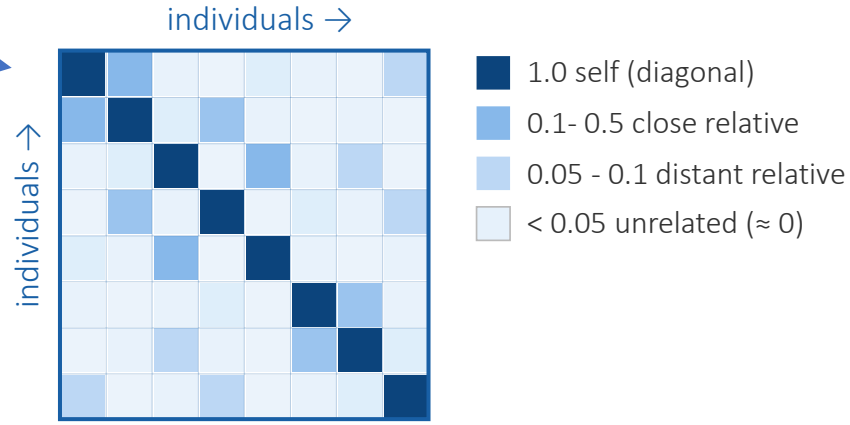
fixed effects
(SNP + covariates)
β_{SNP} = what we are interested in

~ N(0, σ² GRM)

genetic random effect, captures relatedness between all individuals

~ N(0, σ² I)

residual error
(independent)



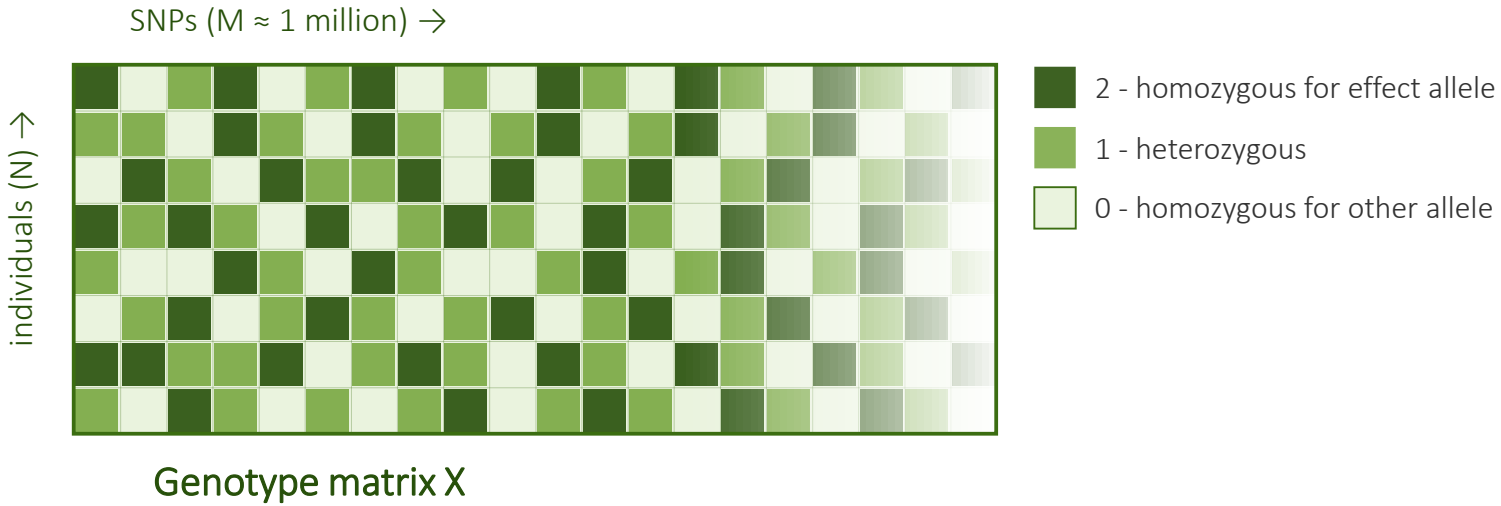
Genetic Relationship Matrix (GRM)

Each cell = relatedness between two individuals estimated from genome-wide SNP data

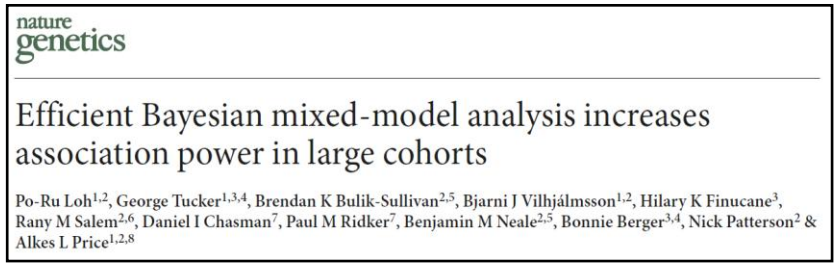
- Good for big biobanks, as they will have many unknown close genetic relationships
- However, can get *very* computationally intensive with big datasets:
  500,000 × 500,000 = **250 billion** cells

BOLT-LMM

- First LMM approach scalable to biobank-scale data (>100k)
- Avoids computing the full N×N GRM by working directly with the genotype matrix
- Still corrects for genome-wide genetic relatedness, but through computationally feasible route
- Designed primarily for quantitative traits



2015

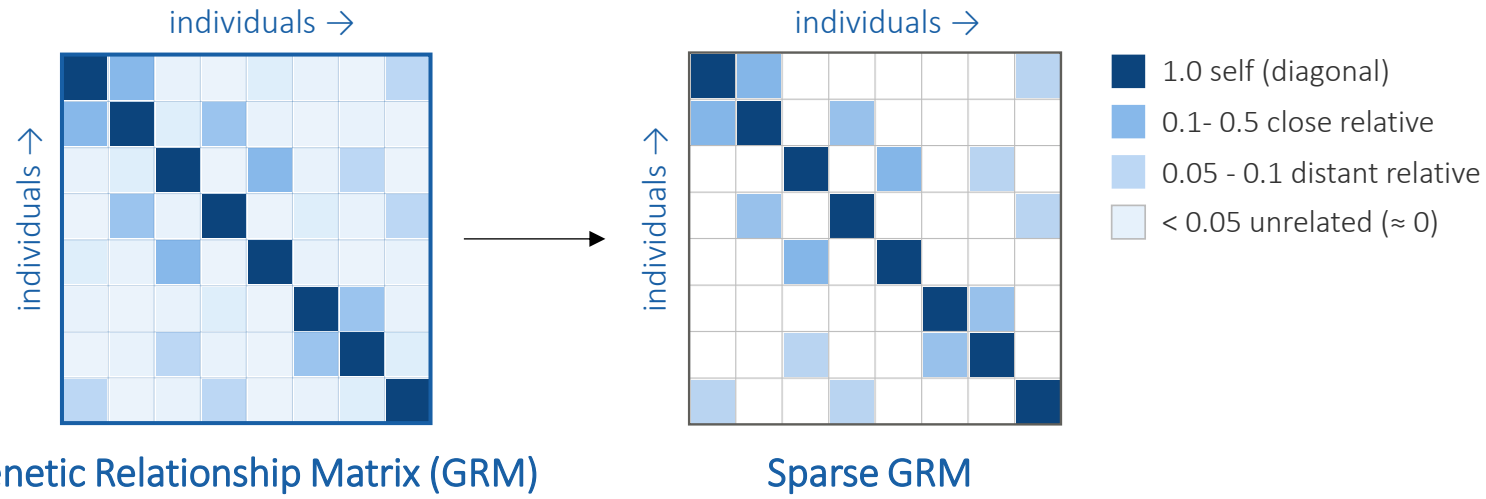

nature genetics

Efficient Bayesian mixed-model analysis increases association power in large cohorts

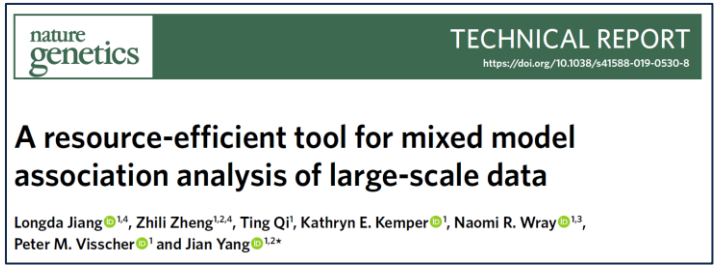
Po-Ru Loh[1,2], George Tucker[1,3,4], Brendan K Bulik-Sullivan[2,5], Bjarni J Vilhjálmsson[1,2], Hilary K Finucane[3], Rany M Salem[2,6], Daniel I Chasman[7], Paul M Ridker[7], Benjamin M Neale[2,5], Bonnie Berger[3,4], Nick Patterson[2] & Alkes L Price[1,2,8]

fastGWA

- Uses a sparse GRM, which only retains relatedness estimates above a threshold (~0.05)
- Handles both quantitative and binary traits
- Part of GCTA

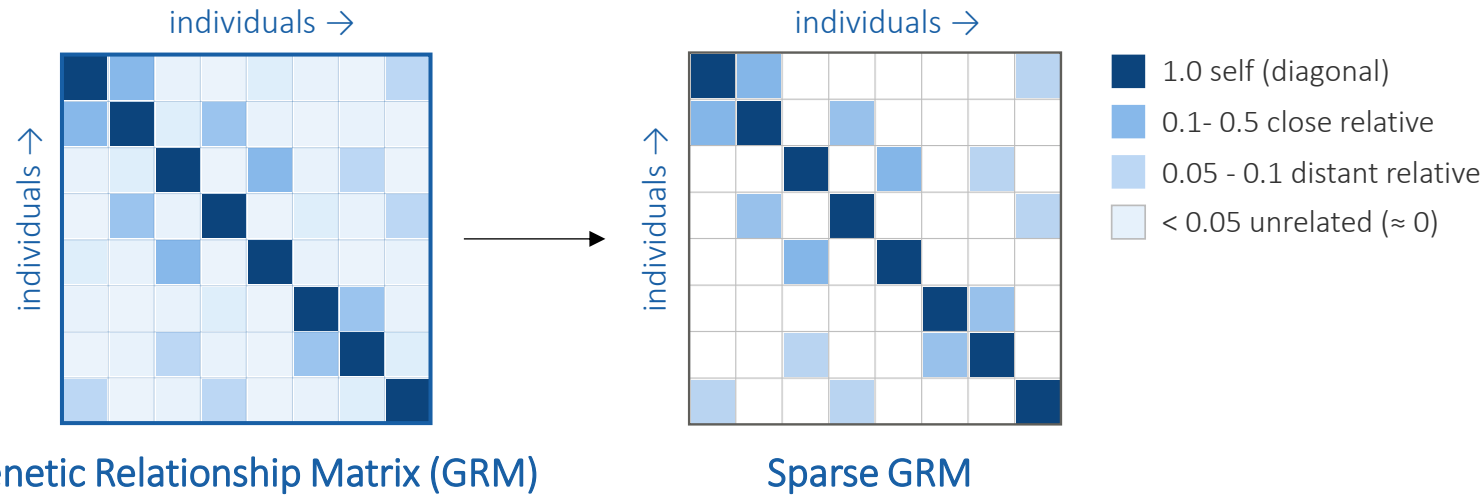


2019



SAIGE

- Extends LMM to binary traits with severe case-control imbalance
- Handles rare cases with an approach that doesn't assume a normal distribution of the test statistic (saddlepoint approximation)
- For computational efficiency, runs a two-step analysis:
  1. Fit null model (= sparse GRM + covariates, *not* the SNP effect) → slow, but needed only once
  2. Test each SNP against null model → fast



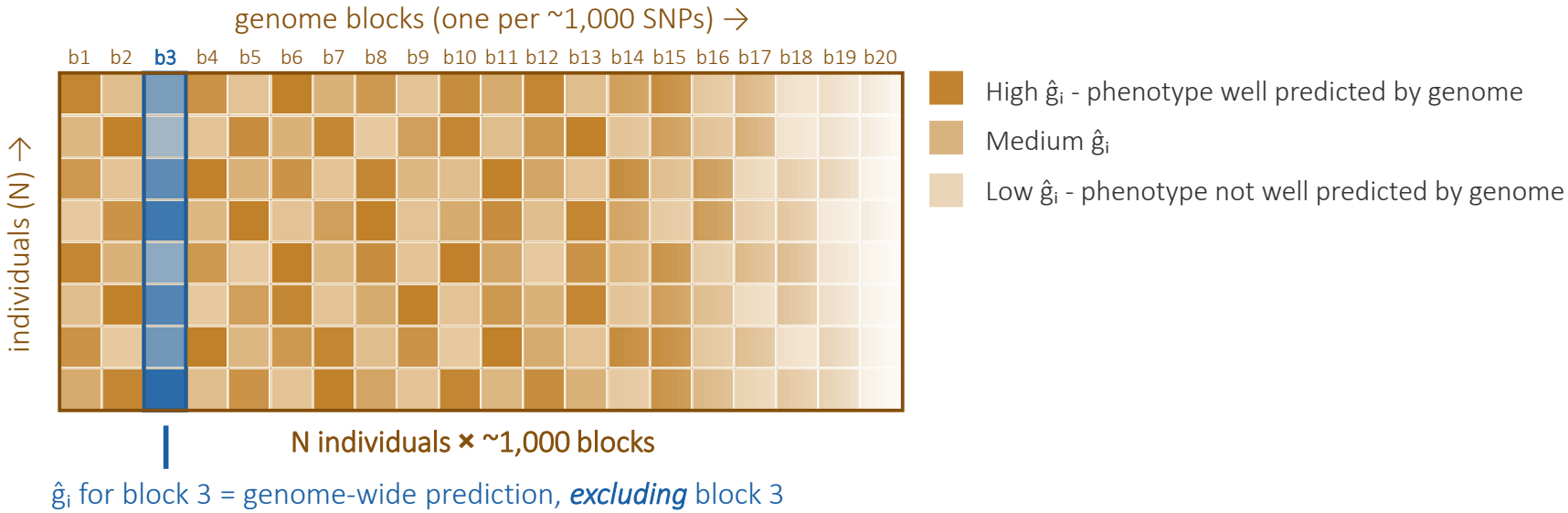
Genetic Relationship Matrix (GRM)

Sparse GRM

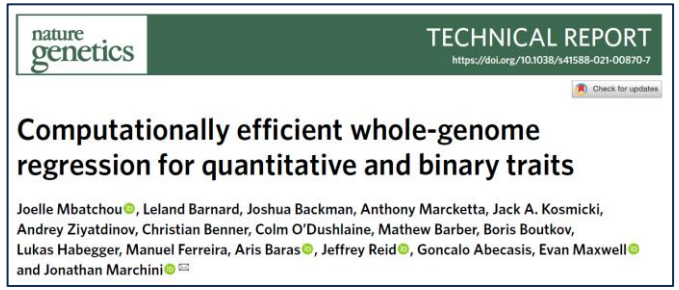
2018

REGENIE

- Control for “genomic background” with two-step procedure:
  1. Predict phenotype from whole genome, except for 1 block of ~1000 SNPs, creating one “genomic background score” $\hat{g}_i$ per person per block
  2. Test each SNP using the $\hat{g}_i$ that excludes its block as a covariate. The SNP effect is estimated above and beyond the genomic background captured by $\hat{g}_i$
- Also robust for rare outcomes (handles binary traits with Firth regression)



2021



UK Biobank Research Analysis Platform (RAP)

- Since 2024, all new UKB project data is only accessible via RAP
- Cloud-based platform (DNAnexus / AWS), data stays in the cloud
- Jobs can be submitted via a browser UI or the dx command-line tool
- All tools from this video are available as RAP apps:
  - PLINK2 for standard analyses
  - BOLT-LMM, fastGWA, REGENIE for large-scale LMM analyses

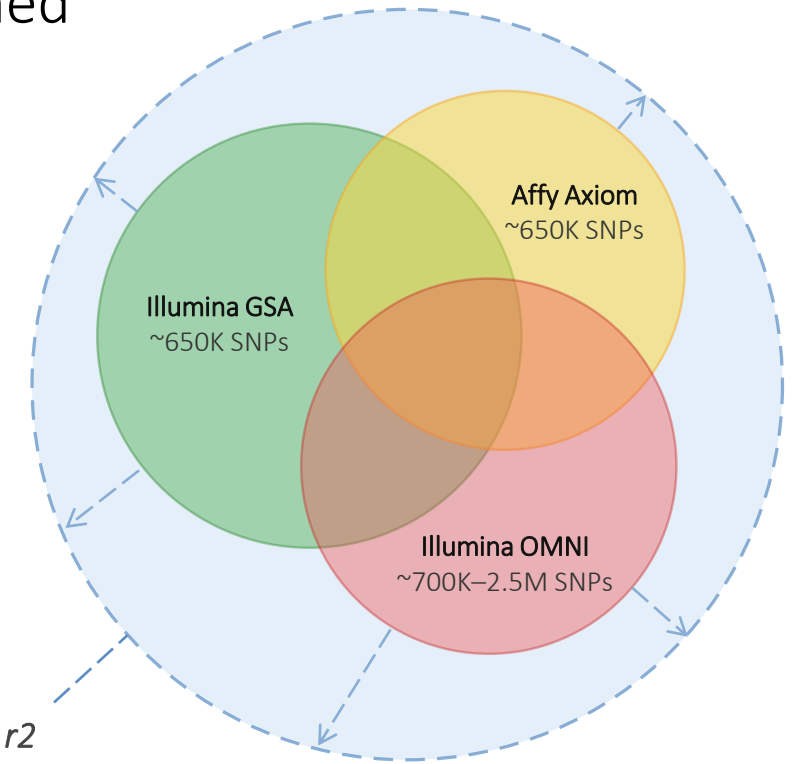
https://dnanexus.gitbook.io/uk-biobank-rap

https://www.dnanexus.com/partnerships/ukbiobank

combining multiple cohorts through meta-analysis

Imputation re-cap

- Why is imputation necessary for meta-analysis:
  - It allows data from different DNA chips to be combined



Imputation re-cap

- Why is imputation necessary for meta-analysis:
  - It allows data from different DNA chips to be combined
  - It allows for untyped variation to be tested

step 1 — genotyping

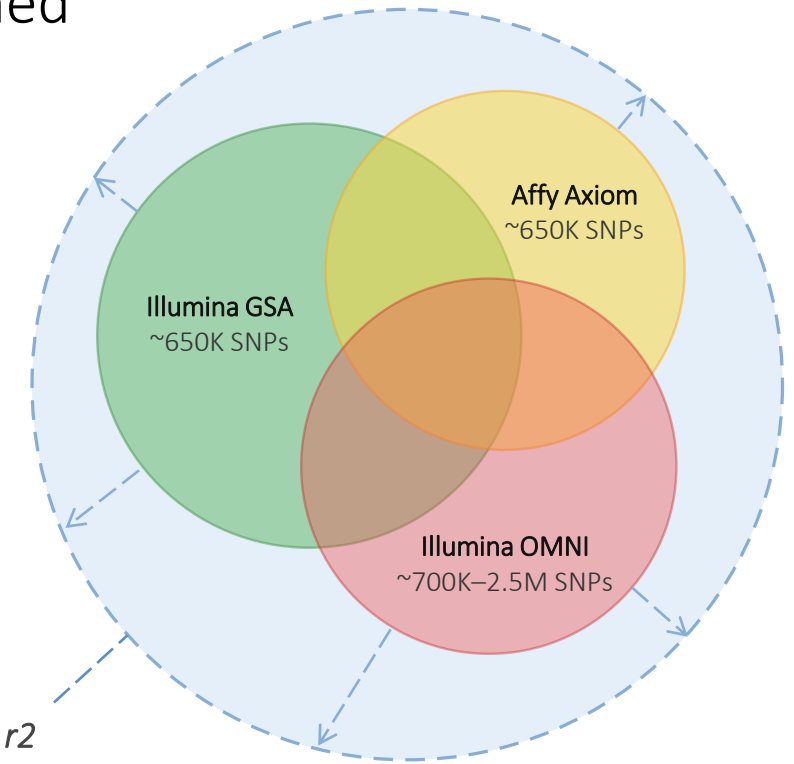
Genotyped sample

. . C . G . . C .

Reference haplotypes

A C T T C T G A C
A G C T C T A G T
T A T C G A C C G
T A A C T T G A T

3 of 9 variants genotyped



TOPMed r2
~300 million variants
~7–15 million SNPs after QC

Imputation re-cap

- Why is imputation necessary for meta-analysis:
  - It allows data from different DNA chips to be combined
  - It allows for untyped variation to be tested

step 2 — haplotype matching

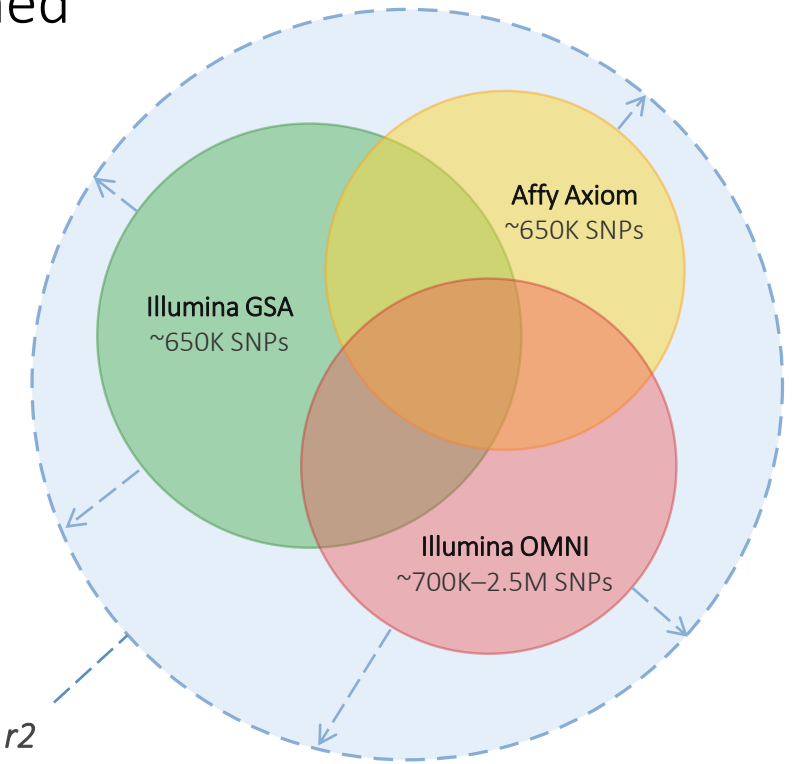
Genotyped sample

·	·	C	·	G	·	·	C	·

Reference haplotypes

A	C	T	T	C	T	G	A	C
A	G	C	T	C	T	A	G	T
T	A	T	C	G	A	C	C	G
T	A	A	C	T	T	G	A	T

two haplotypes explain the typed variants



TOPMed r2
~300 million variants
~7–15 million SNPs after QC

Imputation re-cap

- Why is imputation necessary for meta-analysis:
  - It allows data from different DNA chips to be combined
  - It allows for untyped variation to be tested

step 3 — imputation

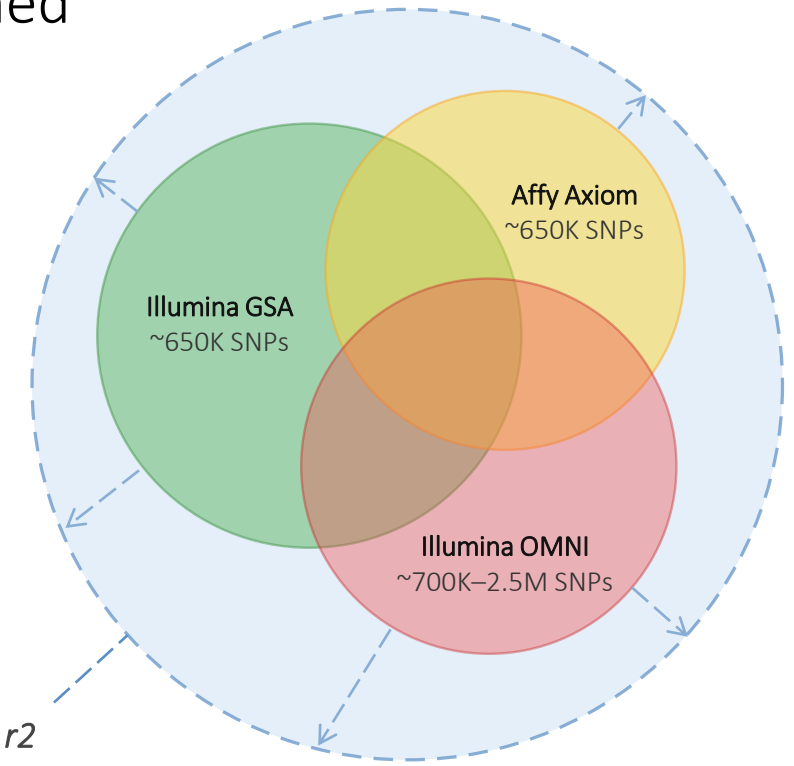
Imputed sample

A	G	C	T	G	A	C	C	G

Reference haplotypes

A	C	T	T	C	T	G	A	C
A	G	C	T	C	T	A	G	T
T	A	T	C	G	A	C	C	G
T	A	A	C	T	T	G	A	T

all 9 variants now known



TOPMed r2
~300 million variants
~7–15 million SNPs after QC

imputed genotypes

imputation is probabilistic:
we're not certain what the true untyped genotype is

imputed genotypes

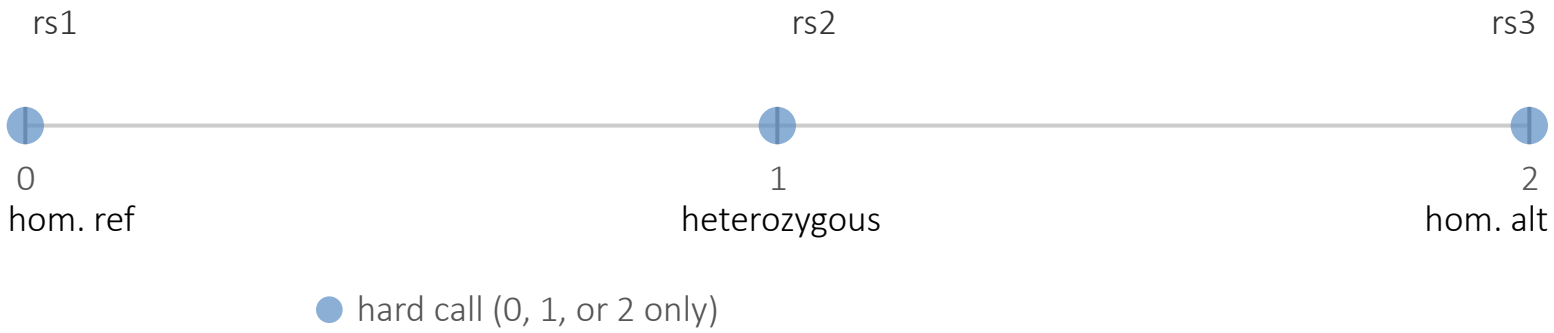
imputation is probabilistic:
we're not certain what the true untyped genotype is

hard calls discard this:
rounding 0.9 → 1 loses information and reduces power

hard calls

SNP	genotype	value
rs1	AA	0
rs2	Aa	1
rs3	aa	2

only values: 0, 1, 2



imputed genotypes

dosages = expected number of effect alleles

imputation is probabilistic:
we're not certain what the true untyped genotype is

hard calls discard this:
rounding 0.9 → 1 loses information and reduces power

a dosage encodes this uncertainty:
expected number of effect alleles, ranging continuously from 0 to 2

hard calls

SNP	genotype	value
rs1	AA	0
rs2	Aa	1
rs3	aa	2

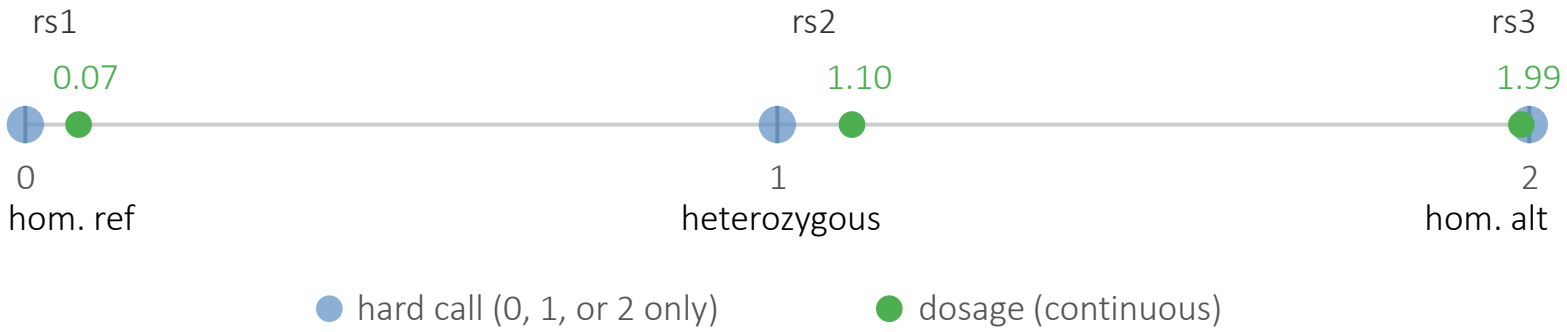
only values: 0, 1, 2

dosages

SNP	P(AA)	P(Aa)	P(aa)	dosage
rs1	0.94	0.05	0.01	0.07
rs2	0.08	0.74	0.18	1.10
rs3	0.01	0.03	0.96	1.99

any value between 0.0 and 2.0

$$\text{dosage} = P(Aa) + 2 \times P(aa)$$



imputed genotypes

dosages = expected number of effect alleles

imputation is probabilistic:
we're not certain what the true untyped genotype is

hard calls discard this:
rounding 0.9 → 1 loses information and reduces power

a dosage encodes this uncertainty:
expected number of effect alleles, ranging continuously from 0 to 2

REGENIE, BOLT-LMM, SAIGE
all work natively with dosages

hard calls

SNP	genotype	value
rs1	AA	0
rs2	Aa	1
rs3	aa	2

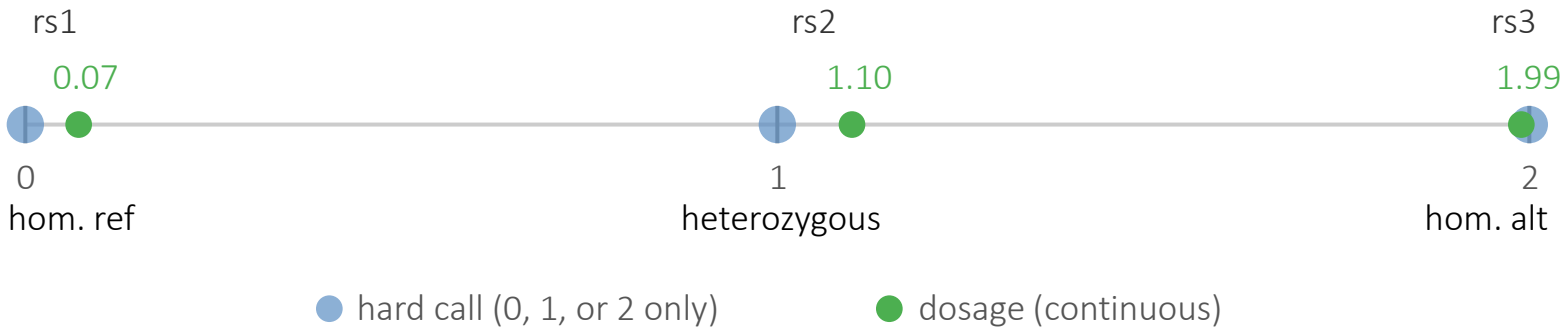
only values: 0, 1, 2

dosages

SNP	P(AA)	P(Aa)	P(aa)	dosage
rs1	0.94	0.05	0.01	0.07
rs2	0.08	0.74	0.18	1.10
rs3	0.01	0.03	0.96	1.99

any value between 0.0 and 2.0

$$\text{dosage} = P(Aa) + 2 \times P(aa)$$



imputed genotypes

dosages = expected number of effect alleles

imputation is probabilistic:
we're not certain what the true untyped genotype is

hard calls discard this:
rounding 0.9 → 1 loses information and reduces power

a dosage encodes this uncertainty:
expected number of effect alleles, ranging continuously from 0 to 2

REGENIE, BOLT-LMM, SAIGE
all work natively with dosages

hard calls

SNP	genotype	value
rs1	AA	0
rs2	Aa	1
rs3	aa	2

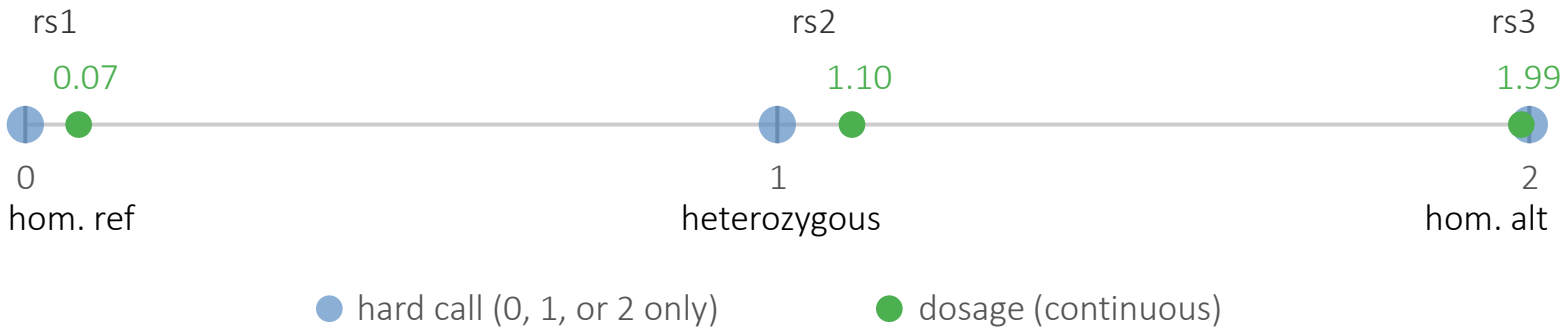
only values: 0, 1, 2

dosages

SNP	P(AA)	P(Aa)	P(aa)	dosage
rs1	0.94	0.05	0.01	0.07
rs2	0.08	0.74	0.18	1.10
rs3	0.01	0.03	0.96	1.99

any value between 0.0 and 2.0

dosage = P(Aa) + 2 × P(aa)



imputation quality (r^2 or INFO score) measures how certain the dosage is

imputed genotypes

dosages = expected number of effect alleles

imputation is probabilistic:
we're not certain what the true untyped genotype is

hard calls discard this:
rounding 0.9 → 1 loses information and reduces power

a dosage encodes this uncertainty:
expected number of effect alleles, ranging continuously from 0 to 2

REGENIE, BOLT-LMM, SAIGE
all work natively with dosages

note: for the practical, we will use hard calls for simplicity; in real analyses, dosages will have more power

hard calls

SNP	genotype	value
rs1	AA	0
rs2	Aa	1
rs3	aa	2

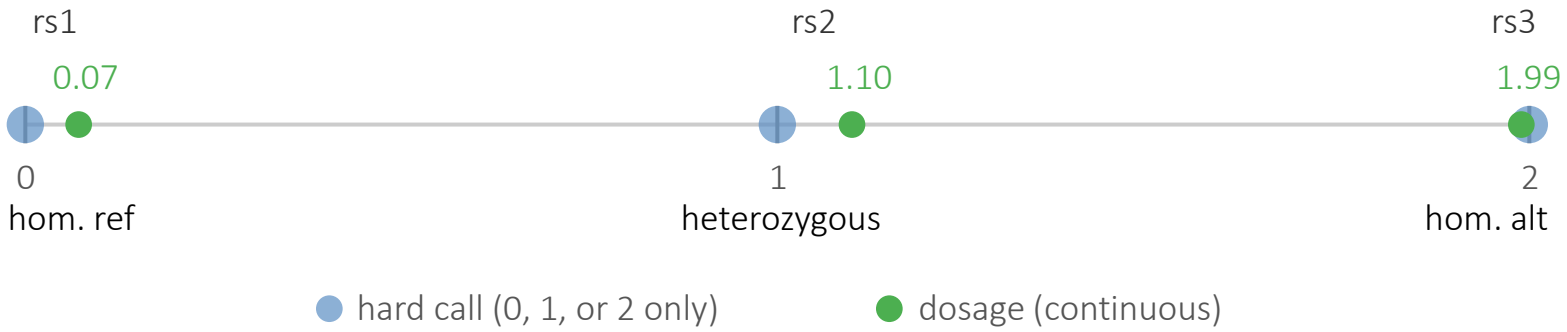
only values: 0, 1, 2

dosages

SNP	P(AA)	P(Aa)	P(aa)	dosage
rs1	0.94	0.05	0.01	0.07
rs2	0.08	0.74	0.18	1.10
rs3	0.01	0.03	0.96	1.99

any value between 0.0 and 2.0

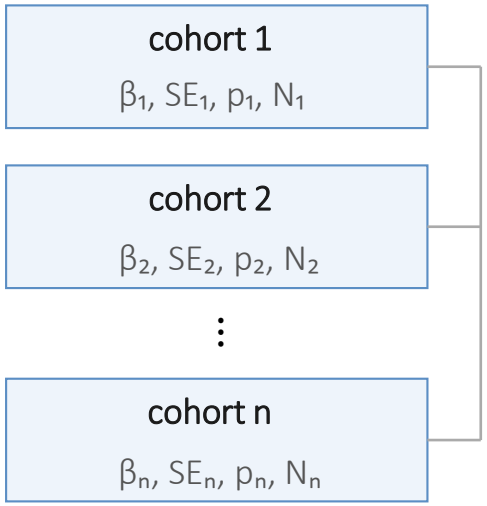
$$\text{dosage} = P(Aa) + 2 \times P(aa)$$



imputation quality (r^2 or INFO score) measures how certain the dosage is

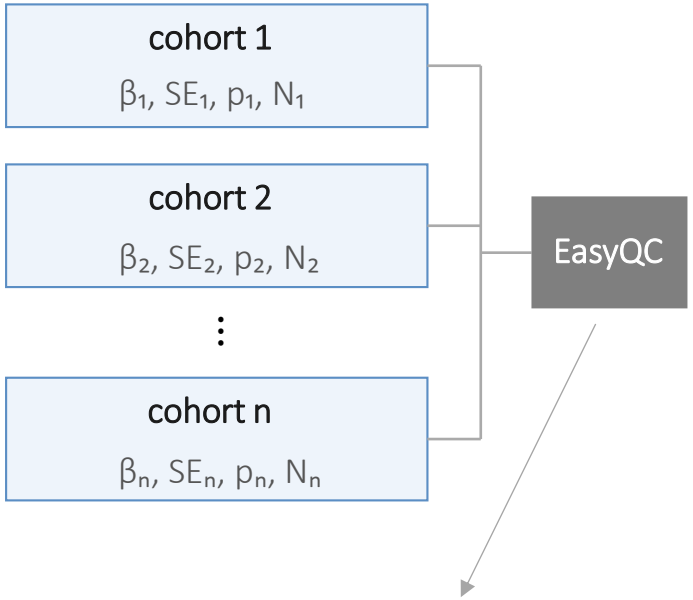
GWAS meta-analysis

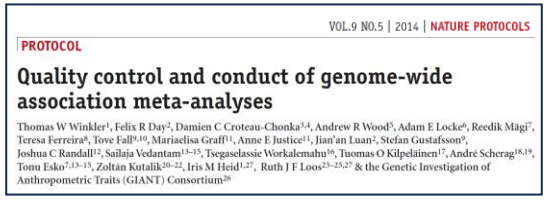
- each cohort runs the same analysis protocol
- shares summary statistics (β, SE, *p*) per SNP with central analyst



GWAS meta-analysis

- each cohort runs the same analysis protocol
- shares summary statistics (β, SE, *p*) per SNP with central analyst
- central analyst does additional QC





VOL.9 NO.5 | 2014 | NATURE PROTOCOLS

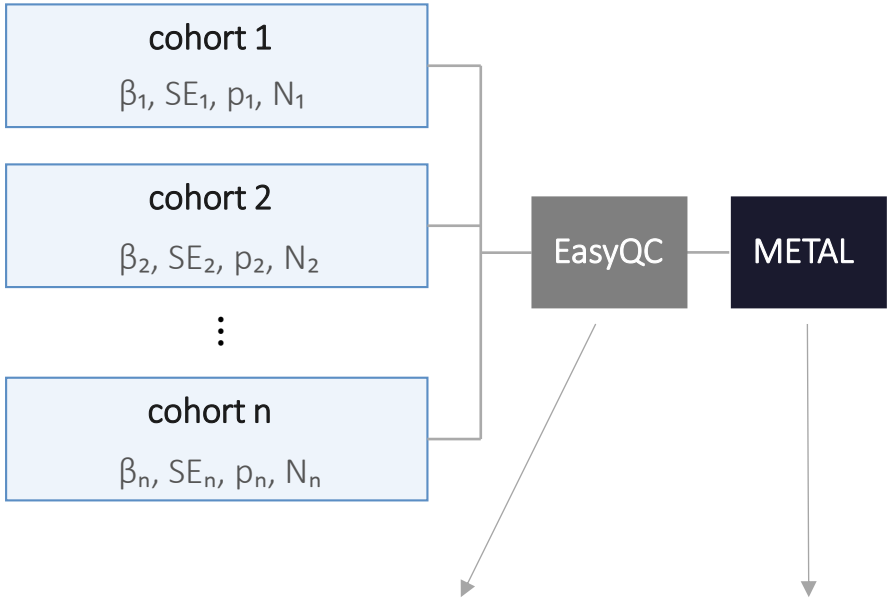
PROTOCOL

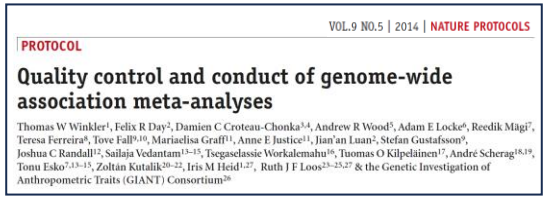
Quality control and conduct of genome-wide association meta-analyses

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[13–15], Tsegaselassie Workalemahu[16], Tuomas O Kilpeläinen[17], André Scherag[18,19], Tonu Esko[7,13–15], Zoltán Kutalik[20–22], Iris M Heid[1,27], Ruth J F Loos[23–25,27] & the Genetic Investigation of Anthropometric Traits (GIANT) Consortium[26]

GWAS meta-analysis

- each cohort runs the same analysis protocol
- shares summary statistics (β, SE, *p*) per SNP with central analyst
- central analyst does additional QC and meta-analyzes



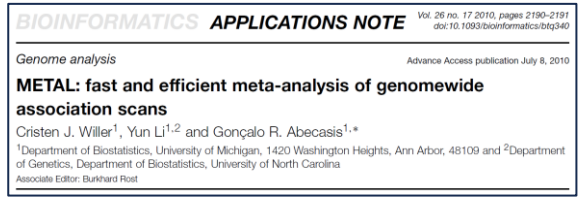


VOL.9 NO.5 | 2014 | NATURE PROTOCOLS

PROTOCOL

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[13–15], Tsegaselassie Workalemahu[16], Tuomas O Kilpeläinen[17], André Scherag[18,19], Tonu Esko[7,13–15], Zoltán Kutalik[20–22], Iris M Heid[1,27], Ruth J F Loos[23–25,27] & the Genetic Investigation of Anthropometric Traits (GIANT) Consortium[26]



BIOINFORMATICS **APPLICATIONS NOTE**

Vol. 26 no. 17 2010, pages 2190–2191
doi:10.1093/bioinformatics/btq340

Genome analysis

Advance Access publication July 8, 2010

METAL: fast and efficient meta-analysis of genomewide association scans

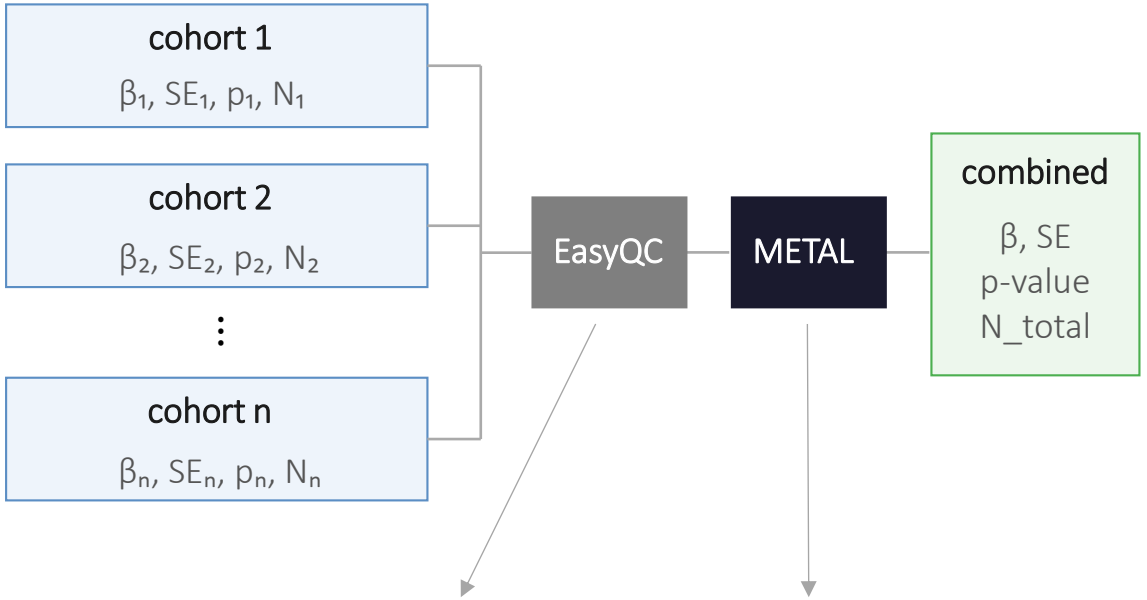
Cristen J. Willer[1], Yun Li[1,2] and Gonçalo R. Abecasis[1,*]

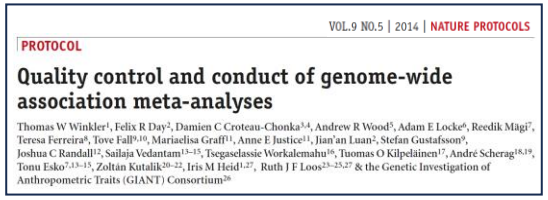
[1]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, 48109 and [2]Department of Genetics, Department of Biostatistics, University of North Carolina

Associate Editor: Burkhard Rost

GWAS meta-analysis

- each cohort runs the same analysis protocol
- shares summary statistics (β, SE, *p*) per SNP with central analyst
- central analyst does additional QC and meta-analyzes

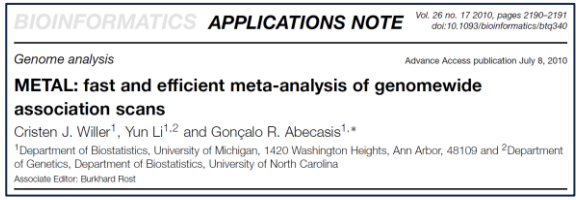



VOL.9 NO.5 | 2014 | NATURE PROTOCOLS

PROTOCOL

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[13–15], Tsegaselassie Workalemahu[16], Tuomas O Kilpeläinen[17], André Scherag[18,19], Tonu Esko[7,13–15], Zoltán Kutalik[20–22], Iris M Heid[1,27], Ruth J F Loos[23–25,27] & the Genetic Investigation of Anthropometric Traits (GIANT) Consortium[26]


BIOINFORMATICS **APPLICATIONS NOTE** Vol. 26 no. 17 2010, pages 2190–2191 doi:10.1093/bioinformatics/btq340

Genome analysis Advance Access publication July 8, 2010

METAL: fast and efficient meta-analysis of genomewide association scans

Cristen J. Willer[1], Yun Li[1,2] and Gonçalo R. Abecasis[1,*]

[1]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, 48109 and [2]Department of Genetics, Department of Biostatistics, University of North Carolina

Associate Editor: Burkhard Rost

Why meta-analysis?

raw data often cannot be shared
consent restrictions, data access agreements, practical barriers

combining cohorts substantially boosts power
equivalent to joint analysis for common variants

reveals whether effects are consistent across cohorts
heterogeneity between studies can be biologically informative or a warning sign

METAL - two main approaches

inverse variance weighted

when to use

phenotype on the same scale across cohorts
(preferred — outputs interpretable β estimates)

input required

β_i (effect size per cohort i)
SE_i (standard error per cohort i)
alleles (A1, A2)

formulas

$w_i = 1 / SE_i^2$

$\beta = \Sigma(\beta_i \cdot w_i) / \Sigma w_i$

$SE = \sqrt{(1 / \Sigma w_i)}$

$Z = \beta / SE$

$p = 2\Phi(-|Z|)$

weight = precision of each cohort estimate

sample size weighted

when to use

different phenotype definitions or transformations
(no interpretable β — outputs Z-score only)

input required

p_i (p-value per cohort i)
Δ_i (direction of effect per cohort i)
N_i (sample size per cohort i)
alleles (A1, A2)

formulas

$w_i = \sqrt{N_i}$

$Z_i = \Phi^{-1}(p_i/2) \cdot \text{sign}(\Delta_i)$

$Z = \Sigma(Z_i \cdot w_i) / \sqrt{\Sigma w_i^2}$

$p = 2\Phi(-|Z|)$

weight = sample size of each cohort

running METAL in practice

METAL runs from a script file specifying column names and input files

input: what each cohort provides (one row per SNP)

SNP	A1	A2	BETA	SE	P	N
rs1234567	C	A	0.031	0.017	0.065	4987
rs7654321	C	G	0.184	0.030	3.2e-09	4991

key output columns

MarkerName	SNP identifier
Effect / StdErr	combined β and SE
P-value	meta-analytic p-value
Direction	+/− per cohort (e.g. +++−)
HetISq	I^2 heterogeneity statistic
HetPVal	p-value for heterogeneity test
TotalSampleSize	N across all cohorts
Freq1	pooled allele frequency

quality checks

Manhattan + QQ plot
inspect for inflation and spurious signals

Direction column
all same sign (+++ or −−−) = consistent effect
mixed signs (++−) = heterogeneous

Heterogeneity: I^2 and HetPVal
$I^2 > 75\%$ = substantial heterogeneity
check for strand issues, analysis protocol errors

Strand flipping
METAL handles automatically via allele freq comparison
remove A/T and C/G ambiguous SNPs beforehand

Next video: gene-environment correlations and assortative mating