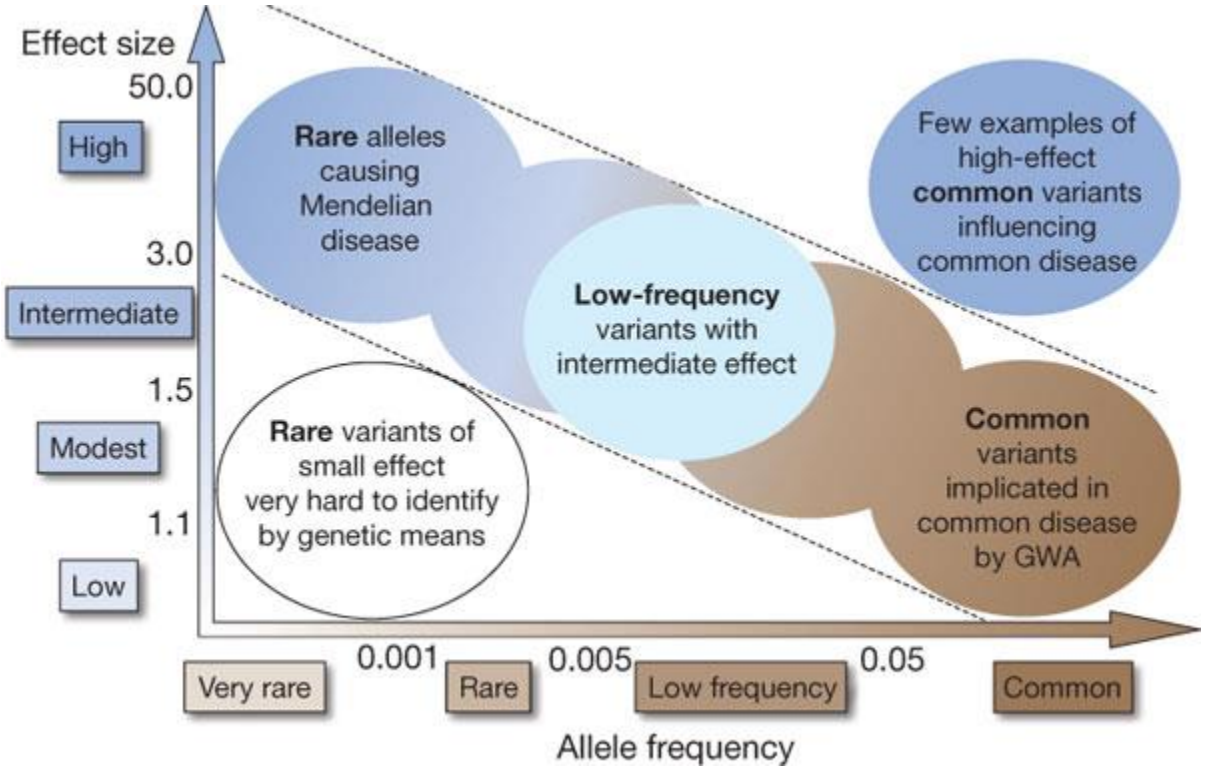
Genome-wide association studies

Dr. Martina Müller-Nurasyid

IMBEI, Genomische Statistik und Bioinformatik
Universitätsmedizin Mainz

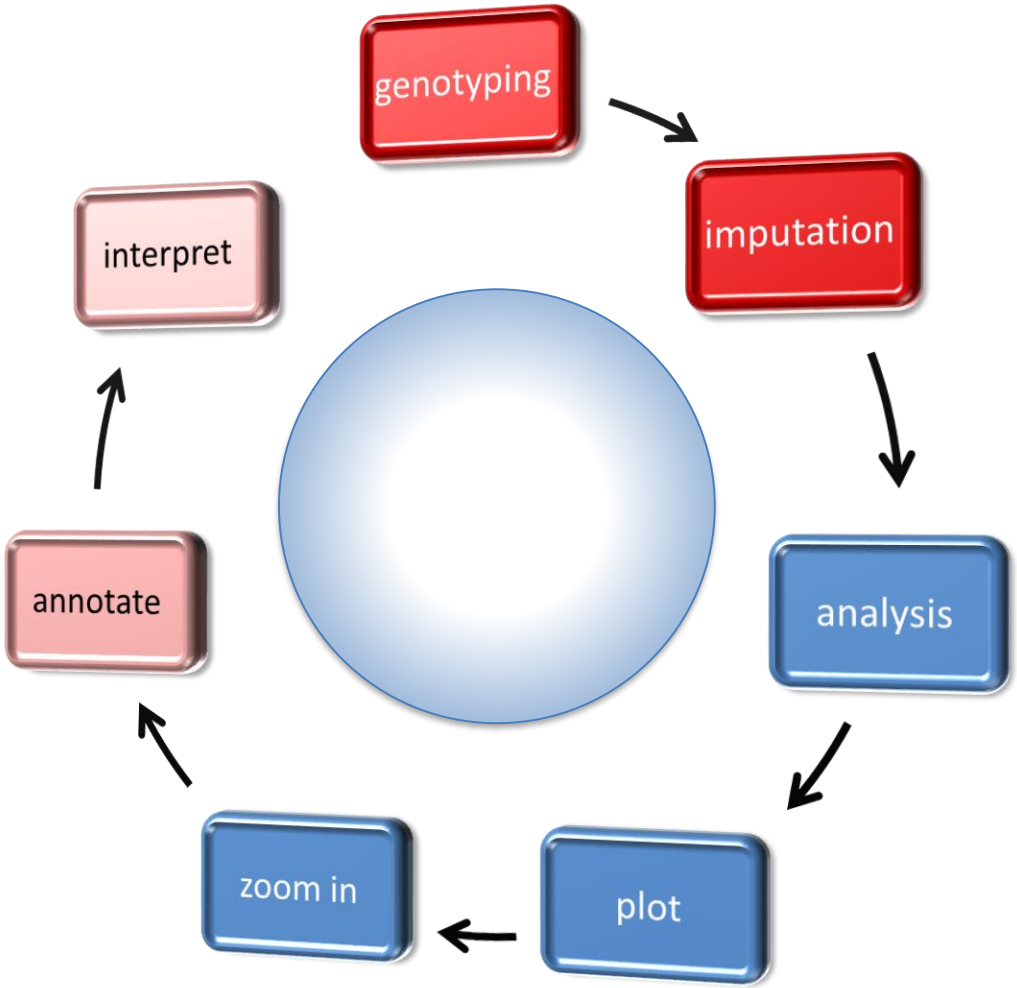
01.06.2022

From DNA to function
Feasibility of identifying genetic effects

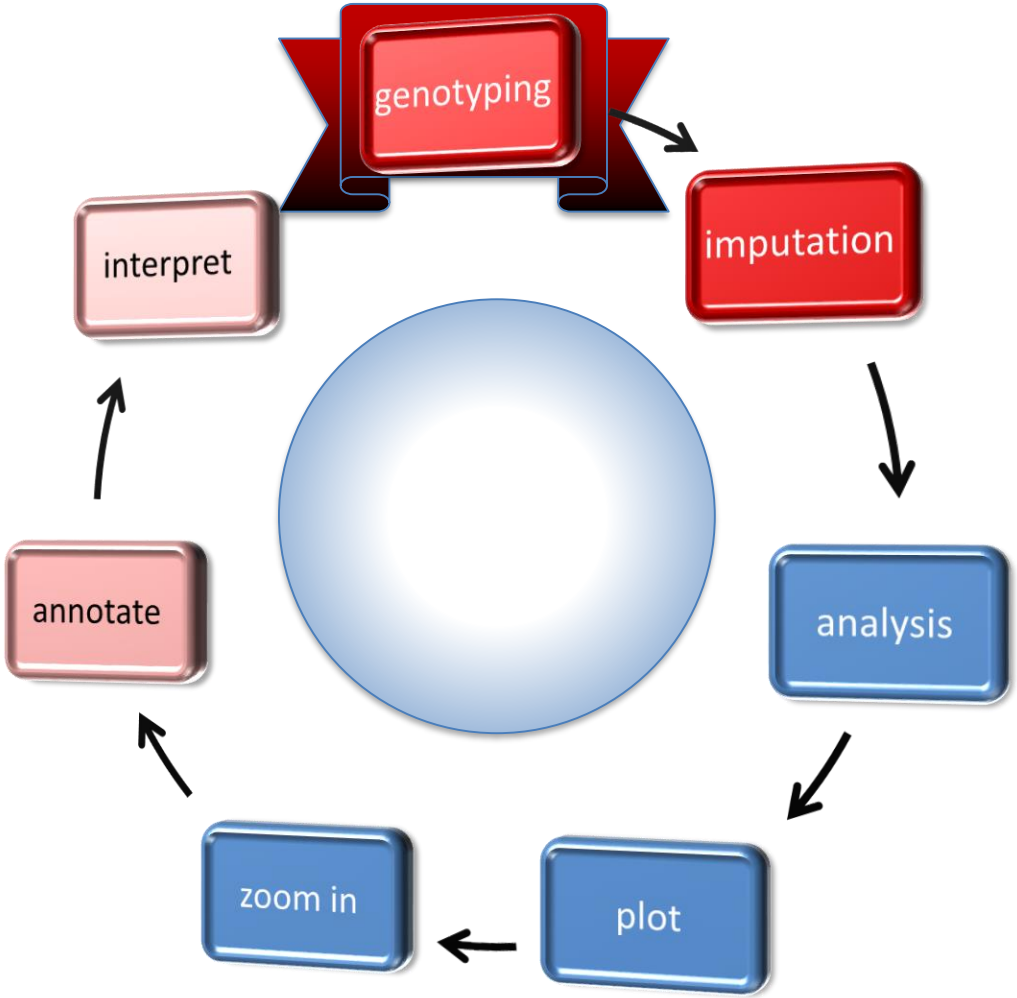


TA Manolio *et al. Nature* **461**, 747-753 (2009)

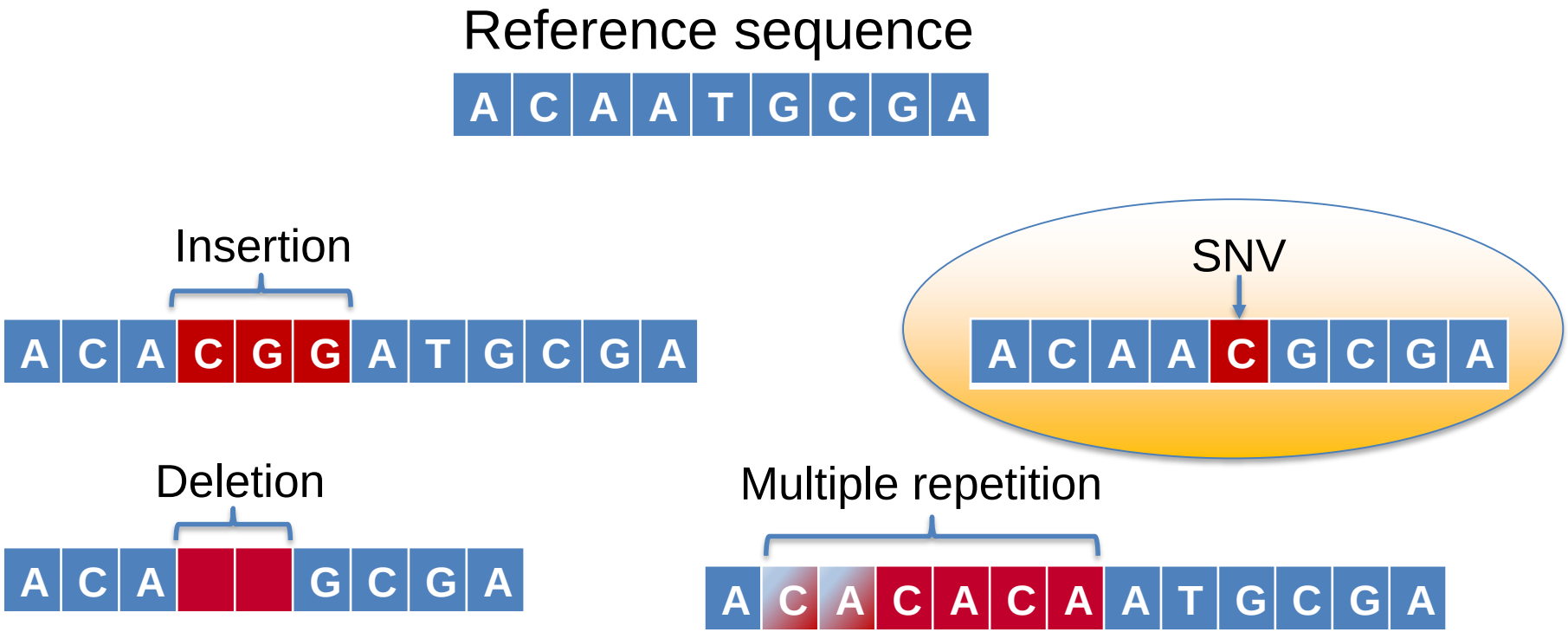
Running a GWAS…



Running a GWAS…



Genetic data
Types of variants



Genetic data
Coding genotypes for analysis

coding for analysis

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

homozygous → 0

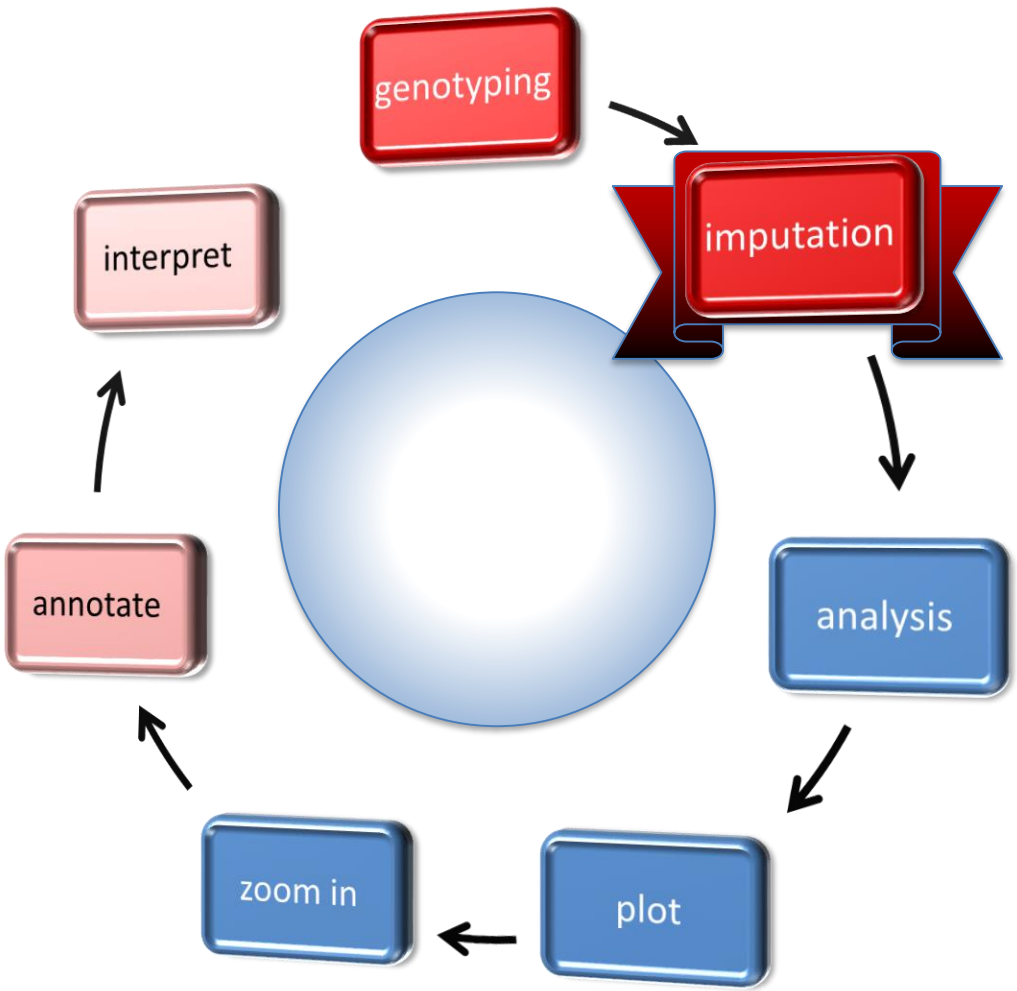
A	C	A	A	T	G	C	G	A
A	C	A	A	C	G	C	G	A

heterozygous → 1

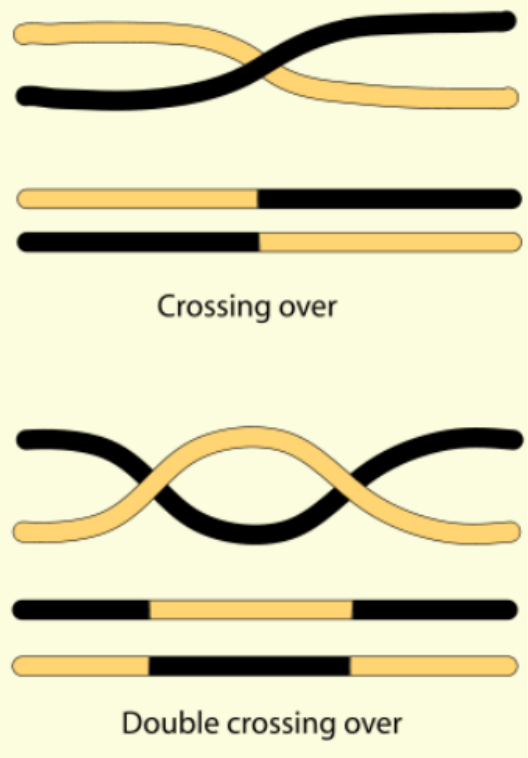
A	C	A	A	C	G	C	G	A
A	C	A	A	C	G	C	G	A

homozygous → 2

Running a GWAS…

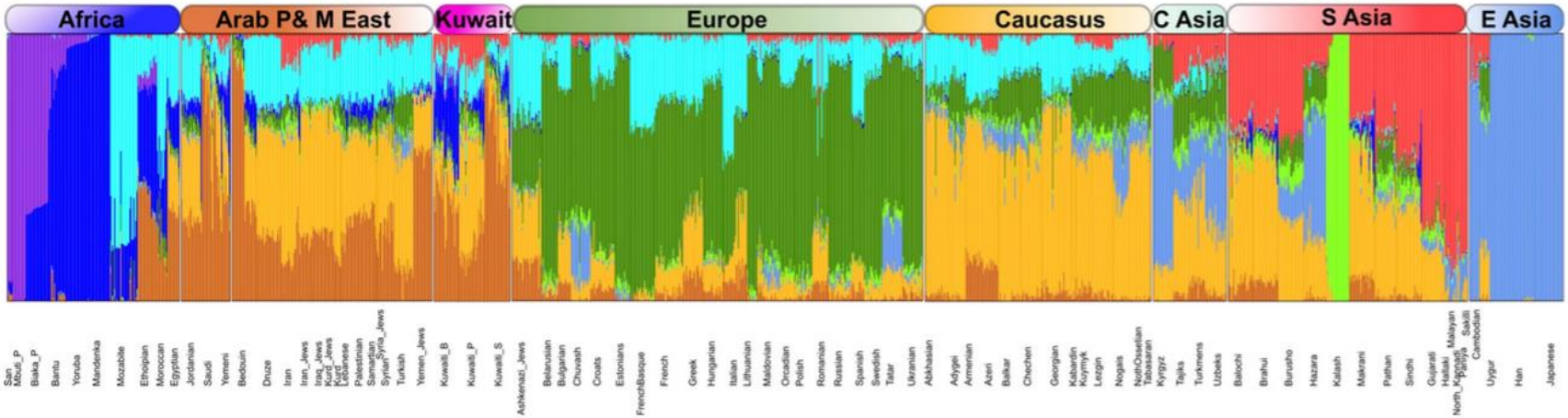


Genetic heterogeneity



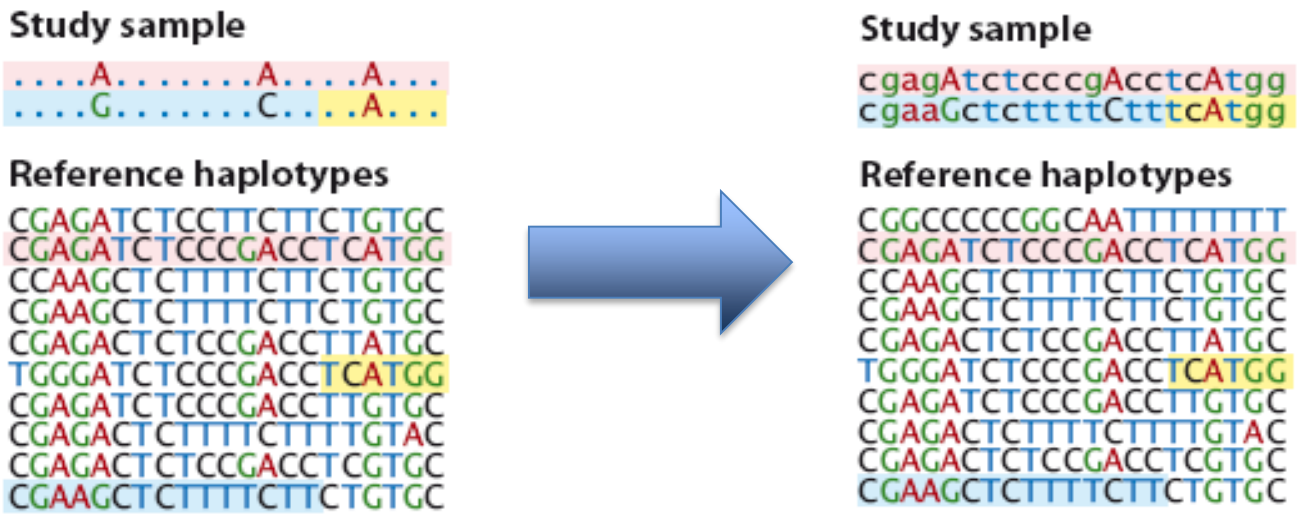
Genetic heterogeneity - example

ADMIXTURE plot for Kuwait study sample



Eaaswarkhanth et al., EJHG, 30, 307-319 (2022)

Imputation


Study sample
....A.....A....A...
....G.....C....A...
Reference haplotypes
CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC
Study sample
cgagAtctcccgAcctcAtgg
cgaaGctcttttCtttcAtgg
Reference haplotypes
CGGCCCCCGGCAATTTTTTTT
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC

Imputationservers

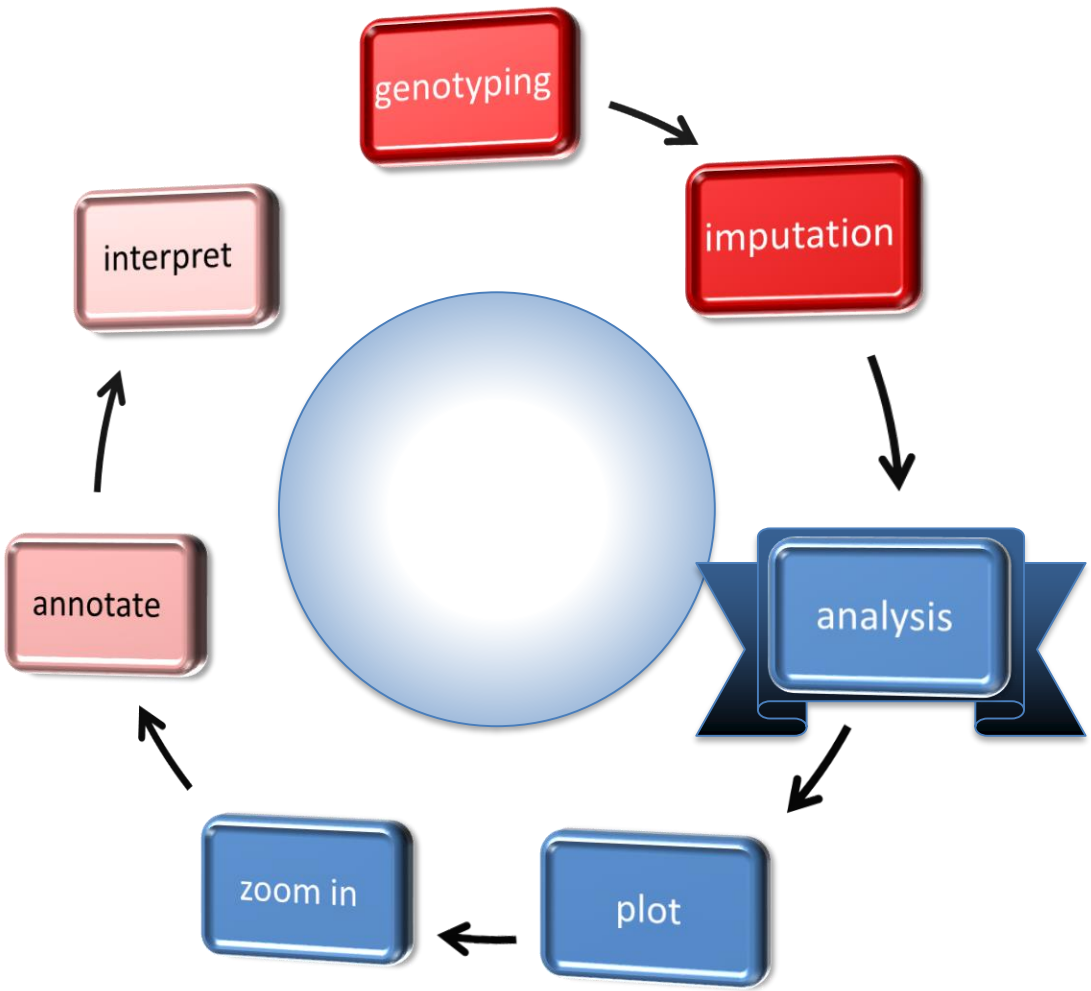
TOPMED Imputation Server
https://imputation.biodatacatalyst.nhlbi.nih.gov

Michigan Imputation Server
https://imputationserver.sph.umich.edu

Sanger Imputation Server
https://imputation.sanger.ac.uk

Various reference panels are available…

Running a GWAS…



Genome-wide association study (GWAS)

- Choose a simple analysis model
- Run your model for each single variant
- Extract results for each variant
- Plot p-values

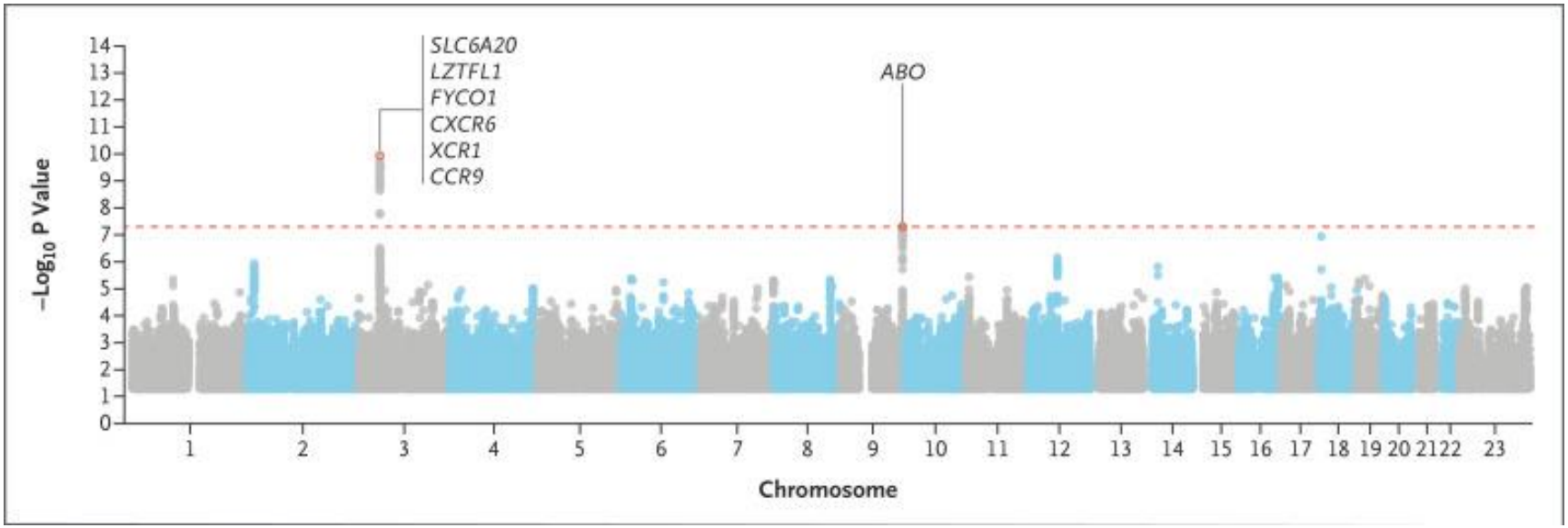
Genome-wide significance level

- Assume $\alpha<0.05$ is considered as significance level
- Probability for at least one false positive result in two independent tests:
  $1-(1-\alpha)^2 = 1-(1-0.05)^2 = 1 - 0.95^2 = 0.0975$
- Error probability for N independent tests:
  $1-(1-\alpha)^N$
- Simplification: $(1 - \alpha)^N \approx 1 - N \cdot \alpha$
  - → $\alpha_{bonf} = \alpha/N$ is the Bonferroni corrected significance level

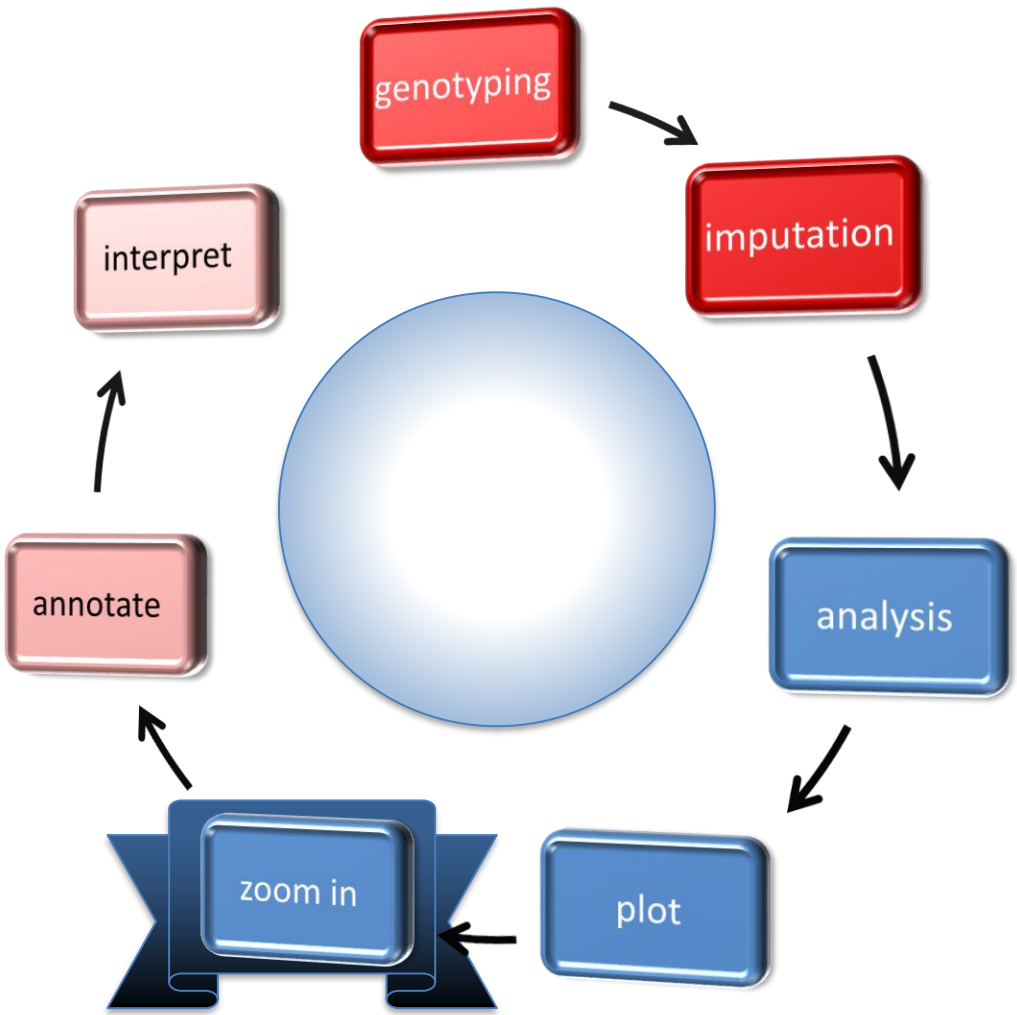
Running a GWAS…

Genomewide Association Study of Severe Covid-19 with Respiratory Failure, NEJM 06/2020 (Ellinghaus et al)

Manhattan plot



Running a GWAS…



http://locuszoom.org

LocusZoom

LocusZoom is a suite of tools to provide fast visualization of GWAS results for research and publication.

Original LocusZoom (R/Python) is ideal for batch generation of static plots.

LocusZoom.js (JavaScript) aims to make LocusZoom plots interactive and scriptable.

Interactive Plots with LocusZoom.js

MY.LOCUSZOOM.ORG
UPLOAD, ANALYZE, AND SHARE

LOCALZOOM
EXPLORE WITHOUT UPLOADING

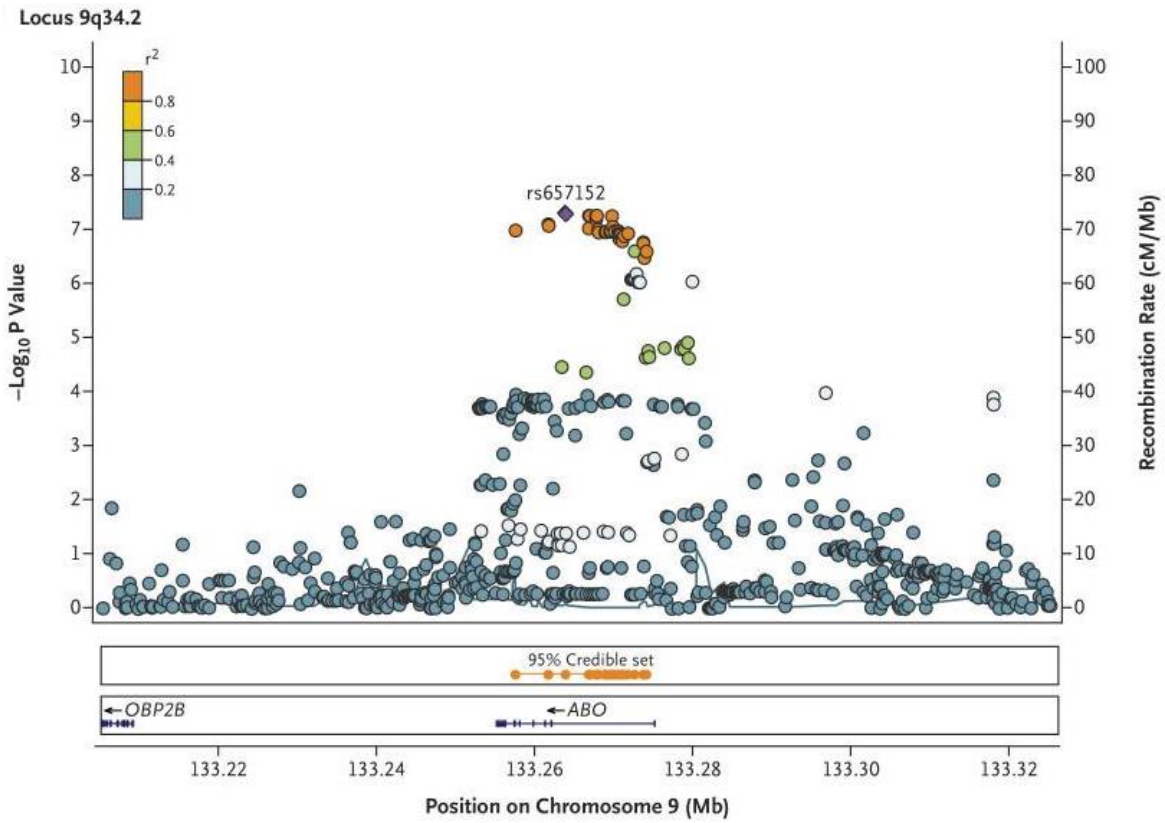
Legacy Services (not actively maintained)

SINGLE PLOT
YOUR DATA - ORIGINAL LOCUSZOOM

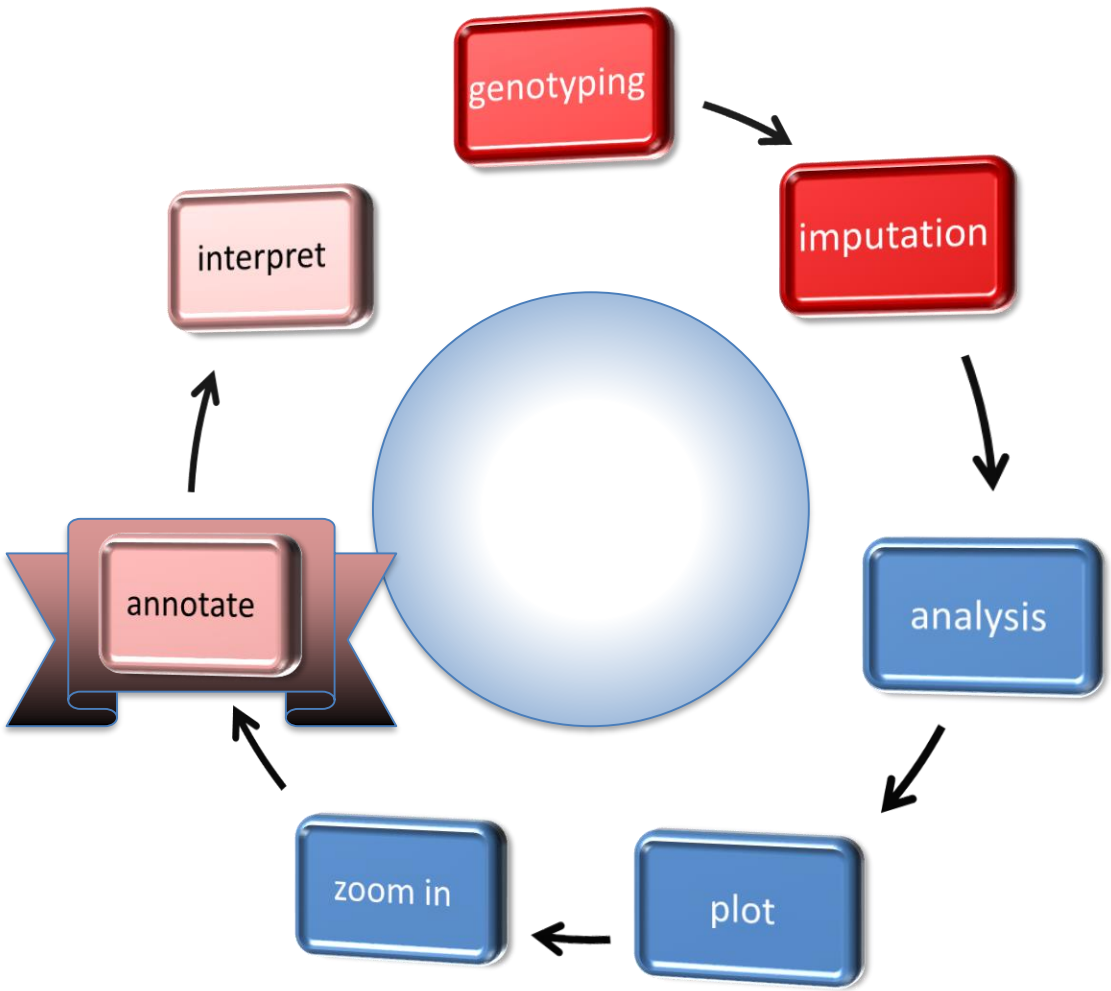
BATCH PLOT WITH HITSPEC
YOUR DATA - ORIGINAL LOCUSZOOM

INTERACTIVE PLOT
PUBLISHED GWAS - LOCUSZOOM.JS

Genomewide Association Study of Severe Covid-19 with Respiratory Failure, NEJM 06/2020 (Ellinghaus et al)



Running a GWAS…



Genetic data
Effects of variants on gene function

Original DNA strand

ACG	GGT	ATA	CGA

Insertion

ACA	GGG	TAT	ACG	A..

Frame shift

Deletion

ACG	GTA	TAC	GA.

Frame shift

Genetic data
Effects of variants on gene function

Original DNA strand

DNA	ACG	GGT	ATA	CGA
RNA	UGC	CCA	UAU	GCU
Amino acid	Cys	Pro	Tyr	Ala

Substitution

ACA	GGT	ATA	CGA
UGU	CCA	UAU	GCU
Cys	Pro	Tyr	Ala

Synonymous change

Genetic data
Effects of variants on gene function

Original DNA strand

DNA	ACG	GGT	ATA	CGA
RNA	UGC	CCA	UAU	GCU
Amino acid	Cys	Pro	Tyr	Ala

Substitution

ACC	GGT	ATA	CGA
UGG	CCA	UAU	GCU
Trp	Pro	Tyr	Ala

Nonsynonymous change

Genetic data
Effects of variants on gene function

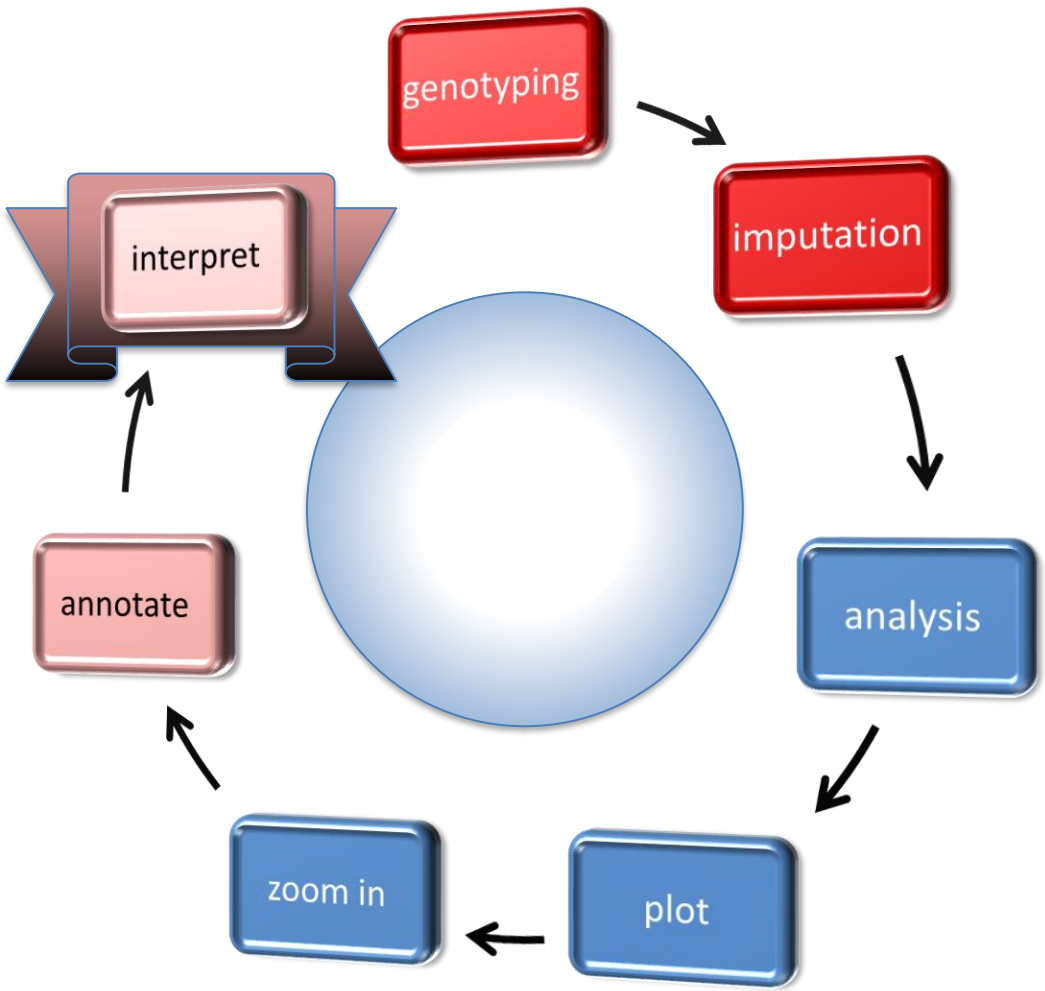
Original DNA strand

DNA	ACG	GGT	ATA	CGA
RNA	UGC	CCA	UAU	GCU
Amino acid	Cys	Pro	Tyr	Ala

Substitution

ACT	GGT	ATA	CGA
UGA	CCA	UAU	GCU
stop	…	…	…

Running a GWAS…



Variant annotation
SNIPA

SNiPA single nucleotide polymorphisms annotator

HelmholtzZentrum münchen
German Research Center for Environmental Health

Weill Cornell Medical College in Qatar

Home

Browse
- Variant Browser
- Association Maps

Annotation
- Variant Annotation
- Block Annotation

Plots
- Regional Association Plot
- Linkage Disequilibrium Plot

Linkage Disequilibrium
- Proxy Search
- Pairwise LD

Help
- Release Notes
- Documentation
- Supplemental Text
- About SNiPA

SNiPA - a tool for annotating and browsing genetic variants

Welcome to SNiPA!

SNiPA offers both functional annotations and linkage disequilibrium information for bi-allelic genomic variants (SNPs and SNVs). SNiPA combines LD data based on the 1000 Genomes Project with various annotation layers, such as gene annotations, phenotypic trait associations, and expression-/metabolic quantitative trait loci. See the documentation for all data sources integrated into SNiPA. For information on updates and new releases, see the Release Notes.

- **Variant Browser** — Explore the functional annotations of variants.
- **Association Maps** — Map trait associations from NHGRI's GWAS Catalog to karyograms.
- **Variant Annotation** — Access detailed variant annotations.
- **Block Annotation** — Summarize variant annotations within LD blocks or chromosomal regions.
- **Regional Association Plot** — Visualize and annotate the results of genetic association studies.
- **Linkage Disequilibrium Plot** — Combine LD data and annotations in an interactive plot.
- **Proxy Search** — Find variants that are linked to other variants by LD.
- **Pairwise Linkage Disequilibrium** — Determine the pairwise LD for two or more variants.
- **Documentation** — See how to use SNiPA and its interactive features.

Current release

***SNiPA* v3.4 (released November 13th, 2020)**

Genome assembly: GRCh37.p13 **Ensembl version:** 87 **1000 genomes:** phase 3 version 5

SNiPA now incorporates **two recent mQTL studies** by Schlosser et al. (Nature Genetics, Feb 2020) and Lotta et al. (preprint, Jul 2020), as well as **pQTLs in the context of SARS-CoV-2** from Pietzner et al. (preprint, Jul 2020). Additionally, over 4 million unique pooled and sex-specific associations from the Neale lab UK Biobank GWASs were integrated. Further updates include the most recent versions of **GTEx** for eQTL data as well as the **GWAS Catalog**, **HGMD public** and **ClinVar**, totalling to **more than 940,000 genetic trait associations**.

For further information, see the release notes or the documentation.

Last update 11/2020: https://snipa.helmholtz-muenchen.de/snipa3/

Variant annotation
SNIPA

SNiPA single nucleotide polymorphisms annotator

HelmholtzZentrum münchen
German Research Center for Environmental Health

Weill Cornell Medical College in Qatar

Home

Browse
- Variant Browser
- Association Maps

Annotation
- Variant Annotation
- Block Annotation

Plots
- Regional Association Plot
- Linkage Disequilibrium Plot

Linkage Disequilibrium
- Proxy Search
- Pairwise LD

Help
- Release Notes
- Documentation
- Supplemental Text
- About SNiPA

Variant Annotation

This module allows you to get detailed annotations for one or more variants. If the results are not what you have expected, please check the "Report" tab for details. close

Variant annotations | Report

rs505922 — show static URL | add to clipboard | save as PDF | delete

SNP properties – Genome Assembly: grch37, Variant set: 1kgpp3v5, Population: EUR

rs505922 (alias rs57823732)

position / outlink		allele info	
physical position	chr9: 136,149,229	alleles	T/C
genetic position [cM]	166.21	frequencies	0.632/0.368
outlink	e!	VEP effect allele	T

Basic features

Conservation/deleteriousness		Linked genes	
phyloP	-0.197	gene(s) hit or close-by	ABO e!
phastCons	0	eQTL gene(s)	GBGT1 e!
GERP++	-0.686	pQTL gene(s)	ABO e!, BCAM e!, CD14 e!, CD200 e!, CD209 e!, CD36 e!, CHST15 e!, FLT4 e!, HO-1 e!, ICAM2 e!, INSR e!, KDR e!, MBL2 e!, MET e!, PLOD2 e!, SELE e!, SELP e!, SVEP1 e!, TIE1 e!, VWF e!, gp130, soluble e!
CADD score	4.514	potentially regulated gene(s)	–
SnpEff effect impact	modifier	disease gene(s)	FLT4 e!, PLOD2 e!, BCAM e!, ABO e!, VWF e!, HMOX1 e!, MET e!, CD36 e!, KDR e!, MBL2 e!, CD209 e!, IL6ST e!, INSR e!

Trait annotations

Show 10 entries — Search:

mQTL association

metabolite	biofluid	p-value	source	source	entry/link
Glycine	blood	6.14×10^{-14}	Lotta et al.	2020.02.03.932541	

Last update 11/2020: https://snipa.helmholtz-muenchen.de/snipa3/

From DNA to function
Public resources: GWAS Catalog

GWAS Catalog | Diagram | Submit | Download | Documentation | About | Blog | EMBL-EBI | NIH National Human Genome Research Institute

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

Home / Downloads / Summary statistics

feedback

Studies with available summary statistics

Users can access all summary statistics from the Catalog FTP site, which is updated nightly following submission. They can also be accessed in the tables below (separate tables for the published and unpublished summary statistics). Metadata associated with summary statistics can be downloaded from Downloads.

If you are an author and have summary statistics you would like to submit to the GWAS Catalog please visit our submission page. These data are made available either through CC0 or EMBL-EBI's standard terms of use, more details can be found here . For licensing information of individual studies, please reveal "Usage License" column.

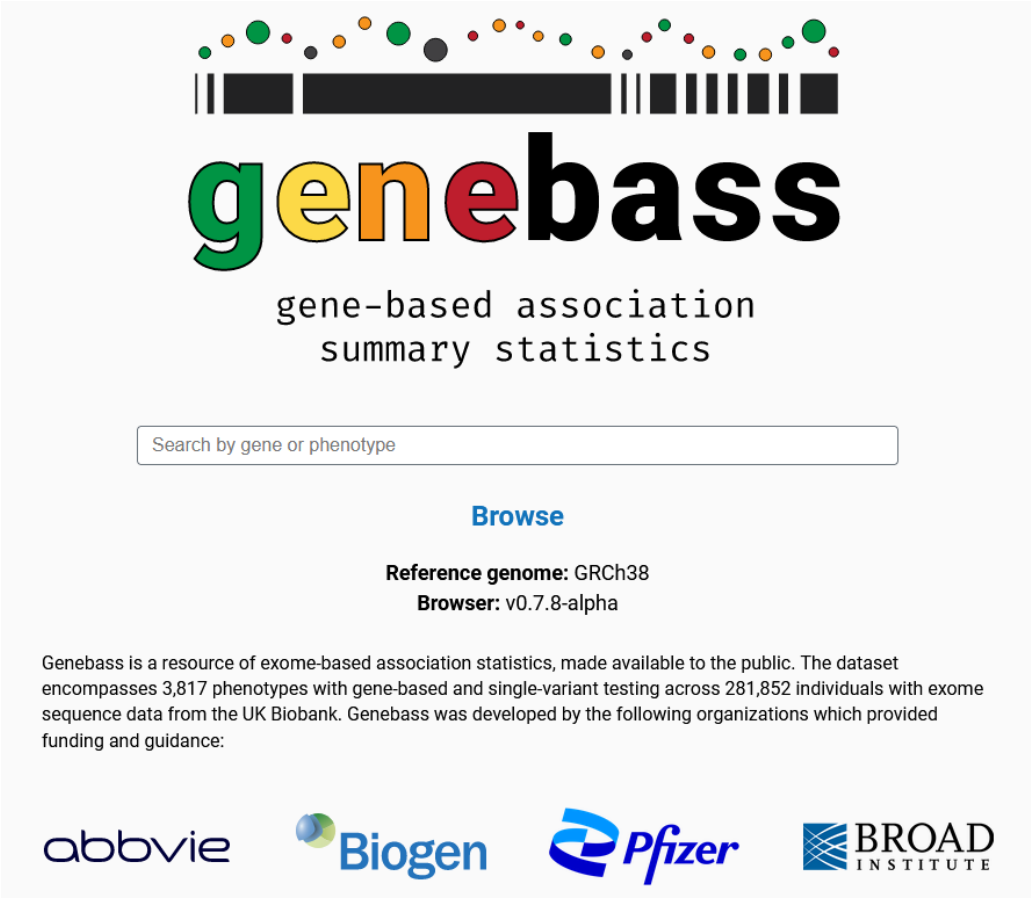
For information on summary statistics and the summary statistics REST API please read the documentation.

List of published studies with summary statistics

Data from ~30,300 published and ~5,800 pre-/unpublished studies

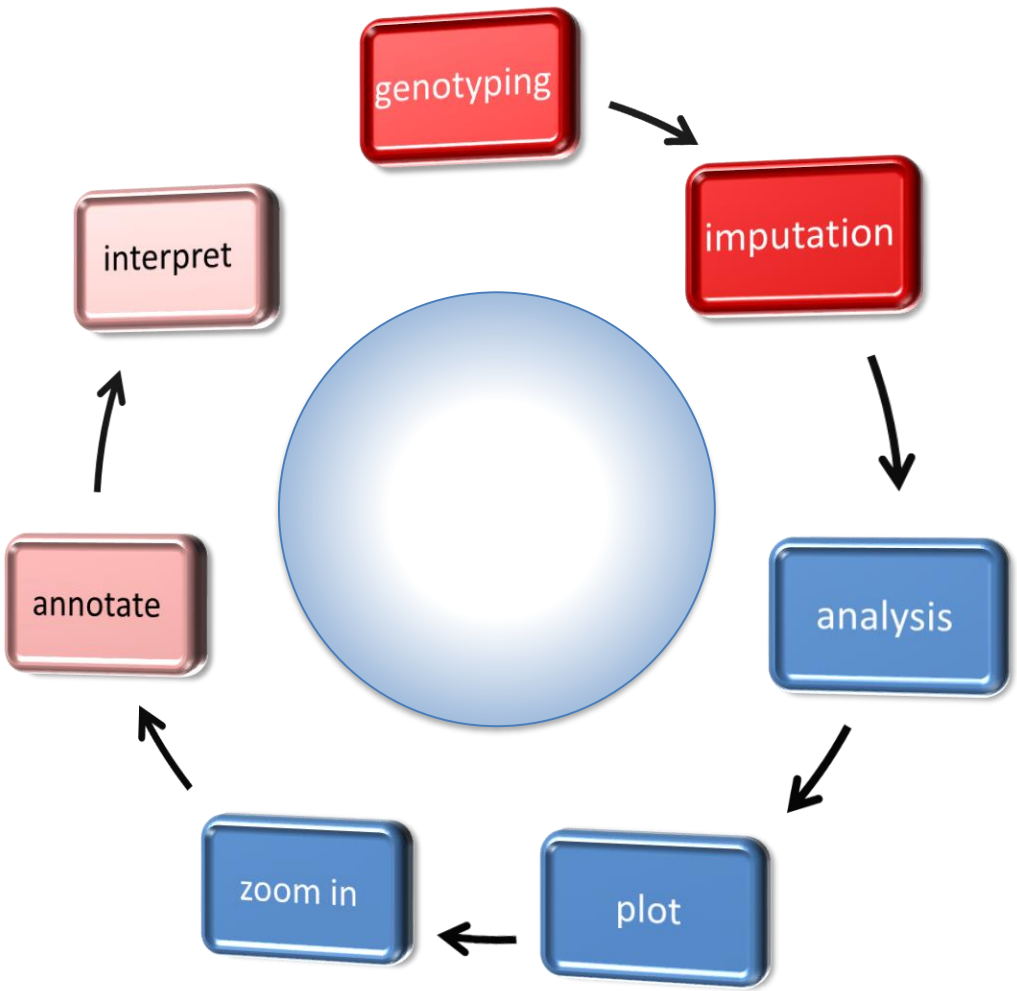
http://www.ebi.ac.uk/gwas/downloads/summary-statistics

From DNA to function
Public resources: GENEBASS (UK Biobank)

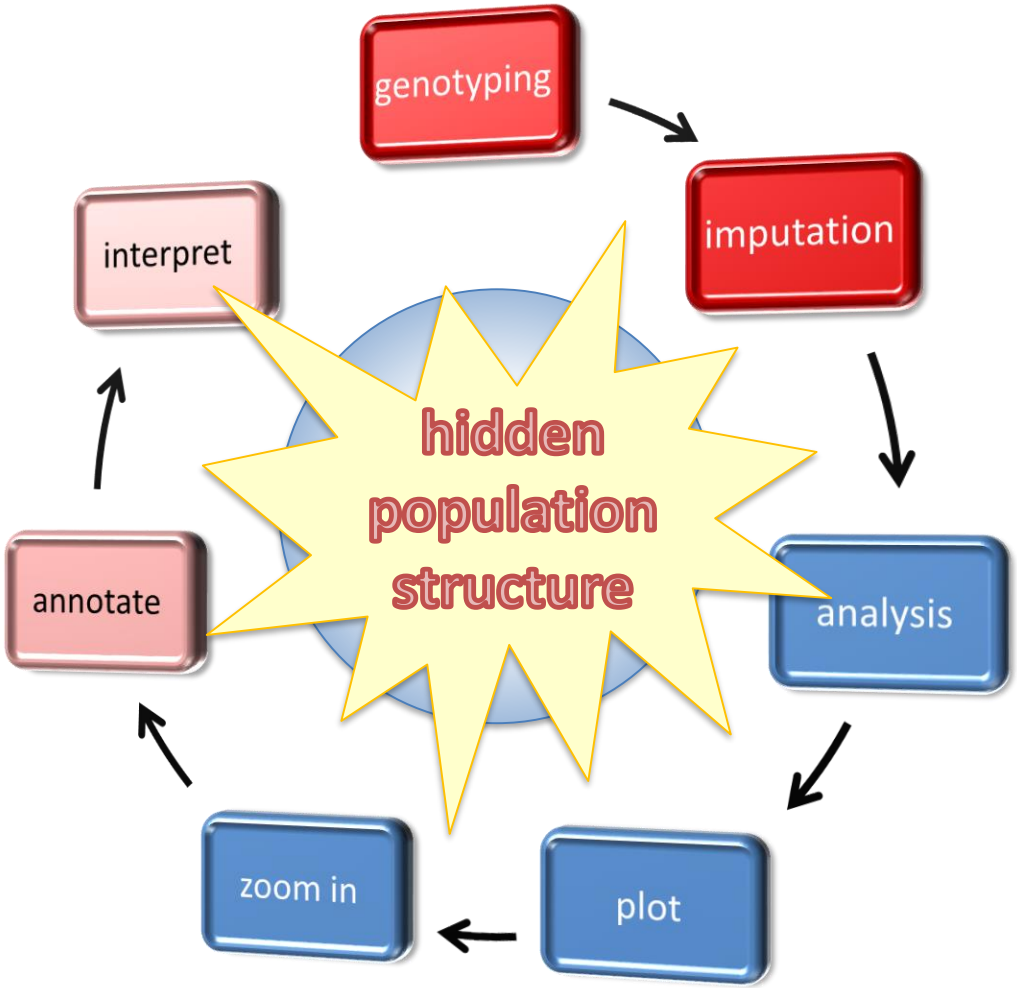


Online since June 2021: https://genebass.org/

Running a GWAS… extensions….



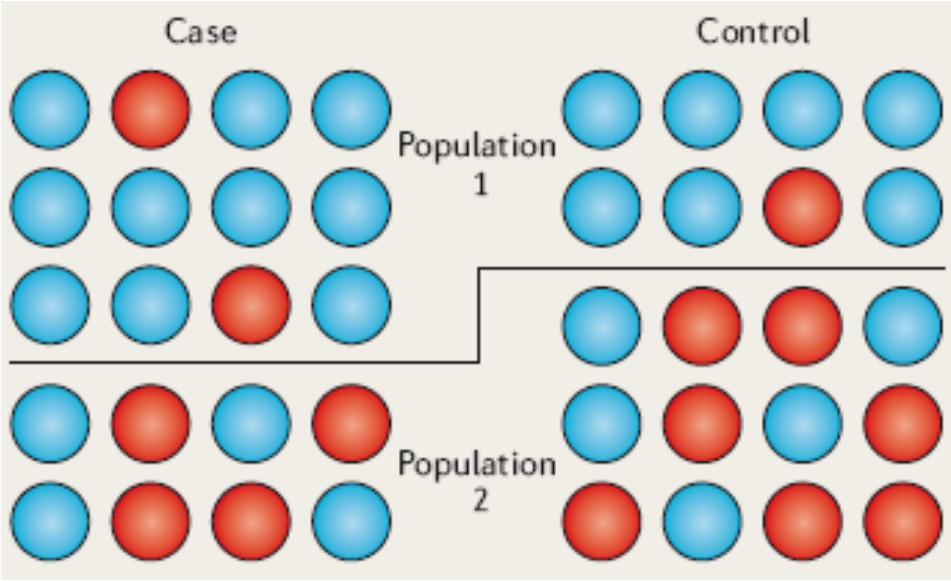
Running a GWAS… extensions…



Genetic variation
Population stratification

Example: chop-stick study
Molecular Psychiatry (2000)

Genetic variation
Population stratification



Overall blue: 14/20 vs. 12/20
Population 1: 10/12 vs. 7/8
Population 2: 4/8 vs. 8/15

Balding, Nature Reviews Genetics 2006

Genetic variation
Population stratification

The problem of population stratification plays a role when:

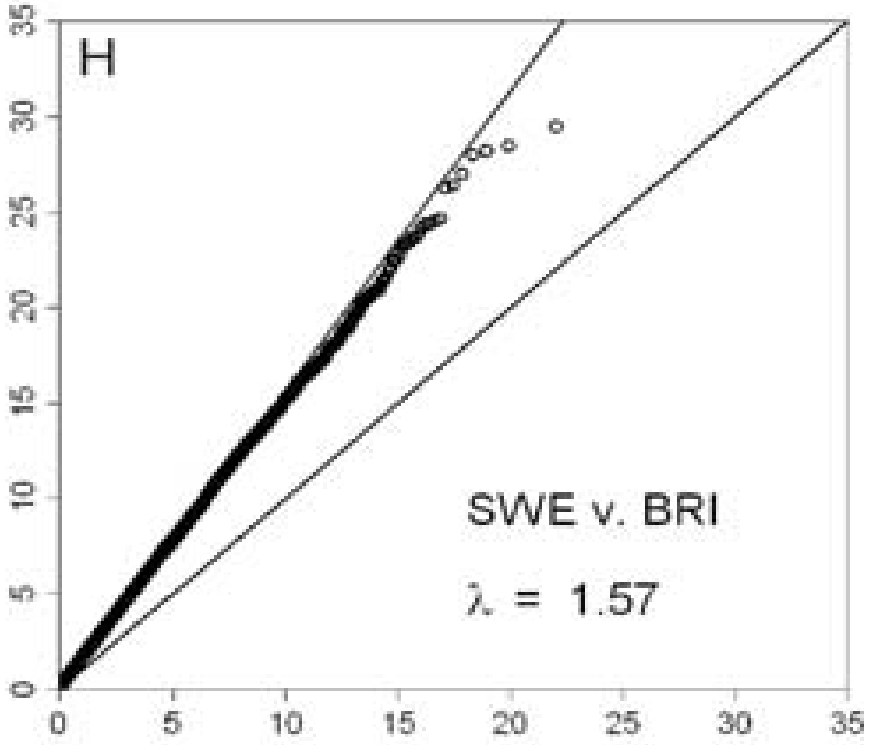
- the study sample consists of two subpopulations
- which are disproportionally represented in cases and controls
- allele frequencies vary between the two subpopulations

Genetic variation
Reasons for population stratification

- Populations with different ancestry
- Hidden family structures
- Higher penetrance of the causal allele in the subgroup because of a different environment
  (e.g. diet)
- Ascertainment bias
  (e.g. the subgroup is more closely monitored by health services than the general population, so that cases from the subgroup are more likely to be included in the study)
- Different genotyping platforms/chips/runs for cases and controls that result in different genotype quality

Check for population stratification

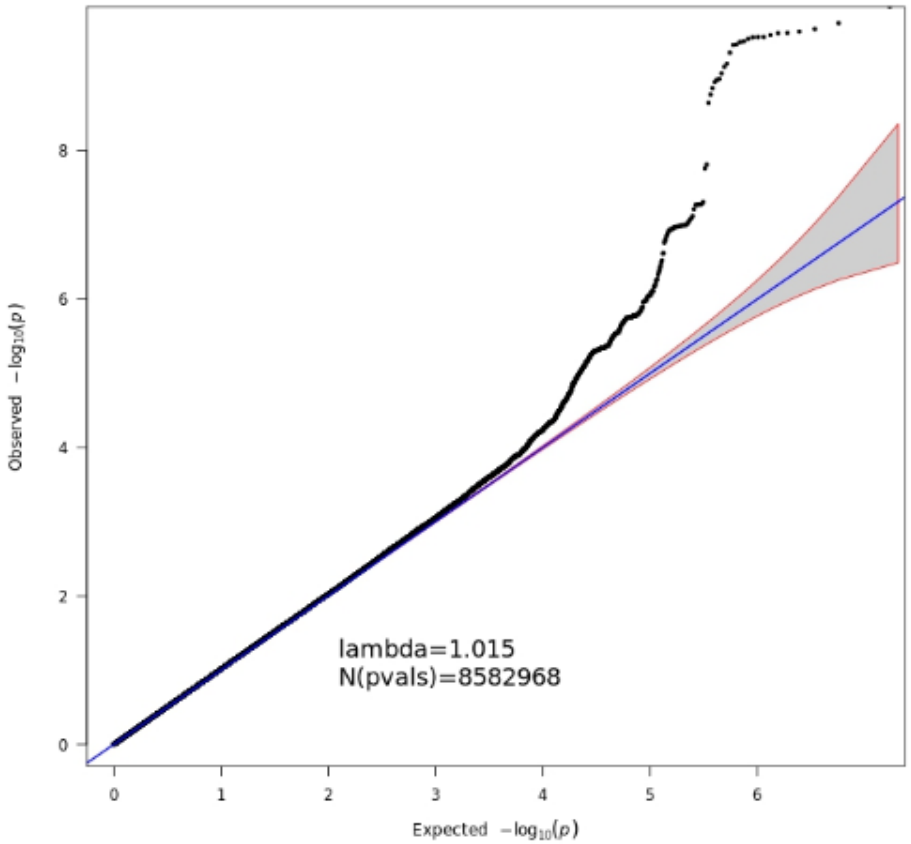
QQ plot



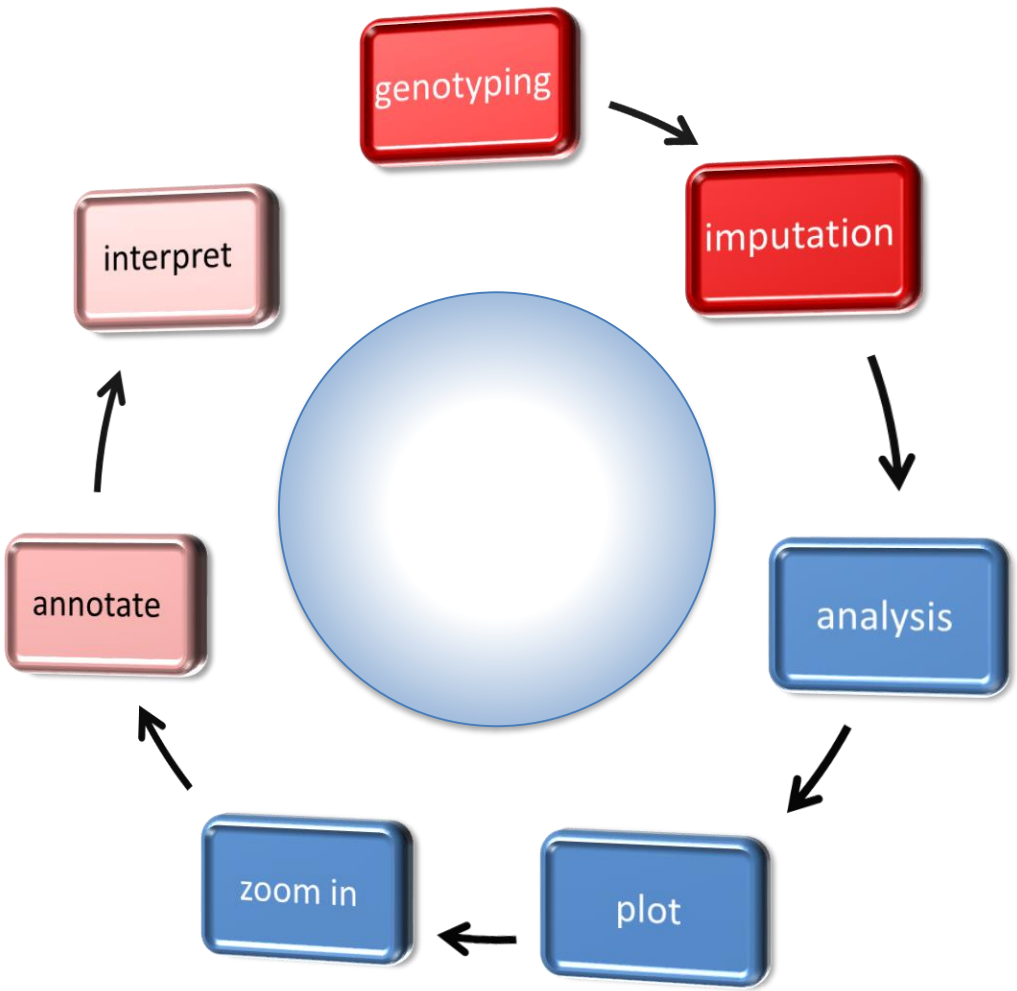
Salmela, PLoS ONE 2008

Genomewide Association Study of Severe Covid-19 with Respiratory Failure, NEJM 06/2020 (Ellinghaus et al)

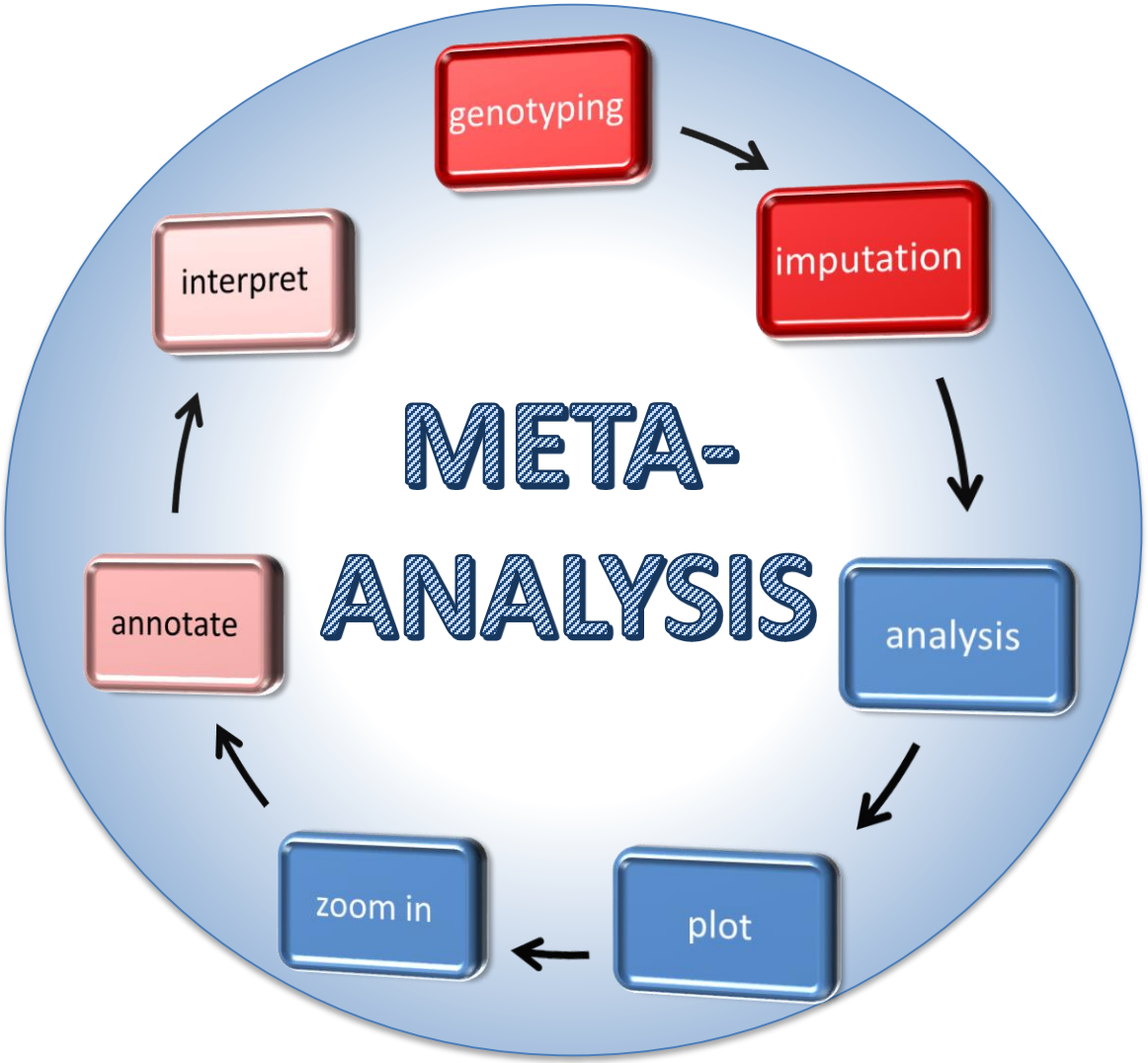
QQ plot



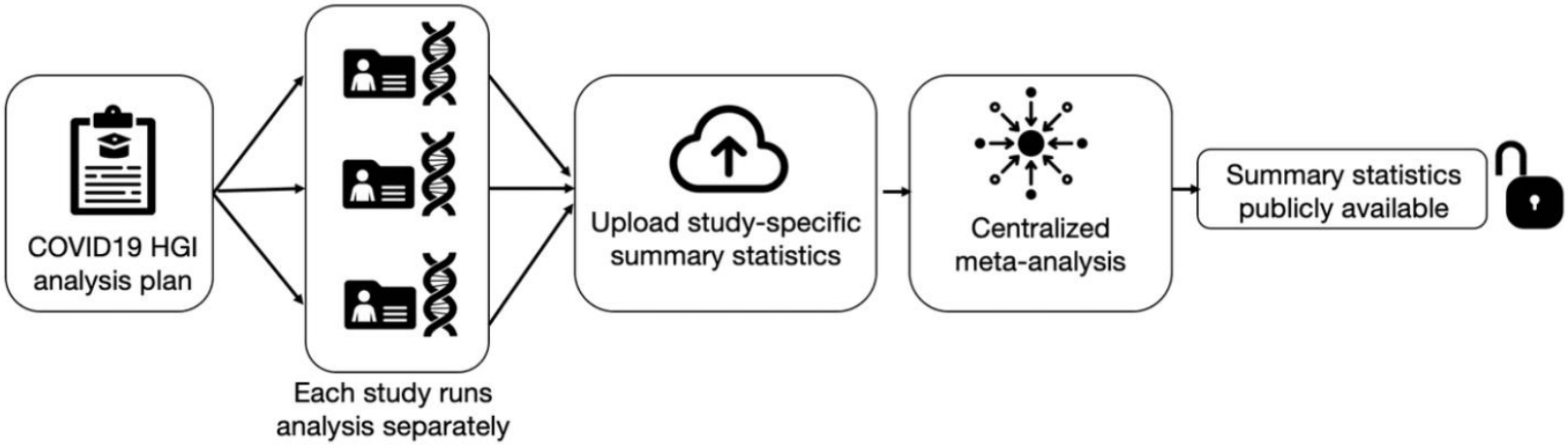
Running a GWAS… extensions….



Running a GWAS… extensions….



Meta-analysis in GWAS



From covid19hg website: https://www.covid19hg.org/data-sharing/

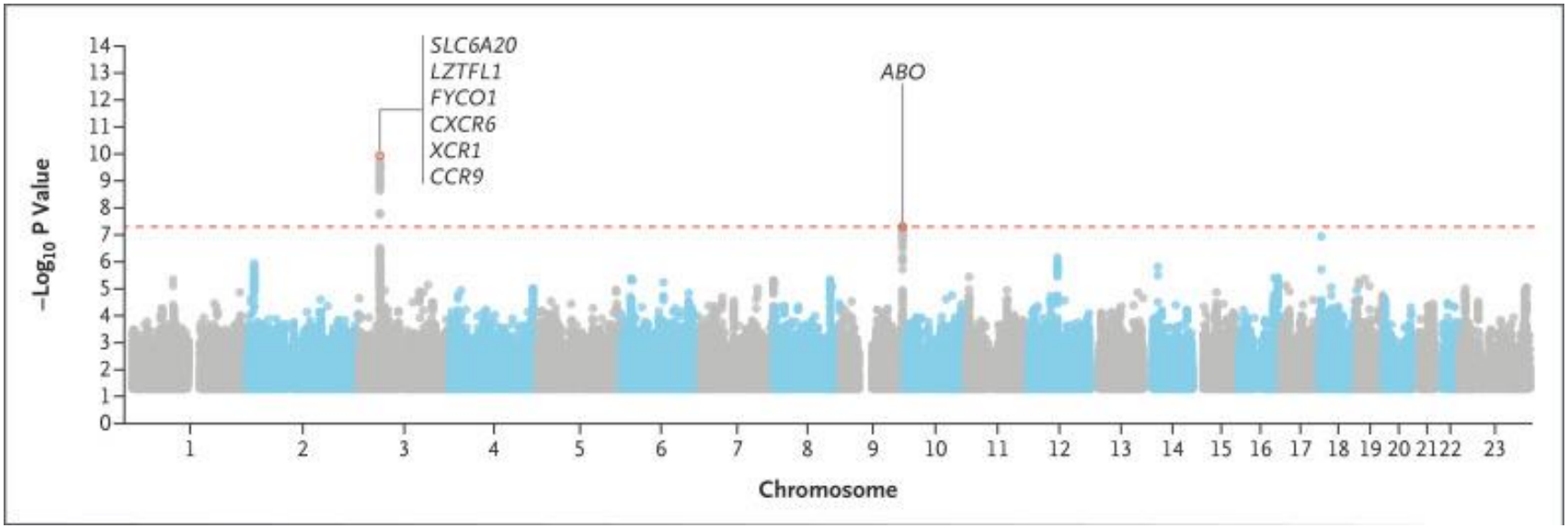
Meta-analysis in GWAS

Common methods for pooling estimates

- z-scores weighted by study size
  - only p-values and sample size are required
  - no pooled estimates are available
- inverse variance weighting
  - effect estimates and standard errors for each study needed
  - pooled estimates are available

Genomewide Association Study of Severe Covid-19 with Respiratory Failure, NEJM 06/2020 (Ellinghaus et al)

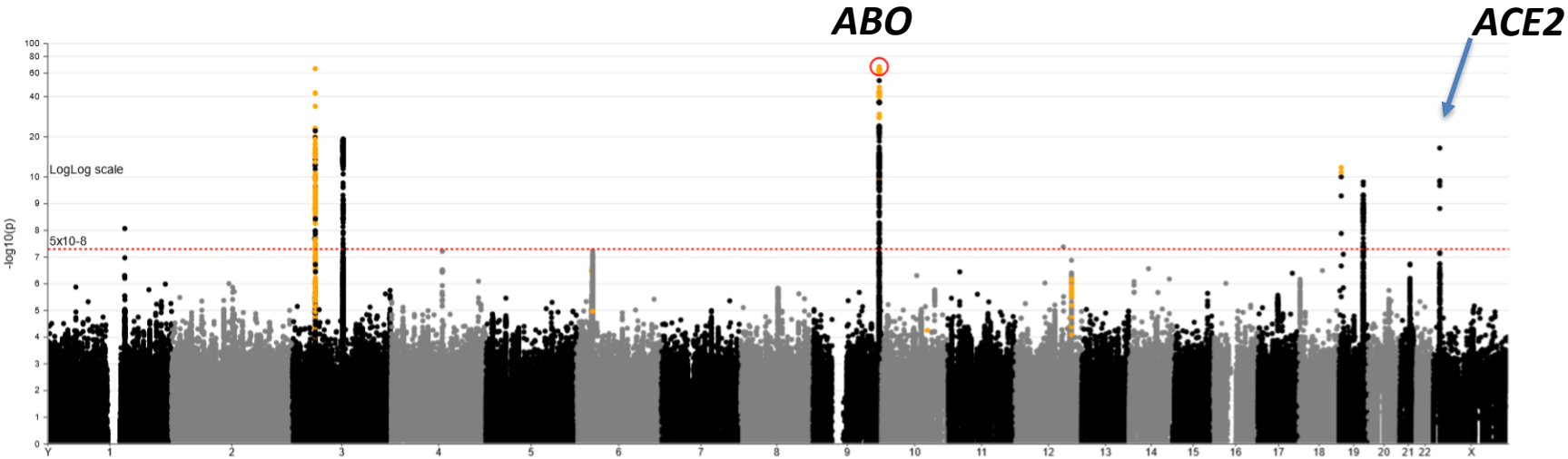
Manhattan plot



total (Italy/Spain)	2090/1725
cases	835/775
controls	1255/950

COVID-19 Host Genetics Initiative, Release 6 (06/2021)

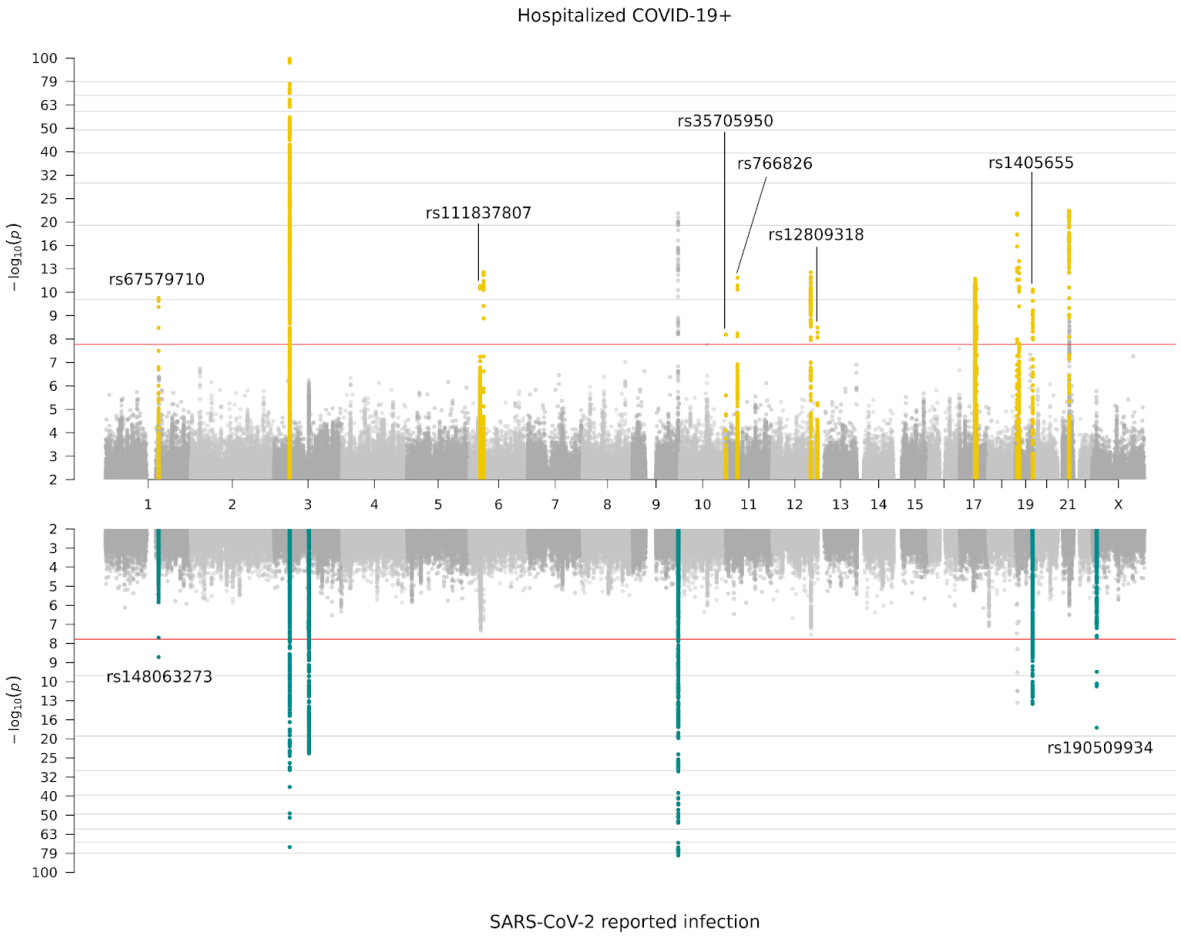
susceptibility



Total	2,586,691
Cases	112,612
Controls	2 474,079

63 studies
25% non-EUR

Mapping the human genetic architecture of COVID-19, Nature 07/2021 (COVID-19 Host Genetics Initiative)



Measures of genetic heterogeneity in meta-analyses

I^2 = percentage of the total variation across studies due to heterogeneity beyond chance

I^2 is based in Cochran's Q (weighted sum of the squared derivation between study and meta-analysis effect estimates)

I^2 >50% is generally considered to indicate heterogeneity between studies

Test statistics are available

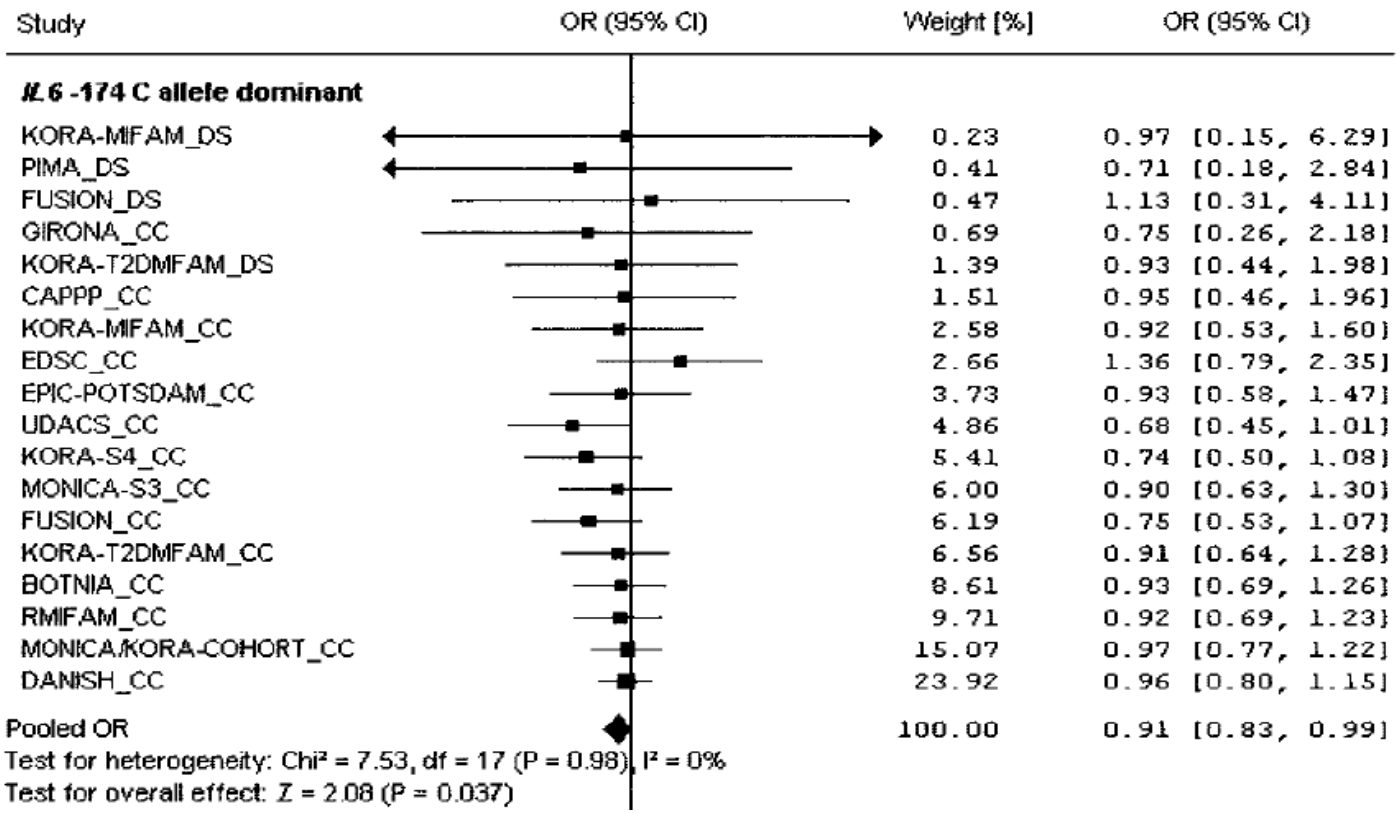
Meta-analysis in GWAS

Fixed-effect metaanalysis assume that in all studies, generally, the same effect is present. Effect estimation varies around this true effect.

Random effect metaanalysis allows effects to vary between studies although a general effect can be described. This should generally be applied, in case heterogeneity between study-specific effect is strong.

Measuring genetic heterogeneity in meta-analyses

IL6 gene promoter polymorphism and type 2 diabetes



Huth et al., Diabetes 2006

Effects of genetic heterogeneity in meta-analyses

GENE	Polymorphism	Q (df)[a] [p]	I^2 (95% CI)	Random effects OR (95% CI)	Fixed effects OR (95% CI)	Random effects p-value	Fixed effects p-value
—	rs9300039	8.38 (3) [0.039]	64% (0–86)	1.29 (1.11–1.50)	1.26 (1.15–1.37)	0.001	2.8×10^{-8}
FTO	rs8050136	12.98 (4) [0.011]	69% (0–86)	1.15 (1.06–1.25)	1.17 (1.12–1.23)	0.001	2.5×10^{-12}
PPARG	rs1801282	6.93 (4) [0.14)	42% (0–76)	1.14 (1.06–1.23)	1.13 (1.08–1.20)	0.0007	3.4×10^{-6}
CDKAL1	rs10946398	8.76 (5) [0.12]	43% (0–76)	1.13 (1.07–1.18)	1.12 (1.08–1.15)	1.2×10^{-6}	1.9×10^{-10}
SLC30A8	rs13266634	3.17 (5) [0.67]	0 (0–61)	1.13 (1.08–1.17)	1.13 (1.08–1.17)	4.1×10^{-9}	4.1×10^{-9}
CDKN2B	rs564398	3.62 (4) [0.46]	0% (0–64)	1.11 (1.06–1.15)	1.11 (1.06–1.15)	5.8×10^{-7}	5.8×10^{-7}
HHEX	rs5015480–rs1111875	6.20 (5) [0.29]	19% (0–68)	1.13 (1.08–1.17)	1.12 (1.08–1.17)	2.2×10^{-8}	3.2×10^{-10}
KCNJ11	rs5215	3.50 (4) [0.48]	0% (0–64)	1.14 (1.09–1.18)	1.14 (1.09–1.18)	9×10^{-11}	9×10^{-11}
IGF2BP2	rs4402960	7.08 (5) [0.21]	29% (0–71)	1.15 (1.10–1.20)	1.15 (1.11–1.19)	2.9×10^{-10}	1.1×10^{-15}
CDKN2B	rs10811661	4.15 (5) [0.53]	0% (0–61)	1.20 (1.15–1.25)	1.20 (1.15–1.25)	2.7×10^{-15}	2.7×10^{-15}
TCF7L2	rs7901695	1.31 (4) [0.86]	0% (0–64)	1.37 (1.32–1.43)	1.37 (1.32–1.43)	1.0×10^{-48}	1.0×10^{-48}

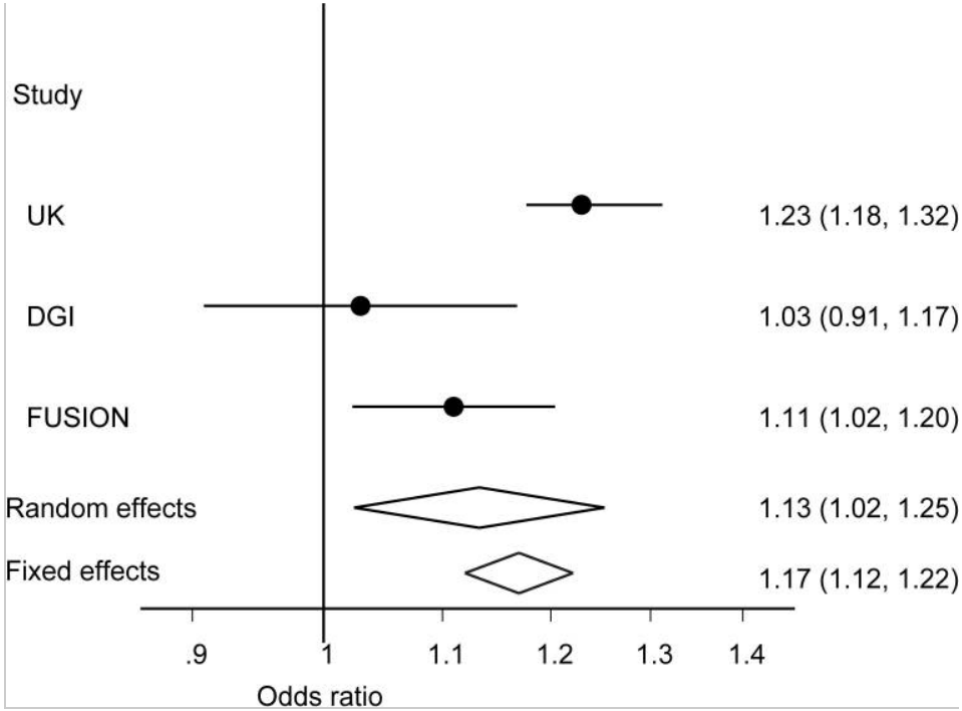
CI: confidence interval; OR: odds ratio

[a]df = degrees of freedom; not all markers were tested by all 3 investigations in their replication efforts, thus even with splitting the discovery and replication phases, there are fewer than 6 datasets (df = 5) for some variants.

Ioannidis, PLoS ONE 2007

Effects of genetic heterogeneity in meta-analyses

Forrest plot for *FTO* variant

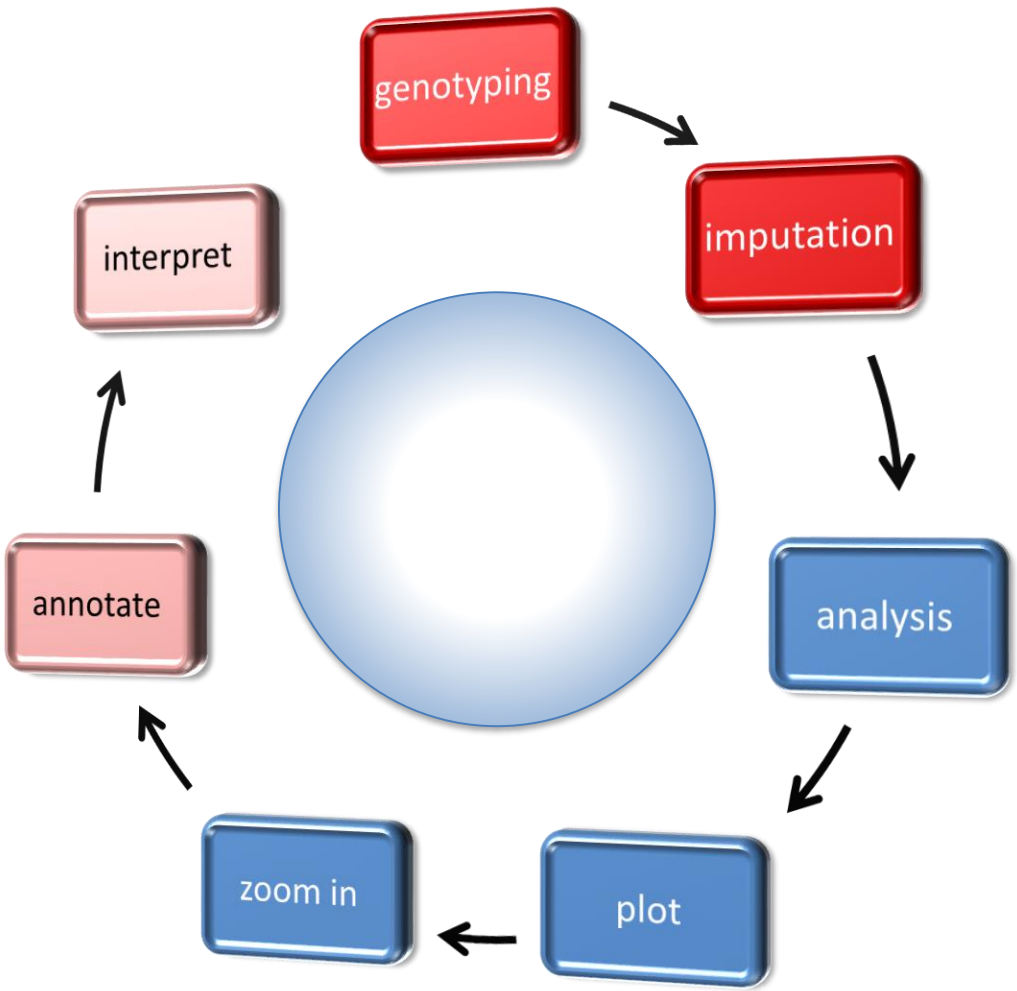


Ioannidis, PLoS ONE 2007

Causes for heterogeneity in meta-analyses

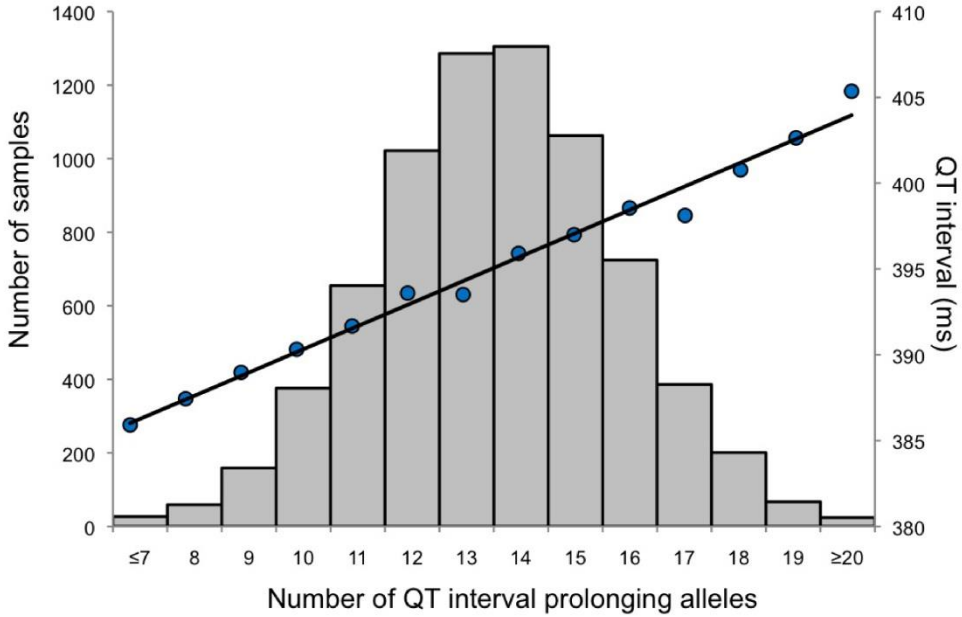
- The identified variant is in **LD with the true causative variant**, while LD can vary across populations
- The true association can be with a **correlated phenotype**, while the correlation between phenotypes can vary across populations
- Association results may be **biased**, e.g. due to population stratification, genotyping error, phenotype misclassification...
- The true effect results from gene-gene or gene-environment **interactions**
- **"Winner's curse"**: initiation of the meta-analysis through single study with overwhelming results

Running a GWAS… extensions….



Running a GWAS… extensions….

Polygenic risk scores for risk stratification
Example: QT interval

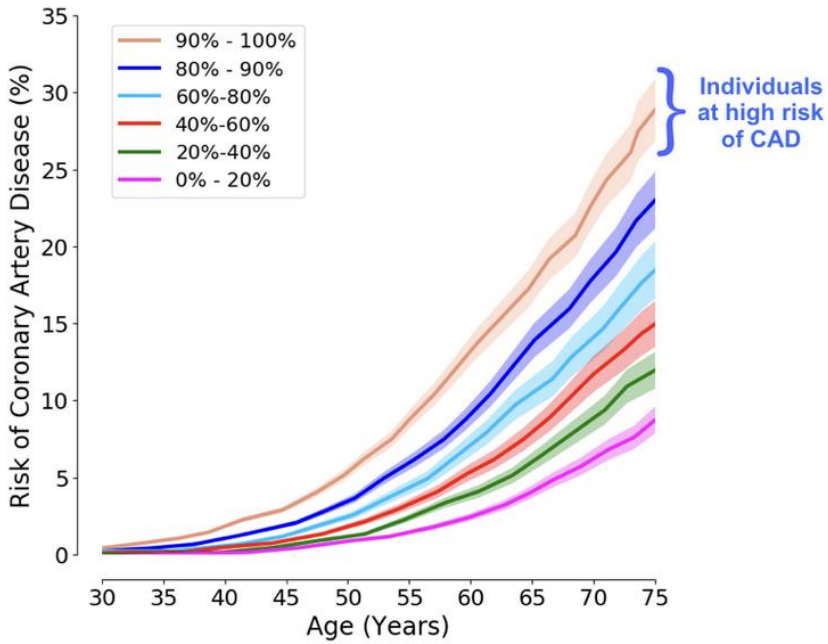


Allelic score of 12 QT-prolonging variants in N=10,563 study participants

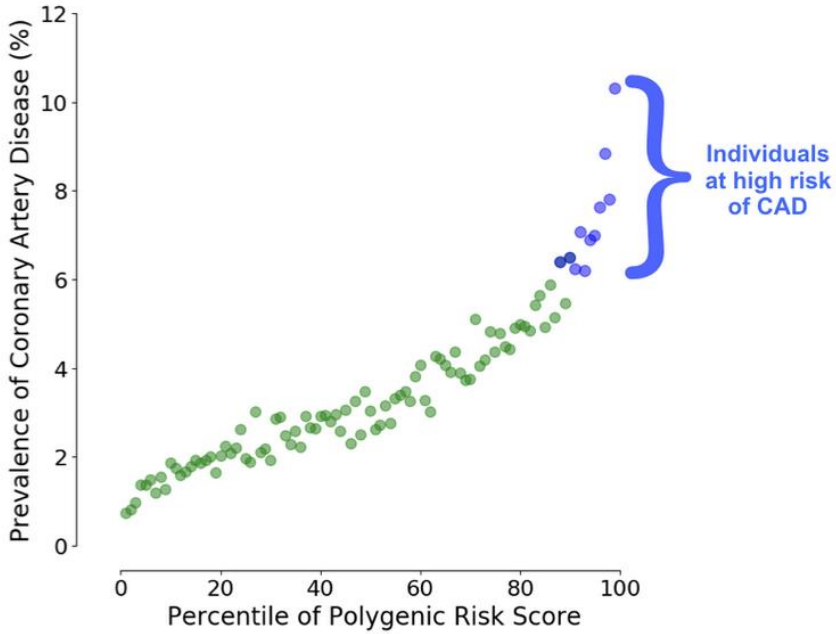
- Effect per QT prolonging allele = 1.53ms ± 0.08ms ($p=1.79*10^{-88}$)
- Several analysis of low vs. high QT score showed strong association to QT

Pfeufer et al. 2009, Nat Genet

Application example: https://www.cardioscore.eu/

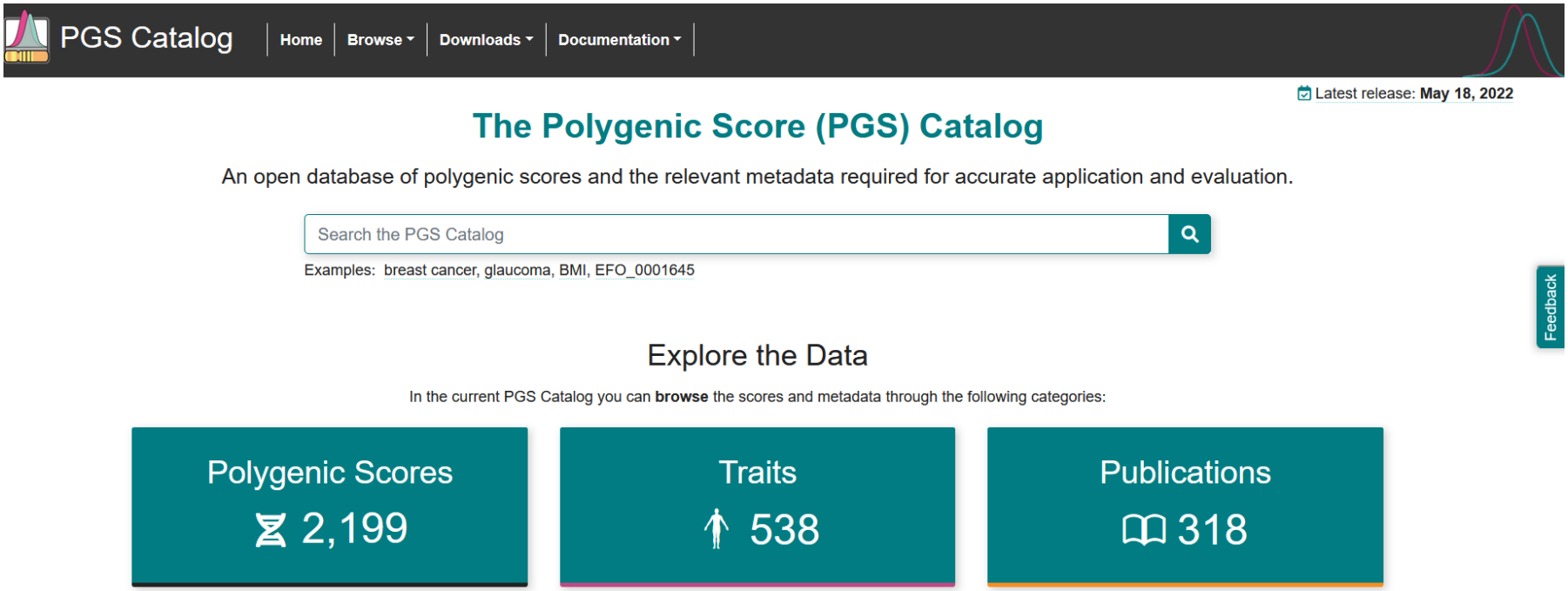


Cumulative absolute risk of developing Coronary Artery Disease (CAD) in UK Biobank population stratified by different percentiles of the Polygenic Risk Score (PRS)



Prevalence gradient for Coronary Artery Disease (CAD) across the distribution of the Polygenic Risk Score in UK Biobank population. Each point represents a different percentile of the PRS distribution

https://www.pgscatalog.org/



Vielen Dank für Ihre Aufmerksamkeit!

Kontakt:
Dr. Martina Müller-Nurasyid
Senior Scientist
IMBEI, Genomische Statistik und Bioinformatik
Universitätsmedizin Mainz
Langenbeckstraße 1, 55131 Mainz
Tel. +49 6131 39 – 38719
E-Mail: martimue@uni-mainz.de

www.unimedizin-mainz.de