ÖAW AUSTRIAN ACADEMY OF SCIENCES

GMI GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

Introduction and hands-on on GWAS

Haijun Liu
haijun.liu@gmi.oeaw.ac.at

Genomic Approaches Practical,
University of Vienna
09/05/2023

Part of Vienna BioCenter

Quick survey

- How familiar are you with the GWAS method?
  - I am fresh to GWAS *(in addition to the course last week)*
  - I know about GWAS but have never carried out a study
  - I am actively working in the GWAS field

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

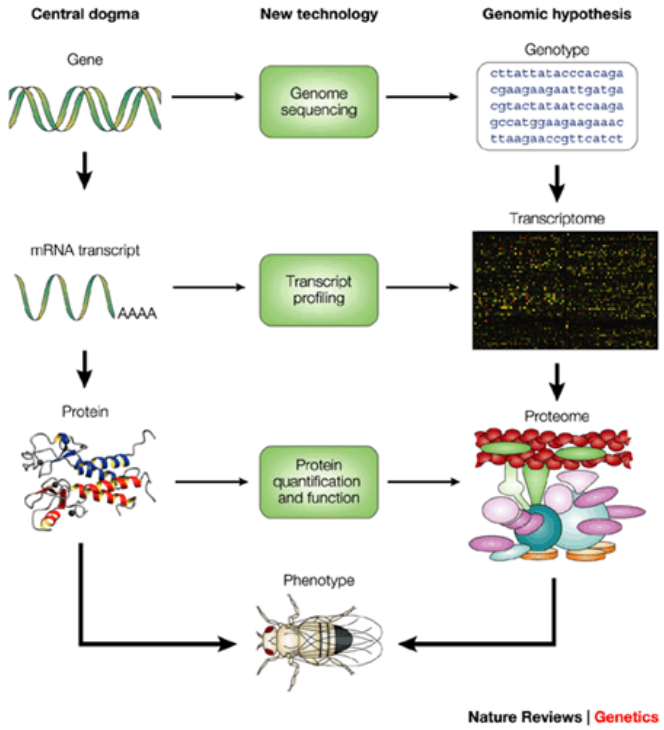
Plan today

- Why GWAS works (general view)
  - Principles, elements,
  - Factors affecting power
- **How to do GWAS ("standard" linear mixed model)**
  - **Populations/phenotyping**
  - **Genotyping**
  - **Models for association mapping**
  - **Hands-on!**
- How to interpret GWAS peaks and following
- Challenges and potential improvements

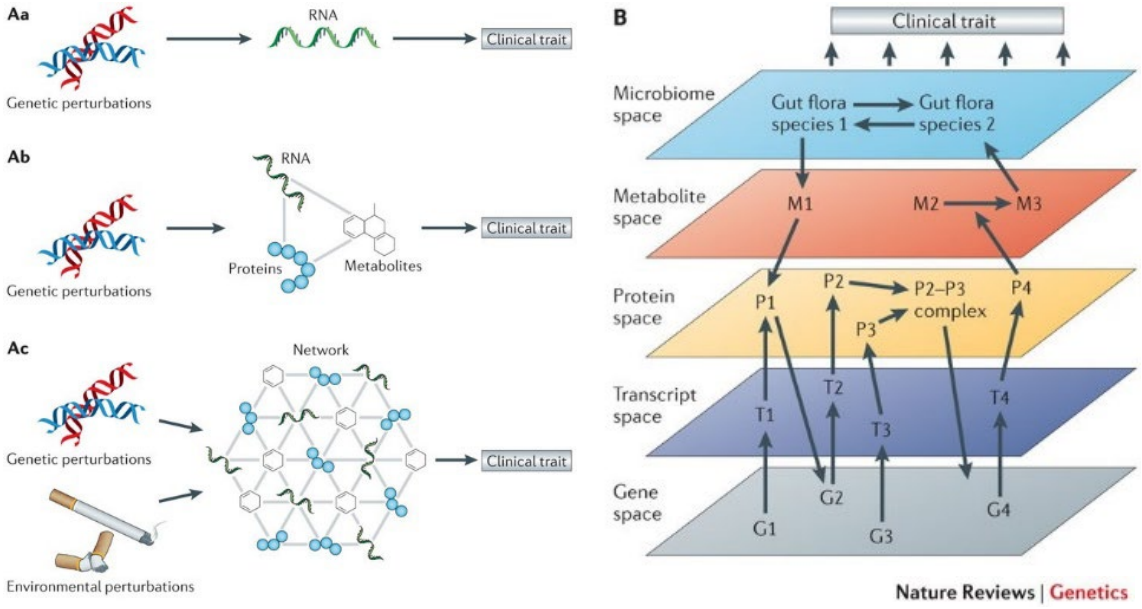
- ***Build a foundation to start GWAS on your own!***

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

The basis of *Biology*: central dogma



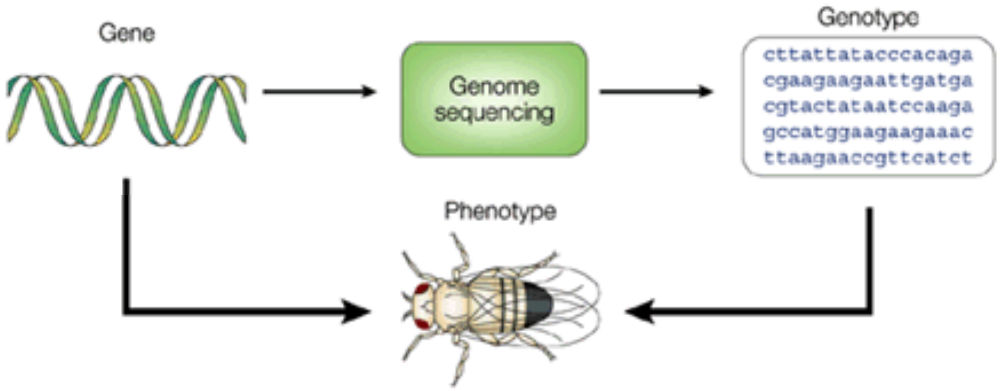
Doerge. *Nat Rev Genet*. 2002

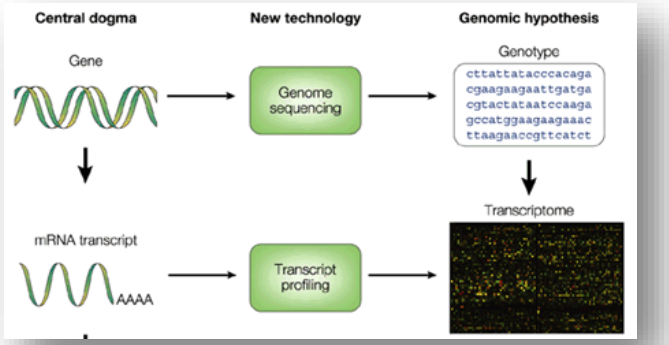


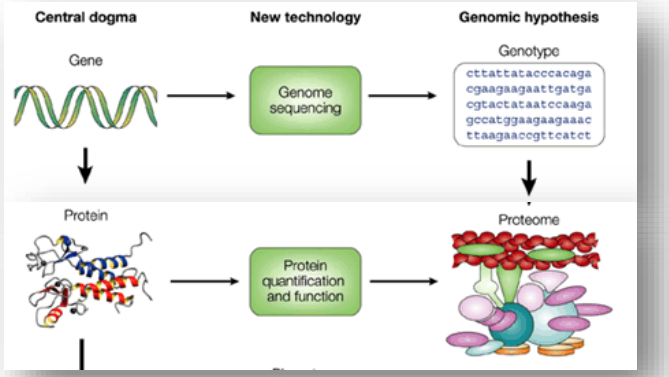
Civelek et al. *Nat Rev Genet*. 2014

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

The basis of *Genetics*: genotype-phenotype links







GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

GWAS success in G-P links

PUBLICATIONS: **5,848**
ASSOCIATIONS: **398,342**

2015-09

Now

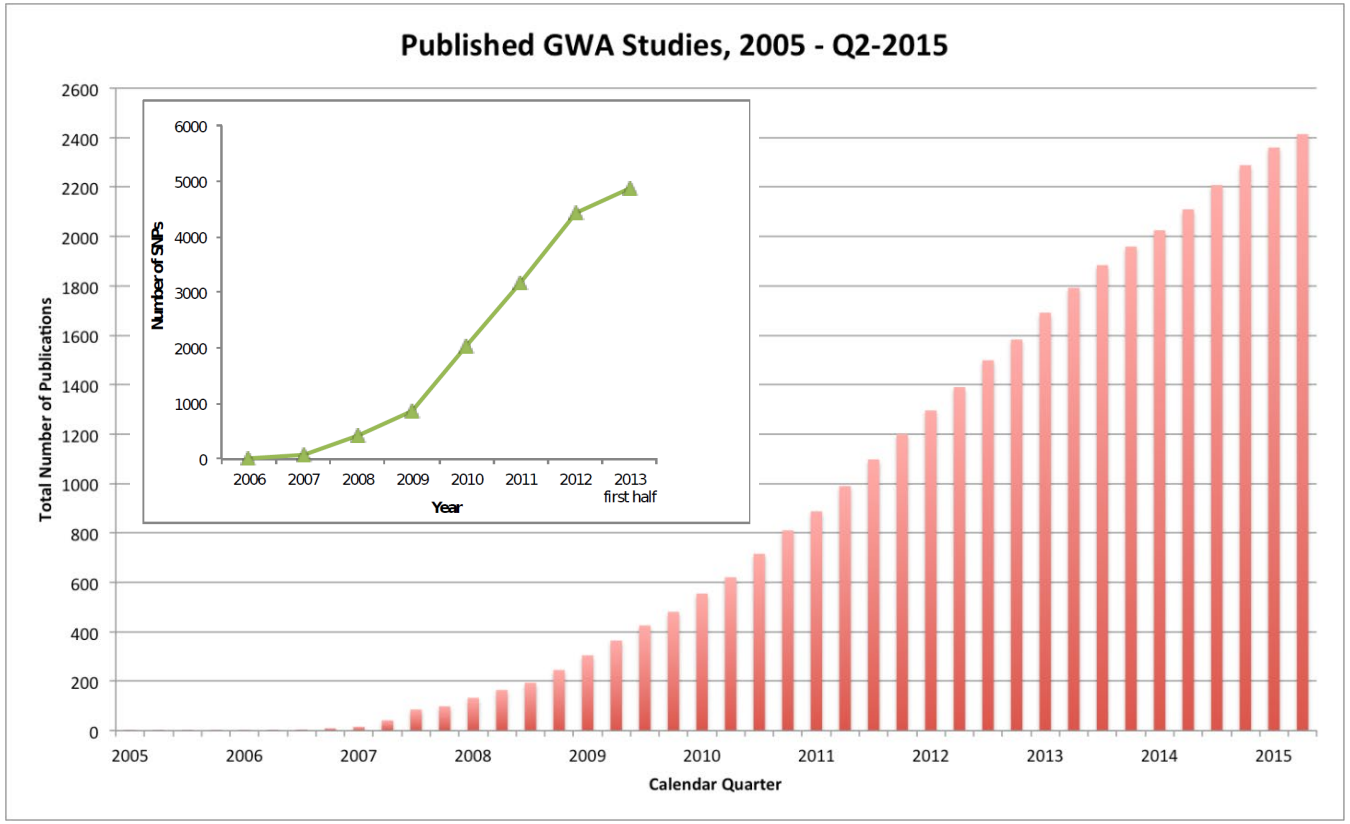
NHGRI-EBI GWAS Catalog
http://www.ebi.ac.uk/gwas/

EMBL-EBI

National Human Genome Research Institute

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

GWAS success in G-P links

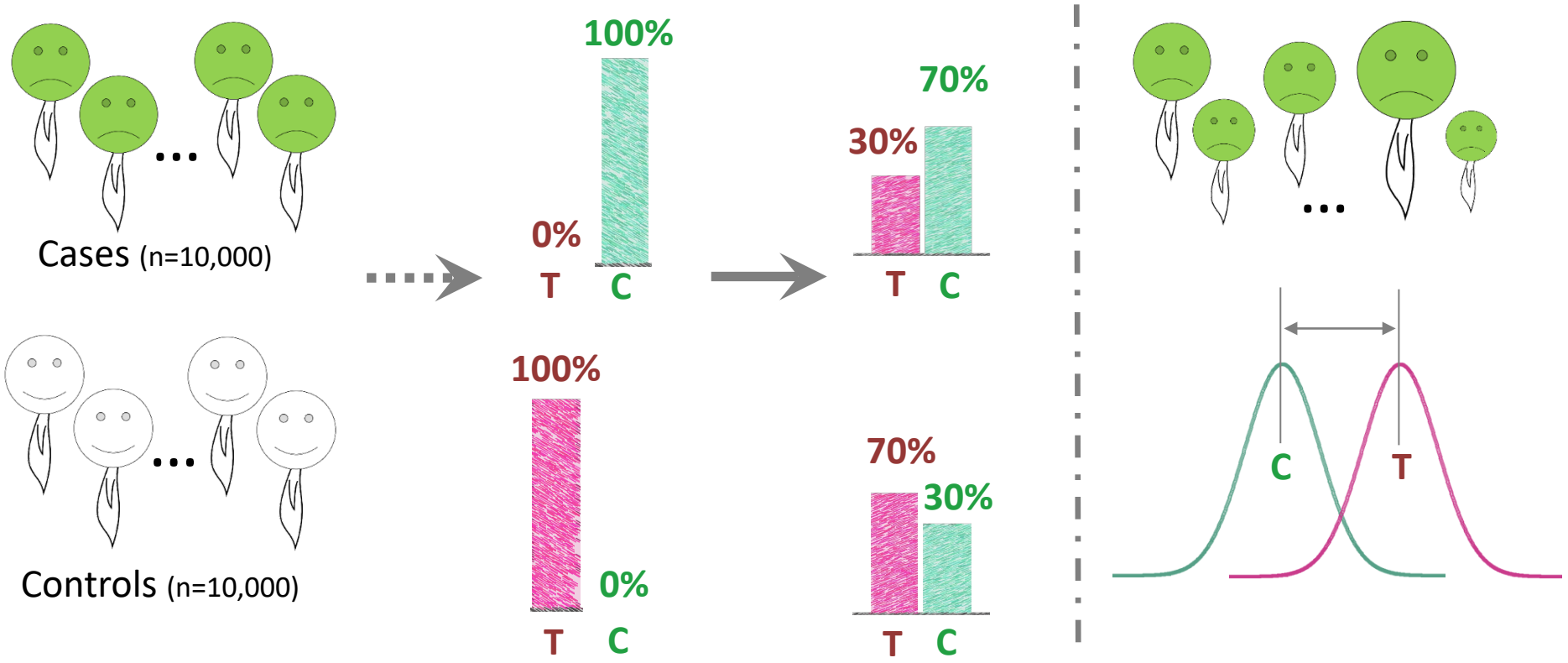


PUBLICATIONS: **5,848**
ASSOCIATIONS: **398,342**

- „Only 8 genes known for human complex traits until 2002”
- *Although there‘s still long distance from associations to causal genes*

- AraGWAS Catalog: 462 phenotypes & 44,680 associations

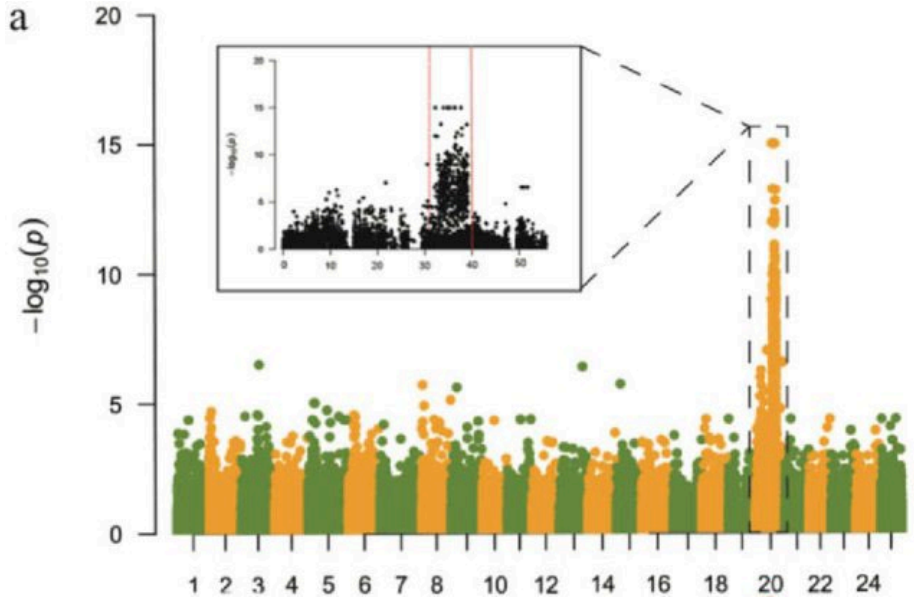
The intuition of genotype-phenotype links



Genome-wide association mapping

- To identify correlations between a phenotypic variation and whole genome genotypes
  - SNPs, INDELs, SVs
  - Test each locus independently*

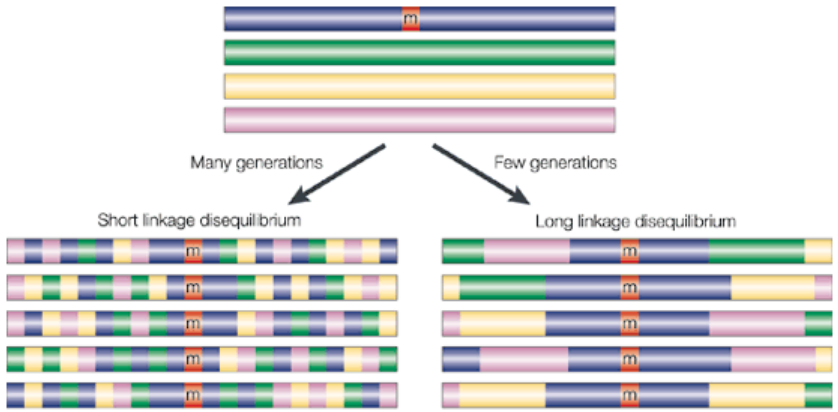
x millions of times



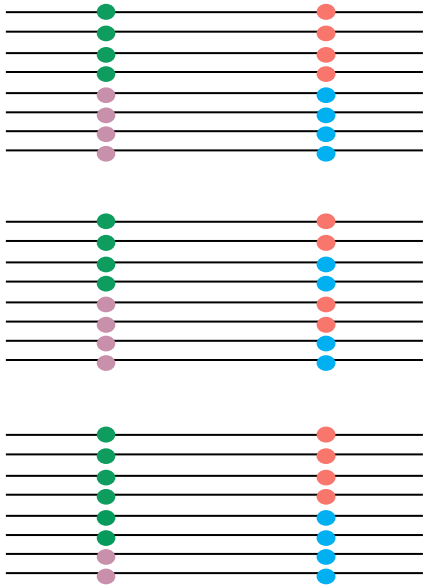
Bian et al. 2020

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

Core concept 1: linkage disequilibrium (LD)



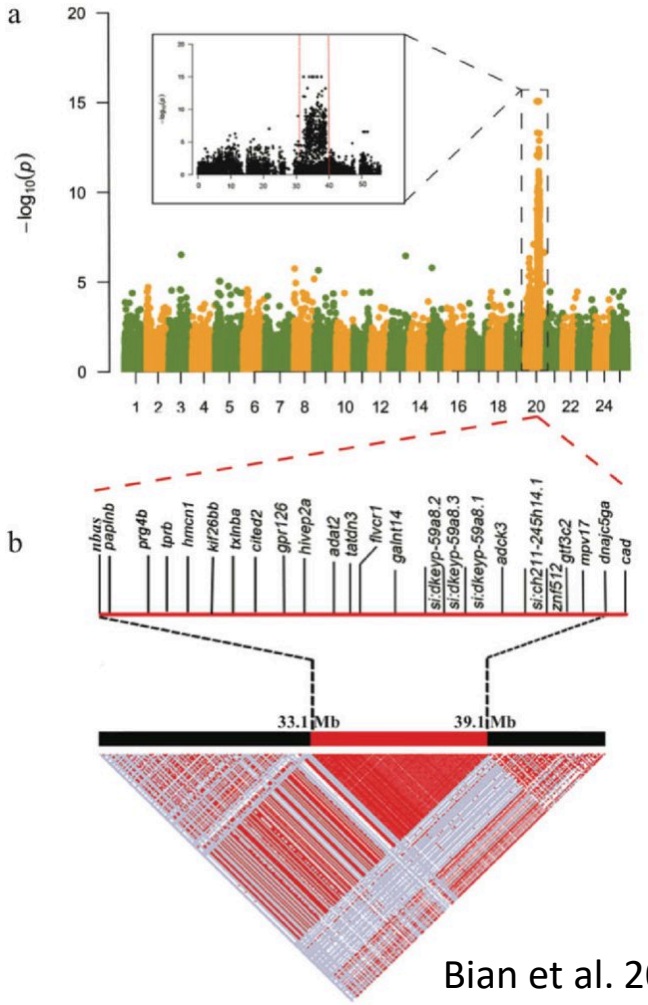
Ostrer, *Nat Rev Genet*, 2001



r^2=1

r^2=0

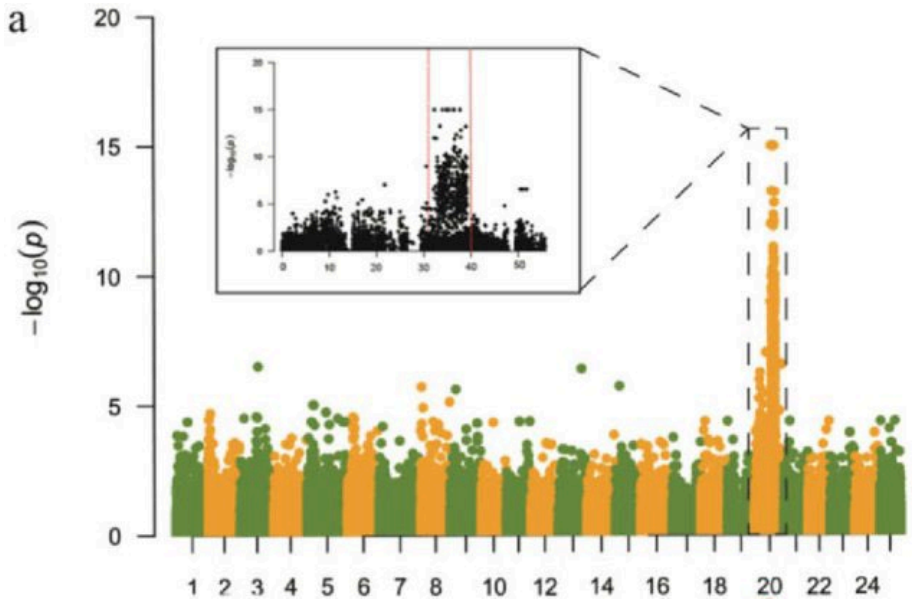
r^2=0.33



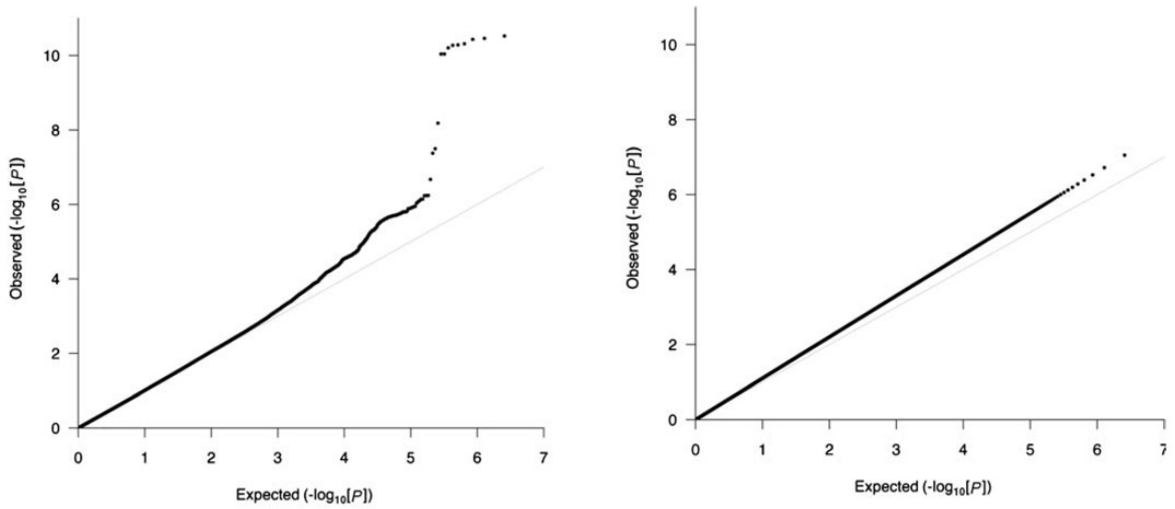
Bian et al. 2020

- LD — the nonrandom association of alleles at different / nearby loci
- No need to genotype all variants
- Power VS. Resolution VS. Causility

Core concept 2: Manhattan & quantile-quantile (QQ) plot



Significance ($-\log_{10}(P)$) along genome



Deviation of the observed Pvalues from the null

(essential for detecting the problems of GWAS)

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

Core concept 3: how to decide a cutoff of significance?

- Type 1 error (**false positive**) is the probability of incorrectly rejecting the null hypothesis
- **Trade-off**

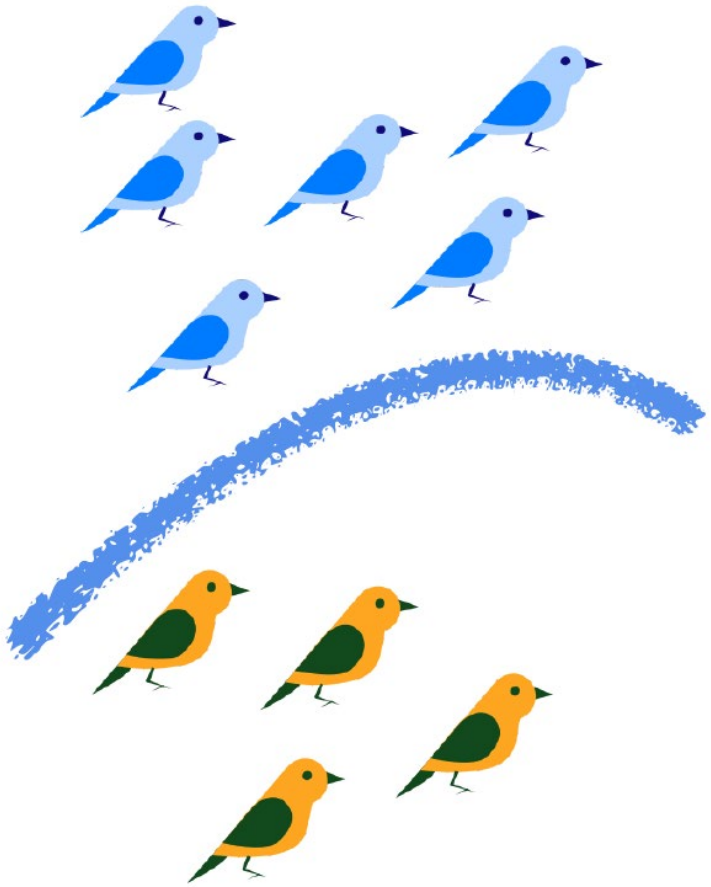
- N hypothesis tests (independently) with a Type 1 error of 0.05: incorrectly reject the null N*0.05
- **If N is large, we would make LOTS of errors**

- ***Multiple testing problem***: the more tests ➔ the greater probability of making a Type 1 error

- (As example) Bonferroni correction: $\alpha_b = \alpha/N$

Core concept 4: population structure

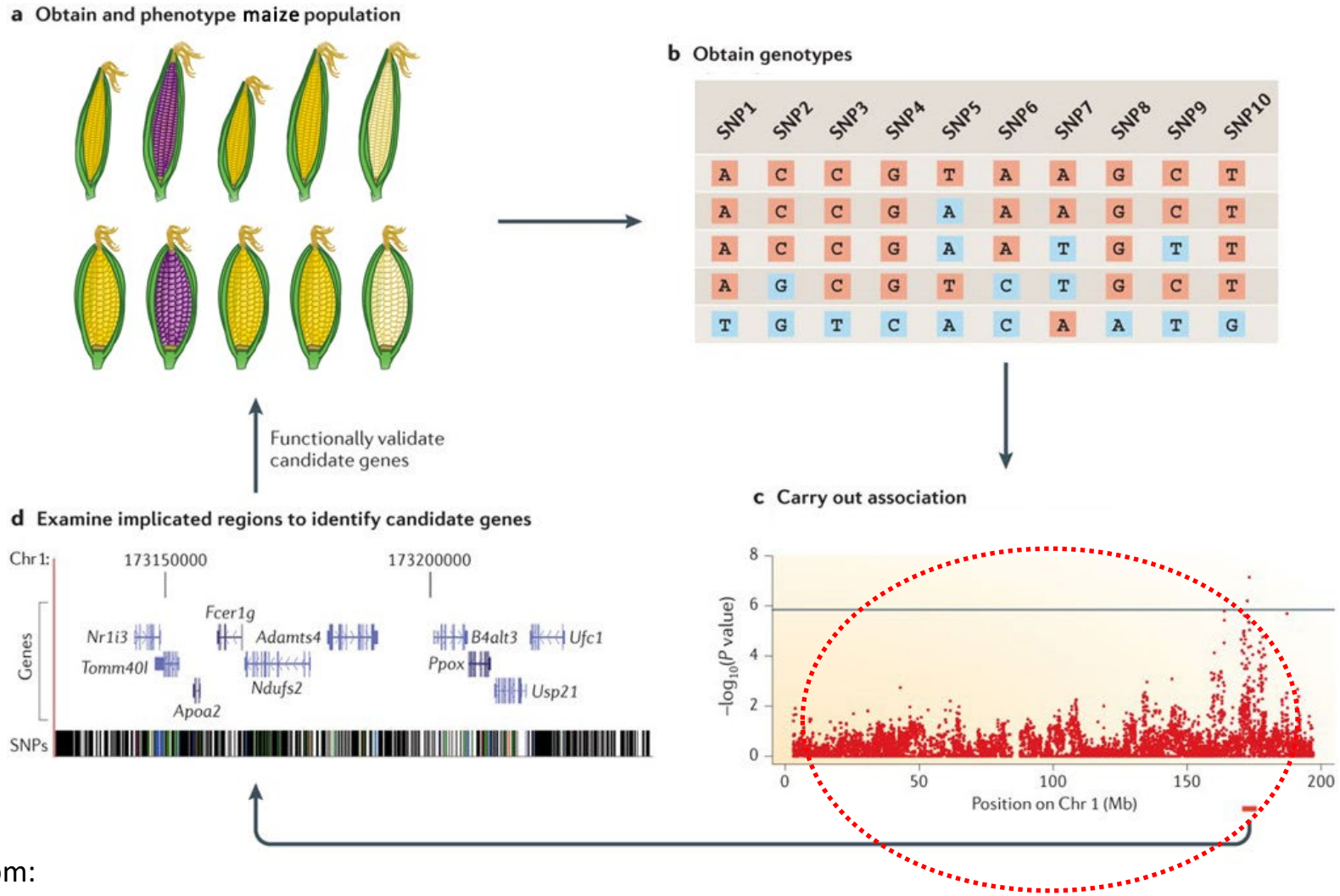
- Individuals within a population are more related than those between populations
- In a population share not only **causative** variants, but also **non-causative** variants that are more common in the population (genetic background)
- The ***K-matrix*** represents this background relatedness, should be took this background into account in GWAS to try to reduce the significance of non-causative variants.



GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

A typical flow of GWAS

Y_n



X_{n*m}

P_m

Modified from:
Flint & Eskin. *Nat Rev Genet*. 2012.

Element 1: population/phenotypes

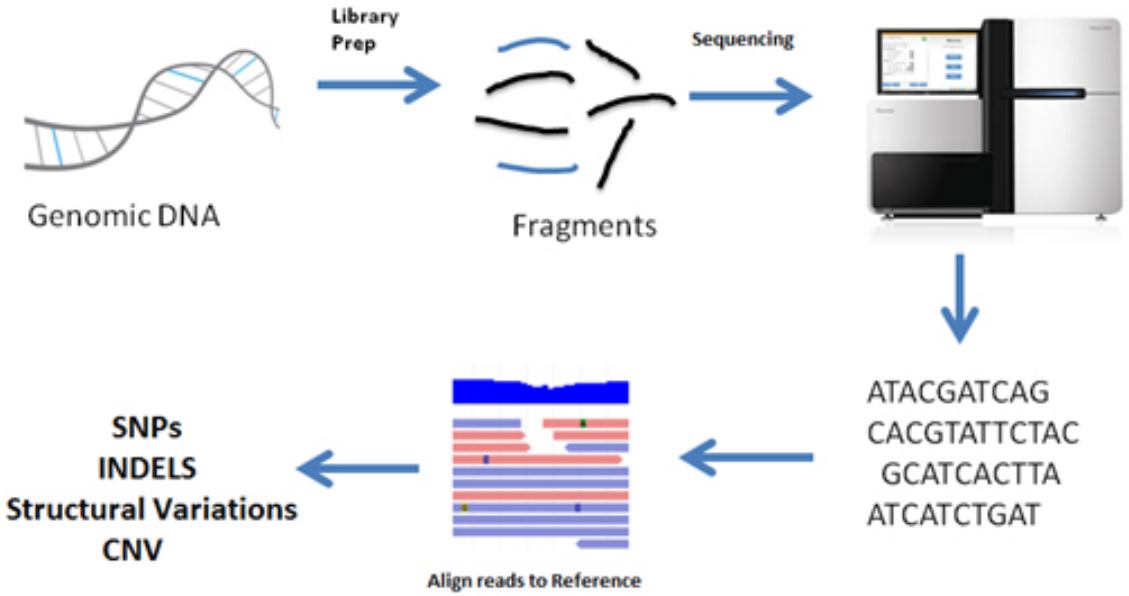
- Natural populations
  - Natural parellel ‚mutants' experiments lasting many many years
  - Once for all
  - Multiple omics data integration

  - Structured, environments, bias, et al (confounding factors)

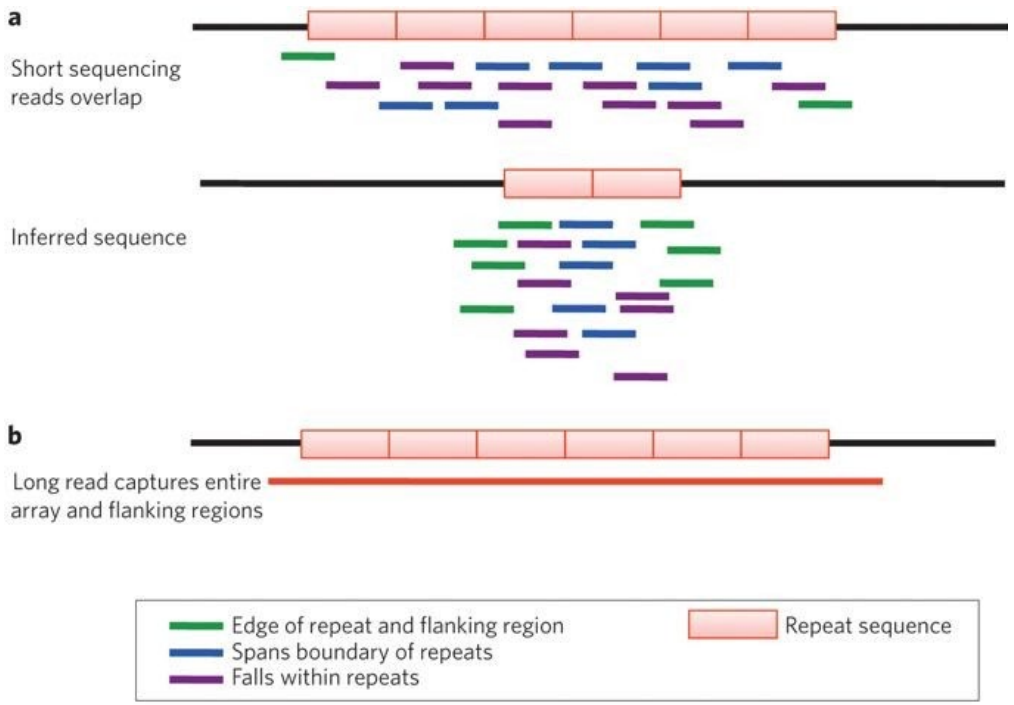
- Genetic design from natural populations
  - Good in plant study!

Element 2: genotyping

Next-GS (short-reads)



Third-GS (long-reads)



Yasir et al., 2022

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

Element 3: GWAS models (An evolution on correction for false positive)

Phenotype ~ Genetic + *non-G*

LM
(Linear Regression)

Phenotype ~ Genotype + e

GLM
(General Linear Model)

Phenotype ~ Genotype + Q + e

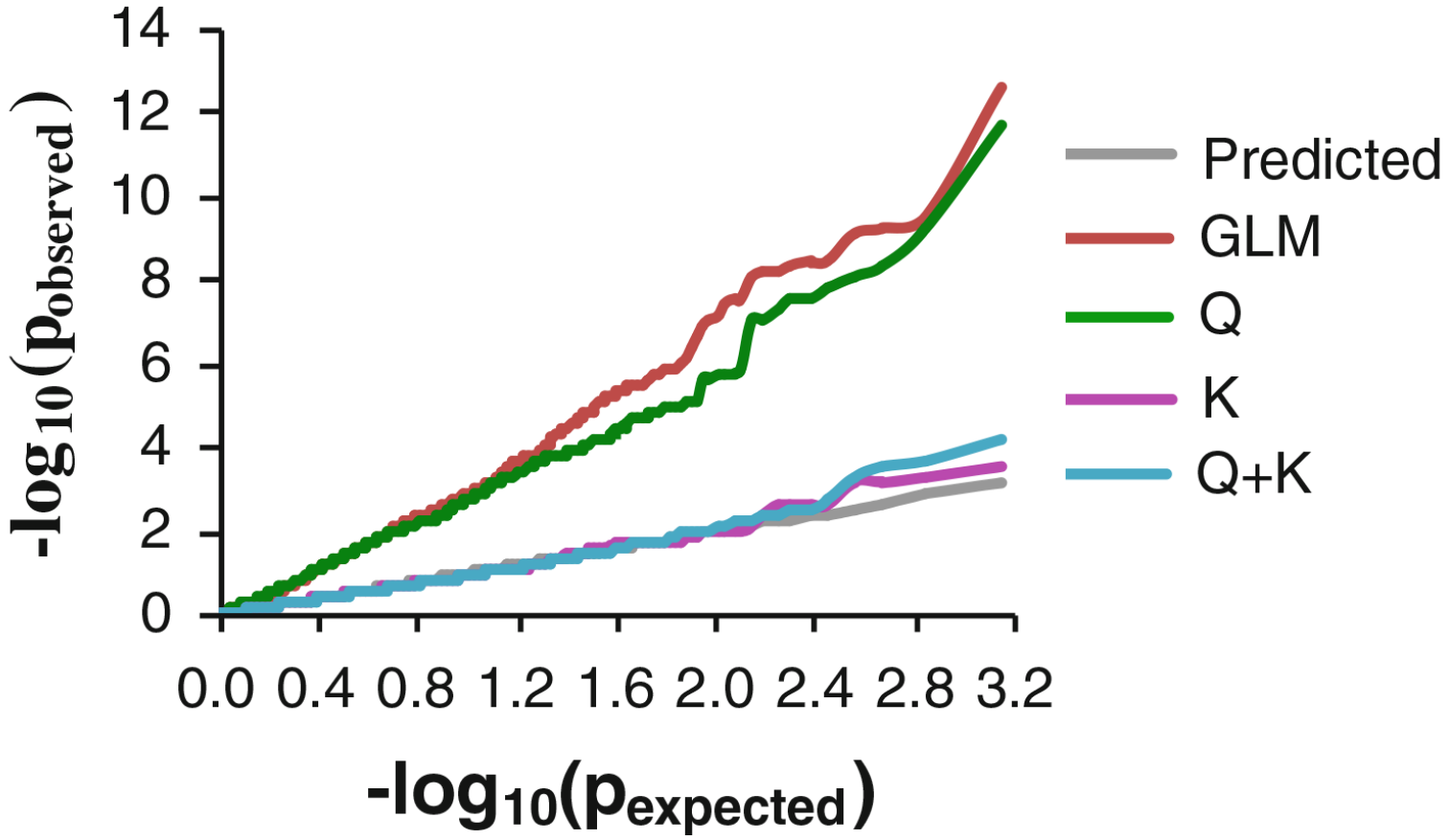
MLM
(Mixed Linear Model)

Phenotype ~ Genotype + Q + K + e

- Population structure (Pritchard et al, *Genetics*, 2000)
- PCA (Price et al, *Nat Genet*, 2006)
- Kinship (Yu et al, *Nat Genet,* 2006)

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

Impact of correction for false positive



Yang et al., 2010

Element 4: Make sense of your peaks!

Interested in candidate genes?

- Use biology, annotation, expression, etc.
- Sequence candidate genes/regions
- Test causality by QTL and/or transgenics in different backgrounds
- Which of my gene of interest would be associated with other phenotypes?

Interested in evolutionary or ecological inferences?

- Consider the consequences of your choice of accessions and phenotypes
- Associations with selective pressures (artificial / environmental)?
- Other signs of selection?
- The reference allele is not necessarily the ancestral allele!

One peak, one Ph.D, Be creative!

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

Questions / Comments?

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

Step-By-Step

https://github.com/heroalone/GWAS_HandsOn_Wien

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

Softwares for conducting GWAS

- TASSEL (https://www.maizegenetics.net/tassel)
- **limix (https://github.com/limix/limix)**
- emmax (https://genome.sph.umich.edu/wiki/EMMAX)
- gemma (https://github.com/genetics-statistics/GEMMA)
- and endless packages/programs …
  - maximum flexibility and options
  - allows for more complicated GWAS analyses
  - requires coding skills

Great to start!

GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

Let‘s do it!

- **Learn how to navigate a Jupyter notebook**
  - 0_running_Jupyter_notebooks.ipynb
- **Explore the phenotype we will use**
  - 1_phenotype_exploration.ipynb
- **Prepare input variables, run GWAS, and output results**
  - 2_GWAS.ipynb
- **Visualize and understand GWAS results**
  - 3_GWAS_interpretation.ipynb

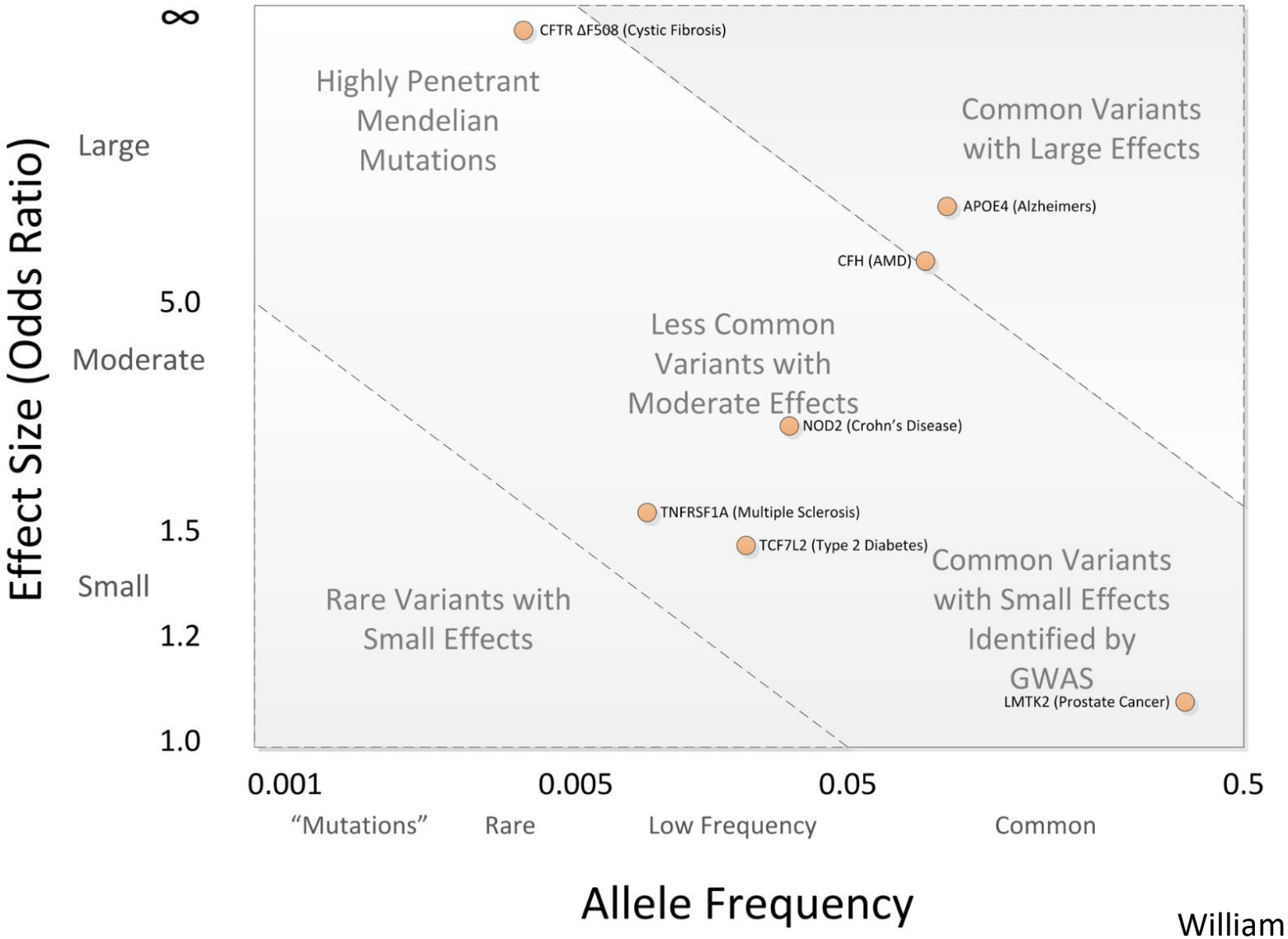
- *Check „Instructions_for_Independent_Work.pdf” if you work more quickly*

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

What can GWAS (basically) do

- Identify key genes/regulators the given phenotype
  - Drive biological hypothesis generation
- Generate insights on genetic architecture of phenotype
  - Key (large effect) regulators?
  - Many genes of small effect?
- Build statistical models to predict phenotype from genotype
  - Plant genome selection
  - Predict disease risk from an individual's genome

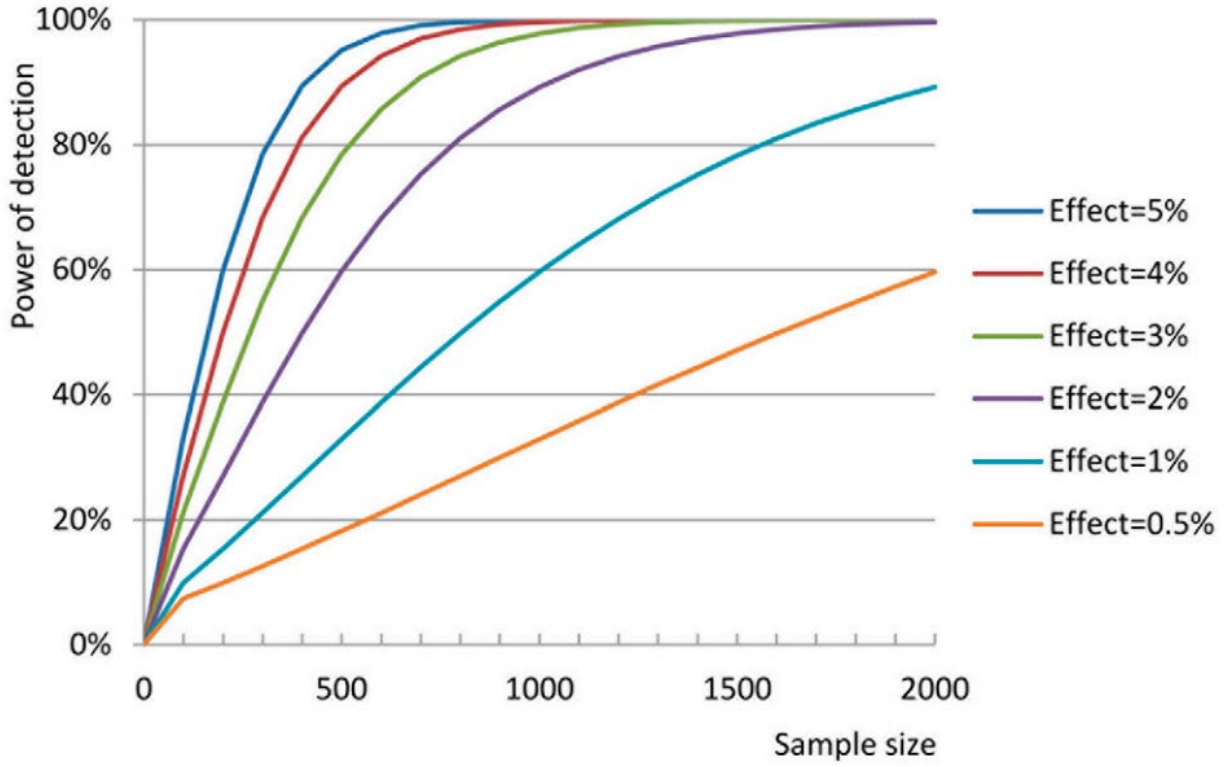
What GWAS limittedly do



William and Jason, 2012

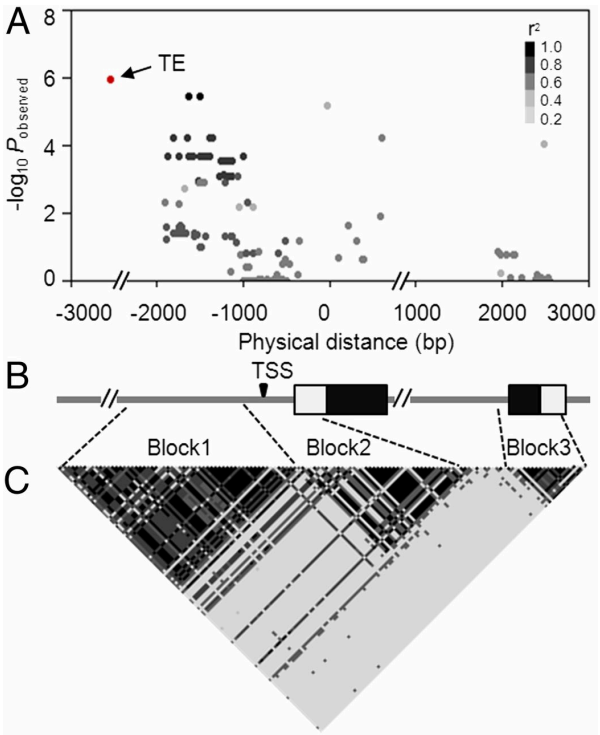
Factors affecting GWAS power

- Sample size
- Genetic architecture of given trait
- Population design & diversity
- Marker density & type
- etc.

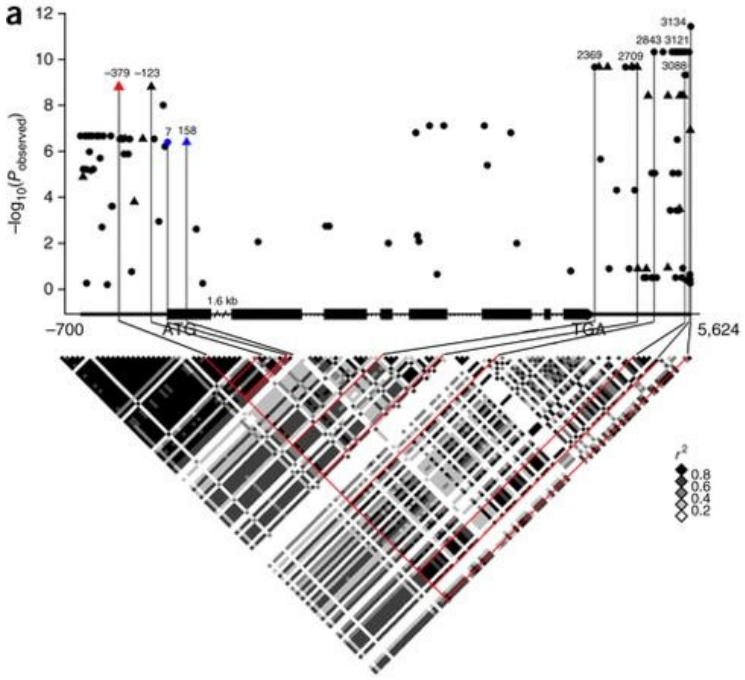


GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

Candidate gene/region association mapping



Yang et al., *PNAS*, 2013

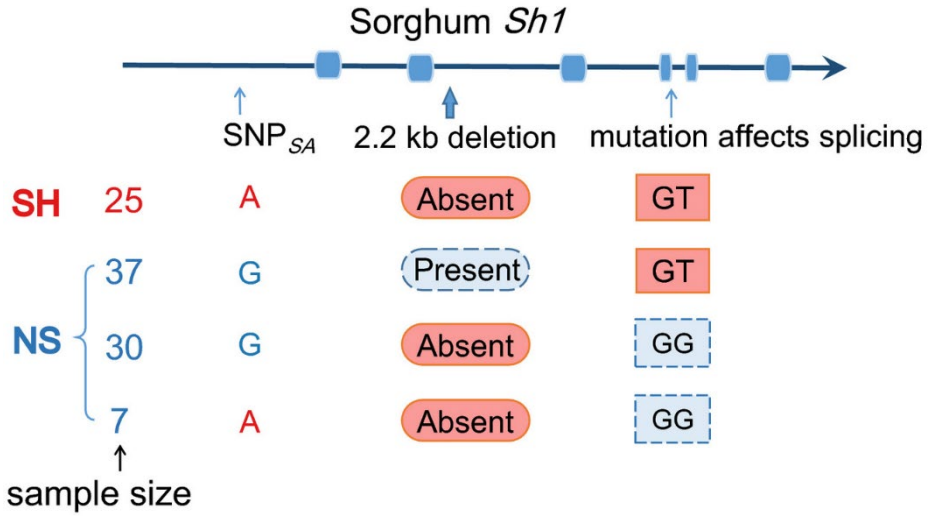


Wang et al., *Nat Genet*, 2016

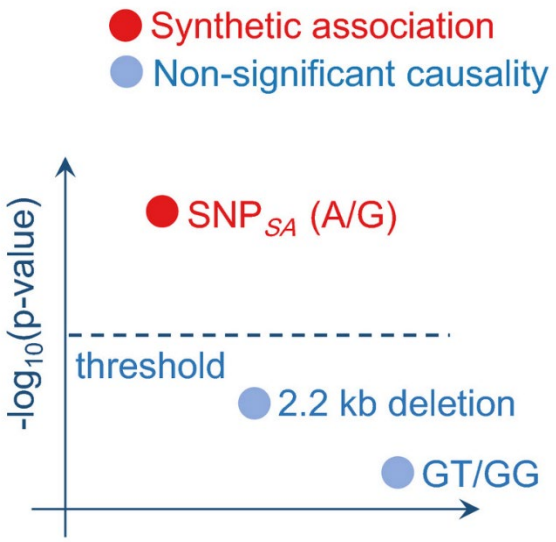
GMI
GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

Last but the most important: **Peaks are not necessarily causal**

- **Incomplete knowledge of variants**
- **Synthetic association**
- Indirect association
- …



- Further study is always required!
- *Good to know how many of exsiting hits are correct*



Liu and Yan, 2018

Why we should care: design, analysis and interpretation

(A well-known but UN-successful or even peril GWAS example)

- **Question:** why some people eat with chopsticks and others do not?
- **Methods:** several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and do association mapping

- **Results**: 'successful-use-of-selected-hand-instruments' (SUSHI) gene was found! And being repeated!

- SUSHI is a histocompatibility antigen gene, has nothing to do with chopstick use but just happens to have different allele frequencies in Asians and Caucasians
- Differ in chopstick use for purely cultural rather than biological reasons!

Hamer and Sirota. 2000

The GWAS future *(personal view)*

- Larger sample size?!
- More advanced design
- More variant types, more omics ‚phenotype/genotypes‘

- New analytical method/strategies !!
- Type 2 error (false negative)
- Interaction
- GxE

- Large-scale causal gene/variant and functional validation
  - Dry and Wet!

- Genomic prediction

Further resources

- Online GWAS Course
- Jason Mezey, BTRY6830
- http://mezeylab.cb.bscb.cornell.edu/Classes.aspx

- Nature Reviews Genetics, Article series, GWAS Collections
  - https://www.nature.com/collections/jpqdqjwqkk
  - (Although the page is no longer updated, since 2013)
- Nature portfolio GWAS subjects:
  - https://www.nature.com/subjects/genome-wide-association-studies
  - Latest Research and Reviews!

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

Questions/Comments?
(Not the final chance!)

haijun.liu@gmi.oeaw.ac.at

GMI
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY