Lecture 3. Genome-wide association studies (GWAS)

Genome-Wide Association Study (GWAS): Method Overview

1. **What is GWAS?**

Dep. Biotechnology Prof. Kenzhebayeva S.S.

Aim of the Lecture. Expected Learning Outcomes

Aim of the Lecture

To introduce the fundamental principles, methodological approaches, and applications of genome-wide association studies (GWAS) for identifying genetic variants associated with complex traits, with emphasis on statistical analysis, candidate gene discovery, and applications in modern biotechnology and crop improvement.

- **Expected Learning Outcomes**
- Define genome-wide association studies (GWAS) and explain their significance in modern genetics and postgenomic research.
- Describe the genetic principles underlying GWAS, including genetic variation, single-nucleotide polymorphisms (SNPs), and linkage disequilibrium (LD).
- Explain the main stages of GWAS, including population selection, genotyping, phenotyping, and data quality control.
- Distinguish between GWAS and traditional linkage-based QTL mapping.
- Describe the basic statistical approaches used to identify marker–trait associations.
- Interpret GWAS results using Manhattan plots and quantile–quantile (Q–Q) plots.
- Explain the importance of population structure, kinship, and multiple-testing correction in GWAS.
- Identify candidate genes and genomic regions associated with complex biological traits.
- Discuss applications of GWAS in plant, medical, animal, and microbial biotechnology.
- Evaluate the advantages, limitations, and potential of GWAS for marker-assisted selection and genomic breeding.

Key learning outcome: Understanding how genome-wide genetic variation can be associated with phenotypic diversity to identify candidate loci and support functional genomics and biotechnology applications.

Lecture Plan

1. Introduction to GWAS: Definition, objectives, historical development, and significance in postgenomic research.
2. Genetic Basis of GWAS: Genetic variation, SNP markers, linkage disequilibrium (LD), and haplotypes.
3. GWAS Experimental Design: Association populations, phenotyping, genotyping, and quality control.
4. Statistical Methods in GWAS: Association models, population structure, kinship, and multiple-testing correction.
5. Visualization and Interpretation of GWAS Results: Manhattan plots, Q–Q plots, and identification of significant marker–trait associations.
6. Candidate Gene Identification and Functional Validation: Genomic annotation, candidate gene prioritization, and experimental validation.
7. Applications of GWAS in Biotechnology: Plant breeding, medical genetics, animal biotechnology, and microbial research.
8. Challenges and Future Perspectives: Limitations of GWAS, integration with multi-omics technologies, and genomic prediction.

Genome-Wide Association Study (GWAS)

is a statistical approach used to identify **genetic variants (primarily SNPs) that are significantly associated with phenotypic traits such as stress tolerance, yield, disease resistance, or grain nutritional quality.**

GWAS leverages natural genetic diversity and historical recombination in populations, without the need to develop biparental mapping populations.

In genetics, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait.

GWA studies typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits that can equally be applied to any other genetic variants and any other organisms.

2. Main Steps of the GWAS Pipeline

- **A. Population selection**

GWAS requires a genetically diverse panel:

1. **natural diversity panels**
2. **breeding germplasm collections**
3. **mutant panels**
4. **MAGIC populations or core collections**
5. **Key requirements:**
6. **high genetic variability**
7. **known or controlled population structure**
8. **sufficient sample size (typically ≥ 200–300 genotypes)**

B. Genotyping

Genotyping platforms include:

1. **high-density SNP arrays (35K–660K)**
2. **GBS (Genotyping-by-Sequencing)**
3. **whole-genome resequencing (WGR)**
4. **Data preprocessing typically includes:**
5. **filtering by minor allele frequency (MAF > 0.05)**
6. **filtering by missingness (<10–20%)**
7. **Hardy–Weinberg equilibrium checks**
8. **linkage-disequilibrium pruning (optional).**

B. Genotyping

Data preprocessing typically includes:

1. filtering by **minor allele frequency (MAF > 0.05)**
2. filtering by **missingness (<10–20%)**
3. Hardy–Weinberg equilibrium checks
4. linkage-disequilibrium pruning (optional)

C. Phenotyping

1. **Accurate phenotyping is essential for reliable GWAS results:**
2. **agronomic traits: yield, plant height, flowering time**
3. **stress-related traits: drought, salinity, heat, disease resistance**
4. **grain nutritional quality: Fe/Zn/Mg content (AAS, ICP-MS, NIRS)**
5. **physiological and biochemical traits: ROS, proline, antioxidant enzymes**
6. **Multi-environment and multi-year phenotyping improves statistical power and reduces noise.**

D. Correcting for population structure

To avoid false positives caused by stratification, GWAS models account for:

1. **Principal Components (PCs)**
2. **Kinship (K-matrix)** ***родство***
3. **Q-matrix** (from STRUCTURE)

These covariates control background genomic relationships.

Statistical models used in GWAS

1. MLM (Mixed Linear Model)

1. Most widely used.
   Model: **y = Xβ + Zu + e**, where
2. *β* = fixed effects (SNP, PCs)
3. *u* = random genetic background (kinship)

2. MLMM (Multi-locus Mixed Model)

- Improved multi-locus detection.

3. FarmCPU

- Separates fixed and random effects → reduces false positives, increases power.

4. BLINK

- Improved version of FarmCPU; very fast for large datasets.

5. Bayesian models (BayesCπ, BayesR)

Useful for highly polygenic traits

3. GWAS Outputs and Interpretation

A. Manhattan plot

Each point = a SNP; peaks indicate

genomic regions (QTL) significantly associated with the trait.

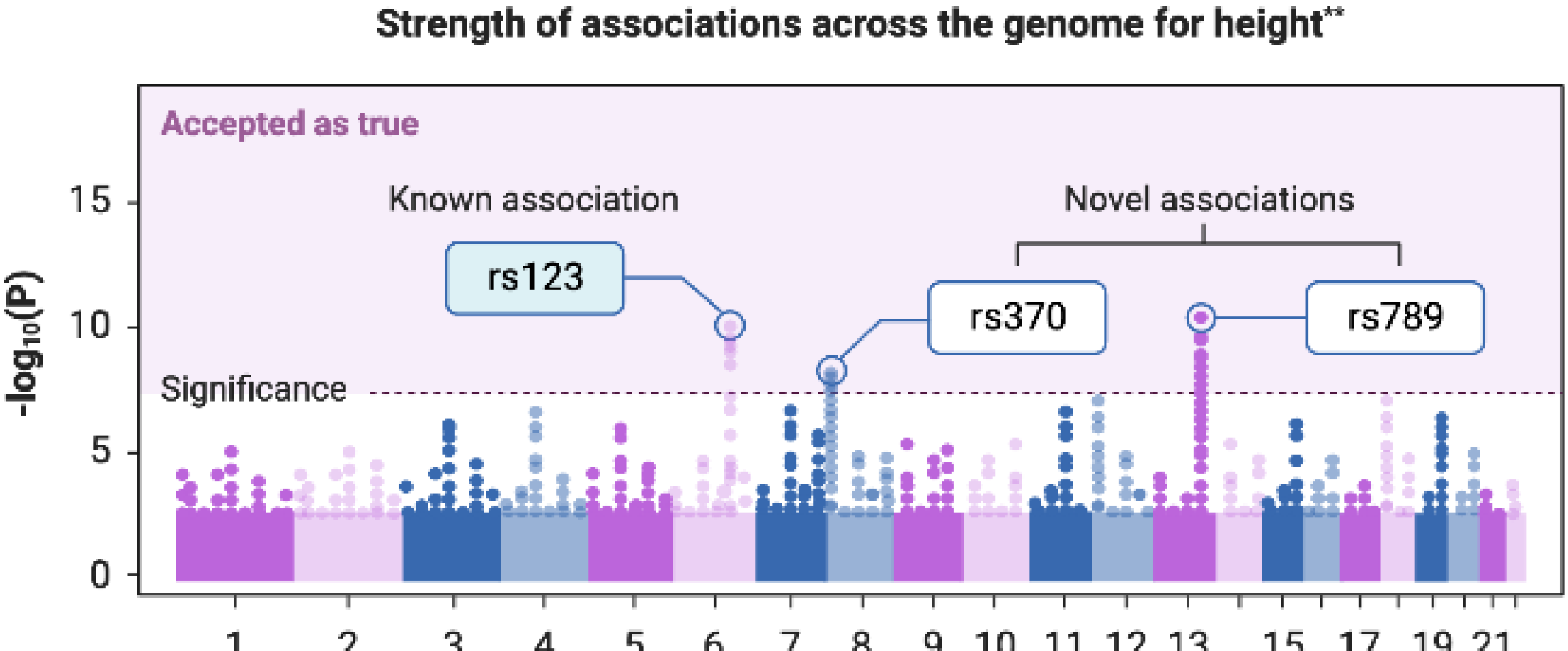
Significance thresholds:

Bonferroni correction

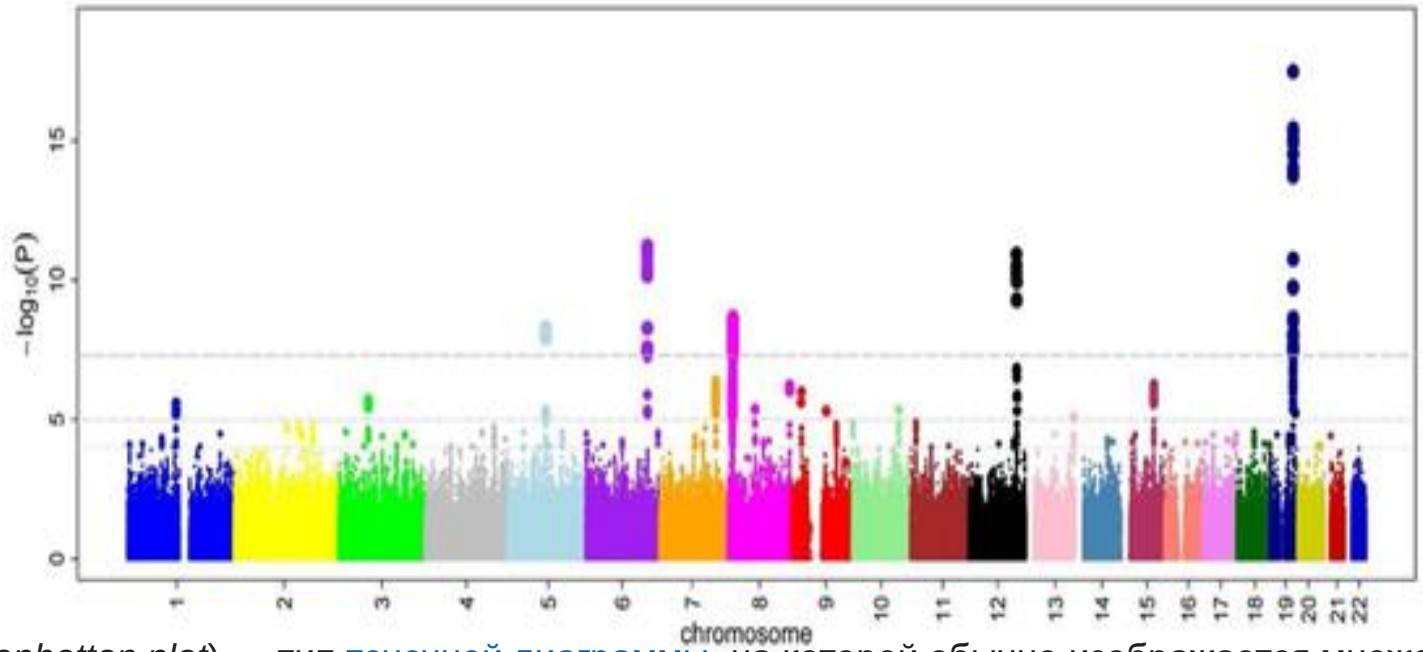
False Discovery Rate (FDR)

permutation tests

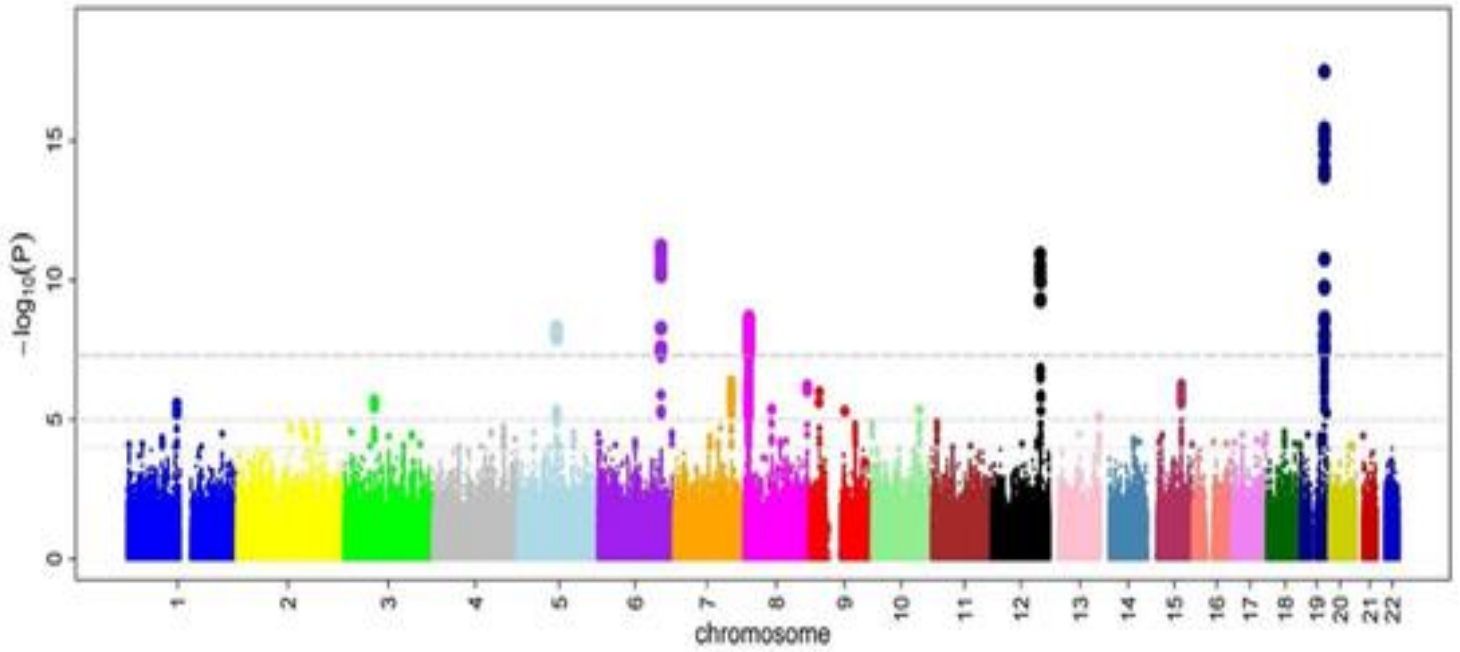
Manhattan Plot for Genome-Wide Association Studies (GWAS)



An illustration of a Manhattan plot depicting several strongly associated risk loci.
Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level.
This example is taken from a GWA study investigating microcirculation, so the tops indicates genetic variants that more often are found in individuals with constrictions in small blood vessels



анхэттенский график (англ. *Manhattan plot*) — тип точечной диаграммы, на которой обычно изображается множество значений, большинство из которых близки к нулю. Часто используется в полногеномных поисках ассоциаций (GWAS). При обработке результатов GWAS на оси X отмечаются геномные координаты. На оси Y отмечается значение, противоположенное десятичному логарифму P-значения для каждого однонуклеотидного полиморфизма. Так как полифорфизмы, которые имеют наибольшую ассоциацию с изучаемым признаком, имеют наименьшее P-значение, то их значение на графике будет иметь максимальное значение.
График получил свое название из-за сходства с панорамой Манхэттена с высокими небоскребами, возвышающимися над



Манхэттенский график (англ. *Manhattan plot*) — тип **точечной диаграммы**, на которой обычно изображается множество значений, большинство из которых близки к нулю. Часто используется в **полногеномных поисках ассоциаций (GWAS).**

При обработке результатов GWAS на оси X отмечаются геномные координаты.

На оси Y отмечается значение, противоположенное **десятичному логарифму P-значения** для каждого **однонуклеотидного полиморфизма.**

Так как полифорфизмы, которые имеют наибольшую ассоциацию с изучаемым признаком, имеют **наименьшее P-значение**, то их значение на графике будет иметь максимальное значение.

График получил свое название из-за сходства с панорамой Манхэттена с высокими небоскребами, возвышающимися над массой более низких зданий.

C. Candidate gene identification

For each significant SNP:

1. inspect **local LD region** (±250 kb or based on LD decay)
2. identify genes from genome annotations (RefSeq v1.0/v2.1)
3. evaluate gene functions (GO, KEGG, Pfam)
4. examine expression profiles (RNA-seq, qPCR)

Often, researchers identify:

- **significant SNPs**
- **candidate QTL**
- **candidate genes**
- **haplotypes and favorable alleles**

4. Validation of GWAS Findings

After identifying significant loci, validation is essential:

- **1. Comparison with known QTL**

Check overlaps with previously reported QTL in literature.

- **2. Haplotype analysis**

Examine the effect of different haplotypes on the trait.

- **3. Expression validation**

qPCR or RNA-seq to confirm transcriptional differences of candidate genes.

- **4. Functional validation**

mutant lines (TILLING, EMS)

CRISPR/Cas knockout or overexpression studies

association tests in an independent population

5. Advantages of GWAS. Limitations of GWAS

1. **high mapping resolution**
2. **uses natural genetic variation**
3. **suitable for complex polygenic traits**
4. **scalable to large datasets**

Limitations

1. population stratification can create false associations
2. requires high-quality phenotyping
3. reduced power for rare alleles (MAF < 0.05)
4. complicated LD patterns in polyploid genomes (e.g., wheat)

6. Typical GWAS Workflow (Panels A–C)

- **Panel A — Genotyping (SNP matrix)**
- **Panel B — Phenotyping (trait measurement)**
- **Panel C — Association Analysis (Manhattan and QQ plots)**

- **When applied to human data, GWA studies compare the DNA of participants having varying phenotypes for a particular trait or disease.**
- **These participants may be people with a disease (cases) and similar people without the disease (controls), or they may be people with different phenotypes for a particular trait, for example blood pressure.**
- **This approach is known as phenotype-first, in which the participants are classified first by their clinical manifestation(s), as opposed to genotype-first.**

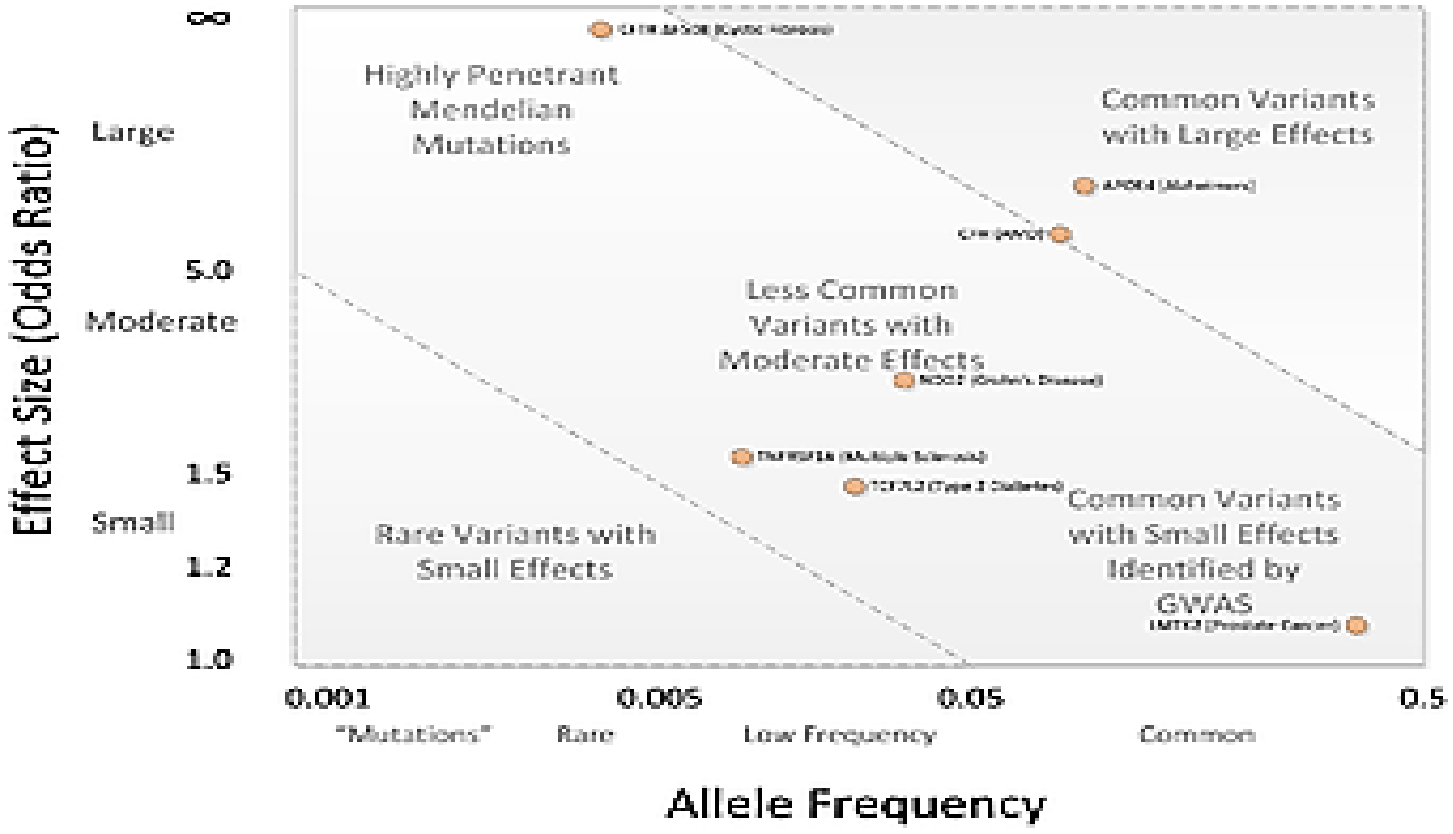
- **Each person gives a sample of DNA, from which millions of genetic variants are read using SNP arrays.**
- **If one type of the variant (one allele) is more frequent in people with the disease, the variant is *associated* with the disease.**
- **The associated SNPs are then considered to mark a region of the human genome that may influence the risk of disease.**

GWA studies

- investigate the entire genome, in contrast to methods that specifically test a small number of pre-specified genetic regions.
- Hence, GWAS is a *non-candidate-driven* approach, in contrast to ***gene-specific candidate-driven studies***.
- GWA studies identify SNPs and other variants in DNA associated with a disease, but they cannot on their own specify which genes are causal

- **The first successful GWAS published in 2002 studied myocardial infarction.**
- **This study design was then implemented in the landmark GWA 2005 study investigating patients with age-related macular degeneration, and found two SNPs with significantly altered allele frequency compared to healthy controls.**
- **As of 2017, over 3,000 human GWA studies have examined over 1,800 diseases and traits, and thousands of SNP associations have been found.**
- **Except in the case of rare genetic diseases, these associations are very weak, but while they may not explain much of the risk, they provide insight into genes and pathways that can be important**

GWA studies typically identify common variants with small effect sizes (*lower right*)

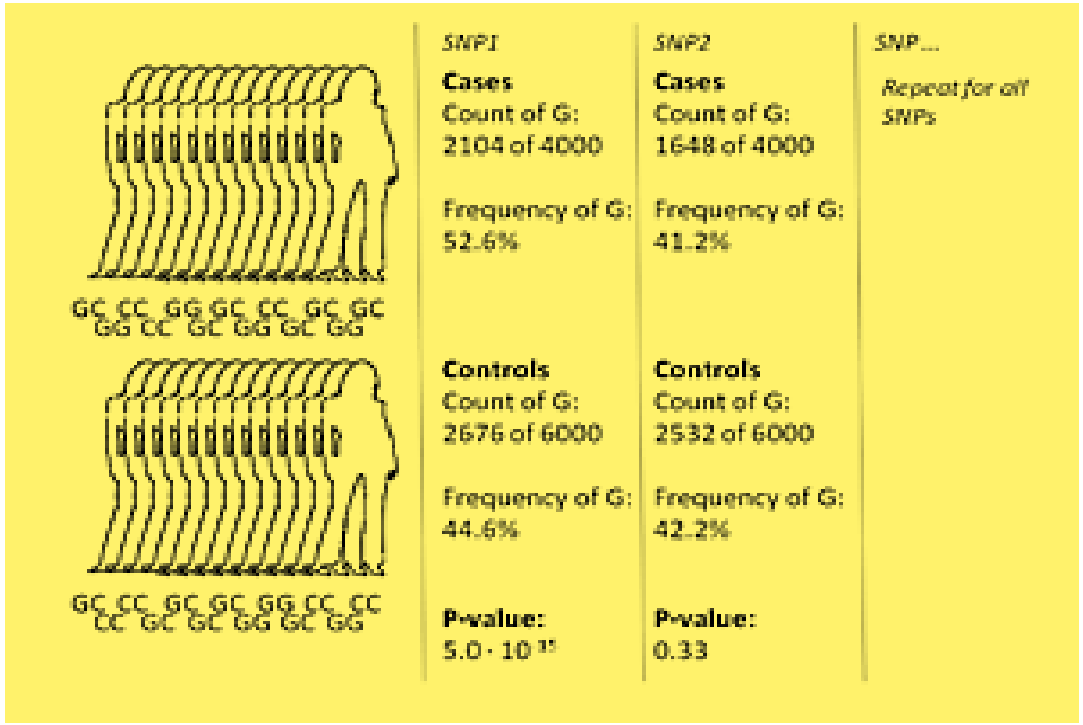


- Any two human genomes differ in millions of different ways.
- **There are small variations in the individual nucleotides of the genomes (SNPs) as well as many larger variations, such as deletions, insertions and copy number variations.**
- Any of these may cause alterations in an individual's traits, or phenotype, which can be anything from disease risk to physical properties such as height.
- Around the year 2000, prior to the introduction of GWA studies, the primary method of investigation was through inheritance studies of **genetic linkage** in families.

- **The approach - genetic linkage had proven highly useful towards single gene disorders.**
- **However, for common and complex diseases the results of genetic linkage studies proved hard to reproduce.**
- **A suggested alternative to linkage studies was the genetic association study.**
-
- **This study type asks if the allele of a genetic variant is found more often than expected in individuals with the phenotype of interest (e.g. with the disease being studied).**
- **Early calculations on statistical power indicated that this approach could be better than linkage studies at detecting weak genetic effects**

- In addition to the conceptual framework several additional factors enabled the GWA studies.
- One was the advent of biobanks, which are repositories of human genetic material that greatly reduced the cost and difficulty of collecting sufficient numbers of biological specimens for study.
- Another was the International HapMap Project, which, from 2003 identified a majority of the common SNPs interrogated in a GWA study.
- The haploblock structure identified by HapMap project also allowed the focus on the subset of SNPs that would describe most of the variation.
- Also the development of the methods to genotype all these SNPs using genotyping arrays was an important prerequisite

Methods



- Example calculation illustrating the methodology of a case-control GWA study. The allele count of each measured SNP is evaluated—in this case with a chi-squared test—to identify variants associated with the trait in question.
- The numbers in this example are taken from a 2007 study of coronary artery disease (CAD) that showed that the individuals with the G-allele of SNP1 (*rs1333049*) were overrepresented amongst CAD-patients

- **The most common approach of GWA studies is the case-control setup (*настройка контроля случаев*) which compares two large groups of individuals, one healthy control group and one case group affected by a disease.**
- **All individuals in each group are genotyped for the majority of common known SNPs.**
- **The exact number of SNPs depends on the genotyping technology, but are typically one million or more.**
- **For each of these SNPs it is then investigated if the allele frequency is significantly altered between the case and the control group.**
- **In such setups, the fundamental unit for reporting effect sizes is the odds ratio (соотношение шансов).**
- **The odds ratio is the ratio of two odds, which in the context of GWA studies are the odds of case for individuals having a specific allele and the odds of case for individuals who do not have that same allele.**

- **As an example, suppose that there are two alleles, T and C.**
- **The number of individuals in the case group having allele T is represented by 'A' and the number of individuals in the control group having allele T is represented by 'B’.**
- **Similarly, the number of individuals in the case group having allele C is represented by 'X' and the number of individuals in the control group having allele C is represented by 'Y’.**
- **In this case the odds ratio for allele T is A:B (meaning 'A to B', in standard odds terminology) divided by X:Y, which in mathematical notation is simply (A/B)/(X/Y).**

- **When the allele frequency in the case group is much higher than in the control group, the odds ratio is higher than 1, and vice versa for lower allele frequency.**
- **Additionally, a P-value for the significance of the odds ratio is calculated using a simple chi-squared test.**
- **Finding odds ratios that are significantly different from 1 is the objective of the GWA study because this shows that a SNP is associated with disease.**
- **Because so many variants are tested, it is standard practice to require the p-value to be lower than 5×10^{-8} to consider a variant significant.**

- There are several variations to this case-control approach.
- A common alternative to case-control GWA studies is the analysis of quantitative phenotypic data, e.g. height or biomarker concentrations or even gene expression.
- Likewise, alternative statistics designed for dominance or recessive penetrance patterns can be used.
- Calculations are typically done using bioinformatics software such as SNPTEST and PLINK, which also include support for many of these alternative statistics.

GWAS

focuses on the effect of individual SNPs.

- **However, it is also possible that complex interactions among two or more SNPs, epistasis, might contribute to complex diseases.**
- **Due to the potentially exponential number of interactions, detecting statistically significant interactions in GWAS data is both computationally and statistically challenging.**
- **This task has been tackled in existing publications that use algorithms inspired from data mining.**
- **Moreover, the researchers try to integrate GWA data with other biological data such as protein-protein interaction network to extract more informative results**

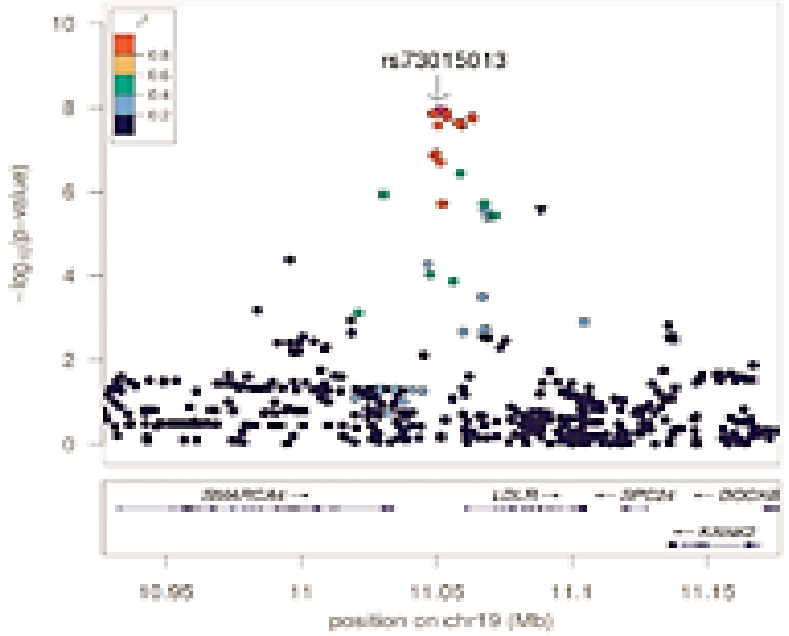
A key step in the majority of GWA studies is the imputation of genotypes at SNPs not on the genotype chip used in the study.

- This process greatly increases the number of SNPs that can be tested for association, increases the power of the study, and facilitates meta-analysis of GWAS across distinct cohorts.
- Genotype imputation is carried out by statistical methods that combine the GWAS data together with a reference panel of haplotypes.
- These methods take advantage of sharing of haplotypes between individuals over short stretches of sequence to impute alleles.
- Existing software packages for genotype imputation include IMPUTE2, Minimac, Beagle and MaCH

- In addition to the calculation of association, it is common to take into account any variables that could potentially confound the results. Sex and age are common examples of confounding variables.
- Moreover, it is also known that many genetic variations are associated with the geographical and historical populations in which the mutations first arose.
- Because of this association, studies must take account of the geographic and ethnic background of participants by controlling for what is called population stratification.
- If they fail to do so, these studies can produce false positive results

- After odds ratios and P-values have been calculated for all SNPs, a common approach is to create a Manhattan plot.
- In the context of GWA studies, this plot shows the negative logarithm of the P-value as a function of genomic location.
- Thus the SNPs with the most significant association stand out on the plot, usually as stacks of points because of haploblock structure.
- Importantly, the P-value threshold for significance is corrected for multiple testing issues.
- The exact threshold varies by study, but the conventional threshold is 5×10^{-8} to be significant in the face of hundreds of thousands to millions of tested SNPs.
- GWA studies typically perform the first analysis in a discoveryю

Results



- Regional association plot, showing individual SNPs in the LDL receptor region and their association to LDL-cholesterol levels.
- This type of plot is similar to the Manhattan plot in the lead section, but for a more limited section of the genome.
- The haploblock structure is visualized with colour scale and the association level is given by the left Y-axis.
- The dot representing the rs73015013 SNP (in the top-middle) has a high Y-axis location because this SNP explains some of the variation in LDL-cholesterol

Conclusion

1. Genome-wide association studies (GWAS) are powerful approaches for identifying genetic variants associated with complex quantitative traits.
2. GWAS uses genome-wide molecular markers, particularly SNPs, to investigate relationships between genetic variation and phenotypic diversity.
3. Accurate phenotyping, high-quality genotyping, and appropriate population selection are essential for reliable association analysis.
4. Statistical models accounting for population structure, kinship, and multiple testing help reduce false-positive associations.
5. Manhattan plots and Q–Q plots facilitate visualization and interpretation of GWAS results.
6. Integration of GWAS with transcriptomics, functional genomics, and genome editing supports candidate gene prioritization and validation.
7. GWAS contributes to gene discovery, marker-assisted breeding, and the genetic improvement of economically important organisms.

Key message: GWAS connects genome-wide genetic variation with complex phenotypes, providing a foundation for candidate gene discovery and biotechnology applications.

Control Question

- **s**

1. What is GWAS, and what are its main objectives?
2. What is the role of SNP markers in genome-wide association studies?
3. What is linkage disequilibrium (LD), and why is it important in GWAS?
4. How does GWAS differ from traditional QTL mapping?
5. What are the main stages of GWAS experimental design?
6. Why are population structure and kinship important in association analysis?
7. What statistical models are commonly used in GWAS?
8. What information is presented in Manhattan and Q–Q plots?
9. Why is multiple-testing correction necessary in GWAS?
10. How are candidate genes identified and functionally validated?
11. What are the main applications of GWAS in plant, medical, animal, and microbial biotechnology?
12. How can GWAS contribute to improving drought tolerance, disease resistance, and grain quality in wheat?

Recommended Literature

- 1. Uffelmann, E., Huang, Q. Q., Munung, N. S., et al. (2021). Genome-wide association studies. Nature Reviews Methods Primers, 1, 59. https://doi.org/10.1038/s43586-021-00056-9
- 2. Tam, V., Patel, N., Turcotte, M., et al. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics, 20, 467–484. https://doi.org/10.1038/s41576-019-0127-1
- 3. Kaler, A. S., & Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. BMC Genomics, 20, 618. https://doi.org/10.1186/s12864-019-5992-7
- 4. Yu, J., Pressoir, G., Briggs, W. H., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics, 38, 203–208. https://doi.org/10.1038/ng170