GWAS 教程

# GWAS 教程 ❤️😍

## Genome Wide Association Studies Definition

Genome-Wide Association Studies, abbreviated as GWAS, represent a scientific methodology aimed at elucidating(阐明) potential associations between genetic loci, namely the genome, and specific traits or diseases. Through meticulous analysis of millions, even tens of millions, of genetic variations within an organism’s genome, GWAS endeavors to identify genetic loci significantly associated with particular phenotypes or diseases.

## The main process of GWAS

### Data collection

DNA samples are systematically collected from individuals, along with recording their phenotypic information, which includes demographic data such as possible disease conditions, age, and gender.

For GWAS, typically large sample sizes are required to identify reproducible genome-wide significant association loci. Determining sample sizes can be facilitated(辅助) using software tools such as CATS or GPC. The phenotypes involved can be either binary traits or quantitative traits. Additionally, the study design can be population-based or family-based.

### Genetyping

Genotyping of each individual’s genes is carried out using existing GWAS arrays or sequencing strategies.

Microarray technology is commonly employed for genotyping **common variants**, or next-generation sequencing methods such as WES or WGS are utilized to cover **rare variants**. Given the high cost of WGS (whole genome sequencing), genotyping based on microarrays is currently the most commonly used method. However, WGS theoretically allows for the determination of nearly every genotype in the entire genome, and thus, with the ongoing development of low-cost WGS technologies, it is expected to become the mainstream method in the coming years.

## The journey from person to data file😎

### stage 1: **The Biological Source and Lab Technology**

1. **The Source Material: DNA**
   * The journey starts with collecting biological samples from thousands of individuals in a study cohort (e.g., 5,000 people with Type 2 Diabetes and 5,000 healthy controls).
   * The most common sources are **blood** or **saliva(唾沫)**.
   * In a lab, DNA is extracted from these samples. This high-purity DNA is the raw material for genotyping.
2. **The Technology: Genotyping Arrays (or “SNP Chips”)**
   * For a GWAS, we don’t typically sequence the entire genome for every person, as that would be too expensive and slow for large cohorts. Instead, we use a technology called a **genotyping array(基因分型技术)**.
   * Think of an array as a small glass slide or “chip” with a grid containing millions of microscopic probes.
   * Each probe is a short, synthetic piece of DNA designed to stick to a specific, known SNP location in the human genome.
   * When an individual’s DNA is washed over this chip, the DNA fragments bind to their matching probes. A laser scanner then reads which version of the SNP (which allele) is present at each location based on a fluorescent signal.
   * The output of this scanning process is a set of raw intensity files (e.g., .IDAT files from Illumina arrays). These files are then processed by specialized software (like Illumina’s GenomeStudio) to “call” the genotype for each person at each SNP.

This “genotype calling” step is what translates the biological signal into the digital data that the bioinformatician receives.

### The Core Data Types: What the Bioinformatician Actually Works With

As a bioinformatician, you will rarely work with the raw intensity files. Your starting point is the “called” genotype data, which is most commonly delivered in the **PLINK file format**. This is not one file, but a set of three files that work together, all sharing the same prefix (前缀)(e.g., my\_gwas\_data).

This three-file system efficiently separates the different kinds of information: the people, the genetic markers, and the actual genetic data.

#### 1. The .fam File: Sample and Phenotype Information

This is a plain-text, space-delimited(空格分隔) file that describes the **individuals** in your study. Each row represents one person. It has **six required columns**

|  |  |  |
| --- | --- | --- |
| **Column** | **Name** | **Description** |
| 1 | **Family ID (FID)** | An ID for the family. In many studies, this is the same as the Individual ID. |
| 2 | **Individual ID (IID)** | A **unique identifier** for each person. This is the most important key. |
| 3 | Paternal ID | The IID of the person’s father (or 0 if unknown). |
| 4 | Maternal ID | The IID of the person’s mother (or 0 if unknown). |
| 5 | **Sex** | 1 = Male, 2 = Female, 0 = Unknown. |
| 6 | **Phenotype** | The trait you are studying. -9 = Missing, 0 = Missing, 1 = Control, 2 = Case. For quantitative traits, this column would hold the measured value (e.g., 175.3 for height in cm). |

**Example .fam file:**

**codeCode**

FAM001 HG001 0 0 1 2  
FAM002 HG002 0 0 2 2  
FAM003 HG003 0 0 1 1  
FAM004 HG004 0 0 2 1

This file tells us we have 4 people. HG001 and HG002 are cases (phenotype=2), while HG003 and HG004 are controls (phenotype=1).

#### 2. The .bim File: Variant (SNP) Information

This is a plain-text file that contains the “map” information for your **genetic markers**. Each row represents one SNP.

|  |  |  |
| --- | --- | --- |
| **Column** | **Name** | **Description** |
| 1 | **Chromosome Code** | The chromosome the SNP is on (1-22, X, Y). |
| 2 | **SNP Identifier** | A unique name for the SNP, often an “rs” ID from the dbSNP database. |
| 3 | Genetic Distance (cM) | Position in centiMorgans (厘摩）(often just 0, as it’s not used in GWAS). |
| 4 | **Base-pair Position** | The physical position of the SNP on the chromosome (e.g., position 752566). |
| 5 | **Allele 1** | The first allele code (e.g., A, C, G, T). Usually the minor allele. |
| 6 | **Allele 2** | The second allele code (e.g., A, C, G, T). Usually the major allele. |

**Example**

inro

1 rs123456 0 752566 G A  
1 rs789101 0 853664 C T

This file tells us about two SNPs on chromosome 1 at specific positions and what their possible alleles are.

#### 3. The .bed File: The Binary Genotype Data

* This is the most important file, containing the actual **genotype data**, but it is **not human-readable**.
* It is a **binary** file, meaning it stores the data in a highly compressed format that computers can read very quickly. This is crucial because this file can be enormous (many gigabytes for a large study).
* It stores the genotype for every individual (from the .fam file) at every SNP (from the .bim file). The order of individuals and SNPs in this binary file exactly matches the order in the .fam and .bim files.
* For a given SNP with alleles G and A, a person can have one of three genotypes: G/G (homozygous(纯合子) for G), A/A (homozygous for A), or G/A (heterozygous(杂合子)). The .bed file stores this information efficiently for millions of SNPs and thousands of people.

These three files together form the complete dataset that is the starting point for your Quality Control pipeline.

### Where Does This Data Come From? (Data Sources)

As a student and future bioinformatician, you will acquire this data in two primary ways:

1. **Direct Collaboration:** You will work directly with a hospital, university, or research consortium that is conducting a study. They will perform the sample collection and genotyping and will provide you with the PLINK filesets to analyze.
2. **Public and Controlled-Access Repositories:** A huge amount of GWAS data is available for secondary analysis. This is critical for research, as it allows scientists to validate findings, combine datasets in meta-analyses, and explore new hypotheses without generating new data. Key repositories include:
   * **dbGaP (The database of Genotypes and Phenotypes):** The primary repository in the United States, run by the NIH. Researchers must apply for access to download datasets, explaining their research plan to a Data Access Committee.
   * **UK Biobank:** A landmark prospective cohort study with incredibly rich genetic and health data on 500,000 UK participants. It is a go-to resource for genetic research. Access is also controlled and requires an application.
   * **EGA (European Genome-phenome Archive):** The European equivalent of dbGaP.

## QC😘

You have just received your raw data files (raw\_data.bed, raw\_data.bim, raw\_data.fam) and you are sitting at a Linux command-line terminal. Your primary tool will be **PLINK**.

The key principle is that **the order of operations matters**. We perform some broad cleaning steps first, then refine our dataset for more sensitive checks. A typical workflow looks like this:

1. **Initial, lenient(宽松) filtering** of both SNPs and individuals to remove the most obvious garbage data.
2. **Filtering based on SNP properties** (MAF, HWE).
3. **Filtering based on individual properties** that require a cleaner set of SNPs to be calculated accurately (Heterozygosity, Relatedness, Ancestry).

### Step 1: Initial Missingness Filters (The First Pass)

Our first goal is to remove the “low-hanging fruit”—individuals and SNPs that failed genotyping so badly they are clearly unusable.

#### 1a. Remove Individuals with Low Call Rate

We start by removing individuals for whom a large fraction of SNPs could not be genotyped. This is often due to poor quality DNA.

* **Command:** *plink --bfile raw\_data --mind 0.05 --make-bed --out qc\_step1*
* **What this does:**
  + --bfile raw\_data: Specifies our input PLINK binary fileset.
  + --mind 0.05: The “–missingness per individual” filter. It removes any person with more than 5% missing genotypes.
  + --make-bed: Tells PLINK to create a new binary fileset as output.
  + --out qc\_step1: Sets the prefix for our new output files (qc\_step1.bed, qc\_step1.bim, qc\_step1.fam).

PLINK will print a report to the screen telling you how many individuals were removed.

#### 1b. Remove SNPs with Low Call Rate

Now, using our slightly cleaner set of individuals, we remove SNPs that failed to genotype across many people. These are unreliable markers.

* **Command:** *plink --bfile qc\_step1 --geno 0.02 --make-bed --out qc\_step2*
* **What this does:**
  + --bfile qc\_step1: Our input is the output from the previous step.
  + --geno 0.02: The “–missingness per **geno**type” (i.e., per SNP) filter. It removes any SNP with more than 2% missing calls.
  + --out qc\_step2: Creates our next set of files.

Again, PLINK will report how many variants (SNPs) were removed.

### Step 2: Core SNP Quality Filters

Now that we’ve removed the worst of the data, we apply two standard SNP-level filters.

#### 2a. Minor Allele Frequency (MAF) Filter（次要等位基因频率 (MAF) 过滤器）

We remove very rare SNPs because we lack the statistical power to test them properly, and they are more likely to be genotyping errors.

* **Command:** *plink --bfile qc\_step2 --maf 0.01 --make-bed --out qc\_step3*
* **What this does:**
  + --bfile qc\_step2: Takes the previous step’s output as input.
  + --maf 0.01: Removes any SNP where the frequency of the less common allele is below 1%.
  + --out qc\_step3: Creates our next set of files.

#### 2b. Hardy-Weinberg Equilibrium (HWE) Filter（哈代-温伯格平衡（HWE）过滤器）

We remove SNPs where the observed pattern of genotypes in the **control group** deviates(差异) significantly from what we’d expect by chance. This is a classic sign of genotyping error.

* **Command:** *plink --bfile qc\_step3 --hwe 1e-6 --make-bed --out qc\_step4*
* **What this does:**
  + --hwe 1e-6: Calculates the HWE p-value for each SNP in controls only by default. It then removes any SNP with a p-value less than our very stringent threshold of 0.000001. We use a tiny p-value because we are testing millions of SNPs and want to avoid removing SNPs that deviate by chance.
  + --out qc\_step4: Creates a fileset with only HWE-compliant SNPs.

### Step 3: Core Individual Quality Filters

At this point, we have a reasonably clean set of SNPs. We can now use this set to perform more sensitive checks on the individuals.

#### 3a. Check for Heterozygosity Outliers(杂合性异常值)

We look for individuals whose DNA might be contaminated (too heterozygous) or come from an inbred background (not heterozygous enough).

* **Command:** *plink --bfile qc\_step4 --het --out heterozygosity\_check*
* **What this does:**
  + --het: This doesn’t filter anything yet. Instead, it calculates the heterozygosity rate for each individual and writes the results to a file named heterozygosity\_check.het.
* **Your Action (Manual Step):**
  1. You will now use a script (in R or Python) to read the .het file.
  2. Calculate the mean and standard deviation (SD)(平均值和标准差) of the F column (which reflects heterozygosity).
  3. **Identify any individuals whose F value is more than 3 SDs above or below the mean.**
  4. Create a simple text file, let’s call it het\_outliers.txt, with two columns (Family ID and Individual ID) for these outlier individuals.

#### 3b. Remove Heterozygosity Outliers

Now you use the list you just created to remove the outliers.

* **Command:** *plink --bfile qc\_step4 --remove het\_outliers.txt --make-bed --out qc\_step5*
* **What this does:**
  + --remove het\_outliers.txt: Removes the individuals listed in your file.
  + --out qc\_step5: Creates the next version of our clean data.

### Step 4: Relatedness and Population Structure(相关性和种群结构)

This is the final and most computationally intensive QC phase. We want to ensure our sample is a set of unrelated individuals from the same broad ancestry. These checks work best on a set of SNPs that are not physically close to each other (i.e., not in “Linkage Disequilibrium”).

#### 4a. Prune SNPs for Independence (LD Pruning) 修剪 SNP 以实现独立性（LD 修剪）

* **Command:** *plink --bfile qc\_step5 --indep-pairwise 50 5 0.5 --out ld\_pruned*
* **What this does:**
  + --indep-pairwise 50 5 0.5: This is a command to create a list of largely independent SNPs. It scans the genome in windows of 50 SNPs, shifts the window by 5 SNPs at a time, and removes one of any pair of SNPs that has an r² > 0.5 (a measure of correlation).
  + --out ld\_pruned: This creates two files: ld\_pruned.prune.in (the list of SNPs to KEEP) and ld\_pruned.prune.out.

#### 4b. Check for Cryptic Relatedness（隐秘关联）

We now use this pruned set of SNPs to check for hidden relatives in our data.

* **Command:** *plink --bfile qc\_step5 --extract ld\_pruned.prune.in --genome --out relatedness\_check*
* **What this does:**
  + --extract ld\_pruned.prune.in: Tells PLINK to perform the next step using only our independent set of SNPs.
  + --genome: This is the main command. It calculates the “identity-by-descent” (IBD)(血统同一性) for all pairs of individuals.
  + --out relatedness\_check: The results are saved in a file called relatedness\_check.genome.
* **Your Action (Manual Step):**
  1. Inspect the .genome file. Look at the PI\_HAT column. A value > 0.1875 indicates a 3rd degree relative or closer.
  2. **Identify all pairs that are too related**. For each pair, decide which individual to remove (a common strategy is to remove the one with the lower call rate).
  3. Create a text file, related\_individuals.txt, with the FID and IID of every person you need to remove.

#### 4c. Remove Related Individuals

* **Command:** plink --bfile qc\_step5 --remove related\_individuals.txt --make-bed --out qc\_step6

## PCA 😉

### Part 1: The “Why” - Understanding the Problem of Population Stratification（人口分层）

Before running any code, you must understand why this step is so critical.

**Population stratification** is the presence of systematic differences in allele frequencies between different subpopulations in your study. If your “case” group (people with the disease) and “control” group (healthy people) have a different mixture of these subpopulations, you can get massive numbers of false-positive results.

**The Classic Analogy(类比): The “Chopsticks Gene”(筷子基因)**

Imagine a hypothetical study looking for genes associated with the ability to use chopsticks.

* You recruit “cases” (people who can use chopsticks) primarily from an *East Asian population.*
* You recruit “controls” (people who cannot) primarily from a *European population.*
* You run a GWAS.

The result? You will find thousands of SNPs that are “significantly associated” with your trait. However, ***these are not “chopstick-skill genes.” They are simply genetic markers that have different frequencies between East Asian and European populations. Your analysis has been completely confounded by ancestry. The association is with ancestry, not the trait itself(关联在于血统，而非特征本身).***

**The Goal of PCA:** PCA is our mathematical tool to detect this hidden ancestral structure in our data so we can either remove outlier individuals or statistically correct for the variation.

### Part 2: The “What” - A Conceptual Understanding of PCA

In a GWAS, we have data for hundreds of thousands or millions of SNPs. You can think of this as a dataset with a million dimensions—far too complex for a human to visualize.

**PCA is a dimensionality-reduction technique.** Its goal is to distill this vast complexity into a few, highly informative new dimensions called **Principal Components (PCs)**.

* **Principal Component 1 (PC1):** This is a new, calculated axis that captures the **single largest source of variation**(最大单一变异源)in your entire genetic dataset. In human genetics, this almost always corresponds to the major axis of global ancestry (e.g., separating individuals of African vs. non-African descent(对比非洲裔和不是非洲裔,概念较大)).
* **Principal Component 2 (PC2):** This is the second axis, which is mathematically uncorrelated with PC1, that captures the **largest amount of the remaining variationndefined（剩余变异最大量)**. This might, for example, separate European and East Asian individuals(区分亚洲人和中东人）.
* **PC3, PC4, etc.:** Each subsequent PC captures progressively smaller and smaller amounts of the remaining variation(剩余变化量越来越小).

By plotting individuals based on their values for PC1 and PC2, we can create a two-dimensional “genetic map” of our study participants, allowing us to visually identify their ancestral relationships.

### Part 3: The “How” - A Practical Step-by-Step Workflow

Here is how you perform PCA on your QC’d GWAS data. The input for this process should be the dataset that has already passed the initial QC steps (missingness, MAF, HWE, etc.). Let’s call our input fileset qc\_step6.

#### Step A: Prepare the Data by Pruning for Linkage Disequilibrium (LD)(修剪连锁不平衡（LD）来准备数据)

* **The Goal:** To get a set of SNPs that are roughly independent of one another.
* **The Reason:** PCA can be biased if it uses all SNPs. A region of the genome with many highly correlated SNPs (a high LD block) is essentially telling the same ancestral story many times. This would give that single region too much influence on the PCA results. By “pruning” the data—removing SNPs that are in high LD with each other—we ensure that SNPs across the whole genome contribute more evenly to the analysis.
* **The PLINK Command:**
* plink --bfile qc\_step6 --indep-pairwise 50 5 0.2 --out ld\_pruned
* **Command Breakdown:**
  + --bfile qc\_step6: Use our QC’d data as input.
  + --indep-pairwise 50 5 0.2: This is the pruning command.
    - 50: Scans the genome in windows of 50 SNPs.
    - 5: Shifts the window forward by 5 SNPs at a time.
    - 0.2: For any pair of SNPs in a window, if their correlation (r²) is greater than 0.2, one of them is removed from the list.
  + --out ld\_pruned: This doesn’t create a new dataset. It creates two files: ld\_pruned.prune.in (a list of SNPs to **keep**) and ld\_pruned.prune.out (a list of SNPs that were removed).

#### Step B: Run the PCA Calculation

* **The Goal:** To calculate the Principal Component values (or “scores”) for every individual in our study.
* **The Reason:** We use the pruned list of independent SNPs to perform the core calculation, as this gives a more robust and unbiased estimate of the underlying ancestry structure.
* **The PLINK Command:**
* plink --bfile qc\_step6 --extract ld\_pruned.prune.in --pca 10 --out gwas\_pca
* **Command Breakdown:**
  + --bfile qc\_step6: We use our full QC’d dataset. We want to calculate PCs for everyone.
  + --extract ld\_pruned.prune.in: This is the critical part. We tell PLINK to perform the calculation **using only the independent SNPs** from the list we just created.
  + --pca 10: This is the main command to run PCA and calculate the top 10 principal components.
  + --out gwas\_pca: Specifies the prefix for the output files.
* **The Key Output File:** This command produces a file named gwas\_pca.eigenvec. This is a plain text file where each row is an individual, and the columns contain their ID followed by their calculated values for PC1, PC2, PC3, etc.

### Part 4: The “What’s Next” - Visualization and Action

Now you have the numerical results. The next step is to visualize them to understand your cohort’s structure.

#### 4a. Visualization

You will use a plotting program like **R (with the ggplot2 library)** or **Python (with matplotlib/seaborn)** to create a scatter plot.

1. **Load the Data:** Read the gwas\_pca.eigenvec file into your R or Python environment.
2. **Create the Plot:**
   * Plot PC1 on the x-axis.
   * Plot PC2 on the y-axis.
   * Each point on the plot represents one individual from your study.

**Pro-Tip (as seen in the tutorial): Merging with a Reference Panel**  
To make your plot truly interpretable, it’s best practice to merge your dataset with a public reference panel containing individuals of known ancestry (like the 1000 Genomes Project) before running the PCA. When you plot the results, you can color the reference samples by their known population (e.g., European=blue, African=green, East Asian=red). Your own study samples will then cluster with the reference populations they are genetically closest to, immediately telling you the ancestral makeup of your cohort.

#### 4b. Interpretation and Action Plan

When you look at your plot, you will see one of a few scenarios:

* **Scenario 1: One Tight Cluster.** This is the ideal result. It means your study population is ancestrally homogeneous(同质的).
* **Scenario 2: A Main Cluster with Outliers.** This is very common. You’ll see a large cloud of points (your primary study population) and a few individual points scattered far away.
* **Scenario 3: Multiple Distinct Clusters.** This indicates your sample contains several different ancestry groups (e.g., a cluster of European ancestry and a separate cluster of East Asian ancestry).

Based on the plot, you take two crucial actions:

1. **Remove Ancestral Outliers(移除祖先离群值):** The individuals who fall far outside the main cluster(s) of your intended study population are considered ancestral outliers. Their inclusion could introduce confounding. You should create a list of their IDs and remove them from your dataset using plink –remove.
2. **Generate Covariates for the Association Analysis(为关联分析添加协变量）:** For the individuals that you keep, there is still subtle variation within the main cluster (it’s a cloud, not a single point). This remaining, continuous variation can still bias your results. Therefore, you **must statistically control for it.**
   * **Action:** You will take the values from the first 5 or 10 principal components (the columns in your gwas\_pca.eigenvec file) and include them as **covariates** in your final GWAS association model. When you run the association test, your model will look like this:
   * Disease\_Status ~ SNP + Age + Sex + \*\*PC1 + PC2 + PC3 + PC4 + PC5\*\*

## Association test and Visualization🤣

### Part 1: The Core Concept - The Statistical Model

You are not just looking for a simple difference. You are performing a **regression analysis** for each SNP. Regression is a statistical method that allows us to model the relationship between a set of variables.

Our model for each SNP looks something like this:

**Trait ~ SNP Genotype + Covariate(协变量) 1 + Covariate 2 + … + Covariate N**

Let’s translate this:

* **Trait:** This is your outcome variable (the “Y”). It could be disease status (case/control) or a continuous measurement (like blood pressure).
* **(tilde(波浪线))**: Means “is modeled by” or “is predicted by”.
* **SNP Genotype:** This is your main predictor of interest (the “X”). The genotype is typically coded numerically (e.g., 0, 1, or 2) based on the number of copies of a specific allele an individual has(基因型通常根据个体特定等位基因的拷贝数以数字形式编码). This is called an **additive model**, and it’s the standard for GWAS.
* **Covariates:** These are the other variables you include in the model to control for confounding effects. You must include covariates to get reliable results.

**The Crucial Role of Covariates:**  
Remember the principal components (PC1, PC2, etc.) we calculated during the PCA step? This is where they become absolutely essential. By including them in the model, **you are statistically controlling for any residual population structure. Your model can then distinguish between a true association with the trait and a false association driven by ancestry.**

Common covariates include:

* **Top 5-10 Principal Components** (to control for ancestry)(控制祖先)
* **Age** (if the trait is age-related)
* **Sex** (if the trait differs between sexes)
* Any other relevant variables from your study (e.g., smoking status, technical batch effects).

### Part 2: Choosing the Right Test

The specific type of regression model you use depends entirely on your phenotype.

* **For Binary Traits (Case vs. Control):** You will use ***Logistic Regression****.*
  + **The Question:** Does the SNP genotype predict the odds of an individual being a case versus a control?
  + **The Output:** The main result is an **Odds Ratio (OR)(比值比)**, which tells you how much the odds of having the disease increase (or decrease) for each copy of the test allele.
* **For Quantitative Traits (Height, BMI, Blood Pressure):** You will use ***Linear Regression****.*
  + **The Question:** Does the SNP genotype predict the value of the trait?
  + **The Output:** The main result is a **Beta coefficient (β)**, which tells you how much the trait value changes for each copy of the test allele (e.g., “each copy of the G allele is associated with a 0.5 cm increase in height”).

### Part 3: The “How” - Practical Implementation with PLINK

Let’s assume you are performing a case-control study for a disease. You will use logistic regression.

* **Input Files:**
  1. FINAL\_QC\_DATA.bed/.bim/.fam: Your fully cleaned dataset from the previous QC steps.
  2. gwas\_pca.eigenvec: The file containing the principal components for each individual. You will need to reformat this slightly to match PLINK’s covariate file format (typically a header row with FID, IID, PC1, PC2, etc.).
* **The PLINK Command:**
* plink --bfile FINAL\_QC\_DATA \  
   --logistic \  
   --covar gwas\_pca.eigenvec --covar-col-name PC1,PC2,PC3,PC4,PC5 \  
   --allow-no-sex \  
   --out my\_gwas\_results
* **Command Breakdown:**
  + --bfile FINAL\_QC\_DATA: Specifies your clean input dataset.
  + --logistic: **This is the key flag that tells PLINK to perform a logistic regression for a case-control trait.** If you had a quantitative trait, you would use –linear instead.
  + --covar gwas\_pca.eigenvec: Specifies your covariate file.
  + --covar-col-name PC1,PC2,PC3,PC4,PC5: Tells PLINK exactly which columns from the covariate file to use in the model. (Note: The exact flag might be –covar-name in newer PLINK versions. Always check the documentation).
  + --allow-no-sex: A practical flag that prevents the analysis from failing if some individuals have missing sex information.
  + --out my\_gwas\_results: Sets the prefix for the output files.

PLINK will now iterate through every SNP in your .bim file, run a logistic regression for that SNP against your phenotype (adjusting for the covariates you provided), and record the results.

### Part 4: Understanding the Output - The Summary Statistics File

The command will produce a file named my\_gwas\_results.assoc.logistic (or similar). This is your **summary statistics file**, and it is the primary result of your entire GWAS. It’s a large text file with one row for every SNP tested.

Here are the most important columns you will find inside:

|  |  |  |
| --- | --- | --- |
| **Column** | **Name** | **Description** |
| CHR | Chromosome | The chromosome the SNP is on. |
| SNP | SNP Identifier | The name of the SNP (e.g., rs123456). |
| BP | Base-pair Position | The physical position of the SNP on the chromosome. |
| A1 | Allele 1 (Effect Allele) | The allele for which the effect is reported. It’s crucial to know which allele is being tested. |
| OR | **Odds Ratio** | **The effect size.** OR > 1 means A1 is the risk allele. OR < 1 means A1 is the protective allele. OR = 1 means no effect. |
| STAT | Test Statistic | The Z-score or Wald statistic from the regression model. A larger absolute value means a stronger association. |
| P | **P-value** | **The statistical significance.** The probability of observing an association this strong or stronger by pure chance, assuming no real effect. |

### Part 5: Defining a “Hit” - The Genome-Wide Significance Threshold

You have millions of p-values. You cannot simply use the old standard of p < 0.05. If you ran a million tests, you would expect 50,000 “significant” results just by random chance! This is the problem of **multiple testing**.

To correct for this, the genomics community has established the **genome-wide significance threshold(基因组显著性阀值**.

* **The Threshold:** **p < 5 x 10⁻⁸** (or 0.00000005)

**Why this specific number?** It’s a sophisticated version of a Bonferroni correction. It accounts for the fact that we are performing roughly one million independent tests on the human genome (the number of “independent” blocks of SNPs in populations of European ancestry). A standard Bonferroni correction would be 0.05 / 1,000,000 = 5 x 10⁻⁸.

Any SNP in your my\_gwas\_results.assoc.logistic file with a p-value **below 5 x 10⁻⁸** i**s considered a “genome-wide significant hit” and is a strong candidate for being genuinely associated with your disease.**

Your next steps in the post-GWAS analysis will be to visualize these results (with Manhattan and Q-Q plots) and investigate the biological function of these significant “hits.”

### Visualization

### Part 1: The Manhattan Plot(曼哈顿图) - Visualizing Your Results

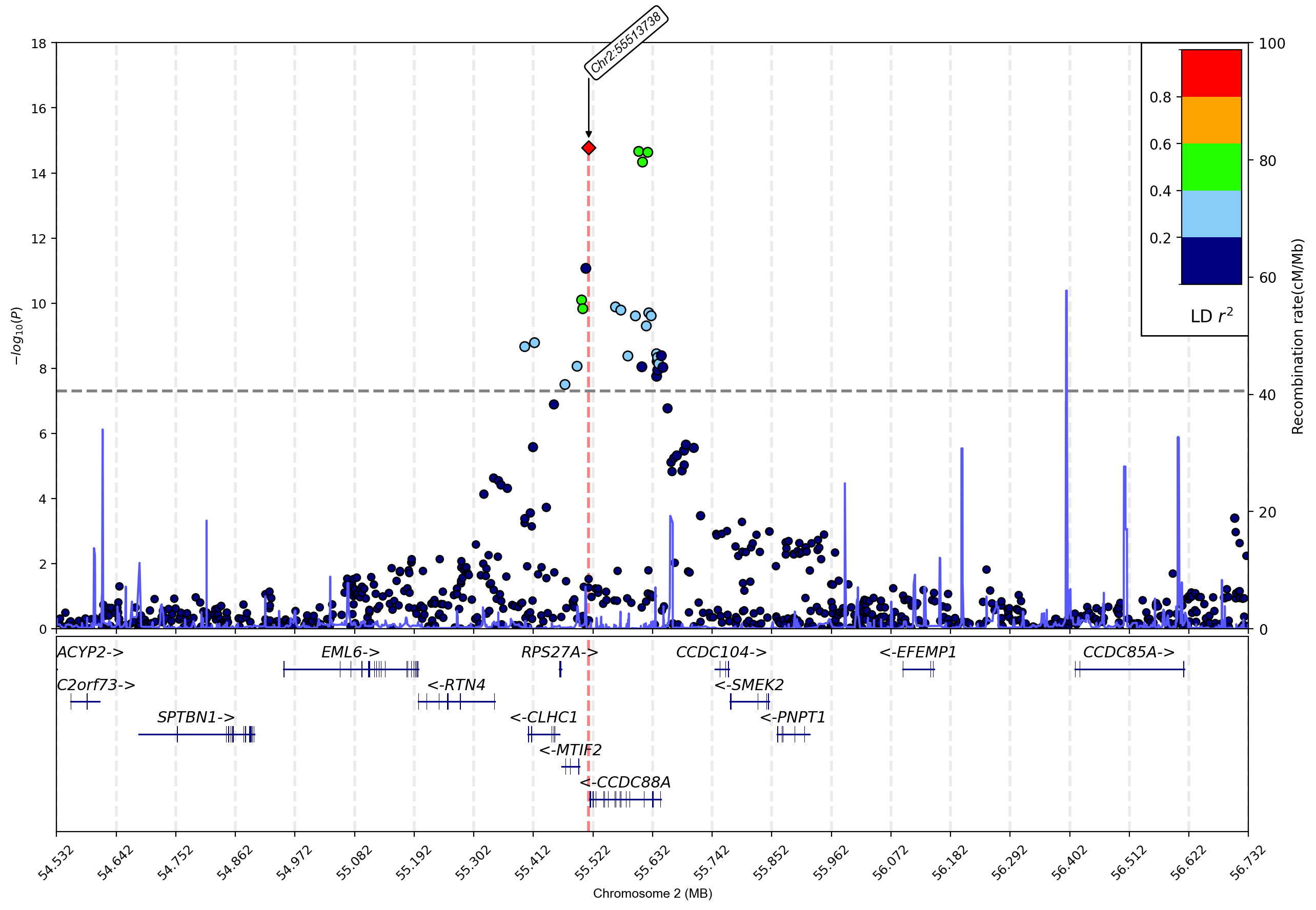
#### The Concept

A Manhattan plot is a scatter plot for all the SNPs in your GWAS.

By the way ,the reason why it was named the Manhattan Plot.why you see this picture



 It was a cloudy and misty day. Those birds formed a significance threshold line. And the skyscrapers above that line resembled the significant signals in your GWAS. I believe you could easily get how the GWAS Manhattan plot was named.



* **The X-axis:** The genomic coordinates. All the chromosomes are laid out end-to-end, from chromosome 1 to 22.
* **The Y-axis:** The level of statistical significance for each SNP. To make the most significant results stand out, we don’t plot the p-value directly. Instead, we plot the **-log₁₀(P-value)**.

**Why -log₁₀(P-value)?**  
This transformation is purely for visualization. A very small, highly significant p-value (e.g., 1x10⁻¹⁰) becomes a very large number (-log₁₀(10⁻¹⁰) = 10), making it appear high up on the plot. A non-significant p-value (e.g., 0.5) becomes a small number (-log₁₀(0.5) = 0.3). This stretches the top of the plot where the interesting results are.

The plot gets its name because the significant results look like “skyscrapers” rising from the city skyline of Manhattan.

#### What to Look For

1. **The Significance Line:** We draw a **horizontal line across the plot at the genome-wide significance threshold.** This line is at **-log₁₀(5x10⁻⁸) ≈ 7.3**.
2. **The “Towers” or “Skyscrapers”:** Any SNP that crosses this line is a **genome-wide significant hit**. You will typically see that it’s not just one SNP, but a whole cluster of nearby SNPs that are highly significant. This “tower” represents a single **genetic locus** associated with your trait. The SNP at the very peak of the tower is your “lead SNP” for that locus.

#### How to Create It (Using R and qqman)

First, you need to install and load the qqman package in R.

# install.packages("qqman")  
# Run this once if you don't have it   
library(qqman)   
library(dplyr) # Useful for data manipulation

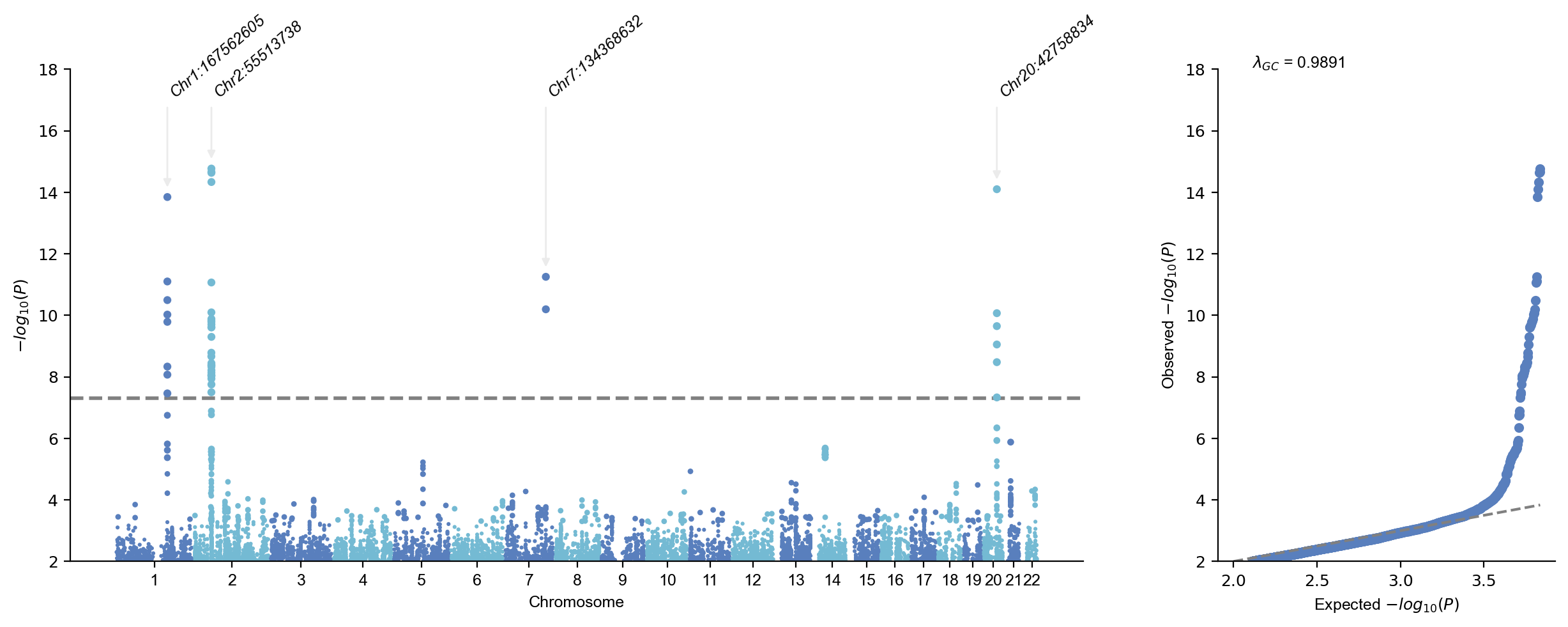
Next, you load your GWAS results file. The qqman package requires specific column names: SNP, CHR, BP, and P. Your PLINK output might have different names, so we often rename them.

# Read the PLINK logistic regression output file  
results <- read.table("my\_gwas\_results.assoc.logistic", header = TRUE)  
  
# The PLINK output columns are CHR, SNP, BP, P. qqman expects these names.  
# If your column names were different, you would rename them here.  
# For example: results\_renamed <- rename(results, SNP = snp\_name, P = p\_value)  
  
# Create the Manhattan plot  
manhattan(results)

This simple command will generate a classic Manhattan plot. You can customize it with titles and highlight specific SNPs.

### Part 2: The Q-Q Plot - A Critical Diagnostic Check

Before you get too excited about the hits on your Manhattan plot, you must perform a sanity check on your overall results using a Q-Q plot. This plot tells you if your statistical test was well-behaved or if it was biased by a hidden confounder.



#### The Concept

A Q-Q (Quantile-Quantile) plot compares the distribution of your observed p-values against the distribution of p-values you would expect to see by pure chance (the “null hypothesis”).

* **The X-axis:** The *expected -log₁₀(P-values) if no SNP was truly associated with the disease.*
* **The Y-axis:** Your *observed -log₁₀(P-values) from your actual data, sorted from smallest to largest.*

#### How to Interpret the Q-Q Plot (This is Crucial)

1. **The Diagonal Line(对角线) (y=x):** A straight diagonal line represents the null hypothesis. The vast majority of SNPs in your study are not associated with the disease, so their p-values should be randomly distributed and fall along this line.
2. **The Ideal “Good” Plot:** A healthy Q-Q plot shows points hugging(紧贴) the diagonal line for almost its entire length. At the very end (the top right), the points will curve upwards, departing from the line. This is the perfect picture: it shows that most of your million p-values are behaving as expected (no association), while a small number of SNPs are far more significant than expected by chance. **These are your true signals.**
3. **The “Bad” Plot (Genomic Inflation):** A major red flag is when the observed p-values **systematically deviate** from the diagonal line early on. The entire cloud of points will appear shifted upwards, off the diagonal. This is called **p-value inflation** and indicates that your p-values are, on average, smaller than they should be across the whole genome.
   * **The Cause:** The number one cause of this is uncorrected population stratification. **It means your PCA covariates were not sufficient to control for all the ancestry differences between your cases and controls.**
   * **The Fix:** If you see this, you cannot trust your results. You must go back to your association analysis and include more principal components as covariates (e.g., use the top 10 PCs instead of the top 5).

#### How to Create It (Using R and qqman)

The qqman package makes this incredibly simple.

# Assuming you have already loaded your 'results' table from the step above  
  
# Create the Q-Q plot  
qq(results$P)

This one command will generate the Q-Q plot and also calculate and display a value called **lambda (λ)**. This “genomic inflation factor” quantifies the deviation. A lambda value very close to 1.0 (e.g., 1.01) is perfect. A value > 1.05 is a strong sign of inflation that needs to be addressed.

## Variant Annotation Test 🤯

This is the process where we move from a simple statistical fact (rs123456 on chromosome 1 is significant) to a **potential biological story** (rs123456 disrupts a binding site for a key transcription factor, which may alter the expression of the nearby gene ABC, a known player in immune response).

The fundamental goal of annotation is to attach biological information to your list of significant SNPs to generate hypotheses about how they might influence the disease or trait.

### Part 1: The “Why” - Most GWAS Hits Are Not Obvious

A common misconception is that a GWAS will point directly to a “smoking gun(确凿证据)” mutation inside a gene that breaks a protein. The reality, revealed by thousands of GWAS, is that **over 90% of significant GWAS hits fall in non-coding regions of the genome** (introns or the vast spaces between genes).

This means they **don’t change the protein sequence directly**. Instead, they are much more likely to affect **gene regulation**—the complex system that controls when, where, and how much of a gene is turned on or off. Annotation is our key to deciphering these regulatory effects.

### Part 2: The Types of Annotation - What Information Are We Looking For?

We layer different kinds of information onto our SNPs. Here are the key questions we ask, from most to least direct in their functional implication.

#### 2.1. Location and Consequence on the Gene

This is the first and most basic level of annotation. We determine where the SNP is located relative to known genes.

* **Intergenic(基因间):** The SNP is in the space between genes.
* **Intronic(内含子):** The SNP is inside a gene but in an intron (a non-coding section that is spliced out before the protein is made).
* **Exonic(外显子):** The SNP is in an exon—a part of the gene that directly codes for the protein. This is a very interesting result and leads to a second, more critical question: what is its consequence?
  + **Synonymous(同义突变):** The SNP changes the DNA code, but the resulting amino acid in the **protein is the same**. Usually considered a low-impact change.
  + **Nonsynonymous (or Missense)(非同义或者错配):** This is a key finding. The SNP changes the amino acid sequence. This could **alter the protein’s structure, stability, or function**.
  + **Nonsense (Stop Gained)(无义):** A very high-impact result. The SNP introduces a **premature(过早) “stop” signal,** causing the protein to be truncated and likely non-functional.
  + **Frameshift(移码):** An insertion or deletion that alters the entire downstream reading frame of the protein. Also a very high-impact, often disease-causing, mutation.

#### 2.2. Regulatory Function

Since most **hits are non-coding**, this is often the most important category for GWAS. We check if our SNP falls within a region of the genome known to have a regulatory role.

* **Promoters/Enhancers(启动子或者增强子):** Does the SNP lie in a DNA region that acts as a switch to turn a nearby gene on or off?
* **Transcription Factor Binding Sites (TFBS)(转录因子结合位点):** Does the SNP alter a specific DNA sequence where a transcription factor (a protein that regulates gene expression) is known to bind? If the SNP changes the “landing pad,” the factor may not be able to bind, altering the gene’s regulation.
* **eQTL (expression Quantitative Trait Locus)(表达数量性状位点):** This is a powerful form of evidence. An eQTL is a SNP that is known to be associated with the expression level of a gene (often a nearby one). If your disease-associated SNP is also a known eQTL for gene ABC, it provides a direct, testable hypothesis: the SNP influences disease risk by changing the amount of ABC protein being made.

#### 2.3. Conservation, Population, and Clinical Data

* **Conservation Score:** Is this specific spot in the genome highly conserved (unchanged) across many different species (e.g., from humans to mice to fish)? High conservation suggests the region is functionally important.
* **Allele Frequencies:** How common is the risk allele in different global populations? (e.g., from the gnomAD database).
* **Clinical Significance:** Has this exact variant already been reported in a clinical setting and cataloged in a database like **ClinVar** as being pathogenic or benign?

### Part 3: The “How” - Tools of the Trade

You do not do this by hand. You use powerful, automated bioinformatics tools that have access to massive databases (like Ensembl, UCSC, ENCODE, and dbSNP).

The most widely used tool for this is the **Ensembl Variant Effect Predictor (VEP)**. Another very popular tool is **ANNOVAR**.

### The Need for Conversion: Why VCF?

First, why do we need to convert? Why can’t we just feed the annotation tool our PLINK files or our summary statistics file?

* **VCF is the Universal Standard:** The Variant Call Format (VCF) is the **gold-standard** format for storing and sharing genetic variant information. Almost every modern genomics tool (including the Ensembl VEP) is designed to read VCF files as its primary input.
* **VCF Contains Essential Allele Information:** A VCF file explicitly states which allele is the **Reference Allele** (the one found in the reference human genome) and which is the **Alternative Allele**. This REF/ALT information is absolutely critical for annotation tools to determine the functional consequence of a variant. PLINK’s format is more focused on the genotypes (A1/A2), which is perfect for association testing but less explicit for annotation.
* **Efficiency:** You do not want to annotate all 8 million SNPs that went into your GWAS. That would be computationally massive and unnecessary. You only want to annotate the handful of SNPs that are statistically significant.

Therefore, the workflow involves two steps:

1. Isolate your significant SNPs.
2. Convert just those SNPs from your final QC’d PLINK fileset into a VCF file.

### Practical Workflow: From PLINK BED to VCF

Let’s assume you have your final, clean PLINK fileset named FINAL\_QC\_DATA and your summary statistics file named my\_gwas\_results.assoc.logistic.

#### Step 1: Create a List of Significant SNPs

Your first task is to create a simple text file that contains only the IDs (rsIDs) of the SNPs that passed the genome-wide significance threshold.

You can do this using a command-line tool like awk or a simple script in R or Python. Here is the awk method, which is very fast:

# This command reads the results file, skips the header (NR>1),  
# checks if the p-value in the last column ($9 for PLINK1.9 logistic) is less than 5e-8,  
# and if it is, it prints the SNP ID from the second column ($2).  
# The output is saved to a new file.  
  
awk 'NR>1 && $9 < 5e-8 {print $2}' my\_gwas\_results.assoc.logistic > significant\_snps.txt

* **my\_gwas\_results.assoc.logistic**: Your input summary statistics file.
* **significant\_snps.txt**: Your output file. This file will be a simple, one-column list of rsIDs, like this:
* codeCode
* rs123456  
  rs789012  
  rs333444  
  ...

#### Step 2: Use PLINK to Extract and Convert the SNPs

Now that you have your “shopping list” of SNPs, you can use PLINK to go back to your high-quality genotype data, pull out just those SNPs, and write them in VCF format.

* **The Tool:** PLINK (version 1.9 or, even better, version 2.0).
* **The Command:**
* plink --bfile FINAL\_QC\_DATA \  
   --extract significant\_snps.txt \  
   --recode vcf \  
   --out my\_gwas\_hits
* **Command Breakdown:**
  + --bfile FINAL\_QC\_DATA: Specifies your final, clean PLINK binary fileset as the source of the genetic data.
  + --extract significant\_snps.txt: This is the key filter. It tells PLINK to only operate on the SNPs listed in this file.
  + --recode vcf: This is the main action command. It tells PLINK to write the output in VCF format.
  + --out my\_gwas\_hits: Specifies the prefix for the output file.

#### Step 3: The Output - Your VCF File

This command will produce the file you need for annotation: ***my\_gwas\_hits.vcf****.*

If you open this file in a text editor, you will see it has two parts:

1. **The Header:** Many lines starting with ## that contain metadata about the file, the reference genome, and definitions for the columns.
2. **The Data:** A single header line starting with #CHROM followed by the data for each of your significant SNPs, one per line.

The crucial columns that VEP will use are:

* *#CHROM*: Chromosome (e.g., 1)
* *POS*: Position (e.g., 752566)
* *ID*: Your SNP Identifier (e.g., rs123456)
* *REF*: The Reference Allele (e.g., A)
* *ALT*: The Alternative Allele (e.g., G)

This my\_gwas\_hits.vcf file is now perfectly formatted to be used as the input for the Variant Effect Predictor (VEP) or any other standard annotation tool.

**An Important Caveat (A Real-World “Gotcha”):**  
PLINK 1.9 sometimes has trouble knowing which allele is REF and which is ALT because the .bim file format doesn’t inherently store that information relative to a reference genome. When it creates the VCF, it might make a guess. Modern tools like **PLINK 2.0** and its .pgen format handle this much more robustly. For precise work, it’s always good practice to double-check that the REF allele in your VCF file matches the actual reference genome (like GRCh38), but for a standard workflow, the PLINK conversion is the accepted and correct first step.

**Workflow using VEP:**

**Step 1: Prepare Your Input File**

First, you need to isolate your top hits. You will filter your summary statistics file to get a list of independent, genome-wide significant SNPs. For each SNP, you need the following information: Chromosome, Position, ID (rsID), Reference Allele, and Alternative Allele. This is often formatted into a standard file type called a **VCF (Variant Call Format)** file.

**Step 2: Run the Annotation Tool**

You can use the VEP web interface for a small number of SNPs, but for a real project, you use the command-line version.

A typical VEP command might look like this:

vep --input\_file my\_gwas\_hits.vcf \  
 --output\_file annotated\_hits.tsv \  
 --cache \  
 --force\_overwrite \  
 --tab \  
 --sift b \  
 --polyphen b \  
 --symbol \  
 --numbers \  
 --regulatory

* This command takes your VCF file (my\_gwas\_hits.vcf) as input.
* It uses a local cache of data (--cache) for speed.
* It outputs a tab-separated file (--tab).
* Crucially, it adds information from various sources:
  + --sift b and --polyphen b: These are two tools that predict how damaging a missense mutation is likely to be to the protein.
  + --symbol: Adds the official gene symbol.
  + --regulatory: This is a key flag! It overlays information about known regulatory elements from the Ensembl Regulatory Build.

**Step 3: Interpret the Output**

The output (annotated\_hits.tsv) will be a rich table with many columns. Your job is to be a detective and sift through this information to find the most compelling biological story for each of your top hits.

You would create a summary table for your paper or report, highlighting the key findings for each significant locus:

* **Lead SNP:** rs123456
* **Gene:** IRF5 (or “intergenic, nearest gene is IRF5”)
* **Consequence:** Intronic
* **Annotation:** “Overlaps with an enhancer element active in T-cells. Is a known eQTL for IRF5.”
* **Hypothesis:** This variant may alter enhancer activity, leading to misregulation of IRF5 expression in T-cells, contributing to disease risk.

## ❤️Advanced Process 🤠

### Part 1: Fine-Mapping - Pinpointing the Causal Variant(精确定位因果变量)

#### Section 1.1: The Core Problem Revisited - Why We Need Fine-Mapping

When your GWAS identifies a significant locus, your Manhattan plot shows a “tower” of SNPs. This happens because of **Linkage Disequilibrium (LD)(连锁不平衡)**.

Imagine a segment of a chromosome is like a “block” of text that is almost always inherited as a single unit without being broken up by recombination.

* Let’s say a single SNP within this block—we’ll call it the **Causal Variant (CV)**—is the one that actually has a biological effect on the disease. For instance, it might disrupt a transcription factor binding site.
* Now, imagine there are 100 other “passenger” SNPs in this same inherited block. These SNPs have no biological function themselves, but because they are always “hitchhiking” with the CV, they will show almost the same statistical association with the disease.

The GWAS result (the p-value) cannot tell these SNPs apart. It’s like seeing a car speeding down the highway; you know the car is moving fast, but you don’t know who is actually pressing the accelerator. The lead SNP from your GWAS (the one with the lowest p-value) is simply the most visible passenger in the car, but it might not be the driver.

**The goal of fine-mapping is to computationally identify the most likely driver.**

#### Section 1.2: The Inputs - What You Need to Get Started

To perform statistical fine-mapping, you need two critical pieces of information for the specific genomic region you want to analyze (e.g., Chromosome 8, positions 12,000,000 to 12,500,000).

1. **Your GWAS Summary Statistics for the Locus:**
   * This is a subset of your main GWAS results file. For every SNP in the region, you need:
     + SNP ID (rsID)
     + Effect allele
     + **Effect size (Beta or Odds Ratio)**
     + **Standard Error (SE) of the effect size** (This is crucial as it represents the uncertainty of the estimate)
   * The Z-score (Beta / SE) is often used as the primary input statistic.
2. **A High-Quality LD Matrix:**
   * This is a large table (a square matrix) that describes the correlation (r²) between every pair of SNPs in your chosen locus.
   * **Crucially, this LD information CANNOT be calculated from your own study cohort.** This is because the correlations in your cohort are distorted by the case-control status (a phenomenon called “non-random sampling”).
   * You **must** use an external **reference panel** of individuals whose ancestry closely matches your study population. Common choices include:
     + **The 1000 Genomes Project:** A good, publicly available resource with diverse populations.
     + **UK Biobank:** An excellent, high-quality reference for European ancestry populations.
   * Tools like ldstore or PLINK can be used to calculate this matrix from the reference panel’s raw genotype data.

#### Section 1.3: The Method - How the Algorithms “Think”

The most popular and powerful fine-mapping methods today are based on a **Bayesian statistical framework**. Let’s conceptually walk through how a tool like **FINEMAP** or **SuSiE** works.

**The Underlying Assumption:** For most GWAS loci, we assume there is likely only **one** (or maybe a small number) of true causal variants within the region.

**The Process:**

1. **Hypothesis Generation:** The algorithm systematically generates thousands of possible hypotheses. For example:
   * Hypothesis 1: rs123 is the causal variant, and all other signals are just due to LD with it.
   * Hypothesis 2: rs456 is the causal variant, and all other signals are just due to LD with it.
   * …and so on for every SNP in the region.
   * More advanced models (like SuSiE) can also test hypotheses with multiple causal variants (e.g., “rs123 and rs789 are both causal”).
2. **Likelihood Calculation:** For each hypothesis, the algorithm calculates a likelihood. It asks: “Given the LD matrix, if Hypothesis 1 were true, what pattern of Z-scores would we expect to see across all the SNPs? And how closely does that expected pattern match the actual Z-scores from our GWAS summary statistics?” A hypothesis that provides a better “fit” to the observed data is given a higher likelihood.
3. **Posterior Probability Calculation(后验概率):** Using Bayesian principles, the algorithm converts these likelihoods into **posterior probabilities**. The **Posterior Inclusion Probability (PIP)** for a given SNP is the sum of the probabilities of all the hypotheses in which that SNP was considered causal. The PIP represents the model’s confidence—on a scale from 0 to 1—that a specific SNP is the causal variant.

#### Section 1.4: The Output - The “Credible Set”

The final, actionable output of the fine-mapping process is the **credible set**.

* **How it’s created:** The software lists all the SNPs in the region, ranked by their PIP. It then starts from the top and sums the PIPs until the total reaches a certain threshold, typically 95%.
* **What it means:** The 95% credible set is the smallest group of SNPs that together have a 95% probability of containing the true causal variant.
* **The Impact:** This is a massive leap forward from the original GWAS result. You might start with a locus containing 300 SNPs that are all statistically significant. A successful fine-mapping analysis might produce a 95% credible set containing only **four** SNPs.

**Example Outcome:**

* **Before Fine-Mapping:** “The locus at 8q24 is associated with the disease (lead SNP rs6983267, P = 1x10⁻¹⁵).”
* **After Fine-Mapping:** “The 95% credible set for the 8q24 locus contains four variants: rs6983267 (PIP = 0.52), rs7000448 (PIP = 0.21), rs10808556 (PIP = 0.15), and rs10505477 (PIP = 0.08).”

This tells experimental biologists that instead of studying hundreds of variants, they should focus their efforts on these four specific candidates, dramatically increasing the efficiency of their research. The next step would be to annotate just these four SNPs to see if one has a particularly compelling biological function.

## Polygenic Risk Scores(PRS)(多基因风险评分)🙄

### Part 2: Polygenic Risk Scores (PRS)

#### Section 2.1: The Core Concept - The Polygenic Architecture of Disease

For decades, genetic research focused on finding ***single genes*** with large effects (e.g., the BRCA1 gene for breast cancer or the CFTR gene for cystic fibrosis). These are called **monogenic(单基因)** diseases.

However, GWAS has definitively shown us that most common, complex diseases and traits (like heart disease, type 2 diabetes, schizophrenia(精神分裂症), height, and blood pressure) are **polygenic**. This means:

* They are influenced by **thousands to millions** of genetic variants across the entire genome.
* The effect of each individual variant is **minuscule(非常小)**. An odds ratio of 1.05 (a 5% increase in odds) is typical for a single SNP.
* No single variant is necessary or sufficient to cause the disease. Instead, it’s the **cumulative burden(累积负担)** of many small-effect risk alleles that determines an individual’s genetic predisposition(易感性).

The goal of a Polygenic Risk Score is to capture and quantify this cumulative genetic burden. It aggregates all these tiny effects into a single, easy-to-understand number that represents an individual’s overall genetic liability.

#### Section 2.2: The “How To” - Calculating a PRS

Imagine you want to calculate a PRS for a new individual, let’s call her Jane. You need two key ingredients:

1. **A “Base” Dataset (The Blueprint):** This is a large GWAS summary statistics file from a study of the disease you’re interested in. This file provides the “weights” for the calculation. For every SNP, it gives you:
   * The SNP ID (e.g., rs123)
   * The risk allele (e.g., ‘G’)
   * The **effect size (β)** of that risk allele (the weight).
2. **A “Target” Dataset (The Individual):** This is Jane’s own genotype data. For every SNP in the base dataset, we need to know how many copies of the risk allele Jane has (0, 1, or 2).

**The Calculation:** The PRS is a simple weighted sum.

PRS\_Jane = (β₁ × Jane’s\_Allele\_Count₁) + (β₂ × Jane’s\_Allele\_Count₂) + … + (βₙ × Jane’s\_Allele\_Countₙ)

Let’s do a tiny example with just three SNPs:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SNP** | **Risk Allele** | **Effect Size (β)** | **Jane’s Genotype** | **Jane’s Risk Allele Count** | **Contribution to PRS** |
| rs111 | A | 0.05 | G/G | 0 | 0.05 \* 0 = 0.0 |
| rs222 | T | 0.02 | T/C | 1 | 0.02 \* 1 = 0.02 |
| rs333 | G | 0.08 | G/G | 2 | 0.08 \* 2 = 0.16 |
| **Total PRS** |  |  |  |  | **0 + 0.02 + 0.16 = 0.18** |

Now, imagine doing this for 2 million SNPs. The final number is Jane’s raw PRS.

#### Section 2.3: A Crucial Challenge - Choosing the Right SNPs and Weights

The simple sum above is a bit too naive. If you just add up all the SNPs from a GWAS, you’ll get a poor result due to LD (counting the same signal multiple times) and noisy, non-zero effect sizes for non-associated SNPs. We need to refine the blueprint. This is where different PRS methods diverge.

**Method 1: Clumping and Thresholding (C+T)** - The Classic Approach  
This is a two-step process to select a smaller, more robust set of SNPs to include in the score.

1. **Clumping(聚类):** A greedy algorithm that iterates through the GWAS results, starting with the most significant SNP. It identifies all other nearby SNPs that are in high LD (e.g., r² > 0.1) with this lead SNP and removes them. It then moves to the next most significant SNP and repeats the process. The result is a set of largely independent SNPs.
2. **Thresholding:** After clumping, you apply a p-value threshold. You might create several PRS versions by including only SNPs with p-values below a certain cutoff (e.g., a PRS using SNPs with p < 5e-8, another with p < 1e-4, another with p < 0.1, etc.). You would then test which of these scores performs best at predicting the disease in an independent validation dataset.

**Method 2: Bayesian Methods (e.g., LDpred2, PRS-CS)** - ***The Modern Standard***  
These methods are more sophisticated and generally more powerful. They don’t just discard SNPs.

* They use an **LD reference panel** to understand how LD spreads the effect of a true causal SNP across its neighbors.
* They *use a statistical model (a “prior”)* that assumes most SNPs have a tiny effect, but some have a slightly larger one.
* The ***algorithm then re-calculates (or “shrinks”) the effect size (β) for every SNP in the genome. It adjusts the original GWAS beta to account for the LD structure.***
* The final PRS is then calculated using all SNPs with these new, more realistic, LD-adjusted weights. This approach has been shown to be more accurate and predictive than C+T.

#### Section 2.4: The Application and Interpretation - From Score to Stratification

The raw PRS number (e.g., 0.18) is meaningless on its own. Its power comes from seeing where an individual falls relative to a population.

1. **Standardization:** You calculate the PRS for thousands of individuals in a reference population. This creates a distribution of scores, which is typically a bell curve (a normal distribution). The raw scores are then standardized (converted to a Z-score) so we can talk about percentiles.
2. **Risk Stratification:** Now we can interpret Jane’s score. If her standardized score places her at the **95th percentile**, it means her calculated genetic risk is higher than 95% of the population.
3. **Clinical Utility:** This is where the real potential lies. Studies have shown that for diseases like coronary artery disease(心冠), individuals in the top 5% of the PRS distribution can have a **3 to 4 times higher lifetime risk** compared to those with an average score.

**Crucially, a PRS is not deterministic.** A high PRS for heart disease does not mean you will get it. It means your baseline genetic risk is high, making it even more important to manage modifiable risk factors like diet, exercise, and cholesterol. The power of a PRS is in identifying these high-risk individuals early so they can be targeted for more intensive screening and preventative care.

## MR🤷‍♀️

### Part 3: Mendelian Randomization (MR)

#### Section 3.1: The Core Problem - Correlation is Not Causation

This is one of the oldest and most difficult challenges in medical research. We frequently observe that a specific risk factor (an “exposure”) is associated with a disease (an “outcome”).

**Classic Example:**

* Observational studies show that people with **higher levels of HDL cholesterol(高密度脂蛋白胆固醇)** (“good” cholesterol) have a **lower risk of heart attack**
* **The obvious Conclusion:** Low HDL cholesterol causes heart attacks.
* **The Drug Development:** This led pharmaceutical companies to spend billions of dollars developing drugs that raise HDL cholesterol levels.
* **The Surprising Result:** The drugs worked—they successfully raised HDL levels. However, they had **zero effect** on reducing heart attacks.

**What went wrong?** The initial observation was a classic case of confounding. The same healthy lifestyle choices (good diet, exercise) that lead to high HDL also protect against heart attacks through other mechanisms. The HDL level itself wasn’t the causal factor; it was just a correlated marker of a healthy lifestyle.

We needed a way to test for causality that wasn’t biased by these lifestyle and environmental confounders. Mendelian Randomization is that method.

#### Section 3.2: The Concept - Nature’s Randomized Controlled Trial

The central idea of MR is to leverage the random process of genetic inheritance. When you are conceived(怀孕), the combination of genes you inherit from your parents is essentially random. This is **Mendel’s Law of Independent Assortment(孟德尔的独立分配定律)**.

This randomization acts like a perfect, lifelong **Randomized Controlled Trial (RCT)**:

* Some people, by pure chance, inherit a set of genetic variants that predispose them to having slightly higher cholesterol levels throughout their lives (the “treatment group”).
* Other people, by pure chance, inherit variants that predispose them to lower cholesterol levels (the “control group”).

Because these genes were assigned randomly at conception, they should not be correlated(混淆) with other lifestyle factors (like diet, smoking, or socioeconomic status) that could confound the relationship between cholesterol and heart disease.

Therefore, if the “treatment group” (genetically higher cholesterol) consistently shows a higher rate of heart attacks than the “control group” (genetically lower cholesterol), we can infer that the higher cholesterol is likely **causally** related to the heart attacks.

#### Section 3.3: The Method - The Instrumental Variable Framework

To perform an MR analysis, we need a **genetic instrument**. This is a SNP (or, more powerfully, a set of SNPs) that serves as a reliable and clean proxy for the exposure we want to test.

This genetic instrument must satisfy three core assumptions:

1. **The Relevance Assumption:** The SNP must ***be strongly and robustly associated with the exposure.*** (e.g., we use SNPs from a GWAS of cholesterol levels that are genome-wide significant).
2. **The Independence Assumption:** The SNP must not be associated with any confounders of the exposure-outcome relationship. (This is generally assumed to be true because of the randomization of genetic inheritance).
3. **The Exclusion Restriction Assumption:** The SNP can only affect the outcome through the exposure. It cannot have an independent, alternative biological pathway to the outcome.

**Violation of Assumption #3** is the biggest potential pitfall(陷阱) of MR. This is called **horizontal pleiotropy(水平基因多效性)** (from pleio, meaning “many effects”). For example, *if a SNP that raises cholesterol also happens to increase blood pressure through a completely separate biological mechanism, then we can no longer be sure if the effect on heart disease is coming from the cholesterol or the blood pressure pathway. There are many advanced statistical tests to check for and correct for potential pleiotropy.*

#### Section 3.4: The Practical Workflow - Two-Sample MR

The real power of MR today comes from the **Two-Sample MR** design. You don’t need one giant dataset with genetic, exposure, and outcome data for all individuals. Instead, you can use the summary statistics from two completely separate, publicly available GWAS.

**The Workflow:**

1. **Define your Question:** Do genetically-predicted **higher levels of LDL cholesterol** cause an increased risk of **Coronary Artery Disease (CAD)**?
2. **Gather Your Data:**
   * **Exposure GWAS:** Download the full GWAS summary statistics for LDL cholesterol (e.g., from the Global Lipids Genetics Consortium).
   * **Outcome GWAS:** Download the full GWAS summary statistics for CAD (e.g., from the CARDIoGRAMplusC4D consortium).
3. **Select Your Instruments:** From the LDL GWAS, select a set of independent, genome-wide significant SNPs. These are your instrumental variables for genetically-predicted LDL.
4. **Extract the Effects:** For each of those LDL-associated SNPs, look up its effect size (beta) and standard error in the CAD GWAS summary statistics file.
5. **Run the Analysis:** Use specialized R packages (like TwoSampleMR) to perform the statistical test.
   * The primary method is **Inverse Variance Weighted (IVW)(逆方差加权) MR**. This is essentially a regression where the slope of the line represents the causal estimate. For each instrument SNP, it plots its effect on the exposure (LDL) on the x-axis and its effect on the outcome (CAD) on the y-axis.
   * If the SNPs that have a large effect on increasing LDL also have a proportionally large effect on increasing CAD risk, the slope will be steep and statistically significant, providing strong evidence for a causal effect.
6. **Perform Sensitivity Analyses:** Run a battery of other MR methods (e.g., MR-Egger, Weighted Median) that are more robust to violations of the pleiotropy assumption. If all the methods point in the same direction, your causal inference is much stronger.