

Aki Asikainen

GENOMIC DATA QUALITY CONTROL IN PHARMACOGENETIC TESTING

Software implementation of data validation tools

ABSTRACT

Aki Asikainen: Genomic data quality control in pharmacogenetic testing; Software implementation of data validation tools

Master's thesis

Tampere University

Master's Degree Programme in Biotechnology and Biomedical Engineering

Supervisors: Professor Matti Nykter, Juho Heliste

Examiners: Professor Matti Nykter, Juho Heliste

2026 January

Pharmacogenetic testing relies on accurate genotype and phenotype data to guide drug selection and dosing. However, clinical labs often encounter samples whose ethnic composition can vary greatly from lab to lab, with complex multi-ethnic backgrounds and genotypes that challenge general-purpose tools. To ensure the quality of incoming data through robust biological methods, a Python-based tool was developed to implement systematic quality control for pharmacogenetic datasets. This prototype tool integrates Hardy-Weinberg equilibrium (HWE) testing, reference frequency comparison, and phenotype-level checks, and is intended for later integration into the Abomics' portal.

Rather than just process structured genotype, haplotype, and phenotype data, the tool supports both biallelic and multi-allelic loci, performing HWE tests on variants using biallelic recoding and on complex haplotypes using allele-wise collapsing. This approach can yield more stable and interpretable trends for quality control by combining rare genotypes, resulting in a clearer overall picture. The tool also supports allele frequency comparisons with external reference databases such as Ensembl and ClinPGx, as well as with internal clinical datasets. In addition, phenotype-level testing adds a new layer of quality control by grouping individuals into broader, more meaningful categories.

To evaluate performance in practice, the tool was first applied to selected tri-allelic SNPs from the 1000 Genomes chromosome 22 data. It identified both typical equilibrium and expected deviations, including missing genotype classes and signals related to population substructure. Analysis using information from the 1000 Genomes dataset construction suggested that strong geographic differentiation among cohorts largely explained these patterns. In anonymized clinical data, the tool detected strong deviations for both major and rare CYP1A2 alleles. Rare alleles showed substantial deviation due to low carrier counts, making them unsuitable for HWE-based validation. For major alleles, comparison of CYP1A2 allele frequencies across ancestries supported the Wahlund effect from population structure as a plausible explanation, rather than genotyping errors. Comparison of CYP2C19 allele frequencies to European and East Asian reference data indicated that the cohort was predominantly of European ancestry. These results revealed that very rare alleles produced statistically significant but minor differences, highlighting the need for minimum count thresholds. Phenotype frequency checks in a random patient subset matched reference distributions for most CYP genes, with some observed differences reflecting expected cohort composition.

However, effective use still requires careful tuning and selection of suitable QC alleles and tests, as well as consideration of whether data from different labs or sample sets with distinct ancestry backgrounds should be analysed separately.

Keywords: Quality Control, Hardy-Weinberg Equilibrium, Pharmacogenetics, Allele Collapsing, Clinical Pipelines

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

TIIVISTELMÄ

Aki Asikainen: Geenidatan laadunvalvonta farmakogeneettisessä testauksessa;
Validointityökalun ohjelmistototeutus

Pro Gradu

Tampereen yliopisto

Bioteknologian ja biolääketieteen tekniikan maisterikoulutus

Ohjaajat: Professori Matti Nykter, Juho Heliste

Tarkastajat: Professori Matti Nykter, Juho Heliste

2026 Tammikuu

Lääkevalintojen ja -annostuksien apuna käytettävä farmakogeneettinen testaus perustuu tarkoihin genotyypin- ja fenotyypitietoihin. Laboratorioihin kuitenkin usein saapuu etnisyydeltään suuresti vaihtelevia näytteitä, joiden monimutkaiset genotyypit aiheuttavat haasteita käytettäville työkaluille. Tutkielman yhteydessä kehitettiin Python-pohjainen työkalu farmakogenetiikan aineistojen laadunvalvontaan. Tämä prototyyppi yhdistää Hardy-Weinbergin tasapainoperiaatteeseen perustuvan testauksen, viiteaineistojen frekvenssivertailun ja fenotyypitason tarkastelun, ja on tarkoitettu myöhemmin jatkokehityksen kautta integroitavaksi Abomicsin tuotantojärjestelmään.

Kehitetty työkalu ei ainoastaan käsittele suoraan annettua genotyyppi-, haplotyyppi- tai fenotyyppidataa, vaan tukee myös monialleelisten varianttien uudelleenkoodausta bialleeliseen muotoon ja monimutkaisten haplotyyppien alleelikohtaista yhdistämistä. Näiden avulla suoritettu HWE-testaus voi tuottaa vakaampia ja helpommin tulkittavia trendejä laadunvalvonnassa, sillä harvinaisemmat genotyypit yhdistyvät selkeämmäksi kokonaiskuvaksi. Työkalu mahdollistaa myös alleelifrekvenssien vertailun ulkoisiin viitetietokantoihin (esim. Ensembl, ClinPGx) ja sisäisiin kliinisiin aineistoihin. Fenotyyppien kautta tehtävä tarkastelu tuo lisänäkökulman laadunvalvontaan ryhmittelemällä näytteet laajempiin ja merkityksellisempiin kategorioihin.

Suorituskykyä arvioitiin ensin 1000 Genomes -projektin aineiston kromosomin 22 kolmialleelillä yksittäisen nukleotidin muutoksilla. Työkalu tunnisti odotettujen tasapainotilanteiden lisäksi poikkeamia, esimerkiksi kokonaan puuttuvia genotyyppisiä ja populaation rakenteesta kertovia testituloksia. Näiden arvioitiin aineiston pohjatietojen perusteella kuvastavan vahvaa maantieteellistä eriytymistä aineiston sisällä. Anonymisoidussa kliinisessä aineistossa havaittiin selkeitä poikkeamia CYP1A2-geenin alleeleilla. Harvinaisemmilla alleeleilla havaitut suuret poikkeamat arvioitiin matalista alleelikantajamääristä johtuviksi, tehden näistä alleeleista epäluotettavia HWE-pohjaiseen laadunvalvontaan. Yleisemmällä alleeleilla havaitut erot arvioitiin Wahlundin ilmiötä vastaaviksi CYP1A2-geenin alleelien sukuuuritarkastelun kautta. Näin ollen erot selittyivät populaatiorakenteella, eivät genotyypityksen virheillä. Eurooppalaisten ja itäaasialaisten viiteaineistojen vertailu kliinisen datan CYP2C19-geenin alleelifrekvensseihin osoitti, että tutkittu kohortti oli pääosin eurooppalaista alkuperää. Hyvin harvinaiset alleelit tuottivat tilastollisesti merkitseviä, mutta käytännössä vähäisiä eroja, mikä puoltaa näytteiden vähimmäisrajan käyttöä. Fenotyyppien jakaumat satunnaisotannassa vastasivat useimpien CYP-geenien viitearvoja, ja havaitut erot heijastelivat odotetusti kohortin koostumusta.

Työkalun ja siinä sovellettujen menetelmien tehokas käyttö vaatii havaintojen perusteella tarkkaa käytettävien testien ja testattavien alleelien valintaa. Myös eri laboratorioista saatujen aineistojen kohdalla tulisi harkita, onko tarpeen analysoida erikseen aineistot, joilla on erilainen etninen tausta.

Avainsanat: laadunvalvonta, Hardy-Weinbergin tasapaino, farmakogenetiikka, alleelien yhdistämien, kliiniset analyysiprosessit

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

USE OF AI-BASED TOOLS

In the preparation of this thesis, I used Microsoft Copilot (GPT-4 / GPT-5) as a supportive tool. Copilot assisted in drafting the structure of the thesis research plan, helping to ensure that all required components were included and formatted according to the university's guidelines. It also provided suggestions for examples and refining the content, improving clarity, coherence, and academic tone. All AI-generated input was critically reviewed and revised by me, and no content was used without my own evaluation and understanding. I acknowledge that I am fully responsible for the entire content of this thesis, including any parts where AI tools were used for support. I accept full accountability for any violations of academic integrity, ethical standards, or publication practices related to the use of AI.

PREFACE

I wish to thank my supervisors for their guidance and constructive feedback throughout the process. Their insights have been invaluable in shaping the direction and clarity of this work. I am particularly grateful to Juho Heliste, whose mentorship was a significant help during this project.

Huge thanks to my family. Their patience and support were essential.

Finally, I thank the academic community, Abomics and everyone whose work made this research possible.

Järvenpää, 21 December 2025

Aki Asikainen

CONTENTS

1. INTRODUCTION	1
2. LITERATURE REVIEW.....	2
2.1 Genetic inheritance, haplotypes, and phenotypes in pharmacogenetic data analysis.....	2
2.2 Genetic data validation	4
2.2.1 Copy number variation	5
2.3 Hardy-Weinberg equilibrium	6
2.3.1 Causes of deviations from the equilibrium.....	7
2.3.2 External validation through reference databases	8
2.3.3 Limitations due to fractional counts and rare genotypes.....	9
2.4 Statistical methods in genetic data analysis.....	14
2.4.1 Chi-square and Fisher's exact test.....	14
2.4.2 Permutation based testing and Monte Carlo simulation	15
2.5 Existing software tools	16
2.6 Finding and Applying Allele Frequency References	17
3. OBJECTIVES	19
4. MATERIALS AND METHODS	20
4.1 Overview	20
4.2 Data Sources.....	20
4.2.1 1000 Genomes Project Data.....	21
4.3 Software Implementation	22
4.4 Data Flow and Functional Overview	23
4.5 Validation use cases.....	24
4.5.1 HWE Assessment at the Variant Level.....	25
4.5.2 Allele-Wise HWE Assessment at the Haplotype Level	25
4.5.3 External Frequency Comparison.....	26
4.5.4 Phenotype Frequency Validation	27
4.5.5 Unit Testing and Verification	27
5. RESULTS AND DISCUSSION.....	29
5.1 Internal HWE Analysis of 1000 Genomes Data at Variant Level	29
5.2 Allele-Wise HWE Assessment at the Haplotype Level	32
5.3 External Frequency Comparison.....	35
5.4 Phenotype Frequency Validation	37
6. CONCLUSIONS.....	39
REFERENCES.....	40

LIST OF SYMBOLS AND ABBREVIATIONS

API	Application Programming Interface
CPIC	Clinical Pharmacogenetics Implementation Consortium
ClinPGx	Clinical Pharmacogenomics (knowledgebase)
CNV	Copy Number Variation
gnomAD	Genome Aggregation Database
HGVS	Human Genome Variation Society
HWE	Hardy-Weinberg Equilibrium
MC	Monte Carlo (simulation)
NM	Normal Metabolizer
PharmVar	Pharmacogene Variation Consortium
PM	Poor Metabolizer
RM	Rapid Metabolizer
REST	Representational State Transfer
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
UM	Ultrarapid Metabolizer

1. INTRODUCTION

Pharmacogenetics looks at how genetic differences affect how people respond to medications and has become a key part of personalized medicine. As pharmacogenetic testing becomes more common in clinical settings, it is crucial to ensure that genetic data is accurate and reliable for patient safety. One important tool for quality control is the Hardy-Weinberg Equilibrium, a concept from population genetics that helps spot genotyping errors, population differences, and other data issues [1]. The Hardy-Weinberg principle, first described in 1908, says that allele and genotype frequencies stay the same from one generation to the next in an ideal population [2]. In practice, most populations do not meet these ideal conditions, so when we see deviations from HWE, it may point to data problems or important biological factors like group differences, inbreeding, or natural selection.

Abomics Ltd runs a pharmacogenetic testing portal that uses genetic data from clinical labs to give medication recommendations based on patient genotypes. As more and more diverse samples are tested, strong quality control is needed to keep clinical results reliable. Pharmacogenetic testing often deals with haplotypes, which are combinations of genetic variants in a gene, possibly with genes that have more than two alleles. These factors make HWE testing and quality control more complicated. Most standard tools do not offer the integration, automation, or interpretation needed in clinical practice [3,4].

To meet these challenges, this thesis introduces a Python-based tool for HWE testing. The tool adapts traditional HWE methods to work with multi-allelic variants and star allele haplotypes. The methods are tuned with allele collapsing and external data usage such as Ensembl and CPIC/PGx. The tool fits smoothly into clinical lab workflows like those at Abomics, and automates quality control and reporting. It highlights unusual data patterns, and improves patient safety by applying population genetics in daily clinical work.

2. LITERATURE REVIEW

2.1 Genetic inheritance, haplotypes, and phenotypes in pharmacogenetic data analysis

Pharmacogenetic data analysis is based on how genetic traits are passed from parents to children. Inheritance means passing genetic material from one generation to the next, following Mendel's laws. Each person gets two copies of every autosomal gene, one from each parent, which make up their genotype [5]. These genotypes shape traits, including how the body handles medications. Genes that influence drug response are called pharmacogenes.

Pharmacogenetics often looks at specific genetic changes, like single nucleotide polymorphisms (SNPs), which can change how genes work. But looking at single variants alone usually isn't enough for clinical use. Instead, these variants are grouped into haplotypes, sets of genetic changes inherited together on the same chromosome. Haplotypes give a fuller picture of genetic variation, especially when several variants together affect how drugs are processed [6].

To make haplotype interpretation consistent, pharmacogenetics uses the star allele system. A frequently used naming convention for pharmacogenes is the so-called star allele naming system. *1 often denotes the wild type, normally functional allele, and other detected haplotypes get a star allele name in ascending order, *2, *3, and so on [7]. Expert groups such as CPIC (Clinical Pharmacogenetics Implementation Consortium) and PharmVar (Pharmacogene Variation Consortium) organize these star alleles, which are the basis for translating genotypes into clinical phenotypes [8,9].

Each person carries two haplotypes per gene, forming a diplotype, a pair of star alleles that together define the individual's genetic status at that locus. For example, the CYP1A2 gene has *1A allele with normal enzyme function and *3 allele with reduced function. Someone with a *1A/*3 diplotype has one normal function and one reduced function allele. This diplotype is then mapped to a phenotype, such as intermediate metabolizer, which predicts how the person will process drugs metabolized by CYP1A2, like caffeine. A person with an intermediate metabolizer phenotype breaks down caffeine more slowly than normal, which may affect their response to caffeinated products [7].

The CYP1A2 gene illustrates the complexity of the process. Over 170 SNPs have been identified in its regulatory region, and many different variant alleles have been recorded.

These changes aren't just in the coding region, many are in non-coding areas that control gene expression. At least 24 haplotypes are linked to regulatory changes, and 17 involve changes in the protein-coding part [10].

The distribution of haplotypes can differ a lot between ethnic groups. For example, the CYP1A2 *1C haplotype, linked to changes in enzyme activity, appears at different rates among Emiratis. Figure 1 shows an example of these rates.

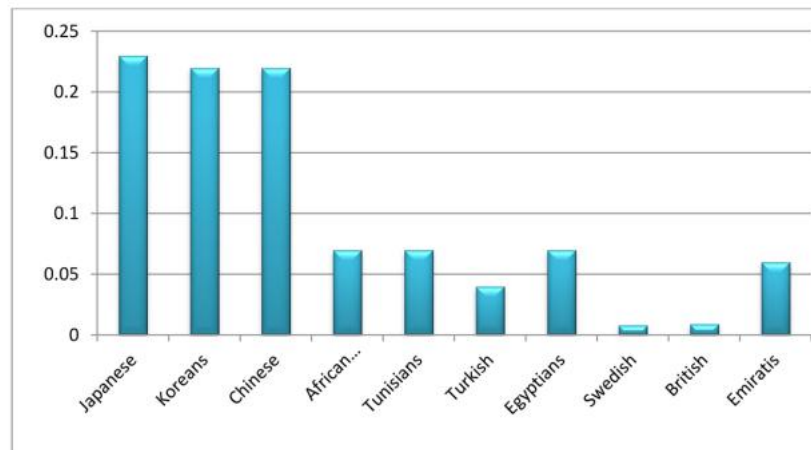


Figure 1. CYP1A2 *1C haplotype frequencies among Emiratis in comparison with other populations [10].

The rates of different CYP1A2 phenotypes also vary between populations worldwide. Figure 2 compares how common poor metabolizers are among Emiratis and other groups, showing why it's important to consider population-specific pharmacogenetic profiles.

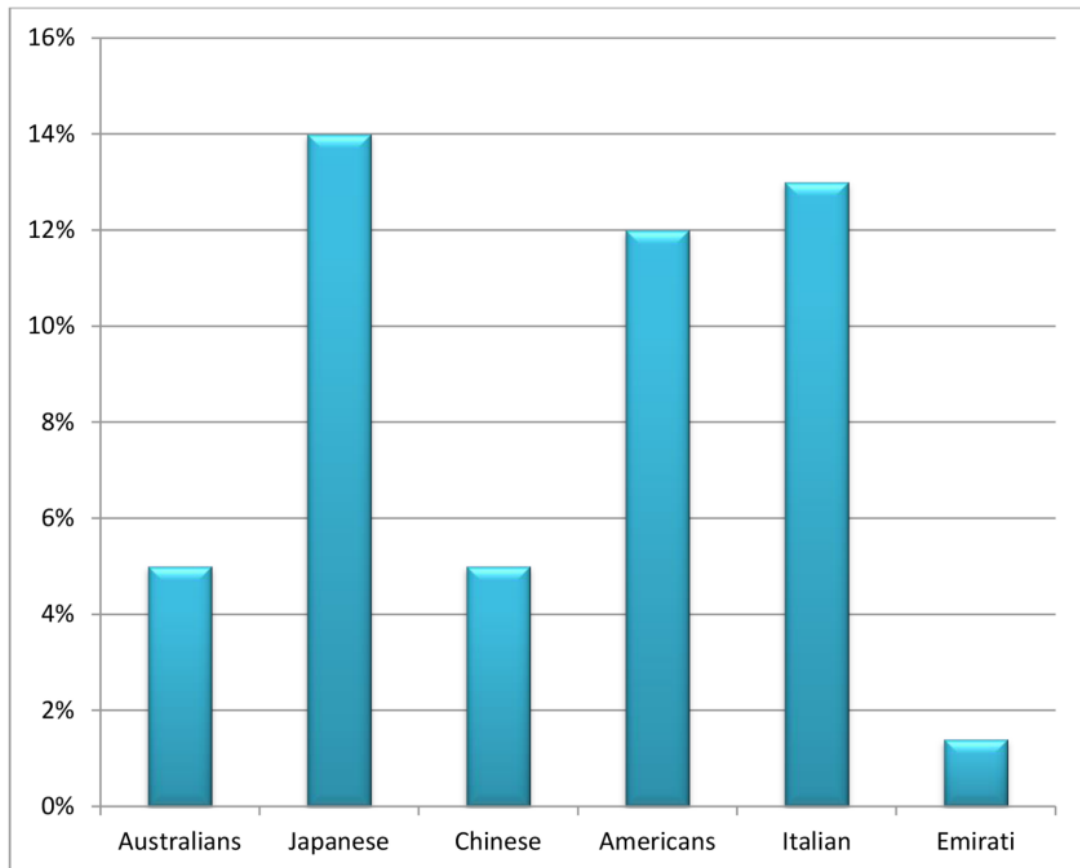


Figure 2. The frequency of CYP1A2 poor metabolizers among Emiratis in comparison with other populations [10].

In clinical pharmacogenetics, turning raw genetic data into haplotypes, diplotypes, and phenotypes uses complex algorithms, reference databases, and quality checks. Mistakes in identifying variants or assigning haplotypes can lead to wrong phenotype predictions and may put patients at risk [1]. That's why strong analytical systems are needed to make sure pharmacogenetic results are reliable.

2.2 Genetic data validation

Reliable analysis of genetic information requires that the input data is systematically validated. Inconsistencies introduced during sample processing, genotype calling, or data ingestion can propagate through statistical tests and result in misleading conclusions [1]. In clinical applications, such as pharmacogenomics, the consequences of such errors can be significant. This chapter reviews key concepts and methods used to assess the quality, integrity, and plausibility of genetic data.

Validation methods depend on the specific situation. For example, at the population level, researchers often use the Hardy-Weinberg Equilibrium to check if observed and expected genotype frequencies match. Differences from HWE can result from biological factors like population structure or inbreeding, as well as technical issues such as genotyping errors or copy number variation [1]. Another way to validate data is by comparing allele frequencies to those in external reference databases like Genome Aggregation Database (gnomAD) or CPIC, which helps identify unexpected patterns based on ancestry or known allele frequencies [11]. It is also important to pay close attention to rare variants, low-frequency alleles, and genotyping artifacts, which may need special statistical methods like collapsing or filtering [12].

Statistical tests commonly used in genetic validation include the chi-square test, Fisher's exact test, permutation-based methods, and simulations such as Monte Carlo (MC). Each has specific assumptions and limitations, particularly in the presence of fractional frequencies or sparse sample sizes [13]. These methods are explored in greater detail in later chapters, accompanied by guidance on their appropriate application across varying conditions.

Finally, existing software tools supporting genetic validation will be reviewed, along with the methodologies and assumptions behind them. Special consideration is given to practical challenges with public databases, allele nomenclature, and integration into automated pipelines.

2.2.1 Copy number variation

Copy Number Variation (CNV) describes DNA segments that differ in copy number from person to person. These changes can be deletions, duplications, or more complex forms, and they often cover large parts of the genome. CNVs can influence gene dosage, gene expression, and disease risk. Since CNVs do not fit the usual diploid model used in many genetic studies, they create special challenges for interpreting results and validating data [14].

CNVs differ from typical single nucleotide polymorphisms (SNPs) because they create more possible genotype combinations than the usual AA, AB, or BB. For instance, a person might have genotypes like AAB or AAA. These extra combinations make it harder to use standard statistical tests, detect equilibrium, estimate allele frequencies, and interpret heterozygosity [15].

Real-world examples help illustrate CNV impact. A well-known case is the deletion of the UGT2B17 gene, which influences androgen metabolism and is linked to variation in testosterone clearance [16]. Another example involves CYP2D6 gene duplications, which affect drug metabolism. Individuals with multiple active copies may experience ultra-rapid processing of medications like codeine or antidepressants, resulting in diminished efficacy or increased toxicity [17].

CNVs also contribute disproportionately to rare variant pools. For instance, a duplication occurring in fewer than 1% of individuals may be flagged during frequency-based filtering unless grouped with related configurations. This makes collapsing strategies essential for avoiding bias and preserving statistical power.

One strategy for addressing copy-number variations (CNVs) during analysis is to exclude affected loci from downstream processing [18]. In the context of this study, such an exclusion approach was adopted to maintain methodological simplicity.

2.3 Hardy-Weinberg equilibrium

The Hardy-Weinberg Equilibrium represents a fundamental concept in population genetics, providing a theoretical framework for describing the stability or change of genetic variation in populations over time. According to this principle, allele and genotype frequencies will stay constant across generations, provided certain conditions are met, like random mating, absence of mutation, migration, selection, and genetic drift. The power of HWE lies in its ability to flag deviations from expected frequencies, which can reveal underlying biological phenomena, such as natural selection or population structure, as well as technical issues, including genotyping errors. As a result, HWE serves as a practical tool for quality control and validation in genetic studies [1].

The Hardy-Weinberg equilibrium is most commonly described for a single locus with two alleles. If the two alleles are denoted as A and a, with respective allele frequencies p and q (where $p + q = 1$), the expected genotype frequencies under HWE are given by the following equation: $p^2 + 2pq + q^2 = 1$ [2].

That is, the proportion of individuals with genotype AA is expected to be p^2 , Aa is $2pq$, and aa is q^2 . These equations provide the baseline for comparison in many practical genetic studies.

These genotype frequencies follow directly from probability theory. When alleles combine at random, the probability of an individual receiving two A alleles (AA) is $p * p = p^2$.

The probability of an Aa genotype can occur in two ways: receiving A from one parent and a from the other ($p * q$), or vice versa ($q * p$), giving a total probability of $2pq$. The probability of receiving two a alleles (aa) is $q * q = q^2$. These calculations assume random mating and no evolutionary influences, which is why the Hardy-Weinberg equation has this specific form.

The Hardy-Weinberg principle can also be generalized to situations with more than two alleles at a locus. If there are k alleles with frequencies p_1, p_2, \dots, p_k (where the sum of all equals 1), then the expected genotype frequencies are determined by expanding $(p_1 + p_2 + \dots + p_k)^2$, resulting in expected frequencies for all homozygous and heterozygous genotypes [19]. For example, with three alleles (A, B, and C) at frequencies $p, q,$ and r , the expected genotype frequencies would be:

- AA: p^2
- BB: q^2
- CC: r^2
- AB: $2pq$
- AC: $2pr$
- BC: $2qr$

This generalization makes HWE a versatile tool for assessing equilibrium in complex genetic systems.

In practical genetic data analysis, HWE is routinely employed to assess the internal consistency of genotype data. Variants that do not conform to HWE expectations can signal a range of issues, including sample contamination, suboptimal probe performance, or misalignment with reference genomes. Nonetheless, it is crucial to interpret HWE deviations within the appropriate context. For example, rare variants may deviate from equilibrium simply due to small sample sizes or loci affected by copy number variations. Thus, while HWE testing is a powerful quality control measure, its results should always be considered together with other biological and technical factors [1,14].

2.3.1 Causes of deviations from the equilibrium

Genetic datasets often show differences from Hardy-Weinberg Equilibrium, and it is important to understand why. These differences usually have clear causes that affect the basic assumptions of the HWE model. In real studies, assumptions like random mating, no selection, no migration, no mutation, and very large populations are often not fully met.

Population structure is a key reason for deviations from equilibrium. If a dataset includes people from different genetic backgrounds, combining them can hide true allele frequencies and make homozygosity seem higher than it is. In studies covering several ethnic or geographic groups, not separating samples before analysis can cause equilibrium violations that reflect actual demographic patterns, not mistakes in analysis.

Non-random mating, such as inbreeding or choosing mates with similar traits, also affects equilibrium. In small or culturally similar groups, relatives may mate more often, which increases the number of homozygous genotypes. Choosing partners based on traits like height or disease can also change the balance between expected and observed heterozygosity.

Natural selection is another important factor. Alleles that help with survival or reproduction become more common over time, which changes genotype frequencies from what is expected under equilibrium. For example, alleles related to drug metabolism, immune response, or resistance to disease can be favored in certain situations.

Technical issues can also cause deviations from equilibrium. Common problems include genotyping errors, especially mistakes in identifying heterozygotes. Sometimes, alleles are missed or probes do not work well, which lowers the observed number of heterozygotes and can look like inbreeding or selection. Other issues like batch effects, sample contamination, and platform-specific problems can make analysis more difficult, especially in large studies where small errors can add up.

All these biological and technical factors show why it is important to consider the study design, population background, and genotyping methods when looking at HWE deviations. Instead of using HWE as a strict rule to exclude variants, it should be seen as a tool to help judge data quality and biological relevance [19].

2.3.2 External validation through reference databases

Internal equilibrium testing is a good starting point for checking genotype consistency, but external validation adds another important perspective. One effective way to check genetic data is to compare it with large population databases. Resources like gnomAD, 1000 Genomes, PharmGKB, and CPIC collect allele and genotype frequencies from many groups [11]. By using these databases, researchers can see if a variant's frequency in their study makes sense or stands out as unusual.

If a variant in a study shows a very different frequency from what is expected in a similar reference group, it could signal a technical problem, a batch effect, or a unique population feature. For instance, if a variant usually appears in 25% of European individuals but only 2% in your European study group, it is worth looking into. This difference might be due to, e.g., missed calls, sample mix-ups, or copy number changes at that location.

Still, researchers need to be cautious. Not every reference database represents all populations equally well. Some may not include enough diversity, may have errors specific to certain technologies, or might miss structural changes like CNVs. Real differences in frequency can also happen due to local history or recent evolution. That's why it's best to compare samples with similar ancestry and to consider biological, demographic, and technical details when interpreting results.

2.3.3 Limitations due to fractional counts and rare genotypes

In studies with small groups, traditional statistical methods often do not work well. Rare variants are hard to validate with statistics, particularly when they appear only a few times in a group or are not found in reference databases. For example, if only one or two people in a group have a certain variant, it is unlikely that any test will detect a meaningful difference from equilibrium. Standard statistical tests may struggle with these variants because their counts are low or because the data include fractional values. For example, as shown in Figure 1, the frequencies of CYP1A2 haplotypes vary significantly between populations. Most count-based statistical tests require whole numbers, so using fractional or non-integer values makes these tests invalid. In these cases, collapsing techniques offer a useful way forward.

For example, Fisher's exact test cannot process fractional inputs, forcing analysts to either round values (introducing bias) or exclude such entries altogether. If data is integer-based, rare genotypes can lead to sparsity in contingency tables. This could result in unstable p-values, for example in the case of the chi-square test, due to its underlying assumptions. In these situations, one option is to use simulation-based methods. Another one is the use of collapsing strategies.

Collapsing means grouping several rare variants together based on shared features, such as function, location in the gene, or how often they appear. For example, all non-synonymous variants in a pharmacogene can be put into a single functional variant group. This lets researchers study their combined effect instead of testing each rare variant one by one [20]. Collapsing can reveal important patterns, reduce the number of

statistical tests needed, and help detect signals that might be missed when looking at each variant separately. This method is especially helpful for rare genotypes or low counts, where traditional statistical tests often do not work well [21]. During validation, researchers should check that grouped variants have frequencies that make sense, match external reference data, and are consistent across batches or platforms. Careful grouping is important. Mixing variants with different effects or from different populations can hide real signals or introduce errors [22].

One practical collapsing approach is to group the site into a target allele versus all others and test the Hardy–Weinberg equilibrium in this simpler, biallelic form. This would help with multi-allelic genetic sites, like CYP2D6, which often lead to genotype tables with many low expected counts [5].

Consider a genetic site with k alleles. Select one allele of interest, A , with frequency p_A , and define q as the frequency of all other alleles combined, so $q = 1 - p_A$. Group genotypes into three categories:

- A/A stays A/A
- A/alt collects all $A/\text{non-}A$ heterozygotes
- alt/alt collects all $\text{non-}A/\text{non-}A$ genotypes

The HWE expected genotype frequencies can be calculated using p_A and q , according to the HWE equations. If the original multi-allelic model is in HWE, the collapsed biallelic distribution will also follow HWE with parameters (p_A, q) .

Consider star allele frequencies:

- $*1 = 0.6$
- $*2 = 0.3$
- $*3 = 0.1$

Full HWE genotype frequency expectations [19]:

- $*1/*1: 0.36$
- $*1/*2: 0.36$
- $*1/*3: 0.12$
- $*2/*2: 0.09$
- $*2/*3: 0.06$
- $*3/*3: 0.01$

When grouped so that $*1$ is the target allele and all others are combined, the frequencies collapse as follows:

- $*1/*1 = 0.36$
- $*1/alt = 0.36 + 0.12 = 0.48 = 2 * 0.6 * 0.4$
- $alt/alt = 0.09 + 0.06 + 0.01 = 0.16 = 0.4^2$

The collapsed expectations exactly match the biallelic HWE formula, confirming validity.

However, after collapsing, for example a chi-square test tends to be less detailed and has fewer degrees of freedom. A Hardy–Weinberg equilibrium goodness-of-fit for 3 alleles has 3 degrees of freedom. When data is collapsed into target versus alternative alleles, the degrees of freedom reduce to 1 [19]. This reduction limits the ability to detect various types of deviations from Hardy–Weinberg equilibrium.

This example illustrates deviation masked by collapsing. Let $n = 1000$, and keep the HWE expectations from the validity example.

Allele frequencies:

- $*1 = 0.6$
- $*2 = 0.3$
- $*3 = 0.1$

Full HWE expected genotype counts:

- $*1/*1: 360$
- $*1/*2: 360$
- $*1/*3: 120$
- $*2/*2: 90$
- $*2/*3: 60$
- $*3/*3: 10$

Observed counts:

- $*1/*1=360$
- $*1/*2=300$
- $*1/*3=180$
- $*2/*2=130$
- $*2/*3=10$
- $*3/*3=20$

This full multi-allelic table shows strong lack of fit. Now collapse to target allele *1:

- $*1/*1 = 360$ (expected 360)
- $*1/alt = 300 + 180 = 480$ (expected 480)
- $alt/alt = 130 + 10 + 20 = 160$ (expected 160)

The collapsed chi-square is 0. Here, the redistribution among non-target cells cancels out in the pooled bins, completely hiding the deviation. This shows the main risk of collapsing.

Next an example of deviation that persists after collapsing. Continue with the setup from the previous example, but use observed counts with again a strong lack of fit:

- $*1/*1 = 420$
- $*1/*2 = 300$
- $*1/*3 = 120$
- $*2/*2 = 110$
- $*2/*3 = 30$
- $*3/*3 = 20$

The chi-square statistic for these non-collapsed values is approximately 46, with a p-value of approximately $1 \cdot 10^{-9}$.

Collapsed HWE expectations are:

- $*1/*1 = 420$ (expected 360)
- $*1/alt = 300 + 120 = 420$ (expected 480)
- $* alt/alt = 110 + 30 + 20 = 160$ (expected 160)

The collapsed chi-square is approximately 10, with a p-value of around 0.002.

Yet another example of deviation masked even after per-allele collapsing.

Allele frequencies:

- $*1 = 0.4$
- $*2 = 0.3$
- $*3 = 0.2$
- $*4 = 0.1$

HWE expected counts:

- $*1/*1 = 160$
- $*2/*2 = 90$
- $*3/*3 = 40$
- $*4/*4 = 10$
- $*1/*2 = 240$
- $*1/*3 = 160$
- $*1/*4 = 80$
- $*2/*3 = 120$

- $*2/*4 = 60$
- $*3/*4 = 40$

Observed counts:

- $*1/*1 = 160$
- $*2/*2 = 90$
- $*3/*3 = 40$
- $*4/*4 = 10$
- $*1/*2 = 280$
- $*1/*3 = 120$
- $*1/*4 = 80$
- $*2/*3 = 120$
- $*2/*4 = 20$
- $*3/*4 = 80$

Allele-wise collapsing gives the following results:

For *1:

- $*1/*1 = 160$
- $*1/alt = 480$
- $alt/alt = 360$

This matches HWE.

For *2:

- $*2/*2 = 90$
- $*2/alt = 420$
- $alt/alt = 490$

This matches HWE.

For *3:

- $*3/*3 = 40$
- $*3/alt = 320$
- $alt/alt = 640$

This matches HWE.

For *4:

- $*4/*4 = 10$
- $*4/alt = 180$
- $alt/alt = 810$

This matches HWE.

Collapsing a multi-allelic genetic region into target-versus-alt and recalculating HWE expectations is mathematically valid and useful for questions focused on a particular allele. It can be used to detect trends of validity from sample data. However, this approach loses information and some multi-allelic differences can be hidden. Also multiple-testing adjustments would be needed for per-allele screening in case of significance claims [25]. This is not necessary for exploratory purposes though.

2.4 Statistical methods in genetic data analysis

Statistical testing is essential for validating genetic data. It helps researchers discover patterns, spot unusual results, and make decisions when results are uncertain. As genetic datasets grow larger and more complex, robust statistical methods are required for both main analyses and for checking consistency within and between datasets. The choice of statistical test should match the data and research question. Using the right method helps ensure that results are valid and can be correctly interpreted.

Every statistical method relies on specific assumptions, such as minimum sample size, expected data distribution, and independence of observations. In pharmacogenomics and rare variant analysis, traditional tests often need to be adjusted or replaced with more flexible, simulation-based approaches. Calculating expected genotype frequencies under Hardy-Weinberg equilibrium, especially in small or scaled sample sizes, often leads to fractional (non-integer) values. These issues require careful handling and adjustment to avoid mistakes and incorrect conclusions.

2.4.1 Chi-square and Fisher's exact test

The chi-square test is a widely used method to compare observed and expected genotype frequencies. In genetic validation, it is commonly used to check for deviations from Hardy-Weinberg proportions and to compare allele frequencies between subgroups. The main advantages of the chi-square test are its simplicity and speed. However, the test works best when sample sizes are large enough and each category has at least five expected counts. When variant counts are low or expected frequencies fall below recommended levels, the test becomes unreliable. This is especially common for rare alleles or small studies, where too few individuals carry certain genotypes to meet the test's requirements.

When using the chi-square test for Hardy-Weinberg equilibrium, it is important to calculate the degrees of freedom correctly. The calculation starts with the total number of genotype categories. For a locus with two alleles, there are three genotypes: AA, Aa, and aa, with alleles A and a having proportions p and q , respectively. One degree of freedom is lost because p (or q) is estimated from the equation $p + q = 1$. A second degree of freedom is lost, as in any goodness-of-fit test, because the counts in all categories must sum to the total sample size, making them not fully independent. As a result, for a biallelic locus, the test typically has one degree of freedom [19]. For multiallelic loci, the principle is the same: start with the total number of genotype categories, then subtract the number of estimated allele frequency parameters and the constraint from the fixed sample size. Being careful with degrees of freedom is important, as it ensures the test result is interpreted correctly.

In these cases, Fisher's exact test is a more accurate choice. It calculates exact p-values and does not require large sample sizes. Fisher's test is well suited for sparse data and simple comparisons, such as checking genotype distributions at sites with low minor allele frequencies or in small subgroups. However, it is not designed to compare observed and expected frequencies like the chi-square test. Instead, Fisher's test looks for differences between two groups or conditions. While some software (such as SciPy) can use Monte Carlo simulation to approximate Fisher's test for larger tables [23], its main use is still to compare group outcomes, not to assess population-level equilibrium. Because of this, Fisher's test does not directly measure deviation from Hardy-Weinberg proportions, which is important for genetic validation. For checking equilibrium and consistency in genotype distributions, the chi-square test is usually preferred [19].

2.4.2 Permutation based testing and Monte Carlo simulation

When sample distributions do not fit standard assumptions or are too sparse for regular tests, permutation and simulation-based methods are good alternatives. Permutation testing shuffles the observed data to create a new baseline for comparison. However, one limitation of permutation is that it relies on the observed data, which may already be biased or affected by systematic errors, so it may not always provide an ideal baseline. Since it does not depend on theoretical distributions, this method is still effective for complex or unusual datasets, including those with rare genotypes or unique allele patterns. Monte Carlo simulations generate datasets using known or estimated parameters, such as allele frequencies or population models. These simulations help estimate genotype distributions, statistical power, and p-values for different scenarios. Monte Carlo methods

are particularly useful in HWE analysis because they can handle fractional and sparse values [24].

A practical example of using Monte Carlo simulation for Hardy-Weinberg equilibrium is to generate many random sets of genotype counts under the null hypothesis that the data follow HWE. You then compare the observed chi-square statistic to the distribution of chi-square values from these simulations. While the chi-square statistic is commonly used, this simulation-based approach also allows you to use other statistics, such as likelihood ratio or exact test statistics, depending on your needs [25]. For instance, you can repeatedly sample genotype counts using the expected HWE proportions and your sample size, then calculate the chi-square or another relevant statistic for each simulated sample. The p-value is estimated as the proportion of simulated values that are greater than or equal to the observed statistic. This approach is especially helpful if expected counts are low or assumptions of standard tests do not hold, as it provides an empirical p-value based on simulated data rather than theoretical tables.

Both techniques offer flexible testing options, but they need a lot of computing power and careful setup. To ensure reproducibility, it is important to use stable random seeds and run enough repetitions. Most importantly, the results are only useful if the simulated data closely match the real biological and technical features of the dataset.

2.5 Existing software tools

General tools like PLINK and R packages offer many features for Hardy-Weinberg Equilibrium testing, allele frequency estimation, and genotype filtering.

PLINK is a widely used, command-line based tool that supports a range of population genetics analyses, including HWE tests, association testing, and extensive data formatting options [3]. It is particularly well suited for large-scale array-based studies and can efficiently analyze thousands of genetic sites. PLINK provides rapid batch processing for HWE testing, allele frequency estimation, and sample-level QC. Its strength lies in scalability and simplicity, making it ideal for large cohort studies and array-based genotyping. However, it assumes diploidy and struggles with fractional genotype values.

R provides a flexible environment for statistical computing with many packages, such as HardyWeinberg and genetics, which support exact tests, permutation testing, custom plots, and tailored filtering for rare variant and population analyses. The open-source ecosystem in R is especially useful for method development and custom workflows, and new genetic analysis packages are being added regularly. These tools are widely used

in research and clinical genetics for robust and reproducible analysis pipelines. They are especially helpful for rare variant analysis or small samples, but users need to be comfortable with statistical scripting and manual data setup.

More researchers now use Python as a flexible foundation for custom validation pipelines. Its scientific libraries, such as pandas, numpy, and scipy.stats, help manage genotype matrices, run standard statistical tests, and build permutation or Monte Carlo simulations for rare variant scenarios. Python is also good at connecting with external tools, automating frequency queries through APIs, and visualizing results with libraries like matplotlib or seaborn. However, when it comes to Hardy-Weinberg equilibrium testing and specialized population genetics analyses, Python's libraries are generally less extensive and mature compared to what is available in R. R offers dedicated packages such as HardyWeinberg, genetics, and genepop [4], which provide robust support for exact tests, permutation procedures, and convenient tools for exploring equilibrium in complex scenarios. In contrast, Python users must often implement custom solutions or rely on more general-purpose libraries like scipy [26], which do not include as many HWE-specific functions or options for rare or multiallelic loci. Despite this, Python's strengths lie in its readable syntax, scalability, and ease of integrating different parts of an analysis workflow, making it a strong choice for producing maintainable and automated pipelines.

2.6 Finding and Applying Allele Frequency References

Researchers often turn to public databases such as Ensembl or CPIC to find allele frequencies by population or ancestry. Ensembl is a comprehensive genome database that integrates genetic variation data from several large-scale projects, including the 1000 Genomes Project and gnomAD. It provides detailed allele and genotype frequency data across many populations, accessible through its web browser and Application Programming Interface (API). Ensembl also offers variant annotation, population stratification, and links to clinical significance and external resources, making it valuable for both research and clinical validation [27].

The Clinical Pharmacogenetics Implementation Consortium (CPIC) provides critical guidance for translating genotypes into drug dosing recommendations. While CPIC itself does not generate raw allele frequency data, it curates star allele frequencies from external resources, such as population-scale datasets and published literature, and makes them accessible via the CPIC Pharmacogenomics (PGx) knowledgebase API. CPIC's

main focus remains on clinical utility and actionable guidelines, but its curated data can assist with HWE-based quality control when supported by additional information or genotype data from other sources [8]

When searching the literature and databases, researchers often encounter missing variant details, absent genotypes (especially wild-type alleles), and inconsistent naming conventions. For example, some pharmacogenomic panels only list frequencies for notable alleles, leaving out the baseline or wild-type genotypes needed for a complete picture. Other sources may use different names for the same variants, such as various star allele labels or a mix of rs numbers and gene-based names.

These differences make it harder to compare data across sources. Inconsistent names and formats complicate automated matching and increase the risk of errors. Even clinical databases may present results in fragments, sometimes missing key details like sample size, ethnicity, or data source. These challenges can significantly impact validation efforts or the estimation of expected genotype patterns.

For SNPs, frequency data are usually more consistent and easier to find across platforms. However, for star alleles, especially in genes like CYP2D6 or TPMT, the information is often incomplete or context-dependent. Researchers may need to infer frequencies from related SNPs or haplotype definitions. This makes validation more difficult and requires careful interpretation of results.

In summary, public databases are important reference sources, but obtaining allele frequency data suitable for validation requires careful attention to format, naming, and population matching. Although automated tools can assist in data extraction, validation still relies on accurately matching study genotypes to reference information. Researchers must balance thoroughness with reliability, as missing or mismatched data can significantly impact subsequent analysis.

3. OBJECTIVES

This study aims to apply established data validation methodologies to genotype distributions in genetic test data from clinical labs. The primary objective is to introduce input data validation steps, such as Hardy-Weinberg equilibrium checks, into Abomics' pharmacogenetic portal and systems. Ensuring new data aligns with trusted reference distributions supports quality control, ultimately improving the reliability of pharmacogenetic interpretations and medication recommendations.

This research investigates how established genetic data validation methods can be effectively integrated into Abomics' existing workflows and systems. It explores the challenges that must be addressed to ensure the reliable application of these methods, such as data compatibility, automation, and scalability.

The thesis delivers a prototype validation workflow compatible with Abomics' systems, demonstrating the feasibility and benefits of established data validation methodologies. This work supports improved quality assurance in pharmacogenetic services and provides a framework for future integration into clinical data pipelines.

4. MATERIALS AND METHODS

4.1 Overview

This chapter outlines the computational framework developed to validate Hardy-Weinberg Equilibrium across both single nucleotide polymorphisms (SNPs) and haplotypes. The pipeline is implemented in Python and supports multiple statistical tests, flexible allele handling, and integration with external frequency databases.

Although the primary goal of the project is to analyze genetic data stored in a test relational database, the pipeline itself has been designed with a modular and database-agnostic interface. This abstraction allows the core logic to remain independent of the underlying database schema, enabling future integration with other data sources that conform to the same input structure.

The test database contains thousands of genetic samples distributed across multiple tables, including variants, genotypes, haplotypes, and phenotypes. These tables are queried dynamically to extract the necessary input for each analysis step. The pipeline processes this input to compute allele frequencies, expected genotype distributions, and statistical test results.

By decoupling the analytical logic from the database structure, the framework ensures both flexibility and scalability. It is optimized to operate on the current test database and is readily adaptable to new datasets formatted in the same relational structure.

4.2 Data Sources

The main data source for the validation pipeline is a relational test database containing thousands of genetic samples. These are distributed across multiple tables, including variants, genotypes, haplotypes, and phenotypes. The database is queried using Structured Query Language (SQL) to extract input data samples for analysis. At the same time, the full database is used as a reference for validation, since it contains the complete history of previously processed samples. This historical context allows the pipeline to detect anomalies and assess whether new data deviate from expected genotype and phenotype distributions.

While the pipeline is designed to process data extracted from this database, its input interface is independent of the database schema. This separation allows the validation logic to remain flexible and adaptable to other data sources that follow the same input format. Also unit testing becomes easier. Before analysis, the extracted data are cleaned to remove invalid entries, filter out incomplete records, and ensure that only usable samples are passed to the pipeline.

External frequency data are retrieved during runtime from two sources: the Ensembl REST (Representational State Transfer) API and the ClinPGx REST API. Ensembl provides genotype frequencies for SNPs across global populations. ClinPGx supplies allele frequencies for pharmacogenes. These are used in statistical tests such as chi-square and Monte Carlo simulations, and also in binomial testing of haplotype distributions.

External data are sometimes incomplete, and some SNPs, genes, or alleles may not appear in API responses. For example, in ClinPGx, the wild type allele may be absent, and some of the alleles listed can be complex, such as $*36+*10x2$. The way copy number alleles are described also varies between clinical labs. For CYP2D6, ClinPGx lists 207 allele records, while PharmVar describes over 135 star alleles, which shows the challenge of matching clinical input with database records [28].

Interpreting genotypes with deletions, such as “-/-”, can be challenging. In star allele notation, this is similar to alleles like CYP2D6 $*5/*5$, where $*5$ means a gene is deleted. Since the -/- format does not match the star allele system, it can be misclassified or missed by automated tools. For SNPs, “-” might show an insertion, where “-” is replaced by a letter like “C”, or a deletion, as in rs121908811. Abomics and other sources, including the Ensembl API [27], use this simpler notation, which is easier to read than formats like Human Genome Variation Society (HGVS) [29].

To address these issues, filtering is applied during data collection in the pipeline. If the wild type frequency is missing, as with TPMT and DPYD in ClinPGx, it is calculated from other allele frequencies and double-checked that the total frequency adds up to one. Although mapping difficult genotypes to their corresponding star allele counterparts could also be helpful, it is beyond the scope of this study.

4.2.1 1000 Genomes Project Data

The 1000 Genomes Project was used as an external, public reference dataset for validating the HWE methods at the variant level. The project aimed to map human genetic variation across diverse populations using large-scale sequencing, and it is now a widely used resource for benchmarking genetic analyses [30]. It provides genotype data from

thousands of individuals across several continental groups, making it a good testbed for quality control methods. Since the data are public and fully anonymized, this thesis can report specific counts and p-values without privacy concerns.

The 1000 Genomes dataset includes 2,504 unrelated individuals from 26 populations, grouped into five main super-populations: African, Admixed American, East Asian, European, and South Asian. Each group represents a specific geographic or ethnic background, such as Yoruba in Ibadan, Nigeria, Han Chinese in Beijing, Finnish in Finland, or Punjabi in Lahore, Pakistan. The project combined low-coverage whole-genome sequencing, deep exome sequencing, and high-density microarray genotyping. This approach identified over 88 million variants, all phased to haplotypes [30].

This thesis used a subset of the 1000 Genomes data: phased genotype calls for chromosome 22 from the Phase 3 release. Chromosome 22 was chosen because it has a moderate number of variants in a small genomic region, making it practical for developing and testing methods while still showing realistic patterns of human variation. The data were obtained as a compressed VCF file and its index from the official 1000 Genomes distribution.

The raw VCF data were converted into the format needed for the validation pipeline using a small helper script. This script reads the chromosome 22 VCF file, scans the variants in order, and applies filters to select suitable sites. Only SNPs are kept. The script can also be set to require a certain number of alleles per site. For this thesis, it was set to select tri-allelic variants and to stop after finding five suitable tri-allelic SNPs from chromosome 22.

4.3 Software Implementation

The validation system is implemented in Python and designed to operate on structured genetic data, which can be extracted, for example, from relational databases. Input data consist of individual-level genotype observations, each associated with a variant or gene identifier. Phenotype information may also be included when available.

The system supports sample size filtering to maintain statistical reliability. For each variant or gene, users can set a minimum sample number, with a default of five. Those with fewer samples are excluded from analysis to prevent unreliable results.

4.4 Data Flow and Functional Overview

The process begins with the input of data, typically sourced from a database. The initial step involves calculating allele frequencies from the input data. These frequencies are then used to compute Hardy-Weinberg Equilibrium genotype frequencies.

At this stage, the workflow introduces a decision point involving allelic recoding. Depending on whether and how recoding is performed, the analysis proceeds in one of three ways. In the case of no recoding, the pipeline uses raw observed genotype or phenotype counts. With biallelic recoding, the pipeline groups one target allele against all others, collapsing the genotype data into a biallelic form. This approach is particularly useful for multi-allelic sites with low counts and allows for HWE testing using simplified genotype categories. For allele-wise biallelic recoding, the pipeline applies the collapsing approach separately for each allele, allowing users to screen for deviations or trends associated with individual alleles while still reducing complexity.

Regardless of the recoding choice, all branches converge on the next decision point: whether to use internal HWE frequencies or external expectations for further analysis. When external expectations are selected, frequencies can be fetched from sources such as Ensembl, ClinPGx, or a full sample database. Otherwise, internal HWE frequencies are used.

The subsequent step is to choose an appropriate validation test for variants, haplotypes, or phenotypes. The available tests include the chi-square test, Monte Carlo simulation, binomial test, and a phenotype-specific chi-square test. If the assumptions for the chi-square test are violated, the workflow allows for a retry using alternative methods.

Finally, the results from the chosen statistical test are compiled and returned, completing the data flow shown in Figure 3.

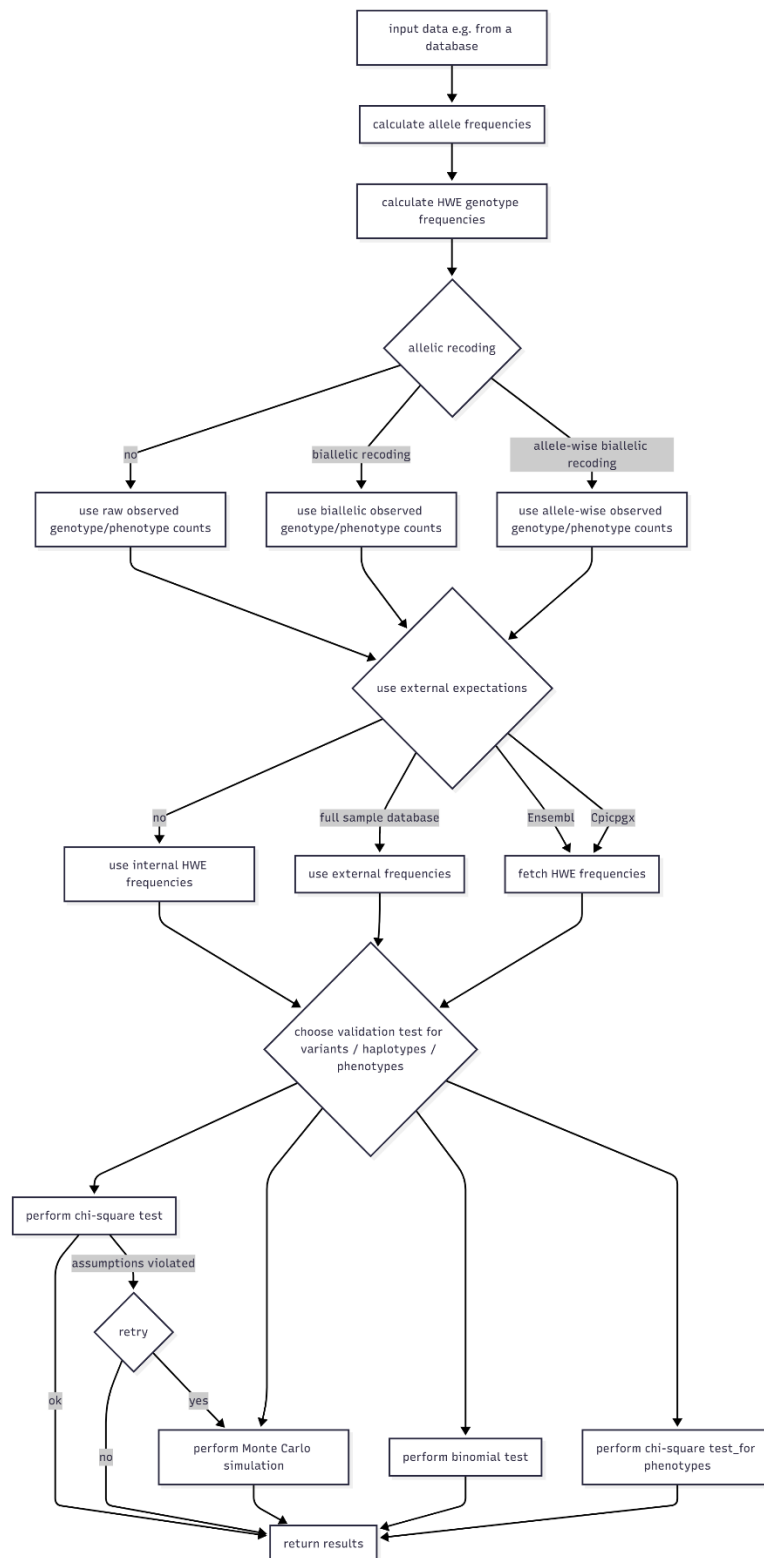


Figure 3. Flowchart for the software implementation.

4.5 Validation use cases

The pipeline is modular, allowing users to selectively apply validation modes depending on the input structure and analytical goals. These include HWE assessment at the variant

or haplotype level, external frequency comparison, validation of phenotype frequencies, and comparison against curated population databases.

4.5.1 HWE Assessment at the Variant Level

This mode checks HWE for each variant by estimating allele frequencies from the study sample or batch, calculating expected genotype distributions under HWE, and comparing them to observed genotype counts. The process starts with input data listing each variant and its two alleles. Genotypes are labeled consistently by sorting the alleles, and the counts for each genotype are totaled for every variant.

Allele frequencies are first counted for each variant. All variants are then recoded to a biallelic form, where the most abundant allele is set as the reference and all other alleles are grouped into an alternate category. Expected genotype frequencies are calculated using standard HWE equations and scaled to the sample size.

A chi-squared test is used to check goodness-of-fit when both expected and observed counts are large enough and the genotype categories match. Degrees of freedom are adjusted as is standard for HWE tests, to account for estimating allele frequencies from the observed data. If the assumptions for the chi-squared test are not met or a count threshold is not reached, a Monte Carlo method is used to get an empirical p-value. This test simulates genotype tables based on the expected HWE frequencies, uses the same chi-squared statistic, and estimates the probability by resampling. The workflow first attempts the chi-squared test, then utilizes Monte Carlo simulation for cases where the chi-squared test is not reliable, and combines the results in a single table. Results are returned as per-variant records, which include the genotype categories, observed and expected counts, test statistic, p-value, and the test type used, with a companion record summarizing any skipped variants and the reasons for their exclusion.

While the primary use case involves biallelic recoding, the method also supports use cases where assessments are performed directly on the original multi-allelic genotypes and frequencies. This provides flexibility for analyses that require full allelic resolution.

4.5.2 Allele-Wise HWE Assessment at the Haplotype Level

In this mode, HWE is checked at the haplotype level by comparing each star allele to all other alleles in the gene. This approach is intended for pharmacogenetic studies that focus on gene-level haplotypes, like CYP2D6 star alleles. It is particularly useful for genes with many different haplotypes.

The analysis starts with a dataset that lists gene names and the two haplotypes found for each person. For each gene, all unique haplotypes in the sample are identified. Each haplotype is tested on its own, while all other haplotypes for that gene are grouped together as the alternate type. This means that a gene with k haplotypes will have k separate HWE tests, each comparing one haplotype to the rest. This method turns a multi-allelic situation into a series of two-allele comparisons. Haplotype frequencies, expected genotypes, and HWE tests follow the same steps as in the variant-level approach.

This method gives more detailed results than gene-level HWE tests because it looks at individual haplotypes. It can show which specific star alleles might cause deviations from equilibrium. However, since each test compares just one haplotype to all others, it may miss patterns involving pairs of haplotypes that are not tested together. For this reason, this approach works best for exploring data and spotting trends, and results should be interpreted carefully.

4.5.3 External Frequency Comparison

This part checks if the frequencies seen in the study sample are different from those found in larger population data. It is mainly used in two ways: to compare haplotype frequencies for pharmacogenetic star alleles with the ClinPGx API, and to compare variant genotype frequencies with Ensembl API population data.

For haplotype data, reference allele frequencies for each gene are fetched from the ClinPGx API, filtered by the chosen population.

The analysis uses an allele-wise biallelic conversion, similar to the haplotype-level assessment. Each unique haplotype in the study is considered as the main allele in turn, and its observed frequency is compared to external reference frequencies with a binomial test. The resulting p-value indicates how much the study data differs from the reference.

For variant data, expected genotype frequencies are obtained from Ensembl for the selected population. These frequencies serve as the expected values for each variant. The observed genotype counts are then compared to these expectations using the same chi-squared or Monte Carlo methods as used in internal HWE testing.

In addition to using Ensembl, the pipeline also supports HWE-based validation using external frequency data from a full sample database. This allows equilibrium testing to be performed using population-level frequencies that reflect the entire reference dataset, providing an alternative perspective to internal HWE checks or Ensembl comparisons.

Identifiers are usually matched to external references by exact name. For external frequencies, allele names are matched by star allele numbers, especially for pharmacogenes. If the wild-type allele is missing and the total frequency does not add up to one, the method adds the wild type to correct the total. If an identifier is missing from the external reference or the frequency is invalid, the test is skipped and the identifier is listed in a companion table with the reason. There is also an optional check to confirm that all frequencies sum to one; if not, the external frequency is excluded. Results are reported in the same format as before, allowing for easy comparison.

HWE testing using reference data from the ClinPGx database is also supported. However, this is not a main use case.

4.5.4 Phenotype Frequency Validation

This step checks if the observed phenotypes in the samples match the expected distribution from external reference data. One important use of this approach is to validate that incoming data looks similar enough to expectations before further analysis [31]. This is especially important in pharmacogenetics, where predicted metabolizer phenotypes are assigned based on genotype and translation rules. Reference frequencies can come from published studies, population databases, or clinical resources. For this thesis, the reference frequencies are taken from a larger sample database.

For each gene, the observed phenotype frequencies are calculated and compared to the reference frequencies using a chi-squared test. In contrast to Hardy-Weinberg analyses, these phenotype frequencies are taken directly from external sources instead of being derived from alleles. To make the comparison fair, any missing phenotype categories are filled with zeros in either the observed or expected set. If the categories still do not match, that gene is excluded from the analysis.

Once again results are reported in the same format as in the previous methods, allowing for easy comparison.

4.5.5 Unit Testing and Verification

Unit testing and validation were essential in building the pipeline. To make sure it was accurate, consistent, and easy to maintain, the pytest library was used for the testing framework.

Each main use case was tested on its own. Each subfunction also had its own unit tests, covering tasks like allele frequency calculations, genotype handling, phenotype parsing, and statistical test wrappers. Both simple synthetic data and real examples were used from the test database to check every function. This helped make sure the logic worked with both perfect and real-world data.

For demonstration purposes, additional longer unit tests were developed to showcase the essential use cases and provide clear examples of the pipeline's expected behavior in practice.

5. RESULTS AND DISCUSSION

The 1000 Genomes Project data was used exclusively in the next sub-chapter, as it is public and available for sharing. In contrast, the Abomics database is not public, and its sample counts or specific values cannot be disclosed due to privacy constraints. While 1000 Genomes data provided a transparent, reproducible foundation for evaluating pipeline methods at the variant level, Abomics data was used in the subsequent chapters to demonstrate real-world results. No numerical values or sample counts from Abomics data are reported, and all analyses were performed using anonymous data without access to personal or identifying information.

5.1 Internal HWE Analysis of 1000 Genomes Data at Variant Level

Data from the 1000 Genomes Project for chromosome 22 was analyzed. Five tri-allelic variants were selected and tested for Hardy-Weinberg equilibrium using the pipeline. The results are summarized below, in Table 1 and Table 2.

Id	Genotypes	Observed Counts	Expected Counts
22:16051453	AA, AC, AG, CC, CG, GG	[2051, 399, 12, 37, 5, 0]	[2033.46, 430.75, 15.32, 22.81, 1.62, 0.03]
22:16055268	AA, AC, AG, CC, CG, GG	[0, 0, 3, 0, 4, 2497]	[0.0, 0.0, 3.0, 0.0, 3.99, 2497.0]
22:16114689	AA, AC, AG, CC, CG, GG	[0, 3, 0, 2495, 6, 0]	[0.0, 2.99, 0.0, 2495.01, 5.99, 0.0]
22:16123252	AA, AG, AT, GG, GT, TT	[0, 0, 35, 0, 321, 2148]	[0.12, 2.24, 32.51, 10.29, 298.18, 2160.65]
22:16155290	CC, CG, CT, GG, GT, TT	[2464, 21, 18, 1, 0, 0]	[2463.17, 22.81, 17.85, 0.05, 0.08, 0.03]

Table 1. Genotype counts for the HWE analysis.

Id	Chi ²	p-value	Test Type
22:16051453	19.09	0.03	mc
22:16055268	0.00	1.00	mc
22:16114689	0.01	1.00	mc
22:16123252	14.66	0.03	mc
22:16155290	17.25	0.08	mc

Table 2. *Statistical summary for the HWE analysis.*

Two of these variants showed slight deviations from equilibrium. Since multiple testing correction was not applied, the reported p-values are provided for reference only and are not used to define significance. Because the sample counts were small, the pipeline used a Monte Carlo simulation to get more reliable results.

Variant 16051453 shows an excess of minor allele C homozygotes and a deficit of AC heterozygotes, in Table 3. This pattern could be explained by the Wahlund effect, as the 1000 Genomes Project data includes multiple subpopulations.

Genotype	Observed	Expected	Difference
AA	2051	2033.46	+17.54 (+0.9%)
AC	399	430.75	-31.75 (-7.4%)
AG	12	15.32	-3.32 (-22%)
CC	37	22.81	+14.19 (+62%)
CG	5	1.62	+3.38 (+208%)
GG	0	0.03	-0.03 (-100%)

Table 3. *The results of variant 16051453.*

Variant 16123252 shows a complete absence of certain genotypes (AA, AG, GG), in Table 4. For example, allele A never co-occurs with G. This pattern can be explained by

strong geographic differentiation and minimal admixture between populations carrying different alleles.

Genotype	Observed	Expected	Difference
AA	0	0.12	-0.12 (-100%)
AG	0	2.24	-2.24 (-100%)
AT	35	32.51	+2.49 (+8%)
GG	0	10.29	-10.29 (-100%)
GT	321	298.18	+22.82 (+8%)
TT	2148	2160.65	-12.65 (-1%)

Table 4. The results of variant 16123252.

The biallelic recoding method gave similar results with the same data, in Table 5 and Table 6. Both methods identified the same two variants with HWE deviations, with p-values being slightly lower after collapsing.

Id	Genotypes	Observed Counts	Expected Counts
22:16051453	A/A, A/alt, alt/alt	[2051,411,42]	[2033.46,446.07,24.46]
22:16055268	G/G, G/alt, alt/alt	[2497,7,0]	[2497.0,6.99,0.0]
22:16114689	C/C, C/alt, alt/alt	[2495,9,0]	[2495.01,8.98,0.01]
22:16123252	T/T, T/alt, alt/alt	[2148,356,0]	[2160.65,330.69,12.65]
22:16155290	C/C, C/alt, alt/alt	[2464,39,1]	[2463.17,40.66,0.17]

Table 5. Genotype counts for the HWE analysis using biallelic recoding.

Id	Chi ²	p-value	Test Type
----	------------------	---------	-----------

22:16051453	15.480135	0.00	chi2
22:16055268	0.000000	1.00	mc
22:16114689	0.010000	1.00	mc
22:16123252	14.660000	0.00	mc
22:16155290	4.190000	0.18	mc

Table 6. Statistical summary for the HWE analysis using biallelic recoding.

The collapsed approach provided stronger statistical evidence, as seen in lower p-values for true deviations. For 22:16155290, the p-value was a bit higher due to rare genotype counts and variation being pooled. This can be interpreted as better reliability from generic validation point of view. With fewer genotypes, interpreting the results was easier.

On the other hand, the non-collapsed approach revealed allele-specific patterns, like missing GG but not TT homozygotes. This method provided more detailed biological insights.

In summary, three variants matched HWE expectations well, and the other two also fit established patterns in human population genetics.

5.2 Allele-Wise HWE Assessment at the Haplotype Level

Anonymous clinical pharmacogenetic data from the Abomics laboratory database were analyzed to validate Hardy-Weinberg equilibrium testing methodology at the haplotype level. For privacy reasons, all counts had to be excluded, but the results are presented as statistical values. The CYP1A2 gene encodes a cytochrome P450 enzyme that has a central role in metabolizing caffeine and several commonly used medicines. Variations in the gene can lead to significant changes in enzyme activity, sometimes resulting in reduced or increased metabolism rates [7]. Because of its broad clinical relevance and the diversity of its alleles, CYP1A2 was selected for detailed analysis in this work.

The primary analysis focuses on allele-wise collapsed HWE testing for CYP1A2, with results from the non-collapsed diplotype-based approach provided later in this chapter for comparison. CYP1A2 data were evaluated using allele-wise collapsing, with the chi-square test applied when assumptions were met and the Monte Carlo simulation used

otherwise, results in Table 7. Seven alleles displayed excellent agreement with HWE expectations, supporting high-quality genotyping. Many alleles, even at low or very low frequencies, showed near-perfect equilibrium, supporting the accuracy of genotyping. The Monte Carlo method offered a reliable way to assess data quality in the control pipeline by successfully handling rare and sparse categories.

Diploypes	Chi ²	p-value	Test Type
*1/*1, *1/alt, alt/alt	63.204184	0.00000	chi2
*1B/*1B, *1B/alt, alt/alt	247.625449	0.00000	chi2
*1F/*1F, *1F/alt, alt/alt	26.259357	0.00000	chi2
*1V/*1V, *1V/alt, alt/alt	0.013853	0.90631	chi2
*1C/*1C, *1C/alt, alt/alt	0.000990	1.00000	mc
*1D/*1D, *1D/alt, alt/alt	0.626010	0.77000	mc
*1E/*1E, *1E/alt, alt/alt	0.000030	1.00000	mc
*1G/*1G, *1G/alt, alt/alt	0.019400	1.00000	mc
*1J/*1J, *1J/alt, alt/alt	1.610910	0.47000	mc
*1K/*1K, *1K/alt, alt/alt	0.002510	1.00000	mc
*1K-1/*1K-1, *1K-1/alt, alt/alt	0.012410	1.00000	mc
*1L/*1L, *1L/alt, alt/alt	94.419320	0.00000	mc
*1L-2/*1L-2, *1L-2/alt, alt/alt	27.825830	0.02200	mc
*1W/*1W, *1W/alt,	895.111110	0.00000	mc

alt/alt			
*7/*7, *7/alt, alt/alt	0.000120	1.00000	mc

Table 7. The allele-wise HWE analysis results at the haplotype level.

There were some rare or very rare alleles that did show a lot of deviation, like *1B, *1L, *1-L2, *1K, *1W [34]. Moreover, phasing of these haplotypes is complex, and their counts in the sample may therefore diverge from the actual counts in the population. This was also indicated by the observed counts, which are not included here. With very few carriers in the dataset, even minor population structure can lead to large apparent HWE violations. Therefore, these alleles are not a good choice for data validation purposes through HWE analysis.

The major allele, *1F, had a high chi-square statistic, which shows a strong deviation from HWE [7]. This allele shows substantial frequency variation among ancestral groups: about 57% in Swedish populations and 7% in Korean populations [35]. The wild type allele *1 also had a high chi-square statistic, with a frequency of 91% among Emiratis and 24% in the Swedish population [9,10,32]. Additionally, partial reason for deviations may be related to challenging phasing of the haplotypes.

The observed HWE deviation is probably caused by combining samples from different ancestral backgrounds without mixing. This is called the Wahlund effect and reflects real population structure, not data quality issues. This pattern does not indicate genotyping errors. Instead, it is expected in population genetics when analyzing diverse patient groups together. For clinical labs serving multi-ethnic populations, these deviations are normal and do not require any changes.

Analyzing the same CYP1A2 data without allele collapsing produced a chi-square statistic that indicated deviation from equilibrium, in Table 8. However, this comprehensive approach could not show which specific alleles or genotype combinations contributed to the deviation, limiting its usage for quality control.

Diploypes	Chi ²	p-value	Test Type
*1/*1, *1/*1B, *1/*1C, *1/*1D, *1/*1E, *1/*1F, *1/*1G, *1/*1J, *1/*1K, *1/*1K-1, *1/*1L, *1/*1L-2, *1/*1V, *1/*1W, *1/*7, *1B/*1B, *1B/*1C, *1B/*1D, *1B/*1E,	11652.25176	0.0	mc

*1B/*1F, *1B/*1G, *1B/*1J, *1B/*1K, *1B/*1K-1, *1B/*1L, *1B/*1L-2, *1B/*1V, *1B/*1W, *1B/*7, *1C/*1C, *1C/*1D, *1C/*1E, *1C/*1F, *1C/*1G, *1C/*1J, *1C/*1K, *1C/*1K-1, *1C/*1L, *1C/*1L-2, *1C/*1V, *1C/*1W, *1C/*7, *1D/*1D, *1D/*1E, *1D/*1F, *1D/*1G, *1D/*1J, *1D/*1K, *1D/*1K-1, *1D/*1L, *1D/*1L-2, *1D/*1V, *1D/*1W, *1D/*7, *1E/*1E, *1E/*1F, *1E/*1G, *1E/*1J, *1E/*1K, *1E/*1K-1, *1E/*1L, *1E/*1L-2, *1E/*1V, *1E/*1W, *1E/*7, *1F/*1F, *1F/*1G, *1F/*1J, *1F/*1K, *1F/*1K-1, *1F/*1L, *1F/*1L-2, *1F/*1V, *1F/*1W, *1F/*7, *1G/*1G, *1G/*1J, *1G/*1K, *1G/*1K-1, *1G/*1L, *1G/*1L-2, *1G/*1V, *1G/*1W, *1G/*7, *1J/*1J, *1J/*1K, *1J/*1K-1, *1J/*1L, *1J/*1L-2, *1J/*1V, *1J/*1W, *1J/*7, *1K/*1K, *1K/*1K-1, *1K/*1L, *1K/*1L-2, *1K/*1V, *1K/*1W, *1K/*7, *1K- 1/*1K-1, ...			
--	--	--	--

Table 8. The HWE analysis results at the haplotype level without collapsing.

5.3 External Frequency Comparison

To further check the reliability of the clinical pharmacogenetic data, CYP2C19 allele frequencies from the anonymized Abomics laboratory database were compared with external population reference data. Rather than reporting direct counts, statistical comparisons were made with the ClinPGx database.

CYP2C19 is a gene with about 20 known allele variants. The most common are *2, *3, and *17. The *2 and *3 alleles are much more frequent in Asian populations than in White or African groups, while the *17 allele is more common in White and African populations than in East Asians [33]. These differences in allele frequencies help provide context for the results, as comparing to both European and East Asian references clarifies the genetic background of the sample population.

Compared to the European baseline, binomial test results from the sample population were close in magnitude but with some low p-values. The wild type allele matched expectations. For common alleles, *17 was a bit lower than expected, while *2 was slightly higher. These small but statistically significant differences are likely detectable due to large sample sizes, thus having limited practical impact.

Rare alleles like 3, 4, and 8 [33] showed statistically significant differences, but the effect sizes were very small. Table 9 suggests that p-values for very rare alleles are unstable and can change a lot with small count differences. For external comparisons, it is best to set minimum count thresholds or group rare alleles together before testing.

Allele	Observed Frequency	Expected Frequency	p-value
*1	0.622654	0.625141	0.51125
*10	0.000060	0.000000	0.00000
*17	0.205005	0.215439	0.00103
*2	0.153934	0.146857	0.01035
*3	0.000782	0.001618	0.00488
*4	0.000662	0.002360	0.00000
*6	0.000060	0.000300	0.07219
*8	0.001504	0.003359	0.00001

Table 9. CYP2C19 allele frequencies vs. ClinPGx European reference.

Compared to the East Asian baseline in Table 10, *17 was much higher than expected, while *2 and *3 were much lower. The reference allele also showed a significant difference, but the actual difference was small. This suggests the group is mostly European.

Allele	Observed Frequency	Expected Frequency	p-value
*1	0.622654	0.595549	0.00000
*10	0.000060	0.000105	1.00000
*17	0.205005	0.020541	0.00000
*2	0.153934	0.283523	0.00000
*3	0.000782	0.072473	0.00000
*4	0.000662	0.000177	0.00025
*6	0.000060	0.000560	0.00160
*8	0.001504	0.000000	0.00000

Table 10. CYP2C19 allele frequencies vs. ClinPGx East Asian reference.

5.4 Phenotype Frequency Validation

A random sample of fifty anonymous individuals was selected from the Abomics database to serve as the sample population for genotype-based validation. Phenotype distributions observed in this sample group for cytochrome P450 genes were compared against data from the full database. The phenotype categories included, for example, Normal Metabolizer (NM), Intermediate Metabolizer (IM), Poor Metabolizer (PM), Rapid Metabolizer (RM), and Ultrarapid Metabolizer (UM). For each gene, a chi-square goodness-of-fit test was performed to assess the distribution of these phenotype groups.

For most CYP genes, observed phenotype distributions in Table 11 were broadly in line with expectations. Specifically, CYP1A2, CYP2B6, CYP2D6, CYP3A4, and CYP3A5 did not show significant differences across groups in this analysis. These findings suggest that the allele translation rules and phenotype grouping for these genes are performing as expected, with no evidence of systematic errors.

CYP2C19, CYP2C8, and CYP2C9 showed significant differences, which are likely due to ancestry differences between the sample and reference groups. For example, the *17 allele of CYP2C19 is associated with increased metabolic activity, and the *17/*17 genotype corresponds to the ultrarapid metabolizer (UM) phenotype. The frequency of the *17 allele varies widely, ranging from 1–4% in Asian populations to 18–27% in Europeans, resulting in considerable variability even within population groups [33]. CYP2C8 is sensitive to small changes in group composition because some phenotype categories are rare in individuals of European ancestry, and large differences in allele frequencies can exist even between neighboring countries [36]. CYP2C9 also shows substantial differences between populations in the frequencies of alleles with reduced function and those with normal function. For example, the *2 allele, which is associated with reduced metabolic capacity, has a frequency of 8–19% in Caucasian populations [34].

This analysis shows that using phenotype distributions is a helpful quality control step alongside Hardy–Weinberg equilibrium tests. HWE analysis often results in categories with very few cases, which can make the results hard to interpret. Phenotype-based methods address this by combining sparse genotype categories into broader, more meaningful groups. This approach checks clinical results directly and can reveal problems in genotyping or category definitions. As a quality assessment tool, it can offer new insights or help explain issues that come up.

gene	phenotypes	chi-square	p-value
CYP1A2	HIGH, IM, NM, OTHER, PM	0.342907	0.98688
CYP2B6	IM, NM, OTHER, PM, RM, UM	2.304534	0.80560
CYP2C19	IM, LIM, NM, PM, RM, UM	14.432216	0.01308
CYP2C8	LDM, NM, OTHER, VM	8.171234	0.04260
CYP2C9	IM AS1, IM AS15, NM AS2, PM AS0, PM AS05	17.480563	0.00156
CYP2C_rs12777823	DECREASED, NORMAL	0.075615	0.78333
CYP2D6	FAIL, IM, NM, OTHER, PM, UM	4.052799	0.54184
CYP3A4	IM, NM, OTHER, PM	4.180415	0.24263
CYP3A5	FAIL, IM, NM, PM	3.320604	0.34479
CYP4F2	DECREASED, NORMAL	1.232118	0.26700

Table 11. Results of the phenotype frequency analysis, without actual count data.

6. CONCLUSIONS

From using this validation pipeline, a few practical lessons stand out. When working with mixed populations, there can be a lot of variation, so it is best to use ancestry-matched references, regional datasets, or a lab's own historical data for validation. For patient groups with different backgrounds, it is more accurate to run HWE tests separately for each ancestry group instead of combining all the data. Labs should use ancestry-informed quality control whenever possible.

Biallelic collapsing increases statistical power but may obscure issues with multi-allelic variants, so both detailed and collapsed approaches are valuable. In automated clinical data validation, collapsing often yields more stable and interpretable trends, as rare genotypes or alleles and their variations are combined, providing a clearer overall view. Additionally, it allows allele-specific analysis, which can be helpful when selecting quality control target alleles.

External frequency comparisons and HWE assessments show that even small differences in population structure can cause large, apparent HWE violations for very rare alleles. Setting minimum count thresholds helps prevent misleading results. This highlights the importance of carefully selecting which alleles and data points are included for validation purposes, as not all variants need to be used. Focusing on those with practical relevance ensures that validation efforts are meaningful and robust. In addition, phenotype frequency validation offers a complementary perspective, especially for rare alleles. By combining sparse categories into broader, more meaningful groups, this method can reveal new insights or help clarify unexpected results that may arise during analysis.

Some limitations remain, such as differences in naming conventions between clinical databases and public references, or missing identifiers in either source. There is also a need for more accurate ancestry determination and better tools to monitor changes over time. Addressing these issues by improving ancestry classification, building systems to track changes, and automating the process of matching and standardizing reference names and identifiers with approaches like rs id or pharmacogenic allele mapping is a logical next step to further enhance the pipeline.

REFERENCES

- [1] B. Chen, J.W. Cole, C. Grond-Ginsbach, Departure from Hardy-Weinberg Equilibrium and genotyping error, *Frontiers in Genetics*, vol.8, no.167, Aug 2017, pp.167–176. Available: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2017.00167/full>
- [2] Edwards AWF. G. H. Hardy (1908) and Hardy-Weinberg Equilibrium. *Genetics* 2008 07;179(3):1143-50.
- [3] Purcell, Shaun; Chang, Christopher (2007). PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis. *American Journal of Human Genetics*, 81.
- [4] ROUSSET F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*. 2008;8(1):103–6.
- [5] Weir BS. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland (MA): Sinauer Associates; 1996.
- [6] Newman WG. *Pharmacogenetics : making cancer treatment safer and more effective*. 1st ed. 2010. New York: Springer; 2010.
- [7] Royal Dutch Pharmacists Association (KNMP). CYP1A2 Pharmacogenetics – English Version [Internet]. The Hague: KNMP; 2023 Nov [cited 2025 Oct 22]. Available from: https://www.knmp.nl/sites/default/files/2023-11/CYP1A2_English.pdf
- [8] Clinical Pharmacogenetics Implementation Consortium. CPIC [Internet]. Stanford University & St. Jude Children's Research Hospital; [cited 2025 Sep 16]. Available from: <https://cpicpgx.org/>
- [9] PharmVar Consortium. CYP1A2 Allele Nomenclature Table [Internet]. PharmVar; [cited 2025 Oct 25]. Available from: <https://www.pharmvar.org/htdocs/archive/cyp1a2.htm>
- [10] Al-Ahmad MM, Amir N, Dhanasekaran S, John A, Abdulrazzaq YM, Ali BR, et al. Genetic polymorphisms of cytochrome P450-1A2 (CYP1A2) among Emiratis. *PLoS one*. 2017;12(9):e0183424.
- [11] K. Chao et al., Genetic ancestry, gnomAD browser, Broad Institute, Nov 2023. Available: <https://gnomad.broadinstitute.org/news/2023-11-genetic-ancestry/>
- [12] J.L. Asimit, A.P. Morris, Collapsing approaches for the association analysis of rare variants, in *Assessing Rare Variation in Complex Traits*, Springer, vol.10,

2016, pp.135–154. Available: https://link.springer.com/chapter/10.1007/978-1-4939-2824-8_10

- [13] J. Graffelman, V. Moreno, The mid p-value in exact tests for Hardy-Weinberg equilibrium, *Statistical Applications in Genetics and Molecular Biology*, vol.12, no.4, Dec 2013, pp.433–448. Available: <https://pubmed.ncbi.nlm.nih.gov/23934608/>
- [14] A. Recke, K.-G. Recke, S. Ibrahim, S. Möller, and R. Vonthein, “Hardy-Weinberg equilibrium revisited for inferences on genotypes featuring allele and copy-number variations,” *Scientific Reports*, vol. 5, p. 9066, Mar. 2015. Available: <https://pubmed.ncbi.nlm.nih.gov/25765626/>
- [15] Su SY, Asher JE, Jarvelin MR, Froguel P, Blakemore AIF, Balding DJ, et al. Inferring combined CNV/SNP haplotypes from genotype data. *BIOINFORMATICS*. 2010;26(11).
- [16] Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, et al. Adaptive Evolution of UGT2B17 Copy-Number Variation. *American journal of human genetics*. 2008;83(3):337–46.
- [17] Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Klein T, Leeder JS. Prediction of CYP2D6 phenotype from genotype across world populations. *Genetics in medicine*. 2017;19(1):69–76.
- [18] Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA. A multi-level model to address batch effects in copy number estimation using SNP arrays. *Biostatistics (Oxford, England)*. 2011;12(1):33–50.
- [19] Garant D. Principles of Population Genetics. *Écoscience (Sainte-Foy)*. 2007;14(4):544–5.
- [20] Pereira CAB, Nakano F, Stern JM, Whittle MR. Genuine Bayesian multiallelic significance test for the Hardy-Weinberg equilibrium law. *Genetics and molecular research*. 2006;5(4):619.
- [21] Asimit JL, Morris A. Collapsing Approaches for the Association Analysis of Rare Variants. In: Zeggini E, Morris A, editors. *Assessing Rare Variation in Complex Traits*. New York, NY: Springer New York; 2015. p. 135–148.
- [22] Dering C, Ziegler A, König IR, Hemmelmann C. Comparison of collapsing methods for the statistical analysis of rare variants. *BMC proceedings*. 2011;5(Suppl 9).
- [23] SciPy 1.11.0 Reference Guide. Fisher's exact test (`scipy.stats.fisher_exact`). SciPy Developers. Available from: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

- [24] Good P. Introduction to Resampling Techniques. Claremont Graduate University; 2016. Available from: <https://wise.cgu.edu/wp-content/uploads/2014/11/Introduction-to-Resampling-Techniques-160727.pdf>
- [25] Mussa Reshid T. Monte Carlo Simulation and Derivation of Chi-Square Statistics. *American Journal of Theoretical and Applied Statistics*. 2023; Available: <https://www.sciencepg.com/article/10.11648/j.ajtas.20231203.13>
- [26] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*. 2020;17(3):261–72.
- [27] Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. *Nucleic acids research*. 2022;50(D1):D988–95.
- [28] Kane M. CYP2D6 Overview: Allele and Phenotype Frequencies. 2021 Oct 15 [updated 2025 Jan 17]. In: Pratt VM, Scott SA, Pirmohamed M, et al., editors. *Medical Genetics Summaries* [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2012–. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK574601/>
- [29] Human Genome Variation Society (HGVS). HGVS Nomenclature Recommendations Summary [Internet]. Available from: <https://hgvs-nomenclature.org/stable/recommendations/summary/>
- [30] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature (London)*. 2015;526(7571):68–74.
- [31] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nature protocols*. 2010;5(9):1564–73.
- [32] Camara MD, Zhou Y, De Sousa TN, Gil JP, Djimde AA, Lauschke VM. Meta-analysis of the global distribution of clinically relevant CYP2C8 alleles and their inferred functional consequences. *Human genomics*. 2024;18(1).
- [33] KNMP Pharmacogenetics Working Group. CYP2C19 English – General background text Pharmacogenetics. The Hague: Royal Dutch Pharmacists Association (KNMP); 2021 Nov 16. Available from: https://www.knmp.nl/sites/default/files/2023-11/CYP2C19_English.pdf
- [34] KNMP Pharmacogenetics Working Group. CYP2C9 – General background text Pharmacogenetics. The Hague: Royal Dutch Pharmacists Association (KNMP); 2023 Nov. Available from: https://www.knmp.nl/sites/default/files/2023-11/CYP2C9_English.pdf

- [35] Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*. 2011;43(3):269–76.
- [36] Busti AJ. Pharmacogenetics: CYP1A2 Genetic Polymorphisms Table [Internet]. *EBM Consult*; 2015 Jun [cited 2025 Oct 22]. Available from: <https://www.ebmconsult.com/articles/pharmacogenetics-cyp1a2-genetic-polymorphisms-table>
- [37] Ghotbi R, Christensen M, Roh HK, Ingelman-Sundberg M, Aklillu E, Bertilsson L. Comparisons of CYP1A2 genetic polymorphisms, enzyme activity and the genotype-phenotype relationship in Swedes and Koreans. *European journal of clinical pharmacology*. 2007;63(6):537–46.