EVALUATING THE IMPACT OF QUALITY CONTROL FILTERS AND IMPUTATION REFERENCE PANELS ON PHARMACOGENOMIC VARIANT CALLING

MDSC 508 Thesis Proposal

Courtney Lenz

30077027

# BACKGROUND, RATIONALE, AND SIGNIFICANCE

As precision medicine transitions from theory to practice, the field of pharmacogenomics (PGx) facilitates this shift by exploring the genetic factors that influence drug response. PGx aims to improve drug safety and efficacy by investigating clinically actionable pharmacogenes. These genes play roles in pharmacokinetic functions, such as drug absorption, distribution, metabolism, and excretion (ADME), as well as in pharmacodynamic processes, such as the molecular interactions between a drug and its target.1 By identifying genotype-phenotype correlations for pharmacogenes, PGx research can be used to predict an individual’s response to a medication.1

One of the most powerful tools for detecting genotype-phenotype associations is genome-wide association studies (GWAS). GWAS scans for genetic variants across populations to identify statistical correlations with a given phenotype. Due to vast improvements in sequencing technology, single nucleotide polymorphism (SNP)-based genotyping arrays have become the preferred tool for use in GWAS as they are highly accurate and considerably more cost-effective than whole genome sequencing.3 Genome-wide genotyping arrays, also referred to as whole-genome arrays, provide more extensive variant coverage in comparison to traditional genotyping arrays.4

As PGx analysis using genome-wide genotyping arrays has become increasingly common, the Association for Molecular Pathology (AMP) has identified a minimum set of alleles (Tier 1) for PGx testing and an optional extended panel of alleles (Tier 2).5 Tier 1 alleles have a significant impact on protein and/or gene function resulting in a noticeable effect on drug response, have an appreciable minor allele frequency (MAF) in a given population, and have available reference materials.5,6 Tier 2 alleles only need to satisfy at least one of the characteristics defined above for Tier 1 alleles.5,6 Due to the well characterization and considerable frequency of these variants, they are commonly screened for in PGx testing; however, the majority of variants in pharmacogenes are rare (MAF < 1%).7,8,9 Calling PGx variants for genome-wide arrays can be complicated by applying standard QC filtering and selecting imputation reference panels.

Quality control (QC) measures are required when processing raw SNP data from genotyping experiments, as variant calling relies on the clustering of signals.10 Poor-quality data and noise signals can cause misclustering, which can result in false positive and false negative results (type I and type II errors, respectively).11 In the context of GWAS analysis, type I errors may lead to the incorrect identification of genetic associations, while type II errors may result in genuine associations being overlooked. To improve the accuracy of variant-calling from GWAS data, genome-wide analysis applications, such as PLINK, are used to filter out poor-quality data.

Following QC, imputation is commonly applied in GWAS analyses to “fill in” missing genotypic data, as SNP arrays have lesser genomic density in comparison to whole genome sequencing.12 The quality, size, and diversity of the reference panel directly influences the accuracy of imputation.12 Therefore, globally focused reference panels, such as the 1000 Genomes Project, Trans-Omics for Precision Medicine (TOPMed), and the Haplotype Reference Consortium (HRC) are often used for genomic imputation due to their robust size and diversity. 13,14,15 Currently, it is unclear how the aforementioned data processing methods impact the ability to call either well documented or rare PGx variants.

The progression of precision medicine depends on accurate genetic data interpretation; however, there lacks a recommended pipeline for processing GWAS data specifically for PGx applications. This the leaves potential for inaccuracies and errors in research results, which may pose risks when applying such research in a clinical setting. Therefore, this project seeks to refine and optimize QC and imputation processes specific to PGx, aiming to enhance the precision and reliability of GWAS research for use in personalized medicine applications.

# LITERATURE REVIEW

Effective QC for GWAS data requires strong genetics, statistics, and bioinformatics knowledge. To simplify this task, Turner et al. (2011) developed a protocol based on the standardized QC strategies established by the eMERGE (electronic MEdical Records and GEnomics).16 The filtering steps recommended by Turner et al. continue to be recommended in more recent GWAS analysis pipelines.17 Current standard QC procedure involves filtering out data based on the following factors:

1. Individual/SNP missingness: High levels of missing data for a particular individual or SNP often indicate poor quality or genotyping errors.16,17 Including this data could result in aberrant genotype calling.16
2. Sex discrepancies: Differences between the assigned sex and genotyped sex of an individual indicate potential sample handling errors.16,17
3. Minor allele frequency (MAF): The frequency at which the second most common allele occurs. SNPs with low MAFs are less reliable, and SNP-phenotype associations are unlikely to be detected. MAF thresholds of 0.01 and 0.05 are commonly used to exclude low frequency variants.16,17
4. Hardy-Weinberg equilibrium (HWE): Genetic variation in one population will remain constant from one generation to the next. Deviations from HWE are likely to be a result of genotyping errors; however, this may indicate evolutionary selection.16,17
5. Heterozygosity rate: High levels of heterozygosity may indicate low sample quality or contamination, while low levels may indicate inbreeding.17
6. Sample relatedness: Given that GWAS operates on the assumption of unrelated subjects, including related individuals without proper adjustment can lead to biased estimations.16,17
7. Population stratification: The presence of multiple subpopulations. Given that allele frequencies might differ among ethnic groups, not accounting for potential population-specific variations may introduce type I and type II errors.16,17

Genotype imputation is a statistical method that utilizes linkage disequilibrium (LD) to predict unobserved genotypes, which can be used to identify potential genetic associations that may otherwise be missed if relying solely on directly assayed variants.12 Linkage disequilibrium (LD) refers to the non-random association of alleles at different genetic loci within a population.18 This occurs because genetic variants are often inherited in groups due to their close proximity on a chromosome.12 Imputation algorithms can detect patterns of LD in a study sample by comparing the observed genetic variants with those in a reference panel.12 When a matching pattern is identified in the reference panel, the genotype of unobserved variants in LD with those observed in the study sample can be predicted.12

Although QC is required in GWAS to ensure the accuracy and relevance of identified associations, it inherently leads to data loss. QC methods that are too strict could result in the loss of valuable data and may introduce type II errors. Previous studies have found that excluding low-quality SNPs instead of assigning them a reduced quality score weight can diminish the power of locus-based methods if the causal variant is of good quality.19,20,21 In 2021, Charon et al. examined the potential consequences of QC filtering and the subsequent data loss on imputation quality.21 This study found that without pre-filtration, which excluded SNPs with MAFs < 0.01, imputation results were reliable, and the absence of pre-filtration improved imputation for all classes of MAFs.21 Additionally, raising the imputation quality score from 0.3 to 0.8 resulted in a 2.5-fold reduction of rare SNPs with a mean MAF < 0.001 and halved the number of very rare SNPs (MAF < 0.0005).21 Therefore, a less conservative approach to QC filtering was recommended to reduce information loss for rare and very rare variants.21

The accuracy of genotype imputation is impacted by two key factors: the reference panel and imputation algorithm.22 Historically, investigation into imputation accuracy has primarily focused on the algorithm.22 However, in a recent preprint, Li et al. (2023) examined the extent to which errors in the reference panel impact imputation performance.22 This study found that introducing perturbations into the reference panel decreased imputation accuracy.22 As MAF decreases, perturbation effects are amplified, while the imputation at common variants remains consistent regardless of the magnitude of the perturbation.22 Genotyping error rates tend to increase as MAF decreases, therefore, this study highlights the importance of accurate variant calls for rare variants within the reference panel.22

Considering that a majority of pharmacogenes are rare, the studies by Charon et al. (2021) and Li et al. (2023) would suggest that strict QC measures and errors in the reference panel are likely to impact the imputation accuracy of PGx variants.

# RESEARCH OBJECTIVES

#### MAIN OBJECTIVE

To evaluate the influence of quality control filters and imputation reference panels on pharmacogenomic variant calling.

#### RESEARCH QUESTION

Do commonly used quality control filters and imputation reference panels impact pharmacogenomic variant calling when processing genome-wide association array data?

#### SPECIFIC AIMS

1. Apply standard and modified quality control filtering to genome-wide association array data, followed by genotype imputation using commonly used reference panels.
2. Cross-reference pre-QC and post-QC imputed datasets with characterized pharmacogene variants.
3. Calculate the pharmacogenomic variant coverage for each dataset and assess intrasample concordance in diplotype and predicted phenotype across the datasets.
4. Perform statistical analyses to compare coverage and intrasample concordance differences between datasets.

#### HYPOTHESES

1. Standard quality control filters will negatively impact variant calling for some pharmacogenes, but not all.
2. Pharmacogenomic variant calling will differ depending on the reference panels used, resulting in intrasample diplotype and the predicted phenotype concordance.

# RESEARCH DESIGN & METHODS

#### METHODS

Dataset

Genome-wide association data from 804 participants enrolled in the CLOZapine INternational (CLOZIN) consortium ([www.clozinstudy.com](http://www.clozinstudy.com)), a multi-center study to detect associations of clozapine response and clozapine-associated adverse drug reactions using extensive phenotypic and genetic data. Details of the study set-up and inclusion criteria can be found elsewhere.23 All CLOZIN participants were genotyped using Illumina’s Infinium Global Screening Array V3.0. This dataset will be used in all of the subsequent processing detailed below.

Quality Control Filtering

PLINK, an open-source whole genome association analysis toolset, will be used to perform quality control filtering.24 Both standard and modified QC filtering methods will include the removal of data with high individual/SNP missingness, sex discrepancies, and closely related individuals. The standard QC filtering will also include the removal of variants that deviate from Hardy-Weinberg equilibrium, multi-allelic variants, and variants with MAF < 0.01. Three QC datasets will be prepared from each filtering method.

Imputation

For both the standard QC and modified QC categories, one dataset from each will be imputed using one of the following three reference panels: TOPMed, 1000 Genomes, and the Haplotype Reference Consortium (HRC).13,14,15 These imputation reference panels were chosen due to their common usage in pharmacogenomic studies.

Pharmacogene Variant Data

Characterized variants for 19 genes known to influence drug response phenotypes will be collected, primarily from the Pharmacogene Variation (PharmVar) Consortium and the Pharmacogenomics Knowledgebase (PharmGKB) databases.25,26 Haplotype tables for genes - *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A4, CYP3A5, CYP4F2, NUDT15,* and *SLCO1B1* will be collected from PharmVar.25 Allele definition tables for genes - *ABCG2, CFTR, DPYD, G6PD, RYR1, TPMT, UGT1A1,* and *VKORC1* will be collected from PharmGKB.26 Tag SNPs for *HLA* from the work of *Erlichster et al.* (2018) will be added to this dataset as traditional *HLA* typing is outside of the scope of this study.27

Analysis

The impact of quality control filtering and imputation reference panels on pharmacogenomic variant calling will be assessed through pharmacogenomic variant coverage and intrasample concordance.

Variant coverage will be assessed by cross-referencing the raw (pre-QC) and post-QC imputed datasets with the PharmVar/PharmGKB dataset to identify which variants can be detected in each dataset. Global variant coverage will be calculated by dividing the number of detected variants by the total number of variants in the PharmVar/PharmGKB dataset. A similar method will be applied to calculate the coverage of Tier 1 (essential panel) and Tier 2 (extended panel) pharmacogenomic variants, as defined by the Association of Molecular Pathology. The coverage differences between the different combinations of quality control and imputation methods will be statistically evaluated using two-sample Z tests. The Z-score will be converted to a p-value using a two-tailed Z-table. A P-value threshold of 0.05 will be used to determine significance.

Intrasample concordance will be assessed for diplotype and predicted phenotype (when applicable) using Cohen’s kappa coefficient to account for chance agreement. The level of agreement represented by the kappa value is as follows: ≤ 0 for none, 0.01-0.20 for slight, 0.21-0.40 for fair, 0.41-0.60 for moderate, and 0.81-1.00 for near perfect.28

SEX AND GENDER CONSIDERATIONS

Neither sex nor gender is anticipated to influence the impact of quality control filters or imputation reference panels on pharmacogenomic variant calling. All samples, regardless of the participant's sex, will be uniformly subjected to the study's protocol. Furthermore, for transparency and to support reproducibility in subsequent research, the sex of the participants will be reported in all publications resulting from this study.

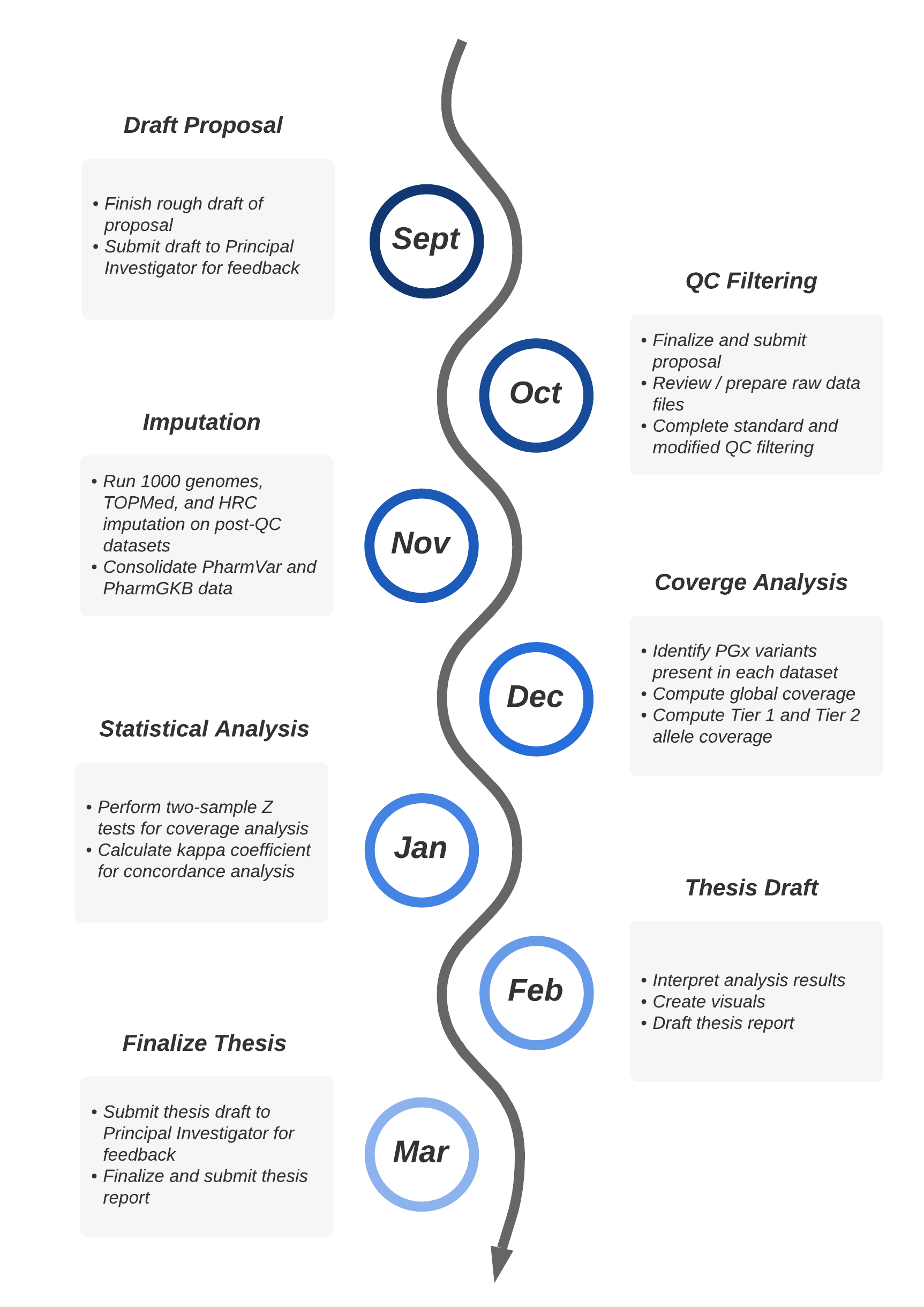
ALIGNMENT WITH MAJOR

The project highlights the interdisciplinary nature of bioinformatics, integrating principles from computer science, statistics, and molecular biology through the preparation and analysis of large genomic datasets. To successfully evaluate the impact of QC filtering and imputation reference panels on PGx variant calling, a comprehensive understanding of genetics/genomics, data processing and management, imputation, computation analysis and statistical interpretation is required. Specifically, this project focuses on tertiary bioinformatics analysis, in which genomic data, once sequenced and aligned, is translated into meaningful/actionable insights.

# KNOWLEDGE TRANSLATION

This novel research systematically evaluates the effects of GWAS quality control filters and imputation reference panels on PGx variant calling. The findings will inform quality control and imputation methods for subsequent PGx studies using genome-wide association arrays, helping to streamline research design and improve PGx outcomes. Results will be shared with CLOZIN consortium members and submitted for publication in a peer-reviewed journal to ensure dissemination of findings to the broader scientific community.

# APPENDIX



**Figure 1: Timeline for Thesis Project.** Proposed seven-month plan to evaluate the impact of quality control filters and imputation reference panels on pharmacogenomic variant calling.

# REFERENCES

1. Adams, J. Pharmacogenomics and Personalized Medicine. Nat. Edu. 2008. 1(1):194
2. McInnes G, Yee SW, Pershad Y, Altman RB. Genomewide Association Studies in Pharmacogenomics. Clin Pharmacol Ther. 2021 Sep;110(3):637–48.
3. Zhao S, Jiang L, Yu H, Guo Y. GTQC: Automated Genotyping Array Quality Control and Report. J Genomics. 2022 Feb 14;10:39–44.
4. Bejjani BA, Shaffer LG. Application of Array-Based Comparative Genomic Hybridization to Clinical Diagnostics. J Mol Diagn. 2006 Nov;8(5):528–33.
5. PharmGKB. AMP’s Minimum Sets of Alleles for PGx Testing. [Internet]. PharmGKB. [cited 2023 Sep 30]. Available from: <https://www.pharmgkb.org/ampAllelesToTest>
6. Pratt VM, Cavallari LH, Del Tredici AL, et al. Recommendations for Clinical CYP2C9 Genotyping Allele Selection. J Mol Diagn. 2019;21(5):746-55.
7. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. Genetics in Medicine. 2017 Jan 1;19(1):20–9.
8. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 2012 Jul 6;337(6090):100–4.
9. Lakiotaki K, Kanterakis A, Kartsaki E, Katsila T, Patrinos GP, Potamias G. Exploring public genomics data for population pharmacogenomics. PLOS ONE. 2017 Aug 3;12(8):e0182138.
10. Lamy P, Andersen CL, Wikman FP, Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. Nucleic Acids Res. 2006;34(14):e100.
11. Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E, D’Agostino N. Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies. Front Genet. 2020 Jun 5;11:447.
12. Li Y, Willer C, Sanna S, Abecasis G. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
13. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.
14. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021 Feb;590(7845):290–9.
15. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.
16. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality Control Procedures for Genome-Wide Association Studies. Current Protocols in Human Genetics. 2011;68(1):1.19.1-1.19.18.
17. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie‐Claire C, et al. A tutorial on conducting genome‐wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Feb 27;27(2):e1608.
18. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun;9(6):477–85.
19. Li JH, Liu A, Buerkle CA, Palmer W, Belbin GM, Ahangari M, et al. The effects of reference panel perturbations on the accuracy of genotype imputation. bioRxiv [Internet]. 2023 [cited 2023 Oct 12]. Available from: <https://www.biorxiv.org/content/10.1101/2023.08.10.552684v1>
20. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. Hum Mol Genet. 2013 Oct 15;22(R1):R16–21.
21. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: Testing for an Accumulation of Rare Variants Using Next-Generation Sequencing Data. Hum Hered. 2012;73(2):84–94.
22. Charon C, Allodji R, Meyer V, Deleuze JF. Impact of pre- and post-variant filtration strategies on imputation. Sci Rep. 2021 Mar 18;11:6214.
23. Okhuijsen-Pfeifer C, van der Horst MZ, Bousman CA, Lin B, van Eijk KR, Ripke S, et al. Genome-wide association analyses of symptom severity among clozapine-treated patients with schizophrenia spectrum disorders. Transl Psychiatry. 2022 Apr 7;12:145.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
25. Gaedigk A, Whirl-Carrillo M, Pratt VM, Miller NA, Klein TE. PharmVar and the Landscape of Pharmacogenetic Resources. Clin Pharmacol Ther. 2020 Jan;107(1):43–6.
26. Whirl-Carrillo M, McDonagh E, Hebert J, Gong L, Sangkuhl K, Thorn C, et al. Pharmacogenomics Knowledge for Personalized Medicine. Clin Pharmacol Ther. 2012 Oct;92(4):414–7.
27. Erlichster M, Goudey B, Skafidas E, Kwan P. Cross-ethnicity tagging SNPs for HLA alleles associated with adverse drug reaction. Pharmacogenomics J. 2019 Jun;19(3):230–9.
28. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012 Oct 15;22(3):276–82.