Supplementary Information

Pharmacokinetic recall study of Estonian Biobank participants carrying novel genetic variants in *CYP2C19* and *CYP2D6*

Authors:

Kristi Krebs[1*], Laura Birgit Luitva[1,2*], Anette Caroline Kõre[3], Raul Kokasaar[4#], Maarja Jõeloo[1], Georgi Hudjashov[1], Kadri Maal[1], Elisabet Størset[5,6], Birgit Malene Wollmann[5], Liis Karo-Astover[1], Krista Fischer[1,2], Estonian Biobank Research Team[1], Volker M Lauschke[7,8,9,10], Magnus Ingelman-Sundberg[7], Espen Molden[5,6], Alar Irs[3], Kersti Oselin[4], Jana Lass[1,3,11‡] and Lili Milani[1‡]

1) Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
2) Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia
3) Tartu University Hospital, Tartu, Estonia
4) Clinic of Oncology and Hematology, North Estonia Medical Center, Tallinn, Estonia
5) Center for Psychopharmacology, Diakonhjemmet Hospital, Oslo, Norway
6) Department of Pharmacy, University of Oslo, Oslo, Norway
7) Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden
8) Dr Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany
9) University of Tübingen, Tübingen, Germany
10) Department of Pharmacy, the Second Xiangya Hospital, Central South University, Changsha, China
11) Institute of Pharmacy, University of Tartu, Tartu, Estonia

* These authors contributed equally as first authors.

‡ These authors contributed equally as last authors.

Current affiliation: West Tallinn Central Hospital, Tallinn, Estonia

Corresponding author E-mail: lili.milani@ut.ee (LM)

Banner author contributors

Estonian Biobank Research Team: Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mait Metspalu, Mari Nelis and Georgi Hudjashov

Table of Contents

Supplementary Note

CYP2C19 Deletion Frequency in the Estonian Biobank

To estimate the frequency of the *CYP2C19* partial gene deletion (*CYP2C19*37*) in the Estonian Biobank (EstBB), deletion carriers were identified using PennCNV[1] applied to genotyping data from the Illumina Global Screening Array (GSA) in 17 batches. Duplicates and samples with call-rate <0.95 were excluded. We only considered deletion calls that (i) fell into the boundaries of an established 61.8k deletion overlapping CYP2C19 exons 1 to 5 (gnomAD structural variants v4.1.0 variant ID: DEL_CHR10_28B50744)[2], and (ii) were at least 5k base pairs long. All other CYP2C19-overlapping deletions were flagged as ambiguous.

Out of a total of 211,299 individuals with genotyping data, 3,859 were classified as deletion carriers, while 204,393 were non-carriers. We excluded 3,047 individuals where the CNV status could not be reliably determined due to ambiguous signals. Among the 208,252 individuals with definitive CNV calls, the estimated frequency of the *CYP2C19*37* partial deletion was 1.9%, suggesting that this structural variant is more prevalent in the Estonian population than previously recognized[2].

Post-recruitment star allele re-assignment with new tools

Since pharmacogenetic star allele calling tools keep evolving, we performed additional star allele assignments for a subset of participants (n = 43) who had short-read whole-genome sequencing data available. We determined star allele diplotypes using two specialised freely available computational tools: Cyrius v1.1.1[3] and Aldy v4.5[4], using default parameters and the GRCh38 reference genome in the calling process.

The diplotype results from these tools were compared against two benchmarks: the UT-tool[5], which was used for star allele calling during the recruitment phase, and pb-StarPhase[6], which we consider the analytical gold standard in this study due to its integration of long-read sequencing, superior phasing, and *CYP2D6-D7* specific reference sequences for accurate alignment and star allele calling[6].

For *CYP2C19*, we expanded the allele calling to include all 114 participants by applying the PharmCAT algorithm[7] on phased genotype data derived from both microarrays and sequencing. The obtained *CYP2C19* star allele assignments were

then systematically compared with the calls produced by the UT-tool and pb-StarPhase to assess general concordance.

All concordance analyses and comparative evaluations were performed using R (version 4.4.3)[8], with custom scripts developed to calculate match rates.

Star Allele Diplotype Assignment and Concordance Analysis

Diplotypes assigned with the pb-StarPhase algorithm[6] were used for all downstream analyses. To assess the similarities of *CYP2D6* diplotypes derived from short-read sequencing data, we compared the diplotype calls obtained using the Cyrius[3] and Aldy[4] tools against the results from pb-StarPhase in a subset of 43 participants (with available genome sequencing data). Both short-read-based tools performed well, with Cyrius demonstrating a concordance rate of 93.0% and Aldy achieving 90.7% when compared to the long-read-based pb-StarPhase calls (Supplementary Figure 2, Supplementary Table 4). The observed mismatches were primarily linked to hybrid allele classifications, highlighting the advantage of pb-StarPhase's long-read approach in resolving structural complexities. The previously used UT-tool[5] exhibited a lower concordance rate of 83.7%, largely due to its limitations in identifying hybrid alleles.

For the *CYP2C19* gene, we also compared the pb-StarPhase results with diplotype calls generated using the UT-tool and the PharmCAT[7] algorithm across all 114 participants (GS+microarray). The overall concordance was notably low, primarily because neither the UT-tool nor PharmCAT could call structural variants for *CYP2C19*. In particular, the partial gene deletion allele *CYP2C19*37* was uniquely identified by pb-StarPhase, leading to significant mismatches with both the UT-tool (concordance: 28.1%) and PharmCAT (concordance: 41.2%, as shown in Supplementary Figure 3, Supplementary Table 4). Furthermore, the more recently characterized CYP2C19*38 allele was not included in UT-tool's original allele database, which contributed considerably to its lower concordance rate.

Long-read sequencing

Genomic DNA extracted from peripheral blood samples was sequenced using PacBio Revio sequencing technology to generate highly accurate circular consensus HiFi

(High-Fidelity) reads. Library preparation and sequencing were performed according to the manufacturer's standard protocols. All samples (n=112) were sequenced at an aimed coverage of 20X (mean=23.7, median=21.7). For each sample, we required a minimum of 57.5 Gbs of raw unmapped sequence to be processed further. HiFi reads were aligned to the human reference genome (GRCh38/hg38) using pbmm2 (v1.17.0). Single nucleotide variants (SNVs) and small insertions/deletions (indels) were called using DeepVariant (v1.6.1) with the PacBio HiFi-specific model. Haplotype phasing was carried out using HiPhase[9] (v1.4.5). SNVs and indels were functionally annotated using the Ensembl Variant Effect Predictor (VEP, v112), with plugins including dbSNP and gnomAD for functional classification and population frequency assessment. For structural variant detection, we employed sawfish (v0.12.10)[10], a tool tailored for sensitive detection of deletions, duplications, and complex rearrangements from long-read data.

Genome-wide screen highlights *CYP2C19* and *CYP2D6* primary contribution

To assess whether genomic regions beyond *CYP2C19* and *CYP2D6* affect the variability of drug metabolism, we conducted genome-wide association analyses (GWAS) using the metabolic ratios of omeprazole and metoprolol.

For omeprazole, the GWAS revealed a genome-wide significant peak on chromosome 10, overlapping with the *CYP2C19* locus (Supplementary Figure 7, Supplementary Table 11). The lead variant is in the CYP2C locus (rs71482318, $P=1.7 \times 10^{-11}$), consistent with prior knowledge that CYP2C19 is the primary enzyme responsible for omeprazole metabolism. This intronic lead variant is not part of any star allele but is in complete linkage disequilibrium (LD) ($r^2=1$) with rs12769205 and rs4244285, which define the poor metabolizer *CYP2C19*2* allele. No other loci neared genome-wide significance ($P < 5 \times 10^{-8}$), indicating no strong evidence for other genomic regions contributing to omeprazole metabolic variability in this dataset. To assess whether the observed association was specifically driven by the *2 allele, we performed a conditional analysis including rs4244285 as a covariate (in LD with rs12769205). As expected, the chromosome 10 signal at the *CYP2C19* locus was markedly attenuated (data not shown), confirming that the association was largely attributable to the *2-defining variants.

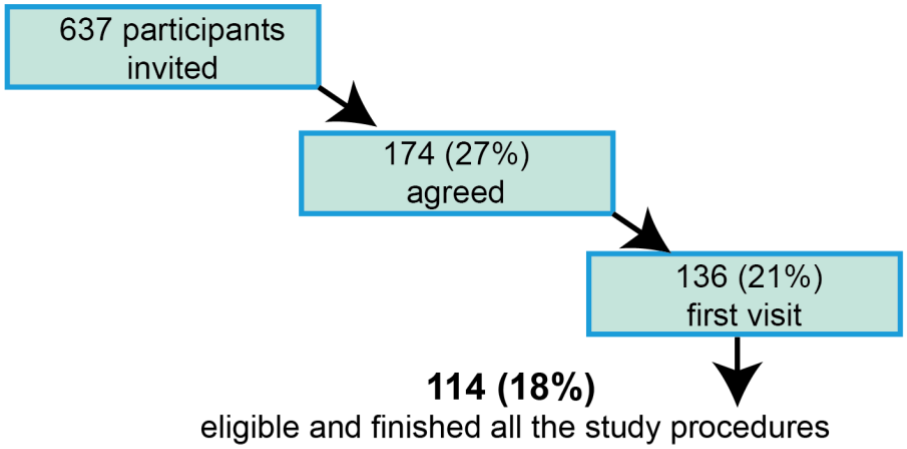
For metoprolol, the GWAS identified a prominent association peak on chromosome 22, next to the *CYP2D6* locus (Supplementary Figure 7, Supplementary Table 11). Six lead variants with the same p-value ($P=6.7 \times 10^{-8}$) are in *WBP2NL* or *SEPTIN3* that are adjacent to the *CYP2D6* gene. In addition, two genome-wide significant loci were identified (Supplementary Figure 7): one on chromosome 10, with the lead variant in *LRMDA* (rs182557066, $P=9.5 \times 10^{-9}$) and another on chromosome 7, with the lead variant in *AKAP9* (rs188755628, $P=1.0 \times 10^{-8}$). Both genes have been reported previously to be linked most significantly to atrial fibrillation and flutter in pooled biobank GWAS (mvp-ukbb.finngen.fi: $P_{LRMDA}=7.9 \times 10^{-22}$ and $P_{AKAP9}=1.1 \times 10^{-11}$). To determine whether the chromosome 22 signal was driven by known functional variants of *CYP2D6*, we performed a conditional analysis including rs3892097, the splice site variant that defines the loss-of-function allele *CYP2D6*4*, as a covariate. After conditioning, the association signal on chromosome 22, as well as the ones on chromosomes 10 and 7, were fully attenuated (data not shown), indicating that the observed signal is largely attributable to the *4 allele. These findings reaffirm the central role of *CYP2D6* in metoprolol metabolism, consistent with previous results[11].

References for the supplementary note

1. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
2. Botton, M. R. *et al.* Structural variation at the CYP2C locus: Characterization of deletion and duplication alleles. *Hum. Mutat.* **40**, e37–e51 (2019).
3. Chen, X. *et al.* Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. *Pharmacogenomics J.* **21**, 251–261 (2021).
4. Hari, A. *et al.* An efficient genotyper and star-allele caller for pharmacogenomics. *Genome Res.* **33**, 61–70 (2023).
5. Reisberg, S. *et al.* Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **21**, 1345–1354 (2019).
6. Holt, J. M. *et al.* StarPhase: Comprehensive Phase-Aware Pharmacogenomic Diplotyper for Long-Read Sequencing Data. 2024.12.10.627527 Preprint at https://doi.org/10.1101/2024.12.10.627527 (2024).
7. Sangkuhl, K. *et al.* Pharmacogenomics Clinical Annotation Tool (PharmCAT). *Clin. Pharmacol. Ther.* **107**, 203–210 (2020).
8. R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

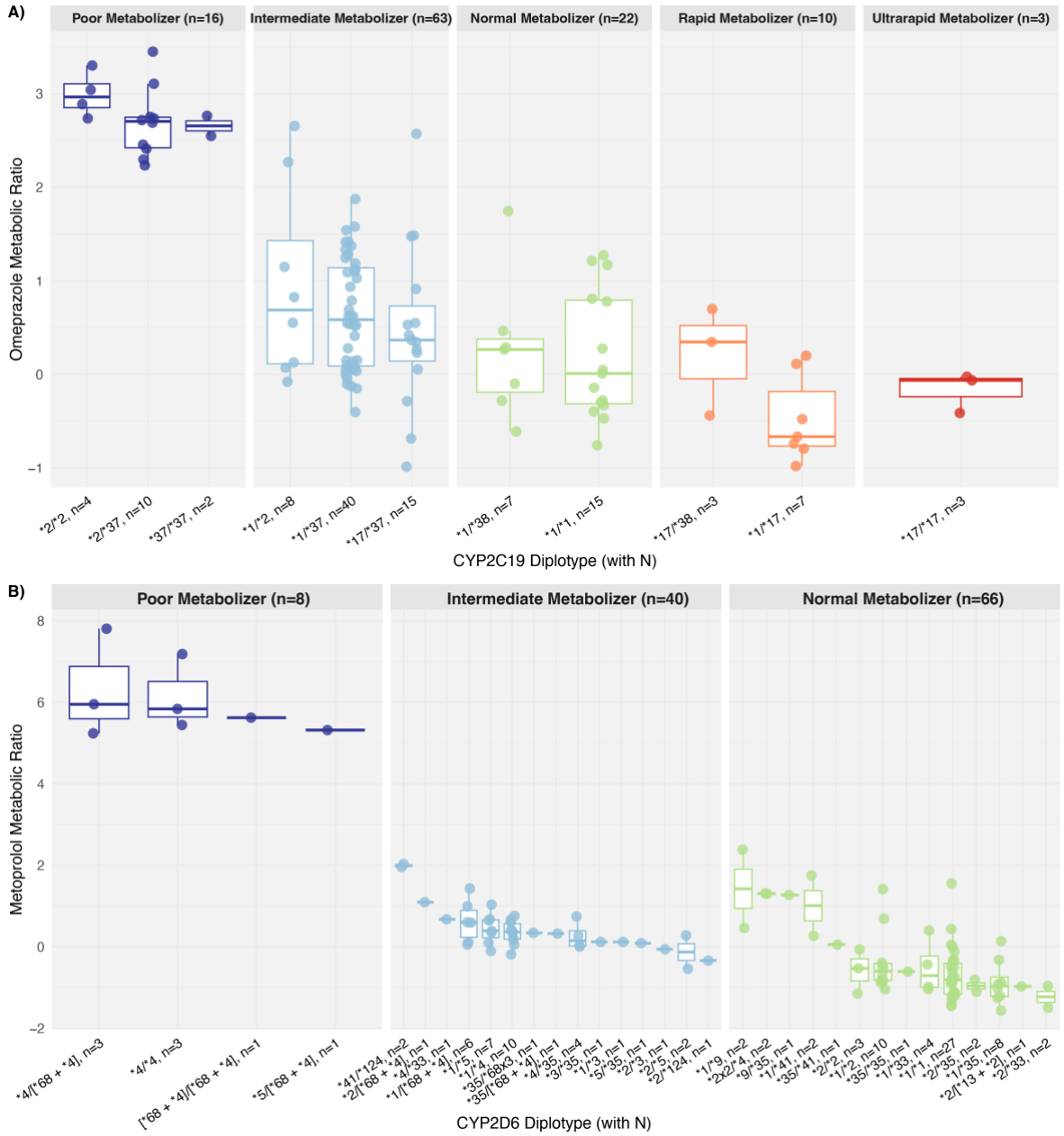
9. Holt, J. M. *et al.* HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinforma. Oxf. Engl.* **40**, btae042 (2024).
10. Saunders, C. T. *et al.* Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. *Bioinformatics* **41**, btaf136 (2025).
11. Laverdière, J. *et al.* Pharmacogenomic markers of metoprolol and α-OH-metoprolol concentrations: a genome-wide association study. *Pharmacogenomics* **24**, 441–448 (2023).

Supplementary Figures

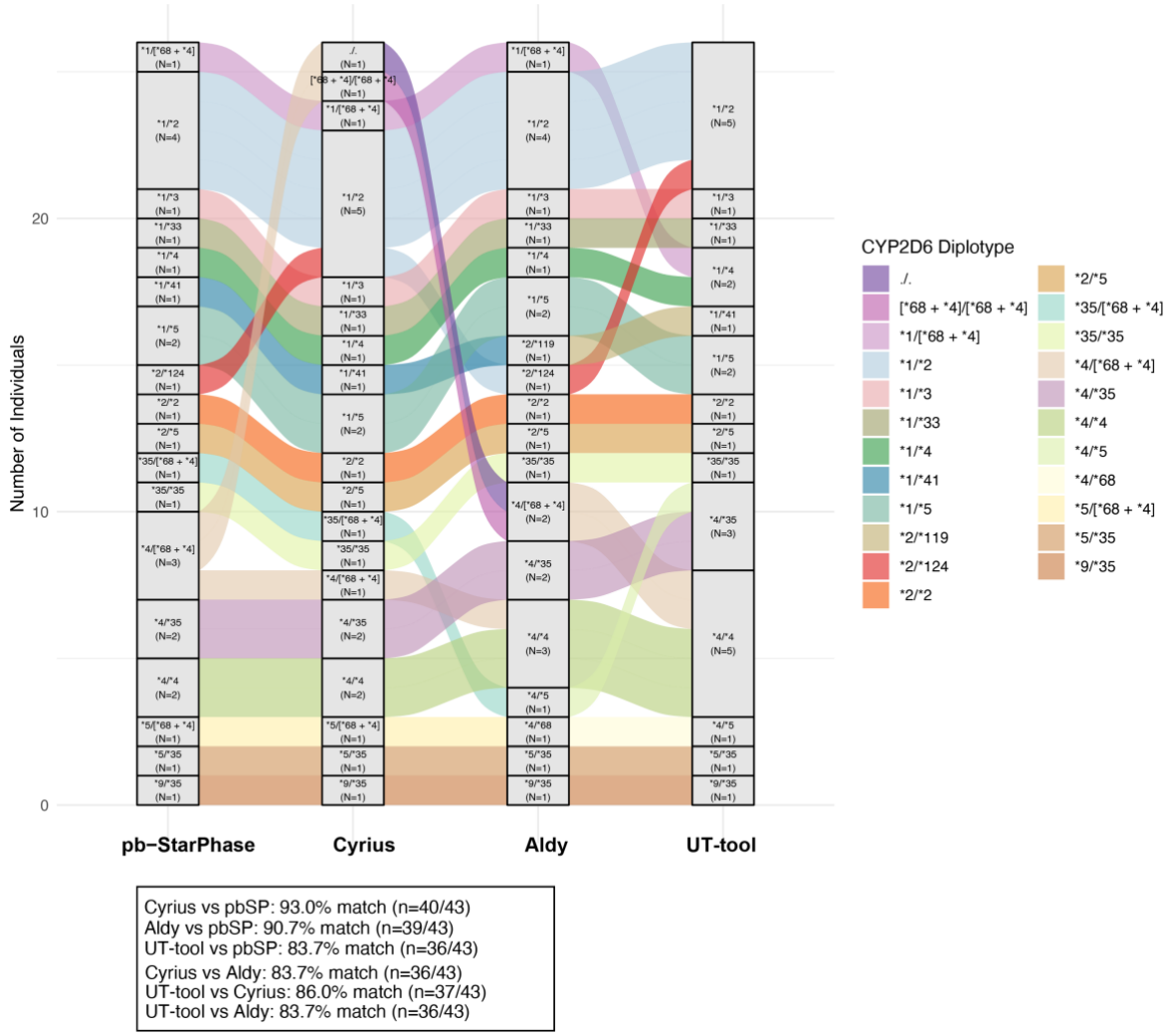


	Female (N=72)	Male (N=42)
Age, mean (SD)	50.39 (12.76)	52.93 (12.90)
Weight, mean (SD)	74.08 (15.11)	88.26 (15.26)
SBP, mean (SD)	129.74 (16.78)	138.00 (15.16)
DBP, mean (SD)	85.46 (10.33)	86.98 (9.82)
Previous diseases, N (%)	37 (51.4)	21 (50.0)
Operations, N (%)	29 (40.8)	10 (23.8)
Drugs used 30 days, N (%)	36 (50.7)	20 (47.6)
White Blood Cells (10^9/L), mean (SD)	7.08 (1.78)	7.29 (2.07)
Hemoglobin (g/L), mean (SD)	135.21 (8.78)	151.07 (9.41)
Hematocrit (%), mean (SD)	40.44 (2.78)	44.52 (2.38)
Platelets (10^9/L), mean (SD)	267.93 (55.43)	221.93 (46.97)
Urea (mmol/L), mean (SD)	8.76 (30.03)	5.91 (1.16)
Creatinine (µmol/L), mean (SD)	67.68 (10.53)	93.56 (15.06)
eGFR (mL/min/1.73m²), mean (SD)	91.46 (16.51)	81.29 (16.22)
Sodium (Na, mmol/L), mean (SD)	140.03 (2.41)	140.29 (1.87)
Potassium (K, mmol/L),mean (SD)	4.21 (0.35)	4.20 (0.32)
Calcium (Ca, mmol/L), mean (SD)	2.39 (0.10)	2.36 (0.09)
Glucose (mmol/L), mean (SD)	5.60 (0.84)	5.86 (2.47)
ALAT (U/L), mean (SD)	22.41 (13.17)	27.68 (11.30)
ASAT (U/L), mean (SD)	23.70 (9.17)	24.38 (4.95)

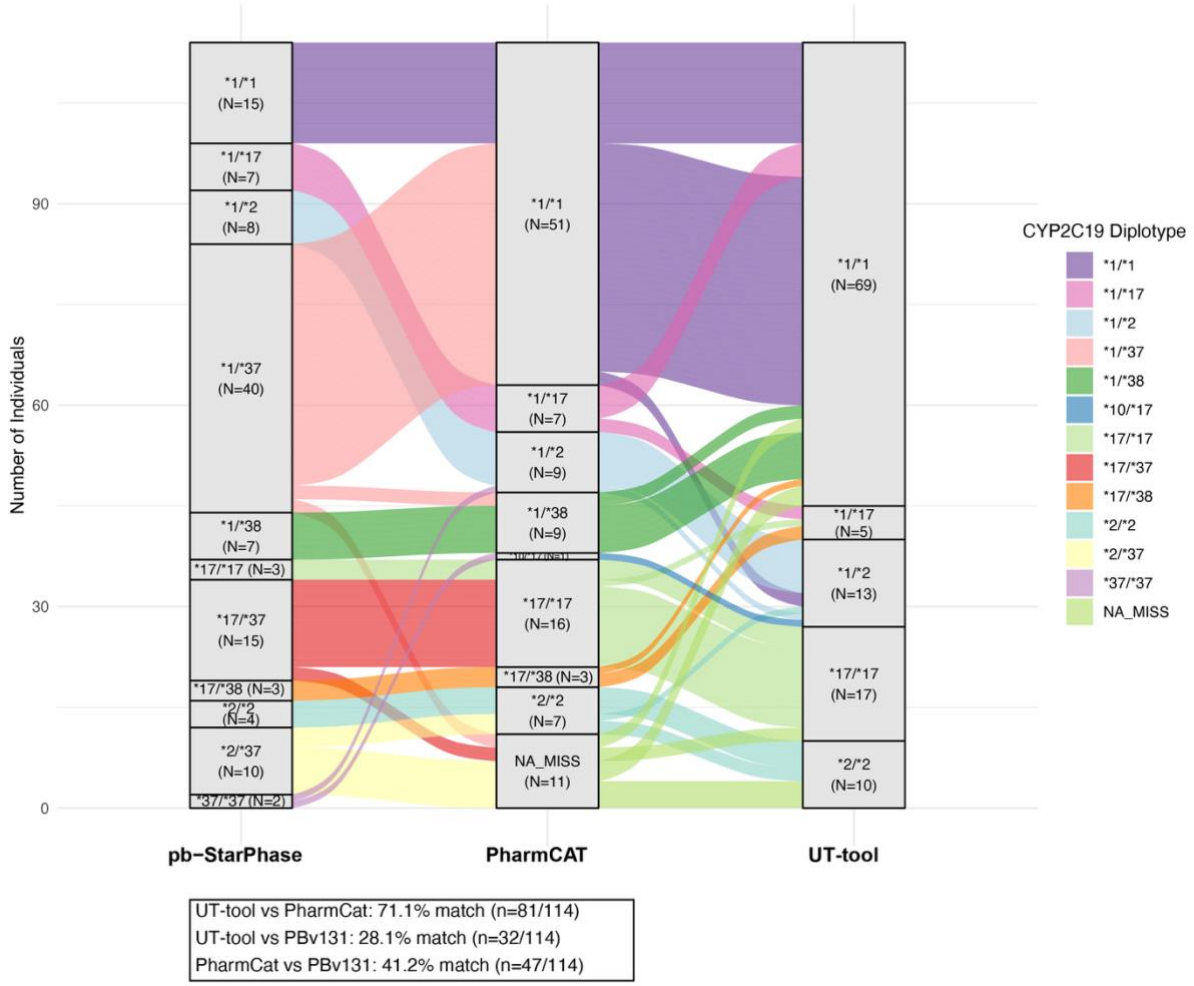
Supplementary Figure 1. Study enrolment flow and the characteristics of final participants. The table summarises the clinical characteristics of all participants who completed the study.



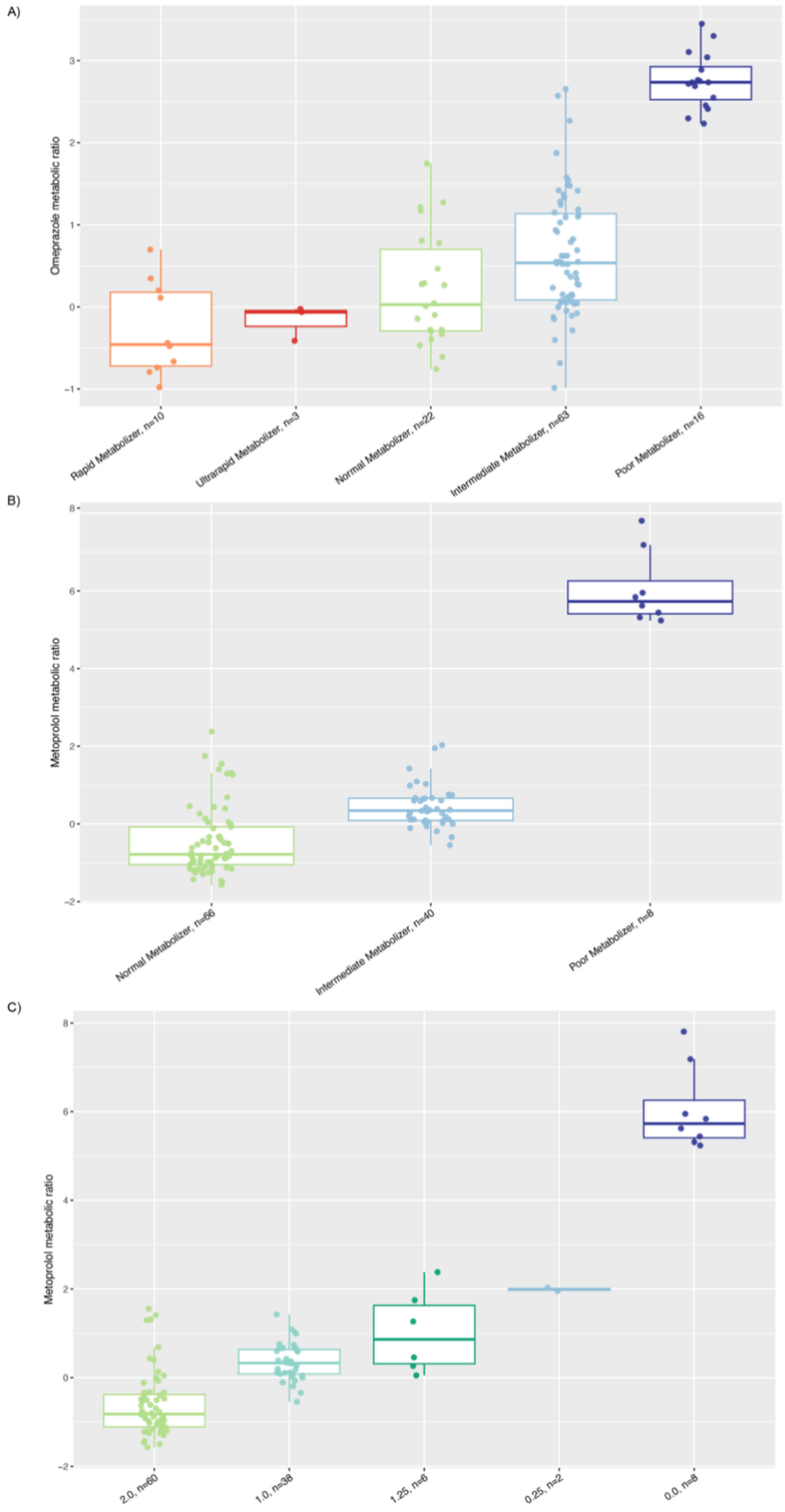
Supplementary Figure 2. Probe-drug metabolic ratios across star allele diplotypes for (A) CYP2C19 and (B) CYP2D6. The x-axis shows diplotypes (with the number of individuals per group), and the y-axis shows the metabolic ratios of omeprazole for CYP2C19 and metoprolol for CYP2D6. Genotypes are coloured by predicted metaboliser phenotype: dark blue for poor metabolisers, light blue for intermediate metabolisers, orange for rapid metabolisers, and red for ultrarapid metabolisers.



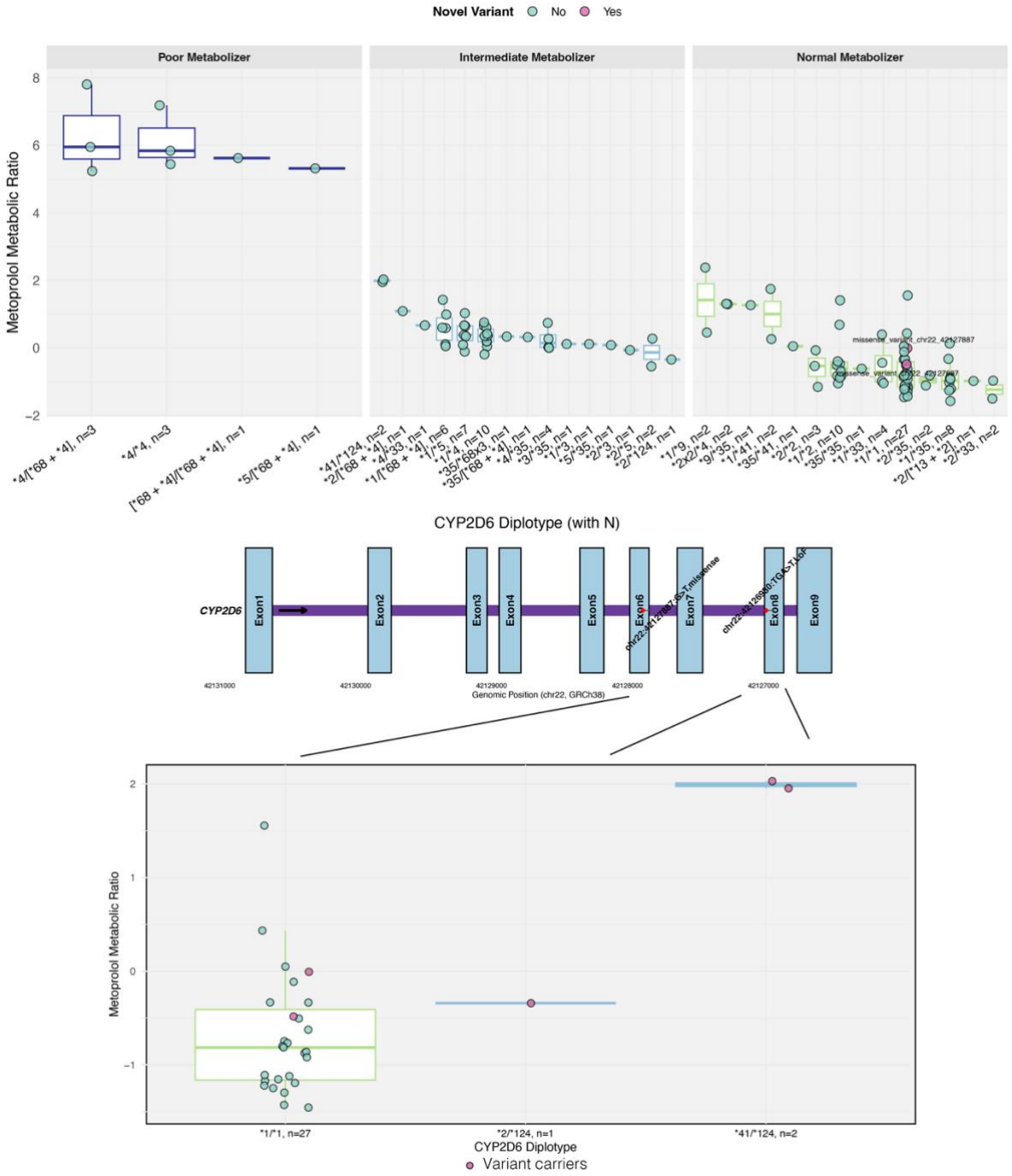
Supplementary Figure 3. Concordance of CYP2D6 star allele calls across four tools. An alluvial plot showing differences in star allele diplotype calls in a subset of 43 participants with short-read genome sequencing data, comparing PacBio StarPhase (pb-StarPhase), Cyrius, Aldy, and the UT-tool. The y-axis indicates the number of individuals, and different colours represent diplotypes. The width of the connecting lines reflects the number of individuals. For clarity, matching *1/*1 calls were excluded from the plot.



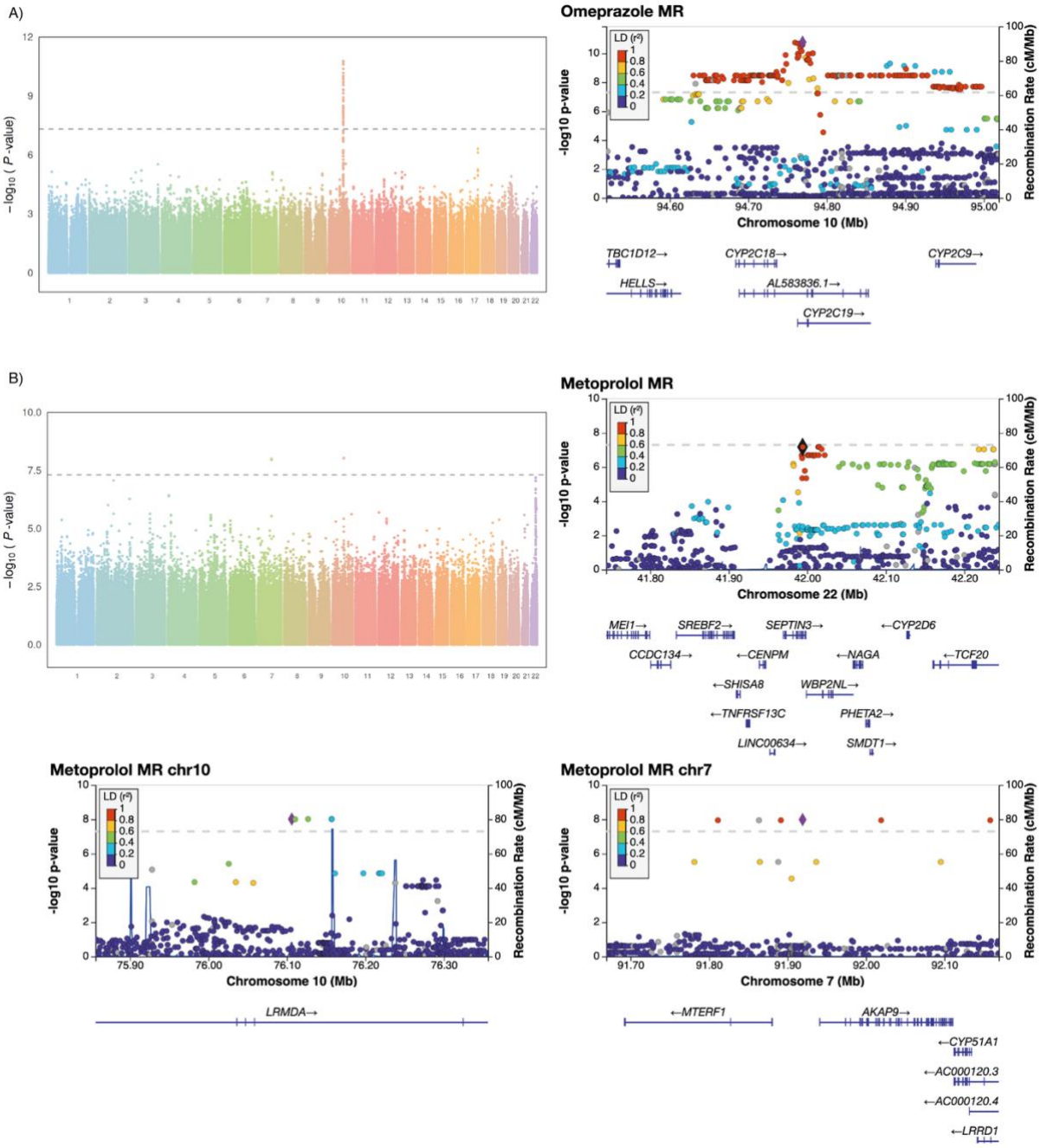
Supplementary Figure 4. Concordance of CYP2C19 star allele calls across three tools. An alluvial plot showing how star allele diplotype calls (n = 114) vary between PacBio StarPhase (pb-StarPhase), PharmCAT, and UT-tool. The y-axis indicates the number of individuals, and different colours represent diplotypes. The width of the connecting lines reflects the number of individuals shared between categories.



Supplementary Figure 5. Probe-drug metabolic ratios across metaboliser phenotypes for (A) CYP2C19 and (B) CYP2D6. The x-axis groups individuals by predicted metaboliser phenotype (A, B) and CYP2D6 activity score (C). The y-axis shows metabolic ratios of omeprazole (A) and metoprolol (B). Colour coding corresponds to phenotype: dark blue for poor metabolisers, light blue for intermediate metabolisers, orange for rapid metabolisers, and red for ultrarapid metabolisers.



Supplementary Figure 6. Carriers of novel CYP2D6 variants by star allele genotype and metabolic ratio. Carriers of novel variants are shown as pink dots; non-carriers are shown as blue dots. The x-axis indicates diplotypes and the number of individuals per group, while the y-axis shows metoprolol metabolic ratio. Text labels on dots indicate the variant positions and predicted functional consequences. Furthermore, a schematic representation of the CYP2D6 gene, with the exons highlighted in blue (coordinates from Ensembl), illustrates the locations of the novel missense variant and the *124 loss-of-function variant.



Supplementary Figure 7. Genome-wide association analysis of omeprazole and metoprolol metabolic ratios. Manhattan plots for omeprazole (A) and metoprolol (B) metabolic ratios are shown on the left. The y-axis represents −log10(P) for the association of SNVs, the horizontal dashed line indicates the genome-wide significance threshold ($P<5\times10^{-8}$). The genomic inflation factor (λ) was 1.0 for omeprazole (A) and 0.9 for metoprolol (B). On the right (and below for metoprolol), corresponding regional association plots display the genomic loci surrounding the top-associated variants. The x-axis shows the genomic position (in megabases, Mb), and the y-axis indicates statistical significance ($-\log_{10}(P)$). The purple diamond marks the most significantly associated SNP within each locus. SNPs are colour-coded according to their linkage disequilibrium (LD, r^2) with the lead SNP, based on data from the European population in the 1000 Genomes Project.