Population-scale predictions of DPD and TPMT phenotypes using a quantitative pharmacogene-specific ensemble classifier

Yitian Zhou, Carolina Dagli Hernandez and Volker M. Lauschke

BACKGROUND: Inter-individual differences in dihydropyrimidine dehydrogenase (DPYD encoding DPD) and thiopurine S-methyltransferase (TPMT) activity are important predictors for fluoropyrimidine and thiopurine toxicity. While several variants in these genes are known to decrease enzyme activities, many additional genetic variations with unclear functional consequences have been identified, complicating informed clinical decision-making in the respective carriers.

METHODS: We used a novel pharmacogenetically trained ensemble classifier to analyse DPYD and TPMT genetic variability based on sequencing data from 138,442 individuals across eight populations.

RESULTS: The algorithm accurately predicted in vivo consequences of DPYD and TPMT variants (accuracy 91.4% compared to 95.3% in vitro). Further analysis showed high genetic complexity of DPD deficiency, advocating for sequencing-based DPYD profiling, whereas genotyping of four variants in TPMT was sufficient to explain >95% of phenotypic TPMT variability. Lastly, we provided population-scale profiles of ethnogeographic variability in DPD and TPMT phenotypes, and revealed striking interethnic differences in frequency and genetic constitution of DPD and TPMT deficiency.

CONCLUSION: These results provide the most comprehensive data set of DPYD and TPMT variability published to date with important implications for population-adjusted genetic profiling strategies of fluoropyrimidine and thiopurine risk factors and precision public health.

British Journal of Cancer (2020) 123:1782–1789; https://doi.org/10.1038/s41416-020-01084-0

BACKGROUND

Adverse drug reactions (ADRs) are a common phenomenon in cancer therapy, and the identification of patients at increased risk thus constitutes an important goal of precision oncology. In the last decade, genetic profiling has identified a multitude of variations that can guide selection and dosing of chemotherapeutic drugs. Two of the most important examples of such pharmacogenetic biomarkers that have transitioned from research into clinical practice are germline variations in the dihydropyrimidine dehydrogenase (DPYD encoding DPD) and thiopurine S-methyltransferase (TPMT) genes.1–4

Fluoropyrimidines are cornerstones of oncological therapy used for the treatment of a wide range of solid tumours. Importantly, DPD deficiency is strongly associated with dose-limiting and sometimes life-threatening toxicity with 60–80% of DPD-deficient individuals experiencing severe ADRs compared to 10–20% of patients with normal enzyme function.5,6 The most extensively studied variation associated with DPD deficiency is DPYD*2A (rs3918290), a splice donor variant that results in truncated protein without catalytic activity.7 Recent meta-analyses moreover confirmed robust associations of DPYD I560S, D949V as well as of the intronic splice variant rs75017182 and the associated haplotype HapB3 with fluoropyrimidine toxicity.8–10 and prospective testing for these variants followed by genotype-guided upfront dose adjustments significantly increased patient safety.11–13

Analogously to DPYD, individuals deficient in TPMT are more susceptible to life-threatening toxicity of thiopurines.14 The most important decreased function alleles are TPMT*2 (rs1800462), *3A (rs1800460 and rs1142345) and *3C (rs1142345).15

In addition to the well-characterised variants illustrated above, DPYD and TPMT harbour hundreds of additional rare genetic variations with unclear effects on enzyme function.16,17 Recent advances in large-scale mutagenesis screens unlock exciting opportunities for the parallel experimental interrogation of the effect of thousands of variants,18 as exemplified by the simultaneous characterisation of the effects of thousands of TPMT variants on intracellular abundance.19 However, without experimental assessments of variant effects on enzyme activity, their interpretation has to rely on computational tools. In the last two decades, a multitude of computational prediction tools have been developed that consider sequence conservation as an indicator of variant deleteriousness, as well as various mechanistic parameters, such as impacts on physiochemical properties, post-translational modifications and structural features, such as protein stability and the disruption of binding interfaces.20,21

These algorithms are mostly trained on pathogenic variants for which evolutionary conservation constitutes a suitable proxy. However, evolutionary constraints for DPYD and TPMT are limited, and conservation scores are thus not the ideal metric to predict variant function. To overcome these problems,
we recently established a computational framework optimised specifically for pharmacogenomic predictions.\textsuperscript{23}

Using this algorithm, we here mapped\textit{DPYD} and\textit{TPMT} variants on an unprecedented scale using whole-genome sequencing (WGS) and whole-exome sequencing (WES) data from 138,842 individuals. Using all variations with known in vivo consequences as benchmark data set, we show that our pharmacogenetically trained ensemble classifier substantially outperforms all previous non-gene-specific prediction methods and achieved predictive accuracy similar to that in in vitro experiments (accuracy 91.4\% compared to 95.3\% for in vitro). Encouraged by these results, we applied the algorithm to the functional interpretation of all 859 identified\textit{DPYD} and\textit{TPMT} variants that affect the amino acid sequence of their respective gene products. Our results reveal that interrogation of only four variants is sufficient to identify >95\% of\textit{TPMT}-decreased function alleles, whereas 174 variations have to be profiled for\textit{DPYD} to explain the same level of genetically encoded functional variability. Furthermore, we demonstrate substantial differences in metabolic activity and the underlying genetic variability across eight different populations with important implications for the design of pharmacogenetic testing strategies and precision public health.

\textbf{METHODS}

Data collections and definitions

In vitro data of missense single-nucleotide variants (SNVs) in\textit{DPYD} and\textit{TPMT} were collected from functional studies conducted using cell lines.\textsuperscript{24–33} As classification criteria for variants differed between studies, we homogenised the definitions and considered all variations as deleterious, which resulted in in vitro activity lower than 70\% of the respective wild-type (WT) enzyme. In vivo data were collected from the ClinVar database.\textsuperscript{34} The deleteriousness of variants was curated based on their annotation as impacting drug response, pathogenic or likely pathogenic. Variant frequencies from 138,842 individuals across eight different populations (12,487 Africans, 17,720 Latinos, 5185 Ashkenazi Jews, 9977 East Asians, 64,603 Non-Finnish Europeans, 12,562 Finnish, 15,308 South Asians and 1000 Swedes) were collected from GnomAD\textsuperscript{35} and SWEGene.\textsuperscript{36} Linkage between the\textit{TPMT} variants rs1800460 and rs1142345 was calculated using LDlink.\textsuperscript{37}

Variant-effect predictions

We used Polyphen-2, SIFT, PROVEAN and CADD for binary predictions of variant functionality. To quantitatively predict the functional impact of all identified variants, we used the ADME-optimised prediction framework (APF) that provides normalised quantitative functionality prediction scores in the range from 1 (neutral) to 0 (deleterious).\textsuperscript{23} The functional impact of frameshift and stop-gain variations was further confirmed by LOFTEE (https://github.com/konradjk/loftee). For qualitative comparisons with binary scores, variants with functionality scores ≤0.5 were considered as deleterious, while scores >0.5 denote functionally neutral variants.

Informedness calculations

Plotting the cumulative number of variants against their cumulative aggregated frequencies reveals the excess of information that can be obtained for each number of variants tested. The informedness (I) of\textit{DPYD} or\textit{TPMT} testing is defined as the maximal vertical difference between this receiver-operating characteristic (ROC) curve and the bisectrix through the origin of the coordinate system. The intronic variant rs75017182 that is linked to haplotype HapB3 was included with a functionality score of 0.5 according to current guidelines.\textsuperscript{38}

Calculation of\textit{DPYD} and\textit{TPMT} metaboliser phenotype frequencies

To calculate the frequencies of poor metabolisers (PM), intermediate metabolisers (IM) and normal metabolisers (NM) for each population analysed, we calculated all diplotype frequencies for\textit{DPYD} and\textit{TPMT} and added the functionality scores of both alleles to yield the corresponding activity score of the individual. PMs, IMs and NMs were defined as activity scores (AS) of ≤0.5, 0.5 < AS ≤ 1.5 and >1.5, respectively, according to the genotype combinations defined by Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines.\textsuperscript{38,39}

\textbf{RESULTS}

Benchmarking of the ADME prediction framework on\textit{DPYD} and\textit{TPMT} variants with known in vivo consequences

To test the predictive power of the APF algorithm, we first defined a gold-standard data set for\textit{DPYD} and\textit{TPMT} that included all variants with well-characterised functional effects in vivo in humans. Extensive literature search revealed a total of 61 variants with known clinical consequences in\textit{DPYD}, of which 12 were classified as neutral and 49 as pathogenic (Supplementary Table 1). For\textit{TPMT}, only nine variants were sufficiently studied for their impact on drug response in vivo, all of which were deleterious. For\textit{DPYD}, these characterised variants resulted in frameshifts (n = 23), followed by missense (n = 21), stop-gain (n = 16) and splice variants (n = 1;\textit{DPYD}*2A), whereas all nine characterised\textit{TPMT} variants were missense (Fig. 1a). Of all 70 clinically annotated\textit{DPYD} and\textit{TPMT} variants, six were common with global minor allele frequencies (MAF) >1\% (C29R, i543V, M166V, V732I and S543N in\textit{DPYD} and A154T in\textit{TPMT}), whereas all other variants were rare (Fig. 1b).

For variants for which in vitro data were available (64 out of 70 variants, 91\%), these experimental categorisations were in good agreement with in vivo phenotypes (95.3\% accuracy, 61 out of 64 variants, Fig. 1c). The APF could analyse all 70 variants and achieved 91.4\% accuracy (64 out of 70), thus closely approximating the accuracy of in vitro testing. In contrast, other commonly used in silico tools, such as SIFT, Polyphen-2, PROVEAN and CADD, had substantially lower predictive power (62.1–78.6\% accuracy) and failed to predict the functional consequences of >50\% of variations. Only one incorrectly classified variant overlapped between in vitro and in silico assessments. R215H in\textit{TPMT} (\textit{TPMT}*8) was incorrectly predicted as benign by APF and in vitro data, while it is associated with reduced\textit{TPMT} activity in vivo.\textsuperscript{40} Notably, however, this allele is very rare (MAF = 0.2\%),\textit{TPMT} enzyme activity in vivo has to the best of our knowledge only been reported for a single carrier.

Population-scale prediction of\textit{DPYD} and\textit{TPMT} variant functionality

As the predictive accuracy of APF exceeded 90\% and was similar to in vitro testing on the gold-standard in vivo data set, we expanded our evaluations and tested\textit{DPYD} and\textit{TPMT} variability on a population scale (Fig. 1d–g). By analysing WGS and WES data from 138,842 individuals, we identified 705 and 154\textit{DPYD} and\textit{TPMT} variants, respectively, of which only 164 (23\%) and 26 (17\%) had been analysed in cell systems (Fig. 1d, e). Of the 164 experimentally tested\textit{DPYD} variants, in vitro tests predicted that 67 (41\%) decreased enzyme function, whereas the fraction was considerably higher using computational tests (n = 114, 69.5\%, Fig. 1d). In contrast, the number of variants predicted to be deleterious by in vitro assays was substantially higher for\textit{TPMT} where 23 out of 26 tested variants (88\%) resulted in functional consequences, while only 19 variants (73\%) were predicted to be deleterious by APF (Fig. 1e).

Lastly, we predicted the functional effects of all identified variants, including those without available in vitro data. In total, 506 out of 705\textit{DPYD} variants (72\%) were predicted to be deleterious with functionality scores below 0.5, of which 311 were LOF variants with <20\% of WT enzyme activity (Fig. 1f). For\textit{TPMT}, 99 and 55 out of 154 variants (64 and 36\%, respectively) were
predicted to be deleterious and neutral, respectively (Fig. 1g). The highest frequency of all alleles was for the loss-of-function allele TPMT*3A with a MAF = 2.8%.

Genetic complexity of DPD and TPMT function

We then evaluated the distribution of DPDY and TPMT genotypes across reduced function alleles. Specifically, we calculated the fraction of reduced function phenotypes that could be explained by selections of variants in the respective genes. For DPDY, 50% of reduced function alleles are explained by two variants (HapB3 and *2A), whereas the numbers of variants that need to be interrogated to explain more of the functional variability increase exponentially (Table 1). For instance, inclusion of six more variants only explains an additional 24.6% of DPD functionality, whereas 88 and 421 variants need to be interrogated to explain 90 and 99% of DPD deficiency, respectively (Fig. 2a and Table 1). The highest excess of information is obtained for the testing of 34 variations, which can explain 84.2% of genetically encoded functional DPD variability while only corresponding to 6.7% of deleterious DPDY variants.

In comparison to DPDY, functional variability in TPMT was characterised by a high excess in information allotted to only few variations. Interrogation of A154T and Y240C (TPMT*3A), corresponding to 2% of all deleterious TPMT variants, was sufficient to explain >70% of functional TPMT functional variability (Fig. 2b and Table 1). Furthermore, >95% of differences in TPMT function were attributed to only four variants (A154T, Y240C, A80P and R163H corresponding to *3A, *3C, *2 and *16), whereas the remaining 95 variants combined only accounted for 4.6%. Notably, while
DISCUSSION

DPD and TPMT deficiency are the major determinants of severe fluoropyrimidine and thiopurine-associated toxicity, and prospective genotyping followed by genotype-guided prescribing has been shown to reduce adverse events. Thus, the identification of genetic variations that result in reduced enzyme function is of central importance to improve patient outcomes. A few dozen variants have well-characterised effects in vivo and can be acted on accordingly once they are identified in a patient's genome. However, previous sequencing efforts, such as the 1000 Genomes Project (n = 2504 individuals) and the Exome Sequencing Project (n = 6503 individuals), identified >100 additional DPD and TPMT variations with unclear functional consequences. Consequently, while rare genetic variations have long been considered important factors to explain at least part of the missing heritability in drug response, there are currently no guidelines for carriers of such uncharacterised
variants as to whether or not to modify doses or switch to alternative agents. There is thus a need for computational tools that can aid in the reliable functional interpretation of such variations.

Our data showed that commonly used non-pharmacogene-specific algorithms that perform well on disease data\(^{65}\) had only moderate predictive power for \(DPYD\) and \(TPMT\) variants (62.1–78.6%). In contrast, \(DPYD\)-Varifer, a machine-learning-based classifier trained exclusively on \(DPYD\) variant data, showed 85% accuracy on a set of novel missense variations compared to in vitro data.\(^{26}\) Here, we find that APF, a quantitative ensemble score recently developed specifically for pharmacogenetic predictions,\(^{23}\) achieved a binary classification accuracy of 91.4% on a set of 70 variations with known in vivo effects. However, we want to emphasise that these results cannot be directly compared to \(DPYD\)-Varifer, as all these variants in question were used for the training of this tool. In contrast, APF has not been trained on any \(DPYD\) variants, suggesting that the underlying framework is broadly applicable to the functional interpretation of variants across pharmacogenes encoding metabolic enzymes. Few discrepancies between the in vitro assessment and in vivo function were reported for some variants, such as Y186C and D974V. Consequently, we argue for the benchmarking of computational tools on variants with known effects in vivo.

While APF constitutes the only tool providing quantitative estimates of variant function across pharmacogenes, it also has notable limitations. APF cannot detect gain-of-function effects and variants that result in increased enzymatic function in vitro, such as \(DPYD\) P1023T,\(^{24}\) which APF classifies as functionally neutral. In addition, APF is designed for the analysis of individual variants. As such, the functionality of variant combinations is driven by the most deleterious variant in the haplotype. However, a recent clinical study showed that a \(DPYD\) haplotype containing three neutral missense variants (C29R, M166V and V732I) is strongly associated with decreased reduced fluorouracil degradation and severe toxicity,\(^{20}\) which would thus be missed by APF. Furthermore, we note that APF results in an excess of false positives for \(DPYD\), such as D829N, A450V and S534N, which are correctly classified as functionally neutral using in vitro assays. By contrast, APF correctly flagged \(DPYD\) Y186C and D974V as deleterious, whereas in vitro studies only detected minor reduction of 15 and 22% of WT \(DPYD\) activity for these variants, respectively.\(^{24}\) Importantly, unlike other algorithms, APF provides an activity score estimate that strongly correlates with measured in vitro activity (\(R^2 = 0.9; P = 2.9 \times 10^{-5}\)).\(^{23}\) While \(DPYD\) V732I (\(DPYD^{*}6\)) is mostly considered as benign by both in vivo and in vitro studies,\(^{7,51,52}\) as well as our APF algorithm, this variant has recently been implicated in increased risk of gastrointestinal and haematological fluoropyrimidine toxicity.\(^{53,54}\) Hence, we find that APF, a quantitative ensemble score recently developed specifically for pharmacogenetic predictions,\(^{23}\) achieved a binary classification accuracy of 91.4% on a set of 70 variants with known in vivo effects. However, we want to emphasise that these results cannot be directly compared to \(DPYD\)-Varifer, as all these variants in question were used for the training of this tool. In contrast, APF has not been trained on any \(DPYD\) variants, suggesting that the underlying framework is broadly applicable to the functional interpretation of variants across pharmacogenes encoding metabolic enzymes. Few discrepancies between the in vitro assessment and in vivo function were reported for some variants, such as Y186C and D974V. Consequently, we argue for the benchmarking of computational tools on variants with known effects in vivo.

While APF constitutes the only tool providing quantitative estimates of variant function across pharmacogenes, it also has notable limitations. APF cannot detect gain-of-function effects and variants that result in increased enzymatic function in vitro, such as \(DPYD\) P1023T,\(^{24}\) which APF classifies as functionally neutral. In addition, APF is designed for the analysis of individual variants. As such, the functionality of variant combinations is driven by the most deleterious variant in the haplotype. However, a recent clinical study showed that a \(DPYD\) haplotype containing three neutral missense variants (C29R, M166V and V732I) is strongly associated with decreased reduced fluorouracil degradation and severe toxicity,\(^{20}\) which would thus be missed by APF. Furthermore, we note that APF results in an excess of false positives for \(DPYD\), such as D829N, A450V and S534N, which are correctly classified as functionally neutral using in vitro assays. By contrast, APF correctly flagged \(DPYD\) Y186C and D974V as deleterious, whereas in vitro studies only detected minor reduction of 15 and 22% of WT \(DPYD\) activity for these variants, respectively.\(^{24}\) Importantly, unlike other algorithms, APF provides an activity score estimate that strongly correlates with measured in vitro activity (\(R^2 = 0.9; P = 2.9 \times 10^{-5}\)).\(^{23}\) V732I was estimated to have an activity score of 0.6, slightly above the selected binary classification threshold for deleteriousness of 0.5, suggesting that also variants with modest decreases in activity might increase toxicity risks, albeit with lower risk scores than clear loss-of-function variants, such as \(DPYD^{*}2A\) and D949V (APF score of 0 for both). This is consistent with the findings by Del Re et al.\(^{54}\) and Boige et al.\(^{53}\) who reported only moderate hazard ratios (HRs) for V732I, D949V of 1.7–19, respectively, whereas HRs for \(DPYD^{*}2A\) were substantially higher (6.3 and 4.2, respectively).

Multiple challenges need to be overcome before fluoropyrimidine and thiopurine dosing based on personalised genomic profiles and their computational interpretations can be implemented in clinical practice. Most importantly, implementation efforts critically rely on the demonstration of clinical utility using stringent prospective trials. In addition, the establishment of population-scale genomic biobanks\(^{53,57}\) with associated longitudinal electronic medical records offers exciting opportunities to test the predictive power of this and other computational
East Asian scale pro intronic variant rs75017182 (start-lost, frameshifts, splicing and stop gain), as well as the amino acid sequence of the respective gene product (missense, Importantly, our analyses considered all variants affecting the prevalence of DPD deperipheral blood mononuclear cell DPD radioassays indicated a prevalence of DPD de
patients to carry at least one functionally relevant allele.12 A previous study indicated that DPD de
worldwide distribution of the major genetic determinants of fluoropyrimidine and thiopurine metabolism with important
deleterious but experimentally uncharacterised variations for
increased monitoring. In addition to the clinical validity of in silico predictions, decisions of gene sequencing require careful evaluations regarding the cost-effectiveness of such measures compared to conventional genotyping, particularly as not even these candidate interrogations are routinely conducted in most countries.

By leveraging APF's scalability, we provide the first population-scale profiles of DPD and TPMT metaboliser phenotypes that consider the entire repertoire of coding genetic variation. Importantly, our analyses considered all variants affecting the amino acid sequence of the respective gene product (missense, start-lost, frameshifts, splicing and stop gain), as well as the intronic variant rs75017182 (DPYD HapB3). Uracil breath tests and peripheral blood mononuclear cell DPD radioassays indicated a prevalence of DPD deficiency in Africans of 8%,58 which aligns very well with our APF estimates of 8.4%. Furthermore, our data suggested a prevalence of reduced DPD function alleles in Europeans of 7.6%, again in good agreement with the results of a prospective study in the Netherlands, which reported 8% of patients to carry at least one functionally relevant DPYD variant allele.12 A previous study indicated that DPD deficiency due to DPYD*2A was very high in Sweden with MAF = 3.5%,59 while our analysis of 1000 Swedish genomes showed lower frequencies of 0.8%. Notably, however, *2A frequencies in neighbouring Finland (n = 12,562) were substantially higher (MAF = 2.4%), corroborating an overall high rate of DPD deficiency in Scandinavia. For TPMT, the available literature indicates PM and IM frequencies of 0.45–0.6% and 9.9–11.9% for Caucasians,15,60–63 again closely corresponding to APF predictions of 0.2% and 9%. Notably, while genotype-based predictions align overall well with measured TPMT phenotypes (97% concordance), concordance is lower for TPMT*1/*2 and *1/*3 heterozygotes (79% concordance).64

In summary, this study demonstrates that the pharmacogenetically trained APF classifier provides accurate predictions of DPYD and TPMT variant functionality outperforming conventionally used algorithms trained on disease data. We show that DPD metaboliser status is genetically complex and requires profiling of tens of variations to explain the majority of phenotypic differences. In contrast, >95% of functional TPMT variability is explained by only four variants. By leveraging population-scale sequencing data, we provide spectra of ethnogeographic variation in DPD and TPMT phenotypes on an unprecedented scale, and reveal unexpectedly large interethnic differences in DPD and TPMT deficiencies. Our results provide a powerful resource for the worldwide distribution of the major genetic determinants of fluoropyrimidine and thiopurine metabolism with important

predicted phenotypic frequencies in panel a: PM, intermediate metabolisers (IM), and normal metabolisers (NM). Panel b: TPMT metaboliser phenotypes (PM, IM, NM). The bar charts illustrate the fraction of poor metabolisers (PM), intermediate metabolisers (IM) and normal metabolisers (NM) for each population. Overlaid dots in panel a indicate predictions based on available variant activity scores of DPYD-Varifier (training and test sets). As this algorithm only provides binary assessments, all deleterious and neutral variants were considered to have functionality scores of 0 and 1,

**Fig. 3** Integrative analysis of population-scale sequencing data reveals striking ethnogeographic differences in DPD and TPMT activity. Line plots (left) show population-specific aggregation of variant functionality scores into spectra of DPD (a) and TPMT (b) functionality. Column plots (right) indicate the fraction of poor metabolisers (PM), intermediate metabolisers (IM) and normal metabolisers (NM) for each population. Overlaid dots in panel a indicate predictions based on available variant activity scores of DPYD-Varifier (training and test sets). As this algorithm only provides binary assessments, all deleterious and neutral variants were considered to have functionality scores of 0 and 1, respectively.
implications for population-adjusted genotyping strategies and precision public health.

ACKNOWLEDGEMENTS
The authors thank ClinVar and the Genome Aggregation Consortium for sharing their data, which were instrumental for this work.

AUTHOR CONTRIBUTIONS
V.M.L. and Y.Z. designed this study. Y.Z. and C.D.H. collected and analysed the data. All authors contribute to the paper writing.

ADDITIONAL INFORMATION
Ethics approval and consent to participate Data analysed in this work are publicly available under the Fort Lauderdale Agreement. Therefore, separated ethical approval was not required.

Consent to publish Not applicable.

Data availability All DPYD and TPMT variants with available in vivo data analysed in this study are provided in Supplementary Table S1. Sequencing data of 138,842 individuals are available at https://gnomad.broadinstitute.org/.

Competing interests The authors declare no competing interests according to the ICMJE Uniform Requirements. However, V.M.L. would like to declare the following financial relationships: co-founder and shareholder of HepaPredict AB; consultancy work for Enginzyme AB. The remaining authors declare no competing interests for this work.

Funding information This work was supported by the Swedish Research Council (grant agreement numbers: 2016-0153, 2016-01154 and 2019-08371), by the EU EFPIA/OCR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking (EUBOOPEN grant number 875510), by the European Union’s Horizon 2020 research and innovation program U-POX [grant agreement number 668353] and by the Strategic Research Programmes in Diabetes (SFO Diabetes) and Stem Cells and Regenerative Medicine (SFO StratRegen). C.D.H. was supported by a fellowship from FAPESP (reference number 2019/19009-4).

Supplementary information is available for this paper at https://doi.org/10.1038/s41416-020-01084-0.

Note This work is published under the standard license to publish agreement. After 12 months the work will become freely available and the license terms will switch to a Creative Commons Attribution 4.0 International (CC BY 4.0).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES
1. Wheeler, H. E., Maitland, M. L., Dolan, M. E., Cox, N. J. & Ratain, M. J. Cancer pharmacogenomics: strategies and challenges. Nat. Rev. Genet. 14, 23–34 (2013).
2. Lauschke, V. M., Milani, L. & Ingelman-Sundberg, M. Pharmacogenomics revisited: the EU biomarkers for improved drug therapy—recent progress and future developments. AAPS J. 20, 4 (2017).
3. Amstutz, U., Froehlich, T. K. & Largiadèr, C. R. Dihydropyrimidine dehydrogenase gene as a major predictor of severe 5-fluorouracil toxicity. Pharmacogenomics 12, 1321–1336 (2011).
4. Roy, L. M., Zur, R. M., Ulenyk, E., Carew, C., Ito, S. & Ungar, W. Thiopurine S-methyltransferase testing for averting drug toxicity in patients receiving thiopurines: a systematic review. Pharmacogenomics 17, 633–656 (2016).
5. Van Kuilenburg, A. B., Haase, J., Richel, D. J., Zoetekouw, L. & Van Lenthe, H., De Abreu, R. A. et al. Clinical implications of dihydropyrimidine dehydrogenase (DPD) deficiency in patients with severe 5-fluorouracil-associated toxicity: identification of new mutations in the DPD gene. Clin. Cancer Res. 6, 4705–4712 (2000).
6. Morel, A., Boldrion-Celle, M., Fey, L., Soulé, P., Craipeau, M. C., Traore, S. et al. Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide polymorphisms on 5-fluorouracil tolerance. Mol. Cancer Ther. 5, 2895–2904 (2006).
7. Offer, S. M., Wegner, N. J., Fossum, C., Wang, K. & Diasio, R. B. Phenotypic profiling of DPYD variations relevant to 5-fluorouracil sensitivity using real-time cellular analysis and in vitro measurement of enzyme activity. Cancer Res. 73, 1958–1968 (2013).
8. Terrazino, S., Cargnin, S., Del Re, M., Danesi, R., Canonic, P. L. & Genazzani, A. A. DPYD IVS14+1G>A and 2846A>T genotyping for the prediction of severe fluoropyrimidine-related toxicity: a meta-analysis. Pharmacogenomics 14, 1255–1272 (2013).
9. Rosmarin, D., Palles, C., Church, D., Domingo, E., Jones, A., Johnstone, E. et al. Genetic markers of toxicity from capecitabine and other fluorouracil-based regimens: investigation in the QUASAR2 study, systematic review, and meta-analysis. J. Clin. Oncol. 32, 1031–1039 (2014).
10. Meulendijks, D., Henricks, L. M., Sonke, G. S., Deenen, M. J., Froehlich, T. K., Amstutz, U. et al. Clinical relevance of DPYD variants c.1679T>C, c.1236G>A/HapA, and c.1610G>A as predictors of severe fluoropyrimidine-associated toxicity: a systematic review and meta-analysis of individual patient data. Lancet Oncol. 16, 1639–1650 (2015).
11. Offer, S. M., Wegner, N. J., Fossum, C., Wang, K. & Diasio, R. B. Phenotypic DPYD genotyping to reduce the risk of fluoropyrimidine-induced severe toxicity: Ready for prime time. Eur. J. Cancer 54, 40–48 (2016).
12. Henricks, L. M., Lauschke, V. M., De Man, F. M., Meulendijks, D., Frederix, G. W. J., Kienhuis, E. et al. DPYD genotype-guided dose individualisation of fluoropyrimidine therapy in patients with cancer: a prospective safety analysis. Lancet Oncol. 19, 1459–1467 (2018).
13. Deenen, M. J., Meulendijks, D., Cats, A., Sechtinger, M. K., Severens, J. L., Boot, H. et al. Upfront genotyping of DPYD*2A to individualize fluoropyrimidine therapy: a safety and cost analysis. J. Clin. Oncol. 34, 227–234 (2016).
14. Weinshilboum, R. M. & Sladek, S. L. Mercaptopurine pharmacogenomics: monogenic inheritance of erythrocithide thiopurine methyltransferase activity. Am. J. Hum. Genet. 32, 651–662 (1980).
15. Schaeffeler, E., Fischer, C., Brockmeier, D., Wernet, D., Moerike, K., Eichelbaum, M. et al. Comprehensive analysis of thiopurine S-methyltransferase phenotype–genotype correlation in a large population of German-Caucasians and identification of novel TPMT variants. Pharmacogenomics 14, 407–417 (2004).
16. Kozya, M., Ingelman-Sundberg, M. & Lauschke, V. M. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. Genet. Med. 19, 20–29 (2017).
17. Ingelman-Sundberg, M., Mikkelsen, S., Zhou, Y. & Lauschke, V. M. Integrating rare genetic variants into pharmacogenomic drug response predictions. Hum. Genomics 12, 26 (2018).
18. Lauschke, V. M. & Ingelman-Sundberg, M. Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. NPJ Genom. Med. 5, 2368–2367 (2020).
19. Matresek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiassiin, M. A., Gray, V. E. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. 50, 874–882 (2018).
20. Peterson, T. A., Doughty, E. & Kann, M. G. Towards precision medicine: advances in computational approaches for the analysis of human variants. J. Mol. Biol. 425, 4047–4063 (2013).
21. Waghil, O., Galardi, M., Busby, B. P., Memon, D., Tynas, A. & Beltrao, P. A resource of variant effect predictions of single nucleotide variants in model organisms. Mol. Syst. Biol. 14, e8430 (2018).
22. Zhou, Y., Fujioka, K., Mikkelsen, S. & Lauschke, V. M. Computational methods for the pharmacogenetic interpretation of next generation sequencing data. Front. Pharmacol. 9, 1437 (2018).
23. Zhou, Y., Mikkelsen, S., Komondai, M., Hiratsuka, M. & Lauschke, V. M. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. Pharmacogenomics J. 19, 115–126 (2019).
24. Offer, S. M., Fossum, C. C., Wegner, N. J., Stuflesser, A. J., Butterfield, G. L. & Diasio, R. B. Comparative functional analysis of DPYD variants of potential clinical relevance to dihydropyrimidine dehydrogenase activity. Cancer Res. 74, 2545–2554 (2014).
25. Ekabayash, T., Jerde, C. R., Shrestha, S., Wu, R., Nie, Q., Giama, H. N. et al. Novel deleterious dihydropyrimidine dehydrogenase variants may contribute to 5-fluorouracil sensitivity in an east African population. Clin. Pharmacol. Ther. 101, 382–390 (2017).
26. Shrestha, S., Zhang, C., Jerde, C. R., Nie, Q., Li, H., Offer, S. M. et al. Gene-specific variant classifier (DPYD-varifier) to identify deleterious alleles of dihydropyrimidine dehydrogenase. Clin Pharmacol. Ther. 104, 709–718 (2018).
27. Ujie, S., Sasaki, T., Mizugaki, M., Ishikawa, M. & Hiratsuka, M. Functional characterization of 23 allelic variants of thiopurine S-methyltransferase gene (TPMT*2–*4). Pharmacogenet. Genomics 18, 887–893 (2008).
28. Salvaggio, G. E., Wang, L., Wiebert, M., Yee, V. C. & Weinshilboum, R. M. Thiopurine S-methyltransferase pharmacogenomics: variant allele functional and comparative genomics. Pharmacogenet. Genomics 15, 801–815 (2005).
29. Hamdan-Khalil, A., Allorge, D., Lo-Guidice, J., Cauff, C., Chevalier, D., Spire, C. et al. In vitro characterization of four novel non-functional variants of the thiopurine S-methyltransferase. Biochem. Biophys. Res. Commun. 309, 1005–1010 (2003).

30. Hamdan-Khalil, R., Gala, J., Allorge, D., Lo-Guidice, J., Horsmans, Y., Houdret, N. et al. Identification and functional analysis of two rare allelic variants of the thiopurine S-methyltransferase gene, TPMT*16 and TPMT*19. Biochem. Pharmacol. 69, 525–529 (2005).

31. Garat, A., Cauff, C., Renault, N., Lo-Guidice, J., M., Allorge, D., Chevalier, D. et al. Characterisation of novel defective thiopurine S-methyltransferase allelic variants. Biochem. Pharmacol. 76, 404–415 (2008).

32. Feng, Q., Vannaprasaht, S., Peng, Y., Angsuthum, S., Avihingsanon, Y., Yee, V. C. et al. Thiopurine S-methyltransferase pharmacogenetics: functional characterization of a novel rapidly degraded variant alosyme. Biochem. Pharmacol. 79, 1053–1061 (2010).

33. Lindqvist Appell, M., Wennerstrand, P., Peterson, C., Hertervig, E. & Mårtensson, L. (2005). In vitro characterization of four novel non-functional variants of the thiopurine S-methyltransferase gene. Pharmacogenet. Genomics 20, 700–707 (2010).

34. Landrum, M. J., Lee, J. M., Riley, G. R., Jiang, W., Rubinstein, W. S., Church, D. M. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, D980-D985 (2014).

35. Lek, M., Karczewski, K. J., Minkel, E. V., Samocha, K. E., Banks, E., Fennell, T. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).

36. Ameur, A., Dahlberg, J., Olason, P., Vezzi, F., Karlsson, R., Martin, M. et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur. J. Hum. Genet. 25, 1253–1260 (2017).

37. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring relationship between single nucleotide polymorphisms and haplotypes in DPYD and toxicity and efficacy of capecitabine in advanced colorectal cancer. Clin. Cancer Res. 17, 3455–3468 (2011).

38. He, Y. F., Wei, W., Zhang, X., Li, Y. H., Li, S., Wang, F. H. et al. Analysis of the DPYD gene implicated in 5-fluorouracil catabolism in Chinese cancer patients. J. Clin. Pharmacol. 53, 307–314 (2008).

39. Boige, V., Vincent, M., Alexandre, P., Teijpar, S., Landolfi, S., Le Malicot, K. et al. DPYD genotyping to predict adverse events following treatment with fluorouracil-based adjuvant chemotherapy in patients with stage III colon cancer: a secondary analysis of the PETACC-8 randomized clinical trial. JAMA Oncol. 2, 655–662 (2016).

40. Del Re, M., Cinieri, S., Michelucci, A., Salvadori, S., Loupakis, F., Schirripa, M. et al. DPYD *6 plays an important role in fluoropyrimidine toxicity in addition to DPYD *2A and c.2846A>T: a comprehensive analysis in 1254 patients. Pharmacogenomics J. 19, 556–563 (2019).

41. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).

42. Fagerberg, L., Hall, M., Quackenbush, J., Tissier, A. & Ohlson, T. The Human Protein Atlas database and web site in 2018. Nucleic Acids Res. 46, D806–D812 (2017).

43. Koistinen, T., Lehtonen, M., Vainio, H., Miettinen, O., Laatikainen, T. et al. The TH exome: whole-genome exome sequencing of 1,000 Finnish subjects. PLoS Genet. 12, e1006462 (2016).

44. Saif, M. W., Lee, A. M., Offer, S. M., McConnell, K., Relias, V. & Diasio, R. B. A DPYD variant (Y186C) in individuals of African ancestry is associated with reduced activity with genotype concordance: a new, simplified pharmacotyping assay. Ann. Clin. Biochem. 43, 354–360 (2006).

45. Zaanan, A., Dumont, L.-M., Loriot, M.-A., Taieb, J. & Narjoz, C. A case of 5-FU-related severe toxicity associated with the p.Y186C DPYD variant. Clin. Pharmacol. Ther. 95, 136–138 (2014).

46. Coenen, M. J. H., de Jong, D. J., van Marrewijk, C. J., Derijks, L. J., Veermeulen, S. H., Wong, D. R. et al. Identification of patients with variants in TPMT and dose reduction reduces hematologic events during thiopurine treatment of inflammatory bowel disease. Gastroenterology 149, 907–917.e907 (2015).

47. Lauschke, V. M. & Ingelman-Sundberg, M. Precision medicine and rare genetic variants. Trends Pharmacol. Sci. 37, 85–86 (2016).

48. Lauschke, V. M. & Ingelman-Sundberg, M. How to consider rare genetic variants in personalized drug therapy. Clin. Pharmacol. Ther. 103, 745–748 (2018).

49. Li, J., Zhao, T., Zhang, Y., Zhang, K., Shi, L., Chen, Y. et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 46, 7793–7804 (2018).

50. Gentile, G., Botticelli, A., Lionetto, L., Mazzuca, F., Simmaco, M., Marchetti, P. et al. Genotype-phenotype correlations in 5-fluorouracil metabolism: a candidate DPYD haplotype to improve toxicity prediction. Pharmacogenomics J. 16, 320–325 (2016).

51. Deen, M. J., Tol, J., Burzy, A. M., Doodeman, V. D., de Boer, A., Vincent, A. et al. Relationship between single nucleotide polymorphisms and haplotypes in DPYD and toxicity and efficacy of capecitabine in advanced colorectal cancer. Clin. Cancer Res. 17, 3455–3468 (2011).

52. Reisberg, S., Kres, K., Lepamets, M., Kals, M., MagGU, M., Metsalu, K. et al. Translating genotypic data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. Genet. Med. 21, 1345–1354 (2019).

53. Mattison, L. K., Fourie, J., Desmond, R. A., Modak, A., Safi, M. W. & Diasio, R. B. Increased prevalence of dihydropyrimidine dehydrogenase deficiency in African-Americans compared with Caucasians. Clin. Cancer Res. 12, 5491–5495 (2006).

54. Olverholm, A., Arkbild, E., Skrtic, S., Albertsson, P., Shubbar, E. & Enerbäck, C. Two cases of 5-fluorouracil toxicity linked with gene variants in the DPYD gene. Clin. Pharmacol. Ther. 43, 331–340 (1988).

55. Tinel, M., Berson, A., Pessayre, D., Letteron, P., Cattoni, M. P., Horsmans, Y. et al. Pharmacogenetics of human erythrocyte thiopurine methyltransferase activity in a French population. Br. J. Clin. Pharmacol. 32, 729–734 (1991).

56. Holme, S. A., Duley, J. A., Sanderson, J., Routledge, P. A. & Anstey, A. V. Erythrocyte thiopurine methyltransferase assessment prior to azathioprine use in the UK. JIMM: Monthly J. Assoc. Physicians 95, 439–444 (2002).

57. Gilsdorf, J. P., Gomollón, F., Cara, C., Luna, M., González-Lama, Y., Pajares, J. M. et al. Thiopurine methyltransferase activity in Spain: a study of 14,545 patients. Digestive Dis. Sci. 52, 1262–1269 (2007).

58. Cooper, S. C., Ford, L. T., Berg, J. D. & Lewis, M. J. V. Ethnic variation of thiopurine methyltransferase activity with genotype concordance: a new, simplified pharmacotyping assay. Ann. Clin. Biochem. 43, 354–360 (2006).