Massively parallel variant characterization identifies *NUDT15* alleles associated with thiopurine toxicity

Chase C. Suiter[a], Takaya Moriyama[a], Kenneth A. Matreyek[b], Wentao Yang[a], Emma Rose Scaletti[c,d], Rina Nishii[a], Wenjian Yang[a], Keito Hoshitsuki[a], Minu Singh[e], Amita Trehan[e], Chris Parish[a], Colton Smith[a], Lie Li[a], Deepa Bhojwani[f], Liz Y. P. Yuen[g], Chi-kong Li[h], Chak-ho Li[i], Yung-li Yang[j], Gareth J. Walker[k,l], James R. Goodhand[k,l], Nicholas A. Kennedy[k,l], Federico Antillon Klussmann[m,n], Smita Bhatia[o], Mary V. Relling[a], Motohiro Kato[p], Hiroki Hori[q], Prateek Bhatia[e], Tariq Ahmad[k,l], Allen E. J. Yeoh[r,s], Pål Stenmark[c,d], Douglas M. Fowler[b,t,u], and Jun J. Yang[a,1]

[a]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105; [b]Department of Genome Sciences, University of Washington, Seattle, WA 98195; [c]Department of Biochemistry and Biophysics, Arrhenius Laboratories for Natural Sciences, Stockholm University, 106 91 Stockholm, Sweden; [d]Department of Experimental Medical Science,Lund University, 221 00 Lund, Sweden; [e]Department of Pediatrics, Advanced Pediatrics Centre, Postgraduate Institute of Medical Education & Research, 160012 Chandigarh, India; [f]Department of Pediatrics, Children's Hospital of Los Angeles, Los Angeles, CA 90027; [g]Department of Pathology, Hong Kong Children's Hospital, Hong Kong; [h]Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong; [i]Department of Paediatrics and Adolescent Medicine, Tuen Mun Hospital, Hong Kong; [j]Department of Laboratory Medicine and Pediatrics, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei 10617, Taiwan; [k]Department of Gastroenterology, Royal Devon and Exeter Hospital NHS Foundation Trust, Exeter EX2 5DW, England; [l]IBD Pharmacogenetics Group, University of Exeter, Exeter EX2 5DW, England; [m]Unidad Nacional de Oncología Pediátrica, Guatemala City 01011, Guatemala; [n]Department of Pediatrics, Francisco Marroquin Medical School, Guatemala City 01011, Guatemala; [o]Division of Pediatric Hematology/Oncology, Institute for Cancer Outcomes and Survivorship, School of Medicine, University of Alabama, Birmingham, AL 35294; [p]Department of Pediatric Hematology and Oncology Research, National Center for Child Health and Development, Tokyo 157-8535, Japan; [q]Department of Pediatrics, Mie University Graduate School of Medicine, Mie 514-8507, Japan; [r]Centre for Translational Research in Acute Leukaemia, Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117599; [s]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117597; [t]Department of Bioengineering, University of Washington, Seattle, WA 98195; and [u]Genetic Networks Program, CIFAR, Toronto, ON M5G 1M1, Canada

Edited by Jasper Rine, University of California, Berkeley, CA, and approved January 29, 2020 (received for review September 9, 2019)

As a prototype of genomics-guided precision medicine, individualized thiopurine dosing based on pharmacogenetics is a highly effective way to mitigate hematopoietic toxicity of this class of drugs. Recently, *NUDT15* deficiency was identified as a genetic cause of thiopurine toxicity, and *NUDT15*-informed preemptive dose reduction was quickly adopted in clinical settings. To exhaustively identify pharmacogenetic variants in this gene, we developed massively parallel NUDT15 function assays to determine the variants' effect on protein abundance and thiopurine cytotoxicity. Of the 3,097 possible missense variants, we characterized the abundance of 2,922 variants and found 54 hotspot residues at which variants resulted in complete loss of protein stability. Analyzing 2,935 variants in the thiopurine cytotoxicity-based assay, we identified 17 additional residues where variants altered NUDT15 activity without affecting protein stability. We identified structural elements key to NUDT15 stability and/or catalytical activity with single amino acid resolution. Functional effects for *NUDT15* variants accurately predicted toxicity risk alleles in patients treated with thiopurines with far superior sensitivity and specificity compared to bioinformatic prediction algorithms. In conclusion, our massively parallel variant function assays identified 1,152 deleterious *NUDT15* variants, providing a comprehensive reference of variant function and vastly improving the ability to implement pharmacogenetics-guided thiopurine treatment individualization.

thiopurines | NUDT15 | pharmacogenetics | massively parallel variant function assay

Thiopurines (e.g., mercaptopurine [MP], 6-thioguanine [TG], and azathioprine) are important antimetabolite drugs with diverse clinical indications. For example, as a potent antileukemia agent, MP-based maintenance therapy is arguably one of the most critical components of the curative treatment regimen for acute lymphoblastic leukemia (ALL) in children and adults (1–5). Thiopurines are also commonly used as immunosuppressive agents for the treatment of rheumatoid arthritis and inflammatory bowel diseases (IBD) (6–8). Extensive intracellular metabolism of thiopurine prodrugs is required for therapeutic efficacy across diseases. Particularly for their cytotoxic effects, thiopurines need to be converted to thioguanosine triphosphate (TGTP) which is incorporated into DNA to form DNA-TG, triggering futile DNA damage repair and ultimately apoptosis (9–12).

Genetic variations in genes encoding thiopurine-metabolizing enzymes can directly influence drug toxicity and antileukemic efficacy (13–16). For example, genetic polymorphisms in thiopurine methyltransferase *TPMT* have been linked to susceptibility to thiopurine-induced hematopoietic toxicity in patients, and preemptive *TPMT* genotype-guided dosing is one of the first examples of genetics-based precision medicine in cancer (17, 18). More recently, we and others have identified inherited NUDT15 deficiency as a major genetic cause for thiopurine intolerance in ALL and IBD patients, most frequently in those of Asian and

Significance

Pharmacogenetics is a prototype of genomics-guided precision medicine. While there is a rapid expansion of novel pharmacogenetic variants discovered by genome sequencing, the lack of variant interpretation in a scalable fashion is a formidable barrier in this field. *NUDT15* polymorphism is a major genetic cause for hematopoietic toxicity during thiopurine therapy. Motivated by the need to understand *NUDT15* variant effects for clinical actions, we developed a massively parallel assay to preemptively characterize 91.8% of all possible missense variants in *NUDT15*. Our function-based variant classification accurately predicted thiopurine toxicity risk alleles in patients. These results vastly improved the ability to implement genotype-guided thiopurine therapy and illustrated the value and potential of a high-throughput variant effect screen in general.

Author contributions: D.M.F. and J.J.Y. designed research; C.C.S., T.M., Wentao Yang, R.N., K.H., C.P., L.L., P.S., and J.J.Y. performed research; C.C.S., T.M., K.A.M., L.L., and J.J.Y. contributed new reagents/analytic tools; C.C.S., T.M., K.A.M., Wentao Yang, E.R.S., R.N., Wenjian Yang, M.S., A.T., C.S., D.B., L.Y.P.Y., C.-k.L., C.-h.L., Y.-l.Y., G.J.W., J.R.G., N.A.K., F.A.K., S.B., M.V.R., M.K., H.H., P.B., T.A., A.E.J.Y., P.S., D.M.F., and J.J.Y. analyzed data; and C.C.S., T.M., K.A.M., Wentao Yang, E.R.S., P.S., D.M.F., and J.J.Y. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jun.yang@stjude.org.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915680117/-/DCSupplemental.

First published February 24, 2020.

Hispanic descent (19–21). *NUDT15* encodes a nucleotide diphosphatase that inactivates TGTP by converting it to thioguanosine monophosphate (TGMP). Thus, NUDT15 functions as a negative regulator of intracellular TGTP, with loss-of-function *NUDT15* variants leading to accumulation of DNA-TG and increased cytotoxicity (20, 22).

With the clinical implementation of *NUDT15*-guided thiopurine dosing (18), this gene is frequently sequenced in thiopurine-treated patients, and novel variants are regularly discovered (23). However, functional consequences of these novel *NUDT15* variants remain largely uncharacterized, thus hampering the implementation of individualized thiopurine therapy. In fact, accurate prediction of the phenotypic effect of genetic variation is a particularly formidable challenge in pharmacogenetics (24). Bioinformatic prediction algorithms assess the essentiality of a given gene (or genetic variant) and thus the impact of those genes on fitness. This assumption is valid for genetic variants related to disease pathogenesis; e.g., deleterious variants in tumor suppressor genes would be under negative selection during evolution (25, 26). However, genetic variants in pharmacogenes are not always subjected to purifying selection because many of them are involved only in xenobiotic metabolism and nonessential in normal physiological conditions (27). Therefore, experimental characterization is needed to determine the function of pharmacogene variants, but traditional methods are low throughput, laborious, and outpaced by the rate at which novel variants are discovered.

To address this challenge, we utilized massively parallel variant function assays to exhaustively identify *NUDT15* variants that alter protein abundance and/or thiopurine sensitivity. In this systematic screen, we scored 91.8% of the 3,097 possible missense variants in *NUDT15*, of which 1,152 variants resulted in loss of activity. Our high-resolution variant-activity map pinpoints structural features essential for NUDT15 activity. Applying the functional effects-based *NUDT15* variant classification, we accurately predict thiopurine toxicity risk variants identified in patients treated with this class of drugs for ALL or IBD.

Results

To comprehensively characterize *NUDT15* variant function, we first constructed a site-saturated mutagenesis library of 3,077 missense variants in this gene, representing 99.3% of all possible amino acid changes across the 163 residues in this protein (Fig. 1). This library of variants was then introduced into an engineered HEK293T landing pad cell line at a defined genomic locus via Bxb1-mediated recombination (28, 29). Each cell in the library harbored only a single *NUDT15* variant, enabling multiplex evaluation of variant function by measuring cellular phenotypes of interest. Additionally, each variant NUDT15 sequence was tagged with a set of random barcodes (at an average of 14, ranging from 1 to 54). Barcodes were measured individually in function assays, thereby representing independent observations of each variant. To functionally characterize *NUDT15* variants, we developed two massively parallel assays that measure the abundance of variant protein and variants' effect on thiopurine cytotoxicity, respectively.

We elected to focus on NUDT15 protein abundance as the functional end point for our first assay because the majority of clinically actionable *NUDT15* variants known thus far exhibit a significant decrease in protein thermostability (20, 30). In this high-throughput screen, individual *NUDT15* variants were fused to the *EGFP* gene, with the fluorescence intensity from the fusion protein as a proxy marker of the steady-state abundance of the variant (29). As shown in *SI Appendix*, Fig. S1*A*, wild-type (WT)

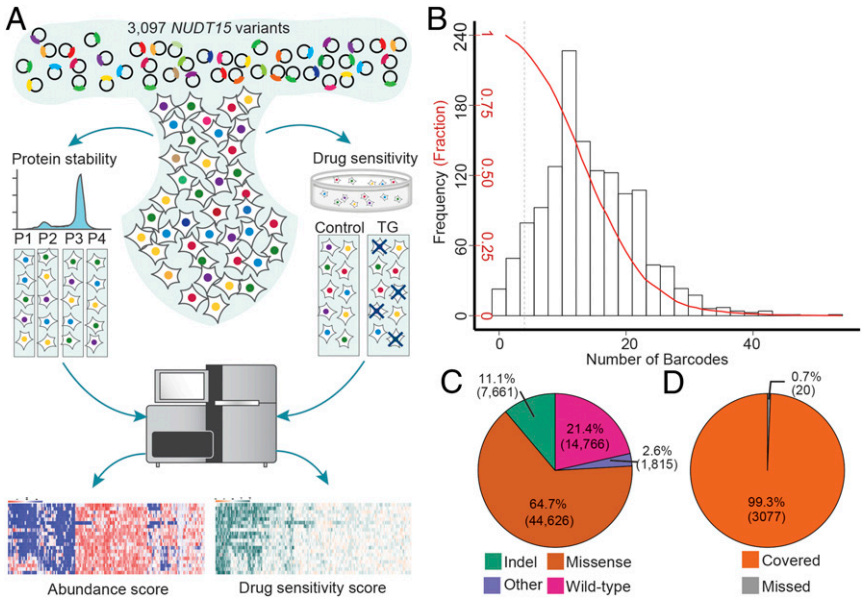

Fig. 1. Massive parallel characterization of variant function in *NUDT15*. (*A*) The *NUDT15* variant library was introduced to HEK293T landing pad cells such that each cell would express only one copy of a specific *NUDT15* variant. Cells were then subjected to two types of phenotyping to determine the effect of a given variant on 1) intracellular NUDT15 protein abundance (measured as the fluorescence intensity of the NUDT15-EGFP fusion protein) or 2) thiopurine cytotoxicity in vitro (measured as the frequency of variant-expressing cells after TG treatment). To estimate a variant's abundance score, cells were flow-sorted into four groups with decreasing fusion protein fluorescence, and variants overrepresented in the low-fluorescence group were associated with NUDT15 protein stability. To assign a variant drug-sensitivity score, each variant was enumerated in library-transfected cells at baseline and after TG treatment in vitro; deleterious variants resulted in low-NUDT15 activity and rendered cells sensitive to thiopurine, which thus became underrepresented after drug exposure. (*B*) Distribution of the number of barcodes in the *NUDT15* library is shown along with the cumulative fraction of barcoded variants (red line). The library included a total of 68,868 unique barcodes, each of which was assigned to a specific variant (median 14 barcodes per variant [ranging from 1 to 54]). (*C*) In the *NUDT15* variant library, there are 44,626 unique barcodes linked to missense variants, as determined by long-read PacBio sequencing. "Other" indicates barcodes linked with more than one variant in the same NUDT15-coding sequence. (*D*) All together, exhaustive mutagenesis of the *NUDT15* gene generated 3,077 missense variants across 163 amino acids, representing 99.3% of all possible missense variations. TG, 6-thioguanine.

NUDT15 tagged with EGFP at the C terminus resulted in a fluorescence signal that was easily detectable by flow cytometry. By contrast, expression of the known low-stability variant Arg139Cys fused in the same manner to EGFP resulted in an approximately three-fold reduction in EGFP signal. Applying this to the *NUDT15* mutagenesis library, we sought to quantify intracellular abundance of all possible variants in parallel. Upon transduction, the population of cells expressing the *NUDT15* library exhibited a left-skewed distribution of normalized EGFP signal, with a predominant peak encompassing WT-expressing cells and a thin tail of cells expressing presumptive destabilizing variants (including Arg139Cys, *SI Appendix*, Fig. S1*B*). Cells were sorted into four equally populated bins with decreasing fluorescence signal, representing variants with decreasing levels of abundance. Subsequent high-throughput sequencing of cells in each bin allowed the calculation of individual variant frequencies from which an abundance score was empirically determined for every variant in the library (ranging from −0.05 to 1.38, Fig. 2*A* and *SI Appendix*, Fig. S2 *A* and *B*). After excluding 154 variants due to low-quality sequencing, we estimated 2,923 abundance scores representing 94.4% of all possible missense variants in this gene, including 858 (95.0%) of 903 possible single-nucleotide variants. On average, each variant was measured 14 times (i.e., 14 independent barcodes per variant, *SI Appendix*, Fig. S2*A*). Notably, 735 variants exhibited an abundance score lower than that of the known toxicity risk variant Arg139Cys, suggesting that they had severe effects on *NUDT15* abundance and possibly thiopurine metabolism.

To validate the results from the high-throughput screen, we selected 14 *NUDT15* variants with a wide range of abundance scores for assessment using orthogonal methods. Their abundance scores from the library screen highly correlated with the EGFP signal of HEK293T cells individually transfected with each variant ($\rho = 0.98$, $P < 2.2 \times 10^{-16}$ by Spearman correlation test, *SI Appendix*, Fig. S3*A*). We also expressed and purified these 14 NUDT15 variant proteins in *Escherichia coli* and subjected them to a thermal stability assay. Variant T_m values normalized to WT protein were strongly correlated with abundance scores ($\rho = 0.85$, $P = 6.0 \times 10^{-5}$ by Spearman correlation test, *SI Appendix*, Fig. S3*B*). Therefore, the high-throughput intracellular abundance screen accurately predicted individual variant protein stability.

This massively parallel variant characterization also revealed biochemical determinants of NUDT15 stability. The active NUDT15 enzyme complex consists of two identical monomers, each with a highly conserved NUDIX motif (22). This motif ($GX_5EX_7REUXEEXGU$) contains amino acids required for coordinating catalytically essential magnesium ions and is formed by an α-helix (α1), a β-sheet (β4), and the connecting loop region. Examining variants across NUDT15 protein, we identified 54 hotspot positions at which genetic variants were more likely to give rise to unstable proteins compared to WT (Fig. 2*B*, $P < 0.01$

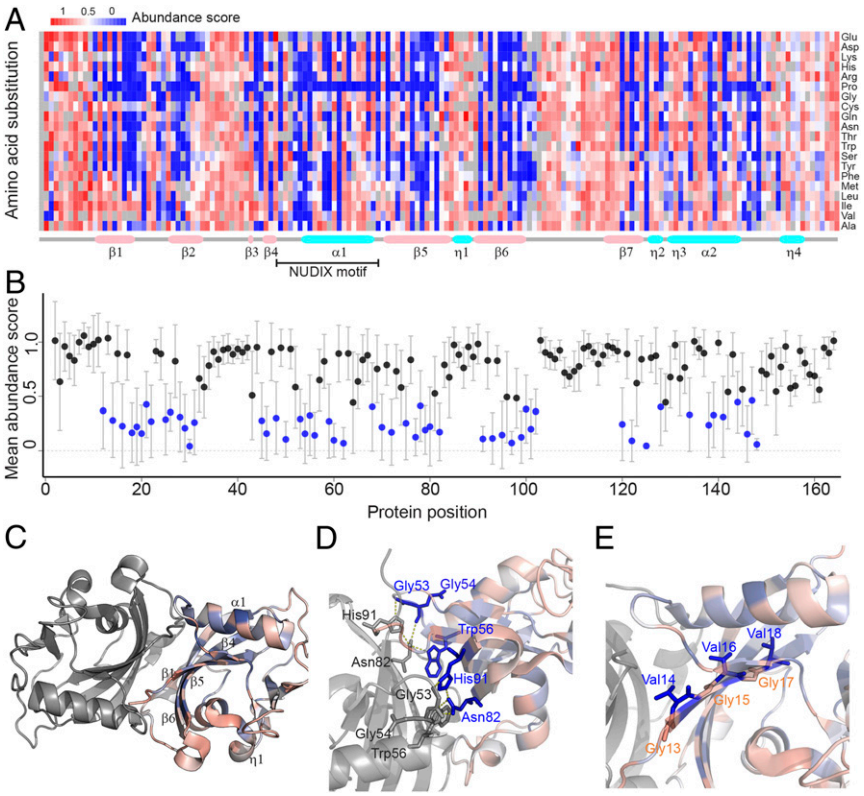

Fig. 2. Effects of genetic variants on NUDT15 protein abundance. (*A*) The NUDT15 abundance score was assigned to 2,922 variants, as plotted in the heatmap. Each column represents an amino acid residue in NUDT15 protein (from 1 to 164), and rows indicate all 19 possible missense changes from the WT sequence at this position. Red to blue denotes high and low protein abundance, respectively. Secondary structures are schematically indicated below the heatmap. (*B*) An average abundance score was calculated for each NUDT15 residue (mean of 19 variants), and positions at which variants consistently encode low-abundance protein were considered as hotspots (54 in total) and highlighted in blue (*Materials and Methods*). Gray lines represent error bars. (*C*) Structural analysis of hotspot residues identified features critical to NUDT15 protein stability. In the three-dimensional structure (Protein Data Bank [PDB] ID code 5LPG), NUDT15 protein is shown as a homodimer with each subunit in either gray or gradient colors representing mean the abundance score, respectively. Residues at the interface between two monomers (Gly53, Gly54, Trp56, Asn82, and His91 in *D*) and α-helix and β-sheets distal to the NUDIX motif (valines at 14, 16, and 18 in GlyVal repeats in *E*) are particularly enriched with hotspot residues. All amino acids listed in *D* and *E* are for WT NUDT15.

compared with all variants, Mann–Whitney–Wilcoxon test). These hotspot residues are unevenly distributed: variations at positions with a small side-chain amino acid (e.g., alanine substitution) were tolerated in general, whereas changes affecting the hydrophobic (valine, proline, leucine, isoleucine, and phenylalanine) or nonpolar aliphatic amino acids (methionine) resulted in large decreases in NUDT15 stability (Fig. 2*A* and *SI Appendix*, Fig. S2*C*). Changes to proline were also poorly tolerated (Fig. 2*A*) because of the unique backbone geometry imposed by its side chain. Some secondary structures were found to be more vulnerable to substitution relative to flexible loop regions. Of the 54 hotspot residues, 68.5% were located in secondary structure elements (α-helices, β-sheets, and 3_{10} helices, Fig. 2*C*), especially beta-strands β1 (odds ratio = 6.19 compared to flexible loop regions, $P = 0.015$ by Fisher exact test), β5 (odds ratio = 3.13 and $P = 0.099$), and β6 (odds ratio = 4.36 and $P = 0.033$). For example, our screen identified a cluster of hotspots in the β1 strand with the valine residues within this glycine/valine repeat consistently vulnerable to genetic variation (Fig. 2*D*). In the crystal structure, these valine side chains strongly interact with the α1-helix to maintain the structure of the NUDIX motif. In fact, a number of indel variants in this region also lead to unstable NUDT15 protein and are associated with thiopurine toxicity in patients (20, 31). Finally, 13 (30.9%) of 42 amino acids in the dimer interface of NUDT15 were extremely vulnerable to genetic variation (Fig. 2*E*), arguing for the importance of intermonomer interaction in maintaining the overall stability of NUDT15 protein.

Although abundance-based screening identified a significant number of deleterious *NUDT15* variants, we postulate that there are other mechanisms by which genetic variants cause loss of function without affecting protein abundance (e.g., alterations of substrate binding). Thus, we performed a secondary massively parallel screen to directly examine the effects of *NUDT15* variants on thiopurine cytotoxicity in vitro. In this assay, HEK293T cells harboring the *NUDT15* variant library were treated with 3 μM TG for 6 d, and high-throughput sequencing was again employed to identify the frequency of each variant prior to drug treatment as well as in cells surviving thiopurine exposure. Because cells expressing loss-of-function NUDT15 variants were more susceptible to thiopurine-induced apoptosis, we estimated a drug-sensitivity score for each variant based on its decrease in frequency after drug treatment (Fig. 1*A*). We successfully evaluated 2,935 variants (94.7% of the library) in this cytotoxicity assay with an average of 14 independent observations per variant, including 866 (95.9%) of all possible single-nucleotide variants (Fig. 3*A* and *SI Appendix*, Fig. S4 *A* and *B*). As an orthogonal validation, we selected nine variants across a range of drug-sensitivity scores and individually tested their impact of thiopurine metabolism in HEK293T cells. Following thiopurine exposure, cells harboring variants with low drug-sensitivity scores

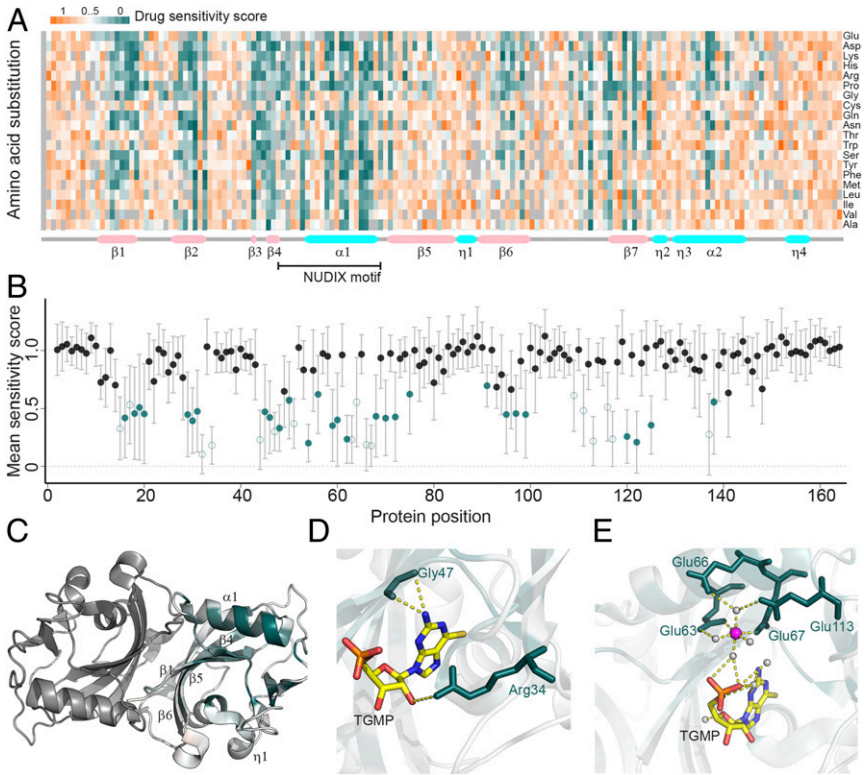

Fig. 3. Effects of *NUDT15* variants on thiopurine cytotoxicity. (*A*) A drug-sensitivity score was assigned for 2,934 variants, as plotted in the heatmap. Each column represents an amino acid residue in NUDT15 protein (from 1 to 164), and rows indicate all 19 possible missense changes from the WT sequence at this position. Red and green denote WT-like and damaging *NUDT15* variants, respectively. Secondary structures are schematically indicated below the heatmap. (*B*) An average drug-sensitivity score was calculated for each NUDT15 residue (mean of 19 variants), and positions at which variants consistently result in increased thiopurine sensitivity were considered as hotspots and highlighted in green (*Materials and Methods*). Among the 45 hotspot residues, 28 were also identified in an abundance-based screen (filled circle), whereas 17 amino acid positions were uniquely sensitive to variation in the drug-sensitivity–based assay (open circle). Gray lines represent error bars. (*C*) Structural analysis of hotspot residues identified variants critical for thiopurine cytotoxicity. In the three-dimensional structure (PDB ID code 5LPG), NUDT15 protein is shown as a homodimer with each subunit in either gray or gradient colors representing a mean drug-sensitivity score, respectively. In particular, residues involved in direct substrate interaction (Arg34 and Gly47 in *D*) and those interacting with the magnesium ion (magenta) or water (gray) (Glu63, Gllu66, Glu67, and Glu113 in *E*) strongly influence catalytic activity without affecting protein stability. TGMP, thioguanosine monophosphate. All amino acids listed in *D* and *E* are for WT NUDT15.

had significantly higher DNA-TG accumulation compared to those expressing high drug-sensitivity score variants (*SI Appendix*, Fig. S5, $\rho = -0.72$, $P = 0.024$, Spearman correlation test).

Similarly, 45 residues were classified as hotspots because variations at each of these positions were more likely to result in increased thiopurine sensitivity compared to all other variants ($P < 0.01$, Mann–Whitney–Wilcoxon test, Fig. 3*B* and *SI Appendix*, Fig. S4*C*). In line with the results from the abundance-based screen, a majority of loss-of-function variants identified from the drug-sensitivity–based assay were also located in secondary structure elements. Thirteen hotspot residues (Gly48, Leu50, Glu51, Glu54, Trp56, Cys59, Ala60, Arg62, Glu63, Thr64, Glu66, Glu67, and Ala68) were located within the highly conserved NUDIX motif (Fig. 3*C*) (22). This motif contains residues responsible for the coordination of magnesium and water molecules and is directly involved in TGTP hydrolysis (30). Interestingly, of the 22 NUDIX motif residues, 10 (Gly48, Leu50, Gly53, Glu54, Thr55, Trp56, Cys59, Ala60, Arg62, and Ala68) were also associated with low abundance. In contrast, *NUDT15* variants affecting Arg34 and Gly47 dramatically altered thiopurine sensitivity with minimal effects on protein stability (Fig. 3*D*). This is also true for residues involved in magnesium coordination (Glu63 and Glu67) or interaction with magnesium-coordinating water molecules (Glu66 and Glu113, Fig. 3*E*).

To summarize the results from the abundance and drug-sensitivity assays detecting different modes of damaging effects, we selected the lower of the two scores as the final NUDT15 activity score (Fig. 4*A*, and *SI Appendix*, Fig. S6 *A*–*C*). We defined variants below 0.45 as damaging after modeling the pattern of activity score distribution, representing 1,152 (40.5%) of all 2,844 variants and 280 (31.0%) of all 903 single-nucleotide variants. Our systematic experimental characterization also allowed for a direct comparison with damaging effects predicted in silico by a number of different algorithms: the combined annotation-dependent depletion (CADD) score, the rare exome variant ensemble learner (REVEL) score (Fig. 4 *B* and *C*), polymorphism phenotyping (PolyPhen2), and sorting intolerant from tolerant (*SI Appendix*, Fig. S6 *D* and *E*). With a CADD score >20 as the criterion for damaging variants (25, 32), 561 variants were predicted as loss-of-activity, of which 293 exhibited a high NUDT15 activity score (i.e., >0.45) with a false discovery rate of 52.2%. Conversely, of 276 CADD-predicted benign variants, 12 had a low NUDT15 activity score with a false omission rate of 4.3%. Similarly, when we performed this analysis using the REVEL prediction [>0.5 as damaging (26)], we observed a false discovery rate and false omission rate of 14.1 and 25.8%, respectively.

To apply our functional effects-based variant classification, we sought to identify *NUDT15* variants in patients treated with thiopurine drugs and ask if the activity score could predict pharmacogenetic variants associated with toxicity. In 2,398 subjects, we identified a total of 10 missense coding variants, of which 6 were exceedingly rare. Cases carrying *TPMT* variants were excluded for further analyses. Five variants were associated with hematopoietic toxicity secondary to thiopurines (Lys33Glu, Arg34Thr, Val75Gly, Arg139Cys, and Arg139His), whereas five were not (Gln6Glu, Arg11Gln, Val18Ile, Ser83Tyr, and Val93Ile, *SI Appendix*, Table S1). The activity score averaged 0.20 for five toxicity variants (range from 0.04 to 0.35), which is significantly lower than that of variants not linked to toxicity (mean of 0.74, ranging from 0.46 to 0.98, $P = 0.0079$ by Mann–Whitney–Wilcoxon test, Fig. 4*D*). With 0.45 as the activity score cutoff, we estimated both sensitivity and specificity at 100% with a 95% confidence interval (CI) from 57 to 100%. In contrast, CADD (>20) or REVEL (>0.5) scores did not accurately predict variants' effect on thiopurine toxicity, with sensitivity and specificity at 100% (CI: 57 to 100%), 40% (CI: 12 to 77%), 20% (CI: 4 to 62%), and

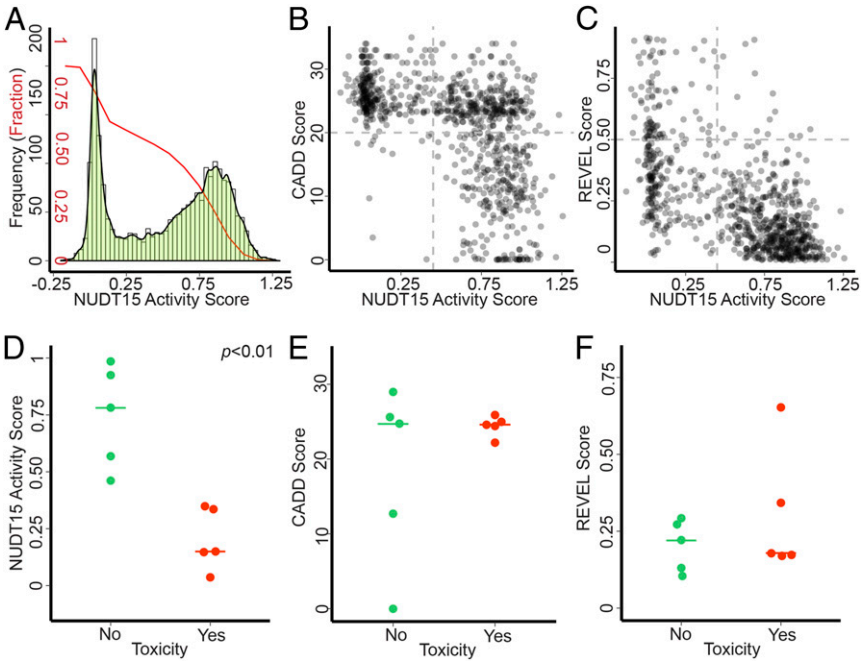

Fig. 4. NUDT15 activity score predicted clinical thiopurine toxicity. (*A*) Combining the abundance score and the drug-sensitivity score, we assigned a final NUDT15 activity score for 2,844 variants, the distribution of which is plotted along with the cumulative fraction of variants (red line). Comparison of the experimentally determined NUDT15 activity score with effects predicted by bioinformatic algorithms (i.e., *B* and *C* for CADD and REVEL scores, respectively). CADD score (>20 as damaging) and REVEL score (>0.5 as damaging) were available for 837 missense variants. Horizontal and vertical dashed lines in *B* and *C* represent the cutoff for the NUDT15 activity score, CADD score, and REVEL score, respectively. *NUDT15* variants were identified by sequencing in 2,398 patients exposed to thiopurine therapy and were classified as toxicity related or nontoxicity related (*Materials and Methods*). NUDT15 activity (*D*), CADD (*E*), or REVEL scores (*F*) were plotted for risk vs. benign variants, with the *P* value estimated by using the Mann–Whitney–Wilcoxon test.

100% (CI: 57 to 100%), respectively (Fig. 4 *E* and *F* and *SI Appendix*, Table S4). The receiver operating characteristic curve analysis projected a probability of 1.0 of accurate prediction of risk variants using the activity score (area under the curve statistic [AUC]), whereas AUC was 0.48 (CI: 0.08 to 0.92) and 0.64 (CI: 0.32 to 1.0) for the CADD- and REVEL-based predictions (*SI Appendix*, Fig. S7 and Table S4).

Querying the publicly available Genome Aggregation Database (gnomAD) of whole-genome/exome sequencing of 141,456 individuals, we identified 108 *NUDT15* missense alleles observed in humans with the population frequency ranging from 0.004 to 2% (Fig. 5*A*). Damaging variants were detected in all populations regardless of ancestry, and the average NUDT15 variant activity score did not differ by race or ethnicity (Fig. 5*B*). Of 8,871 individuals in gnomAD with *NUDT15* variation, 8,323 (93.8%) have a damaging variant and thus are at risk for thiopurine toxicity (Fig. 5*C*). After excluding individuals with the Arg139Cys variant, 931 (63.2%) of 1,472 subjects are at risk for thiopurine toxicity based on their variant activity score, significantly higher than what would be expected by chance ($P < 2.2 \times 10^{-16}$, Fisher exact test).

Discussion

Accurate annotation of the phenotypic effect conferred by genetic variation is of critical importance for the implementation of genomics-guided precision medicine (24). This has become a pressing challenge in recent years with the explosive growth of genome sequencing and the sheer number of novel variants that need to be functionally characterized. In fact, 48.8% of variants cataloged in the ClinVar database are considered of "unknown significance" primarily due to the lack of experimental validation of their functional effects (33). Moreover, the vast majority of human genetic variants are rare (34), for which statistical association with clinical phenotypes would be exceedingly difficult to establish. Thus, functional experiments are essential for predicting the clinical consequences of rare genetic variants. To address this, a number of groups have utilized high-throughput genomics platforms to evaluate variant function at scale (24, 29, 35, 36). Similarly, we report a preemptive large-scale screen of functional variants in *NUDT15* and identify 1,152 loss-of-function variants, and our clinical validation study confirmed the predicted association with thiopurine toxicity. These results provide a comprehensive catalog of all possible missense pharmacogenetic variants in this important drug-metabolizing gene, vastly improving the ability to implement genotype-guided treatment individualization.

Scalable functional assays are of particular importance for pharmacogenetic variants because computational predictions lack sufficient accuracy to be relied upon clinically (37). For example, comparing *NUDT15* variants experimentally identified as damaging vs. those predicted using CADD, we observed that the bioinformatic algorithm classified a substantial proportion of variants as loss-of-function even though they showed no effects on protein stability or thiopurine cytotoxicity (Fig. 4 *B* and *C*). Even though both CADD and REVEL identified some damaging variants in *NUDT15*, their low specificity or sensitivity would render these bioinformatic predictions unsuitable for clinical implementation. One potential explanation for this is that most computational prediction methods rely on (to varying degrees) evolutional conservation of genetic variation with presumable effects on human fitness (e.g., tumor suppressors). However, this assumption that deleterious variants would be selected out during evolution is often invalid for genes involved in drug metabolism. For example, NUDT15 is postulated to degrade oxo-dGTP as a means to mitigate DNA damage, but this endogenous substrate can also be removed by a number of other

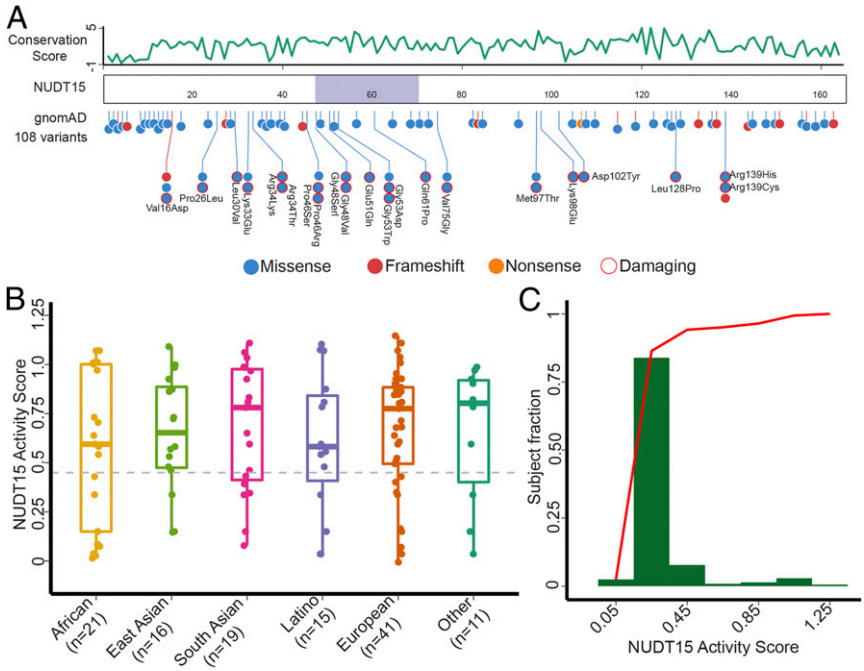

Fig. 5. Population distribution of *NUDT15* variants in humans. (*A*) 108 *NUDT15* variants are identified in whole genome or whole exome seq data of 141,456 individuals in the gnomAD database (https://gnomad.broadinstitute.org/). In the lollipop plot, each circle represents a *NUDT15* variant (blue, red, or orange for missense, frameshift, or nonsense, respectively). Damaging *NUDT15* variants are annotated with the exact amino acid change and open red circles. Average PhastCons scores of trinucleotide are shown by the green line at the top for each amino acid residue to indicate cross-species conservation. The Nudix motif is highlighted in purple. (*B*) Activity score is plotted for NUDT15 variants observed in each of five major race/ethnic groups in the gnomAD cohort. Damaging variants (NUDT15 activity score <0.45, dashed line) were present across populations. The numbers of variants identified in each population are shown in parentheses. (*C*) A percentage of 91.6% of individuals affected by *NUDT15* polymorphism carry a variant that results in significant loss of NUDT15 activity. This is in part explained by the common variant Arg139Cys (with an activity score of 0.15).

NUDIX enzymes (22, 38, 39). As a result, the loss of NUDT15 may be inconsequential under normal physiological conditions. Until the introduction of pharmaceutical agents a few hundred years ago, there was no selection pressure against genetic variants related to drug-induced toxicities, and therefore this type of prediction model performs poorly in this context. That said, it is formally plausible that CADD or REVEL identify modes of damaging effects that are not reflected in our two functional assays or that these variants have subtle effects that our assays were not sufficiently sensitive to detect.

There are a number of caveats with our choice of experimental end points for *NUDT15* variant characterization. In particular, the thiopurine cytotoxicity-based screen, while successful at determining variants that most severely alter cell sensitivity to this drug (presumably via effects on thiopurine metabolism), is likely limited in its sensitivity for two reasons. First, our screen employed HEK293T cells which are known to lack certain components of the mismatch repair system (40) and thus are somewhat more resistant to cell-cycle arrest than mismatch-repair–proficient cell lines. As a result, the high background of thiopurine resistance in HEK293T cells may have masked the effects of variants that modestly influence drug sensitivity. Second, we believe that the degree of variant drop-out after thiopurine exposure is highly dependent on the drug concentration used and the length of incubation. While we did discover variants causing the most deleterious alterations with a probable structural basis, variants with intermediate metabolic activity may be able to metabolize enough drug to persist over the course of our selection. This is illustrated by the Arg139Cys variant which resulted in an 85% reduction in the abundance assay compared to WT but only a 13% decrease in the drug-sensitivity assay.

It should also be noted that our activity-score–based prediction analysis included only 10 *NUDT15* variants and thus had a limited statistical power to accurately estimate its sensitivity and specificity (*SI Appendix*, Table S4). As more *NUDT15* variants are discovered in patients, future studies are warranted to carefully evaluate the performance of activity score using a larger number of variants and sample sizes.

In conclusion, we report the results of a deep mutational scan of *NUDT15* for identifying pharmacogenetic variants, creating a comprehensive reference of risk alleles to enable preemptive tailored thiopurine therapy. Our findings also point to the critical importance and exciting potential of high-throughput variant annotation in pharmacogenes in general.

Materials and Methods

***NUDT15* Mutagenesis Library and the Landing Pad Cellular Model.** The *NUDT15* variant library was synthesized (Twist Biosciences) to exhaustively introduce missense variants across the coding region of this gene, followed by the addition of random sequence barcodes, using procedures previously described (29). This barcoded *NUDT15* library was integrated into the *AAVS1* locus in the HEK293T landing pad cell line such that each cell would express a single *NUDT15* variant fused with EGFP (28, 29). Cells with successful recombination were identified by flow cytometry (BFP-negative and mCherry-positive) with which *NUDT15* variant characterization was performed subsequently. Primers used for library construction and cloning are listed in *SI Appendix*, Table S2. Detailed descriptions of these experiments are provided in *SI Appendix, Supplementary Information Text*.

Protein-Abundance–Based Screen of *NUDT15* Variants. Abundance score of each variant was determined using the VAMP-seq method (29) (Fig. 1*A* and *SI Appendix, Supplementary Information Text*). Briefly, library-expressing HEK293T cells were first sorted into four different bins depending on the level of EGFP normalized to mCherry by flow cytometry. Massive parallel sequencing was then performed to quantify every variant in each of the four bins, from which a variant abundance score was modeled to indicate its intracellular protein abundance (Dataset S1).

Thiopurine-Cytotoxicity–Based Screen of *NUDT15* Variants. For the thiopurine-cytotoxicity–based screen, *NUDT15* library-expressing cells were treated with 3 μM TG or culture media in vitro for 6 d. Cells were then harvested for genomic DNA extraction, and massive parallel sequencing was performed to quantify variant frequency as described above. The final variant-barcode-count table was used as the input for the ABSSeq pipeline (41) to normalize and identify variants with differential frequency between drug-treated and nontreated cells (using an aFold module). Fold change (*s*) in variant frequency was used to estimate the drug-sensitivity score for each variant (Dataset S1) using a similar procedure for the abundance score (*SI Appendix, Supplementary Information Text*). Each drug treatment had four replicates. Because thiopurine drug sensitivity is related to the rate of cell proliferation, we also evaluated variants for their impact on HEK293T cell growth to avoid confounding effects (*SI Appendix, Supplementary Information Text*, Table S5, and Fig. S4 *D* and *E*).

Hotspot Residues and Structural Analysis. Hotspot analysis was performed to identify residues at which genetic variation caused consistent damaging effects as follows: for each amino acid residue, we compared the distribution of the abundance score or the drug-sensitivity score of all 19 variants with that of all *NUDT15* variants in the library; those with lower-than-population average were considered as hotspot residues [$P < 0.05$ using Mann–Whitney–Wilcoxon Test, adjusted by the Benjamini–Hochberg approach (42)]. t-SNE analysis was also performed to visualize similarity across NUDT15 residues in the effects of their genetic variants, using the Rtsne algorithm (43).

For the structural analysis, a color gradient representing the mean abundance score or mean drug-sensitivity score was mapped to the NUDT15 protein structure (5LPG, https://www.rcsb.org/). A total of 42 interface residues were identified by querying Proteins, Interfaces, Structures, and Assemblies (PISA) of the European Bioinformatics webserver (https://www.ebi.ac.uk/pdbe/pisa/) (44). Structural analyses of hotspot residues were performed using PyMOL (version 2.0, Schrödinger).

NUDT15 Activity Score. The smaller value of the abundance score and drug-sensitivity score was then assigned as the final "NUDT15 activity score" for each variant (Dataset S1). The cutoff of the activity score (0.45) was selected by Ckmeans.1d.dp on the basis of bimodal distribution of the scores (45), which identifies clusters by minimizing the within-cluster Euclidian distances (K-means).

Association of *NUDT15* Variants with Thiopurine Toxicity in Patients. To identify *NUDT15* variants and evaluate their effects on thiopurine toxicity in patients, we assembled a cohort of 2,398 patients treated with thiopurine for ALL or IBD, including 1,404 subjects in previously published datasets (20, 21, 31, 46, 47). The ALL cohort included US Children's Oncology Group trial AALL03N1 ($n = 646$), Singapore/Malaysia MaSpore ALL 2003 trial ($n = 140$), Japanese Children's Cancer Group ALL B_{12} ($n = 116$), Guatemalan LLAG-0707 study ($n = 181$), Taiwan TPOG ALL study ($n = 1$), Hong Kong CCCG-ALL2015 trial ($n = 132$), and the Indian Childhood Collaborative Leukemia Group study ($n = 105$), for which targeted sequencing was performed for all three *NUDT15* exons. IBD subjects were from the Exeter pharmacogenetic PRED4 study in the United Kingdom ($n = 1,077$) with *NUDT15* variants identified by whole-exome sequencing (31). *TPMT* risk variants (e.g., rs1800462) were also genotyped as previously reported (20, 21, 31, 46, 47), and cases carrying *TPMT* variants were excluded from further analysis. This study was approved by the respective institutional review boards, and informed consent was obtained from the parents, guardians, and/or patients, as appropriate. Collectively, we identified a total of 10 missense variants (Arg139Cys, Arg139His, Val18Ile, Gln6Glu, Arg11Gln, Lys33Glu, Arg34Thr, Val75Gly, Ser83Tyr, and Val93Ile). To compare the allelic effect across variants, we excluded cases with homozygous or compound-heterozygous *NUDT15* genotype, and the association with toxicity was evaluated based on the difference between heterozygous cases and individuals with WT *NUDT15* (*SI Appendix*, Table S1). Details of this analysis are provided in *SI Appendix, Supplementary Information Text*.

Data Availability. All data discussed in the paper are available in the main text and *SI Appendix*.

ACKNOWLEDGMENTS. We thank the patients and parents who participated in the clinical studies included in this report. This work was supported by the NIH (Grants CA021765, GM115279, GM118578, R01CA096670, U10CA098543, U10CA098413, U10CA095861, U10CA180886, and U10CA180899); the American Lebanese Syrian Associated Charities of St. Jude Children's Research Hospital; the V Foundation for Cancer Research; and Alex's Lemonade Stand Foundation (T.M.). This work was also supported by the Swedish Research Council and the Swedish Cancer Society (P.S.).

1. A. Vora *et al*., Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): A randomised controlled trial. *Lancet Oncol.* **14**, 199–209 (2013).
2. C.-H. Pui, W. L. Carroll, S. Meshinchi, R. J. Arceci, Biology, risk stratification, and therapy of pediatric acute leukemias: An update. *J. Clin. Oncol.* **29**, 551–565 (2011).
3. J. S. Maltzman, G. A. Koretzky, Azathioprine: Old drug, new actions. *J. Clin. Invest.* **111**, 1122–1124 (2003).
4. P. Karran, N. Attard, Thiopurines in current medical practice: Molecular mechanisms and contributions to therapy-related cancer. *Nat. Rev. Cancer* **8**, 24–36 (2008).
5. G. B. Elion, The purine path to chemotherapy. *Science* **244**, 41–47 (1989).
6. W. Reinisch *et al*.; International AZT-2 Study Group, Azathioprine versus mesalazine for prevention of postoperative clinical recurrence in patients with Crohn's disease with endoscopic recurrence: Efficacy and safety results of a randomised, double-blind, double-dummy, multicentre trial. *Gut* **59**, 752–759 (2010).
7. N. K. de Boer, A. A. van Bodegraven, B. Jharap, P. de Graaf, C. J. Mulder, Drug insight: Pharmacology and toxicity of thiopurine therapy in patients with IBD. *Nat. Clin. Pract. Gastroenterol. Hepatol.* **4**, 686–694 (2007).
8. R. Goldberg, P. M. Irving, Toxicity and response to thiopurines in patients with inflammatory bowel disease. *Expert Rev. Gastroenterol. Hepatol.* **9**, 891–900 (2015).
9. P. F. Swann *et al*., Role of postreplicative DNA mismatch repair in the cytotoxic action of thioguanine. *Science* **273**, 1109–1111 (1996).
10. G.-M. Li, The role of mismatch repair in DNA damage-induced apoptosis. *Oncol. Res.* **11**, 393–400 (1999).
11. Y. H. Ling, J. A. Nelson, Y. C. Cheng, R. S. Anderson, K. L. Beattie, 2′-Deoxy-6-thioguanosine 5′-triphosphate as a substrate for purified human DNA polymerases and calf thymus terminal deoxynucleotidyltransferase in vitro. *Mol. Pharmacol.* **40**, 508–514 (1991).
12. S. Yoshida, M. Yamada, S. Masaki, M. Saneyoshi, Utilization of 2′-deoxy-6-thioguanosine 5′-triphosphate in DNA synthesis in vitro by DNA polymerase α from calf thymus. *Cancer Res.* **39**, 3955–3958 (1979).
13. G. Tzoneva *et al*., Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat. Med.* **19**, 368–371 (2013).
14. J. A. Meyer *et al*., Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. *Nat. Genet.* **45**, 290–294 (2013).
15. B. Li *et al*., Negative feedback-defective PRPS1 mutants drive thiopurine resistance in relapsed childhood ALL. *Nat. Med.* **21**, 563–571 (2015).
16. M. V. Relling *et al*., Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus. *J. Natl. Cancer Inst.* **91**, 2001–2008 (1999).
17. E. Y. Krynetski *et al*., A single point mutation leading to loss of catalytic activity in human thiopurine S-methyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 949–953 (1995).
18. M. V. Relling *et al*., Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for thiopurine dosing based on TPMT and NUDT 15 genotypes: 2018 update. *Clin. Pharmacol. Ther.* **105**, 1095–1105 (2019).
19. S.-K. Yang *et al*., A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. *Nat. Genet.* **46**, 1017–1020 (2014).
20. T. Moriyama *et al*., NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. *Nat. Genet.* **48**, 367–373 (2016).
21. J. J. Yang *et al*., Inherited NUDT15 variant is a genetic determinant of mercaptopurine intolerance in children with acute lymphoblastic leukemia. *J. Clin. Oncol.* **33**, 1235–1242 (2015).
22. M. Carter *et al*., Crystal structure, biochemical and cellular activities demonstrate separate functions of MTH1 and MTH2. *Nat. Commun.* **6**, 7871 (2015).
23. J. J. Yang *et al*., Pharmacogene Variation Consortium Gene introduction: NUDT15. *Clin. Pharmacol. Ther.* **105**, 1091–1094 (2019).
24. L. M. Starita *et al*., Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
25. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
26. N. M. Ioannidis *et al*., REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
27. L. F. Jorge, M. Eichelbaum, E.-U. Griese, T. Inaba, T. D. Arias, Comparative evolutionary pharmacogenetics of CYP2D6 in Ngawbe and Embera Amerindians of Panama and Colombia: Role of selection versus drift in world populations. *Pharmacogenetics* **9**, 217–228 (1999).
28. K. A. Matreyek, J. J. Stephany, D. M. Fowler, A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).
29. K. A. Matreyek *et al*., Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
30. N. C. Valerie *et al*., NUDT15 hydrolyzes 6-Thio-deoxyGTP to mediate the anticancer efficacy of 6-thioguanine. *Cancer Res.* **76**, 5501–5511 (2016).
31. G. J. Walker *et al*.; IBD Pharmacogenetics Study Group, Association of genetic variants in NUDT15 with thiopurine-induced Myelosuppression in patients with inflammatory bowel disease. *JAMA* **321**, 773–785 (2019).
32. M. Kircher *et al*., A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
33. M. J. Landrum *et al*., ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
34. J. A. Tennessen *et al*., Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
35. G. M. Findlay *et al*., Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
36. A. O. Giacomelli *et al*., Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).
37. R. Ghosh, N. Oak, S. E. Plon, Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol.* **18**, 225 (2017).
38. Y. Takagi *et al*., Human MTH3 (NUDT18) protein hydrolyzes oxidized forms of guanosine and deoxyguanosine diphosphates: Comparison with MTH1 and MTH2. *J. Biol. Chem.* **287**, 21541–21549 (2012).
39. J. Carreras-Puigvert *et al*., A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family. *Nat. Commun.* **8**, 1541 (2017).
40. E. Cannavo *et al*., Expression of the MutL homologue hMLH3 in human cells and its role in DNA mismatch repair. *Cancer Res.* **65**, 10759–10766 (2005).
41. W. Yang, P. C. Rosenstiel, H. Schulenburg, ABSSeq: A new RNA-seq analysis method based on modelling absolute expression differences. *BMC Genomics* **17**, 541 (2016).
42. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
43. L. van der Maaten, Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res* **15**, 3221–3245 (2014).
44. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
45. H. Wang, M. Song, Ckmeans.1d.Dp: Optimal k-means clustering in one dimension by dynamic programming. *R Journal* **3**, 29–33 (2011).
46. T. Moriyama *et al*., The effects of inherited NUDT15 polymorphisms on thiopurine active metabolites in Japanese children with acute lymphoblastic leukemia. *Pharmacogenet. Genomics* **27**, 236–239 (2017).
47. T. Moriyama *et al*., Novel variants in *NUDT15* and thiopurine intolerance in children with acute lymphoblastic leukemia from diverse ancestry. *Blood* **130**, 1209–1212 (2017).