OPEN ACCESS

ORIGINAL ARTICLE

A candidate gene study of capecitabine-related toxicity in colorectal cancer identifies new toxicity variants at *DPYD* and a putative role for *ENOSF1* rather than *TYMS*

Dan Rosmarin,[1,2] Claire Palles,[1] Alistair Pagnamenta,[3] Kulvinder Kaur,[3] Guillermo Pita,[4] Miguel Martin,[5] Enric Domingo,[1,3] Angela Jones,[1] Kimberley Howarth,[1] Luke Freeman-Mills,[1] Elaine Johnstone,[2] Haitao Wang,[2] Sharon Love,[6] Claire Scudder,[7] Patrick Julier,[7] Ceres Fernández-Rozadilla,[1] Clara Ruiz-Ponte,[8] Angel Carracedo,[8] Sergi Castellvi-Bel,[9] Antoni Castells,[10] Anna Gonzalez-Neira,[4] Jenny Taylor,[3] Rachel Kerr,[2] David Kerr,[11] Ian Tomlinson[1,3]

▶ Additional material is published online. To view please visit the journal (http://dx.doi.org/10.1136/gutjnl-2013-306571).

For numbered affiliations see end of article.

Correspondence to
Professor Ian Tomlinson, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK; iant@well.ox.ac.uk

DR and CP contributed equally.

Received 11 December 2013
Revised 13 February 2014
Accepted 15 February 2014
Published Online First
19 March 2014

To cite: Rosmarin D, Palles C, Pagnamenta A, *et al*. *Gut* 2015;**64**: 111–120.

ABSTRACT

Objective Capecitabine is an oral 5-fluorouracil (5-FU) pro-drug commonly used to treat colorectal carcinoma and other tumours. About 35% of patients experience dose-limiting toxicity. The few proven genetic biomarkers of 5-FU toxicity are rare variants and polymorphisms, respectively, at candidate loci dihydropyrimidine dehydrogenase (*DPYD*) and thymidylate synthase (*TYMS*).

Design We investigated 1456 polymorphisms and rare coding variants near 25 candidate 5-FU pathway genes in 968 UK patients from the QUASAR2 clinical trial.

Results We identified the first common *DPYD* polymorphisms to be consistently associated with capecitabine toxicity, rs12132152 (toxicity allele frequency (TAF)=0.031, OR=3.83, $p=4.31\times10^{-6}$) and rs12022243 (TAF=0.196, OR=1.69, $p=2.55\times10^{-5}$). rs12132152 was particularly strongly associated with hand-foot syndrome (OR=6.1, $p=3.6\times10^{-8}$). The rs12132152 and rs12022243 associations were independent of each other and of previously reported *DPYD* toxicity variants. Next-generation sequencing additionally identified rare *DPYD* variant p.Ala551Thr in one patient with severe toxicity. Using functional predictions and published data, we assigned p.Ala551Thr as causal for toxicity. We found that polymorphism rs2612091, which lies within an intron of *ENOSF1*, was also associated with capecitabine toxicity (TAF=0.532, OR=1.59, $p=5.28\times10^{-6}$). *ENSOF1* is adjacent to *TYMS* and there is a poorly characterised regulatory interaction between the two genes/proteins. Unexpectedly, rs2612091 fully explained the previously reported associations between capecitabine toxicity and the supposedly functional *TYMS* variants, 5′VNTR 2R/3R and 3′UTR 6 bp ins-del. rs2612091 genotypes were, moreover, consistently associated with *ENOSF1* mRNA levels, but not with *TYMS* expression.

Conclusions *DPYD* harbours rare and common capecitabine toxicity variants. The toxicity polymorphism in the *TYMS* region may actually act through *ENOSF1*.

Significance of this study

What is already known on this subject?

- Chemotherapy based on 5-fluorouracil (5-FU) is used for the treatment of several common types of cancer. For colorectal carcinoma, the 5-FU pro-drug capecitabine is often given routinely after surgery for stage II or III tumours.
- The benefits of 5-FU are modest (<5% increased long-term survival) and must be weighed against drug toxicity. Inherited variation in genes involved in 5-FU metabolism can explain some of this toxicity.
- However, only four genetic variants in the 5-FU pathway have good prior evidence of association with toxicity.

What are the new findings?

- We found four new associations between genetic variants and capecitabine toxicity.
- Three of these associations involved variants, two common and one rare, at the dihydropyrimidine dehydrogenase (*DPYD*) locus.
- The fourth association involved a polymorphism within *ENOSF1*, a gene adjacent to thymidylate synthase (*TYMS*).
- Statistically, the ENOSF1 association completely explains two previously reported 5-FU toxicity polymorphisms (5′VNTR and 3′UTR) in the thymidylate synthase gene.
- Most patients with severe myelosuppression carried a rare *DPYD* allele: *2A, 2846T>A, *13 or A551T.

INTRODUCTION

Capecitabine (Xeloda, Roche) is an oral 5-fluorouracil (5-FU) pro-drug commonly given to patients with colorectal cancer (CRC). Capecitabine is activated to 5-FU, which then causes cytotoxicity by inhibiting production of thymidine and by being converted to metabolites that are incorporated into DNA and RNA.[1] As with other 5-FU-based chemotherapy

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Significance of this study

How might it impact on clinical practice in the foreseeable future?

- The genetic architecture of 5-FU/capecitabine toxicity is complex and encompasses rare and common variants.
- Panels of markers in tests used to predict clinically actionable 5-FU/capecitabine toxicity should be updated to include the new *DPYD* and *ENOSF1* alleles, while omitting the *TYMS* 5′VNTR and 3′UTR polymorphisms.
- Rare loss-of-function *DPYD* alleles remain the only genetic variants proven to have large positive predictive values for 5-FU/capecitabine toxicity, and these variants account for the majority of patients with life-threatening myelosuppression induced by capecitabine.

regimens, approximately one-third of capecitabine patients suffer dose-limiting levels of drug-induced adverse events. The rapid onset of toxicity results in mortality for 0.5–2% of patients in monotherapy and combination regimens of infusional and bolus 5-FU,[2] and about half that number for capecitabine schedules.

The most common dose-limiting capecitabine toxicities are hand-foot syndrome (HFS) and diarrhoea. Additionally, an important proportion of patients develop neutropaenia and thromboctytopaenia, and others experience nausea, vomiting, mucositis and stomatitis. Some interpatient differences in toxicity can be explained by clinical factors, such as age, gender, local clinical practice and, possibly, diet.[3–5] However, much variability in toxicity remains unexplained.

The biochemical pathway of capecitabine activation and subsequent 5-FU action and degradation is well established and provides 25 candidate genes in which variation might affect 5-FU toxicity (figure 1, and see online supplementary table 1).[6] Upon absorption in the gut, capecitabine is partially converted to 5-FU in the liver, then preferentially converted to 5-FU at the CRC site. Much 5-FU is degraded in the liver by dihydropyrimidine dehydrogenase (DPYD) prior to activation. As part of the drug's rationally designed activation, 5-FU is further activated in the tumour to cytotoxic compounds that inhibit DNA synthesis by competing with nucleotide precursors for binding with thymidylate synthase (TYMS). Various sources of toxicity may exist, including alternative activation pathways outside the tumour that result in direct DNA/RNA damage through incorporation, undesired transport of activated compounds, variable expression of drug targets, and reduced levels of drug degradation.

Over a decade of publications exists regarding inherited genetic biomarkers of 5-FU-related toxicity, but only a handful of polymorphisms and rare genetic variants associated with toxicity have been identified with high confidence.[7] These include two common polymorphisms in *TYMS* (5′VNTR 2R/3R and 3′UTR 6 bp ins-del) and rare functionally deleterious *DPYD* variants, chiefly *2A and 2846T>A.[7 8] In aggregate, these variants are potentially useful, but suboptimal, for the prediction of toxicity in clinical practice. Furthermore, there is only limited evidence that genetic variants are generalisable as predictors of toxicity across 5-FU regimens.[7]

In this study, we have investigated the 25 candidate genes that act in the capecitabine/5-FU pathway for new common and rare genetic variants that are associated with capecitabine toxicity. Our patient set comprises about 1000 individuals from the QUASAR2 trial of capecitabine±bevacizumab (Avastin, Genentech/Roche). Our results show that comprehensive genetic studies in sufficiently large, homogeneous datasets can identify additional toxicity predisposition variants in known candidate genes.

METHODS

Patients

The QUASAR2 study (http://www.octo-oxford.org.uk/alltrials/infollowup/q2.html; http://www.controlled-trials.com/ISRCTN45133151/) is a phase III randomised controlled trial of adjuvant capecitabine (Xeloda) (1250 mg/m^2 twice daily for 14 days every 3 weeks, total of 8 cycles) ± bevacizumab (Avastin) (7.5 mg/kg every 3 weeks) following resection of stage II/III CRC. Patients were entered into the study between July 2005 and December 2011 at 123 UK and 81 non-UK sites. Survival analyses are scheduled for 2014. Of 1119 patients with blood collected as of July 2010, 1046 were selected for genetic study based on availability of clinical data and informed consent. Patient characteristics are shown in table 1. All work was performed with full UK Ethics Committee approval, according to the tenets of the Declaration of Helsinki.

Toxicity phenotypes

Toxicity phenotype data were collected as part of QUASAR2 according to the NCI Common Toxicity Criteria for Adverse Events (CTCAE) V.3.0. Maximum toxicity (0–4) at any treatment cycle was derived for each of the following individual FU-related toxicities: diarrhoea, nausea and vomiting, mucositis/stomatitis, neutropaenia, thrombocytopaenia and HFS. We then derived a global measure of toxicity, defined as the maximum individual toxicity score measured for each patient. Global and individual toxicities were analysed using two approaches: (1) a binary classification into low toxicity (grade 0–2) versus high (dose-limiting) toxicity (grade 3–4) and (2) a quantitative measure of toxicity. In the latter, if <100 patients experienced a particular grade of toxicity, we combined them into a single bin with an adjacent grade. Specifically, we analysed grades (0–1 v 2 v 3–4) for global toxicity, diarrhoea and HFS and 0v1234 for the other, rarer toxicity phenotypes. Toxicity data by grade are shown in online supplementary table 2.

Genotype data

Exclusion of patients on the basis of availability of DNA and completeness of toxicity data, non-Caucasian ethnicity or genotyping quality control (see online supplementary methods)[9] meant that data were available from 940 patients genotyped using Illumina tagging SNP arrays and a largely overlapping set of 968 QUASAR2 patients genotyped on Illumina HumanExome12v1_A or -12v1-1_A arrays, which were designed to capture uncommon protein-coding variation.[10 11] For each of the 25 capecitabine/5-FU pathway genes (see online supplementary table 1), we identified genetic variants that were present on the arrays and that lay within 25 kb either side of the coding region of the longest isoform of each gene. We used imputation to obtain missing genotypes arising from differences in array content.[12–14] Further genotyping was performed using previously described methods.[15–17]

For loci at which significant or borderline significant associations between genetic variants and toxicity were detected, we performed fine mapping studies by using the methods above to impute all SNPs in a 1.5 Mb flanking region, in order to refine the association signal.

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

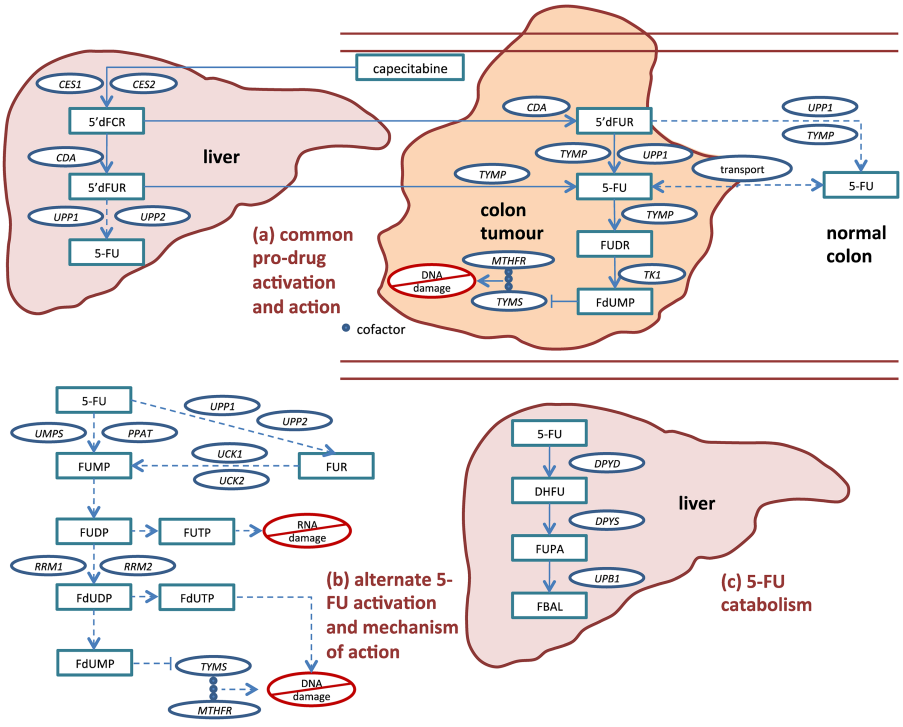


Figure 1 Capecitabine activation and subsequent 5-FU activation, action, transport and catabolism. Capecitabine is an oral 5-FU prodrug that is rationally designed so that concentrations of the cytotoxic metabolites FdUMP, FdUTP and FUTP are higher within malignant cells than within normal cells. After absorption in the gut, most of the drug activation occurs via the common pro-drug activation route (a), starting in the liver and finishing in the tumour. Additionally, 5-FU can be converted to its active compounds via alternate activation routes (b) in colon tumour cells and cells from multiple other tissues. Toxicity may occur if non-target tissue is exposed to activated capecitabine/5-FU (ie, FdUMP, FdUTP and FUTP), either when 5-FU is activated prior to arrival at the tumour or when it exits target tissue and is subsequently activated. 5-FU released in circulation may be quickly metabolised (c) in the liver by enzymes including DPYD. Gene names for candidate enzymes, provided in circles, are defined in online supplementary table S1. Drug catabolites are provided in rectangles. Primary (rationally designed) pathway are shown in solid lines, and alternative pathways shown in dashed lines.

Statistical analysis

For each of the 1456 SNP and exome array variants, our primary analysis was to test associations between global (any 5-FU-related) dose-limiting (grade 012v34) toxicity and genotype. Frequentist tests under a missing data linear or logistic regression model were implemented using SNPTESTv2. Samples were stratified by QUASAR2 treatment with age and gender as covariates. Meta-analysis of the two arms of QUASAR2 was performed using GWAMA, including tests of interarm heterogeneity. A stringent Bonferroni-corrected p value threshold of 3.43×10^{-5} (=0.05/1456) was used to indicate a significant association for the primary analysis of (binary) global dose-limiting toxicity.

For selected SNPs with association signals that reached or approached formal significance, we imputed additional SNPs within 1.5 Mb flanking regions (see online supplementary methods). We performed association tests for global and specific toxicities using the global grade 012v34 measure. Since the genotyped and imputed SNPs were non-independent, we declared associations with imputed SNPs significant using the same threshold of $p=3.43\times10^{-5}$. For any region within which one or more SNPs achieved significant associations with global toxicity, the underlying individual toxicities were investigated at the most strongly associated SNPs using quantitative measures and clinically actionable cut-offs for dose delay or reduction in QUASAR2 (generally, grade 012v34, except for grade 01v234 for diarrhoea).

To test for independent effects of variants within a region, we used logistic regression analysis in R, with age, sex and study arm as covariates. The best-fitting model was determined as that which minimised the Akaike information criterion (AIC) subject to a variant showing an association at p=0.05. Haplotype analyses at *DPYD* and *TYMS/ENOSF1* were performed using the *--hap-logistic* and *--independent-effect* commands in PLINK. Tests to examine multiple genetic variants were performed in PLINK. Receiver operator characteristic (ROC) analysis was performed in Stata using a binary classification of patients to either grade 0/1/2 or grade 3/4 global toxicity, using a genetic score given for each individual by $\Sigma\beta iNi$, where βi is the beta coefficient of the ith SNP significantly associated with global toxicity in a logistic regression model, and Ni is the number of harmful alleles carried by that individual at that locus.

Sequencing and validation of novel variants identified

Amplicon sequencing of the coding regions of *DPYD* and *TYMS* was performed by Roche/454 Titanium GS FLX technology in 100 patients with the highest levels of 5-FU-related toxicity ('HiTox'), specifically grade 3 or grade 4 diarrhoea in the first 4 cycles of treatment and or other grade 3/4 toxicities in the first 4 cycles of treatment. We also sequenced the same amplicons in 100 patients with no adverse toxicity events during the entire duration of treatment ('LoTox'). The missense *DPYD* variant p.Ala551Thr (A551 T) was identified in the HiTox pool. The DNA for each individual comprising the pool was Sanger-sequenced to identify those carrying this variant. Only one heterozygous individual was found. For analysis of the whole sample set, we designed KASPar (http://cshprotocols.cshlp.org/content/2007/9/pdb.prot4841.abstract) allele-specific single-nucleotide variant primers to detect A551 T and included three duplicates of the known variant sample in each run to facilitate genotype clustering (details available upon request). All samples

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Table 1 QUASAR2 patient characteristics

	Number	%
Site		
Colon	930	89
Rectum	116	11
Stage		
II	365	35
III	681	65
Ethnicity		
Caucasian	1046	100
Sex		
Male	593	57
Female	453	43
Age, years		
Median	65	
Minimum	22	
Maximum	85	
WHO performance status		
0–1	1046	100
Treatment		
Capecitabine (cap)	496	47
Cap+bevacizumab	550	53
Grade 3+adverse events		
Global	353	34
Diarrhoea	109	10
HFS	247	24
Mucositis	11	1
Stomatitis	12	1
Vomiting	15	1
Neutropaenia	22	2
Thrombocytopaenia	4	0

that did not cluster with the A allele homozygotes were subsequently examined by bidirectional Sanger sequencing.

RESULTS

A total of 1046 capecitabine-treated patients from the QUASAR2 trial were initially selected. Grade 3+ global toxicity was observed in 34% of these patients, with severe diarrhoea in 10% and HFS in 24% (see online supplementary table 2). Severe toxicity was less common for the other phenotypes, although five patients experienced grade 4 neutropaenia. After exclusions (see online supplementary methods), genotype data were available for 940 patients genotyped on genome-wide tagSNPs arrays and 968 patients genotyped on the Exome array. Across the two types of arrays, we identified 1456 genetic variants that mapped to within 25 kb of the 25 capecitabine/5-FU candidate genes (see online supplementary table 1). In the primary analysis, we searched for associations between each genetic variant and global (grade 012v34) capecitabine toxicity. Rather than regarding the two arms of QUASAR2 as discovery and validation sets, with consequent reduction in statistical power, we performed a meta-analysis of data from the 439 patients in arm A (capecitabine) and the 501 patients in arm B (capecitabine+bevacizumab), and assessed evidence for between-arm heterogeneity. The results of this analysis on a per SNP and SNP set basis are summarised in supplementary tables S3–S5, and the most important findings for all genotyped or imputed SNPs showing a significant association ($p<3.43\times10^{-5}$) are described in detail below.

New associations of common DPYD variants with capecitabine toxicity

We found that the A-allele (freq.=0.03) of SNP rs12132152 was associated with global capecitabine toxicity ($OR_{globalbinary}=3.83$, $p=4.31\times10^{-6}$; table 2). rs12132152 is an intergenic SNP 22 kb downstream of *DPYD* (chr1:97523004, b37). Upon imputation of variants in the region flanking this tag SNP (see online supplementary methods), we identified a few SNPs with marginally more significant associations, notably rs76387818 (chr1:97 539 400; $OR_{binary}=4.05$, $p=2.11\times10^{-6}$, $r^2_{tag}=0.98$; table 2; figure 2A). Results were similar when the quantitative measure of global toxicity was used (table 2). We investigated the individual phenotypes comprising the global toxicity measure. rs12132152 and rs76387818 were principally associated with HFS, under quantitative and binary models (for rs76387818, $OR_{hfsquant}=1.78$, $p=5.51\times10^{-8}$, $OR_{hfsbinary}=6.44$, $p=1.75\times10^{-8}$), albeit with some weaker evidence of associations with diarrhoea (table 2).

We then investigated in silico possible functional mechanisms underlying the rs12132152/rs76387818 association. We examined ENCODE data (http://genome.ucsc.edu/ENCODE/) for the region containing rs12132152 and seven strongly correlated SNPs (figure 2A: approximately chr1:97 475 000-97 562 000, b37). FAIRE and histone K4 methylation data suggested that this is a region of open chromatin and one correlated SNP in particular, rs12123160, lies at a methylated CpG. Although no suitable data from normal liver were available, we found that genotypes at rs12132152 and correlated SNPs were not associated with *DPYD* expression in adipose tissue, lymphoblastoid cells or skin (Genevar database, $p>0.13$)[18] or in colon tissue (TCGA, $p=0.72$).

The second *DPYD* toxicity-associated variant (table 2) was identified following SNP imputation in the region of 1.5 Mb surrounding rs7548189, a tagSNP intronic to *DPYD* (chr1:97 867 713, b37). rs7548189 was borderline associated with the binary meaure global toxicity ($OR_{globalbinary}=1.67$, $p=3.79\times10^{-5}$). Further investigation showed rs7548189 formally to be associated with the quantitative measure of global toxicity, and with diarrhoea under binary and quantitative models ($OR_{globalquant}=1.23$, $p=6.82\times10^{-6}$; $OR_{diarrhoeaquant}=1.18$, $p=1.54\times10^{-5}$; $OR_{diarrhoeabinary}=1.76$, $p=1.72\times10^{-5}$). As can be seen from table 2, HFS also contributed to the association with global toxicity. Following regional imputation, rs12022243 (r^2 with rs7548189=0.95) was found formally to be associated with global toxicity under a binary model ($OR_{globalbinary}=1.69$, $p=2.55\times10^{-5}$). We found that rs12022243 showed excellent imputation quality: of 190 independently assessed individuals, only 4 (2%) genotypes were missing, and all the remaining genotypes were imputed correctly (see online supplementary methods).

We found little evidence from ENCODE data that rs7548189 is functional, but rs12022243 falls in a region of open chromatin that may have enhancer activity. rs7548189 was not associated with *DPYD* expression levels in lymphoblasts, fibroblasts, T-cells, adipose tissue, or skin on the Genevar database ($p>0.07$)[18–20] or in colon tissue from TCGA ($p=0.97$). rs12022243 Was absent from these datasets.

Using logistic regression analysis, we found that the rs12132152 and rs7548189/rs120222243 signals were independent of each other and of the known *DPYD* toxicity variants *2A (rs3918290) and 2846T>A (rs67373796) (details not shown; linkage disequilibrium data in online supplementary figure 1). Further analysis of haplotypes based on 81 tagSNPs in

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Table 2 Selected associations between genetic variants and capecitabine toxicity in QUASAR2

Gene	SNP b37 coordinate	Toxicity-associated allele/other allele	TAF	n Genotyped n imputed	Info score [A]=Hap370 [B]=Hap610 [C]=Omni2.5	Global binary: 012v34 OR (95% CI) p-value	Global quant:01v2v34 OR (95% CI) p-value	HFS binary:012v34 OR (95% CI) p-value	HFS quant:01v2v34 OR (95% CI) p-value	Diarrhoea binary:012v34 OR (95% CI) p-value	Diarrhoea quant:01v2v34 OR (95%CI) p-value	Other clinically actionable associations
DPYD	rs12132152 chr1:97523004	A/G	0.031	456 484	0.993[A]	**3.83 (3.26–4.40)** **4.31×10^{-6}**	**1.61 (1.41–1.82)** **5.89×10^{-6}**	**6.12 (5.48–6.76)** **3.29×10^{-8}**	**1.74 (1.53–1.95)** **1.47×10^{-7}**	0.44 (0–1.32) 0.065	0.85 (0.68–1.02) 0.068	
DPYD	rs76387818 chr1:97539400	A/G	0.031	0 940	0.993[A] 0.999[B] 0.999[C]	**4.05 (3.47–4.62)** **2.11×10^{-6}**	**1.66 (1.45–1.87)** **1.93×10^{-6}**	**6.44 (5.79–7.09)** **1.75×10^{-8}**	**1.78 (1.57–1.99)** **5.51×10^{-8}**	0.44 (0–1.33) 0.071	0.86 (0.68–1.03) 0.083	
DPYD	rs7548189 chr1:97 867 713	A/C	0.196	940 0	N/A	1.67 (1.43–1.91) 3.79×10^{-5}	**1.23 (1.14–1.31)** **6.82×10^{-6}**	1.42 (1.15–1.69) 0.011	1.16 (1.07–1.25) 0.0011	1.21 (0.84–1.58) 0.0015	**1.18 (1.10–1.25)** **1.54×10^{-5}**	**Diarrhoea 01v234** **1.76 (1.50–2.02)** **1.72×10^{-5}**
DPYD	rs12 022 243 chr1: 97 862 780	T/C	0.196	0 940	0.996[A] 0.992[B] 0.998[C]	**1.69 (1.45–1.94)** **$2.55\ x10^{-5}$**	**1.23 (1.14–1.32)** **$4.45\ x10^{-6}$**	1.43 (1.16–1.7) 0.0096	1.16 (1.07–1.25) $8.26\ x10^{-4}$	**1.79 (1.54–2.05)** **$9.86\ x10^{-6}$**	**1.18 (1.11–1.26)** **$1.11\ x10^{-5}$**	
TYMS/ ENOSF1	rs2612091 chr18:683 607	C/T	0.532	940 0	N/A	**1.59 (1.39–1.79)** **5.28×10^{-6}**	**1.19 (0.77–0.91)** **2.35×10^{-6}**	**1.57 (0.45–0.83)** **2.94×10^{-6}**	**1.21 (0.76–0.90)** **3.67×10^{-7}**	1.18 (0.55–1.15) 0.29	1.04 (0.90–1.03) 0.27	**HFS 01v234** **1.57 (0.45–0.83)** **2.94×10^{-6}**
TYMS/ ENOSF1	rs2741171 chr18:700 687	T/C	0.534	0 940	0.960[A] 0.975[B] 0.990[C]	**1.60 (1.39–1.80)** **6.64×10^{-6}**	**1.20 (1.13–1.28)** **9.24×10^{-7}**	**1.74 (1.51–1.97)** **1.64×10^{-6}**	**1.23 (1.16–1.31)** **3.10×10^{-8}**	1.01 (0.70–1.32) 0.92	1.03 (0.97–1.09) 0.37	**HFS 01v234** **1.61 (1.42–1.80)** **1.44×10^{-6}**

The table shows the results of the meta-analysis of global and selected individual toxicities, measured as binary or continuous ('quant') variables, in the two arms of QUASAR2. The frequency of the toxicity-associated allele (TAF) is also shown and ORs are expressed relative to this. Imputation quality Info Score is also shown, as are numbers of samples imputed and directly genotyped. The final column shows results for toxicity phenotype classifications that could lead to treatment change or delay according to the QUASAR2 protocol. There was no evidence of heterogeneity between QUASAR2 arms in any of these analyses $P_{het}>0.2$, $I^2<0.25$).

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

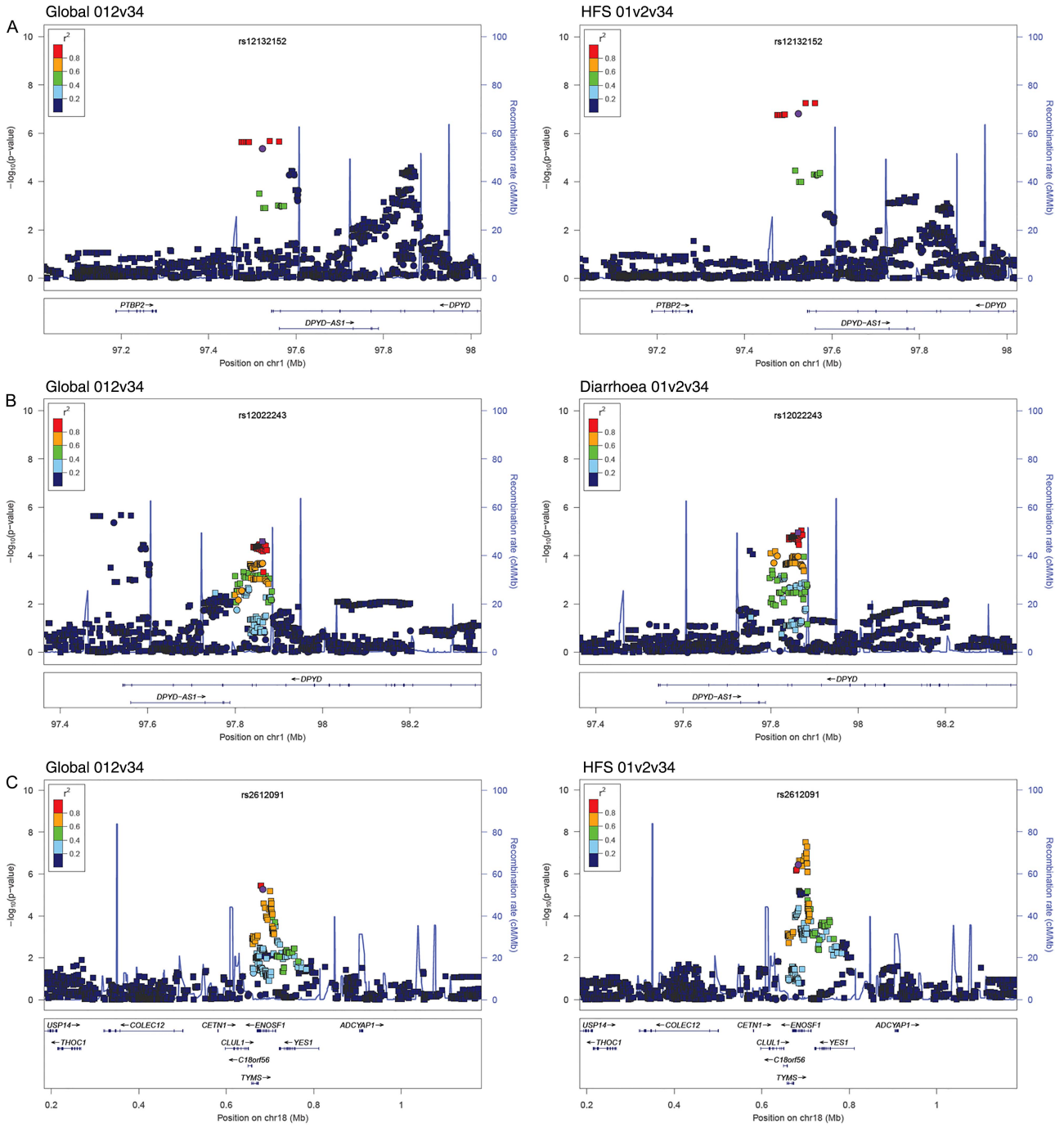


Figure 2 Regional plots of *DPYD* and *TYMS* SNPs for associations with capecitabine-related toxicity. Regional associations between polymorphisms and (left) global grade 012v34 capecitabine toxicity or (right) selected specific toxicities were plotted using LocusZoom and are shown for SNPs 500 kb either side of (a) rs12132152 which maps 3′ of *DPYD*, (b) rs12022243, mapping to an intron of *DPYD* and (c) rs2612091, which maps to an intron of *ENOSF1*. The x-axis shows chromosome position, while the y-axis shows the $\log_{10}$P for the association. Circles represent genotyped SNPs included in the 1456-SNP test panel, with the most significantly associated SNP in purple. Squares represent imputed SNPs. Linkage disequilibrium (calculated using hg19/1000 genomes, March 2012 EUR population) with the most significantly associated SNP is shown by colour as indicated.

a 25 kb window either side of *DPYD* provided no further refinement of the rs12132152 and rs7548189/rs120222243 associations and showed no evidence of additional, independent association signals (details not shown).

Resolving the toxicity associations at the TYMS and ENSOF1 loci

We found that the G-allele (freq.=0.45) of SNP rs2612091 was associated with increased global toxicity ($OR_{globalbinary}$=1.59, $p=5.28\times10^{-6}$; $OR_{globalquant}$=1.19, $p=2.35\times10^{-6}$; table 2). rs2612091 lies 10 kb downstream of *TYMS* within an intron of enolase superfamily member 1 (*ENOSF1*, chr18:683 607). Fine mapping showed somewhat more significant associations for a SNP, rs2741171 (chr18:700 687), in linkage disequilibrium (r^2=0.73) with rs2612091, particularly for the quantitative measure of global toxicity and HFS ($OR_{global\ quant}$=1.2, $p=9.24\times10^{-7}$). The rs2741171 variant is further downstream of *TYMS* (27 kb) and again is intronic for *ENOSF1*, but both

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

SNPs fall between recombination hotspots that flank the entirety of *TYMS* and *ENOSF1* (figure 2B). The rs2612091/rs2741171 effect on toxicity was essentially driven by HFS, with an especially strong association being observed using the quantitative measure (rs2612091: $OR_{hfsquant}=1.21$, $p=3.67\times10^{-7}$; rs2741171: $OR_{hfsquant}=1.23$, $p=3.10\times10^{-8}$).

ENOSF1 is a largely uncharacterised gene that appears to encode a protein and RNAs antisense to *TYMS*. It has been proposed that *ENOSF1* regulates *TYMS* mRNA and/or protein expression.[21] We therefore analysed associations between rs2612091 genotype and *TYMS* and *ENSOF1* expression using the Genevar and TCGA databases. Using Caucasian-matched twin data,[18] the rs2612091 G-allele significantly decreased *ENOSF1* expression in adipose tissue for both twin sets ($p_{set1}=7.0\times10^{-4}$, $p_{tset2}=9.7\times10^{-6}$) and in lymphoblastoid cells for one set ($p_{set1}=0.89$, $p_{set2}=0.0012$). However, rs2612091 genotype was not associated with *TYMS* expression ($p>0.30$ for each of the same analyses). Similarly, in lymphoblastoid cell line expression data,[20] the rs2612091 G-allele was associated with decreased expression of *ENOSF1* ($p=1.9\times10^{-6}$) but not *TYMS* ($p=0.82$). These results were further replicated in the TCGA colon data, in which the rs2612091 G-allele was again associated with decreased *ENOSF1* expression ($OR=0.76$, $p=1.5\times10^{-7}$), but not with *TYMS* ($OR=0.95$, $p=0.45$).

We tested the relationship of rs2612091 to two *TYMS* polymorphisms (5′ VNTR 2R/3R and 3′UTR 6 bp ins-del[22 23] that have previously been reported to alter *TYMS* expression, and hence, to affect 5-FU-related toxicity (see online supplementary table S4). Noting moderate LD between rs2612091 and the 5′VNTR and 3′UTR polymorphisms ($r^2=0.40$ and 0.32; online supplementary figure 1), we tested whether the rs2612091 signal was independent of the 5′VNTR and 3′UTR polymorphisms, using the quantitative measure of HFS (01v2v34) as the toxicity phenotype, because HFS underlies the global toxicity signal observed for all three variants.[7] First, the three variants were incorporated into a covariate-adjusted logistic regression model. Only rs2612091 remained significantly associated when adjusting for the other variants (for rs2612091 $OR_{hfsquant}=1.21$, $p=0.00049$, 5′ VNTR $p=0.19$, 3′UTR $p=0.48$). Using rs2612091 alone in the model minimised the AIC. Second, we tested for evidence of interaction (epistasis) among the variants, but found none between rs2612091 and either the 5′VNTR ($p=0.92$) or the 3′UTR ($p=0.19$). These analyses suggested that rs2612091 and the previously identified variants do not act as a 3-polymorphism tag for unidentified variants, and that rs2612091 alone captures the association signal created by all three variants. Finally, we tested the independence of each individual polymorphism in a 3-polymorphism haplotype (see online supplementary table S4). We found that the G-allele of rs2612091 consistently increased the risk of toxicity, irrespective of the 5′VNTR or 3′UTR genotype ($p=0.0021$). Conversely, neither the 5′VNTR ($p=0.17$) nor the 3′UTR ($p=0.61$) risk-allele consistently increased risk of toxicity when varying the genotype of the other two SNPs. The analysis was repeated as 2-polymorphism haplotypes comprising rs2612091 and either the 5′VNTR or 3′UTR variant. Irrespective of either genotype at either the 5′VNTR or 3′UTR polymorphism, rs2612091 genotype was again significantly associated with HFS, ($p=0.00053$ and $p=1.47\times10^{-6}$ when incorporating 5′VNTRand 3′UTR, respectively). The above analyses were repeated using global toxicity, and all the above results were similar but reduced modestly in significance (data not shown).

Equivalent logistic regression and haplotype analysis using the top fine-mapping SNP rs2741171 showed even stronger evidence that the new SNP signal alone explained the associations at *TYMS/ENOSF1* (data not shown). We further tested the various combinations of rs2741171, rs2612091, 5′VNTR and 3′UTR polymorphisms in a multivariate logistic regression model and found that the model which minimised AIC incorporated rs2741171 alone (details not shown). rs2741171 lies next to a region of open chromatin that may be a p300 binding site.

Identifying and assessing rare susceptibility variants in DPYD and TYMS

We sequenced the coding regions of *DPYD* and *TYMS* in pools of Hi-Tox and Lo-Tox patients (see online supplementary figures S2 and S3). In the HiTox pool, we identified a single missense variant that was not present on SNP and exome arrays (*DPYD* c. G1651A; p.Ala551Thr; chr1:97 981 371). We found no other occurrence of this variant in the full set of 968 patients. A551 T was predicted to be strongly damaging by SIFT, Polyphen, PhyloP and MutationTaster, and the single patient with this allele had experienced grade 4 neutropaenia and thrombocytopaenia. Database searches determined that this variant has been previously reported as causal for DPYD Deficiency Syndrome (OMIM 612779).[24] We confirmed that A551 T was not in linkage disequilibrium with any of the other common or rare *DPYD* toxicity variants.

We then considered only the 19 patients with extreme (grade 4) toxicity at any cycle and determined which alleles they carried at the three toxicity SNPs and their complement of rare *DPYD* alleles from the literature, including 2846 A>T and *2A that were shown to be associated with 5-FU toxicity in our previous meta-analysis[7] (table 3). There was no good evidence that the risk alleles at the three new toxicity SNPs were overrepresented as a group in these 19 patients (table 3). Using the evaluation of Caudle *et al*[8] as a guide, supplemented by data from this study, we then assessed the likely contributions of each rare *DPYD* variant to extreme toxicity. There was insufficient prior evidence[8] to regard six *DPYD* alleles (*4, *5, *6, *9A, M166 V and K259E) as pathogenic (see online supplementary table S6). Inspection of the genotypes of these polymorphisms in the severe toxicity cases (table 3) and the tests of association with binary global toxicity (see online supplementary tables S3 and S6) did not contradict this view. Several other rare *DPYD* alleles (*3, *7, *8, *9B, *10, *11, *12) were not present in our sample set. We denoted four rare *DPYD* alleles as severely functionally deleterious: *2A, 2846T>A, *13 and A551 T. Five of 19 (26%) severe-toxicity patients carried one of these *DPYD* alleles (table 3). Of these five cases, four (80%) had life-threatening bone marrow toxicity (G4 neutropaenia and thrombocytopaenia), whereas the other had G4 diarrhoea. Another individual with G4 neutropaenia, but not thrombocytopaenia, did not carry any of the four *DPYD* alleles. Overall, for prediction of severe myelosuppression, the rare *DPYD* variants had 83% sensitivity, 99% specificity, 29% positive predictive value and 99.9% negative predictive value.

DISCUSSION

Through the analysis of capecitabine-treated patients from the QUASAR2 trial for SNPs and rare genetic variants in 25 capecitabine/5-FU pathway genes and subsequent exon sequencing of *DPYD* and *TYMS*, we have identified new genetic predictors of capecitabine-induced toxicity and clarified the origins of previously reported signals. Two of the new toxicity variants (rs12022243 and rs12132152) map to *DPYD* and are independent of previously reported *DPYD* toxicity alleles. The toxicity-associated allele at rs12022243 is common (freq.=0.22)

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Table 3 Genotypes of QUASAR2 individuals with grade 4 toxicity at selected DPYD variants

Case	D	V	H	N	P	M	S	Possible explanation	2846	*2A	A551T	*13	*4	*5	*6	*9A	M166V	K259E	rs12132152	rs12022243	rs2612091
1	**4**	0	2	0	0	0	0		0	0	0	0	0		0	0	0	0	0	0	2
2	**4**	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	1	2
3	0	**4**	0	0	0	0	0		0	0	0	0	1	1	0	0	0	0	0	1	1
4	**4**	2	2	0	0	0	0		0	0	0	0	0	0	0	2	1	0	0	0	0
5	**4**	0	0	0	0	0	1	2846T>A	1	0	0	0	1	1	0	0	0	0	0	0	0
6	2		1	**4**	**4**	0	2	2846T>A	1	0	0	0	1	0	0	0	0	0	0	2	1
7	3	0	2	**4**	**4**	3	3	A551T	0	0	1	0	0	0	0	0	0	0	0	1	2
8	3	1	3	**4**	2	3	3		0	0	0	0	0	2	0	0	0	0	0	1	2
9	**4**	0	2	0	0	1	0		0	0	0	0	1	0	0	0	0	0	0	1	2
10	**4**	1	3	0	0	1	0		0	0	0	0	0	0	0	0	0	0	0	1	2
11	**4**	2	1	1	0	0	0		0	0	0		0	1	0	0	0		0	1	2
12	0	0	3	**4**	**4**	3	3	*2A	0	1	0	0	0	0	0	1	1	0	0	0	1
13	**4**	0	1	0	1	1	1		0	0	0	0	0	1	0	0	0	0	0	1	1
14	0	**4**	3	0	1	0	0		0	0	0	0	0	0	0	1	1	0	0	1	2
15	**4**	0	1	0	0	1	1		0	0	0	0	0	1	0	0	0	0	0	0	1
16	**4**	0	2	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	1	1
17	1	0	**4**	0	0	0	0		0		0	0	0	0	0	0	0	0	0	1	0
18	2	**4**	3	0	0	0	1		0	0	0	0	0	1	0	0	0	0	0	0	0
19	3	0	2	**4**	**4**	0	**4**	*13	0	0	0	1	0		0			0			

The variants shown are (1) those identified by this study (rs12132152, rs12022243, rs2612091, *DPYD* A551 T), (2) *DPYD* alleles (2846 A>T and *2A) shown to be associated with 5-FU toxicity in the meta-analysis of Rosmarin *et al*[7] and (3) potential *DPYD* toxicity alleles from Caudle *et al*.[8] Genotypes shown are major allele homozygote (0), heterozygote (1) and minor or variant allele homozygote (2). Blank cells denote missing data. The allele that provides a plausible explanation for the severe toxicity is shown. Note that *4 and *5 *DPYD* alleles are in complete linkage disequilibrum (D'=1.0) with 2A or 2846T>A. Overall association statistics and genotype frequencies in cases and controls are shown for the analysis of binary global toxicity in online supplementary table S3.
For the toxicities, D=diarrhoea; H=HFS; M=mucositis; N=neutropaenia; P=thrombocytpaenia; S=stomatitis; V=vomiting.

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

and has a moderate effect ($OR_{globalbinary}$=1.8) on toxicity risk, influencing HFS and diarrhoea. The rs12132152 toxicity allele is less common, but not rare (freq.=0.03), and its effect is greater ($OR_{globalbinary}$=3.8). Although associations between other common *DPYD* polymorphisms and 5-FU toxicity have previously been proposed,[8 23] we failed to confirm these in this analysis and in a previous meta-analysis of multiple datasets.[7]

Through sequencing patients with high capecitabine toxicity, we identified a third new *DPYD* toxicity allele, p.Ala551Thr.[24] Like the other rare *DPYD* variants (*2A and 2846T>A) that have established associations with 5-FU-related toxicity in the heterozygous state, A551 T has been shown to cause the recessive DPYD deficiency syndrome when present as the homozygote or compound heterozygote with another mutant allele. The identification of A551 T in a patient experiencing grade 4 neutropaenia and thrombocytopaenia adds further support to the view that all rare *DPYD* variants that cause DPYD Deficiency Syndrome greatly increase the risk of 5-FU toxicity in heterozygotes. Furthermore, of the three other QUASAR2 patients who experienced grade 4 myelotoxicity—two of whom were the only toxicity-induced deaths in QUASAR2—one carried *DPYD* *2A, one 2846A>C and one *13. Of the 12 carriers of rare, functionally deleterious *DPYD* alleles who did not develop severe toxicity, seven suffered a grade 3 toxicity and may therefore have been spared severe toxicity by capecitabine dose reduction. However, the reason for the remaining five individuals avoiding clinically important toxicity is unclear. Much interpatient difference in 5-FU toxicity remains unexplained, and it remains very plausible that some of this variability is heritable.

Our finding of another capecitabine toxicity SNP, rs2612091, produced some very unexpected results. This common variant is downstream of *TYMS* and intronic to *ENOSF1*. Our detailed analysis showed that the rs2612091 signal appears to explain both the previously reported associations between 5-FU/capecitabine toxicity and *TYMS* polymorphisms (5′VNTR 2R/3R and 3′UTR 6 bp ins-del). This is especially surprising given published evidence that the two *TYMS* polymorphisms directly affect *TYMS* mRNA expression and protein levels. However, recent critical assessments have shown that the data linking *TYMS* expression to 5′VNTR 2R/3R and 3′UTR 6 bp ins-del genotypes are actually very mixed.[25] In fact, our analysis of public mRNA expression data demonstrated rs2612091 to be associated with *ENOSF1* expression and not with *TYMS* expression. Our data imply that the *TYMS* 5′VNTR and 3′UTR toxicity association signals result from LD between these polymorphisms and tagSNP rs2612091 or, more likely, fine-map SNP rs2741171 (see online supplementary figure 1A). We therefore conclude that *ENOSF1* is most likely to be the target of the functional variation tagged by rs2612091. *ENOSF1* and *TYMS* transcripts are overlapping, but toxicity SNP does not appear to act through antisense-mediated down-regulation of *TYMS* mRNA. However, ENOSF1 protein has been proposed as an influence on TYMS activity, and this remains a plausible mechanism of toxicity.

We found no evidence of heterogeneity between the two arms of QUASAR2 in terms of the effects of genetic variants on global toxicity or any specific toxicity. The toxicity profiles of capecitabine and bevacizumab principally overlap through the risk of HFS, although the effects of the former greatly outweigh those of the latter (estimated 18% vs 7% for QUASAR2). In principle, a polymorphism could predispose to HFS caused by either capecitabine and/or bevacizumab. However, such a variant is most likely to act at the level of the HFS target tissue, whereas, our variants were specifically chosen for potential effects on 5-FU metabolism. The most likely effect of the use of bevacizumab on our study was, therefore, a small decrease in statistical power owing to a higher 'background' level of HFS resulting from a non-capecitabine source in arm B.

In summary, we have identified four new variants associated with capecitabine toxicity. Further work is desirable in order to confirm and quantitate these associations in additional datasets, and to understand the mechanistic origins of capecitabine-related toxicity, especially for the *TYMS* and *ENOSF1* loci. While not yet an ideal test for clinical use (see online supplementary figure 4), genetic testing can be used to highlight patients at increased risk of capecitabine toxicity. A two-tier test may be justifiable, comprising (1) a sensitive test for severe, life-threatening toxicity based principally on rare *DPYD* variants and (2) a test additionally incorporating SNPs to highlight the risk of clinically actionable toxicity. It is a moot point as to whether such a strategy would currently be cost effective, and it is not yet known how well our panel of variants predicts toxicity from 5-FU delivered by other means or in combination regimens. Although our dataset was relatively large, and power was good to detect toxicity variants with relatively large effects (typically >80% power for a common variant conferring an OR>1.5), power was very limited to detect variants with lower allele frequency and/or smaller effect size (see online supplementary table S3). Further efforts to identify additional polymorphisms and rare variants associated with 5-FU toxicity remain valid.

URLS

Primer3: http://sourceforge.net/projects/primer3/
LocusZoom: http://csg.sph.umich.edu/locuszoom/
SNPTEST2: https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html
IMPUTE2: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
GWAMA: http://www.well.ox.ac.uk/gwama/download.shtml
TCGA: http://tcga-data.nci.nih.gov/

Author affiliations
[1]Molecular and Population Genetics Laboratory, Oxford, UK
[2]Department of Oncology, University of Oxford, Old Road Campus Research Building, Oxford, UK
[3]Oxford NIHR Comprehensive Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, Oxford, UK
[4]Human Genotyping Unit-CeGen, Human Cancer Genetics Programme, Spanish National Cancer Research Centre, Melchor Fernández Almagro 3, Madrid, Spain
[5]Department of Medical Oncology, Instituto de Investigacion Sanitaria Hospital General Universitario Gregorio Marañón, Universidad Complutense, Madrid, Spain
[6]Centre for Statistics in Medicine, University of Oxford, Botnar Research Centre, Oxford, UK
[7]OCTO, University of Oxford, Old Road Campus Research Building, Oxford, UK
[8]Galician Public Foundation of Genomic Medicine (FPGMX), CIBERER, Genomics Medicine Group, Hospital Clinico, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[9]Genetic Susceptibility to Colorectal Cancer Group, Gastrointestinal & Pancreatic Oncology Team, IDIBAPS/CIBERehd/Hospital Clínic, Centre Esther Koplowitz (CEK), Barcelona, Spain
[10]Institute of Digestive and Metabolic Diseases, Hospital Clínic, Barcelona, Spain
[11]Nuffield Department of Clinical Laboratory Sciences, University of Oxford, Oxford, UK

Collaborators QUASAR2 collaborators, The EPICOLON Consortium—Listed in online supplementary data.

Contributors IT, CP, AGN, GP, MM, CFR, CRP, AC, SCB, AC and DNR planned and designed the study. DNR, CP and IT wrote the manuscript. CP, AJ, AP, KK, KH, ED, EJ, HW, and DNR performed sequencing, genotyping and/or prepared DNA from QUASAR2. Oversight of sequencing was provided by JT. SL, CS and PJ provided QUASAR2 clinical data; RM and DJK ran the QUASAR2 study. DNR, CP and IT analysed the QUASAR2 data and, with LFM, the public functional data.

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Funding QUASAR2 was funded by Hoffman La Roche. Molecular and statistical analyses were funded by the Oxford NIHR Comprehensive Biomedical Research Centre and Cancer Research UK. Core funding to the Wellcome Trust Centre for Human Genetics was provided by the Wellcome Trust (090532/Z/09/Z).

Competing interests None.

Ethics approval Oxfordshire Research Ethics Committee B (Approval No. 05 \Q1605\66).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data are available to bona fide scientific collaborators subject to reporting the primary results of the study and ethical permissions.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 3.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/3.0/

REFERENCES

1 Noordhuis P, Holwerda U, Van Laar JA, *et al*. A non-radioactive sensitive assay to measure 5-fluorouracil incorporation into DNA of solid tumors. *Nucleosides Nucleotides Nucleic Acids* 2004;23:1481–4.
2 Saltz LB, Niedzwiecki D, Hollis D, *et al*. Irinotecan fluorouracil plus leucovorin is not superior to fluorouracil plus leucovorin alone as adjuvant treatment for stage III colon cancer: results of CALGB 89803. *J Clin Oncol* 2007;25:3456–61.
3 Cassidy J, Twelves C, Van Cutsem E, *et al*. First-line oral capecitabine therapy in metastatic colorectal cancer: a favorable safety profile compared with intravenous 5-fluorouracil/leucovorin. *Ann Oncol* 2002;13:566–75.
4 Haller DG, Rothenberg ML, Wong AO, *et al*. Oxaliplatin plus irinotecan compared with irinotecan alone as second-line treatment after single-agent fluoropyrimidine therapy for metastatic colorectal carcinoma. *J Clin Oncol* 2008;26:4544–50.
5 Stein BN, Petrelli NJ, Douglass HO, *et al*. Age and sex are independent predictors of 5-fluorouracil toxicity. Analysis of a large scale phase III trial. *Cancer* 1995;75:11–7.
6 Thorn CF, Marsh S, Carrillo MW, *et al*. PharmGKB summary: fluoropyrimidine pathways. *Pharmacogenet Genomics* 2011;21:237–42.
7 Rosmarin D, Rosmarin D, Palles C, *et al*. Genetic markers of toxicity from capecitabine and other 5-fluorouracil-based regimens: investigation in the QUASAR2 study, systematic review and meta-analysisGenetic markers of toxicity from capecitabine and other 5-fluorouracil-based regimens: investigation in the QUASAR2 study, systematic review and meta-analysis. *J Clin Oncol* Published Online First: 3 Mar 2014. doi:10.1200/JCO.2013.51.1857
8 Caudle KE, Thorn CF, Klein TE, *et al*. Clinical pharmacogenetics implementation consortium guidelines for dihydropyrimidine dehydrogenase genotype and fluoropyrimidine dosing. *Clin Pharmacol Ther* 2013;94:640–5.
9 Dunlop MG, Dobbins SE, Farrington SM, *et al*. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 2012;44:770–6.
10 Huyghe JR, Jackson AU, Fogarty MP, *et al*. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 2013;45:197–201.
11 Goldstein JI, Crenshaw A, Carey J, *et al*. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics (Oxford, England)* 2012;28:2543–5.
12 Howie B, Fuchsberger C, Stephens M, *et al*. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;44:955–9.
13 Delaneau O, Howie B, Cox AJ, *et al*. Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013;93:687–96.
14 Marchini J, Howie B, Myers S, *et al*. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906–13.
15 Dotor E, Cuatrecases M, Martinez-Iniesta M, *et al*. Tumor thymidylate synthase 1494del6 genotype as a prognostic factor in colorectal cancer patients receiving fluorouracil-based adjuvant treatment. *J Clin Oncol* 2006;24:1603–11.
16 Horie N, Aiba H, Oguro K, *et al*. Functional analysis and DNA polymorphism of the tandemly repeated sequences in the 5′-terminal regulatory region of the human gene for thymidylate synthase. *Cell Struct Funct* 1995;20:191–7.
17 Cuppen E. Genotyping by Allele-Specific Amplification (KASPar). CSH protocols 2007;2007:pdb prot4841. doi:10.1101/pdb.prot4841 [published Online First: Epub Date]
18 Nica AC, Parts L, Glass D, *et al*. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 2011;7:e1002003.
19 Dimas AS, Deutsch S, Stranger BE, *et al*. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, NY)* 2009;325:1246–50.
20 Stranger BE, Montgomery SB, Dimas AS, *et al*. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* 2012;8:e1002639.
21 Dolnick BJ, Angelino NJ, Dolnick R, *et al*. A novel function for the rTS gene. *Cancer Biol Ther* 2003;2:364–9.
22 Lecomte T, Ferraz JM, Zinzindohoue F, *et al*. Thymidylate synthase gene polymorphism predicts toxicity in colorectal cancer patients receiving 5-fluorouracil-based chemotherapy. *Clin Cancer Res* 2004;10:5880–8.
23 Schwab M, Zanger UM, Marx C, *et al*. Role of genetic and nongenetic factors for fluorouracil treatment-related severe toxicity: a prospective clinical trial by the German 5-FU Toxicity Study Group. *J Clin Oncol* 2008;26:2131–8.
24 Van Kuilenburg AB, Meinsma R, Beke E, *et al*. Identification of three novel mutations in the dihydropyrimidine dehydrogenase gene associated with altered pre-mRNA splicing or protein function. *Biol Chem* 2005;386:319–24.
25 Ghosh S, Winter JM, Patel K, *et al*. Reexamining a proposal: thymidylate synthase 5′-untranslated region as a regulator of translation efficiency. *Cancer Biol Ther* 2011;12:750–5.

Gut: first published as 10.1136/gutjnl-2013-306571 on 19 March 2014. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Supplementary Table 1. Candidate gene region summary

Gene Symbol	Gene Name	Location – Build 37 (coordinates do not include 25kb flanking region)	Number of Test Panel SNPs
ABCB1	ATP-binding cassette, sub-family B	chr7:87132948-87342564	77
ABCC3	ATP-binding cassette, sub-family C, member 3	chr17:48712218-48769063	64
ABCC4	ATP-binding cassette, sub-family C, member 4	chr13:95672083-95953687	224
ABCC5	ATP-binding cassette, sub-family C, member 5	chr3:183637724-183735727	101
ABCG2	ATP-binding cassette, sub-family G, member 2	chr4:89011416-89152474	57
CDA	cytidine deaminase	chr1:20915444-20945400	26
CES1	carboxylesterase 1 isoform a precursor	chr16:55836764-55867075	24
CES2	carboxylesterase 2 isoform a precursor	chr16:66968347-66978994	59
DPYD	dihydropyrimidine dehydrogenase	chr1:97543300-98386615	239
DPYS	dihydropyrimidinase	chr8:105391652-105479277	69
MTHFR	methylenetetrahydrofolate reductase	chr1:11845787-11866115	38
PPAT	phosphoribosyl pyrophosphate amidotransferase	chr4:57259529-57301845	29
RRM1	ribonucleoside-diphosphate reductase subunit 1	chr11:4115924-4160106	29
RRM2	ribonucleoside-diphosphate reductase subunit 2	chr2:10262735-10270623	19
SLC22A7	solute carrier family 22 member 7 isoform b	chr6:43265998-43273276	26
SLC29A1	equilibrative nucleoside transporter 1	chr6:44187242-44201888	26
TK1	thymidine kinase 1	chr17:76170160-76183285	35
TYMP	thymidine phosphorylase	chr22:50964182-50968258	92
TYMS	thymidylate synthetase	chr18:657604-673499	34
UCK1	uridine-cytidine kinase 1 isoform a	chr9:134399191-134406655	43
UCK2	uridine-cytidine kinase 2 isoform a	chr1:165796890-165877339	22
UMPS	uridine monophosphate synthase	chr3:124449213-124464040	34
UPB1	beta-ureidopropionase	chr22:24890077-24922553	30
UPP1	uridine phosphorylase 1	chr7:48128355-48148330	16
UPP2	uridine phosphorylase 2	chr2:158851691-158992478	43

Supplementary Table 2. Toxicity frequencies in QUASAR2

Adverse Event	CTCAE grade	Patients
Global	0	75
	1	241
	2	375
	3	334
	4	19
	Unreported	2
Diarrhoea	0	370
	1	388
	2	175
	3	99
	4	10
	Unreported	4
Handfoot	0	176
	1	287
	2	331
	3	246
	4	1
	Unreported	5
Mucositis	0	734
	1	248
	2	49
	3	11
	4	0
	Unreported	4
Stomatitis	0	718
	1	244
	2	67
	3	11
	4	1
	Unreported	5
Vomiting	0	817
	1	134
	2	74
	3	12
	4	3
	Unreported	6
Neutropaenia	0	921
	1	71
	2	28
	3	17
	4	5
	Unreported	4
Thrombocytopaenia	0	961
	1	67
	2	9
	3	0
	4	4
	Unreported	5

Supplementary Table 3. Summary of results of association test between 1,456 5-FU pathway genetic variants and binary global toxicity. Note that the unproven possibility of additional, unreported independent risk SNPs at *DPYD* remains (e.g. rs10875047 showed an association at a significance level close to the $P<3.43 \times 10^{-5}$ threshold used).There was very little evidence of between-arm heterogeneity in SNP effects as shown by the I^2 statistic which is <75% in all cases.

rs number	chromosome	position	Risk allele	Alternative allele	OR (95% CI)	P value	I^2 (heterogeneity)	r^2 with one of the lead SNPs (lead SNP #)
310/370CNV/610/Omni2.5 arrays								
rs12132152	1	97523004	A	G	3.83 (2.16-6.79)	$4.31X\ 10^{-6}$	0.16	Lead SNP 1
rs2612091	18	683607	C	T	1.59 (1.30-1.92)	$5.28X\ 10^{-6}$	0	Lead SNP 2
rs7548189	1	97867713	A	C	1.67 (1.31-2.13)	$3.79X\ 10^{-5}$	0	Lead SNP 3
rs4495747	1	97855607	G	A	1.67 (1.31-2.13)	$3.91X\ 10^{-5}$	0	1.0 (3)
rs12021567	1	97856946	T	C	1.67 (1.31-2.13)	$4.09X\ 10^{-5}$	0	1.0 (3)
rs12040763	1	97857061	C	T	1.67 (1.31-2.13)	$4.09X\ 10^{-5}$	0	1.0 (3)
rs12043125	1	97857145	T	G	1.67 (1.31-2.13)	$4.09X\ 10^{-5}$	0	0.9 (3)
rs1112314	1	97850697	T	C	1.67 (1.31-2.13)	$4.20X\ 10^{-5}$	0	1.0 (3)
rs1356917	1	97852258	C	A	1.66 (1.30-2.12)	$5.05X\ 10^{-5}$	0	1.0 (3)
rs10875047	**1**	**97594994**	**C**	**T**	**1.70 (1.31-2.19)**	**$5.28X\ 10^{-5}$**	**0**	**0.1 (1)**
rs11165784	1	97584685	G	C	1.70 (1.31-2.19)	$5.30X\ 10^{-5}$	0	0.2 (1)
rs12566907	1	97862237	C	T	1.54 (1.23-1.93)	$2.05X\ 10^{-4}$	0	0.8 (3)
rs7540201	1	97860321	A	C	1.54 (1.22-1.93)	$2.17X\ 10^{-4}$	0	0.8 (3)
rs11799399	1	97604531	T	C	1.63 (1.26-2.11)	$2.21X\ 10^{-4}$	0	0.1 (1)
rs10875076	1	97837564	T	C	1.52 (1.21-1.91)	$3.07X\ 10^{-4}$	0	0.8 (3)
rs1415683	1	97845053	G	T	1.52 (1.21-1.91)	$3.12X\ 10^{-4}$	0	0.8 (3)
rs1890138	1	97839016	G	A	1.52 (1.21-1.91)	$3.42X\ 10^{-4}$	0	0.8 (3)
rs12039249	1	97603679	C	T	1.58 (1.23-2.03)	$3.48X\ 10^{-4}$	0	0.1 (1)
rs1709409	1	97602726	C	T	1.55 (1.21-1.99)	$4.81X\ 10^{-4}$	0	0.1 (1)
rs1760217	1	97602994	G	A	1.54 (1.20-1.97)	$6.06X\ 10^{-4}$	0	0.1 (1)

rs number	chromosome	position	Risk allele	Alternative allele	OR (95% CI)	P value	I^2 (heterogeneity)	r^2 with one of the lead SNPs (lead SNP #)
rs11165779	1	97564650	T	C	2.00 (1.32-3.01)	$1.03X\ 10^{-3}$	0	0.8 (1)
rs4434871	1	97873007	C	T	1.57 (1.20-2.07)	$1.24X\ 10^{-3}$	0	0.5 (3)
rs2244500	18	661005	A	G	1.37 (1.67-1.14)	$1.26X\ 10^{-3}$	0	0.6 (2)
rs12563828	1	97830721	T	A	1.41 (1.13-1.77)	$2.63X\ 10^{-3}$	0	0.4 (3)
rs10875071	1	97814678	C	T	1.47 (1.14-1.89)	$2.70X\ 10^{-3}$	0	0.6 (3)
rs11165845	1	97819405	A	G	1.39 (1.11-1.75)	$3.41X\ 10^{-3}$	0	0.4 (3)
rs2606246	18	678847	T	C	1.43 (1.79-1.12)	$3.57X\ 10^{-3}$	0	0.3 (2)
rs6678858	1	97878565	T	A	1.44 (1.11-1.86)	$6.08X\ 10^{-3}$	0	0.2 (3)
rs1879375	1	97807335	G	A	1.41 (1.10-1.81)	$6.62X\ 10^{-3}$	0	0.6 (3)
rs4497250	1	97882933	A	G	1.40 (1.10-1.78)	$6.67X\ 10^{-3}$	0	0.3 (3)
rs2847154	18	687270	G	A	1.37 (1.09-1.75)	$7.02X\ 10^{-3}$	0	0.3 (2)
rs7556439	1	97771947	A	C	1.35 (1.08-1.68)	$8.25X\ 10^{-3}$	0	0.2 (3)
rs628959	1	97738860	C	T	1.35 (1.08-1.69)	$8.45X\ 10^{-3}$	0	0.2 (3)
rs507170	1	97738354	G	C	1.35 (1.08-1.69)	$8.47X\ 10^{-3}$	0	0.2 (3)
rs2612081	18	695030	G	A	1.37 (1.09-1.72)	$8.48X\ 10^{-3}$	0	0.3 (2)
rs644428	1	97737704	C	T	1.35 (1.08-1.69)	$8.51X\ 10^{-3}$	0	0.2 (3)
rs553388	1	97737348	C	T	1.35 (1.08-1.69)	$8.51X\ 10^{-3}$	0	0.2 (3)
rs526645	1	97749380	G	A	1.33 (1.08-1.67)	$9.18X\ 10^{-3}$	0	0.2 (3)
rs1609519	1	97781039	G	A	1.34 (1.07-1.67)	$1.01X\ 10^{-2}$	0	0.2 (3)
rs13233308	7	87244960	T	C	1.30 (1.06-1.58)	$1.07X\ 10^{-2}$	0	Lead SNP 4
rs10875061	1	97744048	G	A	1.28 (1.05-1.56)	$1.56X\ 10^{-2}$	0	0.1 (3)
rs11165837	1	97759020	T	A	1.37 (1.09-1.75)	$1.58X\ 10^{-2}$	0	0.1 (3)
rs11165827	1	97739908	A	T	1.28 (1.04-1.56)	$1.65X\ 10^{-2}$	0	0.1 (3)
rs10783058	1	97761972	T	C	1.28 (1.04-1.56)	$1.75X\ 10^{-2}$	0	0.1 (3)
rs10783057	1	97730626	G	A	1.28 (1.04-1.56)	$1.78X\ 10^{-2}$	0	0.1 (3)
rs9782950	1	97803724	C	T	1.30 (1.05-1.62)	$1.78X\ 10^{-2}$	0	0.3 (3)

rs number	chromosome	position	Risk allele	Alternative allele	OR (95% CI)	P value	I^2 (heterogeneity)	r^2 with one of the lead SNPs (lead SNP #)
rs7522938	1	97727820	G	C	1.28 (1.04-1.56)	$1.85X\ 10^{-2}$	0	0.1 (3)
rs12031561	1	97903583	G	A	1.26 (1.04-1.53)	$1.99X\ 10^{-2}$	0.38	0.1 (3)
rs641805	1	97743805	A	T	1.27 (1.03-1.54)	$2.15X\ 10^{-2}$	0	0.1 (3)
rs12726453	1	97750189	C	T	1.26 (1.03-1.54)	$2.17X\ 10^{-2}$	0	0.1 (3)
rs11165875	1	97915213	C	T	1.25 (1.03-1.52)	$2.28X\ 10^{-2}$	0	0.1 (3)
rs6593642	1	97750838	T	C	1.26 (1.03-1.53)	$2.36X\ 10^{-2}$	0	0.1 (3)
rs4148424	13	95931490	G	A	1.31 (1.04-1.67)	$2.36X\ 10^{-2}$	0	Lead SNP 5
rs614664	3	124486993	A	C	1.26 (1.03-1.54)	$2.41X\ 10^{-2}$	0.64	Lead SNP 6
rs1729788	13	95808003	G	A	1.27 (1.03-1.58)	$2.62X\ 10^{-2}$	0	0.04 (5)
rs2725256	4	89050998	G	A	1.25 (1.03-1.52)	$2.70X\ 10^{-2}$	0	Lead SNP 7
rs13336470	16	66999370	G	A	1.36 (1.04-1.79)	$2.73X\ 10^{-2}$	0	Lead SNP 8
rs12028565	1	97894619	C	T	1.24 (1.02-1.51)	$2.87X\ 10^{-2}$	0	0.1 (3)
rs1564481	4	89061265	T	C	1.24 (1.02-1.52)	$2.88X\ 10^{-2}$	0	1 (7)
rs11873007	18	680380	C	T	1.27 (1.02-1.56)	$2.95X\ 10^{-2}$	0	0.3 (2)
rs3819101	18	677240	G	A	1.27 (1.02-1.56)	$2.96X\ 10^{-2}$	0	0.3 (2)
rs12535512	7	87220334	C	T	1.25 (1.02-1.52)	$3.01X\ 10^{-2}$	0	0.8 (4)
rs3786355	18	681962	G	A	1.27 (1.02-1.56)	$3.13X\ 10^{-2}$	0	0.3 (2)
rs7325861	13	95912228	T	G	2.04 (1.06-3.85)	$3.25X\ 10^{-2}$	0	0.04 (5)
rs528455	1	97749198	T	C	1.23 (1.02-1.52)	$3.30X\ 10^{-2}$	0	0.1 (3)
rs4148733	7	87213232	A	G	1.37 (1.03-1.85)	$3.32X\ 10^{-2}$	0.45	0.1 (4)
rs12047910	1	97600039	A	G	1.35 (1.02-1.79)	$3.47X\ 10^{-2}$	0	0.01 (1)
rs4693930	4	89122833	A	G	1.23 (1.01-1.50)	$3.53X\ 10^{-2}$	0	0.1 (7)
rs4949952	1	97886171	T	C	1.23 (1.01-1.52)	$3.55X\ 10^{-2}$	0.30	0.03 (1)
rs4148732	7	87234049	T	C	1.37 (1.02-1.85)	$3.81X\ 10^{-2}$	0.43	0.1 (4)
rs2622629	4	89094064	C	T	1.23 (1.01-1.51)	$4.02X\ 10^{-2}$	0	0.7 (7)
rs2766482	13	95785721	T	G	1.23 (1.01-1.51)	$4.09X\ 10^{-2}$	0	0.04 (5)

rs number	chromosome	position	Risk allele	Alternative allele	OR (95% CI)	P value	I^2 (heterogeneity)	r^2 with one of the lead SNPs (lead SNP #)
rs3821536	3	124483952	C	T	1.30 (1.01-1.67)	$4.20X\ 10^{-2}$	0	0.4 (6)
rs2291081	3	124485235	G	A	1.30 (1.01-1.67)	$4.26X\ 10^{-2}$	0	0.4 (6)
rs899498	13	95804316	A	C	1.25 (1.01-1.54)	$4.36X\ 10^{-2}$	0	0.02 (5)
rs4148432	13	95913082	C	T	1.89 (1.02-3.45)	$4.39X\ 10^{-2}$	0	0.04 (5)
rs7986087	13	95915745	C	T	1.85 (1.02-3.33)	$4.42X\ 10^{-2}$	0	0.04 (5)
rs2853151	8	105396792	T	C	1.85 (1.01-3.33)	$4.56X\ 10^{-2}$	0	Lead SNP 9
rs7550959	1	97926839	G	A	1.22 (1.00-1.47)	$4.59X\ 10^{-2}$	0	0.1 (3)
rs2235035	7	87179086	G	A	1.23 (1.00-1.54)	$4.68X\ 10^{-2}$	0.17	0.2 (4)
rs1922240	7	87183354	T	C	1.23 (1.00-1.54)	$4.72X\ 10^{-2}$	0.17	0.2 (4)
rs1479390	13	95803139	T	G	1.24 (1.00-1.54)	$4.84X\ 10^{-2}$	0	0.06 (5)
rs2651204	6	43259087	T	C	1.41 (1.00-1.96)	$4.85X\ 10^{-2}$	0	Lead SNP 10
Exome array								
rs67376798	1	97547947	A	T	10.0 (2.50-33.3)	$9.74X\ 10^{-4}$	0	0 (1 and 3)
rs11165846	1	97819667	G	C	1.42 (1.14-1.78)	$1.89X\ 10^{-3}$	0	0.4 (3)
rs147266709	9	134398452	T	C	2.79 (1.42-5.47)	$2.82X\ 10^{-3}$	0.68	lead SNP 11
rs9616787	22	50943506	T	C	2.90 (1.34-6.28)	$6.89X\ 10^{-3}$	0.30	lead SNP 12
rs11081251	18	674440	A	C	1.28 (1.04-1.59)	$1.75X\ 10^{-2}$	0	0.4 (1)
rs61122623	7	87196129	T	C	7.19 (1.18-43.7)	$3.21X\ 10^{-2}$	0	lead SNP 13
rs36092077	3	183753777	G	A	1.31 (1.02-1.68)	$3.48X\ 10^{-2}$	0	lead SNP 14

Supplementary Table 4. Testing *TYMS* rs2612091, 5' VNTR and 3'UTR haplotypes for independent effects of one polymorphism

Haplotype analyses were performed in PLINK using the *--independent-effect* command, in which for each polymorphism in turn, alleles are analysed for an association with toxicity whilst keeping the genotypes of the other polymorphisms constant. The test produces a p-value for each such test and then an overall p-value for that polymorphism which shows whether that polymorphism has a consistent association with toxicity regardless of background haplotype genotype. The first three panels show the effects of varying the 5' VNTR allele, 3'UTR allele and rs2612091 allele respectively. Only rs2612091 shows a significant effect overall. The lower two panels show two-polymorphism analyses in which rs2612091 is varied whilst 5'VNTR and 3'UTR alleles are held constant. Note that some rare haplotypes are not shown.

Test SNP	5'VNTR 3'UTR rs2612091	OR for effect of each test SNP allele on haplotypes	OR for pooled effect of both test SNP alleles	p-value
5'VNTR (3 SNP model)	2R/ins/G	1 (ref)	1 (ref)	0.65
	3R/ins/G	1.04		
	2R/del/A	0.89	0.83	0.34
	3R/del/A	0.82		
	2R/ins/A	0.95	0.80	0.081
	3R/del/A	0.77		
	overall			0.17
3'UTR (3 SNP model)	2R/ins/G	1 (ref)	1 (ref)	n/a
	3R/ins/G	1.04	1.04	n/a
	2R/del/A	0.89	0.92	0.67
	2R/ins/A	0.95		
	3R/del/A	0.82	0.80	0.33
	3R/del/A	0.77		
	overall			0.61
rs2612091 (3 SNP model)	2R/ins/G	1 (ref)	1 (ref)	0.66
	2R/ins/A	0.95		
	3R/ins/G	1.04	0.84	0.00068
	3R/ins/A	0.77		
	2R/del/A	0.89	0.88	n/a
	3R/del/A	0.82	0.82	n/a
	overall			0.0021
rs2612091 (2 SNP model)	2R/G	1 (ref)	1 (ref)	0.18
	2R/A	0.92		
	3R/G	1.04	0.84	0.00051
	3R/A	0.79		
	overall			0.00053
rs2612091 (2 SNP model)	ins/G	1 (ref)	1 (ref)	
	ins/A	0.80		n/a
	del/A	0.83	0.90	n/a
	overall			1.47E-06

Supplementary Table 5. Set test analyses of capecitabine/5-FU pathway genes
In order to determine whether there was evidence in QUASAR2 of additional toxicity associations that had not reached formal statistical significance for individual SNPs or rare variants, we performed association tests based on sets of variants. The set tests used SNPs within 25kb of each of the 25 capecitabine/5-FU pathway genes plus *ENOSF1*. Prior to analysis, the known *DPYD* 2846 and *2A variants and the newly identified *DPYD* rs12132152, *DPYD* rs7548189 and *TYMS* rs2612091, as well as anything in linkage disequilibrium of $r^2>0.1$ with these SNPs (including the *TYMS* 5'VNTR and 3'UTR polymorphisms), were removed. Tests were performed by individually testing the association of each SNP under an allelic model using logistic regression adjusted for age, treatment arm and gender, permuting the outcome data and re-testing 10,000 times, then comparing the observed distribution of p-values to those from randomly assigned toxicity data for each set (i.e. per gene or across all SNPs). Using a false discovery rate of q=0.05 (p~0.005), we found no convincing evidence for additional associations at any gene or in the set of variants as a whole. We did, however, note suggestive evidence of associations between variants at the *TYMP* locus and HFS and diarrhoea.

SET	No. SNPs	Global 012v34	Global 01v2v34	HFS 012v34	HFS 01v2v34	Diarrhoea 012v34	Diarrhoea 01v2v34
ABCB1	77	0.29	1	1	1	1	1
ABCC3	64	1	1	1	1	0.55	0.49
ABCC4	221	0.86	0.87	1	1	0.91	0.75
ABCC5	100	0.33	1	1	1	0.13	0.074
ABCG2	57	0.093	1	0.37	0.60	1	1
CDA	25	1	0.24	0.38	0.20	1	0.40
CES1	24	1	0.10	0.19	0.11	1	0.16
CES2	59	0.18	0.28	0.0092	0.046	1	1
DPYD	189	0.31	0.040	0.52	0.52	0.53	0.20
DPYS	69	1	0.32	1	0.58	1	0.70
ENOSF1	22	0.11	0.27	1	0.25	1	1
MTHFR	37	1	1	1	1	1	0.50
PPAT	29	1	1	1	1	1	0.10
RRM1	29	1	1	0.17	1	1	1
RRM2	19	1	1	1	1	1	1
SLC22A7	26	1	1	1	1	0.014	0.22
SLC29A1	26	1	1	1	1	1	1
TK1	35	1	1	1	0.43	1	1
TYMP	92	1	0.24	0.035	0.025	0.055	0.029
TYMS	23	0.12	0.28	1	0.25	1	1
UCK1	43	0.13	0.16	0.25	0.36	0.013	0.0038
UCK2	21	1	1	1	1	0.063	0.31
UMPS	34	0.13	0.085	0.15	0.035	1	0.32
UPB1	30	1	1	1	1	1	1
UPP1	16	1	1	1	1	0.089	1
UPP2	42	1	0.061	0.078	0.25	1	1
As One Set*	**1393**	**0.72**	**0.36**	**0.67**	**0.55**	**0.10**	**0.12**

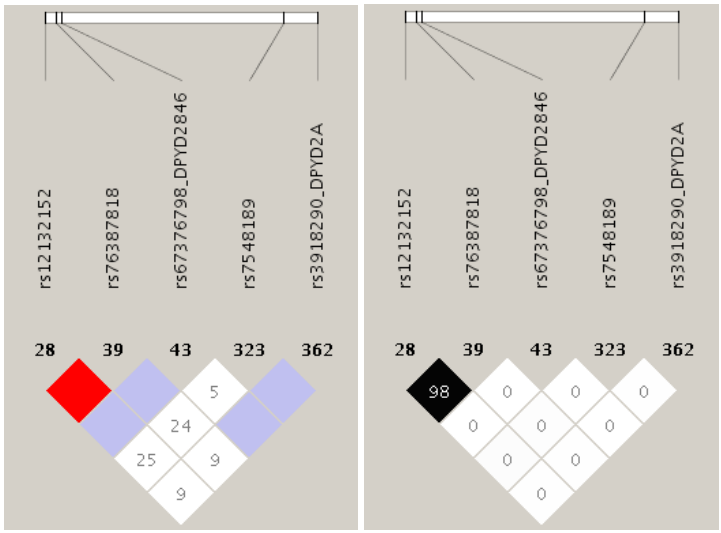
*There is some overlap in the SNPs contained in the *TYMS* and *ENOSF1* set.

Supplementary Table 6. Associations between *DPYD* coding regions variants and capecitabine toxicity in QUASAR2.
The Table shows polymorphisms and rare variants present on the tagSNP or exome arrays, together with summary statistics of association with toxicity in the meta-analysis of the two arms of QUASAR2. MAF=minor allele frequency.

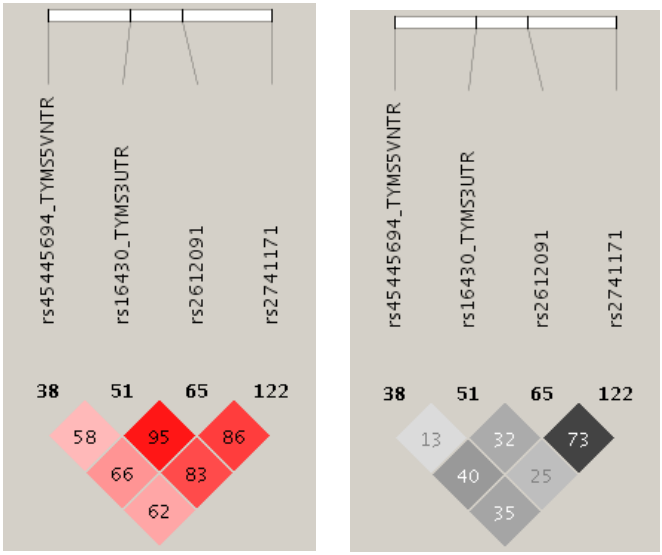
Variant ID						Arm A						Arm B						Overall analysis	
Chr	Position	rs#	Classical ID	Nucleotide	Amino Acid	cases_AA	cases_AB	cases_BB	controls_AA	controls_AB	controls_BB	cases_AA	cases_AB	cases_BB	controls_AA	controls_AB	controls_BB	Meta Beta	Meta P
1	97,547,947	rs67376798	2846A>T	c.T2846A	p.D949V	0	5	137	0	2	328	0	3	190	0	0	340	-2.284433	0.0010
1	97,770,920	rs1801160	*6	c.G2194A	p.V732I	123	11	0	286	19	0	170	13	0	296	20	2	0.058056	0.8277
1	97,770,920	rs1801160	*6	c.G2194A	p.V732I	130	12	0	311	19	0	178	15	0	319	19	2	0.245417	0.3453
1	97,915,614	rs3918290	*2A	c.1905+1G>A	exon skipping	140	2	0	330	0	0	192	1	0	338	2	0	1.308234	0.1793
1	97,981,343	rs55886062	*13	c.T1679G	p.I560S	141	1	0	330	0	0	192	1	0	339	1	0	0.57783	0.6975
1	97,981,395	rs1801159	*5	c.A1627G	p.I543V	89	33	12	198	95	11	116	61	6	208	98	12	0.071863	0.5601
1	97,981,421	rs1801158	*4	c.G1601A	p.S534N	135	7	0	317	13	0	183	10	0	328	12	0	0.302946	0.3688
1	98,039,419	rs56038477		c.G1236A	p.E412E	126	8	0	296	9	0	171	12	0	312	6	0	0.967362	0.0081
1	98,144,726	rs45589337		c.A775G	p.K259E	0	2	140	0	2	328	0	2	191	0	8	332	0.206601	0.7231
1	98,165,091	rs2297595		c.A496G	p.M166V	110	24	0	242	59	4	152	30	1	265	52	1	-0.141077	0.4151
1	98,348,885	rs1801265	*9A	c.C85T	p.C29R	8	40	86	18	110	177	8	59	116	21	117	180	0.204385	0.0781

Supplementary Figure 1. LD between selected variants near (a) *DPYD* and (b) *TYMS/ENOSF1* (left = D'; right=R^2).
Haplotype frequencies from Haploview EM algorithm are shown for *TYMS.*

(a) *DPYD*



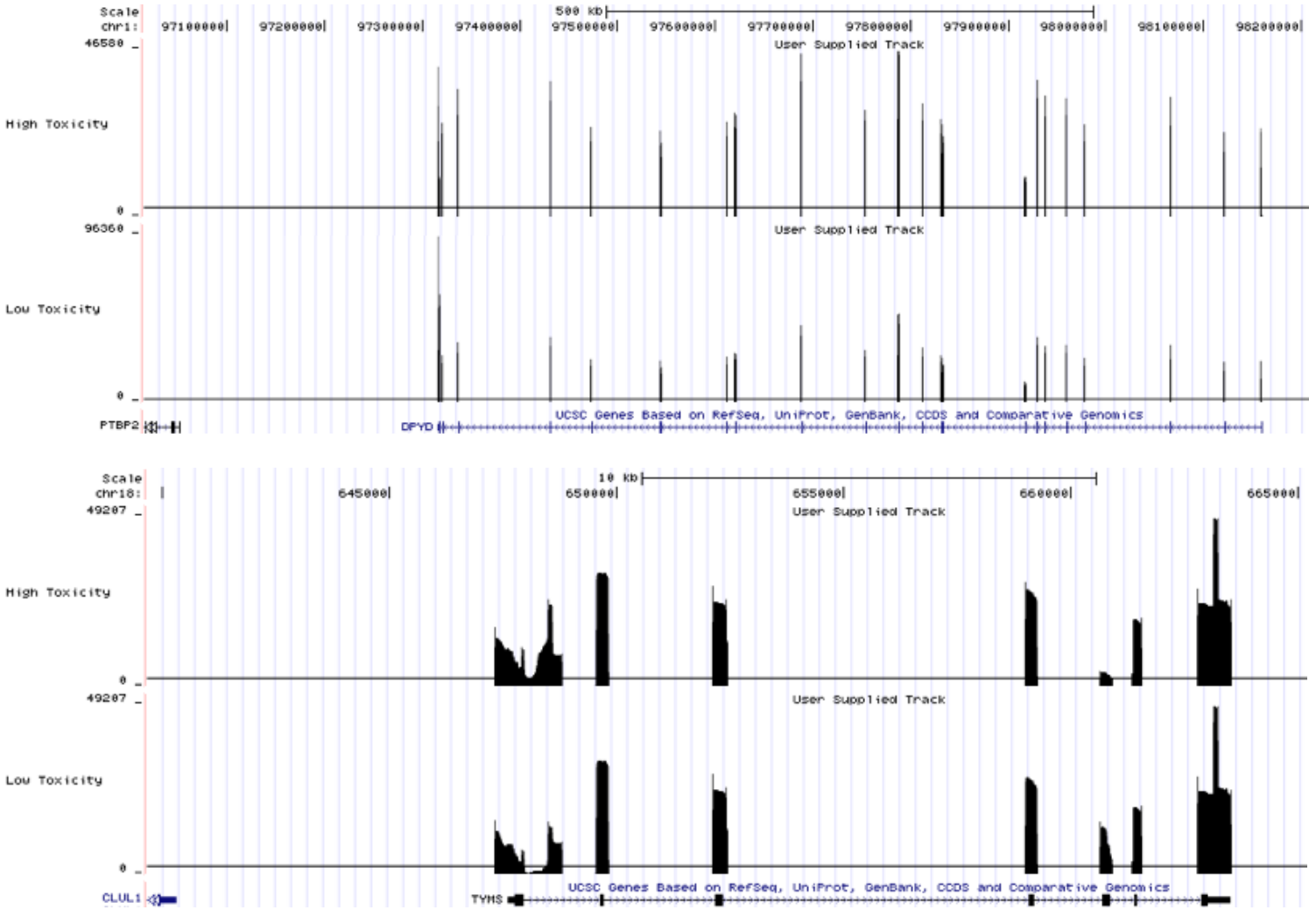
(b) *TYMS*



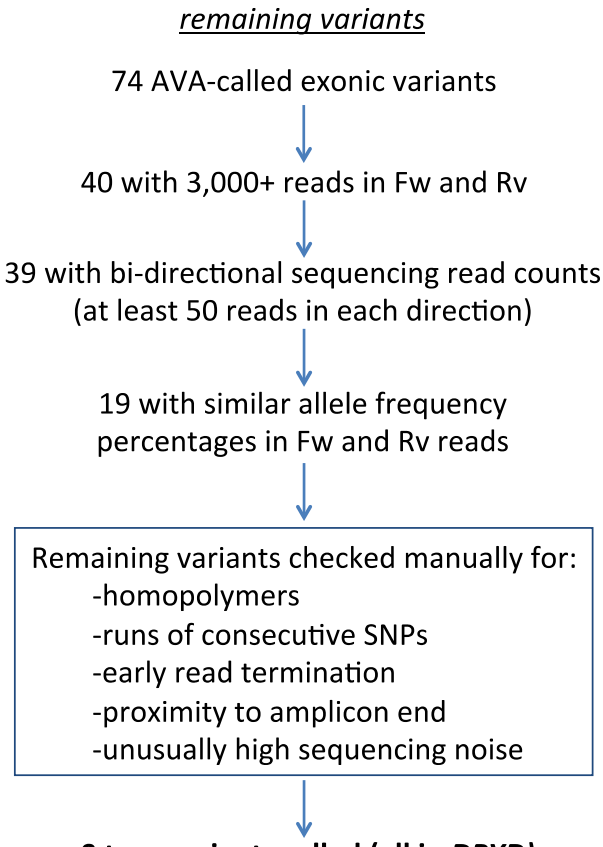
5'VNTR 3'UTR rs2612091		r=risk p=protective
2R/ins/G	0.366	r/r/r
3R/del/A	0.244	p/p/p
3R/ins/A	0.204	p/r/p
3R/ins/G	0.073	p/r/r
2R/del/A	0.061	r/p/p
2R/ins/A	0.045	r/r/p

Supplementary Figure 2. Sequencing coverage achieved

UCSC Bioinformatics Genome Browser user-supplied tracks showing the coverage achieved across *DPYD* (top) and *TYMS* (bottom) with Roche/454 amplicon sequencing. Top panel of each image is the coverage achieved for the 100-patient high toxicity pool; bottom panel of each image is for the 100-patient low toxicity pool. Y-axis range is from 0 coverage to over 40,000 reads per locus per pool; horizontal line on each panel marks 3,000 read coverage per pool (ie, 30x coverage per patient), achieved for all exonic loci except for some in *TYMS* exon 1 and exon 5; loci not reaching coverage goal were filtered prior to analysis.



Supplementary Figure 3. Filtering of variants identified by Roche/454 sequencing of *TYMS* and *DPYD* exons



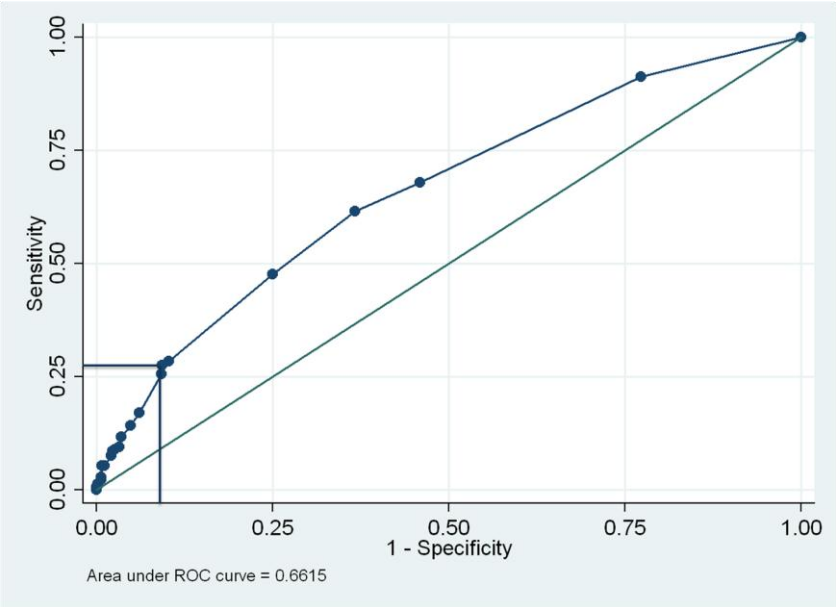
7 dbSNPs on SNP/EXM array	1 dbSNP not on SNP/EXM array	1 novel variant
rs56038477 (1236)	rs17376848 (T1896C; F632F)	1:97981371 (G1651A; A551T)
rs1801158 (1601)		
rs1801159 (1627)		
rs1801265 (85)		
rs1801160 (2194)		
rs67376798 (2846)		
rs2297595 (496)		

Supplementary Figure 4. ROC curve

No additional independent data set was available to test the performance of a model to predict 5-FU toxicity based on the previously-reported capecitabine toxicity variants and our new data. However, in order to provide clues as to the possible clinical utility of our findings, we used the QUASAR2 data set and incorporated *DPYD* 2846T>A (rs67376798), *DPYD* *2A (rs3918290), *DPYD* rs12132152, *DPYD* rs7548189, *DPYD* p.Ala551Thr, and *TYMS* rs2612091 into a ROC analysis for prediction of global grade 012v34 capecitabine-related toxicity. 938 patients were analysed, applying a score for each patient that summed

(number of harmful alleles at each polymorphism) x (beta coefficient per allele)

The three rare *DPYD* variants were assumed to be functionally equivalent and hence combined for the purposes of this analysis into a test of any rare functional allele *versus* no rare allele (OR=7.6, $p=4.5x10^{-4}$). We found the area under curve (AUC) to be 0.66 (95% CI 0.63-0.70). At the cut-off for which the maximum proportion of patients were correctly classified (69%), sensitivity was 27% (95% CI 23-33%), specificity was 91% (95% CI 88-93%), positive predictive value was 60% (PPV: 95% CI 52-68%), and negative predictive value was 71% (NPV: 95% CI 68-74%) (Figure 3). Although this result must be treated cautiously given that it is derived from the same data set used for variant discovery, we note that using just the previously-reported *DPYD* 2846T>A and *2A and *TYMS* 5'VNTR and 3'UTR variants, the equivalent AUC was lower at 0.59.



Supplementary Methods

Genotyping

DNA was extracted from buffy coat samples using conventional methods and samples with sufficient DNA and complete clinical data (N=994) were genotyped on the Hap300/370CNV, Hap610 or Omni2.5 Illumina tagging SNP arrays. 29 samples were excluded following principal component analysis as they did not cluster with CEU HapMap3 samples, 12 samples were excluded because per sample call rates were < 95% and 7 samples were excluded because of gender discrepancies. Quality control procedures were performed to eliminate poorly-performing polymorphisms, as described in [1]. After applying quality control procedures QUASAR 2 patient genotypes were available for SNPs present on the Illumina Hap300/370CNV (N=484), Hap610 (N=364) or Omni2.5 (N=92) tagging SNP arrays. Data were also available for a largely overlapping set of 968 QUASAR2 patients genotyped on the Illumina HumanExome12v1_A or -12v1-1_A arrays, which were designed to capture uncommon protein-coding variation [2]. The additional samples genotyped on the exome arrays but not tagging SNP arrays had missing DNA or clinical data at time of genotyping using tagging arrays. Base calling for all platforms was performed using Illumina Genome Studio and, for exome arrays, additionally by Z-Caller [3], applying a z-score of 7 based on the concordance of calls with Illumina Genome Studio for common variants (99.3%). 8,694 polymorphisms were present on both the SNP tagging arrays and exome array and there was 99.2% genotyping concordance for these SNPs.

For each of the 25 capecitabine/5-FU pathway genes (Supplementary Table 1), we identified genetic variants that were present on one or more of the Hap300/370, Hap610 or exome arrays and that lay within 25kb of the coding region of one of the genes. We used imputation to obtain missing genotypes arising from differences in array content: haplotypes were phased using SHAPEITv2 [4] and imputation performed using IMPUTEv2 [5], employing a 250kb buffer region and the 1000 genomes August 2012 release (all ethnicities) as a reference panel. Only SNPs with an IMPUTEv2 info score of at least 0.95 on each array individually were taken into further analysis by SNPTESTv2 [6]. Further exclusion criteria were a SNPTEST info score below 0.95 on the

pooled score from the three SNP arrays, a minor allele frequency below 0.01 and a Hardy-Weinberg equilibrium p-value below 0.0001. Genotyping and imputation provided a total of 1,456 genetic variants for analysis.

The accuracy of imputation was further tested for specific SNPs using constitutional genotypes from 190 white UK individuals who had been whole genome-sequenced using the Complete Genomics platform. The input for IMPUTE2 was genotype files that represented the SNP content of the arrays used to genotype QUASAR2. The imputed SNP genotype probabilities were converted into genotypes using gtool only if the probability of a particular genotype was ≥0.9. Real genotypes and imputed ones were then compared to determine concordance and missingness.

Further genotyping was performed for the *TYMS* 5'VNTR and 3'UTR variants by previously-described methods [7, 8]. Additional genotyping of the *DPYD* 2846T>A and *DPYD* *2A variants was performed by allele-specific amplification by KASPar [9] for the small number of patients not genotyped using the exome arrays.

For loci at which significant or borderline significant associations between genetic variants and toxicity were detected, we performed fine mapping studies by using the methods above to impute all SNPs in a 1.5Mb flanking region, in order to refine the association signal.

454 Sequencing

Sequencing of the coding regions of *DPYD* and *TYMS* was performed by Roche/454 Titanium GS FLX technology according to the specified amplicon sequencing protocol (see http://454.com/downloads/my454/documentation/gs-junior/method-manuals/GSJunior_AmpliconLibraryPrep-RevJune2010.pdf; further details available from authors). Specifically, we selected the 100 patients with the highest levels of 5-FU-related toxicity ("HiTox"), specifically grade 3 or grade 4 diarrhoea in the first 4 cycles of treatment and or other grade 3/4 toxicities in the first 4 cycles of treatment. We also selected 100 patients with no adverse toxicity events during the entire duration of

treatment ("LoTox"). We used primer3 to design PCR primers and reactions to cover all 23 *DPYD* exons (27 amplicons; 4,784bp) and 7 *TYMS* exons (9 amplicons; 2,276bp) (primers and conditions available upon request). Constitutional DNA samples from each patient were quantitated using PicoGreen, diluted to equal measured concentrations and formed into 10 pools of 20 patients each. Pools were then PCR-amplified for each of the 36 amplicons. Missing or undesired amplicons were identified by an Agilent High Sensitivity DNA Kit. Successful amplicons were quantified by PicoGreen according to the 454 protocol, equalised in concentration and formed into one 100-patient HiTox pool and one 100-patient LoTox pool for sequencing. We aimed to achieve a minimum read depth of 3,000 per target locus per pool (that is about 30x coverage per patient).

Mapping and initial variant calling were performed by Roche/454 software (GS Mapper and AVA). Variants were then filtered to include only those with 3,000+ reads in total, at least 50 reads in each direction and similar allele frequencies in the forward and reverse directions. Variants were then removed if they fell within a homopolymer, a run of consecutive SNPs, an early-terminating read, the end of a full length read, or an area of evidently poor sequence quality. Variant frequencies were determined, per pool, as the proportion of total reads (forward plus reverse) containing the minor allele. Within our targets, we confirmed the presence and allele frequency of known SNPs using our array data. The novel variants were validated with Sanger sequencing of the individual patients who comprised the pool (details available on request).

Functional annotation of variants was performed with ANNOVAR. mRNA expression data were obtained from Genevar [10] and from The Cancer Genome Atlas (TCGA). We analysed these data according to the methods of Li et al [11].

Putative associations with toxicity were determined according to the estimated number of variant and wildtype reads present in the HiTox and LoTox pools (Pearson's Chi Squared or Fisher's exact test).

Supplementary References

1. Dunlop MG, Dobbins SE, Farrington SM*, et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet 2012; **44**: 770-6.
2. Huyghe JR, Jackson AU, Fogarty MP*, et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 2013; **45**: 197-201.
3. Goldstein JI, Crenshaw A, Carey J*, et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics 2012; **28**: 2543-5.
4. Delaneau O, Howie B, Cox AJ*, et al.* Haplotype estimation using sequencing reads. Am J Hum Genet 2013; **93**: 687-96.
5. Howie B, Fuchsberger C, Stephens M*, et al.* Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012; **44**: 955-9.
6. Marchini J, Howie B, Myers S*, et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007; **39**: 906-13.
7. Dotor E, Cuatrecases M, Martinez-Iniesta M*, et al.* Tumor thymidylate synthase 1494del6 genotype as a prognostic factor in colorectal cancer patients receiving fluorouracil-based adjuvant treatment. J Clin Oncol 2006; **24**: 1603-11.
8. Horie N, Aiba H, Oguro K*, et al.* Functional analysis and DNA polymorphism of the tandemly repeated sequences in the 5'-terminal regulatory region of the human gene for thymidylate synthase. Cell Struct Funct 1995; **20**: 191-7.
9. Cuppen E. Genotyping by Allele-Specific Amplification (KASPar). CSH Protoc 2007; **2007**: pdb prot4841.
10. Yang TP, Beazley C, Montgomery SB*, et al.* Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics 2010; **26**: 2474-6.
11. Li Q, Seo JH, Stranger B*, et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 2013; **152**: 633-41.