Human Molecular Genetics, 2012, Vol. 21, No. 8 ***1918–1930***
doi:10.1093/hmg/ddr619
Advance Access published on January 6, 2012

Mapping of the *UGT1A* locus identifies an uncommon coding variant that affects mRNA expression and protects from bladder cancer

Wei Tang[1], Yi-Ping Fu[1], Jonine D. Figueroa[2], Núria Malats[3], Montserrat Garcia-Closas[2,4], Nilanjan Chatterjee[2], Manolis Kogevinas[5,6,7,8], Dalsu Baris[2], Michael Thun[9], Jennifer L. Hall[10], Immaculata De Vivo[11], Demetrius Albanes[2], Patricia Porter-Gill[1], Mark P. Purdue[2], Laurie Burdett[12], Luyang Liu[1], Amy Hutchinson[12], Timothy Myers[12], Adonina Tardón[7,13], Consol Serra[14], Alfredo Carrato[15], Reina Garcia-Closas[16], Josep Lloreta[17], Alison Johnson[18], Molly Schwenn[19], Margaret R. Karagas[20], Alan Schned[21], Amanda Black[2], Eric J. Jacobs[9], W. Ryan Diver[9], Susan M. Gapstur[9], Jarmo Virtamo[22], David J. Hunter[23], Joseph F. Fraumeni Jr[2], Stephen J. Chanock[1], Debra T. Silverman[2], Nathaniel Rothman[2,†] and Ludmila Prokunina-Olsson[1,*,†]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics and [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA, [3]Spanish National Cancer Research Centre, Madrid 28029, Spain, [4]Division of Genetics and Epidemiology, Institute of Cancer Research, London SW7 3RP, UK, [5]Centre for Research in Environmental Epidemiology (CREAL), Barcelona 08003, Spain, [6]Municipal Institute of Medical Research, Barcelona 08003, Spain, [7]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona 08003, Spain, [8]National School of Public Health, Athens 11521, Greece, [9]Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA, [10]Lillehei Heart Institute, Department of Medicine, University of Minnesota, Minneapolis, MN 55455, USA, [11]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA, [12]Core Genotype Facility, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick, MD 21702, USA, [13]Universidad de Oviedo, Oviedo 33003, Spain, [14]Universitat Pompeu Fabra, Barcelona 08002, Spain, [15]Ramón y Cajal University Hospital, Madrid 28034, Spain, [16]Unidad de Investigación, Hospital Universitario de Canarias, La Laguna 38320, Spain, [17]Hospital del Mar-Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra, Barcelona 08003, Spain, [18]Vermont Cancer Registry, Burlington, VT 05401, USA, [19]Maine Cancer Registry, Augusta, ME 04333, USA, [20]Dartmouth Medical School, Hanover, NH 03755, USA, [21]Department of Urology, Washington University School of Medicine, St Louis, MO 63110, USA, [22]National Institute for Health and Welfare, Helsinki 00271, Finland and [23]Department of Epidemiology, Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

Received August 26, 2011; Revised November 10, 2011; Accepted December 30, 2011

A recent genome-wide association study of bladder cancer identified the *UGT1A* gene cluster on chromosome 2q37.1 as a novel susceptibility locus. The *UGT1A* cluster encodes a family of UDP-glucuronosyltransferases (UGTs), which facilitate cellular detoxification and removal of aromatic amines. Bioactivated forms of aromatic amines found in tobacco smoke and industrial chemicals are the main risk factors for bladder cancer. The association within the *UGT1A* locus was detected by a single nucleotide polymorphism (SNP) rs11892031. Now, we performed detailed resequencing, imputation and genotyping in this region. We

*To whom correspondence should be addressed at: Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, 8717 Grovemont Circle, Bethesda, MD 20892-4605, USA. Tel: +1 3014435297; Fax: +1 3014023134; Email: prokuninal@mail.nih.gov
†Co-last author.

clarified the original genetic association detected by rs11892031 and identified an uncommon SNP rs17863783 that explained and strengthened the association in this region (allele frequency 0.014 in 4035 cases and 0.025 in 5284 controls, OR = 0.55, 95%CI = 0.44–0.69, $P = 3.3 \times 10^{-7}$). Rs17863783 is a synonymous coding variant Val209Val within the functional *UGT1A6.1* splicing form, strongly expressed in the liver, kidney and bladder. We found the protective T allele of rs17863783 to be associated with increased mRNA expression of *UGT1A6.1* in *in-vitro* exontrap assays and in human liver tissue samples. We suggest that rs17863783 may protect from bladder cancer by increasing the removal of carcinogens from bladder epithelium by the UGT1A6.1 protein. Our study shows an example of genetic and functional role of an uncommon protective genetic variant in a complex human disease, such as bladder cancer.

INTRODUCTION

With 70 530 new cases and 14 680 deaths in 2010, bladder cancer (MIM 109800) is the fifth most common cancer in the USA (1). The disease is well treatable if detected early, but the high recurrence rates, life-long surveillance and treatment add up to a cost of 4 billion dollars a year, which is estimated to be higher than for other cancers in the USA (2,3).

The involvement of environmental risk factors in bladder cancer etiology was first suggested in 1895 by a German surgeon Ludwig Rehn who reported a high occurrence of bladder cancer among dye industry workers (4). This risk was later attributed to exposures to aromatic amines, such as 2-naphthylamine, 4-aminobiphenyl, 4-nitrobiphenyl, 4,4-diaminobiphenyl and benzidine, found in industrial chemicals (5). The same chemicals are found in tobacco smoke, which is now considered the main risk factor for bladder cancer (6,7). Aromatic amines are converted into biologically active carcinogens during a two-stage cellular detoxification/bioactivation process. The first stage is a hepatic N-hydroxylation of aromatic amines by the CYP1A2 enzyme, which belongs to the cytochrome P450 phase I detoxification system (8). The second stage is an enzymatic conjugation of the N-hydroxylated aromatic amines by phase II detoxification enzymes, such as N-acetyltransferases (NATs), glutathione transferases (GSTs) and UDP-glucuronosyltransferases (UGTs). The conjugation facilitates the excretion of the N-hydroxylated intermediates via stool and urine (9). However, direct exposure to the urine enriched by these highly unstable conjugates can initiate oncogenic transformation of bladder epithelium, and lead to cancer (6,7).

Familial aggregation and twin studies of bladder cancer suggest that genetic factors play a role in its etiology (10,11). Specifically, alterations within the cellular detoxification system can determine individual response to environmental exposures. Genetic variants within the phase II detoxification genes *NAT2* and *GSTM1* have already been identified as risk factors for bladder cancer (12–16). It is not surprising that the *UGT1A* gene cluster on chromosome 2q37.1 has now been linked with bladder cancer susceptibility (17). These findings suggest that cellular detoxification in humans is mediated by several distinct pathways, and alterations within these pathways could affect bladder cancer risk.

In this study, we identified a single nucleotide polymorphism (SNP), rs17863783, which explained and strengthened the genetic association of the *UGT1A* region with the risk for bladder cancer. The associated T allele of rs17863783 is a coding synonymous variant (Val209Val) that affects mRNA expression of a functional splicing form, *UGT1A6.1*. We suggest that the molecular phenotype of this genetic association is related to increased clearance of carcinogens from bladder epithelium by the UGT1A6.1 protein. Our study exemplifies a genetic and functional contribution of an uncommon protective genetic variant to bladder cancer.

RESULTS

Genetic fine-mapping of the *UGT1A* region

The genetic association with bladder cancer within the *UGT1A* gene cluster was detected for a SNP rs11892031 (17). Since multiple coding variants within the *UGT1A* genes have been previously linked with enzymatic activity for different pharmacological and environmental substrates (18), we hypothesized that rs11892031 might be in linkage disequilibrium (LD) with one or more of these functional variants. Thus, we conducted a fine-mapping study to comprehensively catalog genetic variants within the *UGT1A* locus, refine the bladder cancer genetic association and search for a functional link between this genetic association and bladder cancer risk.

The *UGT1A* region includes nine highly similar protein-coding and four non-coding genes, each with a unique alternative first exon followed by a set of common exons 2–5 (19) (Fig. 1A). Rs11892031 localizes to the first intron of both the *UGT1A8* and *UGT1A10* genes and upstream of *UGT1A9*. The activity and specificity of UGT1A proteins are greatly determined by their substrate-binding domains, which are entirely encoded by the nine alternative first exons of the corresponding *UGT1A* genes. Because of the high parology within the *UGT1A* family of genes, some of the 134 non-synonymous and 71 synonymous coding SNPs across these exons included in the current build 132 of the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) might represent misalignment of highly similar genomic sequences, rather than true genetic polymorphisms. To comprehensively catalog and verify coding variants in this region, we generated highly specific long-range amplicons and sequenced alternative first exons of each of the *UGT1A* genes in 44 bladder cancer cases and 30 trios from the HapMap European (CEU) set (www.hapmap.org). From the 156 kb *UGT1A* cluster (chr2:234,191,000–234,347,000, hg18), we sequenced 10 exons that covered 14 358 bp (9.2%) of this region. We reasoned that non-exonic variants located within unique sequences will be well-imputed based on the current reference

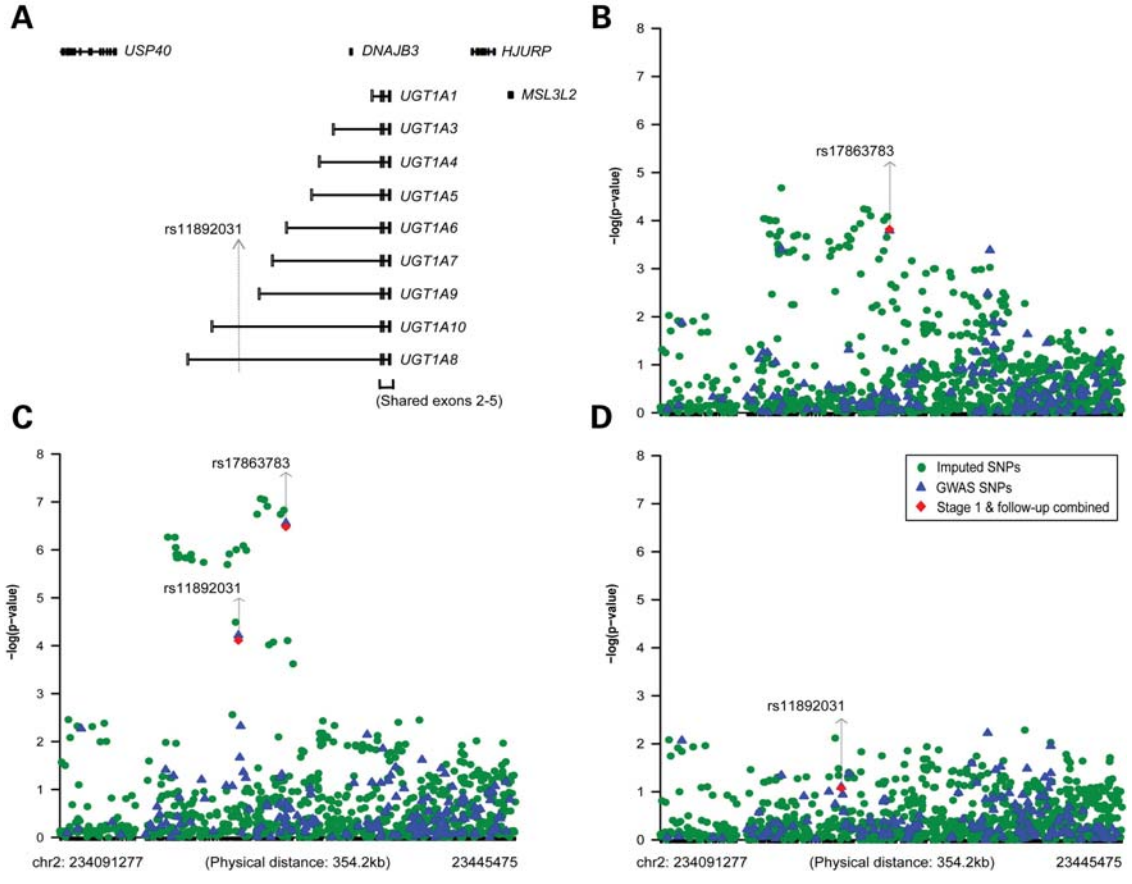

Figure 1. Genomic structure and association results for the *UGT1A* gene cluster ± 100 kb of flanking regions. (**A**) Structure of the *UGT1A* gene cluster consisting of nine genes that combine individual alternative first exons (open rectangles) with shared exons 2–5. (**B**) Association results for 1170 genotyped and imputed SNPs from the *UGT1A* gene cluster. (**C**) Association results for 1170 genotyped and imputed SNPs from the *UGT1A* gene cluster, after adjustment for the GWAS signal, rs11892031. (**D**) Association results for 1170 genotyped and imputed SNPs from the *UGT1A* gene cluster, after adjustment for the novel signal, rs17863783.

sets (HapMap 3 and 1000 Genomes (20,21)), while variants from the highly similar exonic sequences should be refined and cataloged first. We detected 46 known exonic SNPs (27 non-synonymous and 19 synonymous, Supplementary Material, Table S1), but did not identify novel variants found more than in one sample. Based on the LD pattern, we selected 18 SNPs to represent all 46 exonic variants in the *UGT1A* region. These 18 SNPs were genotyped in 1055 cases and 962 controls from the Spanish Bladder Cancer Study (SBCS) used in stage 1 genome-wide association study (GWAS) (17). Genotyping in this large set of samples was mostly done by Sanger sequencing of long-range polymerase chain reaction (PCR) fragments because several variants could be scored from the same amplicons, and/or genotyping by other methods was difficult due to high sequence similarity between alternative first exons of *UGT1A* genes. Our sequencing of 2017 samples did not reveal additional genetic variants. We ignored several genetic variants observed just once and by this we might have missed some very rare variants. For exonic variants with minor allele frequency (MAF)$>$0.01, we detected 46 SNPs, which is similar to 40 variants in the 1000 Genomes project, and 42 variants in the Exome Variant Server (http://snp.gs.washington.edu/EVS/). Based on the SBCS data enriched for coding variants across the locus (Supplementary Material, Table S2), we imputed these variants in the remaining samples in stage 1 GWAS (2477 cases/4158 controls). Using the combined HapMap 3 CEU and 1000 Genomes reference panels, we also imputed all remaining variants within 356 kb (156 kb of the *UGT1A* cluster ± 100 kb, chr2:234,091,000–234,447,000, hg18) in the entire set of stage 1 samples in the bladder cancer GWAS (3532 cases/5120 controls).

The initial GWAS included 166 SNPs in the *UGT1A* region; using imputation, we extended this panel to 1170 SNPs presented on LD plot (Supplementary Material, Fig. S1) and then performed association analysis (Supplementary Material, Table S3). In the combined set of 4035 cases and 5284 controls, the strongest association was observed for a set of 28 uncommon SNPs in high LD with each other ($r^2 > 0.9$) but in moderate LD with rs118920231 ($0.14 < r^2 \leq 0.29$) (Fig. 1B, Supplementary Material, Table S4). Of these 28 markers, only rs17863783 is a coding SNP while no functional significance could be predicted for the remaining 27 variants (Supplementary Material, Table S4). Rs17863783, with MAF of 2.5%, was genotyped in the original GWAS but was excluded from the analysis because of apparent incomplete genotyping and a standard exclusion threshold of MAF $<$ 5% (17). Here, we fully genotyped this marker in all of our samples. To

Table 1. Genetic association results for *UGT1A* SNPs and bladder cancer risk

	MAF[a] Case/control	Cases, $n = 4035$ (n, %) *AA*	*AC + CC*	Controls, $n = 5284$ (n, %) *AA*	*AC + CC*	OR[b] (95%CI)	*P*-value[b]	OR[c] (95%CI)	*P*-value[c]
rs11892031									
All subjects	0.069/0.085	3497 (86.6)	538 (13.3)	4424 (83.7)	860 (16.3)	0.79 (0.70–0.89)	7.75E − 05	0.89 (0.78–1.02)	8.32E − 02
	Case/control	*GG*	*GT + TT*	*GG*	*GT + TT*				
rs17863783									
All subjects	0.014/0.025	3921 (97.2)	114 (2.8)	5022 (95.0)	262 (4.9)	0.55 (0.44–0.69)	3.30E − 07	0.61 (0.47–0.79)	1.52E − 04

[a]Allele frequencies of the C allele for rs11892031 and the T allele for rs17863783.
[b]Estimates from logistic regression under a dominant protective model adjusted for age, gender, study sites and smoking status when applicable.
[c]Estimates from logistic regression under a dominant protective model adjusted for age, gender, study sites, smoking status, with mutual adjustment for rs11892031/rs17863783.
[d]rs11892031 and rs17863783 are separated by 36 994 bp and are in LD ($D' = 0.961$, $r^2 = 0.228$), based on all 9319 study subjects

ensure correct genotyping of this uncommon variant, we cross-validated genotypes in a subset of samples by three methods, Illumina chip, Sanger sequencing and TaqMan genotyping (Supplementary Material, Fig. S2 and Table S5). Association for rs17863783 ($P = 3.3 \times 10^{-7}$; OR = 0.55, 95%CI = 0.44–0.69) was stronger than for the original GWAS marker, rs11892031 ($P = 7.7 \times 10^{-5}$; OR = 0.79, 95%CI = 0.70–0.89) (Table 1 and Supplementary Material, Table S6). Both these SNPs are uncommon variants with frequencies of minor protective alleles in controls of 8.5 and 2.5% for rs11892031 and rs17863783, respectively. There is only moderate LD between these SNPs, $D' = 0.961$ and $r^2 = 0.228$ in the combined GWAS set. To further evaluate whether these SNPs represent the same association signal, we performed a conditional analysis adjusting for the effect of the other variant. Adjustment for rs11892031 attenuated the signal for rs17863783 ($P = 1.52 \times 10^{-4}$; OR = 0.61, 95%CI = 0.47–0.79 after adjustment, Table 1, Fig. 1C), while the loss of signal for rs118920231 after adjustment for rs17863783 ($P = 8.32 \times 10^{-2}$, OR = 0.89, 95%CI = 0.78–1.02 after adjustment, Table 1, Fig. 1D) suggests that these two variants represent the same association. There was no evidence of additional association signal within the *UGT1A* region after adjustment for rs17863783 (Fig. 1D). We also analyzed haplotypes constructed with rs11892031 and 18 selected coding SNPs that represent all the 46 coding SNPs in this region. The protective T allele of rs17863783 was found only on a haplotype with the C allele of rs11892031 and only this haplotype showed a significant protective effect. No association was detected for a haplotype with the C allele of rs11892031 but without the T allele of rs17863783, or any other haplotype (Table 2). Our results suggest that rs17863783, or other variants in strong LD with it, could explain the genetic association initially captured by rs11892031. The protective effect of rs17863783 was stronger among smokers (OR = 0.51; 95%CI = 0.40–0.66, $P = 3.3 \times 10^{-7}$) compared with non-smokers (OR = 0.72, 95%CI = 0.43–1.19, $P = 0.2$), but the interaction between rs17863783 and smoking status was not statistically significant (Table 3). This might be due to low allele frequency of rs17863783, the predominance of smokers among bladder cancer cases, and other causes of bladder cancer in non-smokers. A genetic variant rs1495741 within the *NAT2* gene has previously been associated with bladder cancer and slow acetylation of aromatic amines by the NAT2 enzyme (13). In our samples, the association for rs17863783 was similar in individuals with rapid/intermediate and slow acetylation, classified by rs1495741 genotypes of *NAT2,* and this effect was not modified by smoking status (Supplementary Material, Table S7).

The molecular phenotype of the genetic association: increased expression of the functional splicing form, *UGT1A6.1*

UGT1A6 has two splicing mRNA isoforms, *UGT1A6.1* and *UGT1A6.2*. The bladder cancer-associated rs17863783 is a synonymous variant (Val209Val) located within the long isoform (*UGT1A6.1*, NM_001072) that encodes a full-length protein of 532 amino acids. The short form (*UGT1A6.2*, NM_205862) encodes a protein of 265 amino acids, which is missing a substantial portion of the highly conserved substrate-binding domain, fully encoded by the first exon (Supplementary Material, Fig. S3). UGT1A6 protein expression usually refers to UGT1A6.1 in the literature, because UGT1A6.2 lacks most of the exon 1 and is unlikely to be recognized by antibodies. *UGT1A6* mRNA expression can refer to both *UGT1A6.1* and *UGT1A6.2* splicing forms, depending on the specific method of detection.

We considered the exonic rs17863783 to be the strongest functional candidate from the associated block of 28 linked SNPs, and performed functional evaluation of this variant. Even though synonymous amino acid substitutions do not directly cause protein changes, they may influence disease risk by altering exonic splicing enhancers (ESEs) that bind splicing factors, regulate inclusion of exons or modify expression levels of specific transcripts, without affecting splicing sites (22). Using ESE finder 3.0 software (22), we predicted a differential interaction between rs17863783 alleles and splicing factors (Supplementary Material, Fig. S4). To experimentally evaluate the effect of rs17863783 on splicing and expression of *UGT1A6* transcripts, we created allelic exontrap splicing minigenes that included 2.3 kb genomic fragments surrounding rs17863783 and both alternative first exons of *UGT1A6*. After transient transfection into HeLa (cervical cancer), 293T (normal embryonic kidney), J82 (bladder cancer) and HepG2 (liver cancer) human cell lines, the transcripts produced by the minigenes were analyzed for quantitative mRNA expression of both isoforms. In all cell lines tested,

Table 2. Haplotype analysis of 18 coding SNPs and GWAS signal rs11892031 in the *UGT1A* region among SPBC subjects (n = 2017) and all stage 1 GWAS samples ($n = 8652$)

No	Haplotype Marker order[a]	Frequencies Cases	Controls	Df	OR[b]	*P*-value[b]	OR[c]	*P*-value[c]
In SPBC samples, all genotyped ($n = 2017$)								
—	Omnibus	—	—	10	—	2.26E − 01	—	8.01E − 02
1	CGATAGCGGCGCTGCCCAC	0.0265	0.0257	1	1.04	8.70E − 01	1.16	5.06E − 01
2	CGATAGCGGCGCTGCCCAT	0.2779	0.2610	1	1.09	2.38E − 01	1.12	1.35E − 01
3	CGATAGCGGCGCTGCTTAT	0.0258	0.0208	1	1.27	3.17E − 01	1.32	2.52E − 01
4	CGATAGTTAAGCGGCCTAT	0.0422	0.0386	1	1.11	5.60E − 01	1.14	4.69E − 01
5	CGATAGTTAAGCTGCTTAT	0.1290	0.1371	1	0.93	5.00E − 01	0.90	3.18E − 01
6	CGATATTTAAGCTGCTTAT	0.2310	0.2140	1	1.12	1.83E − 01	1.12	2.07E − 01
7	CGCTAGCGAAGATGCCCAT	0.0123	0.0123	1	1.01	9.88E − 01	1.12	7.26E − 01
8	CGCTAGCGAAGATGCCCGT	0.0335	0.0435	1	0.76	1.39E − 01	0.70	5.67E − 02
9	CGCTAGTGACTCTGCTTAT	0.0120	0.0229	1	0.51	1.68E − 02	0.51	1.91E − 02
10	GGATAGCGGCGCTGCCCAT	0.0135	0.0120	1	1.19	6.29E − 01	1.33	4.54E − 01
11	GGATATTTAAGCTGCTTAT	0.1964	0.2121	1	0.91	2.89E − 01	0.92	3.29E − 01
In stage1 GWAS samples, genotyped and imputed ($n = 8652$)								
—	Omnibus	—	—	12	—	5.43E − 04	—	2.60E − 04
1	CGATAGCGGCGCTGCCCAC	0.0195	0.0209	1	0.93	5.63E − 01	0.95	6.80E − 01
2	CGATAGCGGCGCTGCCCAT	0.3159	0.3110	1	1.03	4.40E − 01	1.04	3.09E − 01
3	CGATAGCGGCGCTGCTTAT	0.0201	0.0212	1	0.94	6.20E − 01	1.04	7.80E − 01
4	CGATAGTTAAGCGGCCTAT	0.0381	0.0325	1	1.21	4.49E − 02	1.19	7.61E − 02
5	CGATAGTTAAGCTGCTTAT	0.1347	0.1381	1	0.97	5.88E − 01	0.97	5.87E − 01
6	CGATATTTAAGCTGCTTAT	0.1933	0.1873	1	1.04	2.96E − 01	1.05	2.42E − 01
7	CGCTAGCGAAGATGCCCAT	0.0115	0.0140	1	0.72	8.44E − 02	0.66	3.49E − 02
9	CGCTAGTGACTCTGCTTAT	0.0073	0.0144	1	0.48	3.16E − 05	0.47	2.88E − 05
11	GGATATTTAAGCTGCTTAT	0.1926	0.1926	1	1.00	9.26E − 01	1.00	9.84E − 01
12	GGACAGTTAAGCGGCTTAT	0.0125	0.0162	1	0.74	4.62E − 02	0.75	6.30E − 02
13	GGATAGTTAAGCGGCCTAT	0.0201	0.0167	1	1.27	7.83E − 02	1.31	5.38E − 02
14	GGATAGTTAAGCGGCTTAT	0.0155	0.0136	1	1.19	2.51E − 01	1.20	2.32E − 01
15	CACTAGCGAAGATGCCCGT	0.0190	0.0216	1	0.83	1.89E − 01	0.79	9.44E − 02

[a]Haplotypes were constructed with the following SNP order: rs1042597|rs17863762|rs11892031|rs72551330|rs56385016|rs17868323|rs11692021|rs6759892|rs2070959|rs1105879|rs17863783|rs6755571|rs2011425|rs45510694|rs45621441|rs3821242|rs6431625|rs17868336|rs45449995. Only haplotypes at >1% frequency in cases or controls were included into analysis.
[b]Estimates from haplotype-specific logistic regression for each haplotype versus all other haplotypes together, and a single omnibus test jointly estimating overall haplotype effect, without adjustment for covariates.
[c]Estimates from haplotype-specific logistic regression analysis, for each haplotype versus all other haplotypes together. A single omnibus test jointly estimating overall haplotype effect was performed, adjusted for age, gender, study sites and smoking status when applicable.

the presence of the protective T allele significantly increased the expression of the *UGT1A6.1* compared with minigenes with the risk G allele. Expression of the *UGT1A6.2* was not affected by rs17863783 alleles (Fig. 2A and B). These minigenes did not include any of other 27 variants in high LD with rs17863783, indicating that the functional effect could be attributed to rs17863783 alone. While this does not exclude the possibility of some other functional variants in this region, our results showed that rs17863783 has critical impact on the function of UGT1A6.1, mechanisms of cellular detoxification and susceptibility to bladder cancer. The UGT1A6.1 protein is primarily expressed in the liver, kidney and bladder tissue (Fig. 3A), in agreement with mRNA expression we detected in a panel of human tissues and cell lines (Fig. 3B, Supplementary Material, Table S8). Expression of both splicing forms, *UGT1A6.1* and *UGT1A6.2*, was similar between normal and tumor bladder samples, suggesting that the functional effect of this gene is not disease specific (Supplementary Material, Fig. S4). In normal human liver samples, *UGT1A6.1* expression was increased 4-fold in carriers of the uncommon protective T allele of rs17863783 ($P = 0.0136$, $n = 88$, Fig. 3C), while no carriers of the uncommon T allele of rs17863783 were found among 44 normal bladder tissue samples available for expression analysis.

The *UGT1A* region and pharmacogenetics of irinotecan toxicity

The *UGT1A* locus is well known for its genetic association with severe toxicity to an anti-cancer drug irinotecan (23,24). Genotyping of the marker UGT1A1*28 (rs8175347), a $(TA)_{5-7}$ repeat within the *UGT1A1* promoter region, is now required by the US Food and Drug Administration (FDA) for adjustment of drug dosage and prevention of irinotecan toxicity in susceptible individuals (25). It is reasonable to hypothesize that genetic variants associated with detoxification of irinotecan may be associated with detoxification of environmental carcinogens, and susceptibility to bladder cancer. There were multiple attempts to identify other markers in this region that could provide similar genetic information and would be easier to genotype than UGT1A1*28 (26–28). Therefore, we used our unique set of 2017 individuals of European descent with complete information for 1170 genetic markers in this region to search for markers in high LD with UGT1A1*28. Four intronic/promoter markers were in a

Table 3. Genetic association results for *UGT1A* SNPs and bladder cancer risk in relation to smoking status

	MAF[a] Case/control	Cases, $n = 4035$ (n, %) *AA*	*AC + CC*	Controls, $n = 5284$ (n, %) *AA*	*AC + CC*	OR[b] (95%CI)	*P*-value[b]	OR[c] (95%CI)	*P*-value[c]	*P*-value[d]
rs11892031										
All subjects	0.069/0.085	3497 (86.6)	538 (13.3)	4424 (83.7)	860 (16.3)	0.79 (0.70–0.89)	7.75E − 05	0.89 (0.78–1.02)	8.32E − 02	—
Never smoker	0.070/0.085	601 (86.5)	94 (13.5)	1214 (83.6)	239 (16.4)	0.81 (0.62–1.05)	1.09E − 01	0.84 (0.63–1.13)	2.52E − 01	Ref.
Ever smoker	0.069/0.085	2886 (86.7)	443 (13.3)	3205 (83.9)	617 (16.1)	0.78 (0.68–0.90)	4.06E − 04	0.91 (0.78–1.06)	2.20E − 01	9.26E − 01
Former smoker	0.067/0.086	1589 (86.9)	240 (13.1)	1699 (83.6)	333 (16.4)	0.78 (0.65–0.94)	7.43E − 03	0.89 (0.73–1.09)	2.63E − 01	8.44E − 01
Current smoker	0.070/0.083	1297 (86.5)	203 (13.5)	1506 (84.1)	284 (15.9)	0.74 (0.60–0.91)	4.98E − 03	0.88 (0.69–1.12)	2.87E − 01	8.81E − 01
	Case/control	*GG*	*GT + TT*	*GG*	*GT + TT*					
rs17863783										
All subjects	0.014/0.025	3921 (97.2)	114 (2.8)	5022 (95.0)	262 (4.9)	0.55 (0.44–0.69)	3.30E − 07	0.61 (0.47–0.79)	1.52E − 04	—
Never smoker	0.015/0.020	674 (97.0)	21 (3.0)	1392 (95.8)	61 (4.2)	0.72 (0.43–1.19)	2.01E − 01	0.83 (0.47–1.46)	5.18E − 01	Ref.
Ever smoker	0.014/0.027	3236 (97.2)	93 (2.8)	3621 (94.7)	201 (5.3)	0.51 (0.40–0.66)	3.30E − 07	0.56 (0.42–0.74)	6.47E − 05	3.07E − 01
Former smoker	0.013/0.024	1782 (97.4)	47 (2.6)	1935 (95.2)	97 (4.8)	0.51 (0.36–0.73)	2.36E − 04	0.56 (0.38–0.84)	4.51E − 03	3.45E − 01
Current smoker	0.016/0.030	1454 (96.9)	46 (3.1)	1686 (94.2)	104 (5.8)	0.50 (0.35–0.73)	2.48E − 04	0.56 (0.37–0.86)	7.38E − 03	3.55E − 01

[a]Allele frequencies of the C allele for rs11892031 and the T allele for rs17863783.
[b]Estimates from logistic regression under a dominant protective model adjusted for age, gender, study sites and smoking status when applicable.
[c]Estimates from logistic regression under a dominant protective model adjusted for age, gender, study sites, smoking status and with mutual adjustment for rs11892031/rs17863783 when applicable.
[d]*P*-value of gene–smoking interaction was estimated from logistic regression under a dominant protective model adjusted for age, gender and study sites.
[e]rs11892031 and rs17863783 are separated by 36 994 bp and are in LD ($D' = 0.961$, $r^2 = 0.228$), based on all 9 319 study subjects.

similarly high LD with UGT1A1*28 ($r^2 = 0.875$). Of these markers, rs6742078 and rs887829 have been reported to be strongly associated with blood bilirubin levels ($P < 10^{-324}$ and $P < 10^{-69}$) (29,30), but we observed no association for these markers and UGT1A1*28 with bladder cancer in our samples (Supplementary Material, Table S9). Interestingly, of 46 coding variants we identified in this region, only 3 variants were in a relatively high LD with UGT1A1*28 ($0.63 < r^2 < 0.67$). All three variants were from the *UGT1A6* gene (rs1105880, Leu105Leu; rs2070959, Thr181Ala; rs11058879, Arg184Ser) and located in the vicinity of our bladder cancer-associated SNP rs17863783 (Ala209Ala), suggesting the functional relevance of UGT1A6.1 for different phenotypes. In fact, according to the pharmacogenomics knowledge database (http://www.pharmgkb.org), UGT1A6.1 metabolizes multiple drugs, including irinotecan, analgetics paracetomol (tylenol), aspirin and naproxen and an anti-convulsant drug phenytoin.

DISCUSSION

UGT1A6.1 and bladder cancer

In the present study, we report the identification of SNP rs17863783 within a cellular detoxification gene, *UGT1A6,* as a protective factor from bladder cancer. Exposure to aromatic amines found in industrial chemicals and tobacco smoke is strongly associated with increased risk of bladder cancer (7). UGTs conjugate UDP-glucuronic acid with N-hydroxylated products of diverse substrates, including aromatic amines (31). The conjugated water-soluble glucuronides can then be excreted via stool and urine (9). Until excretion, the urine is stored in the bladder where it comes in direct contact with bladder epithelium. Urine acidity, which depends on diet, body composition and medications (32–34), is a critical factor that determines the stability of glucuronides. At a low urine pH (< 6.0), glucuronides become unstable and quickly dissociate to release N-hydroxylated oncogenic forms of aromatic amines (35), form DNA adducts and initiate carcinogenesis within the bladder epithelium (36). However, the UGT proteins endogenously expressed in bladder epithelium have the ability to conjugate different substrates (37). Our genetic study suggested that of all *UGT* genes, only *UGT1A6.1* showed genetic association with protection from bladder cancer. Furthermore, the UGT1A6.1 functional protein isoform is strongly expressed in human bladder epithelium (38,39) (Fig. 3A and B, Supplementary Material, Table S8), and conjugates chemicals known to be of risk for bladder cancer (31) (Supplementary Material, Table S10). This suggests that even when the bladder epithelium is exposed to the reactive N-hydroxylated products of aromatic amines generated by dissociation of urine glucuronides, endogenously expressed UGT1A6.1 can reconjugate and remove these intermediates from bladder epithelium, thereby preventing carcinogenesis (Supplementary Material, Fig. S6). By increasing *UGT1A6.1* mRNA expression, the T allele of rs17863783 may help remove carcinogens from bladder epithelium and therefore protect from bladder cancer. Based on the functional role, this variant might be protective only in individuals exposed to particular environmental factors, such as tobacco smoke or chemicals, while remaining neutral in all other situations.

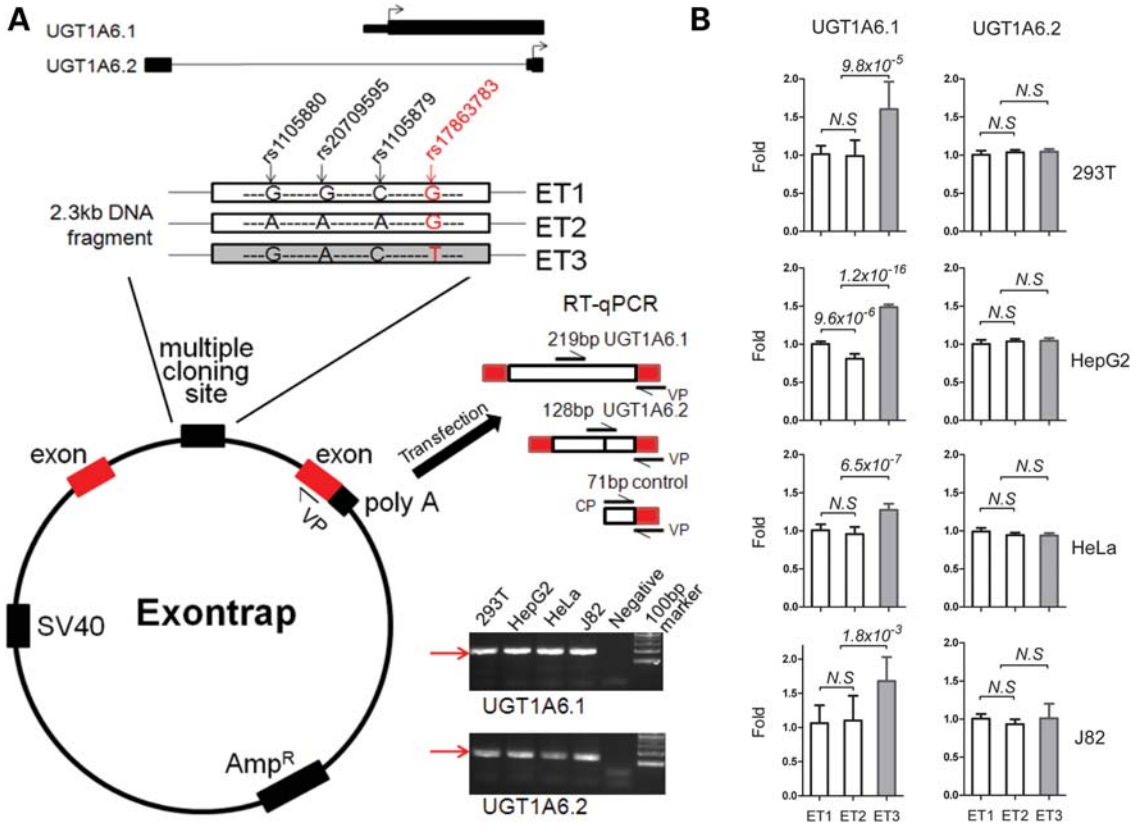

Figure 2. Exontrap experiment for evaluation of allelic effects of the synonymous exonic rs17863783 (Val209Val). (**A**) Genomic structure of the 2.3 kb DNA fragment containing alternative first exons of two splicing forms of *UGT1A6*. Exons are shown as black rectangles and translation start sites as arrows. The first exon of *UGT1A6.1* encodes the entire substrate-binding domain of 287 amino acids, while this domain of *UGT1A6.2* is truncated to 20 amino acids. Sequencing of the 2.3 kb DNA fragment in 90 CEU HapMap samples identified four coding SNPs, rs1105880, rs20709595, rs1105879 and rs17863783, present in three haplotypes with frequencies of 0.265, 0.696 and 0.039. Exontrap minigenes ET1, ET2 and ET3 were constructed to represent each of these haplotypes. The minigenes were transiently transfected into 293T (normal embryonal kidney), HepG2 (liver cancer), HeLa (cervix cancer) and J82 (bladder cancer) cell lines, in 12 biological replicates for each of the constructs and cell lines. The cells were collected 48 h post-transfection, total RNA was extracted and converted into cDNA with a vector-specific primer (VP). For each of the samples, three expression assays were measured by the RT–qPCR assays with SYBR Green detection, in four technical replicates each. All assays specifically detect only RNA generated by exontrap minigenes, but not endogenous *UGT1A6* transcripts. Expression of the *UGT1A6.1* and *UGT1A6.2* splicing forms was normalized by a common assay (primers CP + VP). PCR fragments for the three expression assays were visualized on agarose gel and correct products of 219, 128 and 71 bp were detected. The identity of the PCR fragments was also confirmed by sequencing. (**B**) RT–qPCR expression of *UGT1A6.1* and *UGT1A6.2* splicing forms after transfection of cell lines with minigines ET1, ET2 and ET3, all compared with the ET1 construct. Error bars indicate standard error of the mean with 95% confidence intervals based on 12 individual, independent transfections performed for each of the constructs and cell lines. Difference in expression is indicated as NS (not significant) or with *P*-values for the unpaired two-sided *t*-test.

Rs17863783 and refinement of the GWAS signal

By design, GWAS have been conducted to discover common variants, with MAF > 10%, associated with complex diseases (40), and indeed, most signals detected by cancer GWAS, are loci with SNP markers with MAF > 20% (41). This design strategy is predicated on the 'common disease-common variant' theory postulating that complex traits are caused by combinations of many common alleles with small individual effects (42–44). Compared with common variants, uncommon/rare variants are technically more difficult to genotype with the same level of confidence and completion, partly due to technical issues related to confidence of detection of rare alleles and the necessity of extensive validation studies. Statistical analysis of uncommon/rare variants is also more challenging due to lower power and possible effects of random confounding factors (40,45). As a result, commercial genotyping arrays used in GWAS studies are biased towards variants with MAF > 10% and have a poor representation of variants with MAF < 5% (46), or these latter variants are excluded from the analysis. Among 366 GWAS that reported significant association signals ($P < 10^{-7}$), 275 studies reported association for variants with MAF > 5%, and only 28 GWAS reported 40 SNPs with MAF < 5% (47). The proportion of genetic variation explained in common diseases still appears to be relatively modest (48), in spite of thousands of common variants identified by GWAS (49). Different disease hypotheses have been discussed, and it is now suggested that both common and uncommon/rare variants significantly contribute to genetic susceptibility of common diseases (50–56).

In the original bladder cancer GWAS that analyzed SNPs with MAF > 5%, a common variant at 2q37.1 was reported (17), but due to the standard quality control metrics, the study did not evaluate the uncommon rs17863783 (MAF =

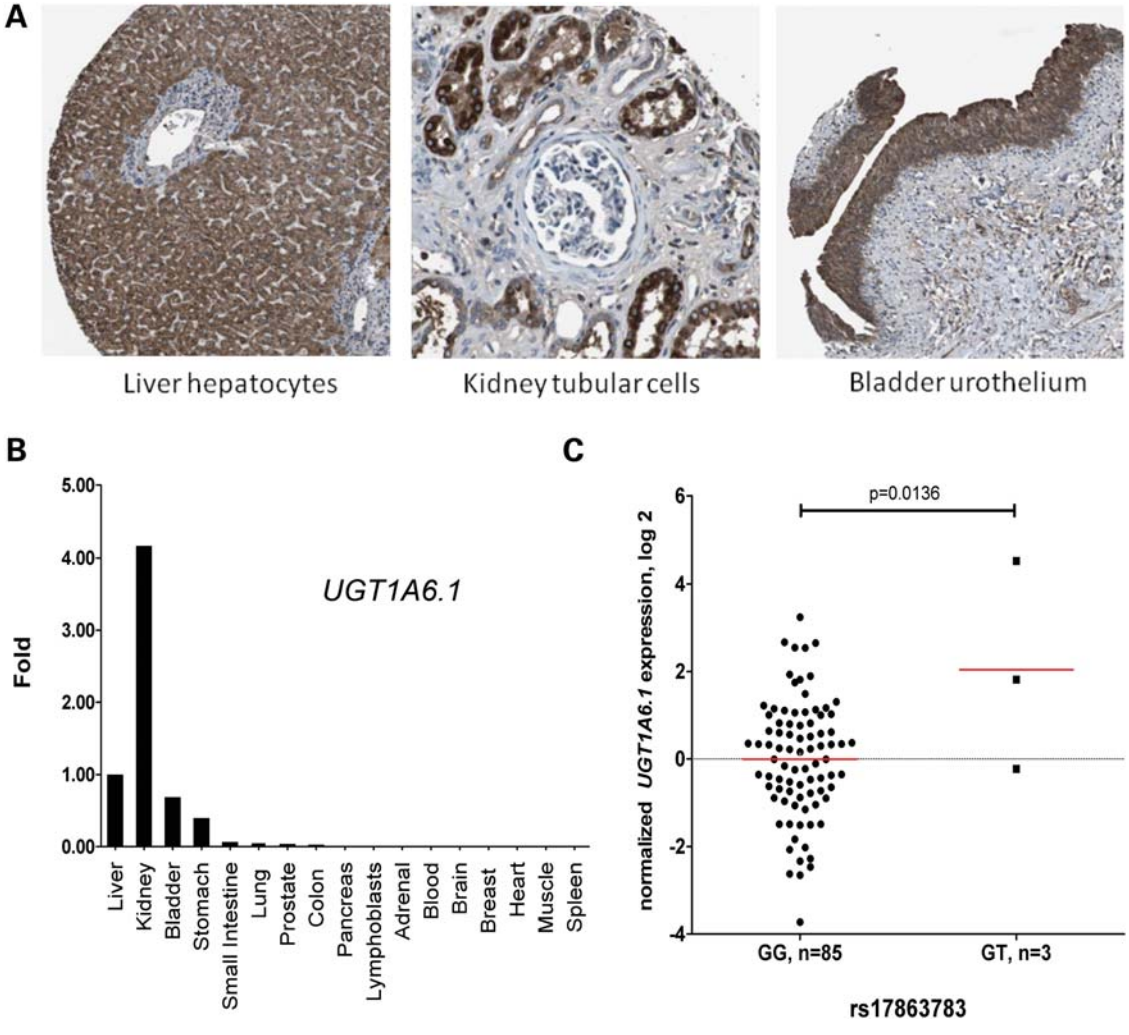

Figure 3. Expression of the UGT1A6 protein and *UGT1A6.1* splicing form in normal human tissues. (**A**) UGT1A6 protein expression in normal human tissues. Tissue microarray analysis in normal human tissues detects UGT1A6 protein expression (as depicted by brown staining) in liver hepatocytes, kidney tubular cells and bladder epithelium. The images and annotations are courtesy of the Human Protein Atlas project (http://www.proteinatlas.org/search/UGT1A6). (**B**) mRNA expression in a panel of normal human tissues. Expression values were normalized by two endogenous controls, beta-2-microglobulin (*B2M*) and cyclophilin (*PPIA*). Expression values are presented on log2 scale in relation to the the expression level in the liver. (**C**) mRNA expression in 88 normal liver tissue samples from healthy controls. Expression values of the total set passed the normally test and were analyzed with the unpaired two-sided *t*-test. The results are presented on the log2 scale in relation to the mean of the GG group.

2.5%), which we now identified to be responsible for the association originally detected by a more common SNP rs118920231 (MAF = 8.5%). This might be considered 'synthetic' association (53,57,58), because a more common variant rs11892031 captures the signal of an uncommon linked SNP rs17863783 ($D' = 0.96$). However, the less common rs1786383 falls on the backbone that contains the rs11892031 alleles ($r^2 = 0.228$), resulting in the detection of the association signal. It is postulated that in the case of a 'synthetic' association, the association signal should become stronger when the right variant is interrogated (53). In fact, we detected stronger association for the less common variant rs17863783, and it could explain the original association for rs11892031, but not vice versa (Table 1). Our unbiased search through all variants in this region, not limited by variants in high LD ($r^2 > 0.8$) with rs118920231, has been instrumental in identification of a probable causal variant, rs17863783. Our GWAS identified the *UGT1A* region for bladder cancer susceptibility, but the fine-mapping has identified a variant that explained and strengthened the original genetic association and provided a plausible functional mechanism for its effect. The risk G allele of rs17863783 is conserved in 33 of 41 species (Supplementary Material, Fig. S7), while the protective T allele is a derived allele found only in a small percentage of humans, 4.9% of controls and 2.8% of bladder cancer cases. The protective T allele is clearly functional, as it is associated with increased mRNA expression of *UGT1A6.1*. A recent study concluded that rarer derived variants, with MAF < 8–10%, are more likely to be functional than the more common variants (59). This can be explained by the likely deleterious selective pressure on the derived risk alleles that keep them at low allele frequencies. Here, the functional derived T allele of rs17863783 is a protective allele. It is possible that the newly derived protective

variants in detoxification genes, such as UGTs, may be favored by positive selection in modern environment, substantially altered by humans. Low frequencies of these alleles may be a reflection of the short evolution period after introduction of tobacco smoking and industrial chemicals into human environment. This can also indicate that the human-specific environmental factors, such as chemicals, drugs and dietary components, might have weak deleterious effects that result in minor positive selection pressure on genetic variants that regulate metabolism of these substrates. By expanding our analysis to the broader *UGT1A* region, we tested and excluded the possibility that the same genetic variants underlie mechanisms responsible for bladder cancer susceptibility and detoxification of anti-cancer drug irinotecan.

In conclusion, we performed a detailed fine-mapping analysis of the *UGT1A* locus reported in our recent bladder cancer GWAS, identified an uncommon protective functional genetic variant, rs17863783, that greatly accounted for the initial GWAS signal, and provided the first link to the underlying molecular phenotype of this association. Although we provide compelling genetic and functional evidence for rs17863783, this does not exclude the possibility of existence of other functionally important variants in this region. The combination of common, uncommon and rare variants will eventually extend our understanding of human disease and begin to map the genomic architecture of a complex disease, such as bladder cancer. Furthermore, understanding the impact of environmental exposures should be instrumental in the functional interpretation of genetic associations identified by GWAS.

MATERIALS AND METHODS

Study subjects

Stage 1 GWAS bladder cancer cases and controls of European descent were drawn from five studies in the USA and Europe, as previously described (17): SBCS (1106 cases/1050 controls), Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO, 708 cases/1874 controls), The American Cancer Society Cancer Prevention Study II Nutrition Cohort (CPS-II, 687 cases/730 controls), New England Bladder Cancer Study (NEBCS-ME,VT, 630 cases/759 controls) and Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC, 401 cases/707 controls). Additional GWAS follow-up samples were drawn from: Health Professionals Follow-up Study (HPFS, 113 cases/115 controls), New England Bladder Cancer Study (NEBCS-NH, 355 cases/374 controls) and Nurse's Health Study (NHS, 63 cases/57 controls). HapMap DNA samples from 30 European trios (CEU) used for sequencing and genotyping were purchased from the Corriell Institute for Medical Research (Camden, NJ, USA). As previously described (17), each participating study obtained informed consent from study participants and approval from its respective Institutional Review Board for this study. For stage 1 only, participating studies obtained institutional certification permitting data sharing in accordance with the NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS).

Tissue samples and cell lines for functional studies

Paired (normal/tumor) bladder tissue samples from 44 anonymous bladder cancer patients were purchased from Asterand (Detroit, MI, USA) under exemption #4715 by the NIH Office of Human Subject Research. Previously described liver samples (60) were provided by the University of Minnesota. DNA from normal tissue samples was prepared with Gentra kit (Qiagen) and used for sequencing and genotyping. Samples of total RNA from 17 non-cancerous human tissues (skeletal muscle, spleen, adrenal gland, kidney, brain, pancreas, heart, small intestine, stomach, bladder, colon, prostate, liver, lung and breast) were purchased from Clontech (Mountain View, CA, USA) or BioChain (Hayward, CA, USA). Samples of total RNA from the NCI-60 set of cell lines (61) were provided by the Molecular Targets Team, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis (DCTD/NCI/NIH). All other cell lines were purchased from the American Type Culture Collection (ATCC) and were maintained according to the recommended conditions. For each sample, 1–2 μg of DNAase-treated total RNA was converted into cDNA with random hexamers and SuperScript III reverse transcriptase (Invitrogen). cDNA samples were diluted with nuclease-free water and 5 ng of total RNA was used for each quantitative reverse transcriptase PCR (qRT–PCR).

Sequencing and genotyping of the *UGT1A* region

Long-range amplicons of ~1.3 kb covering each of the *UGT1A* exons and flanking intronic sequences were generated with specific primers and conditions (Supplementary Material, Table S11). PCR fragments were confirmed by agarose gel, and sequenced with 3730xl DNA Analyzer (Applied Biosystems). Sequence analysis was performed with Sequencher 4.2 software (Gene Code, MI, USA) and all genetic variants were scored manually by two people, independently. The DNA samples from cases and controls were mixed on genotyping plates, and the sample status was blinded to the laboratory investigators. Although rs17863783 was present on the Illumina chip, the genotyping was incomplete (~75%). For this study, we genotyped the marker in all samples in stage 1 GWAS plus 1077 additional samples from three of the follow-up sets (HPFS, NEBCS-NH and NHS) (17). The default genotyping method for this marker was by a TaqMan allelic discrimination assay, in 384-well format. For 5 μl reactions we used 5 ng DNA, 2× genotyping buffer and a genotyping assay C__25972736_20 (all from Applied Biosystems), according to the instructions. To ensure correct genotype clustering and scoring for rs17863783, each genotyping plate contained control samples with known genotypes, NA19194 (T/T) and NA19116 (T/T) from the HapMap YRI panel. The TaqMan genotyping results were validated by two other platforms (Illumina chip and Sanger sequencing). A concordance rate of 99.2–100% confirmed the high quality of genotyping by the three methods (Supplementary Material, Fig. S2 and Table S5). Four additional SNPs were genotyped by Illumina chip and confirmed by sequencing of ~2000 samples and used as additional controls for genotyping concordance (Supplementary Material, Fig. S2 and Table S6).

Imputation

We used IMPUTE2 software (62) to estimate genotypes of SNPs not directly genotyped in the *UGT1A* region. Genotypes of 166 SNPs from this region (chr2:243,091,000–234,447,000) have been generated by the stage 1 bladder cancer GWAS in 3461 cases and 4694 controls (17). We imputed 1004 additional SNPs in this region for the entire stage 1 GWAS samples using a combined set of reference panels: 1000 Genomes Project [June 2010 release (21)], HapMap Phase 3 CEU [second February 2009 release (20)] and a subset of the stage 1 GWAS samples (SBCS, $n = 2{,}017$) in which 18 exonic SNPs were completely genotyped by sequencing. We evaluated the imputation performance using the average posterior probability for the best-guessed genotypes, and the IMPUTE2-info score, which is associated with the imputed allele frequency estimate ranging from 1 to 0 (high to low confidence). Markers with posterior probability <0.9 or IMPUTE-info score <0.9 were excluded from the association analysis.

Statistical analysis

Fisher's exact tests of the Hardy–Weinberg equilibrium (HWE) for controls and for the entire set were conducted for all markers. There was only one marker showed significant deviation from HWE ($P < 0.001$), and it was flagged but retained in the analysis. LD measures (D' and r^2) were estimated using Haploview (63). GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html) was used to combine all the imputed variants (with >90% imputation certainty) and actual genotyping data. Association with bladder cancer risk was tested under a dominant protective model (one risk allele is sufficient for protective effect) using PLINK (64) and SAS/STAT system version 9.2 (SAS Institute Inc., Cary, NC, USA), with the adjustment for age (in 5-year categories), gender, study sites and smoking habit (current, former or never). In the original bladder cancer GWAS (17), it was found that study sites best approximate eigenvalue of principle component analysis to control for population stratification. Thus, we used study sites for similar adjustment in our analyses. To test for the presence of independent association signals for bladder cancer risk in the 2q37.1 region, we conditioned on the original GWAS signal (rs11892031) in a logistic regression model for the additive effect, with adjustment for the same covariates. Genotype–smoking interactions were assessed by stratifying individuals as current, former, ever or never smokers for association testing, as well as adjusted for the same covariates in the logistic regression models, including other interaction terms. Genotyping data of SNP rs1495741 in the *NAT2* gene were retrieved from the original GWAS (17) to stratify individuals as rapid/intermediate (rs1495741 AG/GG) and slow (rs1495741 AA) acetylators. *NAT2–UGT1A* interactions were tested in a logistic regression model with the adjustment for the same covariates along with interaction terms. Haplotype-specific odds ratios and *P*-values were estimated using PLINK (64) for each haplotype (>1%) versus all other haplotypes together, as well as a single omnibus test jointly estimating overall haplotype effects.

mRNA expression analysis

Expression of *UGT1A6* mRNA in human tissues and cell lines was measured with TaqMan expression assays Hs01592477_m1 for *UGT1A6.1* (NM_001072.3) and Hs01651483_m1 for *UGT1A6.2* (NM_205862. 1). Endogenous controls Beta-2-microglobulin (*B2M,* assay Hs00187842_m1) and Cyclophilin (*PPIA,* assay 4326316E) were used for normalization of expression. For all assays, reactions with water and 10 ng of genomic DNA from pooled HapMap samples were used as negative controls. The expression detection was performed on the ABI PRISM 7900HT SDS (Applied Biosystems) with cDNA prepared from 5 ng of total RNA, 0.25 μl of 20× TaqMan gene expression assays or 2.5 μl of 2× Gene Expression Master Mix in 5 μl reaction volume. The expression was measured in four technical replicates and average values were used for the analysis.

ESE prediction

Screening for ESEs (http://rulai.cshl.edu/cgibin/tools/ESE3) was performed with a web-based bioinformatic tool using a 50 bp DNA sequence with alleles T and G of rs17863783.

Exontrap splicing assays

A 2.3 kb genomic DNA fragment surrounding rs17863783 and containing alternative first exons of *UGT1A6.1* and *UGT1A6.2* was generated with specific primers (Supplementary Material, Table S11) in 60 HapMap individuals from a European population (CEU). Sequencing of these fragments detected four exonic SNPs in three haplotypes. The PCR products representing the haplotypes were cloned into an Exontrap vector (MoBiTec, Gottingen, Germany), using *Xho*I and *Bam*HI restriction sites. After validation by sequencing, the constructs were transfected into 293T, HeLa, J82 and HepG2 cell lines. Transfections were performed with LTX and PLUS transfection reagents (Invitrogen) for HeLa, J82 and HepG2 and Lipofectamine 2000 transfection reagent (Invitrogen) for 293T cell lines, in 12 biological replicates for each of the cell lines and constructs. The cells were seeded in a 96-well plate at a cell density of 1×10^5, transfected next day with 200 ng of constructs and harvested 48 h post-transfection. Total RNA was extracted with QIACube with RNAeasy protocol combined with DNAse treatment (Qiagen). For each sample, 0.5–1 μg of total RNA was converted into cDNA with SuperScript III reverse transcriptase (Invitrogen) using a vector-specific primer (Supplementary Material, Table S11). cDNA samples were diluted with nuclease-free water and 10–20 ng of total RNA was used for each quantitative SYBR Green qRT–PCR. Three assays were measured for each of the samples—a common assay and two assays for specific splicing forms (Supplementary Material, Table S11). All expression assays were designed to uniquely quantify transcripts generated *in vitro* during the Exontrap experiment, but not endogenous *UGT1A6* transcripts.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *HMG* online.

ACKNOWLEDGEMENTS

The NCI bladder cancer GWAS and follow-up studies are supported by the intramural research program of the National Institutes of Health, National Cancer Institute.

Following individuals are acknowledged for their support: Francisco Real (Molecular Pathology Programme, Centro Nacional de Investigaciones Oncológicas, Madrid, Spain). Marie-Joseph Horner (DCEG, NCI/NIH, Rockville, MD, USA). Adam Mumy (DCEG, NCI/NIH, Rockville, MD, USA). Natalia Orduz (DCEG, NCI/NIH, Rockville, MD, USA). Leslie Carroll (Information Management Services, Silver Spring, MD, USA). Gemma Castaño-Vinyals (Institut Municipal d'Investigació Mèdica, Barcelona, Spain). Fernando Fernández (Institut Municipal d'Investigació Mèdica, Barcelona, Spain). Paul Hurwitz (Westat, Inc., Rockville, MD, USA). Charles Lawrence (Westat, Inc., Rockville, MD, USA). Marta Lopez-Brea (Marqués de Valdecilla University Hospital, Santander, Cantabria, Spain). Anna McIntosh (Westat, Inc., Rockville, MD, USA). Angeles Panadero (Hospital Ciudad de Coria, Coria (Cáceres), Spain). Fernando Rivera (Marqués de Valdecilla University Hospital, Santander, Cantabria, Spain). Robert Saal (Westat, Rockville, MD, USA). Maria Sala (Institut Municipal d'Investigació Mèdica, Barcelona, Spain). Kirk Snyder (Information Management Services, Inc., Silver Spring, MD, USA). Anne Taylor (Information Management Services, Inc., Silver Spring, MD, USA). Montserrat Torà (Institut Municipal d'Investigació Mèdica, Barcelona, Spain). Jane Wang (Information Management Services, Silver Spring, MD, USA).

Conflict of Interest statement. The authors have declared that no competing interests exist.

FUNDING

This project has been funded in part with federal funds from the National Cancer, Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Support for individual studies that participated in the effort is as follows: SBCS (D.T.S.)—Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics and intramural, contract number NCI N02-CP-11015. FIS/Spain 98/1274, FIS/Spain 00/0745, PI061614 and G03/174, Fundació Marató TV3, Red Temática Investigación Cooperativa en Cáncer (RTICC), Consolíder ONCOBIO, EU-FP7-201663; and RO1-CA089715 and CA34627. NEBCS (D.T.S.)—Intramural research program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics and intramural, contract number NCI N02-CP-01037, PLCO (M.P.P.)—The NIH Genes, Environment and Health Initiative (GEI) partly funded, DNA extraction and statistical analyses (HG-06-033-NCI-01 and RO1HL091172-01), genotyping at the Johns Hopkins University Center for Inherited Disease Research (U01HG004438 and NIH HHSN268200782096C) and study coordination at the GENEVA (N.C.)—The NIH Genes, Environment and Health Initiative [GEI] partly funded DNA extraction and statistical analyses (HG-06-033-NCI-01 and RO1HL091172-01), genotyping at the Johns Hopkins University Center for Inherited Disease Research (U01HG004438 and NIH HHSN268200782096C) and study coordination at the GENEVA Coordination Center (U01 HG004446) for EAGLE and part of PLCO studies. Genotyping for the remaining part of PLCO and all ATBC and CPS-II samples were supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The PLCO is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, ATBC (D.A.)—This research was supported in part by the Intramural Research Program of the NIH and the National Cancer Institute. Additionally, this research was supported by US Public Health Service contracts N01-CN-45165, N01-RC-45035 and N01-RC-37004 from the National Cancer Institute, Department of Health and Human Services. NHS & HPFS (I.D.V.)—CA055075 and CA087969.

REFERENCES

1. Jemal, A., Siegel, R., Xu, J. and Ward, E. (2010) Cancer statistics, 2010. *CA Cancer J. Clin.*, **60**, 277–300.
2. Botteman, M.F., Pashos, C.L., Redaelli, A., Laskin, B. and Hauser, R. (2003) The health economics of bladder cancer: a comprehensive review of the published literature. *Pharmacoeconomics*, **21**, 1315–1330.
3. Svatek, R.S., Sagalowsky, A.I. and Lotan, Y. (2006) Economic impact of screening for bladder cancer using bladder tumor markers: a decision analysis. *Urol. Oncol.*, **24**, 338–343.
4. Dietrich, H. and Dietrich, B. (2001) Ludwig Rehn (1849–1930)—pioneering findings on the aetiology of bladder tumours. *World J. Urol.*, **19**, 151–153.
5. Case, R.A. and Pearson, J.T. (1954) Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. II. Further consideration of the role of aniline and of the manufacture of auramine and magenta (fuchsine) as possible causative agents. *Br. J. Ind. Med.*, **11**, 213–216.
6. Hecht, S.S. (2003) Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat. Rev. Cancer*, **3**, 733–744.
7. Shirai, T. (1993) Etiology of bladder cancer. *Semin. Urol.*, **11**, 113–126.
8. Poupko, J.M., Hearn, W.L. and Radomski, J.L. (1979) N-Glucuronidation of N-hydroxy aromatic amines: a mechanism for their transport and bladder-specific carcinogenicity. *Toxicol. Appl. Pharmacol.*, **50**, 479–484.
9. Bock, K.W. (1991) Roles of UDP-glucuronosyltransferases in chemical carcinogenesis. *Crit. Rev. Biochem. Mol. Biol.*, **26**, 129–150.
10. Murta-Nascimento, C., Silverman, D.T., Kogevinas, M., Garcia-Closas, M., Rothman, N., Tardon, A., Garcia-Closas, R., Serra, C., Carrato, A., Villanueva, C. *et al.* (2007) Risk of bladder cancer associated with family history of cancer: do low-penetrance polymorphisms account for the increase in risk? *Cancer Epidemiol. Biomarkers Prev.*, **16**, 1595–1600.
11. Kiemeney, L.A. (2008) Hereditary bladder cancer. *Scand. J. Urol. Nephrol. Suppl.*, **218**, 110–115.
12. Dong, L.M., Potter, J.D., White, E., Ulrich, C.M., Cardon, L.R. and Peters, U. (2008) Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *JAMA*, **299**, 2423–2436.
13. Garcia-Closas, M., Malats, N., Silverman, D., Dosemeci, M., Kogevinas, M., Hein, D.W., Tardon, A., Serra, C., Carrato, A., Garcia-Closas, R. *et al.* (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder

cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*, **366**, 649–659.
14. Hein, D.W. (2002) Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. *Mutat. Res.*, **506–507**, 65–77.
15. McGrath, M., Michaud, D. and De Vivo, I. (2006) Polymorphisms in GSTT1, GSTM1, NAT1 and NAT2 genes and bladder cancer risk in men and women. *BMC Cancer*, **6**, 239.
16. Sanderson, S., Salanti, G. and Higgins, J. (2007) Joint effects of the N-acetyltransferase 1 and 2 (NAT1 and NAT2) genes and smoking on bladder carcinogenesis: a literature-based systematic HuGE review and evidence synthesis. *Am. J. Epidemiol.*, **166**, 741–751.
17. Rothman, N., Garcia-Closas, M., Chatterjee, N., Malats, N., Wu, X., Figueroa, J.D., Real, F.X., Van Den Berg, D., Matullo, G., Baris, D. *et al.* (2010) A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.*, **42**, 978–984.
18. Nagar, S. and Remmel, R.P. (2006) Uridine diphosphoglucuronosyltransferase pharmacogenetics and cancer. *Oncogene*, **25**, 1659–1672.
19. Gong, Q.H., Cho, J.W., Huang, T., Potter, C., Gholami, N., Basu, N.K., Kubota, S., Carvalho, S., Pennington, M.W., Owens, I.S. *et al.* (2001) Thirteen UDPglucuronosyltransferase genes are encoded at the human UGT1 gene complex locus. *Pharmacogenetics*, **11**, 357–368.
20. Consortium, T.I.H. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
21. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
22. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q. and Krainer, A.R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
23. Nagar, S. and Blanchard, R.L. (2006) Pharmacogenetics of uridine diphosphoglucuronosyltransferase (UGT) 1A family members and its role in patient response to irinotecan. *Drug Metab. Rev.*, **38**, 393–409.
24. Tukey, R.H., Strassburg, C.P. and Mackenzie, P.I. (2002) Pharmacogenomics of human UDP-glucuronosyltransferases and irinotecan toxicity. *Mol. Pharmacol.*, **62**, 446–450.
25. Marques, S.C. and Ikediobi, O.N. (2010) The clinical application of UGT1A1 pharmacogenetic testing: gene-environment interactions. *Hum. Genomics*, **4**, 238–249.
26. Innocenti, F., Grimsley, C., Das, S., Ramirez, J., Cheng, C., Kuttab-Boulos, H., Ratain, M.J. and Di Rienzo, A. (2002) Haplotype structure of the UDP-glucuronosyltransferase 1A1 promoter in different ethnic groups. *Pharmacogenetics*, **12**, 725–733.
27. Saeki, M., Saito, Y., Jinno, H., Sai, K., Ozawa, S., Kurose, K., Kaniwa, N., Komamura, K., Kotake, T., Morishita, H. *et al.* (2006) Haplotype structures of the UGT1A gene complex in a Japanese population. *Pharmacogenomics J.*, **6**, 63–75.
28. Minami, H., Sai, K., Saeki, M., Saito, Y., Ozawa, S., Suzuki, K., Kaniwa, N., Sawada, J., Hamaguchi, T., Yamamoto, N. *et al.* (2007) Irinotecan pharmacokinetics/pharmacodynamics and UGT1A genetic polymorphisms in Japanese: roles of UGT1A1*6 and *28. *Pharmacogenet. Genomics*, **17**, 497–504.
29. Johnson, A.D., Kavousi, M., Smith, A.V., Chen, M.-H., Dehghan, A., Aspelund, T., Lin, J.-P., van Duijn, C.M., Harris, T.B., Cupples, L.A. *et al.* (2009) Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.*, **18**, 2700–2710.
30. Sanna, S., Busonero, F., Maschio, A., McArdle, P.F., Usala, G., Dei, M., Lai, S., Mulas, A., Piras, M.G., Perseu, L. *et al.* (2009) Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum. Mol. Genet.*, **18**, 2711–2718.
31. Tukey, R.H. and Strassburg, C.P. (2000) Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu. Rev. Pharmacol. Toxicol.*, **40**, 581–616.
32. Elliot, J.S., Sharp, R.F. and Lewis, L. (1959) Urinary pH. *J. Urol.*, **81**, 339–343.
33. Remer, T. and Manz, F. (1995) Potential renal acid load of foods and its influence on urine pH. *J. Am. Diet Assoc.*, **95**, 791–797.
34. Echeverry, G., Hortin, G.L. and Rai, A.J. (2010) Introduction to urinalysis: historical perspectives and clinical application. *Methods Mol. Biol.*, **641**, 1–12.
35. Alguacil, J., Kogevinas, M., Silverman, D., Malats, N., Real, F.X., Garcia-Closas, M., Tardon, A., Rivas, M., Tora, M., Garcia-Closas, R. *et al.* (2011) Urinary pH, cigarette smoking and bladder cancer risk. *Carcinogenesis.*, **32**, 843–847.
36. Kadlubar, F.F., Ketterer, B., Flammang, T.J. and Christodoulides, L. (1980) Formation of 3-(glutathion-S-YL)-N-methyl-4-aminoazobenzene and inhibition of aminoazo dye-nucleic acid binding *in vitro* by reaction of glutathione with metabolically-generated N-methyl-4-aminoazobenzene-N-sulfate. *Chem. Biol. Interact.*, **31**, 265–278.
37. Moore, B.P., Hicks, R.M., Knowles, M.A. and Redgrave, S. (1982) Metabolism and binding of benzo(a)pyrene and 2-acetylaminofluorene by short-term organ cultures of human and rat bladder. *Cancer Res.*, **42**, 642–648.
38. Nakamura, A., Nakajima, M., Yamanaka, H., Fujiwara, R. and Yokoi, T. (2008) Expression of UGT1A and UGT2B mRNA in human normal tissues and various cell lines. *Drug Metab. Dispos.*, **36**, 1461–1464.
39. Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C.A., Amini, B., Andersen, E., Andersson, A.C., Angelidou, P., Asplund, A., Asplund, C. *et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics*, **4**, 1920–1932.
40. Wang, W.Y., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.
41. Chung, C.C. and Chanock, S.J. (2011) Current status of genome-wide association studies in cancer. *Human Genetics.*, **130**, 59–78.
42. Chakravarti, A. (1999) Population genetics—making sense out of sequence. *Nat. Genet.*, **21**, 56–60.
43. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
44. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.*, **17**, 502–510.
45. Anderson, C.A., Pettersson, F.H., Barrett, J.C., Zhuang, J.J., Ragoussis, J., Cardon, L.R. and Morris, A.P. (2008) Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.*, **83**, 112–119.
46. Barrett, J.C. and Cardon, L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.
47. Panagiotou, O.A., Evangelou, E. and Ioannidis, J.P. (2010) Genome-wide significant associations for variants with minor allele frequency of 5% or less—an overview: a HuGE review. *Am. J. Epidemiol.*, **172**, 869–889.
48. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
49. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
50. Polychronakos, C. (2008) Common and rare alleles as causes of complex phenotypes. *Curr. Atheroscler. Rep.*, **10**, 194–200.
51. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
52. Campbell, H. and Manolio, T. (2007) Commentary: rare alleles, modest genetic effects and the need for collaboration. *Int. J. Epidemiol.*, **36**, 445–448.
53. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.*, **8**, e1000294.
54. Schork, N.J., Murray, S.S., Frazer, K.A. and Topol, E.J. (2009) Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, **19**, 212–219.
55. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R. and Amos, C.I. (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **82**, 100–112.
56. Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.
57. Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B. and Hakonarson, H. (2010) Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 730–742.
58. Goldstein, D.B. (2011) The importance of synthetic associations will only be resolved empirically. *PLoS Biol.*, **9**, e1001008.

59. Zhu, Q., Ge, D., Maia, J.M., Zhu, M., Petrovski, S., Dickson, S.P., Heinzen, E.L., Shianna, K.V. and Goldstein, D.B. (2011) A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.*, **88**, 458–468.
60. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M. *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, 714–719.
61. Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., Major, S., Morita, D., Chary, K.K., Reimers, M.A., Scherf, U., Kahn, A., Dolginow, D. *et al.* (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol. Cancer Ther.*, **6**, 820–832.
62. Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
63. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
64. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.