Human Molecular Genetics, 2005, Vol. 14, No. 12 ***1621–1629***
doi:10.1093/hmg/ddi170
Advance Access published on April 27, 2005

Global gene expression as a function of germline genetic variation

Deborah French[1,†], Mark R. Wilkinson[1,†], Wenjian Yang[1], Luc de Chaisemartin[1], Edwin H. Cook[5,6], Soma Das[6], Mark J. Ratain[7], William E. Evans[1,4], James R. Downing[2], Ching-Hon Pui[2,3] and Mary V. Relling[1,4,*]

[1]Department of Pharmaceutical Sciences, [2]Department of Pathology and [3]Department of Hematology–Oncology, St Jude Children's Research Hospital, Memphis, TN, USA, [4]University of Tennessee, Memphis, TN, USA and [5]Department of Psychiatry, [6]Department of Human Genetics and [7]Department of Medicine, University of Chicago, Chicago, IL, USA

Received January 5, 2005; Revised and Accepted April 19, 2005

Common, functional, germline genetic polymorphisms have been associated with clinical cancer outcomes. Little attention has been paid to the potential phenotypic consequences of germline genetic variation on downstream genes. We determined the germline status of 16 well-characterized functional polymorphisms in 126 children with newly diagnosed acute lymphoblastic leukemia (ALL). We assessed whether global gene expression profiles of diagnostic ALL blasts from the same patients differed by these germline polymorphic genotypes. Gene expression values were adjusted for ALL-subtype-specific patterns. Of the 16 loci, only the *UGT1A1* promoter repeat polymorphism [A(TA)nTAA] (*UGT1A128) and *GSTM1* deletion were significant predictors of global gene expression in a supervised approach, which divided patients based on their germline genotypes [*UGT1A1*: 124 probe sets, false discovery rate (FDR) = 13%, $P \leq 0.0031$; *GSTM1*: 112 probe sets, FDR = 42.5%, $P \leq 0.0084$]. Genes whose expression distinguished the *UGT1A1* (TA) 7/7 genotype from the other *UGT1A1* genotypes included *HDAC1*, *RELA* and *SLC2A1*; those that distinguished the *GSTM1* null genotype from non-null genotype included *NBS1* and *PRKR*. In an unsupervised approach, the gene expression profiles using the entire array delineated two major clusters of patients. The only germline genotype frequency that differed between the two clusters was *UGT1A1* ($P = 0.002$; Fisher's exact test). Although their expression is limited to specific tissues, both *GSTM1* and *UGT1A1* are involved in the conjugation (and thus transport, excretion and lipophilicity) of a broad range of endobiotics and xenobiotics, which could plausibly have consequences for gene expression in different tissues.**

INTRODUCTION

There is considerable interest in the possible impact of common, functional, germline polymorphisms on clinical outcomes among patients with cancer (1–6). Direct mechanistic studies can attribute differences in tissue-specific enzyme activity or substrate selectivity in gene products to germline genetic variation. However, relatively little attention has been paid to the possible consequences of germline variation on genome-wide phenotypic variation, which may have distinct effects on a variety of human tissues. For example, germline genetic variation of gene products involved in hepatic metabolism of a substrate (e.g. cytochrome P450 mediated synthesis of steroids) affects not only the liver tissue where metabolism localizes but also the distant tissues (vasculature, skeletal muscle, central nervous system, lymphoid tissue, kidney, etc.) that are responsive to the downstream effects of the circulating substrate (e.g. transcription regulated by steroid-sensitive nuclear hormone receptors).

The development of microarrays provides an efficient method of interrogating the possible broad effects of germline variation by enabling analysis of phenotype at the transcript

*To whom correspondence should be addressed at: Department of Pharmaceutical Sciences, St Jude Children's Research Hospital, 332 N Lauderdale, Memphis, TN 38105-2794, USA. Tel: +1 9014952348; Fax: +1 9015256869; Email: mary.relling@stjude.org
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

level. Gene expression profiling has been shown to identify molecular subtype and the risk of relapse in acute lymphoblastic leukemia (ALL) (7). In addition, it may be used to determine genetic risk factors in irradiation-induced brain tumors (8) and treatment-related myeloid leukemia (7). Within an individual, gene expression differs substantially among different tissue types (9). However, constitutive gene expression in brain, liver and lymphoid tissue demonstrates significant heritability (10–12), suggesting that germline genetic polymorphisms might affect gene expression and function across multiple tissue types, although this has yet to be studied.

Herein, we have compared global gene expression among unrelated individuals based on their germline genotypes at 16 functionally important polymorphic loci in genes involved in endobiotic and xenobiotic uptake, metabolism and detoxification. Our findings demonstrate that germline polymorphisms can affect gene expression profiles.

RESULTS

Genotype distributions were consistent with Hardy–Weinberg equilibrium within each race. However, distributions of genotypes between white and black patients were quite different, with genotype frequencies at nine of the 16 loci differing significantly by race ($P < 0.05$). In addition, within our sample population that had both genotype and gene expression data available, only 25 out of 165 patients were black. Because of the substantial racial differences in the genotype frequencies and in the functional consequences of some polymorphisms (13) and also because of the small number of black patients, we performed the gene selection analyses (described subsequently) within white patients, the largest racial group ($n = 126$).

None of the germline polymorphisms was significantly associated with the major molecular and immunophenotypic subtypes of ALL ($P = 0.12–0.86$; Supplementary Material, Table S1).

Supervised analysis of the association between polymorphisms and gene expression

One-way analysis of variance (ANOVA) revealed that the adjusted gene expression profiles were significantly associated with germline polymorphisms for two of the 16 loci: *GSTM1* and *UGT1A1* (Fig. 1; Table 1). We also treated the three genotypic categories as ordered categorical variables to enforce the effect of the heterozygous genotype as intermediate and preserve the assumption of no heterosis. *UGT1A1* and *GSTM1* remained the only polymorphisms that significantly clustered gene expression (data not shown).

For the *GSTM1* deletion polymorphism, two genotypes were possible (null or non-null). When the *t*-test was applied to the adjusted expression levels to order the probe sets, at $P < 0.001$, 19 probe sets distinguished *GSTM1* genotype with a false discovery rate (FDR) = 30.8% (Table 1). To determine the optimal number of probe sets to distinguish the null from the non-null *GSTM1* genotype, two-means clustering was performed. The optimal number of probe sets to distinguish *GSTM1* genotypes was 112, $P \leq 0.0084$, but with a high FDR of 42.5% (Supplementary Material, Fig. S1). Permutation analysis was used to evaluate the significance of these selected probe sets and on the basis of 500 permutations, the average number of significant ($P \leq 0.0084$) probe sets was 40. Only 10 permutations (2%) yielded a higher number of significant ($P \leq 0.0084$) probe sets than the observed 112 probe sets. The top distinguishing probe sets are indicated in Supplementary Material, Table S3. *GSTM1* and *GSTM4* probe sets, both of which are likely to anneal to *GSTM1* (deleted from the germline in *GSTM1* null individuals), were the most significant probe sets that distinguished *GSTM1* genotypes. Because our objective was to discover additional genes affected by the 16 germline polymorphisms, these two probe sets were excluded from the estimates of FDR and hierarchical clustering. Two-third versus one-third cross validation using the selected 112 probe sets provided an estimated prediction accuracy of 83%. The distinguishing probe sets included *FYN*, *WEE1*, *NBS1* and *PRKR* (Supplementary Material, Table S3).

Gene expression signals that were not adjusted for ALL subtype also differed by *GSTM1* genotype (Supplementary Material, Table S2), with the top 100 selected probe sets overlapping by 70% between the analyses using the adjusted and the unadjusted expression signals.

For *UGT1A1*, the initial ANOVA included all three genotypes (6/6, 6/7 and 7/7) for gene selection. At $P < 0.001$, 94 genes distinguished *UGT1A1* genotype, with FDR = 5% (Table 1). To further explore which of the three main genotypic categories accounted for the primary differences among the gene expression profiles, we performed pair-wise comparisons between the three *UGT1A1* genotypes (Supplementary Material, Fig. S2), and found that the 7/7 genotype differed from both the 6/7 and 6/6 genotypes, whereas the 6/6 genotype did not differ substantially from the 6/7 genotypic group. Thus, the two genotypes with six (TA) repeats (i.e. 6/6 and 6/7) were pooled as one genotypic group (6/*) and compared with the 7/7 group to generate the final probe set selection (Supplementary Material, Table S4). The *t*-test was applied to the adjusted expression levels to order the probe sets and at $P < 0.001$, 149 probe sets distinguished the two *UGT1A1* genotypic categories with an FDR of 3.2% (Table 1). To determine the optimal number of probe sets to distinguish *UGT1A1* 7/7 genotype from the other genotypes, two-means clustering was performed. The optimal number of probe sets was 124, $P \leq 0.0031$, with an FDR of 13% (Fig. 2). Permutation analysis was performed to evaluate the significance of these selected probe sets and on the basis of 500 permutations, the average number of significant probe sets was 13 ($P \leq 0.0031$). Only two permutations (0.4%) had more significant probe sets than the observed 124 probe sets. Leave-one-out cross validation using the selected 124 genes provided an estimated prediction accuracy of 87% for *UGT1A1* genotypes. Probe sets distinguishing *UGT1A1* genotypes included *HDAC1*, *TOP2B*, *RELA* and *SLC2A1* (Supplementary Material, Table S4).

Gene expression also differed by *UGT1A1* genotype when gene expression signals were not adjusted for ALL subtype (Supplementary Material, Table S2). The top 100 selected probe sets overlap 81% between the analyses using the adjusted and unadjusted gene expression signals.

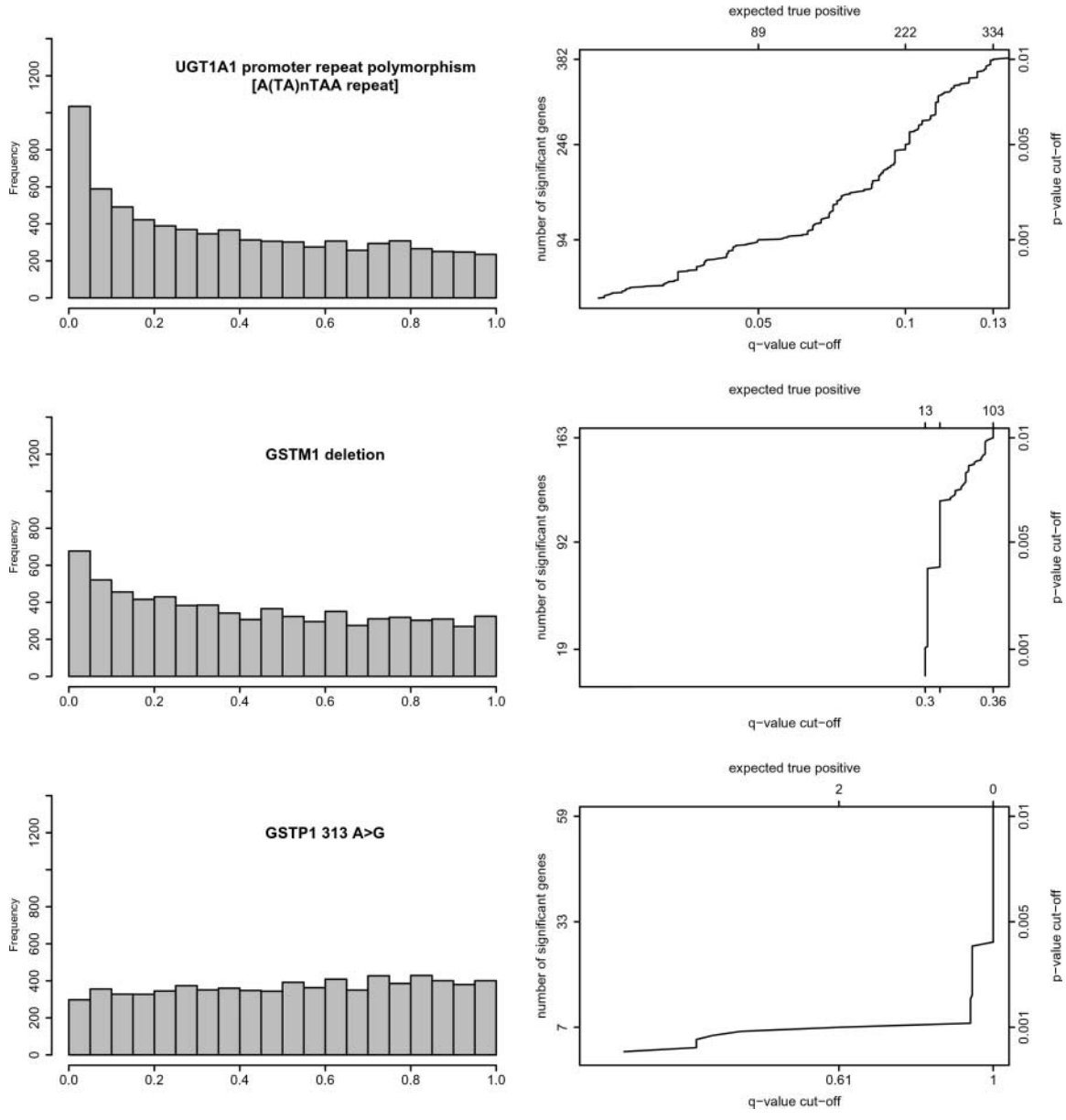

Figure 1. Distributions of *P*-values (left panels) for genotype versus gene expression profiles for the two loci [*UGT1A1* (top) and *GSTM1* (middle)] that were associated with gene expression profiles and one [*GSTP1* (bottom)] that was not significantly associated. Histograms illustrate observed *P*-values from ANOVA, where the *y*-axis represents the number of genes at a given *P*-value (alpha) depicted on the *x*-axis. Over-representation of small *P*-values (for *UGT1A1* and *GSTM1*) suggests a significant relationship between genotypes and gene expression. The corresponding graphs on the right panel depict the number of genes (left *y*-axes) with *P*-value less than any given *P* value cut-off (right *y*-axes), the estimated *q*-value/FDR (lower *x*-axes) and the estimated number of true positive genes (top *x*-axes).

Unsupervised analysis of the association of polymorphisms and gene expression

We applied hierarchical clustering to the adjusted gene expression signals of all 7369 probe sets, which defined two major clusters, designated cluster A and cluster B, with sizes of 85 and 41 patients, respectively. The distributions of genotypes for each of the 16 polymorphic loci were then compared between the two clusters. The only locus whose genotype frequency differed significantly between clusters A and B was *UGT1A1* ($P = 0.002$) (Supplementary Material, Table S5).

Expression of polymorphic genes

Of the 13 polymorphic genes, only *NR3C1*, *GSTM1*, *GSTP1*, *TPMT*, *MTHFR*, *RFC* and *TYMS* were themselves represented on the array with probe sets that passed the detection filter. Of these seven loci, *GSTM1* was the only gene that was differentially expressed between genotypes corresponding to that locus, with median expression levels of 5885 in patients with the non-null genotype and 2174 in those with the null genotype (*t*-test, $P < 0.0001$). The fact that expression levels in the null genotype were as high as they were may be due

Table 1. Number of significant probe sets and FDR estimations for all polymorphisms using expression levels adjusted for ALL subtypes

	P-value cut-off = 0.001				*P*-value cut-off = 0.005			
	N	FDR (from *q*-value)	Empirical FDR (from permutation)	Permutation *P*-value	*N*	FDR (from *q*-value)	Empirical FDR (from permutation)	Permutation *P*-value
*CYP3A4*1B*	18	38.6	100	0.52	68	51.6	76.5	0.34
*CYP3A5*3*	8	75.7	93.8	0.5	37	81.6	89.2	0.42
GSTM1 deletion	19	30.8	26.3	0.020	92	32.2	32.6	0.01
GSTP1 313 A>G	7	61.2	63.6	0.22	33	100	97	0.47
GSTT1 deletion	5	98	100	0.74	27	98	100	0.67
MDR1 exon 26 3435 C>T	5	99.9	100	0.70	32	99.9	100	0.52
MDR1 exon 21 2677 G>T/A	10	65.7	60	0.24	33	99.9	100	0.50
MTHFR 1298 A>C	3	99.5	100	0.91	21	99.5	100	0.89
MTHFR 677 C>T	11	64.6	100	0.53	28	99.9	100	0.68
NR3C1 1088 A>G	23	30.2	47.8	0.15	69	50.6	58	0.14
RFC 80G>A	8	78.2	75	0.31	37	90.3	81.1	0.32
TPMT	9	74.5	100	0.73	39	93.6	100	0.61
TYMS enhancer repeat	7	70.9	85.7	0.45	28	100	100	0.63
VDR Fok 1 start site T>C	5	99.9	100	0.78	25	99.9	100	0.82
VDR intron 8 G>A	3	100	100	0.90	13	100	100	0.98
*UGT1A1**28 (6/* versus 7/7)	149	3.2	6	0.002	366	6.6	9.7	0.002
*UGT1A1*28* (6/6 versus 6/7 versus 7/7)	94	5	8.5	0.004	246	9.5	13.8	0.004

N, Number of probe sets whose *P*-value for association with the genotype of interest is less than the cut-off *P*-value.
6/* represents 6/6 and 6/7 genotypes combined.

to possible cross-hybridization of the *GSTM1* probe set with *GSTM4*, as they share homology. We examined whether it was possible to predict *GSTM1* genotype based on *GSTM1* expression level. We determined that an absolute expression level of 3750 maximized accuracy of genotype prediction in the original 'training' set of the 126 children, with a prediction accuracy of 88%.

Independent validation of genes distinguishing *GSTM1* genotype

We used the expression array results and *GSTM1* genotypes from an additional 81 Caucasian children with ALL from the Total XIIIA treatment protocol at St Jude Children's Research Hospital as a test set (7). Two-means clustering using the 112 probe sets that distinguished *GSTM1* genotypes in the original training set generated two clusters in the test set, between which the *GSTM1* null genotype significantly differed in frequency (Fisher's Exact test, $P = 0.036$). Linear discriminant analysis using the 112 probe sets correctly predicted 62% of the *GSTM1* genotypes among these 81 patients ($P = 0.038$).

DISCUSSION

Many studies have successfully used microarray technology to interrogate the association between genome-wide expression profiles and various phenotypic endpoints (14–22). Acquired genomic abnormalities and mutations in somatic cells have been correlated with gene expression (7,23,24). In addition to directly affecting the encoded gene, germline polymorphisms could influence the phenotype of gene expression by virtue of downstream effects of the encoded gene, and could thereby regulate global gene expression in distant tissues. However, to date, little attention has been paid to the possible consequences of germline genetic variation on genome-wide expression in human tissues.

We found that a common germline genetic polymorphism in the *UGT1A1* promoter (*UGT1A1*28*) was the most significant predictor of global gene expression, although its product is not expressed in bone marrow. UGT1A1 is the major UDP-glucuronosyl transferase isoform expressed in the liver and is the principal isoform to catalyze bilirubin glucuronidation. Six TA repeats [A(TA)$_6$TAA (*UGT1A1*1*)] correspond to the higher activity, 'wild-type' allele, whereas the variant allele of seven TA repeats [A(TA)$_7$TAA (*UGT1A1*28*)] is associated with Gilbert's syndrome (25). Functional studies have shown that glucuronidation in *UGT1A1* 6/7 heterozygotes is closer to that seen in 6/6 homozygotes than the low activity 7/7 homozygotes (25–28), consistent with our finding that gene expression in the 6/7 heterozygotes was similar to the 6/6 homozygotes, whereas gene expression in the 7/7 genotype was distinct.

One mechanism by which *UGT1A1* genotype could affect downstream gene expression is through glucuronidation of 17β-estradiol (29). It is a ligand for estrogen receptor-alpha, which regulates transcription in diverse target cells (30). Patients with the *UGT1A1* 7/7 genotype would be expected to have higher concentrations of the unconjugated ligand and thus greater activation of the receptor (31,32).

The levels of several of the genes which differentiated the two *UGT1A1* genotypic groups could be related to altered 17β-estradiol levels. These include *RELA*, a component of the NFκB transcription factor complex that regulates many genes involved in immunity and inflammation (33); *HDAC1*, a component of the histone deacetylase complex that regulates

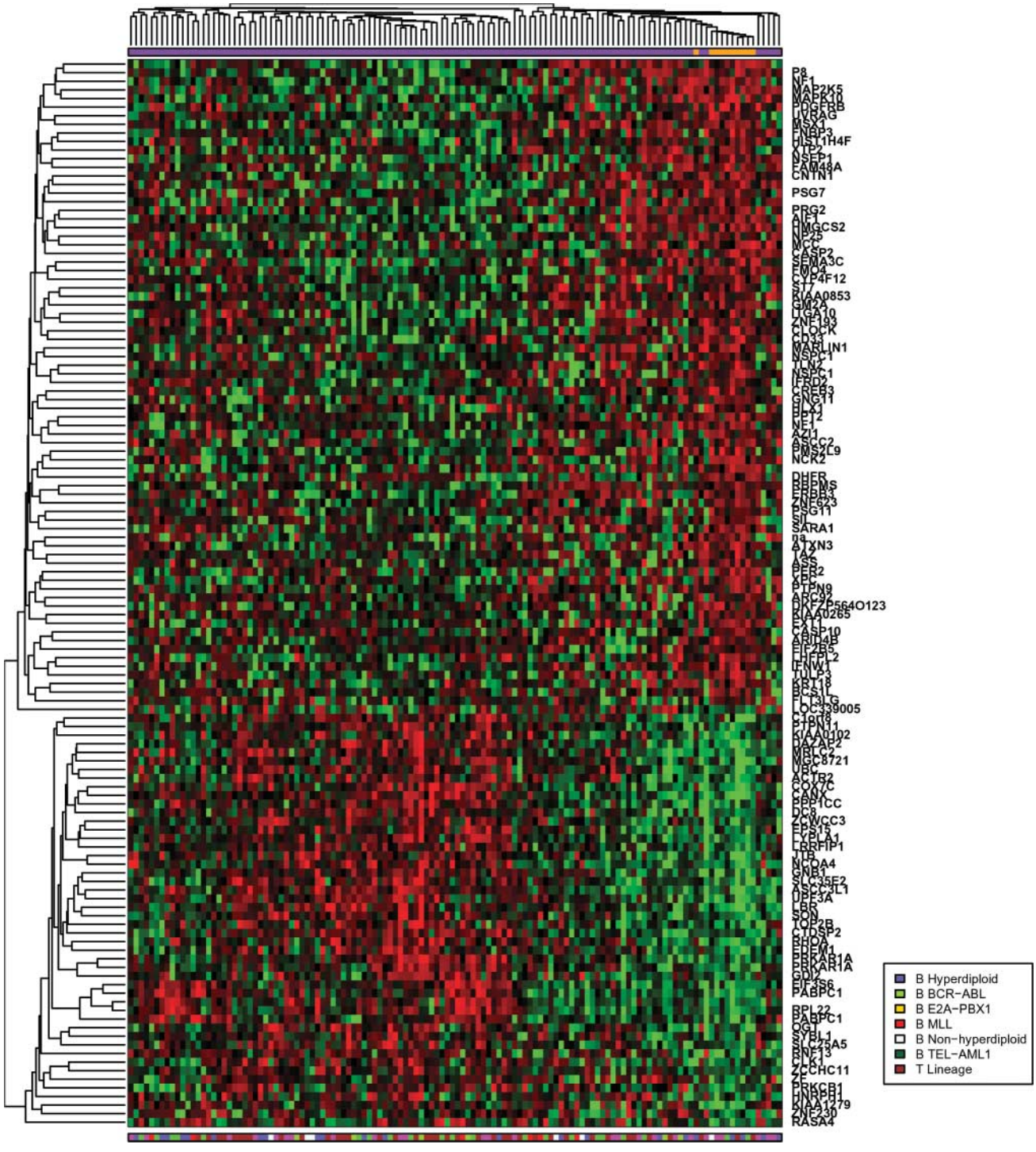

Figure 2. Hierarchical clustering of the 124 genes (rows) that optimally differentiated *UGT1A1* genotypes. Along the top of the diagram, orange indicates the individuals with the 7/7 genotype and purple indicates the individuals with the 6/6 and 6/7 genotypes. Along the bottom of the diagram, the ALL subtype of each individual is indicated by the colors shown in the legend. As expected, the adjusted expression levels (red = high expression, green = low expression) of these genes did not distinguish ALL subtypes.

eukaryotic gene expression (32,34) and *SLC2A1*, a glucose transporter in the blood–brain barrier (35,36).

Glucuronidation of several other endobiotics, such as thyroxine and leukotriene B4, could also plausibly regulate gene expression in lymphoid tissue (37–41).

The only other polymorphism to cluster gene expression levels was the glutathione-*S*-transferase (GST) M1 deletion, although this was only statistically significant in the supervised and not in the unsupervised analysis (Table 1). GSTs catalyze the conjugation of xenobiotics and endogenous

compounds to glutathione (42). Approximately 50% of whites carry a homozygous deletion of this gene (43–45). Several of the genes that differentiated non-null from null *GSTM1* genotypes are involved in response to oxidative stress, which would plausibly differ compensatorily in patients in response to low or to high GST activity. Included were *NBS1*, a member of the *MRE11/RAD50* complex involved in DNA double-strand break repair (46,47) and *PRKR*, a kinase that controls several stress response pathways (48). Because the mechanism of the *GSTM1* polymorphism involves total gene deletion, it is not surprising that expression levels of *GSTM1* are lower in patients with the germline homozygous deletion. Our data in an independent test set suggest that gene expression signatures can be used to predict the germline genotype for *GSTM1*.

Substantial constitutive variation exists in gene expression within and between populations, and this variation shows significant heritability (10-12,49–51). Approximately 60% of genes are expressed in most tissues (9), and therefore it is plausible that germline genetic variation may affect gene expression levels in distant tissues. A challenge is to sample tissues across individuals such that there is uniformity in the tissue type, to minimize the impact of cell type heterogeneity on gene expression. Although ALL blasts suffer from the disadvantage that they represent cells that have acquired genetic changes somatically (and thus differ from the germline), they have the advantage that at diagnosis, the vast majority of samples are >90% clonal for a single tissue type (blasts). Thus, we hypothesized that by adjusting gene expression signatures for the variability known to be associated with molecular subtype of ALL, gene expression in these samples is at least somewhat informative for germline effects; in fact, even unadjusted gene expression signals differed by germline polymorphisms.

Inter-individual variation in the expression, tissue-specific enzyme activity or substrate specificity of gene products can be directly attributable to germline genetic variation. To date, little attention has been paid to the possible consequences of germline genotype on distant tissues. Although the expression of both *GSTM1* and *UGT1A1* is concentrated in liver, they catalyze the conjugation and therefore transport and ultimately excretion of various endogenous and exogenous compounds, thereby affecting systemic levels of circulating regulatory small molecules. We have demonstrated that germline polymorphisms in these genes affect global gene expression profiles in lymphoid tissue.

MATERIALS AND METHODS

All children with newly diagnosed childhood ALL enrolled on the St Jude Children's Research Hospital treatment protocol Total XIIIB who had diagnostic bone marrow blasts available for expression array analysis were evaluated ($n = 165$) (7) for the primary analysis. The major subtypes of ALL were represented in this set, including t(9;22)[*BCR-ABL*], t(1;19)[*E2A-PBX1*], t(12;21)[*TEL-AML1*], *MLL* rearrangements, hyperdiploid >50 and T-cell ALL. An independent set of 81 patients with ALL from the predecessor protocol Total XIIIA served as a test set for the findings from the primary analysis.

DNA was extracted from normal blood cells. Genotyping was performed for 16 polymorphic loci. The three most common *TPMT* inactivating mutations, which define the *2, *3A, *3B and *3C alleles (52,53), were genotyped and taken together to classify each patient as wild-type or heterozygote as previously described (no homozygous variant patients were observed). *CYP3A4*1B* and *CYP3A5*3* (54), the *UGT1A1* promoter repeat polymorphism [A(TA)nTAA] (*UGT1A1*28*), *MDR1 (ABCB1)* exon 21 2677G>T/A, *MDR1 (ABCB1)* exon 26 3435C>T, *VDR* intron 8 G>A, *VDR* start site FokI, *GSTP1* 313A>G (13), the thymidylate synthase (*TYMS*) 5′-UTR repeat, the *GSTT1* deletion, the *GSTM1* deletion, the *MTHFR* 1298A>C polymorphism (55,56), *NR3C1* 1220A>G (57), *RFC (SLC19A1)* 80G>A and the *MTHFR* 677C>T polymorphisms (58) were all genotyped as described earlier.

The observed frequencies of some genotypes were quite low, and it was not clear how to 'group' these rare genotypes. Hence, the following genotypes were excluded from further analysis: the *TYMS* 4 repeat allele ($n = 3$; among whites, $n = 1$); the *MDR1 (ABCB1)* exon 21 'A' allele ($n = 7$; among whites, $n = 5$); *UGT1A1* A(TA)$_5$TAA allele ($n = 6$; among whites, $n = 0$) and the *CYP3A5*3* A/A genotype ($n = 15$; among whites, $n = 1$).

High quality total RNA (7) was extracted with TriReagent (MRC, Cincinnati, OH, USA) from cryopreserved mononuclear cell suspensions from bone marrow at diagnosis. RNA integrity was determined by the use of Agilent 2100 Bioanalyzer for concentration and size fractionation, and reproducibility was tested by processing 10% of the samples in duplicate on independent chips (7). The Affymetrix HG-U95Av2 GeneChip (Affymetrix Inc., Santa Clara, CA, USA) comprised 12 625 probe sets representing around 9600 unique genes and was used to interrogate the expression of RNA as described earlier (7). Signals (level of gene expression) and detection calls (presence of transcript) were reported based on MicroArray Suite version 5.0 (MAS5.0, Affymetrix®). To reduce the FDR, probe sets were filtered out if called 'present' in <5% of the patient samples, leaving 7369 probe sets for the analysis. Signals were $\log_2$-transformed for data analysis. Using expression data from the same patients (7,59) we previously showed that expression of specific genes by quantitative PCR was highly correlated with the assessment of expression by the Affymetrix HG-U95 Av2 chip.

Gene expression profiles vary significantly by the major ALL molecular subtypes (7,17,59–61). Unsupervised hierarchical clustering using unadjusted signals for 7369 probe sets confirmed this subtype/ploidy partitioning within our data set (data not shown). In order to discern possible relationships between gene expression patterns and germline host characteristics, we first adjusted the gene expression levels for the major ALL molecular subtypes. Because the specific acquired clonal genetic abnormalities found in leukemic blasts define the molecular ALL subtypes, and gene expression levels vary by ALL subtype, we reasoned that the variation in gene expression attributable to germline genetic variation would be better discerned if the gene expression levels were first

adjusted for ALL subtype. However, ALL subtype may itself be influenced by germline variation. Thus, we also analyzed the unadjusted expression levels (Supplementary Material, Table S2). To adjust the expression levels for ALL subtype, we applied ANOVA to the $\log_2$-transformed gene expression data, using the seven molecular subtypes of ALL as the independent factor. The residual levels of expression, after adjusting for subtype, were used as the adjusted gene expression levels for subsequent analyses.

Genotypes were treated as unordered categorical variables in the analyses. Fisher's exact test was used to test whether there was a confounding association between each polymorphism and the major ALL molecular subtypes (Supplementary Material, Table S1).

In a supervised analysis, for each polymorphic locus we applied ANOVA or *t*-test (for three or two genotypic categories, respectively) to the gene expression levels (dependent variables), to assess whether gene expression differed by germline genotype (independent variable) and to rank order the differentially expressed probe sets. We estimated the FDR based on the *q*-value method (62) and by an empirical procedure based on permutation to evaluate the significance of the probe sets selected by the *t*-test. For genotypes that showed a low FDR, we performed *k*-means clustering based on a varying number of top selected probe sets and recorded the misclassification rate compared with that obtained using the true genotypes. The optimal probe set list that distinguished between the genotypes was defined as that producing the minimal number of misclassifications. Hierarchical clustering was then applied based on the selected probe sets.

We performed cross validation to classify the original set of 126 patients into different genotypes using the expression of selected genes. For *GSTM1*, we randomly split the patients into a 2/3 training set and 1/3 test set. A linear discriminant analysis model (63) was built on the training set and applied to predict the *GSTM1* genotypes of the test set. We repeated the procedure 500 times and reported the average prediction accuracy. For *UGT1A1*, we performed leave-one-out cross validation due to the small number of patients of *UGT1A1* 7/7 genotype. Linear discriminant analysis was also used to predict the *GSTM1* genotype for patients in the independent test set of 81 additional patients.

In an unsupervised analysis, global hierarchical clustering was performed on the gene expression data and Fisher's exact test was used to test whether there was an association between each polymorphism and the major gene expression clusters. All statistical analyses were performed using the statistical environment R1.9.1 [R Development Core Team, http://www.r-project.org].

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Pam McGill, Nancy Duran, Jean Cai, Erick Vasquez and Peixian Chen for technical assistance; our clinical and research faculty and staff and the patients and their families for participating. This work was supported by NCI CA 51001, CA 78224, CA 21765 and the NIH/NIGMS Pharmacogenetics Research Network and Database (U01 GM61393, U01 GM61374 (www.pharmgkb.org) from the National Institutes of Health; by a Center of Excellence grant from the State of Tennessee and by the American Lebanese Syrian Associated Charities (ALSAC). C.-H.P. is the American Cancer Society F.M. Kirby Clinical Research Professor.

Conflict of Interest statement. None declared.

REFERENCES

1. Davies, S.M., Bhatia, S., Ross, J.A., Kiffmeyer, W.R., Gaynon, P.S., Radloff, G.A., Robison, L.L. and Perentesis, J.P. (2002) Glutathione *S*-transferase genotypes, genetic susceptibility, and outcome of therapy in childhood acute lymphoblastic leukemia. *Blood*, **100**, 67–71.
2. Krajinovic, M., Lemieux-Blanchard, E., Chiasson, S., Primeau, M., Costea, I. and Moghrabi, A. (2004) Role of polymorphisms in *MTHFR* and *MTHFD1* genes in the outcome of childhood acute lymphoblastic leukemia. *Pharmacogenomics. J.*, **4**, 66–72.
3. Relling, M.V. and Dervieux, T. (2001) Pharmacogenetics and cancer therapy. *Nat. Rev. Cancer*, **1**, 99–108.
4. Desai, A.A., Innocenti, F. and Ratain, M.J. (2003) Pharmacogenomics: road to anticancer therapeutics nirvana? *Oncogene*, **22**, 6621–6628.
5. de Jong, F.A., Marsh, S., Mathijssen, R.H., King, C., Verweij, J., Sparreboom, A. and McLeod, H.L. (2004) ABCG2 pharmacogenetics: ethnic differences in allele frequency and assessment of influence on irinotecan disposition. *Clin. Cancer Res.*, **10**, 5889–5894.
6. Bertucci, F., Nasser, V., Granjeaud, S., Eisinger, F., Adelaide, J., Tagett, R., Loriod, B., Giaconia, A., Benziane, A. Devilard, E. *et al.* (2002) Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. *Hum. Mol. Genet.*, **11**, 863–872.
7. Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.
8. Edick, M.J., Cheng, C., Yang, W, Cheok, M., Wilkinson, M., Pei, D., Evans, W.E., Kun, L.E., Pui, C.H. and Relling, M.V. (2005) Lymphoid gene expression as a predictor of risk of secondary brain tumors. *Genes Chromosomes Cancer*, **42**, 107–116.
9. Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V., Shmoish, M., Ophir, R., Benjamin-Rodrig, H., Safran, M., Domany, E. and Lancet, D. (2003) GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biol.*, **326**, 1067–1072.
10. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.
11. Cheung, V.G., Jen, K.Y., Weber, T., Morley, M., Devlin, J.L., Ewens, K.G. and Spielman, R.S. (2003) Genetics of quantitative variation in human gene expression. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 403–407.
12. Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R. *et al.* (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.
13. Kishi, S., Yang, W., Boureau, B., Morand, S., Das, S., Chen, P., Cook, E.H., Rosner, G.L., Schuetz, E., Pui, C.H., and Relling, M.V. (2004) Effects of prednisone and genetic polymorphisms on etoposide disposition in children with acute lymphoblastic leukemia. *Blood*, **103**, 67–72.
14. Cario, G., Stanulla, M., Fine, B.M., Teuffel, O., Neuhoff, V., Schrauder, A., Flohr, T., Schafer, B.W., Bartram, C.R., Welte, K. *et al.* (2005) Distinct gene expression profiles determine molecular treatment response in childhood acute lymphoblastic leukemia. *Blood*, **105**, 821–826.

15. Fine, B.M., Stanulla, M., Schrappe, M., Ho, M., Viehmann, S., Harbott, J. and Boxer, L.M. (2004) Gene expression patterns associated with recurrent chromosomal translocations in acute lymphoblastic leukemia. *Blood*, **103**, 1043–1049.
16. Holleman, A., Cheok, M.H., den Boer, M.L., Yang, W., Veerman, A.J., Kazemier, K.M., Pei, D., Cheng, C., Pui, C.H., Relling, M.V. *et al.* (2004) Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *N. Engl. J. Med.*, **351**, 533–542.
17. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
18. Ferrando, A.A., Neuberg, D.S., Staunton, J., Loh, M.L., Huard, C., Raimondi, S.C., Behm, F.G., Pui, C.H., Downing, J.R., Gilliland, D.G. *et al.* (2002) Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell*, **1**, 75–87.
19. Szakacs, G., Annereau, J.P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K.J., Reinhold, W., Guo, Y., Kruh, G.D., Reimers, M. *et al.* (2004) Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. *Cancer Cell*, **6**, 129–137.
20. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.
21. Davis, R.E. and Staudt, L.M. (2002) Molecular diagnosis of lymphoid malignancies by gene expression profiling. *Curr. Opin. Hematol.*, **9**, 333–338.
22. Kakiuchi, S., Daigo, Y., Ishikawa, N., Furukawa, C., Tsunoda, T., Yano, S., Nakagawa, K., Tsuruo, T., Kohno, N., Fukuoka, M. *et al.* (2004) Prediction of sensitivity of advanced non-small cell lung cancers to gefitinib (Iressa, ZD1839). *Hum. Mol. Genet.*, **13**, 3029–3043.
23. Huang, J.Z., Sanger, W.G., Greiner, T.C., Staudt, L.M., Weisenburger, D.D., Pickering, D.L., Lynch, J.C., Armitage, J.O., Warnke, R.A., Alizadeh, A.A. *et al.* (2002) The t(14;18) defines a unique subset of diffuse large B-cell lymphoma with a germinal center B-cell gene expression profile. *Blood*, **99**, 2285–2290.
24. Armstrong, S.A., Golub, T.R. and Korsmeyer, S.J. (2003) MLL-rearranged leukemias: insights from gene expression profiling. *Semin. Hematol.*, **40**, 268–273.
25. Bosma, P.J., Chowdhury, J.R., Bakker, C., Gantla, S., de Boer, A., Oostra, B.A., Lindhout, D., Tytgat, G.N., Jansen, P.L. and Oude Elferink, R.P. (1995) The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *N. Engl. J. Med.*, **333**, 1171–1175.
26. Duguay, Y., McGrath, M., Lepine, J., Gagne, J.F., Hankinson, S.E., Colditz, G.A., Hunter, D.J., Plante, M., Tetu, B., Belanger, A. *et al.* (2004) The functional UGT1A1 promoter polymorphism decreases endometrial cancer risk. *Cancer Res.*, **64**, 1202–1207.
27. Miners, J.O., McKinnon, R.A. and MacKenzie, P.I. (2002) Genetic polymorphisms of UDP-glucuronosyltransferases and their functional significance. *Toxicology*, **181–182**, 453–456.
28. Iyer, L., King, C.D., Whitington, P.F., Green, M.D., Roy, S.K., Tephly, T.R., Coffman, B.L. and Ratain, M.J. (1998) Genetic predisposition to the metabolism of irinotecan (CPT-11). Role of uridine diphosphate glucuronosyltransferase isoform 1A1 in the glucuronidation of its active metabolite (SN-38) in human liver microsomes. *J. Clin. Invest.*, **101**, 847–854.
29. Cheng, Z., Rios, G.R., King, C.D., Coffman, B.L., Green, M.D., Mojarrabi, B., MacKenzie, P.I. and Tephly, T.R. (1998) Glucuronidation of catechol estrogens by expressed human UDP-glucuronosyltransferases (UGTs) 1A1, 1A3, and 2B7. *Toxicol. Sci.*, **45**, 52–57.
30. McDonnell, D.P. and Norris, J.D. (2002) Connections and regulation of the human estrogen receptor. *Science*, **296**, 1642–1644.
31. Guillemette, C., Deivo, VI., Hankinson, S.E., Haiman, C.A., Spiegelman, D., Housman, D.E. and Hunter, D.J. (2001) Association of genetic polymorphisms in UGT1A1 with breast cancer and plasma hormone levels. *Cancer Epidemiol. Biomarkers Prev.*, **10**, 711–714.
32. Sparks, R., Ulrich, C.M., Bigler, J., Tworoger, S.S., Yasui, Y., Rajan, K.B., Porter, P., Stanczyk, F.Z., Ballard-Barbash, R., Yuan, X. *et al.* (2004) UDP-glucuronosyltransferase and sulfotransferase polymorphisms, sex hormone concentrations, and tumor receptor status in breast cancer patients. *Breast Cancer Res.*, **6**, R488–R498.
33. Valentine, J.E., Kalkhoven, E., White, R., Hoare, S. and Parker, M.G. (2000) Mutations in the estrogen receptor ligand binding domain discriminate between hormone-dependent transactivation and transrepression. *J. Biol. Chem.*, **275**, 25322–25329.
34. Kawai, H., Li, H., Avraham, S., Jiang, S. and Avraham, H.K. (2003) Overexpression of histone deacetylase HDAC1 modulates breast cancer progression by negative regulation of estrogen receptor alpha. *Int. J. Cancer*, **107**, 353–358.
35. Wang, D.Y., Fulthorpe, R., Liss, S.N. and Edwards, E.A. (2004) Identification of estrogen-responsive genes by complementary deoxyribonucleic acid microarray and characterization of a novel early estrogen-induced gene: EEIG1. *Mol. Endocrinol.*, **18**, 402–411.
36. Shi, J. and Simpkins, J.W. (1997) 17 beta-Estradiol modulation of glucose transporter 1 expression in blood–brain barrier. *Am. J. Physiol.*, **272**, E1016–E1022.
37. Findlay, K.A., Kaptein, E., Visser, T.J. and Burchell, B. (2000) Characterization of the uridine diphosphate-glucuronosyltransferase-catalyzing thyroid hormone glucuronidation in man. *J. Clin. Endocrinol. Metab.*, **85**, 2879–2883.
38. Bocher, V., Pineda-Torra, I., Fruchart, J.C. and Staels, B. (2002) PPARs: transcription factors controlling lipid and lipoprotein metabolism. *Ann. N Y Acad. Sci.*, **967**, 7–18.
39. Turgeon, D., Chouinard, S., Belanger, P., Picard, S., Labbe, J.F., Borgeat, P. and Belanger, A. (2003) Glucuronidation of arachidonic and linoleic acid metabolites by human UDP-glucuronosyltransferases. *J. Lipid Res.*, **44**, 1182–1191.
40. Garcia-Silva, S. and Aranda, A. (2004) The thyroid hormone receptor is a suppressor of ras-mediated transcription, proliferation, and transformation. *Mol. Cell Biol.*, **24**, 7514–7523.
41. Yen, P.M. (2001) Physiological and molecular basis of thyroid hormone action. *Physiol. Rev.*, **81**, 1097–1142.
42. Board, P., Coggan, M., Johnston, P., Ross, V., Suzuki, T. and Webb, G. (1990) Genetic heterogeneity of the human glutathione transferases: a complex of gene families. *Pharmacol. Ther.*, **48**, 357–369.
43. Daly, A.K., Thomas, D.J., Cooper, J., Pearson, W.R., Neal, D.E. and Idle, J.R. (1993) Homozygous deletion of gene for glutathione *S*-transferase M1 in bladder cancer. *Br. Med. J.*, **307**, 481–482.
44. Townsend, D.M. and Tew, K.D. (2003) The role of glutathione-*S*-transferase in anti-cancer drug resistance. *Oncogene*, **22**, 7369–7375.
45. Xu, S., Wang, Y., Roe, B. and Pearson, W.R. (1998) Characterization of the human class Mu glutathione *S*-transferase gene cluster and the GSTM1 deletion. *J. Biol. Chem.*, **273**, 3517–3527.
46. Matsuura, S., Kobayashi, J., Tauchi, H. and Komatsu, K. (2004) Nijmegen breakage syndrome and DNA double strand break repair by NBS1 complex. *Adv. Biophys.*, **38**, 65–80.
47. Demuth, I., Frappart, P.O., Hildebrand, G., Melchers, A., Lobitz, S., Stockl, L., Varon, R., Herceg, Z., Sperling, K., Wang, Z.Q. and Digweed, M. (2004) An inducible null mutant murine model of Nijmegen breakage syndrome proves the essential function of NBS1 in chromosomal stability and cell viability. *Hum. Mol. Genet.*, **13**, 2385–2397.
48. Williams, B.R. (1999) PKR; a sentinel kinase for cellular stress. *Oncogene*, **18**, 6112–6120.
49. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
50. Pastinen, T., Sladek, R., Gurd, S., Sammak, A., Ge, B., Lepage, P., Lavergne, K., Villeneuve, A., Gaudin, T., Brandstrom, H. *et al.* (2004) A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics*, **16**, 184–193.
51. Whitney, A.R., Diehn, M., Popper, S.J., Alizadeh, A.A., Boldrick, J.C., Relman, D.A. and Brown, P.O. (2003) Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. U S A*, **100**, 1896–1901.
52. Krynetski, E.Y., Schuetz, J.D., Galpin, A.J., Pui, C.H., Relling, M.V. and Evans, W.E. (1995) A single point mutation leading to loss of catalytic activity in human thiopurine S-methyltransferase. *Proc. Natl Acad. Sci. USA*, **92**, 949–953.
53. Tai, H.L., Krynetski, E.Y., Yates, C.R., Loennechen, T., Fessing, M.Y., Krynetskaia, N.F. and Evans, W.E. (1996) Thiopurine *S*-methyltransferase deficiency: two nucleotide transitions define the most prevalent mutant allele associated with loss of catalytic activity in Caucasians. *Am. J. Hum. Genet.*, **58**, 694–702.

54. Blanco, J.G., Edick, M.J., Hancock, M.L., Winick, N.J., Dervieux, T., Amylon, M.D., Bash, R.O., Behm, F.G., Camitta, B.M., Pui, C.H. *et al.* (2002) Genetic polymorphisms in CYP3A5, CYP3A4 and NQO1 in children who developed therapy-related myeloid malignancies. *Pharmacogenetics*, **12**, 605–611.
55. Marsh, S., Ameyaw, M.M., Githang'a, J., Indalo, A., Ofori-Adjei, D. and McLeod, H.L. (2000) Novel thymidylate synthase enhancer region alleles in African populations. *Hum. Mutat.*, **16**, 528.
56. Relling, M.V., Yang, W., Das, S., Cook, E.H., Rosner, G.L., Neel, M., Howard, S., Ribeiro, R., Sandlund, J.T., Pui, C.H. and Kaste, S.C. (2004) Pharmacogenetic risk factors for osteonecrosis of the hip among children with leukemia. *J. Clin. Oncol.*, **22**, 3930–3936.
57. Koper, J.W. (1997) Lack of association between five polymorphisms in the human glucocorticoid receptor gene and glucocorticoid resistance. *Hum. Genet.*, **99**, 663–668.
58. Kishi, S., Griener, J.C., Cheng, C., Das, S., Yang, P., Cook, E.H., Hudson, M., Rubnitz, J., Sandlund, J.T., Pui, C.H. and Relling, M.V. (2003) Homocysteine, pharmacogenetics, and neurotoxicity in children with leukemia. *J. Clin. Oncol.*, **21**, 3084–3091.
59. Cheok, M.H., Yang, W., Pui, C.H., Downing, J.R., Cheng, C., Naeve, C.W., Relling, M.V. and Evans, W.E. (2003) Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat. Genet.*, **34**, 85–90.
60. Kohlmann, A., Schoch, C., Schnittger, S., Dugas, M., Hiddemann, W., Kern, W. and Haferlach, T. (2003) Molecular characterization of acute leukemias by use of microarray technology. *Genes Chromosomes Cancer*, **37**, 396–405.
61. Ross, M.E., Zhou, X., Song, G., Shurtleff, S.A., Girtman, K., Williams, W.K., Liu, H.C., Mahfouz, R., Raimondi, S.C., Lenny, N. *et al.* (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.
62. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. U S A*, **100**, 9440–9445.
63. Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Section 3.1, p. 92.