nature genetics

Genome-wide analysis of transcript isoform variation in humans

Tony Kwan[1,2], David Benovoy[1,2], Christel Dias[1], Scott Gurd[2], Cathy Provencher[2], Patrick Beaulieu[3], Thomas J Hudson[1,2,4], Rob Sladek[1,2] & Jacek Majewski[1,2]

We have performed a genome-wide analysis of common genetic variation controlling differential expression of transcript isoforms in the CEU HapMap population using a comprehensive exon tiling microarray covering 17,897 genes. We detected 324 genes with significant associations between flanking SNPs and transcript levels. Of these, 39% reflected changes in whole gene expression and 55% reflected transcript isoform changes such as splicing variants (exon skipping, alternative splice site use, intron retention), differential 5′ UTR (initiation of transcription) use, and differential 3′ UTR (alternative polyadenylation) use. These results demonstrate that the regulatory effects of genetic variation in a normal human population are far more complex than previously observed. This extra layer of molecular diversity may account for natural phenotypic variation and disease susceptibility.

Alternative pre-mRNA processing increases the complexity of eukaryotic transcriptomes, allowing multiple transcripts and protein isoforms with distinct functions to be produced from a single genomic locus[1]. Within an organism, tissue specific gene isoforms are known to have important functions in development and proper functioning of diverse cell types[2]. Across individuals, changes in normal isoform structure have phenotypic consequences and have been associated with disease[3,4]. Splicing defects in a number of genes, such as the cystic fibrosis transmembrane conductance regulator, *CFTR*, result in several known mendelian disorders[5]. More subtle changes, such as alternative 3′ processing and polyadenylation, have recently been associated with complex disorders: *OAS1* in severe acute respiratory syndrome[6], *TAP2* in type I diabetes[7], and *IRF5* in susceptibility to systemic lupus erythematosus[8,9].

Several recent studies have suggested that natural variation at the level of whole-gene expression is common in humans and is associated with genetic variants, such as SNPs or copy number variants (CNVs)[10–13]. Studying variation in gene expression is becoming increasingly important because of its contribution to phenotypic differences among individuals and its possible regulatory and functional relationships to diseases. However, little is known at present about the genetic variation at the sub-transcript level or about differences in multiple transcript isoforms of the same gene. Here, we interrogated transcripts across their entire length, using the Affymetrix GeneChip Human Exon 1.0 ST Array, which can detect splicing differences between various types of samples[14–16].

Exons within a gene are represented on the microarray by individual probe sets, and were considered discrete units for our analysis of transcript isoform-processing differences. We used triplicate samples of lymphoblastoid cell lines (LCLs) derived from 57 unrelated Centre d'Etudes du Polymorphisme Humain (CEPH) CEU individuals (Utah residents with northern and western European ancestry) genotyped by the HapMap consortium[17], allowing us to establish a possible genetic basis for any observed variations in transcript isoforms with associated SNPs. A linear regression analysis under a codominant model was carried out to associate probe set expression intensities with the genotypes of all SNP markers within a window of 50 kb flanking the boundaries of the transcript cluster (meta–probe set) containing the probe set. We assessed the statistical significance of the variation using the *t*-statistic, and used the regression equation to estimate the fold change in expression between the two homozygous genotypes. We used permutation testing[18] to determine empirical *P*-values corresponding to the asymptotic *P*-values obtained from the regression. Subsequently, we applied the false discovery rate (FDR) correction to establish a cutoff *P*-value of 9.73×10^{-9}, corresponding to the 0.05 FDR level (see Methods). This yielded 757 unique probe sets showing significant SNP associations, belonging to 317 unique meta–probe sets (**Supplementary Table 1** online). Although the most significant SNPs may not be the causative polymorphisms responsible for these differences in probe set expression, they are very probably in linkage disequilibrium with the causative polymorphism(s). This is reflected in the distance distribution of associated polymorphisms, most of which are in close proximity to the probe sets (**Supplementary Fig. 1** online). The association analysis at the transcript (meta–probe set) level resulted in a 0.05 FDR cutoff of 6.02×10^{-7}, yielding 127 unique transcripts with significant genetic association at the gene expression

[1]Department of Human Genetics, McGill University and [2]McGill University and Génome Québec Innovation Centre, 740 Dr. Penfield, Room 7210, Montréal, Québec H3A 1A4, Canada. [3]Division of Hematology-Oncology, Research Centre, Sainte-Justine Hospital, Montréal, Québec H3T 1C5, Canada. [4]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 800, Toronto, Ontario M5G 1L7, Canada. Correspondence should be addressed to J.M. (jacek.majewski@mcgill.ca).

Received 9 July 2007; accepted 31 October 2007; published online 13 January 2008; doi:10.1038/ng.2007.57

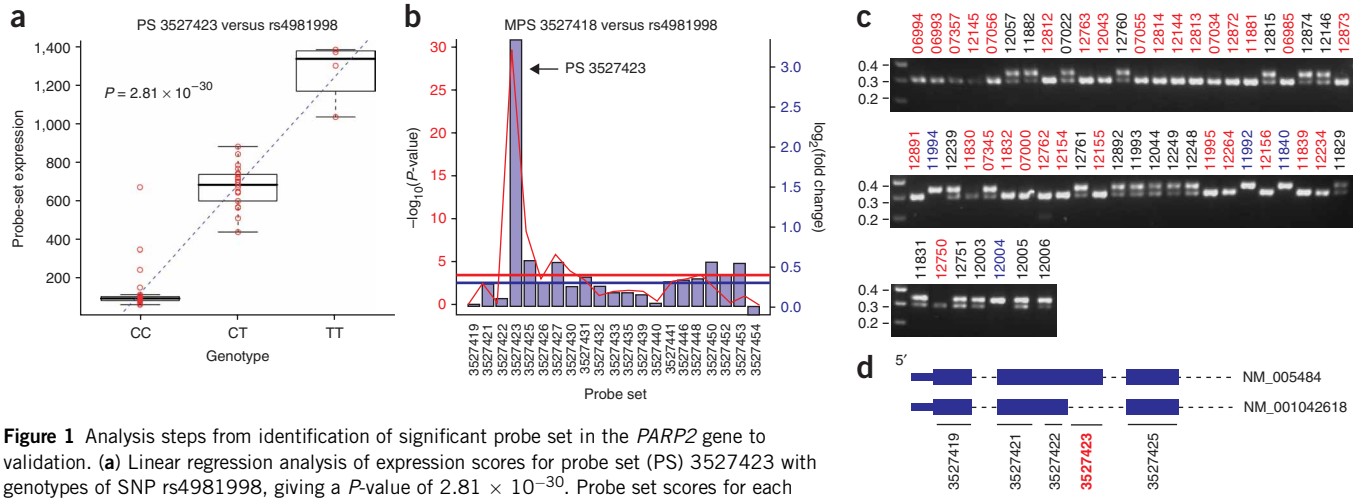

Figure 1 Analysis steps from identification of significant probe set in the *PARP2* gene to validation. (**a**) Linear regression analysis of expression scores for probe set (PS) 3527423 with genotypes of SNP rs4981998, giving a *P*-value of 2.81×10^{-30}. Probe set scores for each individual are shown in red and regression line is indicated with blue dashes. (**b**) Visualization of probe set 3527423 in the context of all other probe sets belonging to the same transcript (meta–probe set 3527418). For each probe set, the significance level (*P*-value) is graphed (red line), along with fold change expression between the mean scores of the two homozygous genotypes ($mean_{TT}$ / $mean_{CC}$) (vertical blue bars). The solid horizontal red and blue lines represent the significance and fold change expression for the regression analysis at the meta–probe set level against SNP rs4981998. Arrow, probe set 3527423. (**c**) RT-PCR validation of probe set 3527423 using flanking exon-body primers. Individuals are highlighted by color according to their genotype for SNP rs4981998: CC (red), CT (black), TT (blue). (**d**) Schematic of 5′ end of two isoforms of *PARP2* with exon array probe sets shown below the exons. The significant probe set 3527423 is highlighted in red and corresponds to alternative 5′ splice site use resulting in a larger second exon for NM_005484.

level. Of these 127 transcripts, all but seven were common to the 317 transcripts derived from the regression analysis at the probe-set level; therefore, our final dataset comprised 324 transcripts predicted to have expression changes at the meta–probe set and/or probe set level.

We examined the 324 transcripts in greater detail (**Fig. 1**; examples in **Fig. 2**) to determine the nature of the isoform changes on a transcript level (summarized in **Supplementary Table 2** and **Supplementary Fig. 2** online). Expression changes were automatically classified on the basis of the positions of the variable probe sets, followed by manual curation based on visualization of the entire transcript (**Supplementary Fig. 2**). A large number of genes (127, or 39%) showed whole-gene expression changes. However, an even larger proportion (55%) of genes showed transcript-isoform changes only, without an accompanying change in the expression of the entire locus. Nearly half of these transcript variations were at the splicing level (85, or 26%), with the remaining changes at the level of transcript termination (57, or 18%) and initiation (35, or 11%) (**Fig. 3**). It should be noted that some of the genes showing changes in the expression level of the whole gene also showed further changes in splicing, transcript termination and/or transcript initiation, suggesting that transcript isoform variation constitutes a large part of the genetic variation we have observed. A small number (20, or 6%) of genes showed very complex patterns of isoform variation that were difficult to interpret. Notably, when we compare the proportion (18%) of significant probe sets within the 3′ untranslated regions (UTRs) with the proportion of all 3′ UTR core probe sets (13%) on the array, we found a significant over-representation (Pearson's chi-squared test, $P = 5.73 \times 10^{-6}$) of probe sets in this region, indicating that transcript termination variations may occur more frequently than expected. Because predicted changes to the 3′ UTR may affect mRNA stability and subcellular localization, this type of isoform variation may have important regulatory roles. These findings illustrate a very complex pattern of expression changes associated with genetic variation, encompassing alterations at the whole-gene expression level and/or differences in transcript isoforms.

We proceeded, using two different methods, to validate 32 of our top candidate events distributed among the coding (16), 5′ UTR (6), and 3′ UTR (10) regions. For alternative splicing events of internally located probe sets, we performed RT-PCR on our entire panel of cell lines using exon-body primers in the two exons flanking the candidate probe set (**Fig. 1c**). We confirmed 15 probe sets showing SNP association to splicing of a cassette exon or intron (**Table 1**) and classified them as follows: eight probe sets corresponded to splicing of a coding exon, four probe sets were located in the 5′ UTR and resulted in the removal of potential promoter sequences or alternative start codon use, two probe sets were found within intronic regions and resulted in intron retention, and the remaining probe set was located in the 3′ UTR and altered its length. The second, more sensitive validation method using quantitative real-time RT-PCR was applied to differentially expressed probe sets within the 5′ or 3′ UTR and to those in which one of the flanking probe sets was missing in one of the alternative isoforms. We designed sets of primers to amplify the differentially expressed probe set itself and compared the resulting PCR products to ones corresponding to adjacent probe sets showing no association to the SNP and also expected to have similar expression levels across all cell lines. Quantitative PCR data was used to perform a linear regression fit with the original associated SNP and confirm the significance and direction of the association analysis with the microarray data at a nominal *P*-value of $0.05/N$, where N is the number of candidates tested in the real-time RT-PCR. Using this method, we validated six UTR-located probe sets showing SNP association: four in the 3′ UTR (alternative polyadenylation) and two in the 5′ UTR (differential transcriptional initiation). We also used this method on the candidate probe sets that failed our initial validation method owing potentially to low sensitivity of endpoint PCR of minor isoforms, and we were able to validate another four probe sets: two within coding regions and two within the 3′ UTRs. In total, 25 of 32 candidate probe sets were validated, for a success rate of 78%. The remaining 7 probe sets failed validation, which can be partially accounted for by unannotated SNPs located within the probe sets possibly leading to altered

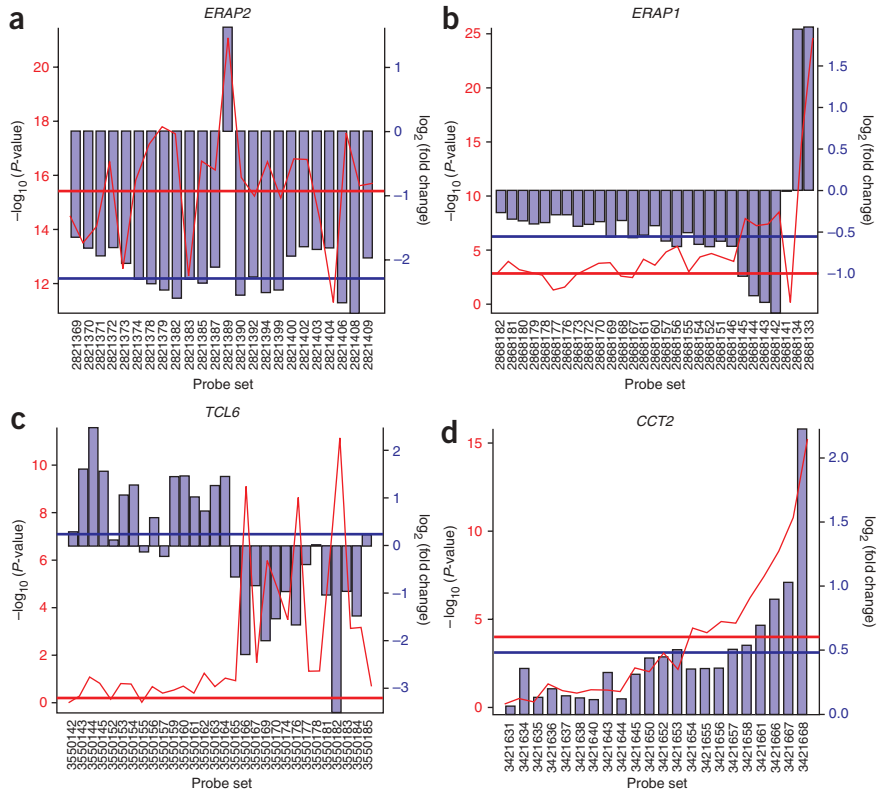

Figure 2 Examples of different types of transcript isoform events observed. Data is graphed as in **Figure 1b**. (**a**) Gene expression level changes of *ERAP2*, including alternative splicing of a cassette exon. (**b**) Differential 3′ UTR change of *ERAP1* resulting in long and short isoforms with alternative stop codon use. (**c**) Expression of two *TCL6* transcript isoforms that contain different 5′ and 3′ ends. (**d**) Increasing significance and fold change in expression levels toward the 3′ end of the *CCT2* gene, suggesting genetic variation associated with mRNA stability.

hybridization signals[19] (see Methods), suboptimal primer design, limited sensitivity of our validation methods, and/or noise from the microarray. We also validated several differentially spliced exons under a more relaxed stringency below our estimated cutoff, indicating that the frequency of genes showing SNP-associated changes is probably greater than what can be estimated from our current analysis. A recent estimate suggests that ∼21% of annotated alternatively spliced genes are associated with SNPs that determine the relative abundances of the alternative transcript isoforms[20].

A recent study used Illumina arrays to capture gene expression information within the CEU population[13]. The Illumina design, along with many other expression platforms, targets probes to the 3′ end of genes and cannot identify specific isoform changes. Our present results demonstrate that the nature of the changes is qualitatively different than previously reported for several genes in that study. For example, our analysis shows that *IRF5*, implicated in susceptibility to systemic lupus erythematosus, shows differences in the 3′ UTR (**Fig. 4**), where the A allele of rs10954213 creates a functional polyadenylation site, shortening its 3′ UTR[8,9]. This result for *IRF5* contrasts the original predicted change at the gene expression level[10,13] and occurs because the Illumina array interrogates *IRF5* with a probe in the 3′ UTR specific to the long isoform. Other examples previously classified as expression changes include *PTER*, which we show to have a variation in the 3′ UTR, and *C17orf81* (also known as *DERP6*), which shows alternative splicing of a cassette exon. Another interesting example is *ERAP2*, which has been reported as having an expression change[10]. Our results confirm this variation in expression; however, we additionally detect alternative splice-site use in one of the exons (**Fig. 2a**). Many platforms have been used so far in these population-wide expression analyses, and although there is substantial overlap between the studies, significant discordance also exists. A recent paper identified 374 gene-expression phenotypes associated with SNP markers from a study of 3,554 genes[10]. Differences in statistical stringency and false discovery rate most likely explain the higher proportion of SNP associations in their study. However, their set of 3,554 genes was preselected for the most variable expression phenotypes among an original set of >8,000 genes. This restricted set of genes may exclude examples of isoform changes without an accompanying change in whole-gene expression, which we observed in our study. In future expression association studies, comparative meta-analyses across different microarray designs may help eliminate platform-specific technical artifacts and allow the elucidation of true isoform and gene-level variations.

We show that tools such as the exon array, targeting probes to many regions of the gene, give a more complete picture of the true complexity of variation in gene expression than previously believed. This variation exists at all levels of transcript processing, beginning with initiation of transcription, through pre-mRNA splicing[16,20,21], to alternative polyadenylation, and it has the potential to exert diverse cellular responses and phenotypic effects.

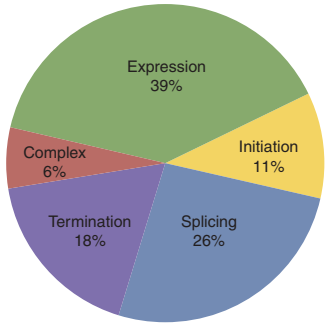

Figure 3 Classification of genes showing expression changes at the exon and/or transcript level. The 324 genes were classified into separate categories depending on the nature of the isoform change occurring: expression changes at the whole transcript level (green), transcription initiation changes (yellow), alternative splicing of a cassette exon (blue), transcription termination changes (purple), and complex changes of multiple event types (red). The percentages shown assume a uniform false-positive rate for all results. To obtain a lower bound for the relative frequency of isoform variants, we have also recalculated the frequencies of the isoform changes (but not whole-gene expression and complex changes) based on our current false positive rate estimate of ∼20% (from validation experiments). Thus, we obtained the following ranges for each of the changes: whole gene expression, 39–44%; initiation, 10–11%; splicing, 24–26%; termination, 16–18%; and complex events, 6–7%.

Table 1 Validation of candidate probe sets

Gene	Probe set	SNP	*P*-value	Chromosomal location	Probe set location	Type of event	RefSeq/EST evidence
CEP192	3779862	rs482360	3.71×10^{-19}	chr18:13047770–13048132	Coding	Intron retention	Yes
ZNF83	3869658	rs1012531	2.72×10^{-10}	chr19:57808794–57808830	Coding	Intron retention	Yes
C17orf57	3724617	rs3760372	5.54×10^{-12}	chr17:42793744–42793848	Coding	Exon skipping	Yes
CAST	2821249	rs7724759	7.17×10^{-16}	chr5:96102207–96102239	Coding	Exon skipping	Yes
CD46	2377476	rs4844390	1.06×10^{-14}	chr1:204329527–204329556	Coding	Exon skipping	Yes
ATP5SL	3863093	rs1043413	9.38×10^{-11}	chr19:46631033–46631167	Coding	Exon skipping	Yes
ERAP2	2821389	rs2255546	8.37×10^{-22}	chr5:96261677–96261705	Coding	Alternative splice site use	Yes
POMZP3	3057764	rs2005354	3.77×10^{-22}	chr7:75892151–75892256	Coding	Exon skipping	Yes
ULK4	2670619	rs1717020	5.99×10^{-11}	chr3:41932478–41932514	Coding	Exon skipping	No
PARP2	3527423	rs2297616	2.81×10^{-37}	chr14:19883099–19883123	Coding	Alternative splice site use	Yes
ATPIF1	2327383	rs2481974	4.26×10^{-11}	chr1:28248451–28248478	Coding	Alternative splice site use	Yes
MRPL43	3303658	rs12241232	1.24×10^{-11}	chr10:102731257–102731290	Coding	Exon skipping, differential stop codon use and 3′ UTR length	Yes
DKFZp451M2119	2588913	rs10930785	1.93×10^{-28}	chr2:178022380–178022482	5′ UTR	Exon skipping	Yes
RNH1	3358076	rs11821392	4.34×10^{-15}	chr11:494826–494888	5′ UTR	Exon skipping	Yes
SNX11	3725089	rs7224014	4.20×10^{-9}	chr17:43543086–43543116	5′ UTR	Exon skipping	Yes
USMG5	3304753	rs7911488	2.66×10^{-24}	chr10:105143981–105144095	5′ UTR	Exon skipping	Yes
SEP15	2421300	rs1407131	7.57×10^{-13}	chr1:87091818–87092018	5′ UTR	Differential 5′ UTR length	Yes
SLC35B3	2941033	rs3799255	2.12×10^{-10}	chr6:8380460–8380572	5′ UTR	Differential 5′ UTR length	Yes
C17orf81	3708382	rs2521985	2.55×10^{-13}	chr17:7100907–7100934	3′ UTR	Exon skipping, differential 3′ UTR length	Yes
ERAP1	2868133	rs7705827	6.09×10^{-19}	chr5:96123330–96124483	3′ UTR	Differential 3′ UTR length	Yes
TAP2	2950168	rs3763355	1.98×10^{-13}	chr6:32897620–32897880	3′ UTR	Alternative splice site use, differential 3′ UTR length	Yes
IRF5	3023264	rs6969930	8.27×10^{-22}	chr7:128183412–128183723	3′ UTR	Differential 3′ UTR length	Yes
PPIL2	3938301	rs5999098	1.46×10^{-12}	chr22:20374916–20375108	3′ UTR	Differential 3′ UTR length	Yes
PTER	3236819	rs1055340	5.25×10^{-18}	chr10:16595519–16595641	3′ UTR	Differential 3′ UTR length	No
WARS2	2430765	rs1325933	3.53×10^{-8}	chr1:119285989–119286236	3′ UTR	Differential 3′ UTR length	Yes

List of candidate probe sets validated by qualitative or quantitative RT-PCR. The gene name and the significant probe set are indicated along with the SNP and *P*-value from the linear regression analysis. The chromosomal location of the probe set is also shown, including its relative location within the gene. The nature of the isoform change is indicated, as is any existing RefSeq or EST evidence of this change.

Transcript alterations within coding regions of the gene, such as the addition or removal of sequences coding for functional domains or the introduction of premature stop codons, may greatly alter the protein sequence, structure and function[22,23]. Changes outside the coding regions can also have wide-ranging regulatory consequences. Differential exon selection within the 5′ and 3′ UTRs may alter mRNA stability and translational efficiency by the addition or removal of regulatory sequences. In some genes (for example, *ATPIF1* and *TAP2*), selection of an alternative splice site for the terminal exon resulted in differential stop codon use and, consequently, changes in the length and composition of the 3′ UTR. Alterations in the 3′ UTR can also be effected by alternative use of polyadenylation sites, and approximately half of human genes are predicted to contain several polyadenylation sites, resulting in transcripts with different 3′ UTR lengths[24,25]. Altering a functional polyadenylation site through a single polymorphism may lead to isoform switching. The 3′ UTR is also involved in post-transcriptional regulation through the targeting of specific UTR sequences by microRNAs (miRNA)[26,27]. Expression of multiple isoforms may be indirectly controlled through the differential expression of miRNAs or by polymorphisms in these miRNA-specific sequences. The end consequence of many of these alterations in the UTRs affects a cascade of downstream processes such as stability, localization and translation efficiency, and it directly contributes to phenotypic diversity and possible disease states. A systematic characterization of the polymorphisms to determine the true causative SNPs resulting in these changes will lead to the possible identification of new regulatory motifs and is currently being undertaken.

Earlier studies suggested that gene expression constituted an important piece of human variation, and although it remains a significant aspect, the added complexity of transcript-processing variations and the potential outcome of these differences greatly alter our earlier perceptions. We estimate that between 50 and 55% of gene expression variation is isoform based. Our results constitute an important change in way we view the effects of common genetic variation in humans and highlight the need for broader investigation into the causes of differential gene expression, as well as previously found and new disease associations that lack clear functional variants.

METHODS

Cell line preparation. We obtained triplicate RNA samples from LCLs derived from the parents of 30 CEPH (CEU) trios (60 individuals) that had been genotyped for approximately 4 million SNPs by the International HapMap Project[17]. Cells were grown at 37 °C and 5% CO_2 in RPMI 1640 medium (Invitrogen) supplemented with 15% (vol/vol) heat-inactivated FCS (Sigma-Aldrich), 2 mM L-glutamine (Invitrogen) and penicillin/streptomycin (Invitrogen). Cell growth was monitored with a hemocytometer and cells were collected at a density of 0.8×10^6 to 1.1×10^6 cells/ml. Cells were then resuspended and lysed in TRIzol reagent (Invitrogen). Three successive growths were performed (corresponding to the second, fourth and sixth passages) after thawing frozen cell aliquots. Three cell lines showed extremely poor growth and were not used in the study, leaving 57 LCLs for subsequent analyses.

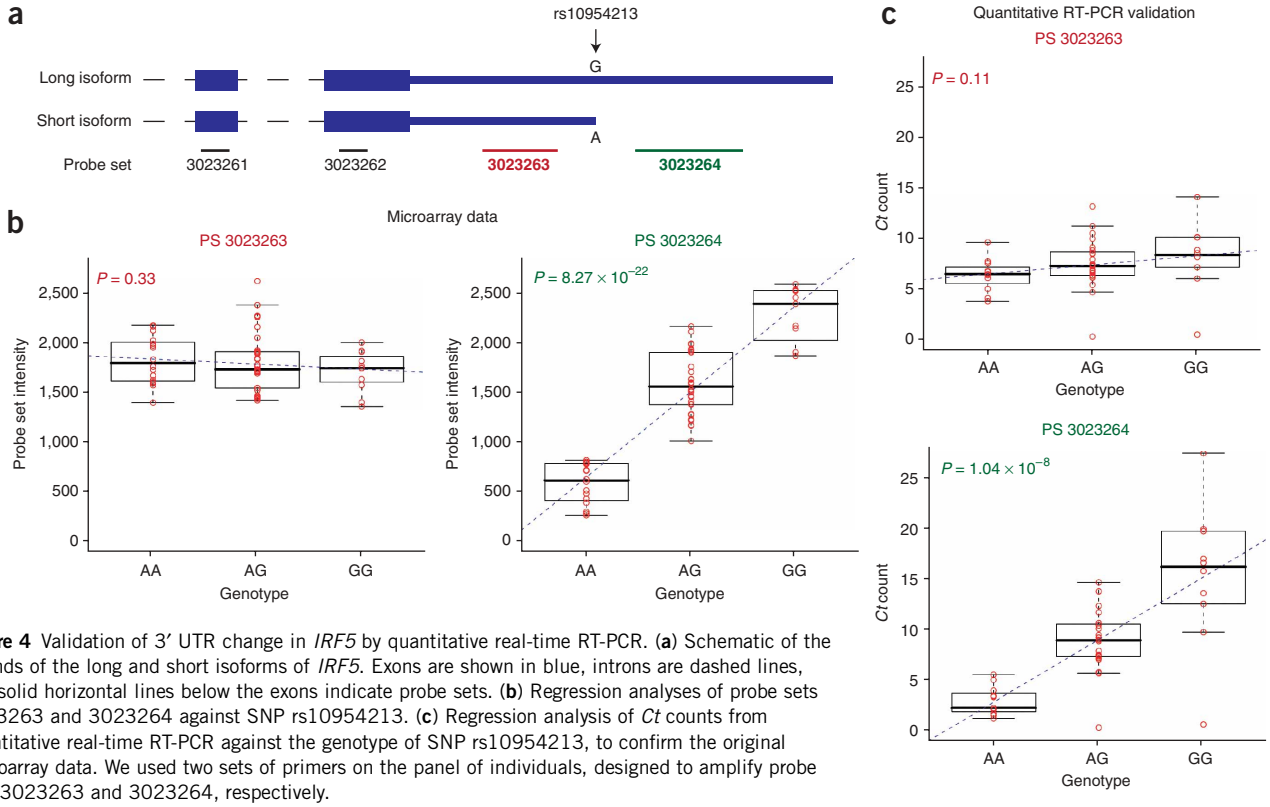

Figure 4 Validation of 3′ UTR change in *IRF5* by quantitative real-time RT-PCR. (**a**) Schematic of the 3′ ends of the long and short isoforms of *IRF5*. Exons are shown in blue, introns are dashed lines, and solid horizontal lines below the exons indicate probe sets. (**b**) Regression analyses of probe sets 3023263 and 3023264 against SNP rs10954213. (**c**) Regression analysis of *Ct* counts from quantitative real-time RT-PCR against the genotype of SNP rs10954213, to confirm the original microarray data. We used two sets of primers on the panel of individuals, designed to amplify probe sets 3023263 and 3023264, respectively.

Affymetrix exon arrays. We isolated RNA using TRIzol reagent following the manufacturer's instructions (Invitrogen) and assessed the RNA quality using RNA 6000 NanoChips with the Agilent 2100 Bioanalyzer (Agilent). Biotin-labeled targets for the microarray experiment were prepared using 1 µg of total RNA. Ribosomal RNA was removed with the RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen) and cDNA was synthesized using the GeneChip WT (Whole Transcript) Sense Target Labeling and Control Reagents kit as described by the manufacturer (Affymetrix). The sense cDNA was then fragmented by uracil DNA glycosylase and apurinic/apyrimidic endonuclease-1 and biotin-labeled with terminal deoxynucleotidyl transferase using the GeneChip WT Terminal labeling kit (Affymetrix). Hybridization was performed using 5 micrograms of biotinylated target, which was incubated with the GeneChip Human Exon 1.0 ST array (Affymetrix) at 45 °C for 16–20 h. After hybridization, nonspecifically bound material was removed by washing and specifically bound target was detected using the GeneChip Hybridization, Wash and Stain kit, and the GeneChip Fluidics Station 450 (Affymetrix). The arrays were scanned using the GeneChip Scanner 3000 7G (Affymetrix) and raw data was extracted from the scanned images and analyzed with the Affymetrix Power Tools software package (Affymetrix).

Preprocessing and analysis of array hybridization data. The Affymetrix Power Tools software package was used to quantile-normalize the probe fluorescence intensities and to summarize the probe set (representing exon expression) and meta–probe set (representing gene expression) intensities using a probe logarithmic-intensity error model (see URLs below). High false-positive rates are common in microarray studies, and previous studies have suggested that a major factor arises from probes overlapping SNPs that result in changes to hybridization intensity[28], potentially influencing the apparent association between the SNP genotype and probe intensities. To reduce potential influences of SNPs on false positives, all probes containing known SNPs (dbSNP release 126) were masked out before summarizing probe set and meta–probe set scores. The presence of unannotated SNPs affecting probe hybridization will remain (see below), but these cannot be detected by any statistical methods except for the impractical solution of resequencing all probes across the panel used in the study. We also filtered probe intensity levels by magnitude of response, removing probes that seemed to be in the background. Probe intensities were extracted for a series of 16,934 antigenomic probes targeted to nonhuman sequences and averaged by their relative G+C content. The threshold for background expression was defined as the average intensity for a given G+C content plus 2 s.d. For any given genomic probe on the array, if the intensity across all samples was below the threshold for the same G+C percentage, then it was considered background and masked from the analysis. In total, 670,809 probes corresponding to core annotated probe sets were masked from the analysis, reducing the number of core probe sets in the analysis to 244,027 probe sets.

Association analysis and multiple test correction. We examined probe set expression levels for association with flanking SNPs. For each of the 244,027 core probe sets and 17,653 meta–probe sets, we tested for association of the expression levels to HapMap phase II (release 21) SNPs with a minor allele frequency of at least 5% within a 50-kb region flanking either side of the gene containing the probe set, using a linear regression model in the R software package. Raw *P*-values were obtained from the regression using the standard asymptotic *t*-statistic.

To correct for testing of associations between multiple probe sets and SNPs, we carried out permutation tests followed by FDR correction. Within each expression-versus-genotype matrix, we randomly permuted the expression values for all probe sets belonging to the same meta–probe set (to preserve the haplotype block structure). For each expression measurement, we computed and retained only the highest asymptotic *P*-value and produced the distribution of maximum *P*-values within the permuted dataset. The maximum asymptotic *P*-values from the experimental data were then converted into empirical *P*-values by mapping onto the permuted distribution. The above procedure corrects for testing multiple SNPs against each expression value. Subsequently, we performed an FDR correction[29] on the empirical *P*-values, to control the FDR across multiple expression values. The procedure was applied separately to measurements at the probe set and meta–probe set levels. We used a 0.05 FDR criterion as a significance cutoff in our analysis. For the sake of clarity, all of the values and cutoffs quoted in the results correspond to the raw, uncorrected *P*-values.

Classification of transcript isoforms. We developed an automated method to categorize the transcriptional and isoform changes. The algorithm first classifies transcripts as expression variants if there is an association of the entire meta–probe set significant at the $P < 6.02 \times 10^{-7}$ level (see above for explanation of the cutoffs). Subsequently, the algorithm identifies all individual probe sets significant at the $P < 9.73 \times 10^{-9}$ level that do not belong to the expression variants detected above. All such significant probe sets are then grouped into blocks corresponding to exons, according to their RefSeq annotation. Each significant block is classified as an initiation, splicing or termination change according to its position within the transcript (3′, internal, or 5′, respectively). Cases with two or more of the above events occurring in a single transcript are classified as complex. Finally, all results were manually curated. To visualize the potential nature of the isoform changes on a gene level, the probe sets were examined in the context of their transcript, mRNA, and EST information. For each gene predicted to have SNP-associated transcript- or exon-level expression changes, we plotted the *P*-values of all the corresponding probe sets and overlaid the fold change expression levels between the two homozygous genotypes for the significant SNP identified in the association analyses (**Supplementary Fig. 2**). We made minor adjustments (23 of 324 events) to the automated classifications, mostly in cases where the designations were not consistent with annotated alternative isoform structures or where the Affymetrix transcript annotation was incorrect.

Validation of transcript isoform changes. Total RNA was treated with 4 U of DNase I (Ambion) for 30 min to remove any remaining genomic DNA. First-strand complementary DNA was synthesized using random hexamers (Invitrogen) and Superscript II reverse transcriptase (Invitrogen). All primers used for RT-PCR reactions (**Supplementary Table 3** online) were designed using Primer3 (ref. 30) software. Candidate probe sets showing association were validated in two ways, depending on their location within the gene. For all probe sets located within coding exons and possessing flanking exons in all known RefSeq isoforms, we designed locus-specific primers within the adjacent flanking exons. Approximately 20 ng of total cDNA was then amplified by PCR using Hot Start Taq Polymerase (Qiagen) with an activation step at 95 °C (15 min) followed by 35 cycles at 95 °C (30 s), 58 °C (30 s) and 72 °C (40 s) and a final extension step at 72 °C (5 min). Amplicons were visualized by electrophoresis on a 2.5% agarose gel.

For probe sets located within 5′ or 3′ untranslated regions or within exons that did not have a flanking exon, we designed a set of primers to amplify the differentially expressed candidate probe set itself. For comparison, other primer pairs were designed to amplify products that corresponded to the adjacent probe sets and were not significantly associated with the same SNP. Total expression measurements were carried out using real-time PCR with Power SYBR Green PCR Master Mix (Applied Biosystems) following the manufacturer's instruction on an ABI 7900HT (Applied Biosystems) instrument. The reaction was set up in 10 µl final volume applying the following conditions: 8 ng of total cDNA and 0.32 µM of gene-specific primers; cycling, 95 °C (15 min) and 95 °C (20 s), 58 °C (30 s), 72 °C (45 s) for 40 cycles. Relative quantification of each amplicon was evaluated on RNA from 57 cell lines in triplicate. For each amplicon, a standard curve was established using dilution series of a mix of cDNA samples with known total cDNA concentration. Human 18S rRNA was also quantified using TaqMan probes as a control for well-to-well normalization (TaqMan Pre-Developed Assay Reagents for Gene Expression – Human 18S rRNA, 4319413E, Applied Biosystems). The cycle threshold (*Ct*) values for each replicate were transformed to relative concentrations using the estimated standard curve function (SDS 2.1, Applied Biosystems) and normalized based on 18S real-time data from the same samples to account for well-to-well variability. The quantitative data was used in regression analyses with the same SNP identified in the original association to confirm the significance, using a *P*-value threshold of $0.05/N$ where N is the number of candidate genes tested using this method. The regression line was required to be in the same direction as the original association. Quantitative RT-PCR of the control probe sets showing no association with the SNP were also required to be nonsignificant at this threshold.

Effect of unannotated SNPs on the analysis. We have previously shown that SNPs located within probes may affect their hybridization to target DNA[16], and have therefore conservatively masked out all probes containing SNPs to circumvent this problem. However, probes containing unannotated SNPs are not accounted for; therefore, we wanted to assess the effect of these unknown SNPs on our analysis. We selected 83 genes, each of which contained only a single significant probe set. Many (63) of these probe sets are supported by a single independent, nonoverlapping probe, and such probe sets are the most susceptible to the effect of SNPs, because every probe could potentially be affected by a single SNP. We sequenced the probe sets from the cell lines of six individuals, three from each of the two homozygous genotypes of the associated SNP. We observed that the sequences for 56 probe sets (67.5%) were identical in all samples tested, suggesting that these are more likely to be true events and not an artifact of one or more SNPs located in the individual probes representing the probe set. In the remaining 27 probe sets (32.5%), we identified previously unknown SNPs or indels overlapping one or more of the probes of the probe set, and in most cases, these polymorphisms segregated with one of the two homozygous sample groups, most likely giving rise to the apparent false-positive hit. We excluded these 27 probe sets from our candidate list presented in the manuscript. All of the remaining candidates are supported by two or more independent probes, and are much less susceptible to the effect of unknown SNPs. Only 2 out of the 32 candidates from the final dataset selected for validation (6%) contained previously unidentified SNPs and hence failed validation, showing that the effect of SNPs on the final results presented here is small.

URLs. Results from regression analyses at the probe set and meta–probe set levels, including gene-level plots of expression changes, and other relevant information can be found at the GRiD (Genetic Regulators in Disease) website (http://www.regulatorygenomics.org). For the probe logarithmic-intensity error model, see http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

Accession codes. US National Center for Biotechnology Information, Gene Expression Omnibus: The data discussed in this publication are accessible through the GEO Series accession number GSE9372.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors would like to thank D. Serre, T. Pastinen, E. Harmsen and H. Zuzan for helpful discussions and D. Sinnett for technical assistance. This work is supported by Genome Canada, Genome Québec and the Canadian Institutes of Health Research (CIHR). T.J.H. is the recipient of a Clinician-Scientist Award in Translational Research by the Burroughs Wellcome Fund and an Investigator Award from CIHR. J.M. is a recipient of a Canada Research Chair.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Kim, H., Klein, R., Majewski, J. & Ott, J. Estimating rates of alternative splicing in mammals and invertebrates. *Nat. Genet.* **36**, 915–917 (2004).
2. Black, D.L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).
3. Faustino, N.A. & Cooper, T.A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
4. Nissim-Rafinia, M. & Kerem, B. The splicing machinery is a genetic modifier of disease severity. *Trends Genet.* **21**, 480–483 (2005).
5. Zielenski, J. Genotype and phenotype in cystic fibrosis. *Respiration* **67**, 117–133 (2000).
6. Field, L.L. *et al. OAS1* splice site polymorphism controlling antiviral enzyme activity influences susceptibility to type 1 diabetes. *Diabetes* **54**, 1588–1591 (2005).
7. Qu, H.Q. *et al.* Genetic control of alternative splicing in the *TAP2* gene: possible implication in the genetics of type 1 diabetes. *Diabetes* **56**, 270–275 (2007).
8. Cunninghame Graham, D.S. *et al.* Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. *Hum. Mol. Genet.* **16**, 579–591 (2007).
9. Graham, R.R. *et al.* Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. USA* **104**, 6758–6763 (2007).
10. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
11. Spielman, R.S. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* **39**, 226–231 (2007).

12. Stranger, B.E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).
13. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
14. Clark, T.A. *et al.* Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.* **8**, R64 (2007).
15. Gardina, P.J. *et al.* Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7**, 325 (2006).
16. Kwan, T. *et al.* Heritability of alternative splicing in the human genome. *Genome Res.* **17**, 1210–1218 (2007).
17. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
18. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
19. Alberts, R. *et al.* Sequence polymorphisms cause many false *cis* eQTLs. *PLoS ONE* **2**, e622 (2007).
20. Nembaware, V., Wolfe, K.H., Bettoni, F., Kelso, J. & Seoighe, C. Allele-specific transcript isoforms in human. *FEBS Lett.* **577**, 233–238 (2004).
21. Hull, J. *et al.* Identification of common genetic variation that modulates alternative splicing. *PLoS Genet* **3**, e99 (2007).
22. Liu, S. & Altman, R.B. Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res.* **31**, 4828–4835 (2003).
23. Lewis, B.P., Green, R.E. & Brenner, S.E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. USA* **100**, 189–192 (2003).
24. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
25. Yan, J. & Marr, T.G. Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res.* **15**, 369–375 (2005).
26. Valencia-Sanchez, M.A., Liu, J., Hannon, G.J. & Parker, R. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.* **20**, 515–524 (2006).
27. Wu, L., Fan, J. & Belasco, J.G. MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl. Acad. Sci. USA* **103**, 4034–4039 (2006).
28. Naef, F. & Magnasco, M.O. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E* **68**, 011906 (2003).
29. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
30. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).