Haplotype and diplotype analyses of variation in ERCC5 transcription cis-regulation in normal bronchial epithelial cells

Xiaolu Zhang,1 Erin L. Crawford,1 Thomas M. Blomquist,2 Sadik A. Khuder,3 Jiyoun Yeo,1 Albert M. Levin,4 and James C. Willey1,2
1Division of Pulmonary/Critical Care and Sleep Medicine, Department of Medicine, University of Toledo Health Sciences Campus, Toledo, Ohio; 2Department of Pathology, University of Toledo Health Sciences Campus, Toledo, Ohio; 3Departments of Medicine and Public Health and Homeland Security, University of Toledo Health Sciences Campus, Toledo, Ohio; and 4Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan

Submitted 25 February 2016; accepted in final form 26 May 2016

Zhang X, Crawford EL, Blomquist TM, Khuder SA, Yeo J, Levin AM, Willey JC. Haplotype and diplotype analyses of variation in ERCC5 transcription cis-regulation in normal bronchial epithelial cells. Physiol Genomics 48: 537–543, 2016. First published May 27, 2016; doi:10.1152/physiolgenomics.00021.2016.—Excision repair cross-complementation group 5 (ERCC5) gene plays an important role in nucleotide excision repair, and dysregulation of ERCC5 is associated with increased lung cancer risk. Haplotype and diplotype analyses were conducted in normal bronchial epithelial cells (NBEC) to better understand mechanisms responsible for interindividual variation in transcript abundance regulation of ERCC5. We determined genotypes at putative ERCC5 cis-regulatory SNPs (cis-rSNP) rs751402 and rs2296147, and marker SNPs rs1047768 and rs17655. ERCC5 allele-specific transcript abundance was assessed by a recently developed targeted sequencing method. Syntetic relationships among alleles at rs751402, rs2296147, and rs1047768 were assessed by allele-specific PCR followed by Sanger sequencing. We then assessed association of ERCC5 allele-specific expression at rs1047768 with haplotype and diplotype structure at cis-rSNPs rs751402 and rs2296147. Genotype analysis revealed significantly (P < 0.005) higher interindividual variation in allelic ratios in cDNA samples relative to matched gDNA samples at both rs1047768 and rs17655. By diplotype analysis, mean expression was higher at the rs1047768 alleles syntenic with rs2296147 T allele compared with rs2296147 C allele. Furthermore, mean expression was lower at rs17655 C allele, which is syntenic with G allele at a linked SNP rs873601 (D′ = 0.95). These data support the conclusions that in NBEC, T allele at SNP rs2296147 upregulates ERCC5, variation at rs751402 does not alter ERCC5 regulation, and that C allele at SNP rs17655 downregulates ERCC5. Variation in ERCC5 transcript abundance associated with allelic variation at these SNPs could result in variation in NER function in NBEC and lung cancer risk.

lungen cancer; allele-specific expression; single-nucleotide polymorphism; RNA-sequencing

Excision repair cross-complementation group 5 (ERCC5) gene, also known as Xeroderma Pigmentosum complementation group G (XPG) (26), plays an important role in nucleotide excision repair (NER). In addition, ERCC5 is among a set of key antioxidant, DNA repair, and cell cycle control genes identified by this laboratory to associate with lung cancer risk (2, 22). Furthermore, variation in ERCC5 regulation is reported to be associated with treatment response and outcome in bronchogenic carcinoma as well as other cancers (21, 30, 31, 36, 38).

Known transcription regulators of ERCC5 in normal bronchial epithelial cells (NBEC) include CCAAT/enhancer binding protein gamma (CEBPG), E2F Transcription Factor 1 (E2F1), and YY1 (a transcription factor belonging to the GLI-Kruppel class of zinc finger proteins) (8, 22). CEBPG is a truncated C/EBP isofrom that lacks a transcription activation domain and therefore functions through heterodimerization with other C/EBP members (32). Knockout of CEBPG or its binding partner CEBPA results in emphysema, a condition associated with lung cancer risk (9, 15). E2F1 is a critical regulator of cell cycle progression (24). A study in a series of 58 lung tumors of all histological types supported a pivotal role of E2F1 in tumorigenesis (11).

The common single nucleotide polymorphic (SNP) sites rs751402 and rs2296147 reside in the ERCC5 5′-untranslated region (UTR), rs751402 within a known CEBPG binding site based on chromatin immunoprecipitation studies (33) and rs2296147 within an experimentally confirmed binding site for E2F1 and YY1 (8). Variation at rs2296147 is predicted to alter binding of the TP53 transcription factor (19). Both rs751402 and rs2296147 are associated with lung cancer risk in molecular epidemiologic studies (29, 38).

In a previous study, based on genotype analysis we found that SNPs rs751402 and rs2296147 were associated with interindividual variation in allelic imbalance in ERCC5 expression in NBEC (6). This observation suggests that one or both of these SNPs affects ERCC5 cis-regulation. However, based on genotyping analysis it was not possible to sort out with confidence the independent roles of rs751402 and rs2296147. Furthermore, the patterns of allelic imbalance variation observed indicated that one or more cis-regulatory SNPs in addition to rs751402 and rs2296147 also played a role. One candidate is rs17655, a common polymorphic site in the ERCC5 3′-UTR reported to be associated with lung cancer risk (20). Genetic variation in the 3′-UTR of a gene can play a role in cis-regulation by influencing microRNA (miRNA) binding activity (25, 28, 35).

In an effort to better characterize putative cis-acting genetic variants that serve as expression quantitative trait loci (eQTL) responsible for interindividual variation in ERCC5 transcription regulation, we used recently developed methods to assess in more detail the role of the previously studied 3′-UTR sites rs751402 and rs2296147 (2, 22) and additional 3′-UTR site rs17655. Specifically, using allele-specific PCR amplicon libraries prepared for next-generation sequencing (NGS) accord­
ing to a recently described method (5), haplotype and diplotype structure of the ERCC5 5′-UTR region containing rs751402 and rs2296147 were assessed by allele-specific polymerase chain reaction (PCR) followed by direct sequencing. We determined allele-specific expression (ASE) as a measurement of allelic ratio at the marker site rs1047768 in the ERCC5 coding region close to 5′-UTR and at rs17655 in the coding region nearby 3′-UTR. We then evaluated the association of ASE with each rs751402-rs2296147-rs1047768 haplotype and diplotype.

MATERIALS AND METHODS

Study subjects. Bronchoscopic brush biopsy samples of NBEC and matched peripheral blood samples were obtained as previously described (22) from 80 subjects: 60 subjects without cancer and 20 subjects with lung cancer. Demographic characteristics of these newly recruited subjects used in current study are presented in Supplementary Table S1.1 This study was conducted under University of Toledo Institutional Review Board-approved protocol #106894. All the subjects in current study were subset of the Lung Cancer Risk Test trial, a prospective cohort study (7). Individuals were recruited at 13 locations in the United States and provided informed consent to participate. Each subject agreed to the banking of residual nucleic acids for use in future studies under University of Toledo Biomedical Institutional Review Board protocol #108538. Then all samples were de-identified, and links to identifying information were kept at each site for subjects recruited at that site. Inclusion criteria required subjects to be at high demographic risk for lung cancer based on age (50–90 yr) and smoking history (≥ 20 pack-yr). Both current and former smokers were eligible. Subjects had to have been without a diagnosis of lung cancer prior to or at enrollment. Subjects were excluded if they were previously diagnosed or treated for lung cancer or had a high pretax likelihood of lung cancer, if they were positive for hepatitis B or C or human immunodeficiency virus or had active tuberculosis or if the physician deemed them to be medically inappropriate for safety concerns. Also excluded were children, pregnant women, prisoners, mentally disabled, those that had received a double lung transplant, radiation, or chemotherapy of any kind within the last month, and those scheduled to receive either radiation or chemotherapy.

DNA and RNA extraction. NBEC samples obtained at bronchoscopic brush biopsy were shipped to ResearchDX (Irvine, CA), where RNA was extracted, treated with DNase I (Qiagen, Valencia, CA) to eliminate contaminating genomic DNA (gDNA), and frozen in aliquots. One aliquot of each frozen RNA sample was shipped to the University of Toledo and tested for gDNA contamination with a pair of primers designed to span an intron-exon junction in Secretoglobin, family 1A, member 1 gene (Qiagen, Valencia, CA) to eliminate contaminating genomic DNA (gDNA), and frozen in aliquots. One aliquot of each frozen RNA sample was shipped to the University of Toledo and tested for gDNA contamination with a pair of primers designed to span an intron-exon junction in Secretoglobin, family 1A, member 1 gene and thereby amplify only gDNA (4). Total RNA was reverse-transcribed to cDNA using Moloney murine leukemia virus reverse transcriptase and oligo-dT primers as described previously (22). Matched gDNA was extracted from whole blood using FlexiGene DNA Kit (Qiagen) according to the manufacturer’s protocol.

Genotyping and allelotyping. Genotype at each polymorphic site was determined by TaqMan SNP genotyping assay (Applied Biosystems) according to the manufacturer’s protocol. Direct assessment of the syntenic relationship of alleles at rs751402, rs2296147, and rs1047768 in individuals homozygous for rs1047768 was accomplished by allele-specific PCR amplification followed by Sanger sequencing (The University of Michigan DNA Sequencing Core, Ann Arbor, MI) as described previously (6). An overview of polymorphic sites and primers relative to ERCC5 gene coordinates is depicted in Fig. 1. The sequences and design of allele-specific primers were described previously (6).

It was not possible to conduct analysis in 11 of the 80 samples. In three samples, there was no amplification due to poor quality of gDNA and cDNA; four subjects were heterozygous at rs751402 but homozygous at rs2296147, and the spanned region between rs751402 and rs1047768 was too long to assess synteny by direct sequencing; four samples were not amplified with primers determining rs2296147-rs1047768 synteny, likely due to variation in the transcription start site as previously reported (6).

Measurement of ERCC5 allele-specific and total expression. ERCC5 allelic imbalance was measured in cDNA and matched gDNA samples from each subject by a modified version of previously described methods (5). Allelic imbalance in gDNA was measured as a control for possible allelic difference in PCR efficiency and/or genomic copy number variation. Using the gDNA allelic ratio as an internal control is a robust design to control for these confounding factors (14). In brief, a custom, multiplex competitive PCR amplicon library was prepared for targeted NGS, then the library was sequenced at the University of Michigan DNA Sequencing Core using the Sequenom MassArray 2000 platform. To prepare the library, cDNA or matched gDNA samples from each subject was combined with a mixture containing 1 primers spanning SNP rs1047768 in the ERCC5 5′-UTR region, SNP rs17655 in the 3′-UTR region, and the ACTB loading control gene, and 2) a known number of internal standard molecules for each of these targets. Each primer was designed with a universal tail sequence (similar to that used for arrayed primer extension: APEX-2) not present in the human genome to allow for multitemplate PCR addition of barcode and platform-specific sequencing adapters. The internal standards mixture was prepared as described previously (5).

Three sequential PCR amplifications were conducted to prepare the library prior to sequencing. In the first reaction each target sequence was amplified with 5 μM APEX-tailed primers using an air thermal cycle (RapidCycler; Idaho Technology, Idaho Falls, ID). PCR conditions were 95°C/3 min (Tsq DNA polymerase activation); 35 cycles of 94°C/5 s (denaturation), 58°C/10 s (annealing), 72°C/15 s (extension). Each product from the first PCR was purified using QIAquick PCR purification kit (Qiagen) to remove residual primers, and primer dimers then used as template for barcoding PCR. Each barcoding reaction was cycled in the air cycle under the following conditions: 95°C/3 min (Tsq DNA polymerase activation); 15 cycles of 94°C/5 s (denaturation), 58°C/10 s (annealing), 72°C/15 s (extension). The final concentration of each forward and reverse barcoding primer was 1 μM. The third PCR for adding Illumina platform-specific adapters was conducted with the same PCR conditions and primer concentrations as the barcoding PCR. Representative PCR products were checked for quality and quantity with Bioanalyzer 2100 (Agilent Technologies) following each of the three amplifications. All products from the third-step PCR were combined at equal volumes and then purified by QIAquick PCR purification kit (Qiagen) to remove residual primers. The concentration of purified products was checked by Bioanalyzer 2100 and then sent for sequencing. ASE was presented as allelic ratio, which was calculated as the ratio of sequencing counts for each allele and filtered as described below.

Data processing pipeline. The University of Michigan Illumina Sequencing services provided raw sequencing data in FASTQ format. Practical Extraction and Reporting Language (PERL) scripts were used to combine Read 1 (forward) and Read 2 (reverse) sequence reads for each template sequenced. These “joined” reads were then de-multiplexed based on dual-index barcoding on each template, and the locus was identified based on the region representing the primer sequences. Intervening amplicon sequence that was “captured” between the primer sequences was aligned using custom alignment with Approximate String matching algorithm as previously described (3, 5). These alignment calls then provided relative abundance in the form of sequence “counts” for each allele at each locus. The ratio of these

1 The online version of this article contains supplemental material.
allele-specific sequence counts at each locus then represents the allele-specific expression ratio.

Filtering for stochastic sampling error. To control for stochastic sampling error, we implemented a previously developed equation that identifies the minimum acceptable input of target gene molecules into library preparation, obtained by measurement relative to known number of input IS and minimum allowable number of amplicons from library loaded into sequencer measured as sequencing counts (3).

A filter for stochastic sampling-determined analytical variation was applied to both allele-specific expression and total transcript abundance. Only values with stochastic sampling-dependent coefficient of variation (CV) expected to be ≤10% were subjected to subsequent analysis. Total transcript abundance was presented as target gene NT molecules/10^6 ACTB molecules.

Statistical analysis. Ratios of allele-specific expression values from individuals heterozygous at the marker SNP were normalized to the mean allelic ratio in genomic DNA samples and then log2-transformed prior to analysis. Variance in allelic ratio measured in cDNA samples was tested for difference from that in matched gDNA samples by Fisher’s distribution test (F-test). Difference in mean allelic ratio between cDNA samples and matched gDNA samples was determined by paired sample t-test. One-way analysis of variance (ANOVA) was used to compare the mean of allelic ratios associated with genotype or diplotype. All statistical tests were two-sided with a statistical significance level of P < 0.05, using either the R statistical programming language (v. 3.2.0) or SAS program (v. 9.3). All graphs were plotted using GraphPad Prism 6.

RESULTS

Interindividual variation in allelic imbalance at 5’-UTR site rs1047768. ERCC5 allele-specific transcript expression was measured as allelic ratio in NBEC cDNA samples from individuals heterozygous at the rs1047768 marker polymorphic site located in the ERCC5 5’-UTR. For analysis of allelic imbalance at rs1047768, data from the 80 newly enrolled subjects were combined with data from 81 previously enrolled subjects (6). Among the newly enrolled subjects, 33 individuals were heterozygous at the marker SNP rs1047768, while among the previously enrolled subjects 22 were heterozygous (6).

Fig. 1. Schematic overview of ERCC5 gene putative cis-regulatory polymorphic sites and orientation for allele-specific expression measurement. ERCC5 gene coordinate represents National Center for Biotechnology Information (NCBI) Gene NC_000013.11. All positions noted here are relative to the reported NCBI mRNA RefSeq TSS for ERCC5 NM_000123.3. Gray or white arrowhead indicates the direction of transcription relative to gene orientation. TSS, transcript start site. A: the syntenic relationship of alleles at rs751402, rs2296147, and rs1047768 in individuals heterozygous for rs1047768 was assessed by allele-specific PCR amplification followed by direct sequencing. C allele Reverse Primer, a C allele-specific primer used in combination with Forward Primer for specific amplification from C allele at rs2296147 to determine synteny between alleles at rs2296147 and rs751402. T allele Forward Primer and C allele Forward Primer were used in combination with Reverse Primer in exon 2 for specific amplification from T allele or C allele, respectively, at rs2296147 to determine the synteny between alleles at rs2296147 and rs1047768. cDNA instead of gDNA was used for this amplification to avoid a large intron 1. The depicted cluster region of TSS represents highly variable ERCC5 transcription initiation sites as discussed previously (6). B: the allele-specific expression of ERCC5 was measured at 2 polymorphic sites, rs1047768 and rs17655. Generally, native templates in cDNA sample were amplified after mixture with a known number of internal standard molecules for each respective native template. Each internal standard contained 6 nucleotides altered relative to the respective native template but identical priming sites and amplified with the same efficiency as the native template. Barcodes that allowed for multiplexing and adapters specific for Illumina HiSeq platform were added by PCR. The products were quantified and purified then sent for sequencing on Illumina HiSeq. ***Nucleotide alteration in internal standard (IS) relative to native template (NT).
Fig. 2. Allelic ratios measured at rs1047768 and rs17655. The base 2 of logarithm transformation was applied to allelic ratios, which were normalized to the mean allelic ratio in gDNA (allelic ratio/average of gDNA allelic ratios) for each of the 2 polymorphic sites, rs1047768 and rs17655, in cDNA and matched gDNA samples and used for statistic tests. The dashed line at 0 is reference line for allelic ratio of 1.

A: interindividual variation in T/C allelic ratios measured at polymorphic site rs1047768 located in ERCC5 coding region exon 2 is significantly higher in cDNA samples relative to matched gDNA controls (F-test, \( P < 0.0001 \)). The mean log2(T/C ratio) in cDNA (M = 0.11, SD = 0.34) was not significantly different from that for matched gDNA (M = 0.03, SD = 0.11) according to t-test (\( P = 0.1500 \)). B: similarly to rs1047768, significantly higher interindividual variation in allelic ratios in cDNA compared with matched gDNA (F-test, \( P = 0.0005 \)) was also observed at polymorphic site rs17655 located in ERCC5 exon 15 close to 3′-UTR. The mean log2(G/C ratio) is significantly higher in cDNA (M = 0.25, SD = 0.35) than in gDNA (M = −0.14, SD = 0.16) (t-test, \( P < 0.0001 \)). M, mean; SD, standard deviation.

**Table 1.** Summary of haplotype structures in ERCC5 5′-UTR region from 70 newly recruited subjects and 80 previously enrolled subjects

<table>
<thead>
<tr>
<th></th>
<th>rs751402</th>
<th>rs2296147</th>
<th>rs1047768</th>
<th>Haplotype Count</th>
<th>Frequency, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>112</td>
<td>37</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>63</td>
<td>21</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>55</td>
<td>18</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>51</td>
<td>17</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

**DISCUSSION**

Known sources of inherited risk for lung cancer include variation in cis-acting regulatory single nucleotide polymorphisms (cis-rSNPs) and/or key transcription factors that regulate antioxidant, DNA repair, and cell proliferation control...
Table 2. Summary of diplotype structures in ERCC5 5'-UTR region from subjects heterozygous at rs1047768 site

<table>
<thead>
<tr>
<th>Parental Chromosome 1</th>
<th>Parental Chromosome 2</th>
<th>Diploptype Count</th>
<th>Frequency, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-T-T</td>
<td>G-C-C</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>A-T-T</td>
<td>A-T-C</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>G-C-T</td>
<td>G-C-C</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>G-C-T</td>
<td>G-T-C</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>G-T-T</td>
<td>G-C-C</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>G-T-T</td>
<td>G-T-C</td>
<td>9</td>
<td>20</td>
</tr>
</tbody>
</table>

n = 44.

Fig. 3. Allelic ratios measured at rs1047768 sorted by various diplotype. ANOVA was used to assess the difference in T/C allelic ratios among groups. All effects were statistically significant at the 0.05 significance level. Allelic ratios in relationship with 6 presented diplotypes at rs751402, rs2296147, and rs1047768 demonstrated that higher abundance of transcript from rs1047768 marker site C or T allele was associated with T allele at putative cis-regulatory SNP rs2296147 and not associated with variation at rs751402 (Fig. 3). Notably, rs2296147 T allele participates in formation of an in silico predicted TP53 transcription factor-binding site (19) and that site is predicted to be lost when C allele is present. In previous studies TP53 upregulates ERCC5 transcription (17). Therefore, it is reasonable to hypothesize that TP53 upregulates ERCC5 transcription more effectively when T allele is present at rs2296147. In contrast to strong evidence for the cis-regulatory role of rs2296147 in ERCC5 regulation, haplotype and diplotype data do not support a similar role for rs751402. Haplotype-based analyses presented here demonstrated that rs2296147 acts as an eQTL, while rs751402 does not. This is new information that it was not possible to obtain based only on genotype-based analyses (6).

Fig. 4. Increased lung cancer risk through suboptimal normal bronchial epithelial (NBEC) regulation of protective genes. This schematic indicates the putative genetic basis for hereditary increased lung cancer risk in 3 individuals. Single nucleotide polymorphisms (SNPs) that affect transcript abundance regulation are indicated numerically and as diamonds (trans-regulatory SNPs) or circles (cis-regulatory SNPs). As indicated, each individual is at increased risk due to suboptimal regulation of a different combination of genes. Furthermore, when the same gene is suboptimally regulated in multiple individuals (e.g., gene C in individuals 1 and 3), a different set of SNPs may be responsible in each individual.
tional questions remain, some brought into focus by these studies. For example, the data reported here support the hypothesis that rs2296147 and rs17655 or SNPs linked to them influence transcription of ERCC5. A testable hypothesis to explain these results is that there is (1) higher TP53-mediated ERCC5 transcription rate from rs2296147 T allele and (2) higher miRNA-mediated ERCC5 transcript degradation at rs873601 G allele. In addition, we observed that mean T/C ratio at marker SNP rs1047768 and mean G/C ratio at SNP rs17655 were significantly higher than 1.0. However, we also observed significant variation around the mean allelic ratio at each marker SNP. This raises a question because, if rs2296147 and rs873601 each acted as independent cis-regulatory variants without any contribution from the other (for example, in regulation of separate alternative transcripts) or from any other cis-acting SNP, we would expect very little interindividual variation around these mean ratios. One likely explanation is that the predominantly expressed ERCC5 transcripts incorporate both marker SNPs (rs1047768 and rs17655). Because the putative cis-regulatory sites (rs2296147 and rs873601) are unlinked, another testable hypothesis is that the variation in allelic ratio measured at each marker SNP is due to variation in structure of haplotypes comprising these sites. Testing this hypothesis will require long-range allele-specific PCR.

**Additional possible mechanisms for ERCC5 cis-regulation.**

Although the variants assessed in this study likely are major factors in cis-regulation of ERCC5, variants not included in this study also may play a role. This might include SNPs that have a large biological effect but are rare in the population or common SNPs that have less biological effect. Identification of either will require analysis of a larger cohort. As possible additional mechanisms, the cis-regulatory sites rs751402 and rs2296147 assessed in present study are linked to intronic polymorphic sites in the ERCC5 gene (34). These intronic variants might affect different splicing regulatory elements, leading to aberrant allele-specific splicing of ERCC5 pre-miRNA (10, 23, 27).

**Value of transcript abundance regulation as intermediate lung cancer risk marker.** Consistent with a complex genetic mechanism of lung cancer risk, the effect size of each DNA variant associated with lung cancer risk is very small. Consequently, thousands of subjects are needed to directly assess the association of individual genetic variants and lung cancer risk (16). The data presented here support the conclusion that inherited variation in gene regulation is a powerful intermediate phenotypic marker for lung cancer risk, as presented schematically in Fig. 4. As we report here and previously (6), it is possible to assess this type of intermediate risk factor with far fewer patients than the thousands typically necessary for a genome-wide association study aiming to determine association of each individual SNP with risk (1). Specifically, the association of a single genetic variant with transcription regulation (e.g., rs2296147 with ERCC5 regulation) or the association of interindividual variation in transcript abundance pattern with lung cancer risk may be assessed with hundreds of subjects (2). For example, starting with 161 subjects we observed significant association of rs2296147 genotype with ERCC5 ASE (Fig. 3), and with fewer than 100 subjects we observed significantly altered ERCC5 regulation with lung cancer (22) (data not shown). In contrast, there was not a clear association of rs2296147 T allele dosage with lung cancer risk among the subjects enrolled for this study (data not shown).

**Conclusion**

Based on the findings in the current study, we conclude that rs2296147 is an eQTL for ERCC5 and that the T allele at rs2296147 is associated with higher ERCC5 transcript abundance, possibly through increased responsiveness to TP53 transcription factor. Genotype at rs17655 also is associated with variation in ERCC5 transcript abundance, likely due to the effect on miRNA binding affinity at the linked SNP rs873601. These effects on ERCC5 transcription likely result in variation in nucleotide excision DNA repair function. These findings provide a plausible explanation for the association of genotype at rs2296147 and rs17655 with lung cancer risk.

**GRANTS**

Significant portions of this work were funded by National Institutes of Health Grants RC2-CA-148572, IMAT R21-CA-138397, and RO1-HL-108016 and the University of Toledo Medical Center George Isaac Cancer Research Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**DISCLOSURES**

J. C. Willey has 5–10% equity interest in and serves as a consultant to Accugenomics, Inc. Technology relevant to this manuscript was developed and patented by J. C. Willey, and is licensed to Accugenomics. These relationships do not alter our adherence to all Physiological Genomics policies on sharing data and materials.

**AUTHOR CONTRIBUTIONS**

X.Z., T.M.B., and J.C.W. conception and design of research; X.Z. and J.Y. performed experiments; X.Z., S.A.K., A.M.L., and J.C.W. analyzed data; X.Z., E.L.C., S.A.K., and J.C.W. edited and revised manuscript; X.Z. and J.C.W. approved final version of manuscript.

**REFERENCES**

6. Blomquist TM, Crawford EL, Willey JC. cis-acting genetic variation at an E2F1/YY1 response site and putative p53 site is associated with altered