# Allelic mRNA expression imbalance in C-type lectins reveals a frequent regulatory SNP in the human surfactant protein A (SP-A) gene

**Abul K. Azad**[1], **Amanda Curtis**[2], **Audrey Papp**[2], **Amy Webb**[2], **Daren Knoell**[3], **Wolfgang Sadee**[2,4], and **Larry S. Schlesinger**[1,4]

[1]Department of Microbial Infection and Immunity, Center for Microbial Interface Biology, College of Medicine, The Ohio State University

[2]Program in Pharmacogenomics, Departments of Pharmacology, Psychiatry, and Medical Genetics, College of Public Health, The Ohio State University

[3]College of Pharmacy and Davis Heart and Lung Research Institute, The Ohio State University

## Abstract

Genetic variation in C-type lectins influences infectious disease susceptibility but remains poorly understood. We employed allelic mRNA expression imbalance (AEI) technology for *SP-A1, SP-A2, SP-D, DC-SIGN, MRC1,* and *Dectin-1,* expressed in human macrophages and/or lung tissues. Frequent AEI, an indicator of regulatory polymorphisms, was observed in *SP-A2, SP-D*, and *DC-SIGN*. AEI was measured for *SP-A2* in 38 lung tissues using four marker SNPs and was confirmed by next generation sequencing of one lung RNA sample. Genomic DNA at the *SP-A2* DNA locus was sequenced by Ion Torrent technology in 16 samples. Correlation analysis of genotypes with AEI identified a haplotype block, and, specifically, the intronic SNP rs1650232 (30% MAF); the only variant consistently associated with an approximately two-fold change in mRNA allelic expression. Previously shown to alter a NAGNAG splice acceptor site with likely effects on SP-A2 expression, rs1650232 generates an alternative splice variant with three additional bases at the start of exon 3. Validated as a regulatory variant, rs1650232 is in partial LD with known *SP-A2* marker SNPs previously associated with risk for respiratory diseases including tuberculosis. Applying functional DNA variants in clinical association studies, rather than marker SNPs, will advance our understanding of genetic susceptibility to infectious diseases.

## Keywords

C-type lectin; marker SNP; regulatory variant; *SP-A2*; allelic expression imbalance

---

[4]These authors contributed equally to this work.

## INTRODUCTION

There is ample pervasive evidence that host genetics play a major role in infectious diseases [1–3]. To date, hundreds of genetic variants have been associated with susceptibility or resistance to a variety of infectious agents including bacteria, viruses and parasites. Genes encoding C-type lectins, a subset of innate immune molecules involved in pathogen-host cell interactions, have been implicated in susceptibility to tuberculosis [4], pneumonia, respiratory syncytial virus infection, and other infectious diseases. However, the majority of published association studies suffer from lack of mechanistic or functional validation of candidate variants that have been implicated by statistical means [4]. The present study focuses on identifying causative variants in C-type lectin genes.

The major human C-type ($Ca^{2+}$-dependent) lectins include cell membrane-bound receptors such as the macrophage mannose receptor (MR or MRC1), dendritic cell-specific ICAM-3-grabbing non-integrin (DC-SIGN) and Dectin-1, and soluble collectins (collagen like-containing C-type lectins) such as lung surfactant proteins A and D (SP-A and SP-D) [reviewed in [5]]. MRC1, DC-SIGN, and Dectin-1 are also known as macrophage pattern recognition receptors (PRRs) which bind microbial carbohydrate ligands in a $Ca^{2+}$-dependent manner and act as both cell adhesion and phagocytic receptors serving at the interface between host and pathogens [6]. These PRRs and the soluble collectins constitute an important subset of host innate immune molecules relevant to many infectious and non-infectious diseases.

Regulation of gene expression is a main path guiding the evolution of both pathogen and host. We therefore test here the hypothesis that regulatory polymorphisms have risen to high frequency in a key host gene family influencing response to infectious agents. The search for such regulatory variants that can exist anywhere in a gene locus has proven difficult. Developed recently, analysis of differential mRNA expression of two alleles in a candidate gene has emerged as a robust method to discover the presence of regulatory variants in relevant cells and tissues. Allelic RNA expression imbalance (AEI) can be determined by accurately measuring the ratios of the two alleles in both genomic DNA and mRNA, using frequent, heterozygous, single nucleotide polymorphisms (SNPs) as markers. Defined by RNA allelic ratios deviating from measured DNA ratios, AEI is an indicator of regulatory polymorphisms residing within the same gene locus [7–9]. Such *cis*-acting regulatory polymorphisms can then be identified by scanning the gene locus for causative SNPs in multiple target tissues, with AEI ratios as the phenotype [10]. Results from our laboratories show that these types of variants can be unexpectedly frequent and can be involved in disease susceptibility and treatment outcomes [9–14]. Use of functionally validated variants, rather than marker SNPs of unknown effect on the gene's function, can guide the search for true associations with diseases in clinical studies.

In this study, we used the AEI approach to screen several C-type lectin genes in search of *cis*-acting regulatory variants, identifying frequent and robust AEI in three of six genes studied. Detailed SNP scanning of the gene locus of *SP-A2* showing a promising AEI profile in lung tissues identified the likely main regulatory polymorphism that can be linked to previous clinical association studies of infectious diseases using only marker SNPs. This

work provides evidence that the AEI method represents a new platform for identifying frequently occurring functional polymorphisms, broadly applicable to genes involved in the pathogenesis of infectious and non-infectious diseases.

## RESULTS

### C-type lectin expression in human macrophages and lung tissue

Since mRNA expression and processing are tissue-specific events, mRNA measurements including allelic mRNA ratio analysis must be performed in target cells or tissues where mRNA is robustly expressed. We used real-time PCR to measure targeted gene mRNA expression (in the form of cDNA) from 20 human monocyte-derived macrophage (MDM) samples and 30 lung tissue samples. Transcript expression differed substantially from donor to donor, ranging from 4–20-fold within each gene. Expression of C-type lectins in MDMs stimulated with IL-4 was: DC-SIGN  MRC1 > Dectin-1. SP-A1 and SP-A2 were at the limits of detection by real-time PCR, consistent with very low or no expression of these proteins by macrophages. Rank ordered transcript expression of C-type lectins in lung tissue was SP-A2> SP-D > SP-A1> MRC1> Dectin> DC-SIGN.

### Genotyping of Marker SNPs

We genotyped 20 human MDM and 30 lung tissue samples for *MRC1*, *DC-SIGN*, *Dectin-1*, *SP-A1*, *SP-A2,* and *SP-D* marker SNPs (Table 1). Samples heterozygous at any one of the selected marker SNPs were used for allelic mRNA expression assays.

### Allelic mRNA expression in *DC-SIGN*, *SP-A2* and *SP-D*

Representative tracings of the SNaPshot assay performed in lung tissues heterozygous for the marker SNPs are shown in Figure 1A, documenting substantial differences in allelic ratios of the mRNAs and respective gDNA ratios in *DC-SIGN*, *SP-A2*, and *SP-D,* which demonstrate the presence of AEI. None of the lectins we studied showed AEI in MDM samples, while lung tissue samples lacked AEI for *SP-A1*, *MRC1* and *Dectin-1* (data not shown). Allelic mRNA expression ratios (corrected for allelic gDNA ratios) are shown in Figure 1B. The AEI ratios for *DC-SIGN* were as high as 6 in lung tissue samples, while AEI was evident in only one (rs11465396) of the two marker SNPs used for *DC-SIGN*. Moreover, strong AEI ratios were present in four of seven tissues; these results indicate that a regulatory variant exists in relatively high linkage disequilibrium (LD) with marker SNP rs11465396, but in negative LD with marker SNP rs4804801. The allelic ratios in SP-D were significant but rather small and were not investigated further. Detailed analysis focused on *SP-A2* because of its significant and consistent AEI ratios.

*SP-A2* has extensive homology to *SP-A1*; therefore, considerable care was taken to design primers that would specifically amplify each gene individually. In addition, *SP-A2* is expressed at much higher levels in the lung than *SP-A1*, mitigating any possible interference between the two genes at the level of mRNA. *SP-A1* expression was too low for reliable AEI analysis (>28 PCR cycles). Since the *SP-A2* AEI ratios occurred in each sample analyzed, and only in one direction, the marker SNP itself could be responsible for the difference in expression, or, instead, a regulatory variant in LD with the marker SNP could be

responsible. Because of the robust and unidirectional AEI ratios, we chose *SP-A2* for further characterization of functional polymorphisms, following the lead that any functional SNP is likely to reside within the same haplotype block as the marker SNP. Moreover, genetic variants in *SP-A2* have been proposed to mediate differential mRNA expression and alternative splicing [15–17], reinforcing the possible presence of a regulatory variant.

### Extended AEI analysis of *SP-A2*

To increase the number of tissues eligible for *SP-A2* AEI analysis, DNA from 38 lung samples was genotyped for four *SP-A2* marker SNPs: rs1059046 (Thr9Asn), rs17886395 (Ala91Pro), rs1965707 (synonymous, amino acid position 140) and rs1965708 (Gly223Lys) (for genomic location see Supplementary Figure 1). The SNPs corresponding to amino acids 9, 91, 140 and 223 have been used in previous reports to define *SP-A2 1A* haplotypes [18, 19]. Twenty seven of the samples were heterozygous for at least one of the marker SNPs; therefore they were informative for allelic mRNA expression measurement. Twenty of 27 testable samples manifested AEI, indicating the presence of a frequently occurring functional SNP. In these samples, AEI ranged from unity (1.0) to two-fold, with similar AEI ratios when more than one marker SNP was heterozygous in the same sample (Figure 2, Table 2). This latter result validates the analytical procedure used to measure AEI and suggests the presence of only one main regulatory variant. In all but one tissue with one marker SNP (rs1059046), the major/minor allele ratios were <1.0, indicating that the minor allele is expressed more highly compared to the major allele – a potential gain of function. When AEI occurs often, and largely in one direction, one would expect that any one of the marker SNPs could have been responsible for AEI. Indeed, two of the marker SNPs were significantly associated with AEI (rs1059046 with *p*=0.0005 and rs1965707 with *p*=0.003). However, neither of these two marker SNPs correlated perfectly with presence or absence of AEI (being homozygous in all tissues lacking AEI and heterozygous in all others). This result indicated that there is one or more distinct functional SNP(s) in high LD with the marker SNPs.

### SOLiD RNA deep-sequencing of *SP-A2* in lung RNA

We used RNA sequencing (RNA-Seq) to characterize the *SP-A2* transcripts in one of the lung tissues. *SP-A2* mRNA was highly expressed, at 124 fragments per kilobase of gene per million mapped (FPKM). RNA-Seq expression of other candidate lectins was considerably lower, with SP-D (18.5 FPKM) > SP-A1 (9.5 FPKM)    MRC1 (9.4 FPKM) > Dectin-1 (5.9 FPKM).

The *SP-A2* transcript was well defined, with no evidence of extended 5′ or 3′ transcripts (Supplementary Figure 2). However, the *SP-A* region is difficult to map because of multiple gene homologs, while the 5′ end of the transcript had decreased coverage, consistent with typical RNA-Seq of human tissues. Therefore, transcripts at the 5′ end of the gene were insufficiently abundant for detailed analysis. The alignments indicate the absence of abundant splice variants or non-annotated untranslated regions (UTRs), which could have affected the interpretation of AEI analysis at exons that are shared between distinct *SP-A2* mRNA transcript isoforms. Use of an algorithm to detect polyadenylation sites revealed several peaks in the 3′-UTR; this finding suggests the presence of multiple mRNAs with

different lengths of the 3′-UTR consistent with genome browser annotations. The regulatory effect of varying 3′-UTR length on mRNA stability and translation remains to be studied. Annotated SNPs detected in *SP-A2* mRNA in this sample are also shown in Supplementary Figure 2. Several of these have similar high population abundance (~40%) as they mostly reside on one frequent haplotype. This result supports the notion that *SP-A2* exists in two main haplotypes each occurring with high abundance (also known as "yin-yang haplotypes" which is defined as two high-frequency haplotypes with exactly opposite allelic configurations). Allelic expression ratios were detectable in heterozygous loci at the 3′ end of the transcript (Supplementary Table 1), overall supporting the presence of AEI. However, AEI ratios varied for each SNP, presumably owing to the complex process of amplification and library processing. Also, some of the detected SNPs may not be in phase with the other SNPs, or they may reside in different transcripts with distinct polyadenylation sites; therefore an accurate estimate of the AEI ratio was not feasible in this sample.

**Analysis of the regulatory variant accounting for AEI using Ion Torrent DNA sequencing of *SP-A2***

To perform extended genetic scanning in the *SP-A2* gene, we focused on lung tissue samples that either lacked or displayed significant AEI (8 lung samples each). Genomic DNA was sequenced including the entire transcribed *SP-A2* gene region (exons and introns), plus 1 kb upstream and downstream, using an Ion Torrent semiconductor sequencer (Supplementary Figure 3). This analysis revealed a large number of additional SNPs, listed in Supplementary Table 2. This technique also confirmed the presence of the four marker SNPs we studied that are highlighted in bold and listed in this table. A one-by-one scan of all the detected SNPs was then performed to search for a SNP that matches perfectly with presence or absence of AEI – even a single mismatch argues against a causative role of any given SNP because of the robust AEI ratio analysis (while some analytical error needs to be allowed for). This search led to only a single SNP in the 5′ end of intron 2 (rs1650232 G/A; Supplementary Figure 1) being heterozygous in each AEI-positive tissue and homozygous in all AEI-negative tissues. To further validate this result, the genomic region containing rs1650232 was analyzed by Sanger sequencing in all lung tissues with AEI ratios available (Table 2). The association of rs1650232 with AEI was highly significant, with a *p* value of $1 \times 10^{-5}$ (Table 2). In three tissues, the SNP-AEI association did not match (#24, 26, and 27) (Table 2, with two AEI ratios close to the cut-off value). Assignment of AEI-positive or negative status would change with different cut-off criteria, while the presence of additional rare regulatory variants also cannot be excluded as confounding factors. By process of elimination, rs1650232 remained the only variant found in these 38 tissues that optimally met all criteria for a causative regulatory variant, with the minor (A) allele causing increased expression as measured with the allelic mRNA expression assays. This result supports the notion that rs1650232 alone is sufficient to account for the measured AEI ratios.

**Predicted function of rs1650232**

We performed bioinformatics analyses to assess the possible function of rs1650232. In *SP-A2* the rs1650232 SNP lies next to a NAGNAG 3′ splice acceptor site found in ~ 30% of human genes [20]. Alternative splicing occurs at these sites if both *AG* alleles of the NAGNAG acceptor can be used by the spliceosome [21]. The rs1650232 minor *A* allele

creates an additional tandem NAG site (Figure 3). The alternative transcript produced is 3 bases longer, and is catalogued in the expressed sequence tag (EST) database (Santa Cruz genome browser). This 5′-untranslated region of the *SP-A2* gene had previously been associated with alternative splicing of the *SP-A2* transcript [16], and more recently with mRNA expression and translation of the *SP-A2* gene [22]. Thus, according to our allelic RNA expression analysis, the rs1650232 NAGNAG splice acceptor variation promotes alternative splicing and increased abundance of the *SP-A2* transcript.

### Effect of the rs1650232 *A* allele on formation of an alternatively spliced *SP-A2* mRNA variant with three additional bases at the beginning of exon 3

We extracted RNA from 16 lung tissues, converted the mRNA to cDNA with a primer targeting exon 3 in *SP-A2*, and prepared bar-coded libraries representing the region spanning the exon 2/exon 3 junction, for sequencing on the Ion Torrent. Of these samples, two were homozygous for the rs1650232 *G* allele, 3 heterozygous *G/A,* and 10 homozygous for the *A* allele. Shown in Figure 4 are the sequence read alignments onto the intron 2/exon 3 region obtained for the three genotypes, demonstrating the appearance of the three predicted additional bases in the spliced mRNA before exon 3, in an additive fashion for heterozygotes and homozygotes. Carriers of the main *GG* genotype have 7±3% of the RNA transcripts containing the extra three NAG bases, whereas *AG* heterozygotes have 20±6%, and *AA* homozygotes have 58±5%. This result confirms the proposed splicing mechanism in lung tissues, and together with the AEI results, indicates that rs1650232 is the main regulatory variant in *SP-A2*.

## DISCUSSION

Genetic variation plays a key role in human phenotypic variability, including susceptibility or resistance to diseases and response to therapies. In this context, expression genetics has become an emerging area that addresses a main source of phenotypic variability. We have observed that mRNA expression of human innate immune genes (C-type lectin genes in this study) varies up to several-fold from donor to donor. There is strong evidence that genetic regulatory factors contribute to this variability, but the degree and mechanism of genetic contribution to this phenotype has been explored only to a limited extent. Confounding the analysis of genetic factors, high variability of total mRNA levels in target tissues from different subjects can have multiple causes, including environmental influence on gene expression (regulation in *trans*). The use of allelic mRNA analysis is a powerful tool for the discovery of *cis*-acting regulatory polymorphisms [9–14], an approach that has lagged behind genome-wide scanning studies using numerous marker SNPs, with no prior applications to infectious diseases. Therefore, we screened selected C-type lectin genes known to be implicated in the pathogenesis of both infectious and non-infectious diseases from healthy human donor macrophages and lung tissues, measuring allelic mRNA ratios.

### Allelic expression imbalance in *DC-SIGN, SP-D, and SP-A2*

An indicator of the presence of regulatory variants, frequent AEI was present in 3 of the 6 genes studied: *SP-A2, SP-D, DC-SIGN*, suggesting that regulatory variants may be abundant in genes involved in the innate immune system. Substantial and even near mono-allelic

expression was observed in *DC-SIGN* (Figure 1) in 4 out of 7 heterozygous lung samples, with AEI ratios suggesting robust effects on mRNA expression. We suspect the presence of a relatively frequent regulatory variant in partial LD with rs11465396, as the AEI ratios were in the same direction for all 4 samples (Figure 1B). This result informs further research in identifying the causative regulatory variant. The AEI ratios observed in *SP-D* mRNA did deviate significantly but only modestly from unity; thus any physiological significance cannot be surmised from these results. Consistent and robust AEI ratios were observed in *SP-A2* mRNA in lung tissues, which was further studied to identify the causative variant. We acknowledge that lung tissues are composed of several cell types, which could confound interpretation of the results. However, detection of significant AEI permits the search for any regulatory variant even if several cell types are contributing to the AEI ratios.

## A regulatory SNP in *SP-A2*

The AEI ratios consistently ranging below unity indicated the presence of a regulatory variant within the haplotype block shared by all four marker SNPs used for AEI analysis (Figure 2), with the minor allele apparently being expressed more robustly (gain of function); however, none of the marker SNPs fully accounted for the observed ratios. In addition, in one tissue we observed an AEI ratio >1, arguing against this marker SNP being responsible, but rather a variant allele that in this particular individual had undergone a recombination event, thereby, residing on the main haplotype. To identify the causative regulatory variant, heterozygous in all samples with AEI and homozygous in all others, we employed rapid gDNA sequencing (Ion Torrent technology) across the entire *SP-A2* gene locus. This approach revealed a single eligible SNP, rs1650232, in intron 2 residing in the 5′-UTR, largely fulfilling this stringent criterion. An apparent mismatch in three tissues (Table 2) could have been low analytical accuracy in two samples with AEI ratios near the cut-off value. The remaining mismatch in sample 24 (AEI ratio 0.98, rs1650232 heterozygous) could have resulted from compensatory regulatory variants present at low rates (a likely scenario when 38 tissues are analyzed). Since rs1650232 alters a NAGNAG site affecting splicing, we sequenced the generated mRNAs from each genotype (*GG, AG, GG*), demonstrating that formation of an mRNA species including three bases before exon 3 indeed depends on the presence of the minor *A* allele of rs1650232 in human lung tissues.

SNP rs1650232 had previously been associated with differential *SP-A2* RNA expression and alternative splicing [16, 22, 23], but had not been utilized in clinical association studies. Taken together, these results identify rs1650232 as a main regulatory variant in *SP-A2*, accounting for nearly all observed AEI. Since rs1650232 alters the length of the 5′-UTR by 3 nucleotides without affecting the amino acid sequence, it is likely that the increased abundance of the alternatively spliced transcript leads to increased protein activity, as reported [22].

The *SP-A2* gene locus consists of large haplotype blocks, often with exact opposite allelic configuration involving multiple SNPs – also referred to as yin-yang haplotypes. As a result, allelic mRNA ratios measured at four different marker SNPs were all in the same direction below 1.0, except in one single case (Figure 2), indicating each minor allele resides on the same haplotype. The presence of long yin-yang haplotypes may be related to evolutionary

selection pressure maintaining more than one haplotype across populations, while neutral processes cannot be ruled out [24]. Parallel selection of alleles with varying activity may be prevalent in genes supporting the innate immune system, balancing the need for protection against infections with autoimmune disorders [4]. Any haplotype carrying the minor allele of rs1650232 may therefore be associated with increased activity of SP-A2.

### Previous clinical association studies with SNPs and haplotypes containing rs1650232

*SP-A2* rs1650232 is in strong but varying LD with 4 other common coding SNPs that have been associated with several different aspects of host susceptibility in other studies. The haplotypes consist of one synonymous (rs1965707) and three non-synonymous (rs1059046, rs17886395 and rs1965708) SNPs, the same SNPs that we have used here as marker SNPs in our allelic expression studies. Generating several haplotypes, one of these (haplotype *1A3*) has previously been associated with tuberculosis [18, 19], a particular interest of our laboratories. This *SP-A2* haplotype alone or in combination with an *SP-A1* haplotype was also associated with increased risk of meningococcal disease and respiratory symptoms in infants at risk for asthma [25, 26].

Association of this 4-SNP *SP-A2* haplotype seemed to be more pronounced than that of individual SNPs, suggesting that multiple markers may better predict the risk for diseases. However, two distinct non-synonymous SNPs, rs17886395 and rs1965708, were also found to be associated with tuberculosis [19], and rs1965708 alone may confer susceptibility to meningococcal disease [25] or high-altitude pulmonary edema [27]. The results presented here indicate that a single SNP, rs1650232, can largely account for genetic effects on expression regulation, while any effect of non-synonymous SNPs on SP-A2 protein function cannot be addressed with the approach taken here. Previous association studies regarding susceptibility to tuberculosis and other infectious diseases using various SNPs need to be interpreted with respect to rs1650232, which alone may account for most of the genetic variability of *SP-A2* expression and other associated phenotypes.

In conclusion, our study demonstrates the "proof of principle" that regulatory polymorphisms can be successfully identified from a relatively small number of target samples through the AEI approach. Such polymorphisms can then be further validated using cellular, molecular, biochemical and bioinformatics approaches. The identification of validated functional polymorphisms can then be included in a more robust tool kit for clinical association studies in the field.

## MATERIALS AND METHODS

### Primary human macrophages and lung tissues as sources of C-type lectin genes

Human monocyte-derived macrophages (MDMs) from 20 healthy human donors of different ethnic origin were used as sources of DNA and RNA for *MRC1*, *DC-SIGN* and *Dectin-1*. MDMs do not express SP-A or SP-D. Peripheral blood mononuclear cells (PBMCs), composed of monocytes and lymphocytes, were isolated from donor blood using a Ficoll-Hypaque cushion, and MDMs were prepared from PBMCs using standard laboratory techniques [28]. Briefly, the monocytes present in PBMCs were allowed to differentiate into

MDMs in the presence of IL-4 (to induce or enhance the expression of certain C-type lectins of interest) in RPMI containing 20% autologous serum in Teflon wells for 5 days. MDMs were purified by adherence to 6-well tissue culture plastic plates for 2 hand non-adherent lymphocytes washed away. MDMs were harvested in cold PBS from the monolayer (pre-chilled on ice for 30 min) by using a rubber policeman.

Thirty human lung biopsy specimens were obtained from subjects undergoing surgical cardiothoracic procedures at the OSU, using normal adjacent lung tissue, from the OSU Tissue Procurement Pathology Core. Tissues were immediately snap frozen upon removal in RNA*later* (a RNA stabilizing solution from Qiagen Inc., Valencia, CA), as the source of DNA and RNA for analysis of *SP-A1, SP-A2* and *SP-D*, in addition to *MRC1*, *DC-SIGN* and *Dectin-1*. An approved IRB protocol by the OSUMC was used for both MDM and lung tissue acquisition.

### DNA and RNA extraction and Reverse Transcriptase (RT)-PCR

Genomic DNA and total RNA were prepared from both MDMs and frozen lung tissue samples. Lung tissues were first minced by a tissue chopper (Mickle Laboratory, Surrey, UK). For DNA extraction, one portion of the cells or lung tissue was digested with proteinase K in ATL buffer and processed step-wise to obtain purified DNA by using the DNeasy Tissue kit from Qiagen, according to the manufacturer's instructions. For total RNA isolation, another portion of the cells or lung tissue was homogenized in RLT buffer by a Mini-Bead Beater Cell Disruptor (BioSpec Products, Bartlesville, OK) and then processed step-wise to obtain purified RNA by using the RNeasy Mini kit from Qiagen, according to the manufacturer's instructions. cDNA was generated from mRNA present in total RNA (1 μg) by reverse transcription using Superscript II Reverse Transcriptase, primed with oligo-dT (Invitrogen), plus gene-specific primers targeting regions immediately 3′ of the marker SNPs. Quantitative Real-Time PCR (qRT-PCR) was performed on the cDNA to determine mRNA levels. Primers used for qRT-PCR were the same as those selected for genotyping and primer extension (Supplementary Table 3), with PCR conditions optimized for each primer on an ABI 7000 cycler with SYBR Green dye as a detector. Expression levels were compared to β-actin expression in the same sample. To assure absence of gDNA in the cDNA preparations, qRT-PCR was also performed in the absence of added reverse transcriptase, showing no significant contamination of RNA with remnant gDNA.

### Genotyping of the SNPs

SNaPshot [TM] (Life Technologies, Foster City, CA) primer extension methodology was used for genotyping of selected marker SNPs. Gene-specific primers were used to PCR amplify the target stretch of DNA (50–150 bp) in the selected C-type lectin gene (Supplementary Table 3). The targeted polymorphism was evaluated using fluorescently-labeled terminator nucleotides [29]. The products were analyzed on an Applied Biosystems (ABI) 3730 capillary electrophoresis DNA instrument, and peak ratios were calculated with Gene Mapper [TM] 3.0 software (ABI / Life Technologies). Additional genotyping was accomplished with either Sanger, or Ion Torrent [TM] (Life Technologies) semiconductor sequencing of specific gene regions of interest.

## Determination of Allelic mRNA Expression Imbalance (AEI)

To detect the presence of *cis*-acting functional polymorphisms in target human samples, allelic ratios were measured in both genomic DNA and mRNA with the SNaPshot fluorescent primer extension method. In heterozygous samples, one allele serves as the control for the other, and differences between allele ratios in mRNA (in the form of cDNA) as compared to genomic DNA indicate the presence of AEI [29, 30]. Selection criteria to determine suitability of an indicator SNP are as follows: 1) the SNP must be located in coding or un-translated exons; 2) Ideally the minor allele frequency should be high (between 0.15 and 0.50) to allow for a sufficient number of heterozygous samples to be assayed. SNPs with smaller allele frequencies can be used, but it would then be necessary to select several indicator SNPs in one gene for obtaining enough heterozygous samples to reach statistical power; and 3) The SNP should preferably not lie too close to an exon boundary (at least 20 or 25 bases) so that the same set of primers used to amplify the surrounding sequence can be used in both DNA and RNA.

In AEI analysis, the same primers applied to genotyping were used to amplify regions of genomic DNA or cDNA, containing indicator SNPs for the initial six C-lectin genes. An extension primer was designed to hybridize adjacent to the indicator SNP, enabling extension with fluorescently-labeled nucleotide complementary to the SNP alleles. Peak heights were determined to calculate the AEI ratio of each heterozygote allele, normalized to the ratio of the same alleles in genomic DNA. In all assays of gDNA allelic ratios, there was no indication of copy number variation; therefore, to minimize experimental error within individual samples, for each marker SNP we averaged the allelic gDNA ratios and normalized to 1 (assuming a 1:1 allele ratio in gDNA). Standard deviations were in the range of ±5%. Mean mRNA allelic ratios were adjusted accordingly.

## Transcriptome Sequencing (RNA-Seq)

Ten nanograms of total lung RNA was used as starting input for RNA sequencing. We used the NuGen Ovation RNA-Seq kit (NuGen Technologies, San Carlos, CA) to prepare double-stranded cDNA for sequencing. The NuGen kit begins with low input quantities of total RNA, then uses a linear isothermal amplification process to amplify the RNA and converts it to cDNA (as opposed to a PCR based amplification). This amplification process uses random hexamers and oligo-dT to prime the reaction, thereby priming all RNA sequences across the entire gene. In addition, the NuGen method reduces ribosomal RNA to approximately 5% in the RNA preparations. cDNA synthesis was verified and quantified with qRT-PCR. The double-stranded cDNA was sheared to 140-bp fragments with the Covaris sonicator. After end repair, SOLiD 4 (Life Technologies) sequencing adaptors were ligated to both ends of the library fragments. The ligation product was amplified for 8 cycles of PCR, size selected, then quantified again using qRT-PCR. Emulsion PCR and templated bead enrichment were performed on EZ bead instruments (Life Technologies). The lung RNA was sequenced on a SOLiD4 HQ instrument from Life Technologies. The Bioscope program (Life Technologies) was used for alignment of transcript reads. We used the SOLiD system for allelic sequencing because each DNA base is queried twice in the sequencing process, with two different probes, thus reducing the effect of mismatches on identification and quantification of minor alleles.

### Extended *SP-A2* gene sequencing

The Ion Torrent Personalized Genome Machine (PGM)™ (Life Technologies, Foster City CA) was used to sequence the *SP-A2* gene loci from 16 lung tissues. The starting library material was generated by PCR amplification of six different 1–4 kb DNA fragments, designed to include the entire *SP-A2* genomic sequence, plus ~1,000 base pairs 5′ and 3′ of the transcribed region. *SP-A2* primers used for generating *SP-A2* specific amplicons are listed in Supplementary Table 3. For each sample, the amplicons were treated with exonuclease I to degrade PCR primers, combined in equimolar proportions, then sheared into ~ 100 bp fragments using a Covaris sonicator. After shearing, the DNA fragments were centrifuged over a Centricon YM-30 membrane column (50 bp MW cut-off), and the remaining DNA fragments were retained. The recovered fragments were end-polished; then Ion Torrent bar-coded adaptors were ligated onto 100 ng of each sample using an Ion Torrent Library Preparation Kit (Life Technologies). Bar-coded libraries were quantified with qRT-PCR; then six or ten libraries were combined in equimolar amounts to produce templated beads for sequencing. Automated emulsion PCR was performed on an Ion One Touch instrument using the 100 bp Ion One Touch preparation kit (Life Technologies). Preparations containing either six or ten combined bar-coded libraries were loaded onto 314 Chips™ (Life Technologies) and sequenced on the PGM.

### Sequencing of the exon 2/intron 2/exon 3 region in *SP-A2* mRNA

mRNA extracted from lung tissues from 2 subjects homozygous for the rs1650232 *G* allele, 3 heterozygous *G/A,* and 10 homozygous for the *A* allele, was converted to cDNA using a primer targeting a region in exon 3 (CAGATGGAGTGAGTC) and amplified with the same primer and a reverse primer targeting exon 2 (TGGTGTTTCTCCAGGCGG). The amplicons were converted to 15 bar-coded libraries for sequencing according to a protocol provided by Life Technologies for sequencing on an Ion Torrent instrument. Reads of ~100 bases were then aligned onto the genomic *SP-A2* sequence and displayed as cumulative reads for each base.

### Data analysis

Association between genotypes and discrete variables (presence or absence of AEI, defined as allelic mRNA ratios significantly deviating from allelic gDNA ratios by >3 S.D.) was analyzed with Helix-tree software. Statistical significance of data was expressed as *p* values by using the student *t*-test.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

# References

1. Sorensen TI, Nielsen GG, Andersen PK, Teasdale TW. Genetic and environmental influences on premature death in adult adoptees. N Engl J Med. 1988 Mar 24; 318(12):727–732. [PubMed: 3347221]

2. Cooke GS, Hill AV. Genetics of susceptibility to human infectious disease. Nat Rev Genet. 2001 Dec; 2(12):967–977. [PubMed: 11733749]

3. Hill AV. Aspects of genetic susceptibility to human infectious diseases. Annu Rev Genet. 2006; 40:469–486. [PubMed: 17094741]

4. Azad AK, Sadee W, Schlesinger LS. Innate immune gene polymorphisms in tuberculosis. Infect Immun. 2012 Oct 1.80:3343–3359. [PubMed: 22825450]

5. Torrelles JB, Azad AK, Henning LN, Carlson TK, Schlesinger LS. Role of C-type lectins in mycobacterial infections. Curr Drug Targets. 2008 Feb; 9(2):102–112. [PubMed: 18288961]

6. McGreal EP, Miller JL, Gordon S. Ligand recognition by antigen-presenting cell C-type lectin receptors. Curr Opin Immunol. 2005 Feb; 17(1):18–24. [PubMed: 15653305]

7. Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. Cis-acting variation in the expression of a high proportion of genes in human brain. Hum Genet. 2003 Jul; 113(2):149–153. [PubMed: 12728311]

8. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. Science. 2002 Aug 16.297(5584):1143. [PubMed: 12183620]

9. Johnson AD, Gong Y, Wang D, Langaee TY, Shin J, Cooper-DeHoff RM, et al. Promoter polymorphisms in ACE (angiotensin I-converting enzyme) associated with clinical outcomes in hypertension. Clin Pharmacol Ther. 2009 Jan; 85(1):36–44. [PubMed: 18946466]

10. Zhang Y, Bertolino A, Fazio L, Blasi G, Rampino A, Romano R, et al. Polymorphisms in human dopamine D2 receptor gene affect gene expression, splicing, and neuronal activity during working memory. Proc Natl Acad Sci U S A. 2007 Dec 18; 104(51):20552–20557. [PubMed: 18077373]

11. Johnson AD, Wang D, Sadee W. Polymorphisms affecting gene regulation and mRNA processing: broad implications for pharmacogenetics. Pharmacol Ther. 2005 Apr; 106(1):19–38. [PubMed: 15781120]

12. Johnson AD, Zhang Y, Papp AC, Pinsonneault JK, Lim JE, Saffen D, et al. Polymorphisms affecting gene transcription and mRNA processing in pharmacogenetic candidate genes: detection through allelic expression imbalance in human target tissues. Pharmacogenet Genomics. 2008 Sep; 18(9):781–791. [PubMed: 18698231]

13. Lim JE, Pinsonneault J, Sadee W, Saffen D. Tryptophan hydroxylase 2 (TPH2) haplotypes predict levels of TPH2 mRNA expression in human pons. Mol Psychiatry. 2007 May; 12(5):491–501. [PubMed: 17453063]

14. Wang D, Guo Y, Wrighton SA, Cooke GE, Sadee W. Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs. Pharmacogenomics J. 2010 Apr 13.

15. Wang G, Umstead TM, Phelps DS, Al Mondhiry H, Floros J. The effect of ozone exposure on the ability of human surfactant protein a variants to stimulate cytokine production. Environ Health Perspect. 2002 Jan; 110(1):79–84. [PubMed: 11781168]

16. Karinch AM, DeMello DE, Floros J. Effect of genotype on the levels of surfactant protein A mRNA and on the SP-A2 splice variants in adult humans. Biochem J. 1997; 321:39–47. [PubMed: 9003399]

17. Wang G, Guo X, Floros J. Human SP-A 3′-UTR variants mediate differential gene expression in basal levels and in response to dexamethasone. Am J Physiol Lung Cell Mol Physiol. 2003 May; 284(5):L738–L748. [PubMed: 12676764]

18. Floros J, Lin HM, García A, Salazar MA, Guo X, DiAngelo S, et al. Surfactant protein genetic marker alleles identify a subgroup of tuberculosis in a Mexican population. J Infect Dis. 2000; 182:1473–1478. [PubMed: 11023470]

19. Malik S, Greenwood CM, Eguale T, Kifle A, Beyene J, Habte A, et al. Variants of the SFTPA1 and SFTPA2 genes and susceptibility to tuberculosis in Ethiopia. Hum Genet. 2006 Feb; 118(6): 752–759. [PubMed: 16292672]

20. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet. 2004 Dec; 36(12):1255–1257. [PubMed: 15516930]

21. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, et al. Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. Am J Hum Genet. 2006 Feb; 78(2):291–302. [PubMed: 16400609]

22. Silveyra P, Raval M, Simmons B, DiAngelo S, Wang G, Floros J. The untranslated exon B of human surfactant protein A2 mRNAs is an enhancer for transcription and translation. Am J Physiol Lung Cell Mol Physiol. 2011 Nov; 301(5):L795–L803. [PubMed: 21840962]

23. Wang G, Guo X, Floros J. Human SP-A 3′-UTR variants mediate differential gene expression in basal levels and in response to dexamethasone. Am J Physiol Lung Cell Mol Physiol. 2003 May; 284(5):L738–L748. [PubMed: 12676764]

24. Liu J, Zhou Z, Hodgkinson CA, Yuan Q, Shen PH, Mulligan CJ, et al. Haplotype-based study of the association of alcohol-metabolizing genes with alcohol dependence in four independent populations. Alcohol Clin Exp Res. 2011 Feb; 35(2):304–316. [PubMed: 21083667]

25. Jack DL, Cole J, Naylor SC, Borrow R, Kaczmarski EB, Klein NJ, et al. Genetic polymorphism of the binding domain of surfactant protein-A2 increases susceptibility to meningococcal disease. Clin Infect Dis. 2006 Dec 1; 43(11):1426–1433. [PubMed: 17083016]

26. Pettigrew MM, Gent JF, Zhu Y, Triche EW, Belanger KD, Holford TR, et al. Respiratory symptoms among infants at risk for asthma: association with surfactant protein A haplotypes. BMC Med Genet. 2007; 8:15. [PubMed: 17407567]

27. Saxena S, Kumar R, Madan T, Gupta V, Muralidhar K, Sarma PU. Association of polymorphisms in pulmonary surfactant protein A1 and A2 genes with high-altitude pulmonary edema. Chest. 2005 Sep; 128(3):1611–1619. [PubMed: 16162765]

28. Kang BK, Schlesinger LS. Characterization of mannose receptor-dependent phagocytosis mediated by *Mycobacterium tuberculosis* lipoarabinomannan. Infect Immun. 1998; 66(6):2769–2777. [PubMed: 9596746]

29. Zhang Y, Wang D, Johnson AD, Papp AC, Sadee W. Allelic expression imbalance of human mu opioid receptor (OPRM1) caused by variant A118G. J Biol Chem. 2005 Sep 23; 280(38):32618–32624. [PubMed: 16046395]

30. Wang D, Johnson AD, Papp AC, Kroetz DL, Sadee W. Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. Pharmacogenet Genomics. 2005 Oct; 15(10):693–704. [PubMed: 16141795]
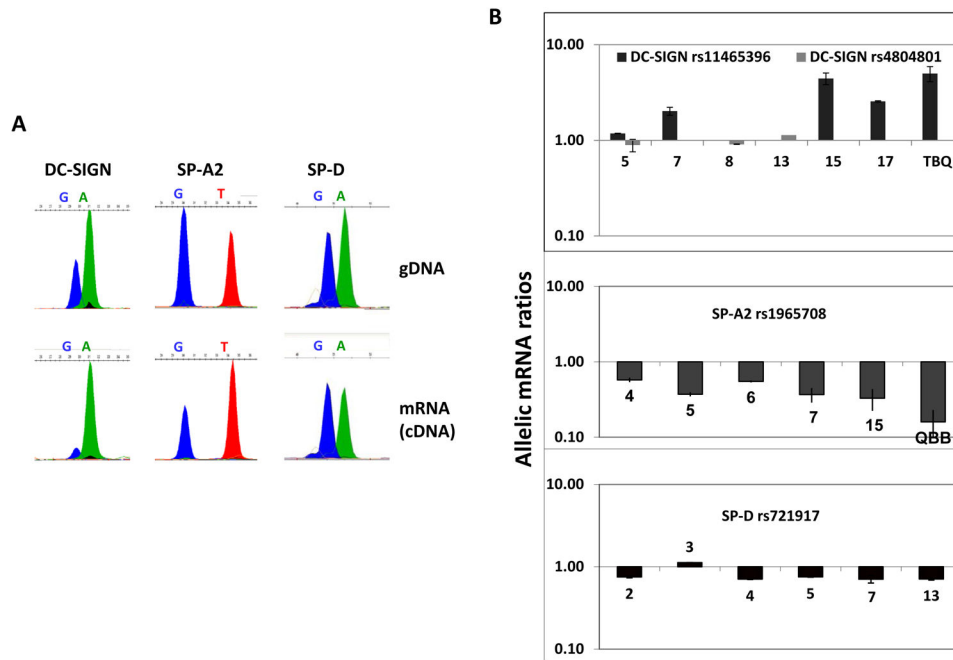
**Figure 1.**
mRNA and gDNA allelic ratios at marker SNPs for *DC-SIGN* (at rs11465396 *G>A*), *SP-A2* (at rs1965708 *G>T*) and *SP-D* (at rs721917 *G>A*) in lung tissues, measured with SNaPshot methodology. (**A**) Chromatographic peaks representing relative allelic abundance in gDNA and mRNA (the latter shown as cDNA). (**B**) Allelic mRNA expression ratios of *DC SIGN*, *SP-A2* and *SP-D*, with mRNA ratios (major/minor allele) normalized to gDNA ratios set at unity (1.00). For *DC-SIGN*, results with two marker SNPs are shown (rs11465396 and rs4804801), but significant deviation from unity (presence of AEI) was observed only with rs11465396. Results are the mean ± SD (n  3). The number or letters labeled on the top or bottom of each bar represent the individual donor of lung tissue.
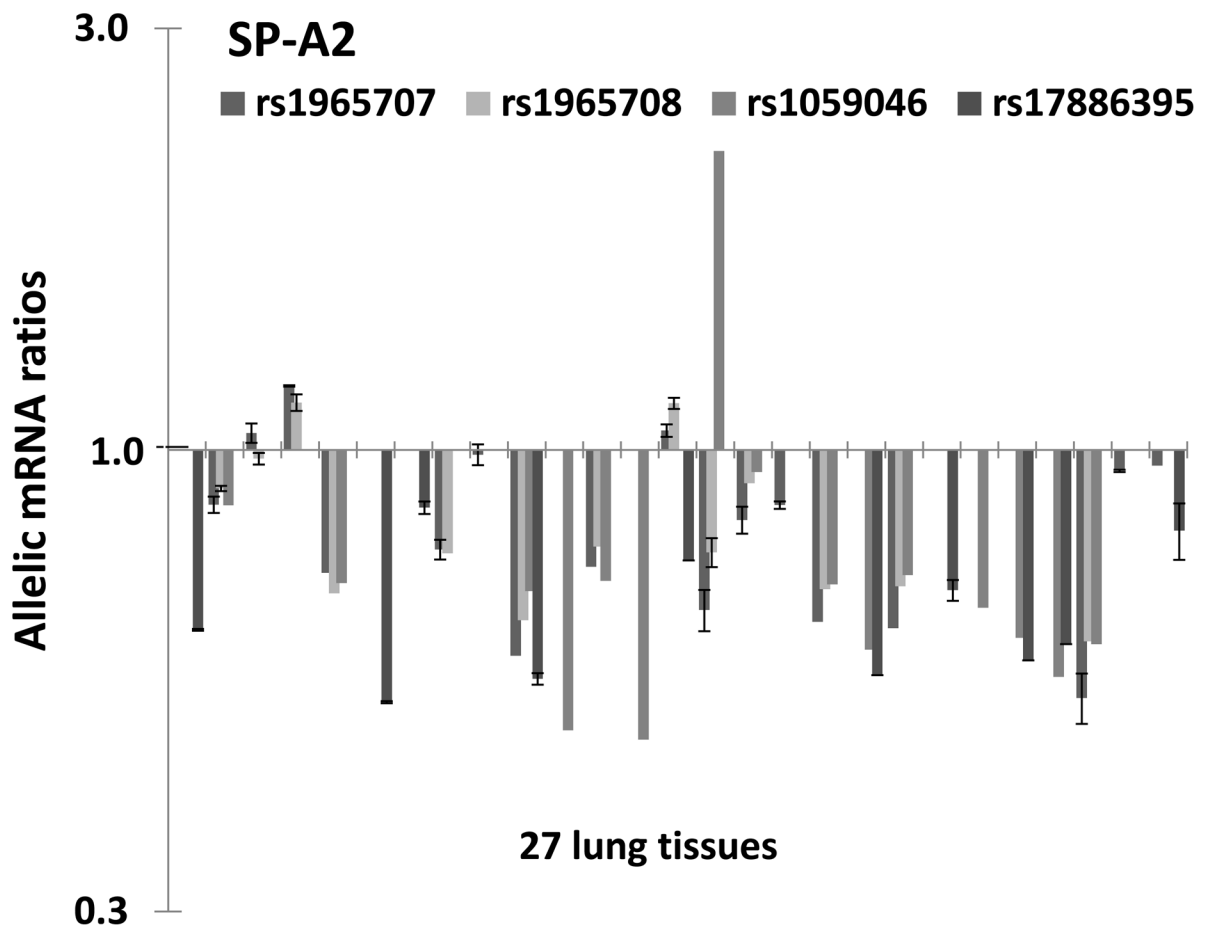
**Figure 2.**
Allelic mRNA expression ratios of *SP-A2* in lung tissues heterozygous for one or more of the four marker SNPs, which are rs1059046 (exon 2), rs17886395 (exon 3), rs1965707 (exon 4) and rs1965708 (exon 4). All four SNPs were genotyped in all tissues; where allelic mRNA ratios are lacking, the SNP position was homozygous. AEI was assumed to be present when the ratios were <0.8 or >1.25 (approximately 3×SD). None of the marker SNPs are heterozygous in all tissues showing AEI, or homozygous in all lacking AEI; therefore none of the four marker SNPs fully accounted for the observed AEI. mRNA ratios (major/minor allele) were normalized to gDNA ratios set at 1.0. Results are mean ± SD (n=3).

**Figure 3.**
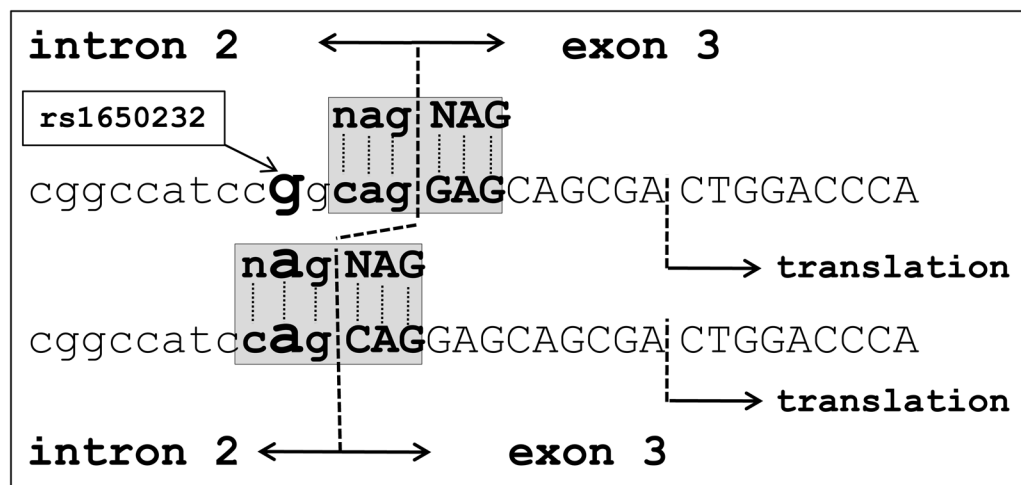The intron 2 SNP rs1650232 generates a new NAGNAG splice acceptor site shifted by three bases. Exon 2 is thereby extended by 3 bases. Human ESTs in the UCSC genome browser contain sequence fragments both including and excluding the three additional bases at 5′ of the annotated intron 2/exon 3 splice junction. A schematic diagram of the *SP-A2* gene locus is provided in Supplementary Figure 1.
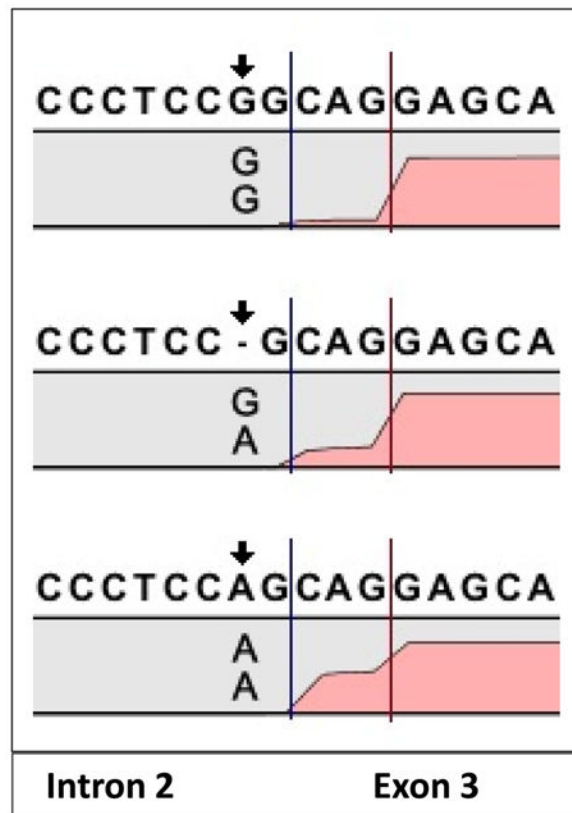
**Figure 4.**
Detection of the splice variant with three additional bases generated with the minor *A* allele of rs1650232. A region spanning from exon 2 into exon 3 was PCR amplified and sequenced on an Ion Torrent instrument. Shown are the cumulative reads for each base for one representative lung tissue from homozygous *GG*, heterozygous *GA* and homozygous *AA* carriers. Vertical lines bracket the three additional bases, indicating the exon 2/exon 3 boundary either before or after the three bases depending upon the allele present (*G* or *A*), and the Y axis represents cumulative aligned sequences. Substantial read counts for the three bases were observed only in *A* allele carriers, with double the counts in homozygous versus heterozygous carriers.

**Table 1**

Marker SNPs used in genotyping and initial AEI analysis for C-type lectin genes from MDM and lung tissues

| Gene | Marker SNP | Heterozygotes Detected [a] |
|---|---|---|
| *MRC1* | rs2296414 | 11 |
| *MRC1* | rs941 | 20 |
| *DC-SIGN* | rs11465396 | 7 |
| *DC-SIGN* | rs4804801 | 4 |
| *Dectin-1* | rs16910526 | 0 |
| *SP-A1* | rs4253527 | 4 |
| *SP-A2* | rs1965708 | 7 |
| *SP-D* | rs721917 | 7 |
| *SP-D* | rs17851984 | 0 |

[a]After genotyping, the number of heterozygote samples eligible for AEI analysis is provided. Both MDM (N = 20) and lung tissue (N = 30) samples were genotyped for *MRC1*, *DC-SIGN* and *Dectin-1*; and only lung tissue (N = 30) samples were genotyped for *SP-A1*, *SP-A2* and *SP-D*.

**Table 2**

*SP-A2* genotypes and allelic mRNA expression (AEI) ratios

| Lung sample | rs1059046 (p=0.0005) | rs17886395 (p=0.91) | rs1965707 (p=0.003) | rs1965708 (p=0.92) | rs1650232 (p=0.00001) | Average AEI Ratio [a] | AEI SD |
|---|---|---|---|---|---|---|---|
| 1 | A_A | G_G | --- | C_C | A_A | | |
| 2 | C_C | G_C | --- | A_A | A_G | 0.63 | |
| 3 | C_C | C_C | --- | C_C | A_A | | |
| 4 | A_C | G_G | C_T | C_A | A_G | 0.88 | 0.02 |
| 5 | C_C | C_C | C_T | C_A | G_G | 1.01 | 0.05 |
| 6 | A_A | G_G | C_T | C_A | --- | 1.16 | 0.04 |
| 7 | A_C | G_G | C_T | C_A | A_G | 0.71 | 0.02 |
| 8 | A_A | G_G | --- | C_C | A_A | | |
| 9 | A_C | G_C | C_C | C_C | A_G | 0.52 | |
| 10 | A_A | G_G | --- | C_C | A_A | | |
| 11 | A_A | G_G | --- | C_C | A_A | | |
| 12 | A_C | G_G | --- | C_C | A_G | | |
| 13 | A_A | G_G | --- | C_C | A_A | | |
| 14 | C_C | G_C | --- | A_A | A_G | 0.86 | |
| 15 | A_C | G_G | C_T | C_A | A_G | 0.77 | 0.01 |
| 16 | A_A | G_G | C_C | C_C | A_A | | |
| 17 | A_A | --- | C_T | C_C | G_G | 0.99 | 0.02 |
| 18 | A_C | G_C | C_T | C_A | A_G | 0.62 | 0.06 |
| 19 | A_A | G_G | C_C | C_C | A_A | | |
| 20 | A_A | G_G | C_C | C_C | A_A | | |
| 21 | A_C | G_G | C_C | C_C | A_G | 0.48 | 0.03 |
| 22 | A_C | G_G | C_T | C_A | A_G | 0.74 | |
| 23 | A_C | G_G | C_C | C_C | A_G | 0.47 | |
| 24 | C_C | G_C | C_T | C_A | A_G | 0.98 | 0.2 |
| 25 | A_C | | C_T | C_A | A_G | 0.63 | 0.15 |
| 26 | A_C | G_G | C_T | C_A | A_G | 0.9 | 0.06 |
| 27 | C_C | G_G | C_T | --- | A_A | 0.87 | 0.03 |
| 28 | A_C | G_G | C_T | C_A | A_G | 0.68 | 0.04 |

| Lung sample | rs1059046 (p=0.0005) | rs17886395 (p=0.91) | rs1965707 (p=0.003) | rs1965708 (p=0.92) | rs1650232 (p=0.00001) | Average AEI Ratio [a] | AEI SD |
|---|---|---|---|---|---|---|---|
| 29 | --- | G_G | C_C | C_C | --- | | |
| 30 | A_C | G_C | C_C | C_C | A_G | 0.57 | 0.03 |
| 31 | A_C | G_G | C_T | C_A | A_G | 0.68 | 0.05 |
| 32 | C_C | G_C | C_C | C_C | A_G | 0.69 | |
| 33 | A_C | G_G | C_C | C_C | A_G | 0.66 | |
| 34 | A_C | G_C | C_C | C_C | A_G | 0.6 | 0.02 |
| 35 | A_C | G_C | C_C | C_C | A_G | 0.58 | 0.03 |
| 36 | A_C | G_G | C_T | C_A | A_G | 0.58 | 0.05 |
| 37 | A_A | G_G | C_T | C_C | A_A | 0.95 | 0.03 |
| 38 | C_C | G_C | C_T | C_C | A_G | 0.88 | 0.11 |

[a] The AEI ratios are major/minor allele, determined by the SNaPshot assay. The first four SNPs listed were used as marker SNPs; the fifth SNP (rs1650232) represents the likely regulatory SNP accounting for all cases of AEI. Presence of AEI was determined if the average AEI ratio (n 3; data combined for allelic ratios obtained from each of three marker SNPs heterozygous in a given individual) was 0.88 or lower, or 1.24 or higher (this narrow range was chosen because AEI ratios determined for *SP-A2* mRNA had small SDs).