scientific reports

OPEN

Genetic information from discordant sibling pairs points to ESRP2 as a candidate trans-acting regulator of the CF modifier gene SCNN1B

Tim Becker[1,2], Andreas Pich[3], Stephanie Tamm[4], Silke Hedtfeld[4], Mohammed Ibrahim[3], Janine Altmüller[5], Nina Dalibor[5], Mohammad Reza Toliat [5], Sabina Janciauskiene[6,7], Burkhard Tümmler[4,6] & Frauke Stanke[4,6]✉

SCNN1B encodes the beta subunit of the epithelial sodium channel ENaC. Previously, we reported an association between SNP markers of SCNN1B gene and disease severity in cystic fibrosis-affected sibling pairs. We hypothesized that factors interacting with the SCNN1B genomic sequence are responsible for intrapair discordance. Concordant and discordant pairs differed at six *SCNN1B* markers (Praw = 0.0075, Pcorr = 0.0397 corrected for multiple testing). To identify the factors binding to these six SCNN1B SNPs, we performed an electrophoretic mobility shift assay and captured the DNA–protein complexes. Based on protein mass spectrometry data, the epithelial splicing regulatory protein ESRP2 was identified when using *SCNN1B*-derived probes and the ESRP2-*SCNN1B* interaction was independently confirmed by coimmunoprecipation assays. We observed an alternative SCNN1B transcript and demonstrated in 16HBE14o– cells that levels of this transcript are decreased upon ESRP2 silencing by siRNA. Furthermore, we confirmed that mildly and severely affected siblings have different ESPR2 genetic backgrounds and that ESRP2 markers are linked to the response of CF patients' nasal epithelium to amiloride, indicating ENaC involvement (Pbest = 0.0131, Pcorr = 0.068 for multiple testing). Our findings demonstrate that sibling pairs clinically discordant for CF can be used to identify meaningful DNA regulatory elements and interacting factors.

Genetic variation in humans contributes significantly to phenotypic variation. The question of which single nucleotide polymorphism (SNP) determines disease outcome and/or severity, has been addressed in more than 2000 genome-wide association studies (GWAS)[1]. However, the results have revealed that more than 90% of the polymorphisms identified by GWAS do not directly alter the gene's coding sequence. This has led to the conclusion that clinically relevant variation of the human genome mediates gene regulation, i.e., the transcript expression level or the composition of transcript isoforms[2]. Recent genome and epigenome studies have substantiated this hypothesis[3]. Enrichment of elements known for covalent modifications of DNA bases or their associated nucleosomes[4] among disease- and trait-associated genetic variants determined by GWAS has also been noted.

The gene affected in cystic fibrosis (CF), cystic fibrosis transmembrane conductance regulator (CFTR), encodes a chloride- and bicarbonate channel of epithelia[5–7] that localizes with the epithelial sodium channel ENaC at the apical membrane of many[8], albeit not all[9], epithelial cells. Both CFTR and ENaC act synergistically to regulate salt and fluid transport across the epithelium[8,10], and the SCNN1B gene, encoding the beta subunit of ENaC, is a highly plausible modifier gene of CF.

[1]Institute for Community Medicine, Ernst Moritz Arndt University, Greifswald, Germany. [2]xValue GmbH, Villich, Germany. [3]Research Core Unit Proteomics, Hannover Medical School, Hannover, Germany. [4]Department of Paediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Hannover, Germany. [5]Cologne Center for Genomics, University of Cologne, Cologne, Germany. [6]German Center for Lung Research (DZL), Partner site BREATH, Hannover, Germany. [7]Department of Pneumology, Hannover Medical School, Hannover, Germany. ✉email: mekus.frauke@mh-hannover.de
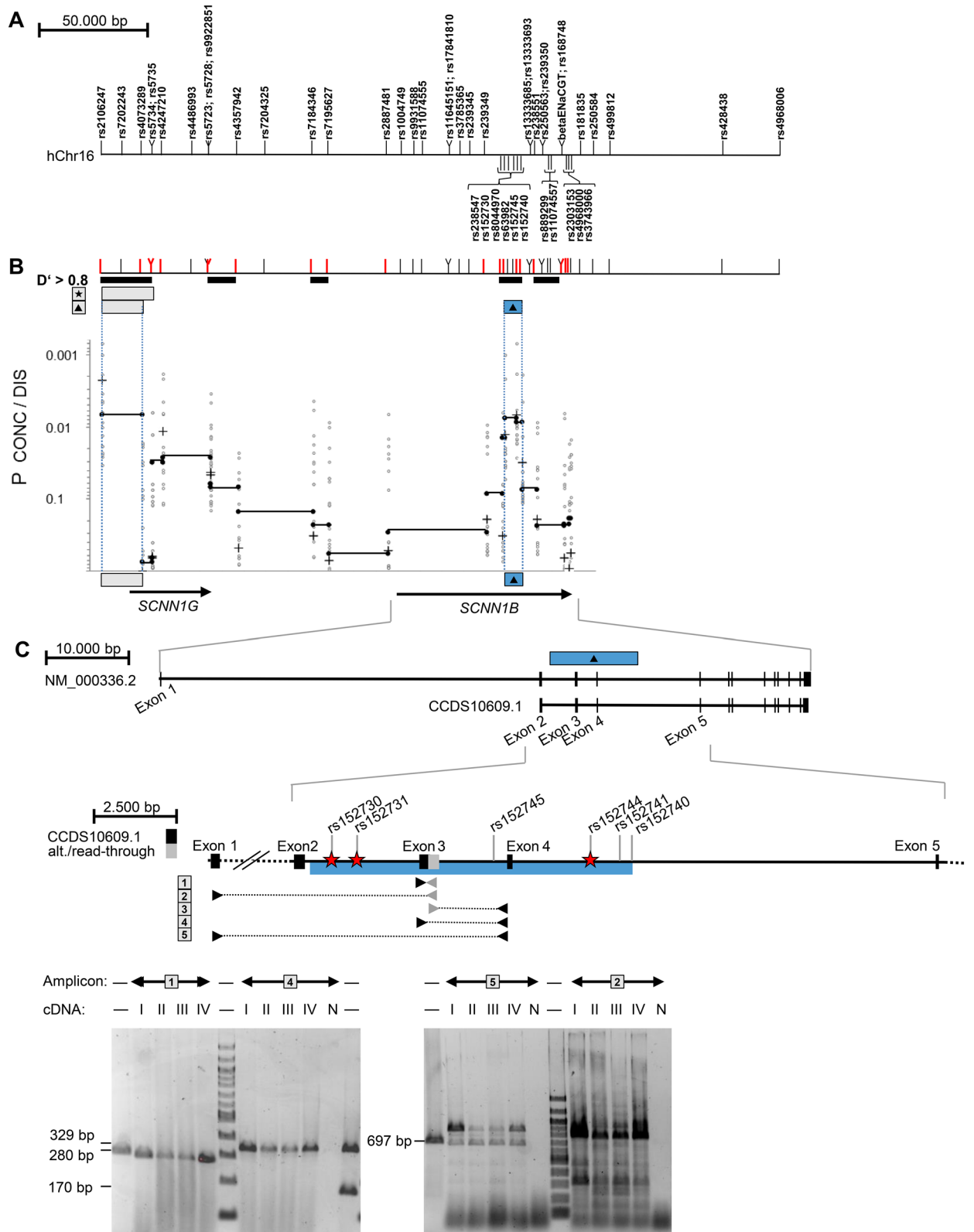
◀**Figure 1.** Mapping of the association signal determining intrapair discordance among F508del-CFTR homozygous CF sibling pairs, identification of an alternative SCNN1B transcript and position of SNPs analyzed by EMSA-PSeq. (**A**) Position of SNP and microsatellite markers on the genomic sequence of the *SCNN1G/SCNN1B* region. The position of genetic markers was derived from the genomic sequence on NC_000016 (assembly 03-Feb-2014; GRCh38; region 23302270–23381299). Nomenclature of *SCNN1B* exons was derived from Voilley et al. 1995[59] and Saxena et al. 1998[60], who have described exon/intron borders and the coding sequence for a 640 amino acid SCNN1B protein encoded by 13 exons (NP_000327.2; GI: 124301196; CCDS10609.1). The position of intragenic primers and expected sizes of amplicons covering exon 1–5 of *SCNN1B* are based on this coding sequence. SNPs rs5735, rs5723, rs1004749, rs238547, rs152730, rs250536 and on microsatellite betaENaCGT were previously typed[11] and further 49 SNPs were genotyped for fine-mapping (this work). (**B**) Haplotype blocks, informative markers and association signals in the SCNN1B/SCNN1G region. Markers depicted in red are informative (minimal allele frequency > 0.4) and were used to describe haplotype blocks (a measure of linkage disequilibrium D′ > 0.8) and map association signals as previously described[13]. Briefly, to map association signals, we defined genomic segments by adjacent ancestral informative markers that isolate the fragment that carries the causative variant(s)[13]. *p* values were generated by the software package FAMHAP[55–57]. *p* values refer to the best signal observed for a two-marker-combination of the 20-marker-set (Pbest) and the corrected value for multiple testing of 20 markers (Pcorr). Symbols refer to: Star: signal obtained by transmission disequilibrium test in the set of 37 sibling pairs families with extreme phenotypes (Pbest = 0.00101 for rs2106247–rs4073289; Pcorr = 0.0561 corrected for multiple testing of 20 informative markers); Filled triangle: association with intrapair discordance (Pbest = 0.0007 for rs2106247–rs152745; Pcorr = 0.0397 corrected for multiple testing of 20 informative markers) comparing 14 discordant and 23 concordant patient pairs. Uncorrected raw *p* values are shown for single markers (+), 2-marker-haplotypes of adjacent informative markers (filled circles; genomic segments spanned by two adjacent markers are linked by a black line) and 2-marker-haplotypes of non-adjacent markers (open circles). The SCNN1B association signal observed for intrapair discordance (filled triangle) was seen on the neighboring segments rs152730–rs152745 (Praw = 0.0075) and rs152745–rs152740 (Praw = 0.00869). A second independent association signal with intrapair discordance in the SCNN1G locus colocalizes with the previously noted survivor effect at SCNN1G[11] and was excluded from further analysis because of this confounding factor. (**C**) SNPs targeted by EMSA-PSeq and combinatorial PCR describing an alternative transcript generated by exon-read-through. Interaction partners of SNPs rs152730, rs152731 and rs152744, indicated by red star symbols, have been characterized by EMSA-PSeq. Equivalent amounts of cDNA from T84 (I), 16HBE14o-(II), CFTE29o-(III) and CFBE41o-(IV) or no template (negative control, N) were analyzed by combinatorial PCR with 5 amplicons. The grey intronic sequence (amplicon 1, 2, 3) was derived from ESTs BM694355 and BU730506 (see text and supplement for details). Intron-spanning primers targeting the SCNN1B reference mRNA (amplicons 4 and 5) as well as amplicon 1 targeting the alternative transcript and the 5′ UTR were observed in all cell lines. The size of amplified products from all amplicons was in accordance with the expected length. Size markers are loaded in lanes indicated by "—" and contain either a 100-bp-ladder or PCR products of 697 bp, 329 bp, 280 bp and 170 bp as indicated adjacent to the gel. Sanger sequencing of amplicon 1 confirmed the exon3-read-through and Sanger sequencing of amplicon 2 products verified its identity derived from a spliced mRNA as it showed a joined exon1 and exon2 sequence.

Previously, we focused on the three genes encoding the subunits of ENaC as candidate genes in the European CF twin and sibling study[11]. In an association study on extreme phenotypes, anthropometry and lung function data were used to select patients whose clinical data fell below the 25th centile (severely affected) or above the 75th centile (mildly affected) for both clinical parameters[12]. Three groups of affected patient pairs were defined as follows by a ranking algorithm used to describe the severity of CF: concordant mildly affected sibling pairs, concordant severely affected sibling pairs and discordant sibling pairs. Discordant sibling pairs were composed of one mildly and one severely affected sibling[12].

When discordant sibling pairs were compared to concordant sibling pairs, one *SCNN1B* haplotype defined by SNPs rs238547–rs152730–rs250563 occurred more frequently among discordant than among concordant siblings[11]. We concluded that this signal cannot be fully explained by a variant observed within *SCNN1B* because discordant siblings have a dissimilar phenotype by definition[12], and yet these siblings share an *SCNN1B* intragenic haplotype[11]. Our working hypothesis relied on the idea that the association signal in *SCNN1B* delineates a functional regulatory element, whereby a DNA-binding protein encoded *in trans* interacts with this regulatory element, stably binding to the haplotype of the regulatory element that is predominant among discordant siblings. Hence, the genetic variation of the interaction partner can determine the phenotype causing intrapair discordance in affected sibling pairs.

In association studies involving affected patient pairs, interaction between a regulatory element and a DNA-binding protein may result in a paradoxical situation. The regulatory element is recognized through an INTER-pair comparison by an association with the phenotype "discordance of sibs". However, genetic information at the regulatory element is shared by both siblings within a pair, which provides an opportunity to identify the DNA-binding protein encoded *in trans* by an INTRApair comparison. If the phenotype is caused by interaction of the DNA-binding protein with the regulatory element, mildly and severely affected siblings of discordant pairs must carry different genetic information at the locus encoding the DNA-binding protein.

	Freq. among 14 discordant pairs	Freq. among 23 concordant pairs	
rs152730–rs8044970–rs63982–rs152745–rs152740 haplotype			
TTAGA	0.839	0.599	
GTCAT	0.051	0.185	Pcorr = 0.0397
GGAAT	0.036	0.129	
Other pooled[a]	0.074	0.087	
rs152730–rs8044970–rs63982–rs152745–rs152740 haplotype			
TTAGA/TTAGA[b]	0.714	0.400	
TTAGA/GTCAT	0.071	0.201	
TTAGA/GGAAT	0.071	0.127	Pcorr = 0.049920
GGAAT/GGAAT[b]	0.000	0.020	
GTCAT/GTCAT[b]	0.000	0.059	
Other pooled	0.144	0.193	

Table 1. Haplotype and diplotype distribution observed among concordant and discordant cystic fibrosis F508del homozygous sibling pairs for markers rs152730, rs8044970, rs63982, rs152745, rs152740. [a]Four rare haplotypes (Freq. < 0.05) were observed among discordant pairs, 8 rare haplotypes (Freq. < 0.05) were observed among concordant pairs. [b]To identify all genetic variants associated with intrapair discordance on the genomic fragment rs12730–rs152740, the entire 8000 bp genomic fragment was compared by Sanger re-sequencing for three homozygotes for TTAGA, two homozygotes for GGAAT and one homozygote for GTCAT (see Table 2).

Haplotype at rs152730–rs8044970–rs63982–rs152745–rs152740	Associated with intrapair disease manifestation	rs152730[a]	rs62029384	rs8044970[a]	rs8044984	rs80443907	rs152731	rs152732	rs180878	rs152733
TTAGA	Discordant	T	C	T	T	C	T	C	T	T
GGAAT	Concordant	G	T	G	G	T	C	C	T	T
GTCAT	Concordant	G	C	T	T	C	C	T	G	C
Different between TTAGA and GGAAT as well as GTCAT?		Yes	No	No	No	No	Yes	No	No	No
Haplotype at rs152730–rs8044970–rs63982–rs152745–rs152740	**Associated with intrapair disease manifestation**	**rs63982[a]**	**rs152745[a]**	**rs8062922**	**rs152744**	**rs62029385**	**rs152743**	**rs152741**	**rs57406669**	**rs152740[a]**
TTAGA	Discordant	A	G	C	A	C	G	C	C	A
GGAAT	Concordant	A	A	T	G	T	G	T	G	T
GTCAT	Concordant	C	A	C	G	C	A	T	C	T
Different between TTAGA and GGAAT as well as GTCAT?		No	Yes	No	Yes	No	No	Yes	No	Yes

Table 2. Variants observed on contrasting haplotypes identified after Sanger re-sequencing of the 8000 bp genomic fragment defined by rs152730–rs152740 associated with intrapair discordance. [a]rs152730–rs8044970–rs63982–rs152745–rs152740 were used to map the fragment associated with intrapair discordance and define the contrasting haplotypes TTAGA (associated with intrapair discordance) and GGAAT as well as GTCAT (both associated with intrapair concordance); see Table 1.

Results

Six SNPs within SCNN1B differ between concordant and discordant CF patient pairs.

We previously reported that intrapair discordance for CF disease severity is associated with three intragenic markers spanning *SCNN1B* from codon 3 to codon 293. To describe the genomic fragment for which concordant and discordant pairs carry different genetic information, we analyzed 7 previously typed markers[11] and 49 SNPs genotyped for fine-mapping in the 16p12 region, encompassing the entire *SCNN1G/SCNN1B*-locus (Fig. 1). Next,

we employed a haplotype-based fine-mapping strategy previously used to identify causative variants within this cohort[13]. To determine whether concordant and discordant pairs carried the same or different genetic information, we employed the software package FAMHAP to construct two-marker haplotypes composed of two informative markers. By using this approach, we found a significant difference in two-marker-haplotype distributions for two adjacent genomic fragments defined by markers rs152730–rs152745 and rs152745–rs152740 (Praw = 0.0075 and Praw = 0.00869, respectively; corrected for multiple testing of all informative markers at the *SCNN1G/SCNN1B*-locus Pcorr = 0.0397, Fig. 1). We concluded that the variant(s) that determine intrapair discordance are located on the genomic fragment between rs152730 and rs152740. Based on the allele frequency distribution among concordant and discordant pairs, we selected representatives for the contrasting haplotypes for Sanger resequencing of the mapped genomic fragment (Table 1). We chose three homozygotes for the haplotype TTAGA, two homozygotes for the haplotype GGAAT and one homozygote for the haplotype GTCAT for sequencing of the rs152730–rs152740 genomic fragment on contrasting alleles at markers rs152730–rs8044970–rs63982–rs152745–rs152740. We used long-range PCR to amplify an 8269 bp and an 8856 bp product encompassing the sequence of interest (Table 2). Sanger sequencing was performed using internal primers positioned every 500 bp on the forward and reverse strands. Using the software CodonCode Aligner, 476 primary reads with a median length of 737 bp were aligned to the reference sequence, assuring coverage of at least 4 reads per haplotype at each genomic position. Based on this alignment, we identified 6 SNP positions for carriers of the contrasting haplotypes for which concordant and discordant pairs had different genetic information. At the six SNPs rs152730–rs152731–rs152745–rs152744–rs152741–rs152740, alleles associated with intrapair concordance carried the haplotype GCAGTT; in contrast, alleles associated with intrapair discordance carried the haplotype TTGACA. None of these six SNPs reside within the coding sequence of *SCNN1B*. However, according to in silico analyses, they possibly alter the secondary structure of the pre-mRNA (SupplTab. 1, SupplTab. 2, SupplFig. 1).

An uncommon alternative *SCNN1B* transcript generated by intron retention in epithelial cell lines.

Because none of the six identified SNPs affect the amino acid sequence of the SCNN1B protein, we next aimed to determine whether *SCNN1B* undergoes alternative splicing. Alternative transcripts were inferred from mapped expressed sequence tags (ESTs, SupplFig. 2). As a source for polyA + RNA, we used T84 cells, which are derived from colon carcinoma, 16HBE14o-cells, which are virus-transformed non-CF respiratory epithelial cells, and CFBE41o- and CFTE29o-cells, both of which are immortalized respiratory epithelial cells derived from F508del-CFTR homozygous CF patients. Primers for combinatorial reverse-transcription PCR were designed to reflect ESTs reported for SCNN1B in the area of interest defined by SNPs rs152730–rs152740 (SupplFig. 2A). By using primers located in exons 3 and 4 or exons 3 and 5, we detected wild-type SCNN1B in all four epithelial cell lines (Fig. 1C). Additionally, we amplified a 280 bp product using one primer located within exon 3 and one primer located 100 bp 3′ of the splice site at the end of exon 3 (Fig. 1C and SupplFig. 2). We specifically investigated this intronic sequence because it has been reported to be retained in EST clones BM694355 and BU730506, which are generated from a cDNA library prepared from retinal pigment epithelium of a healthy adult male. Primers designed to detect ESTs AW844136, CV337204 and BX485038 did not amplify a product (SupplFig. 2). The 280 bp product, derived from the alternative SCNN1B transcript generated by exon read-through, was reliably amplified from T84 derived cDNA even when the RNA was pretreated with DNAse (SupplFig 2). Moreover, Sanger sequencing of the alternative product confirmed its identity at the exon 3/intron 3 border. Hence, the alternative mRNA was generated by an exon read-through event and matched the genomic sequence by the base (SupplFig 2E). Furthermore, primers placed upstream in exon 1 encoding the 5′ UTR of SCNN1B in combination with a primer placed on the retained intron sequence yielded a product from cDNA (Fig. 1C). Conversely, no signal was observed when using a primer located in the downstream exon 5 in combination with the retained intron sequence (data not shown). To summarize, cancer-derived intestinal epithelial cells as well as virus-immortalized respiratory epithelial cells expressed an alternative *SCNN1B* transcript in which intron 3 was partially retained. If translated, this alternative transcript would preserve the reading frame at the end of exon 3 and would be translated into a protein that terminates prematurely after an 26 additional amino acids derived from the retained intron sequence, producing a severely truncated SCNN1B protein of 221 amino acids.

The *SCNN1B* haplotype found in discordant pairs is enriched for predicted transcription factor binding sites.

We next assumed that an allelic association with a discordant manifestation of CF severity is mediated by factors that recognize the allele enriched among discordant pairs. To test our hypothesis, we assessed whether the 6-marker-haplotype that is associated with intrapair discordance for CF severity, i.e., whether TTGACA at the six SNPs rs152730–rs152731–rs152745–rs152744–rs152741–rs152740 attracts different DNA-binding proteins compared to the GCAGTT allele that is observed among concordant sibling pairs. To identify potential transcription factor binding sites, we used the tool "Match" (available at http://www.gene-regulation.com/)[14], which is based on a library of mononucleotide weight matrices from TRANSFAC6.0. The settings were restricted to vertebrate transcription factors and limited to minimize false negatives (estimated error rate of 10% for training data set). As an input sequence, we used both alleles at each of the six divergent SNPs and +/− 20 bp flanking sequences. Next, we compared the list of putative transcription factor binding sites between the input sequences derived from haplotypes associated with concordance and discordance and noted those predicted to interact with only one of the two contrasting alleles at each SNP. Surprisingly, only 6 binding sites were predicted for the haplotype observed among concordant sibling pairs; 21 predicted interactions were exclusively related to the six-marker-haplotype associated with intrapair discordance (Fig. 2, SupplTab 3a). Different from concordant sibling pairs, the haplotype observed among discordant sibling pairs had significantly
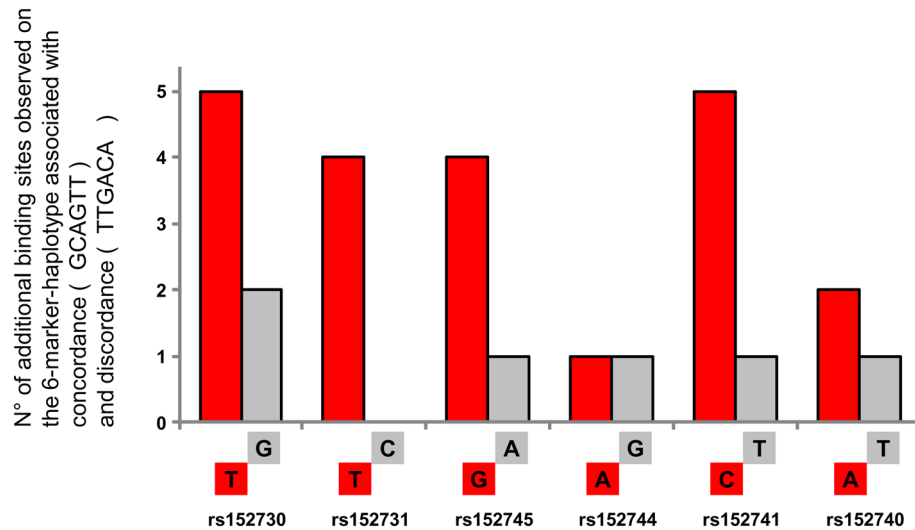
Figure 2. No. of predicted interaction partners unique to the six-marker-haplotype observed among concordant and discordant CF sibling pairs. Using the software Match at http://www.gene-regulation.com/[14], binding sites for transcription factors were predicted for the two contrasting haplotypes composed of the six SNPs rs152730–rs152731–rs152745–rs152744–rs152741–rs152740. The figure shows the number of additional, allele-specific transcription factor binding sites predicted on the haplotype associated with intrapair concordance (grey) and for intrapair discordance (red). Please note that six allele-specific binding sites were predicted for haplotype GCAGTT while for TTGACA, associated with discordance, a total of 21 additional binding sites for transcription factors are predicted. Binding sites predicted to be shared for both alleles are not shown.
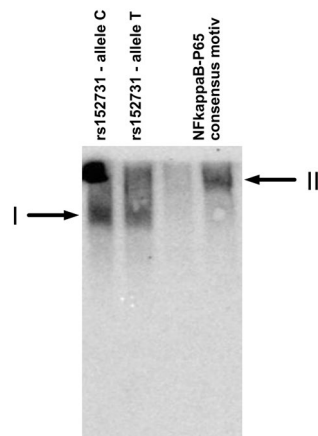
Figure 3. Electrophoretic mobility shift assay and subsequent protein sequencing (EMSA-PSeq). The dsDNA probes for SCNN1B SNP rs152731 allele C, SCNN1B SNP rs152731 allele T and the NFkappaB-P65 consensus motif[16] were incubated with T84 nuclear extract. After non-denaturing 7% polyacrylamide electrophoresis, the samples were transferred by electrotransfer in a denaturing, SDS-containing buffer onto a membrane assembly whereby the gel was covered with an uncharged nylon membrane (Hybond C, Amersham), followed by two charged nylon membranes (Hybond N+, Amersham). Unbound oligonucleotides were transferred through the uncharged membranes and could be visualized on the charged nylon membranes (SupplFig4, SupplFig5). Probes derived from rs152731 gave rise to a high molecular weight complex I which differs from complex II generated with an NFkappaB-P65 consensus motif. P65, STAT3 and STAT6 were identified in the high molecular weight complex II, whereby for this experiment, the position corresponding to the visualized signal was excised from the second gel half loaded identically without electroblotting and subjected to protein mass spectrometry.

more opportunities to interact with DNA-binding proteins ($p = 0.048$; in comparison to the expectancy value derived from 26 binding sites distributed equally between both haplotypes).

Next, we evaluated whether predicted interacting proteins of the *SCNN1B* haplotype TTGACA (SupplTab 3a) are associated with the response to amiloride upon superfusion of the nasal epithelium. The function of ENaC in vivo can be assessed based on the potential difference between the nasal epithelium and the subcutaneous space[15]. According to the nasal potential difference with the use of amiloride (indicative of ENaC-mediated sodium transport), except for GATA2Sat (Praw = 0.0456), none of the genes encoding predicted interaction partners showed an association with ENaC function (SupplTab 3b).

Interaction partners of double-stranded DNA sequences can be captured with an electrophoretic mobility shift assay following protein sequencing (EMSA-PSeq). As our in silico analysis did not extend to DNA-binding proteins with unknown binding motifs, we aimed to identify interacting proteins by performing a modified electrophoretic mobility shift assay followed by protein sequencing of the DNA–protein complex (EMSA-PSeq; see supplement for experimental details). Briefly, we used nuclear extracts derived from epithelial cells and biotinylated 35-mer dsDNA probes that centrally carry one of the contrasting alleles of the SNPs as bait. To separate the unbound probe from the probe-protein-complexes, we performed native poly-acrylamide gel electrophoresis, and the probe-protein-complexes were visualized after transfer of the separated samples to a membrane. The gel fragment corresponding to the signal generated by the probe-protein-complex was excised, and proteins within the excised gel fragment were identified by protein mass spectrometry. We used the NFkappaB-P65-consensus sequence[16] for optimization of the experimental setup. Protein mass spectrometry and MASCOT analysis identified several hundred proteins per high-molecular weight complex. To enable the recognition of proteins that incidentally comigrate together with the probe/protein complex and/or that bind to any DNA unspecifically, and/or are introduced to the sample as contaminants during handling of the gel fragment, a set of 25 EMSA-PSeq experiments were performed in parallel for noise filtering (SupplFig 3). In the EMSA-PSeq sample obtained with the NFkappaB-P65-consensus probe (Fig. 3), we detected P65 (score 78, tagged by three specific peptides) as a unique signal within a total of 25 evaluated EMSA-PSeq data sets. Additionally, the EMSA-PSeq sample baited with the NFkappaB-P65-consensus probe uniquely attracted STAT3 and STAT6, both of which were not observed in any other of the 25 EMSA-PSeq datasets.

Nucleic acid binding proteins attracted by probes of SNPs rs152730, rs152731 and rs152744 and identified as unique by EMSA-PSeq. To detect interaction partners of SCNN1B SNPs associated with intrapair discordance, we used certain experimental conditions to detect NFkappaB-P65 using a probe with a p65 consensus sequence as bait (Fig. 3). Although we varied the conditions for electrophoresis, electrotransfer and detection (see supplement for details), we were not able to obtain a reproducible high-molecular-weight complex for either allele of rs152745, rs152741 and rs152740. However, high-molecular-weight DNA–protein-complexes were observed with probes representing the two contrasting alleles at SNPs rs152730, rs152731 and rs152744. As described in detail within the Supplemental material, we have filtered the raw data set of 25 EMSA-PSeq experiments for low-expressed proteins annotated to have nucleic acid binding capabilities and attracted to only one SNP (SupplFig 3).

Even when considering the inaccuracy of complex sizes after separation on the native polyacrylamide gel, high-molecular-weight complexes were incompatible with the interaction of a single protein found by EMSA-PSeq (SupplTab 4). Thus, several independent comigrating protein–protein and protein–protein-nucleic acid complexes were likely subjected to protein sequencing. Among the proteins identified as specific for the P65 consensus probe, P65 was one of 11 proteins (Fig. 4, SupplTab 4). Sixteen, five and eight proteins were identified uniquely for SNPs rs152730, rs152731 and rs152744, respectively, by EMSA-PSeq and data mining (Fig. 4, SupplTab 4). Since P65 was found with the same experimental and data evaluation strategy used for the data sets for the SCNN1B SNPs, we assume that our true-positive protein of interest has been captured as well.

The ESRP2 genetic background is associated with the manifestation of the amiloride-sensitive sodium current in the nasal epithelium. To prove or disprove that a protein identified by EMSA-PSeq can influence ENaC function, we performed a candidate-gene-based association study among CF patients in which one of the studied phenotypes addresses ENaC function in the nasal epithelium in vivo[17]. To select a plausible genetic locus based on the list of 29 candidate proteins (Fig. 4, SupplTab 4), we excluded components of the spliceosome multiprotein complex. Among the remaining EMSA-PSeq-derived proteins, the epithelial-specific splicing regulatory protein 2 ESRP2[18,19] was the most reasonable candidate. First, CF is an epithelial disease and ESRP2 is consistent with this feature[20,21]. Second, an alternative SCNN1B transcript was observed (Fig. 1C, SupplFig 2) and ESRP2 is plausible based on its role in transcript processing.

We found ESRP2 exclusively on probes representing rs152731 whereas ESRP1 was detected on probes derived from the three EMSA-PSeq SNPs (Fig. 5). In the association study, the *ESRP1* marker ESRP1-Sat1 showed no association with the phenotype or severity of CF (Praw > 0.2). In contrast, ESRP2-Sat1 exhibited an association signal (Pbest = 0.04) that was confirmed with a second microsatellite and 5 SNPs (Pbest = 0.0131, Pcorr = 0.068 for multiple testing of 7 markers; Fig. 5) for the manifestation of amiloride-sensitive sodium conductance, a hallmark of ENaC function.

The two rs152731 alleles are nonequivalent concerning ESRP2 binding. ESRP2 was detected as a factor binding to the C to T SNP rs152731 by EMSA-PSeq (Fig. 4) by using polyacrylamide gels with a separating distance of 4 cm. To detect ESRP2-rs152731 binding complexes, we employed EMSAs using polyacrylamide gels with a separating distance of 20 cm (Fig. 6A,B). To exclude nonspecific binding, signals obtained with an

Column groups — Probe / Allele / Sample:

		Neg Ctrl.	Neg Ctrl.	Neg Ctrl.	Neg Ctrl.	Neg Ctrl.	Neg Ctrl.	Neg Ctrl.	NFkappaB-P65 -----	IFNGR1-SNP1 C	IFNGR1-SNP1 C	IFNGR1-SNP1 T	IFNGR1-SNP1 T	IFNGR1-SNP2 C	IFNGR1-SNP2 T	rs152730 G	rs152730 G	rs152730 T	rs152730 T	rs152731 C	rs152731 T	rs152731 T	rs152744 A	rs152744 A	rs152744 G	rs152744 G
Protein	Exp. ppm	#102	#104	#56	#57	#240	#242	#244	#245	#146	#152	#147	#153	#27	#28	#148	#154	#149	#155	#241	#243	#26	#150	#156	#151	#157

lowly expressed & enriched on NFkappaB-P65 probe only

Protein	Exp. ppm	#102	#104	#56	#57	#240	#242	#244	#245	#146	#152	#147	#153	#27	#28	#148	#154	#149	#155	#241	#243	#26	#150	#156	#151	#157
CARHSP1	132.0								74																	
CHAF1B	17.2								49																	
EYA3	6.0								61																	
NCOA7	1.1								66																	
PKN1	2.8								85																	
PSIP1	74.1								73																	
RELA	76.0								78																	
STAT3	174.0								45																	
STAT6	93.5								30																	
USP39 ***	28.6								81																	
YAP1	19.2								230																	

lowly expressed & enriched on rs152730 probes only

Protein	Exp. ppm	#148	#154	#149	#155
BAZ2B	0.7	32			
CHD5	2.1				46
DIEXF	6.5				129
GTF2F2	28.6			50	44
HNRNPUL2	116.0				62
PCBP4	5.7	31			
PRMT6	3.6		90	95	
PTBP3	102.0	32			
REPIN1	9.2		71		
RRP36	7.4	42			
TBL1X	7.5	31		33	
TBL1Y	2.1	31		33	
THOC1 ***	6.5			30	
TLX2	0.5			32	
TSR1	9.4	36		33	
UTP3	30.0	39			

lowly expressed & enriched on rs152731 probes only

Protein	Exp. ppm	#241	#243	#26
ESRP2	6.3	119	92	
GAR1	25.4			46
RBMXL1	80.2			37
RNASEH2A	21.2	69	96	
VSIG8	0.4			93

lowly expressed & enriched on rs152744 probes only

Protein	Exp. ppm	#150	#156	#151	#157
ACIN1 ***	10.8	51		78	
DDB2	5.3		79	106	88
DDX19B	57.1			57	
EXOSC9	2.3				63
PARN	3.1	61	75		133
PRPF39	2.6			46	57
PUF60 ***	145.0			40	
SUPT5H	11.4				70

components of KEGG spliceosome * includes** USP39, THOC1, ACIN1, PUF60 , RBMX and SRSF1,2,3,4 and 6 and hnRNPs A1, A2, A3, C, K, M and U (AQR, BCAS2, BUD31, CCDC12, CDC40, CHERP, CRN, CWC15, DDX23, DDX42, DDX46, DDX5, DHX15, DHX16, DHX38, DHX8, FUSIP1, ISY1, LSM7, LSM8, NCBP2, PHF5A, PLRG1, PPIE, PPIH, PPIL1, PQBP1, PRP2, PRP38, PRPF18, PRPF19, PRPF3, PRPF31, PRPF38A, PRPF38B, PRPF4, PRPF40A, PRPF40B, PRPF6, PRPF8, RBM17, RBM22, SART1, SF3A1, SF3A2, SF3A3, SF3B1, SF3B14, SF3B2, SF3B4, SF3B5, SFSR7, SLU7, SMNDC1, SNRNP200, SNRNP27, SNRNP40, SNRP70, SNRPA1, SNRPB, SNRPB2, SNRPC, SNRPD3, SNRPG, SNU114, SNU13, SNU23, SNW1, SR140, SYF1, SYF2, TCERG1, THOC2, THOC3, THOC4, TRA2, TXNL4A, U2AF1, UAP56, WBP11: not found)

Protein	Exp. ppm	#245	#146	#152	#147	#153	#27	#28	#148	#154	#149	#155	#241	#243	#26	#150	#156	#151	#157
ESRP1	11.8		103		142					208		140	141			100	143	187	206
CDC5L	15.8							38											
CTNNBL1	12.8			65		165				195		96					146		100
EFTUD2	104.0		53	268	380		36	34	52	400		315				177		156	73
EIF4A3	169.0		191	95	174	112	187		97	195	83					112		156	153
HSPA1A	518.0	1059	408	447	472	518	32	229	393	494	447	336				371	665	539	747
LSM2	239.0			36		148			51	145		100	144						
LSM3	73.5		35			35			54		32								
LSM4	205.0														62				
LSM5	44.9					46													
LSM6	352.0			61	48	50			50	75	50	91	71	47					63
MAGOH	87.9		81	43					46	42									
NCBP1	39.8		136	48	45	58			45		45						118		174
PCBP1	584.0		77	189	129	148			136	180	163	213	74	78		121	183	192	206
RBM25	19.5		42						36										
RBM8A	81.5		92	49	97				82	118	103	132				83		107	
RP9	5.3		40		42					42									
SF3B3	169.0		135	444		536			38	496		461	232	133		413	523	263	470
SNRPA	147.0			39	43				44	46	41								
SNRPD1	308.0	86	142	113	77	105			145	185	32	142				80		117	
SNRPD2	433.0									44									
SNRPE	148.0		35		45	67			61	53	37	51							
SNRPF	264.0					38				37	30								
U2AF2	109.0		103	50	42	89			54	57	51	49				141		140	44

serine/arginine-rich splicing factors (SRSF5,7,8,9,10,11: not found)

Protein	Exp. ppm	#146	#152	#147	#153	#148	#154	#149	#155	#241	#243	#150	#156	#151	#157
SRSF1 ***	337.0		102	49	64	95	33		112		69				
SRSF2 ***	475.0		113		93		115		149						91
SRSF3 ***	603.0	30							47			70		93	
SRSF4 ***	93.6		91		83		107		123						
SRSF6 ***	101.0	189			133			172	233	53		106	135	103	

hnRNPs

Protein	Exp. ppm	#56	#57	#240	#242	#245	#146	#152	#147	#153	#27	#28	#148	#154	#149	#155	#241	#243	#26	#150	#156	#151	#157
HNRNPA1 ***	486.0					255	348	220	260	203			308	243	406	213	197	99		281	186	454	281
HNRNPA1L2	216.0						325		269				308		373			99					
HNRNPA2B1 ***	1202.0					93	383	323	350	393			404	437	516	473	253	234		162	482	346	387
HNRNPA3 ***	733.0						38	145	31	142			36	167	45	143		113		180		149	200
HNRNPAB	417.0					76	252		265	230			135	323	170	305	158	206	94	222	185	252	225
HNRNPA0	250.0								31														
HNRNPC ***/CL1	150.0						60						58		65				60				
HNRNPD0	1017.0					159	341	284	336	279			284	385	209	318	197	255	65	343	221	332	277
HNRNPF	305.0					145	402	357	502	400			420	406	439	449	320	304		394	220	365	
HNRNPH1	350.0					187	252	312	265	347			287	390	275	484	245	281		267	430	223	277
HNRNPH2	286.0						143		141				164	145									
HNRNPH3	336.0								97	125				115		96		51			127		119
HNRNPK ***	863.0		93			303	394	389	363	535			468	569	401	507	358	359	31	535	525	500	482
HNRNPL	371.0					112			537	40		552	42	459	48	600	180	238		343	404	350	485
HNRNPLL	23.2						30															51	
HNRNPM ***	376.0									62				79		97					83		67
HNRNPQ	282.0			85			150	340	240	493		58	291	472	304	459	475	521		430	332	373	354
HNRNPR	153.0					282	178	290	166	383		58	178	420	220	347	231	217		343	195	331	205
HNRNPU ***	656.0	87				96	616	759	689	732		33	517	850	624	973	76	137		895	597	990	676
HNRNPUL1	75.8										65	38		64		91	147	119					
HNRNPUL2	116.0															62							

◀**Figure 4.** Proteins specific for rs152730, rs152731 and 152744 and abundantly present components of the spliceosome identified by EMSA-PSeq. Numbers correspond to MASCOT scores obtained for protein sequencing (primary data provided in: raw data xls supplement R3) which is a probability score that describes $-10LOG_{10}(P)$, with P being the probability that the detected protein represents a significant match in this data set. ESRP2 was captured on rs152731 probes with a MASCOT score of 119 (C-allele, $p = 10^{-12}$) and 92 (T-allele, $p = 10^{-9}$), respectively. Data is shown for proteins specific for one SNP (top 40 rows), proteins that exhibit partial specificity (spliceosomal proteins) and proteins that were unspecifically found in the majority of probe-derived samples but were absent from the majority of negative controls (hnRNPs). Proteins specific for one probe or one SNP which were identified from the dataset derived for 25 EMSA-PSeq samples based on 1. Their absence from empty negative control samples 2. Their absence from EMSA-PSeq samples obtained with probes for other analyzed SNPs 3. Their expression levels of less than 200 ppm according to pax-db and 4. Their annotated capabilities to interact with nucleic acids (see SupplFig 3 for details). Data on ESRP1 is provided for comparison as both, ESRP1 and ESRP2 were interrogated in an association study (see Fig. 5). As many RNA-binding proteins were found among the proteins identified as SNP-specific, and as an alternative SCNN1B transcript has been observed, EMSA-PSeq data was systematically screened for constituents of the spliceosome. Exp. ppm denotes expression values retrieved from the Protein Abundances Across Organisms-database pax-db[61] for the data set "colon integrated", except for EYA3, BAZ2B, DIEXF, PRMT6, TLX2, UTP3, LSM3, LSM5, RP9 [Whole organism, (PeptideAtlas, Aug 2014)] and TLX2 [Whole organism (Integrated)]. Samples #26, #27 and #28 have been derived from denaturing SDS-polyacrylamide gels, hence these lack the multiprotein complexes observed in all other samples derived from native polyacrylamide electrophoresis. Details on experimental conditions are described in SupplTab 9.

....

antibody directed against ESRP2 were compared to protein-DNA-complexes probed with IgG (isotype control). Signal intensities for rs152731-C were comparable in IgG and anti-ESRP2-Ab lanes. In contrast, signals obtained for rs152731-T were stronger with anti-ESRP2-Ab than with IgG. Normalized signal intensities for rs152731-T were higher than those for rs152731-C ($p = 0.013$).

Next, we addressed whether ESRP2 recognizes rs152731 directly. We used a coimmunoprecipitation (co-IP) experiment using a biotinylated rs152731 EMSA probe as bait with a raw nuclear extract. After precipitation with an anti-biotin-Ab, detection of ESRP2 by Western blotting in co-IP samples verified that all components sufficient for ESRP2 to bind rs152731 are present within the nuclear extract and/or the components used for the EMSA preceding co-IP (Fig. 6C,D). In three independent co-IP experiments comparing the two contrasting rs152731 alleles, ESRP2 signals derived from probes with rs152731-T yielded stronger signals than probes for rs152731-C. In summary, the results from two different techniques confirmed that rs152731-T binds ESRP2 better than does rs152731-C. Thus, the allele rs152731-T associated with intrapair discordance (Tables 1, 2) attracts ESRP2 to a greater extent and is more vulnerable to its genetic variations (Fig. 5).

ESRP2 knockdown alters global SCNN1B expression in 16HBE14o-cells.

To determine whether ESRP2 influences the transcript species generated from SCNN1B, we downregulated ESRP2 by siRNA in T84 and 16HBE14o-cells and quantified the wild-type and the alternative transcripts by qPCR (Fig. 7). Silencing of ESRP2 in T84 cells resulted in highly variable changes in SCNN1B transcripts. Moreover, the observed changes were comparable to those using scrambled control siRNA. In contrast, no systematic effect of scrambled control siRNA was detected in 16HBE 14o-cells ($p = 0.50$ for wild-type, $p = 0.32$ for alternative SCNN1B transcript; Wilcoxon signed-rank test). Moreover, the amounts of wild-type and alternative SCNN1B were increased in 16HBE14o-cells upon treatment with siRNA directed against ESRP2 ($p = 0.015$ for wild-type; $p = 0.054$ for alternative SCNN1B transcript). Comparison of changes in wild-type and alternative SCNN1B transcript levels assessed by paired ΔΔCt levels for both amplicons indicated that downregulation of ESRP2 induced expression of functional wild-type SCNN1B in 16HBE14o-cells ($p = 0.081$, Wilcoxon signed rank test).

The upper respiratory tract is the origin of 16HBE14o-cells while T84 cells are derived from intestinal cells; 16HBE14o-cells are virus-immortalized and T84 are colon cancer cells. Because ESRP2 plays a prominent role in cancer[18,19], it is not surprising that these two cell lines behaved differently in the ESRP2 gene silencing assay. Furthermore, the baseline expression of both SCNN1B transcripts was lower in 16HBE14o than in T84 cells. Interestingly, T84 cells carry only one, but 16HBE14o-cells carry two of the ESRP2-receptive alleles T–G–A–C at markers rs152731–rs152745–rs152744–rs152741. This may explain the observed differences between 16HBE14o and T84 cells in response to ESRP2 silencing. Homozygosity for T–G–A–C is associated with intrapair discordance among CF twins and siblings.

Discussion

The design of the European CF twin and sibling study was inspired by Risch and Zhang[22] who proposed that the use of sibling pairs with extremely concordant or discordant phenotypes will advance the discovery of quantitative trait loci in humans[22–24]. Due to the high power of this approach, it was estimated that the genotyping load for studies undertaken with an extreme sib-pair design, selecting for patient pairs who exhibit phenotypes below the 30th or above the 70th centile, can be reduced by up to 40-fold[22]. Based on this strategy for patient recruitment, 37 F508del-CFTR homozygous sibling pairs of 318 cystic fibrosis affected patient pairs were selected for the association study by a ranking algorithm[12]. The selected sibling pairs were comparable in terms of their birth cohort[11]. This strategy helped us to minimize the influence of a major nongenetic confounder[25], i.e., complex therapeutic management which has improved the life expectancy of CF patients by several decades. Pulmonary and gastrointestinal disease manifestations were assessed quantitatively by CF population centiles for the normalized forced expiratory volume in 1 s (FEV1) and by weight as a percentage of predicted weight for height. For
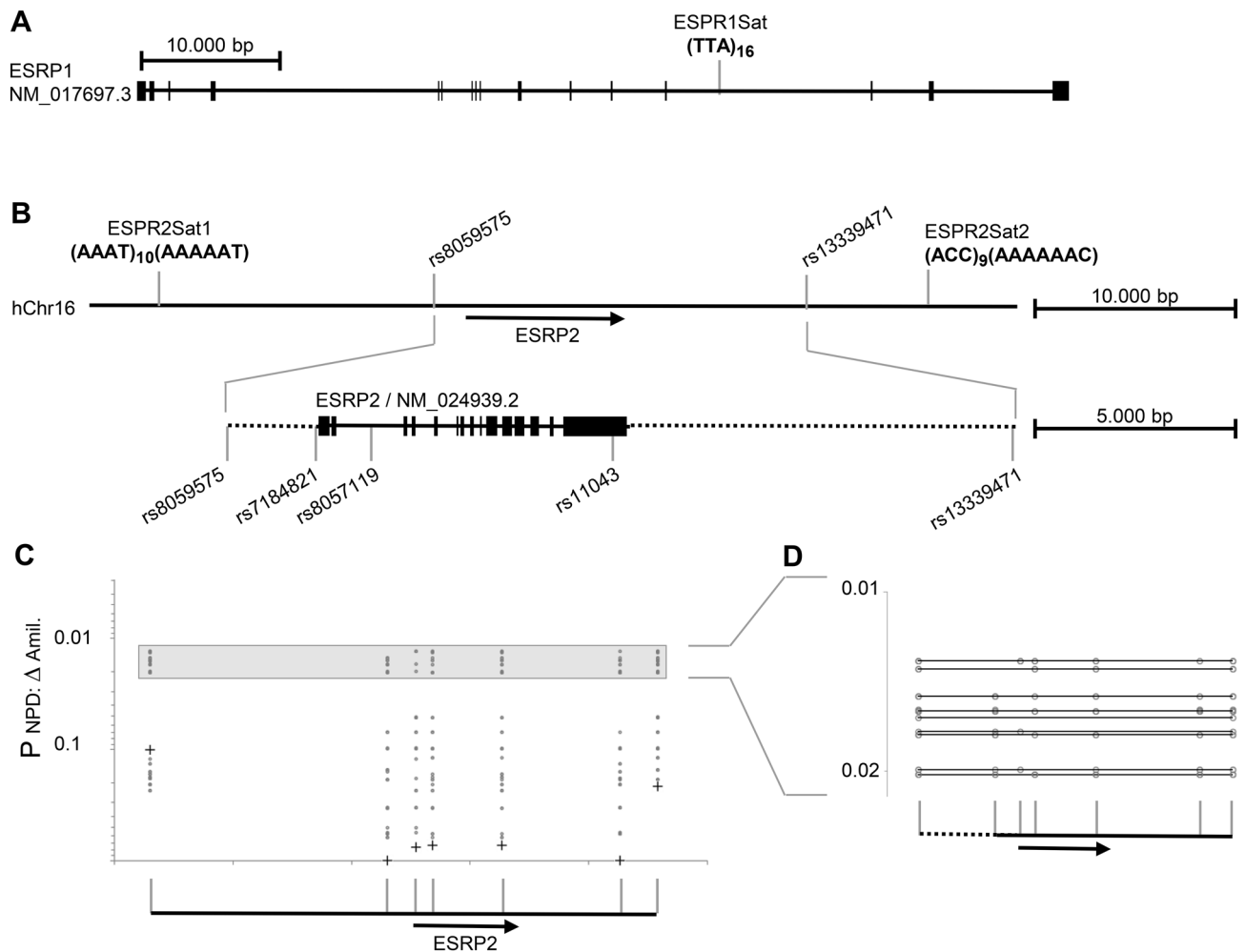
Figure 5. Association study on ESRP1 and ESRP2. (**A**) Position of intragenic microsatellite marker ESRP1Sat within ESRP1. ESRP1 was selected as control as this protein was identified at most probes used for EMSA-PSeq. In other words, no enrichment of ESRP1 with any probe used for EMSA-PSeq has been observed. No association of the ESRP1Sat allele distribution with CF disease severity or manifestation of the basic defect was seen (Praw > 0.2; data not shown). (**B**) Five SNPs and two intergenic microsatellite markers were typed on ESRP2. Markers on the ESRP2 locus were in strong linkage disequilibrium recognized by D′ values of 0.884, 0.918, 0.884 and 0.977 on the four segments between two adjacent SNP markers spanning ESRP2. (**C,D**) Analyzing markers in ESRP2, two association signals were observed: intrapair comparison of mildly versus severely affected sibling of discordant pairs was skewed (Praw = 0.1092 for ESRPSat2; Praw = 0.104 for the combination of ESRP2Sat1 and rs8057119; Pcorr = 0.323 corrected for multiple testing of 7 markers, data not shown) and an allelic association with the response to superfusion of the nasal epithelium with amiloride assessed by nasal potential difference measurement was seen (Pbest; raw = 0.0131; Pcorr = 0.068 corrected for multiple testing of 7 markers, data shown in (**C,D**). For this case-reference association study, F508del homozygotes with contrasting responses upon superfusion of the nasal epithelium with amiloride were compared in their ESRP2 genetic background. Cases were 16 patients with an amiloride response of 31 mV, references were 15 patients with 21 mV or less[17]. Raw *p* values of single markers are shown as "+" in (**C**), for raw *p* values obtained for multimarker haplotypes, marker positions are shown as "◯" in (**C,D**). For 0.02 > Praw > 0.01, multimarker combinations are displayed in (**D**).

these two parameters, we selected sibling pairs with extreme phenotypes in the upper and lower 25%[17,26]. Our selection criteria were in line with recommendations proposed by Risch and Zhang[22], however, these criteria resulted in a small study population, thus limiting the power of the genetic association study. Moreover, since our study population is of white European descent, a group in which F508del-CFTR is the most common mutation causing CF, we cannot be certain that our findings can be applied to other populations.

Risch and Zhang concluded from their simulation studies that "extremely discordant sibling pairs represent a powerful design for the association studies of candidate genes"[22], and our findings fully support this idea. The

use of sibling pairs with extreme clinical phenotypes has been applied before[27,28], and our data support the notion that gene–gene interactions mediated by factors encoded *in trans* of the studied locus can be distinguished in an association study when mildly and severely affected siblings of discordant pairs are compared (Fig. 8).

Hence, in this study we identified a haplotype within SCNN1B associated with intrapair discordance in CF sibling pairs (Fig. 1, Tables 1, 2). We investigated in silico (Fig. 2) and experimentally (Figs. 3, 4) the occurrence of DNA-binding proteins interacting differentially with single or multiple SNPs within the haplotype. We were able to recognize ESRP2 as a candidate for validation among nucleic acid binding proteins, which showed an association with SCNN1B/ENaC function (Fig. 5). We further employed EMSA and co-IP as two different experimental approaches to support the allele-dependent interaction between rs152731 in SCNN1B and the nucleic acid binding protein ESRP2 (Fig. 6). Moreover, we demonstrated that siRNA mediated silencing of ESRP2 in respiratory epithelial cells causes an alteration in global SCNN1B expression (Fig. 7). Altogether, our findings consistently support the idea that SCNN1B and ESRP2 are interacting partners and that ESRP2 is capable of altering the SCNN1B transcript repertoire. It is plausible that this interaction can alter ENaC function and has an influence on the manifestation of CF.

Our proof-of-principle study has several limitations. Specifically the findings cannot fully explain whether the regulatory element within *SCNN1B* leads to the alternative SCNN1B transcript, from which a severely truncated, 221 amino acid SCNN1B may be translated. Similar truncated SCNN1B isoforms of 217 and 306 amino acids have been observed in patients with systemic pseudohypoaldosteronism[48]. In addition, heterologous expression of these truncated SCNN1B mutants in *Xenopus* oocytes showed that they can assemble with wild-type alpha- and gamma ENaC subunits[48]. These SCNN1B mutants resulted in lower ENaC activity (by 3–7%) than the wild-type protein[48]. Under physiological conditions, parallel expression of two SCNN1B transcript species, one of which yields a truncated SCNN1B isoform upon translation, can reduce SCNN1B function but the extent remains unclear.

Using EMSA-Pseq for the identification of DNA-binding proteins has previously been conceived[29–31]. To examine whether our EMSA conditions allowed the formation of high-molecular-weight multiprotein complexes with coherent DNA–protein-interactions in vivo, we investigated protein-DNA-complexes using an NFkappaB-P65-consensus as bait. In line with published data[32–35], this probe attracted NFkappaB-p65 and its known interaction partners, such as STAT3 and STAT6. From the EMSA-Pseq of SCNN1B probes, we selected ESRP2 as a candidate for further validation experiments. This selection was based on the fact that this protein is characteristically expressed in epithelial cells and that similar to other SNP-specific proteins recognized by EMSA-PSeq, ESRP2 has been well-characterized as an RNA-binding protein[18,19].

The defining border between RNA- and DNA-binding proteins has recently softened because typical DNA-binding proteins have become known to target long noncoding RNAs, defining dual-recognition nucleic acid binding proteins[36,37]. A growing number of nucleic acid binding proteins have been recognized to interact with both nucleic acid species[37–44] and genomic DNA[45]. The ability to bind to DNA and RNA simultaneously designates a dual recognition protein capable of shuttling between both nucleic acid types during transcription[38]. During transcription, DNA and RNA are physically close, and therefore, cotranscriptional processes, such as pre-mRNA splicing, can be mediated by putative dual-recognition proteins, such as hnRNP splicing regulatory factors[46,47].

Nevertheless, we cannot exclude that SCNN1B may have other important interacting partners in addition to ESRP2 that were not discovered in this study. In this work, we analyzed only three of six SNPs by EMSA-PSeq. Furthermore, while we filtered our primary protein sequencing data using a positive control (NFkappaB-P65) and several technical controls to recognize contamination, we did not incorporate a protein–probe-interaction with low binding affinity, which might enable the detection of weak interacting partners. In the future, the resolution of the EMSA-Pseq can be improved by using scrambled probes to control for nonspecific binding and by incorporating the false-positives captured in the data evaluation strategy. Moreover, EMSA-PSeq utilizes mass spectrometry to identify proteins. In contrast, nucleic acids such as long noncoding RNAs with the potential of interacting with the DNA sequence cannot be identified in this experimental setting, and thus, their relevance needs to be verified by other methods.

The haplotype associated with intrapair discordance covers a genomic segment of 8 kb, implying that a synergistic relationship of more than one interaction partner is responsible for the selective advantage that underlies the maintenance of linkage disequilibrium over such a distance. Thus, it is unlikely that the SCNN1B function can be fully understood based on studying single SNPs. Regardless, we are convinced that the methodology proposed herein—analysis of clinically discordant sibling pairs in combination with EMSA-PSeq—aids in our understanding of how some of the 10,000 SNPs identified by GWAS as being meaningful (www.genome.gov/gwastudies. Accessed at 03.02.2015) contribute to the manifestation of phenotypes in humans. As gene–gene interactions have been suggested to account for the phenomenon termed "missing heritability"[49], the discovery of regulatory interactions such as those between ESRP2 and *SCNN1B* might help to annotate existing GWAS data sets that have been performed with sibling pairs[50–52].

Methods
Details on the experimental procedures are provided in the supplement.

Cell culture.
Biomaterials were derived from T84 colon cancer cells and immortalized respiratory epithelial 16HBE14o-, CFBE41o- and CFTE29o-cells.
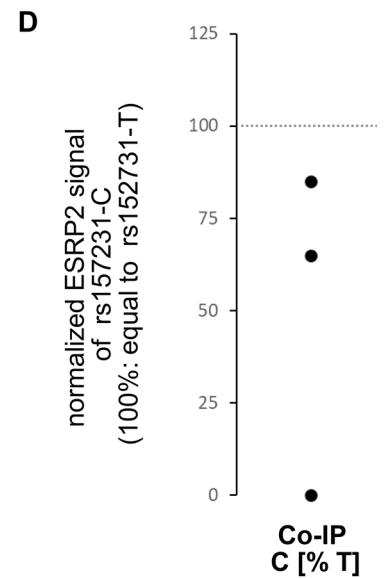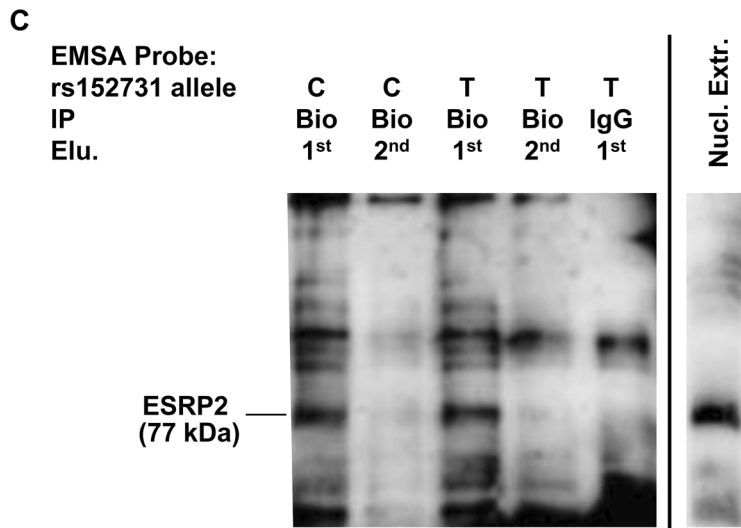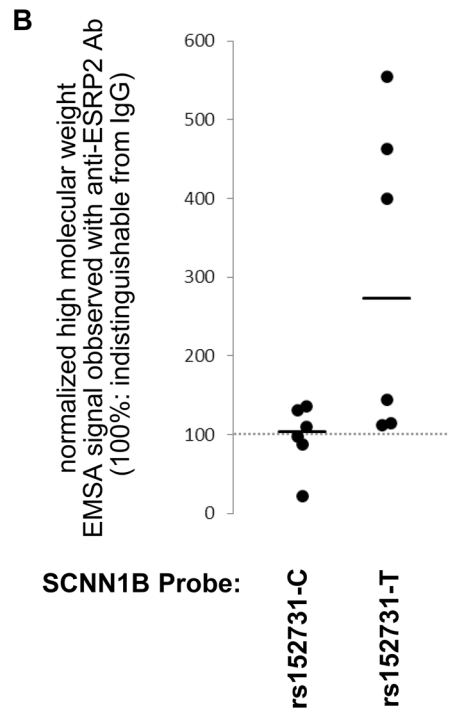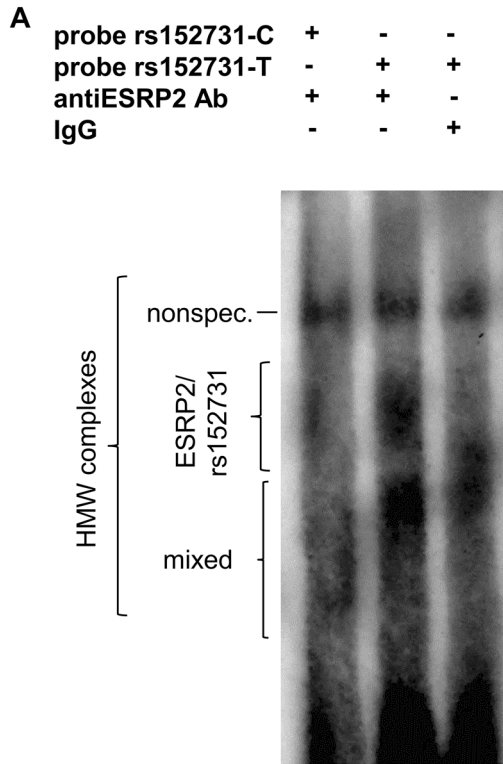
◀ **Figure 6.** Binding of ESRP2 to rs152731-C and rs152731-T in EMSA and co-immunoprecipitation. (**A**) EMSA experiment using nuclear extract proteins derived from 16HBE14o. Signals marked "HMW" (high molecular weight) encompass different entities that can be differentiated by comparing the lane with anti-ESRP2-Ab to the control lane using IgG. Presence of ESRP2 in these HMW multi-protein complexes is deduced from the increased signal in the presence of an anti-ESRP2-antibody in comparison to the neighboring IgG control lane. Nonspec.: nonspecific signal, observed similarly with both, IgG control and anti-ESRP2-Ab. ESRP2/rs152731: signal observed predominantly in the presence of anti-ESRP2-Ab. Mixed: protein–DNA-complexes partially containing ESRP2 and partially nonspecific. The number of independent experiments is N = 6 whereby nuclear extracts were from 16HBE14o (N = 5) and T84 (N = 1). The blot shown is an exemplificative blot out of the 6 performed replicates (source data is provided with this manuscript). (**B**) Signals obtained with rs152731 probes and an antibody directed against ESRP2 were compared to signals obtained when IgG was used instead of the specific antibody. Data from such paired samples are shown from six independent experiments that differ concerning nuclear extract and electrophoresis conditions. Signals labelled "ESRP2/rs152731" in Fig. 6A. Could be distinguished clearly from the DNA–protein-complex labelled "mixed" in three experiments. Normalized data obtained for rs152731-C and rs152731-T probes were compared by Mann–Whitney U test whereby signals for high molecular weight complexes observed for rs152731-T were stronger ($p = 0.013$). (**C**) 150 µg nuclear extract from T84 cells was incubated with 500 nmol dsDNA probes for SCNN1B SNP rs152731 allele C and SCNN1B SNP rs152731 allele T in an EMSA experiment of a total reaction volume of 40 µl for 3 h at 37 °C. Protein–DNA-complexes were captured with an antibody directed against biotin (ab19221, Abcam, Cambridge, UK), or with IgG as a negative control (indicated by "IP"), and precipitated using protein-G-agarose-beads at 4 °C for 16 h. The supernatant was discarded and proteins were dissociated from agarose beads using Laemmli buffer in two elution steps (indicated by "Elu." 1st and 2nd) of 30 µl each (10 min, 50 °C), loaded on a denaturing 11% polyacrylamide gel and transferred to a protean supported membrane with 20 µM pore size by tank-blot for 50 min at 100 V. ESRP2 immunoreactive bands were visualized with an antibody directed against ESRP2 (ab113486, Abcam, Cambridge, UK). A lane loaded with nuclear extract only was used as a control to identify ESRP2 (predicted molecular weight: 77 kDa). The number of independent experiments is N = 3 whereby nuclear extracts were from T84. The blot shown is an exemplificative blot out of the 3 performed replicates (source data is provided with this manuscript). (**D**) In the example shown in (**C**), ESRP2 signals derived from the probe with rs152731 allele C correspond to 85% of the signal observed for rs152731 allele T. Two replicate experiments, not shown here, have demonstrated 0% and 65% intensity on the allele rs152731-C, respectively. While the C-to-T-ratios observed in three co-IP experiments were highly variable, these three experiments consistently indicate that rs152731-C binds ESRP2 less tightly than rs152731-T.

RNA preparation. For RNA isolation, cells were cultured in plates, grown to confluency, snap-frozen in the gaseous phase of liquid N_2 and stored at − 80 °C. RNA was extracted using QIAamp RNA Blood Mini Kit (52,304, Qiagen, Hilden, Germany) and RNase-free DNase Set (79,254, Qiagen, Hilden, Germany).

Oligonucleotides for PCR. Sequences of primers used for genotyping and combinatorial PCR are listed in the supplement (SupplTab 5).

Extraction of nuclear proteins. Nuclear extracts were prepared according to published standard methods[53] (SupplTab 6). To prevent carryover of the high-salt buffer used for lysis of nuclei, nuclear proteins were dialyzed against low-salt HEPES buffer. The completeness of dialysis was verified by measuring the conductivity of the nuclear extract with a needle probe (customized, Technische Forschungswerkstätten of the Hannover Medical School). The quality of nuclear proteins was ascertained by verifying the conductivity of the final extract after dialysis and by noting the absence of degradation by SDS-electrophoresis followed by Coomassie staining.

EMSA-PSeq. To identify proteins that interact with a specific DNA sequence, we performed an EMSA experiment, visualized the shifted band, captured the DNA–protein complexes by excising the corresponding region of the polyacrylamide gel and then performed protein mass spectrometry. The composition of the EMSA binding buffer was adjusted to reflect the nuclear milieu[54] (SupplTab 7). All experimental details and data analysis methods are provided in the supplement (SupplTab 8, SupplTab 9, SupplFig4, SupplFig5, SupplFig6).

Coimmunoprecipitation. Biotinylated 35-mer double-stranded DNA probes for the rs152731 allele C and allele T were incubated with nuclear extract in an EMSA experiment. The DNA–protein-complexes were precipitated using an anti-biotin antibody and protein G agarose beads. The protein–DNA-complexes were eluted from the beads in three consecutive steps, and Western blotting with anti-ESRP2 was used to detect ESRP2 in the precipitated protein–DNA-complexes. IgG instead of the anti-biotin antibody served as a negative control in all experiments. ESRP2 was identified using a signal from an unpurified nuclear extract developed in parallel in each Western blot experiment.

siRNA-mediated downregulation of ESRP2 in epithelial model cell lines and SCNN1B transcript analysis by real-time RT-PCR. siRNA directed against ESRP2 and scrambled control siRNA was purchased from GE Healthcare (mixture of four siRNAs, on-target plus pool, GE Healthcare). T84 and 16HBE14o-cells were transfected for 24 h and 48 h using a protocol supplied by the manufacturer with 10 µl of Dharmafect 1 and 100 pmol siRNA in 2 ml of cell culture medium per well of a 6-well plate. Commercially
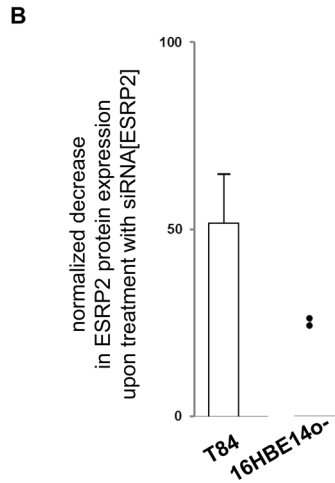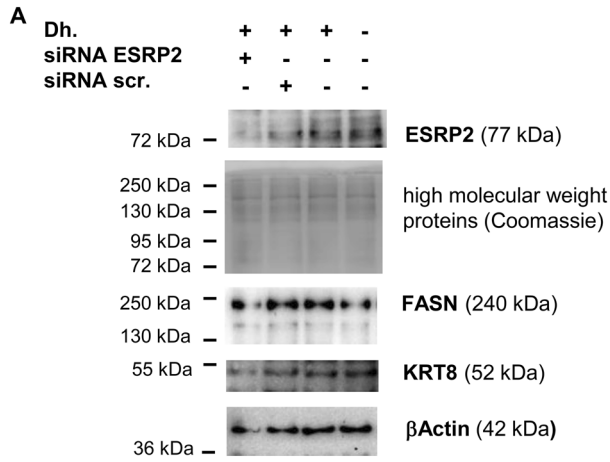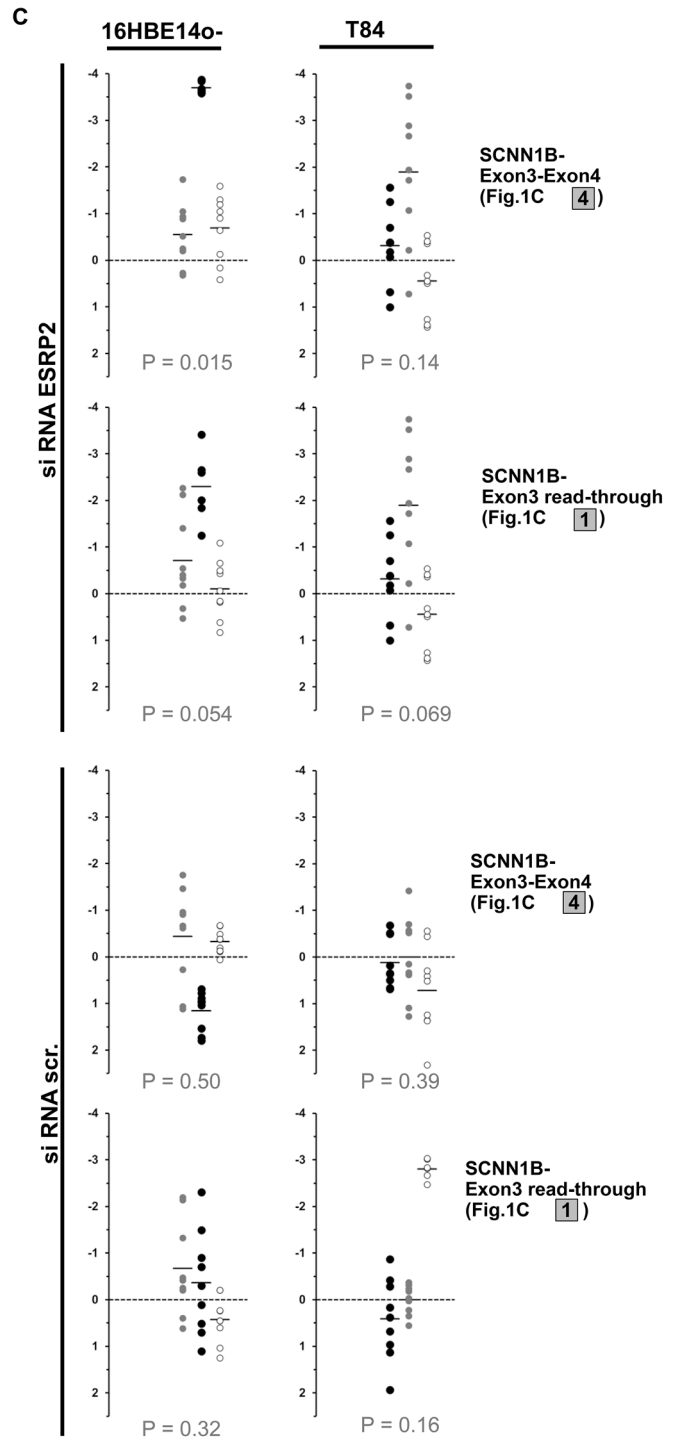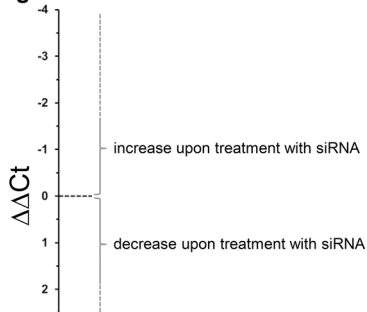
◀ **Figure 7.** siRNA-mediated downregulation of ESRP2 affects the SCNN1B transcript spectrum in 16HBE14o-cells. siRNA directed against ESRP2 (siRNA ESRP2; mix of four siRNAs, on-target plus pool, GE Healthcare) or scrambled control (siRNA scr., GE Healthcare); was provided to T84 and 16HBE14o-cells. ESRP2 protein expression and SCNN1B mRNA expression levels were compared between cells supplied with the transfection agent Dharmafect 1 (Dh.) only versus cells that received Dharmafect and siRNA directed against ESRP2. Experiments in both cell lines were conducted from two (T84) or three (16HBE14o) different passages as biologically independent replicates whereby in each of these experiments, cell culture plates were treated in parallel for the following conditions: growth control (no Dharmafect transfection agent or siRNA; 8 plates), transfection agent control (with Dharmafect but without siRNA; 8 plates), treated with siRNA directed against ESRP2 (8 plates) and treated with scrambled siRNA (8 plates). Four out of eight plates were harvested at 24 h post-transfection and the remaining four plates were analyzed after 48 h post-transfection. Nuclear extracts were analyzed for ESRP2 protein expression from one plate in these sets of four (**A,B**). RNA was isolated from the remaining three plates (**C**). (**A**) Nuclear extracts derived from 16HBE14o-cells treated with 100 pmol siRNA directed against ESRP2 (siRNA ESRP2) or scrambled control (siRNA scr.), transfection control and growth control were analyzed for ESRP2 protein by western blot. The intensity of Coomassie-stained high molecular weight bands was used as a western blot loading control. Signals for fatty-acid synthase (FASN), keratin 8 (KRT8) and βActin are provided as control detections. (**B**) Reduction of protein expression levels of ESRP2 by siRNA directed against ESRP2 was judged from nuclear extracts. All but two sample sets from 16HBE14o-showed very low or absent ESRP2 protein signals and were excluded from quantitative analysis. (**C**) SCNN1B wild-type (exon3–exon4, see Fig. 1C, amplicon 4) and read-through alternative transcript (see Fig. 1C, amplicon 1) were visualized from cDNA derived from control and siRNA-treated cells. Data was evaluated based on the threshold cycle Ct provided for an SCNN1B amplicon in comparison to the housekeeping gene aldolase from technical duplicates for each qPCR reaction. Independent experiments derived from 12 cell culture plates, three each for the conditions growth control (no Dharmafect or siRNA), transfection control (Dharmafect only), siRNA ESRP2 and siRNA scr. were included into the analysis if at least two out of three samples from all conditions gave valid results for all three amplicons (n = 3 for T84 and n = 3 for 16HBE14o). All samples for 16HBE14o 24 h after transfection had to be excluded as SCNN1B transcripts could not be detected in the majority of qPCR reactions. ΔCt values were calculated for each sample as Ct[SCNN1B]-Ct[aldolase]. To compare differential expression between two triplicate sets, ΔΔCt was calculated for all possible nine combinations comparing three case and three reference samples. To assess the influence of Dharmafect, ΔΔCt[Dharmafect] was calculated as ΔΔCt[Dharmafect] = ΔCt[growth control] – ΔCt[transfection control]. The influence of an siRNA ΔΔCt[siRNA ESRP2] and ΔΔCt[siRNA scr.] treatment was calculated as ΔΔCt[siRNA ESRP2] = {ΔCt[siRNA ESRP2] – ΔCt[transfection control] – {mean ΔΔCt[Dharmafect]} and ΔΔCt[siRNA scr.] = {ΔCt[siRNA scr.] – ΔCt[transfection control]} – {mean ΔΔCt[Dharmafect]}. ΔΔCt[siRNA ESRP2] and ΔΔCt[siRNA scr.]. qPCR Data is shown for three independent experiments in each cell line (N = 3) whereby values derived from one biological replicate are displayed using the same color in black, grey or white circles, respectively. *p* values were calculated by Wilcoxon signed-rank test to test against the hypothesis that equal proportions of samples show an increase or decrease of SCNN1B transcript species upon treatment with siRNA. Only unrelated values—i.e. ΔΔCt derived from unshared ΔCt values within biological replicates—were included into the rank test. Significance in this test indicates a systematic effect of siRNA on SCNN1B transcription.

available kits were used according to the manufacturer's instructions. RNA was isolated using the RNA-easy-mini-kit (Qiagen), transcribed into cDNA with the High-Capacity cDNA Reverse Transcription kit with RNase inhibitor (Applied Biosystems) and used as a template for real-time PCR with PowerUp SYBR Green Master Mix (Applied Biosystems) to target wild-type and read-through alternative SCNN1B transcripts with the StepOne-Plus real-time PCR system (ThermoFisherScientific). The housekeeping gene aldolase was amplified from 5 ng cDNA with 400 nM forward and reverse primers. SCNN1B transcripts were amplified from 30 ng of cDNA with 400 nM (read-through alternative transcript) and 700 nM (wild-type transcript) forward and reverse primers, respectively. Amplification was carried out using annealing at 60 °C. Threshold cycle (Ct) values were retrieved using StepOne-Software (Thermo FisherScientific).

Genetic markers. Except for 7 previously typed markers[11], genetic markers were developed de novo for this project. Genotyping was performed by the SNPstream assay (technology by Beckman Coulter, used at Cologne Center of Genomics, Cologne, Germany), by microsatellite genotyping using direct blotting electrophoresis[17] or by PCR–RFLP (see SupplTab 5a).

Evaluation of genetic data in the association study on European cystic fibrosis twins and siblings. The work presented here derived data from an association study on European CF twins and siblings[17]. The study was approved by the ethics committee of Hannover Medical School and written informed consent was obtained from all participants or their parental guardians. All methods were performed in accordance with relevant guidelines and regulations. The clinical characteristics of the patients have been described in detail elsewhere[12,15,17]. Briefly, the 12% most informative pairs from the entire sample of 318 CF twin and sibling pairs for whom pulmonary function data and weight and height were available in 1996 were selected by a ranking algorithm[12]. To study genetic modifiers, we aimed to reduce the effect of the disease-causing CFTR gene on the disease phenotype, thus deciding to study only one CFTR mutation genotype. F508del-CFTR, present on 70% of CF chromosomes from white populations of Central and West-European countries, is the only CFTR mutation for which such an approach is feasible. Moreover, patient subsamples were examined to assess the manifestation
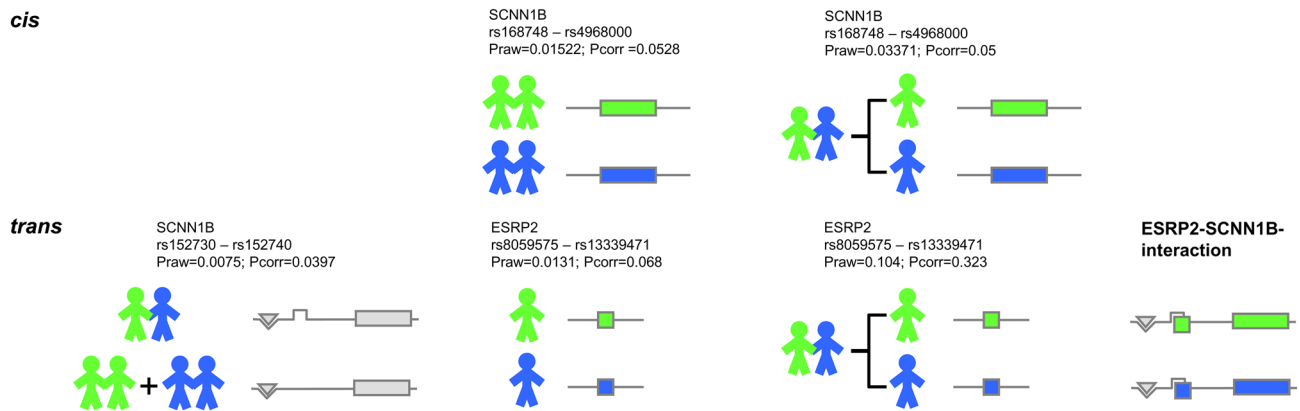
Figure 8. Identifying *cis* and *trans* regulatory elements in SCNN1B. Concordant mildly (two siblings who are shown in green), concordant severely (two siblings who are shown in blue) and discordant siblings (pairs composed of one green and one blue colored sib) have been selected from a total of 318 sibling pairs[12] and analyzed in a series of case-reference association studies. *Cis*: The genomic segment rs168748–rs4968000 determines SCNN1B functionality directly: the observed interpair and matching intrapair association with CF disease severity can be explained by the SCNN1B variants *in cis* (data not shown). *Trans*: To explain the association with the discordant phenotype, variants at SCNN1B genomic segment rs152730–rs152740 require an interacting partner encoded *in trans* to SCNN1B. The genetic variation at ESRP2 contributes to the manifestation of SCNN1B functionality. In this model, binding sites for DNA interacting partners are shown as a triangular groove (for interaction partners shaped as a triangle) and as a square groove (for interaction partners shaped as a square). In the context of this work, ESRP2 is represented by a square and fits into its binding side at the SCNN1B locus. As a consequence, the phenotype of a sib within a discordant pair is shaped by the interaction of ESRP2 and SCNN1B: the allele rs152731-T, associated with intrapair discordance (Tables 1,2), is more vulnerable to ESRP2 (Fig. 6) and its genetic variations (Fig. 5). Functionally divergent ESRP2 alleles, visualized as green and blue squares, thus determine the phenotype of carriers with ESRP2-receptive SCNN1B alleles that are shown with a square groove. These ESRP2-receptive SCNN1B alleles include rs152731-T and are overrepresented among discordant, but rare among concordant pairs. The five displayed association signals correspond to: SCNN1B, rs168748–rs4968000, Praw = 0.01522; Pcorr = 0.0528: interpair comparison of allele distribution between 11 concordant mildly and 10 concordant severely affected sibling pairs (data not shown). SCNN1B, rs168748–rs4968000, Praw = 0.03371; Pcorr = 0.05: intrapair comparison of allele distribution between mildly affected siblings and severely affected siblings from discordant pairs (data not shown). SCNN1B, rs152730–rs152740, Praw = 0.0075; Pcorr = 0.0397: interpair comparison of allele distribution between 14 discordant sibling pairs and 21 concordant sibling pairs (see Fig. 1B). ESRP2, rs8059575–rs13339471, Praw = 0.0131; Pcorr = 0.068: comparison of allele distribution of unrelated patients stratified for the response of the nasal epithelium upon superfusion with amiloride assessed by nasal potential difference measurement (see Fig. 5). ESRP2, rs8059575–rs13339471, Praw = 0.104; Pcorr = 0.323: intrapair comparison of allele distribution between mildly affected siblings and severely affected siblings from discordant pairs.

of the basic defect of impaired ion conductance in the respiratory tissue, as determined in vivo by nasal potential difference measurement[15], and in intestinal tissue, as determined ex vivo by intestinal current measurement[15]. Genetic information obtained from the case and reference populations with contrasting phenotypes was compared using the software package FAMHAP[55], which allows family-based analysis[56,57], accepts data evaluation in association studies on unrelated individuals as well as on affected sibling pairs[55] and is adapted to handle intrapair comparison of genotype data in sibling pairs[55]. Correction for multiple testing at loci typed with more than one marker was performed by haplotype permutation[56]. For this purpose, the entire data set of cases and references was used to estimate haplotype frequencies[55]. To ensure a consistent assignment of rare haplotypes in small subsamples, the genotype data of 101 families with a total of 171 patients from the European CF twin and sibling study were used as a training set in all comparisons. Haplotype, or, in cases of noninformative phase or haplotype uncertainty, weighted haplotype explanation lists were assigned to each individual whereby the haplotype frequencies of the entire data set were taken into account to compute conditional likelihood weights[55]. Permutation was performed by randomly assigning the affection status to the individuals in each replication[55]. For the comparison of case sibling pairs to reference sibling pairs, the affection status was permuted or not with an equal chance for both siblings simultaneously[55–57]. For all comparisons described herein, the phenotypes and sample sizes of the case and reference populations are detailed within the legends, in Figs. 1, 5 and 8 as well as in Table 1.

Statistical analyses. The algorithms of Sham and Curtis[58] were used to compare the observed occupancy of SCNN1B haplotypes associated with concordance vs discordance with unique transcription factors to the expectancy value derived from binding partners distributed equally between both haplotypes.

The EMSA band intensity between rs152731-C and rs152731-T probes was compared using a Mann–Whitney-U-Test in technically (electrophoresis) and biologically (cell culture and preparation of nuclear extract) independent experiments.

Changes in the expression levels of SCNN1B transcripts were judged from threshold cycle Ct values obtained by qPCR using the $\Delta\Delta$Ct method comparing siRNA or treated cells to untreated controls. To test against the hypothesis that no change in the SCNN1B transcript was observed, the Wilcoxon signed rank test was used to assess whether or not equal proportions of samples showed an increase or decrease in SCNN1B transcript species upon treatment with siRNA. For this analysis, technically (qPCR) and biologically (cell culture and experimental intervention) independent Ct values were used. For statistical analysis, $\Delta\Delta$Ct values derived from independent siRNA-treated or control samples were used.

Data availability

Supplemental Information is provided with this manuscript. Primary data will be shared with interested parties upon reasonable request.

References

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-1006 (2014).
2. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102-110 (2015).
3. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
4. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).
5. Ratjen, F. *et al.* Cystic fibrosis. *Nat. Rev. Dis. Primer* **1**, 15010 (2015).
6. Strausbaugh, S. D. & Davis, P. B. Cystic fibrosis: a review of epidemiology and pathobiology. *Clin. Chest Med.* **28**, 279–288 (2007).
7. Elborn, J. S. Cystic fibrosis. *Lancet Lond. Engl.* **388**, 2519–2531 (2016).
8. Berdiev, B. K., Qadri, Y. J. & Benos, D. J. Assessment of the CFTR and ENaC association. *Mol. Biosyst.* **5**, 123–127 (2009).
9. Enuka, Y., Hanukoglu, I., Edelheit, O., Vaknine, H. & Hanukoglu, A. Epithelial sodium channels (ENaC) are uniformly distributed on motile cilia in the oviduct and the respiratory airways. *Histochem. Cell Biol.* **137**, 339–353 (2012).
10. Tarran, R., Trout, L., Donaldson, S. H. & Boucher, R. C. Soluble mediators, not cilia, determine airway surface liquid volume in normal and cystic fibrosis superficial airway epithelia. *J. Gen. Physiol.* **127**, 591–604 (2006).
11. Stanke, F. *et al.* The TNFalpha receptor TNFRSF1A and genes encoding the amiloride-sensitive sodium channel ENaC as modulators in cystic fibrosis. *Hum. Genet.* **119**, 331–343 (2006).
12. Mekus, F. *et al.* Categories of deltaF508 homozygous cystic fibrosis twin and sibling pairs with distinct phenotypic characteristics. *Twin Res. Off. J. Int. Soc. Twin Stud.* **3**, 277–293 (2000).
13. Stanke, F. *et al.* Hierarchical fine mapping of the cystic fibrosis modifier locus on 19q13 identifies an association with two elements near the genes CEACAM3 and CEACAM6. *Hum. Genet.* **127**, 383–394 (2010).
14. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
15. Bronsveld, I. *et al.* Chloride conductance and genetic background modulate the cystic fibrosis phenotype of Delta F508 homozygous twins and siblings. *J. Clin. Invest.* **108**, 1705–1715 (2001).
16. Schreck, R., Zorbas, H., Winnacker, E. L. & Baeuerle, P. A. The NF-kappa B transcription factor induces DNA bending which is modulated by its. *Nucleic Acids Res.* **18**, 6497–6502 (1990).
17. Stanke, F. *et al.* Genes that determine immunology and inflammation modify the basic defect of impaired ion conductance in cystic fibrosis epithelia. *J. Med. Genet.* **48**, 24–31 (2011).
18. Warzecha, C. C., Shen, S., Xing, Y. & Carstens, R. P. The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. *RNA Biol.* **6**, 546–562 (2009).
19. Warzecha, C. C. *et al.* An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J.* **29**, 3286–3300 (2010).
20. Haq, I. J., Gray, M. A., Garnett, J. P., Ward, C. & Brodlie, M. Airway surface liquid homeostasis in cystic fibrosis: pathophysiology and therapeutic targets. *Thorax* **71**, 284–287 (2016).
21. Stoltz, D. A., Meyerholz, D. K. & Welsh, M. J. Origins of cystic fibrosis lung disease. *N. Engl. J. Med.* **372**, 351–362 (2015).
22. Risch, N. & Zhang, H. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589 (1995).
23. Eaves, L. & Meyer, J. Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behav. Genet.* **24**, 443–455 (1994).
24. Risch, N. J. & Zhang, H. Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *Am. J. Hum. Genet.* **58**, 836–843 (1996).
25. Stanke, F. *et al.* An informative intragenic microsatellite marker suggests the IL-1 receptor as a genetic modifier in cystic fibrosis. *Eur. Respir. J.* **50**, 1700426 (2017).
26. Mekus, F., Laabs, U., Veeze, H. & Tummler, B. Genes in the vicinity of CFTR modulate the cystic fibrosis phenotype in highly concordant or discordant F508del homozygous sib pairs. *Hum. Genet.* **112**, 1–11 (2003).
27. Rogus, J. J. *et al.* High-density single nucleotide polymorphism genome-wide linkage scan for susceptibility genes for diabetic nephropathy in type 1 diabetes: discordant sibpair approach. *Diabetes* **57**, 2519–2526 (2008).
28. Santangelo, S. L. *et al.* A discordant sib-pair linkage analysis of age-related macular degeneration. *Ophthalmic Genet.* **26**, 61–67 (2005).
29. Jiang, D., Jia, Y. & Jarrett, H. W. Transcription factor proteomics: identification by a novel gel mobility shift-three-dimensional electrophoresis method coupled with southwestern blot and high-performance liquid chromatography–electrospray-mass spectrometry analysis. *J. Chromatogr. A* **1218**, 7003–7015 (2011).
30. Stead, J. A., Keen, J. N. & McDowall, K. J. The identification of nucleic acid-interacting proteins using a simple proteomics-based approach that directly incorporates the electrophoretic mobility shift assay. *Mol. Cell. Proteomics MCP* **5**, 1697–1702 (2006).
31. Woo, A. J., Dods, J. S., Susanto, E., Ulgiati, D. & Abraham, L. J. A proteomics approach for the identification of DNA binding activities observed in the electrophoretic mobility shift assay. *Mol. Cell. Proteomics MCP* **1**, 472–478 (2002).
32. Dhar, K., Rakesh, K., Pankajakshan, D. & Agrawal, D. K. SOCS3 promotor hypermethylation and STAT3-NF-kappaB interaction downregulate SOCS3 expression in human coronary artery smooth muscle cells. *Am. J. Physiol. Heart Circ. Physiol.* **304**, H776-785 (2013).

33. Shen, C. H. & Stavnezer, J. Interaction of stat6 and NF-kappaB: direct association and synergistic activation of interleukin-4-induced transcription. *Mol. Cell. Biol.* **18**, 3395–3404 (1998).
34. Yoshida, Y. *et al.* Interleukin 1 activates STAT3/nuclear factor-kappaB cross-talk via a unique. *J. Biol. Chem.* **279**, 1768–1776 (2004).
35. Yu, Z. & Kone, B. C. The STAT3 DNA-binding domain mediates interaction with NF-kappaB p65 and inducible nitric oxide synthase transrepression in mesangial cells. *J. Am. Soc. Nephrol. JASN* **15**, 585–591 (2004).
36. Connolly, K. M., Wojciak, J. M. & Clubb, R. T. Site-specific DNA binding using a variation of the double stranded RNA binding motif. *Nat. Struct. Biol.* **5**, 546–550 (1998).
37. Hudson, W. H. & Ortlund, E. A. The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.* **15**, 749–760 (2014).
38. Gaillard, C., Cabannes, E. & Strauss, F. Identity of the RNA-binding protein K of hnRNP particles with protein H16, a sequence-specific single strand DNA-binding protein. *Nucleic Acids Res.* **22**, 4183–4186 (1994).
39. Dempsey, L. A., Hanakahi, L. A. & Maizels, N. A specific isoform of hnRNP D interacts with DNA in the LR1 heterodimer: canonical RNA binding motifs in a sequence-specific duplex DNA binding protein. *J. Biol. Chem.* **273**, 29224–29229 (1998).
40. Chennathukuzhi, V. M., Kurihara, Y., Bray, J. D., Yang, J. & Hecht, N. B. Altering the GTP binding site of the DNA/RNA-binding protein, Translin/TB-RBP, decreases RNA binding and may create a dominant negative phenotype. *Nucleic Acids Res.* **29**, 4433–4440 (2001).
41. Zhao, H., Yang, Y. & Zhou, Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res.* **39**, 3017–3025 (2011).
42. Shkreta, L. & Chabot, B. The RNA splicing response to DNA damage. *Biomolecules* **5**, 2935–2977 (2015).
43. Norman, M., Rivers, C., Lee, Y.-B., Idris, J. & Uney, J. The increasing diversity of functions attributed to the SAFB family of. *Biochem. J.* **473**, 4271–4288 (2016).
44. Ghosh, P. & Sowdhamini, R. Genome-wide survey of putative RNA-binding proteins encoded in the human proteome. *Mol. Biosyst.* **12**, 532–540 (2016).
45. Mikula, M., Bomsztyk, K., Goryca, K., Chojnowski, K. & Ostrowski, J. Heterogeneous nuclear ribonucleoprotein (HnRNP) K genome-wide binding survey reveals its role in regulating 3′-end RNA processing and transcription termination at the early growth response 1 (EGR1) gene through XRN2 exonuclease. *J. Biol. Chem.* **288**, 24788–24798 (2013).
46. Busch, A. & Hertel, K. J. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip. Rev. RNA* **3**, 1–12 (2012).
47. Shipman, K. L., Robinson, P. J., King, B. R., Smith, R. & Nicholson, R. C. Identification of a family of DNA-binding proteins with homology to RNA splicing factors. *Biochem. Cell Biol. Biochim. Biol. Cell.* **84**, 9–19 (2006).
48. Edelheit, O. *et al.* Truncated beta epithelial sodium channel (ENaC) subunits responsible for multi-system pseudohypoaldosteronism support partial activity of ENaC. *J. Steroid Biochem. Mol. Biol.* **119**, 84–88 (2010).
49. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1193–1198 (2012).
50. Beekman, M. *et al.* Genome-wide linkage analysis for human longevity: genetics of healthy aging study. *Aging Cell* **12**, 184–193 (2013).
51. Hemani, G. *et al.* Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *Am. J. Hum. Genet.* **93**, 865–875 (2013).
52. McQueen, M. B. *et al.* The National Longitudinal Study of Adolescent to Adult Health (Add Health) sibling pairs genome-wide data. *Behav. Genet.* **45**, 12–23 (2015).
53. Pollock, R. M. DNA–protein interactions. In *Current Protocols in Molecular Biology* (ed. Ausubel, F. M.) 12.0.1-12.11 (Wiley, New Work, 1997).
54. Century, T. J., Fenichel, I. R. & Horowitz, S. B. The concentrations of water, sodium and potassium in the nucleus and cytoplasm of amphibian oocytes. *J. Cell Sci.* **7**, 5–13 (1970).
55. Herold, C. & Becker, T. Genetic association analysis with FAMHAP: a major program update. *Bioinforma. Oxf. Engl.* **25**, 134–136 (2009).
56. Becker, T. & Knapp, M. A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am. J. Hum. Genet.* **75**, 561–570 (2004).
57. Knapp, M. & Becker, T. Family-based association analysis with tightly linked markers. *Hum. Hered.* **56**, 2–9 (2003).
58. Sham, P. C. & Curtis, D. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann. Hum. Genet.* **59**, 97–105 (1995).
59. Voilley, N. *et al.* Cloning, chromosomal localization, and physical linkage of the beta and gamma subunits (SCNN1B and SCNN1G) of the human epithelial amiloride-sensitive sodium channel. *Genomics* **28**, 560–565 (1995).
60. Saxena, A. *et al.* Gene structure of the human amiloride-sensitive epithelial sodium channel beta subunit. *Biochem. Biophys. Res. Commun.* **252**, 208–213 (1998).
61. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168 (2015).

Acknowledgements

Author contributions

B.T. and F.S. conceived and designed the project. Funding was acquired by F.S., T.B and B.T. A.P., S.T., S.H., M.I., J.A. and N.D. performed the experiments. A.P, J.A., M.R.T, S.J. and F.S. analyzed the primary data. T.B. developed and implemented the program to perform the interpair- and intrapair comparisons of genetic data of affected sib pairs. Formal analysis of genetic data was done by F.S. and T.B. T.B., A.P, S.T., S.H., M.I., J.A., M.T., S.J., B.T. and F.S. drafted the manuscript and/or revised the draft critically for content. F.S. visualized the data and wrote, edited and revised the final manuscript.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79804-y.

Correspondence and requests for materials should be addressed to F.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.