Supplementary Methods and Results

Distinct known and novel SAS have more annotated isoforms

We analyzed 5,169 known SAS genes (including loci with more than 2 gene partners), 7,823 novel SAS genes, and 7,929 genes without any evidence for antisense transcription, and found that known and novel SAS genes had more isoforms than non SAS genes (average of 2.3 and 2.3, versus 1.8, respectively; Welch T-test P = 3.2 x 10^{-84}; Supplementary Fig. S1; Welch T-test P values in Supplementary Table 2). However, we also found that on average, known and novel SAS genes were significantly longer (83.7 kb and 70.6 kb, versus 22.9 kb), had longer mRNAs (2.3 kb and 2.6 kb, versus 1.8 kb), and had more introns (9.4 and 9.9 versus 6.6; Supplementary Fig. S1; P values in Supplementary Table 2). Thus, the multiple alternative isoforms found in known and novel SAS genes could simply reflect the increased chance of observing alternative transcription in longer genes.

To address this possibility, we binned non-SAS genes by size quantiles: 0%-10% (genes of length 100 bp – 1.87 kb), 10 %- 20% (1.87 kb – 3.49 kb), 20%-30% (3.49 kb – 5.57 kb), 30% - 40% (5.57 kb – 8.13 kb), 40%-50% (8.13 kb – 11.44 kb), 50%-60% (11.44 kb – 15.95 kb), 60%-70% (15.95 kb – 22.23 kb), 70%-80% (22.23 kb – 32.74 kb), 80%-90% (32.74 kb – 54.71 kb), 90%-100% (54.71 kb – 780.14 kb). Next, all known SAS genes with sizes corresponding to each bin were indentified from the total pool of known SAS genes. To test if these known SAS genes had the same gene size distribution as the non-SAS genes, we compared the size distributions of the known SAS and non-SAS genes in each bin (differences assessed using a T-test). The six bins with genes of comparable lengths (T-test p-value > 0.05; bins spanning 20%-70%, and bin 80%-90%) were further considered. In the first 3 of these bins (i.e. the smallest genes) there was no significant difference between the number of gene isoforms in known SAS and non-SAS genes. In contrast, for the largest three bins (starting at the 50th percentile), there was a significant enrichment of isoforms in the known SAS versus non-SAS genes. Thus, for genes of length > 11.4kb, there was a positive association between antisense transcription and alternative splicing (Supplementary Table 5).

Alternative splicing patterns in individual genes

The splicing of individual exons in sense genes was either positively or negatively correlated to antisense expression (**Supplementary Table 1, Supplementary Figs. 1** and **3**). At the majority of known (68.6% of 191) and novel (62.5% of 168) SAS loci, individual genes contained subsets of exons with splicing indexes that were positively, negatively, and un-correlated to antisense transcription. Un-correlated exons are likely constitutively spliced and therefore present in the majority of alternatively spliced transcripts. In contrast, antisense-correlated exons represent changes in the level of expressed mRNA isoforms containing positively-correlated exons and excluding negatively-correlated exons. On average, antisense-correlated splicing events occurred at 32.3% of known SAS gene exons and 23.1% of novel SAS gene exons, indicating that alternative isoforms differed from each other by multiple exons (**Supplementary Fig. 2**).

Functional annotation of genes with known, novel, or no antisense transcripts

We used the David Bioinformatics Resources[(Huang et al. 2009), (Dennis et al. 2003)] to functionally annotate the 4,792 known SAS, 7,648 novel SAS and 7,137 non SAS genes whose IDs could be converted to DAVID IDs, specifically focusing on Gene Ontology[(Michael et al. 2000)] terms (GO), and Uniprot Keywords (http://www.uniprot.org/manual/keywords).

The novel SAS gene category was the most enriched in functional categories, including 58 GO categories and 45 Keywords **(Supplementary Table 3b)**. Notably, 31 of 39 GO terms in the Biological Process category pertained to regulation (i.e. “Regulation of Apoptosis”, “Negative Regulation of Transcription, “Regulation of Signal Transduction”, etc). Highly enriched UniProt Keywords included “Phosphoprotein”, “Alternative Splicing”, “Kinase”, “Apoptosis”, and “Proto-oncogene”. Overall, these GO terms and Keywords are consistent with the strong enrichment of Cancer Gene Census (CGC) genes observed in the novel SAS class relative to the set of all protein coding genes (215 of 389 CGC genes, Chi-square test, $P = 4.2\times10^{-9}$). CGC data was obtained from the Wellcome

Trust Sanger Institute Cancer Genome Project web site,
http://www.sanger.ac.uk/genetics/CGP.

The known SAS gene category was enriched in 14 GO terms and 8 Keywords **(Supplementary Table 3a)**, notably also including the Keywords “Phosphoprotein” and “Alternative Splicing”. Thus, one potential defining characteristic of both known and novel SAS genes is significant processing at both the transcriptional as well as the post-translational stages.

In contrast, genes without antisense transcription **(Supplementary Table 3c)** were enriched in only 5 GO terms (including “Defence Response to Bacterium”, and “Chemotaxis”), and 22 Keywords, including “Secreted”, “Antimicrobial”, “Antibiotic”, “Immune Response” and “Inflammation”.

Structural consequences of antisense-correlated splicing

Alternative splicing is a mechanism by which cells can increase the number of proteins encoded by a set number of genes. To determine the extent to which antisense-correlated splicing events may alter protein domains in known SAS genes, we investigated splicing events that affected Superfamily**(Gough et al. 2001)** domains. Of the 259 known SAS gene pairs expressed in the CEU-YRI samples, 87 genes had 123 expressed exons that encoded 131 protein domains (69 unique domains, of which a subset was observed multiple times). Over a third (33) of these genes had exons encoding 35 protein domains spliced in an antisense-correlated manner. The remaining genes (54) had 60 protein domains that were not affected by antisense-correlated splicing. Thus, antisense-correlated splicing events have putative consequences on protein function.

Population differences in splicing events

To determine if we could detect population-specific splicing events, we re-analyzed the 87 CEU and 89 YRI datasets separately, and found a total of 155 known and 149 novel

SAS genes with antisense-correlated splicing events in total (**Supplementary Fig. 4**). Approximately half of these known (47.1%) and novel (55.7%) SAS genes had antisense-correlated splicing events unique to either the CEU or the YRI populations, with the remainder being observed in both groups. Overall, a remarkable increase in the number of novel SAS genes undergoing antisense-correlated splicing (35.6% versus 24.1%), was observed to occur solely in the YRI population, indicating that a considerable proportion of antisense-correlated splicing events are indeed population specific.

A similar pattern emerged when analyzing exons rather than genes, with a much larger proportion of exons having antisense-correlated splicing patterns in the YRI individuals **(Supplementary Fig. 4)**. Thus, the YRI population was enriched relative to the CEU population in the number of novel SAS loci and of exons with antisense-correlated splicing.

There were three potential explanations for this observation. The first was that the array probesets were designed using EST and cDNA libraries that were biased toward a high representation of European sequences; this would introduce bias when analyzing non-European genomes. Consequently, we would expect both known and novel SAS genes to be influenced by such a bias, since there are thousands of genes in each category (and hundreds of expressed genes in the LCL samples). This was not the case, as we did not observe any differences in the known SAS gene category, but only in the novel SAS gene category. This indicates that array design does not explain the population differences observed.

The second possibility was that a larger number of novel versus known SAS genes were expressed in the YRI population, facilitating detection of additional antisense-correlated splicing events. These genes would be detectable as differentially expressed between the CEU and YRI samples. Thus, we ascertained the overlap between differentially expressed genes and novel SAS genes with antisense-correlated splicing events unique to the YRI individuals. Although we did detect differentially expressed genes between the YRI and CEU populations (data not shown), they were nearly absent from the set of genes with antisense-correlated splicing events unique to the YRI individuals. Thus, differential expression does not explain the population differences observed.

Third, it was possibile that individuals of Yoruban descent have a greater variability in the splicing of novel SAS genes (but not known SAS genes); i.e. novel SAS genes undergo more extreme exon exclusion and inclusion events in YRI compared to CEU individuals. If that was the case, then these exons would be more likely to be detected as expressed above threshold (i.e. expression above background in > 20% of samples), and consequently, to be detected as antisense-correlated splicing events in the YRI individuals. We tested this possibility by comparing the standard deviation (SD) of splice index values for probesets in novel SAS genes in the CEU and YRI samples. Considering 3,735 expressed probesets in the 189 novel SAS genes, there was a small but significant increase in probeset SI values (i.e. inclusion and exclusion) in the YRI versus the CEU individuals (mean SD values: 0.0283 in CEU, 0.0299 in YRI, Kolmogorov-Smirnov test p-value = 8.3 x 10^{-8}). In contrast, for known SAS genes, there was no significant difference in probest SI value variability between YRI and CEU individuals (n = 2,771 probesets in 244 known SAS genes; mean SD values: 0.0286 in CEU, 0.0295 in YRI, Kolmogorov-Smirnov p-value > 0.01). Thus, compared to known SAS genes, the novel SAS gene probesets had a higher level of splicing variance (i.e. SI values) in the YRI population, indicating that the alternative splicing of these genes is distinct (i.e. they had more frequent or more extreme inclusion in YRI vs CEU samples). Given this observation, it is not surprising that the subset of antisense-correlated splicing events is also higher in this population.

Intron retention

Due to the presence of intronic probesets in the dataset, we were able to track intron retention events, and specifically those that were correlated with the expression of an antisense gene. For known SAS genes, there were 5,327 intronic probesets in 228 genes, while for novel SAS genes there were 10,063 intronic probesets in 637 genes. Intronic probesets could indicate not only intron retention events, but also cryptic exons that have eluded annotation using current methods, or embedded genes. A total of 116 intronic probesets passing expression thresholds were found in 9 known SAS genes, of which 29 were significantly correlated with the expression of 8 known SAS genes. In Novel SAS genes, 208 intronic probesets passed expression thresholds and mapped to in 14 genes. Of these, 33 were significantly correlated with the expression of 12 antisense constructs.

Regions of SAS overlap are enriched in exons

The frequency of exons per kilobase (exons/kb) was calculated for all known SAS gene pairs. For each gene pair, the number of exons/kb in the overlapping region (including exons from both strands) was compared to the number of exons/kb in non-overlapping regions of both genes. A total of 1,694 known SAS gene pairs (96.0% of 1,765 pairs) had a significantly increased exon frequency in the overlapping regions (Chi-square test, Bonferroni corrected $P < 0.05$). For this analysis, overlapping alternative exons (ie. sharing the same genomic location, but differing in 5´ or 3´ ends) were only counted once. Considering all SAS gene pairs, the average frequency of exons in overlapping regions was 3.1/kb, over seven times greater than that in non-overlapping regions (0.43/kb), and this difference was highly significant (Welch's t-Test, $P < 2.2\times10^{-16}$).

Nucleosome enrichment in exons

To confirm that SAS genes harbored more nucleosomes in exons than introns, we re-analyzed the publicly available activated T-cell ChIP-seq and microarray data (Schones et al. 2008). We used the microarray data to identify the subset of known SAS genes expressed in activated T-cells, and the ChIP-seq data to determine if relative to introns, the exons of expressed SAS genes were indeed enriched in nucleosomes.

Activated T-cell ChIP-seq and microarray expression data were downloaded from GEO (Barrett et al. 2009) (GSE10437). ChIP-seq data were processed as previously described (Schwartz et al. 2009). T-cell microarray data[(Schones et al. 2008)] data were processed using the Affymetrix Expression Console Software (http://www.affymetrix.com/), and MAS5[(Hubbell et al. 2002)] p-values were used to identify 8,627 expressed genes in activated T-cells. Of these, 189 belonged to the set of 2,995 known SAS genes, and a smaller subset of 122 genes had both antisense-correlated splicing events and corresponding nucleosome occupancy ChIP-seq data. Mean nucleosome occupancy was calculated separately for intronic and exonic regions in SAS gene partner regions. Areas of exon:intron sequence overlap were considered exonic sequence. Overall, there was an average 1.2-fold enrichment of nucleosome peaks in exonic rather than intronic regions of individual genes (Student's t-Test P = 5.7×10^{-9}; **Supplementary Fig. 5**).

SAS overlaps are enriched in PolII occupancy and alternative exons

A total of 557 known SAS genes were expressed in GM12878 and harbored at least one significant PolII peak. For 8 of these genes, the length of the SAS sequence overlap spanned the whole gene, and these were excluded from further analysis. Of the remaining 549 genes, we analyzed 248 genes with at least one non promoter-associated PolII peak, indicating the presence of the elongating form of PolII in the gene body.

We examined the likelihood that enrichment of PolII peaks in areas of SAS overlap was due to transcriptional termination, which causes a decrease in PolII speed**(Nag et al. 2006)**. Of the 248 known SAS genes with PolII peaks, we found 31 genes with PolII peaks in the terminal 100bp of the gene. For 9 of these genes, the area of SAS overlap did not span the 3´ end, leaving 22 (8.9%) genes with terminal PolII peaks. Of these, 10 genes formed 5 pairs that overlapped over the last 100bp of their length. The remaining 12 genes had 3´-terminal PolII peaks spanning the SAS overlap, but the partner gene did not. In 8 of the cases, this was because the partner gene was not in the starting list of 248 genes (possibly due to a lack of expression), and in another 4 cases the partner gene had PolII peaks in the SAS overlap but they did not overlap the 3´-most 100bp of the sequence. In general, the complete 3´ exon of these 22 genes harbored one PolII peak, showing a local PolII enrichment. It was not possible to determine whether the local PolII accumulation was, in these cases, due to transcriptional termination rather than due to effects on chromatin structure, but we note that this occurs in only a small proportion of the genes studied. Furthermore, whether PolII accumulation is caused by increased nucleosome frequency, by termination, or by transcriptional interference from the antisense strand, its presence likely results in attenuated transcriptional speed that is expected to affect the splicing of exons in the 3´-end of these genes.

Multiple species analysis

The correspondence between known and novel antisense transcription was calculated as described in the Methods section of the main text. Supplementary Table 4 enumerates the p-values for all comparisons. Note that although there were over 1.4 milion ESTs available for chimp, the spliced EST table downloaded from UCSC was nearly un-populated, and very likely under-represents the true number of spliced ESTs in this organism. When considering all chimp ESTs (including un-spliced ESTs), there was a significant enrichment (1.68) of antisense ESTs in the multiple versus single transcript gene categories ($p = 1.98 \times 10^{-77}$). These numbers are comparable to the results from all human ESTs (data not shown). The positive association between novel antisense transcription (measured using spliced ESTs) and genes with multiple annotated transcripts also remained significant when considering the subset of genes with highly expressed antisense ESTs (i.e. genes with antisense EST counts above the median in each species, compared to those with EST counts greater than 1). Thus, the association between antisense transcription and alternative splicing remained robust to increases in gene expression.

siRNA and H3K27me3 levels at exons with antisense-correlated splicing

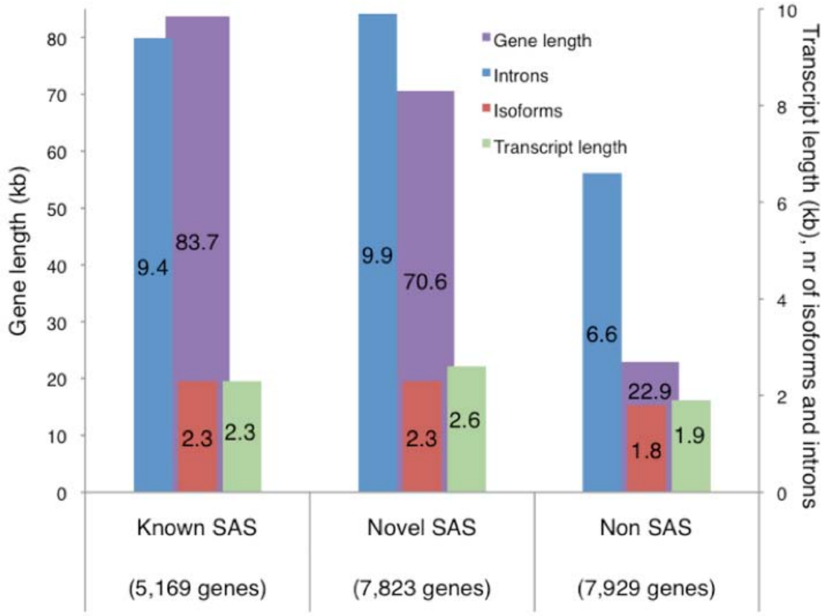
We tested the hypothesis that endogenous antisense transcription leads to siRNA formation, and consequent deposition of silencing marks near alternatively spliced exons, as recently demonstrated at the Fibronectin 1 (FN1) gene ((Allo et al. 2009)). At the FN1 locus, siRNAs derived from endogenous antisense transcripts cause increased levels of H3K27me3 and H3K9me2 histone modifications, and consequent inclusion of the alternatively spliced ED1 exon. If this mechanism was a prevalent cause for exon inclusion in the splicing events identified in this work, H3K29me3 and H3K9me2 modifications should preferentially mark exons that are included in the mRNA. To test this hypothesis, we used publicly available data for H3K27me3 levels generated for one of the 176 samples (GM12878; primary UCSC table: wgEncodeBroadChipSeqPeaksGm12878H3k27me3; H3K9 methylation data were not available for GM12878). We asked whether exons whose splicing index was negatively correlated with antisense transcription had higher H3K27me3 levels compared to exons whose splicing index was not correlated to antisense transcription. According to the hypothesis, our expectation was that siRNAs produced from SAS loci would cause H3K27me3 marks that lead to exon inclusion (i.e. a large SI value). In GM12878, we flagged those exons whose splicing was negatively correlated to the antisense gene, and which had high SI values (i.e. were included) in the GM12878 sample. The threshold for inclusion was set as the 50th percentile of the SI values of all probesets in GM12878. We then identified a second category of probesets whose SI values were not correlated to antisense transcription, and which were relatively excluded in GM12878 (ie. with SI values smaller than the 50th percentile). In total, there were 144 known and novel SAS genes with probesets in both categories. Probesets were mapped to exons, and the regions further investigated spanned 400bp of flanking sequence centered on the 5' boundaries of the shortlisted exons. To test the hypothesis, we compared H3K27me3 levels at both types of exons in all 144 genes, expecting to see higher levels of H3K27me3 in the included exons with antisense-correlated splicing, and lower levels in the excluded exons without antisense-correlated splicing. However, there was no significant difference in the H3K27me3 levels of included and excluded exons in these regions (total integral coverage: 0.83 in the included vs 0.75 in the excluded exons; T-test P = 0.18). Overall, H3K27me3 levels were quite low at the regions of interest; based on these results, we can neither conclusively rule out the hypothesis, nor provide evidence in support of it.

Since there was a lack of measurable changes in H3K27me3 levels between the two types of exons (i.e. those for which splicing is correlated to antisense transcription, and those for which it is not), we sought to directly ascertain whether siRNAs were enriched in exons with antisense-correlated splicing. We used RNA-seq data generated by the Cold Spring Harbour Laboratories (CSHL) for GM12878, which profiled RNAs between 20-200bp in length, assuming that this RNA fraction include siRNAs. These small RNAs were directionally cloned from the nuclear fraction, and sequenced from the 5' end, thus maintaining strand-specificity. We mapped the reads to the exons defined above, and asked whether included exons with antisense-correlated splicing (and with relative inclusion in GM12878) were enriched in reads relative to excluded exons that did not have antisense-correlated splicing. Once again, we saw low levels of small RNAs mapping to the region of interest, and no significant differences in abundance between the different types of exons. Specifically, only 36 genes had mapped reads, and these occurred at an average of 2.6 reads. This low coverage suggests that deeper RNA-seq libraries may be needed to address this question.

In conclusion, using publicly available data, we did not find conclusive evidence in support of the possibility that siRNA abundance (or the repressive histone mark H3K27me3) is associated with the antisense-correlated alternative splicing events identified in this study. The data available for this analysis may not be sensitive enough to detect a subtle effect, so other methods (such as those used by (Allo et al. 2009)) could be used to address this question in future studies.

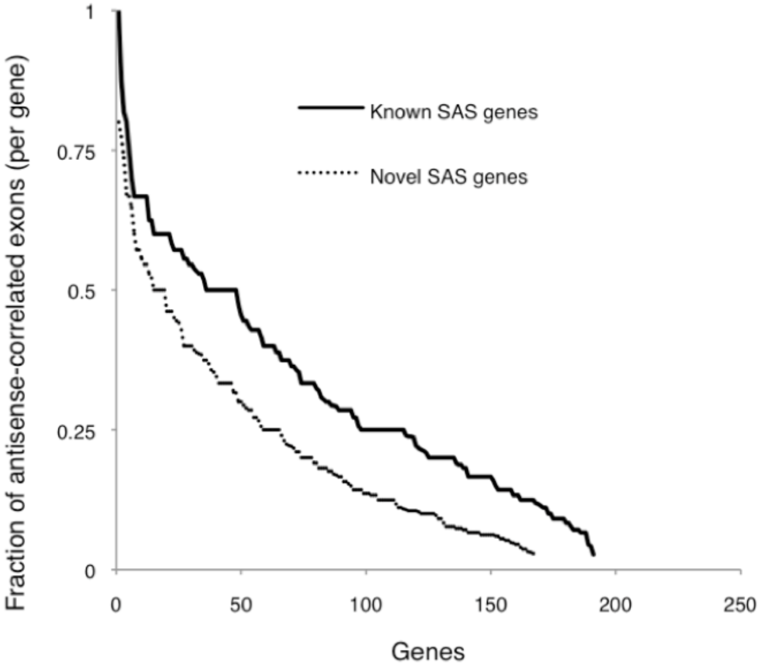
Supplementary Figures and Tables

Supplementary Figure 1. On average, known and novel SAS genes have significantly longer gene (left vertical axis) and transcript lengths than genes with no antisense transcription, as well as significantly more introns and isoforms (right vertical axis; for P values see Supplementary Table 2.). The total number of protein coding known SAS genes, novel SAS, and non SAS genes in the human genome is shown on the x-axis. Black arrows denote transcriptional direction.



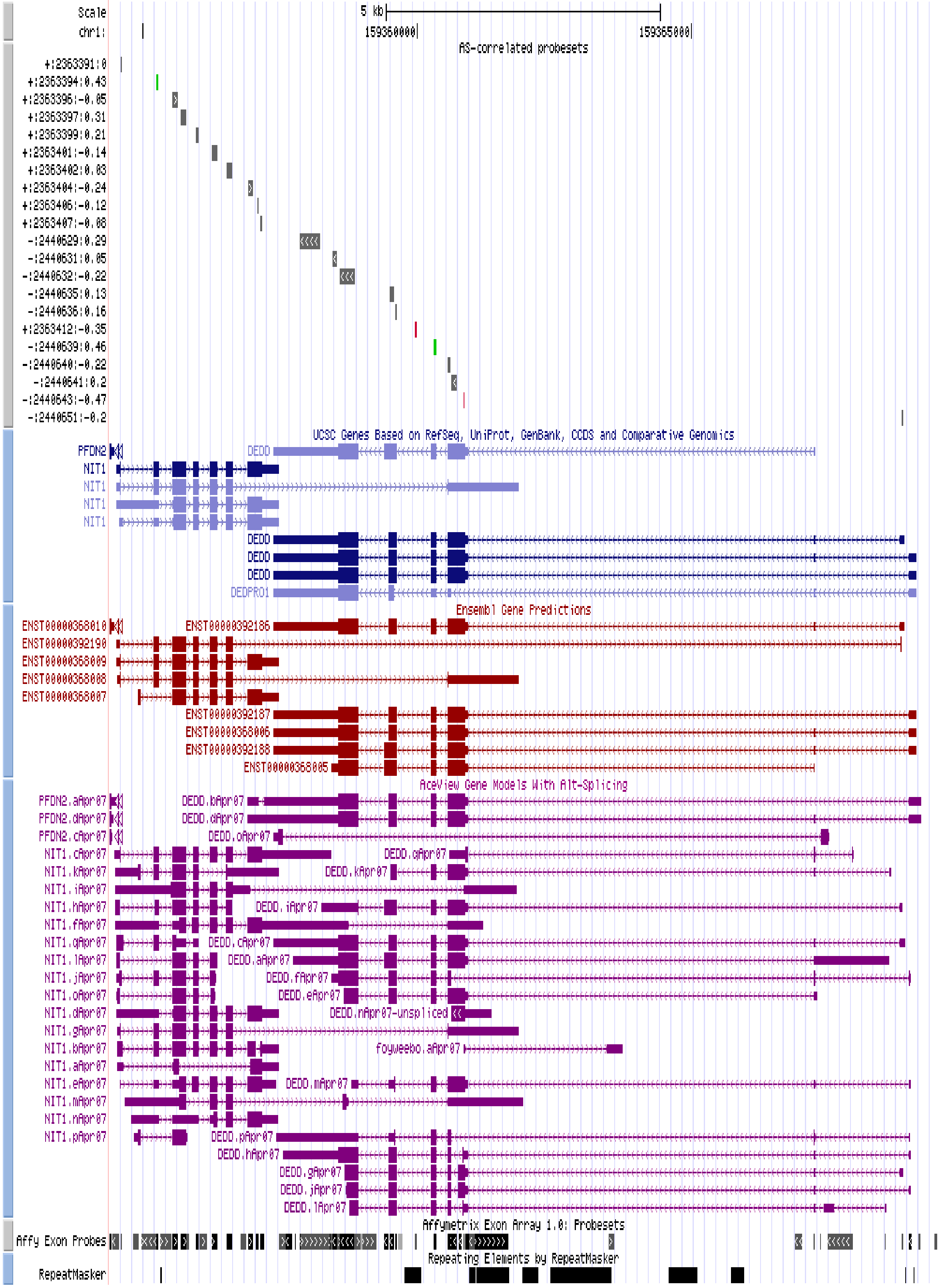
Supplementary Figure 2. Proportion of antisense-correlated exons.

Antisense-correlated exons in known (solid line) and novel (dashed line) SAS genes.



Supplementary Figure 3. **UCSC track view of a known SAS gene locus with two gene members, DEDD and NIT1, shows antisense-correlated splicing events.**

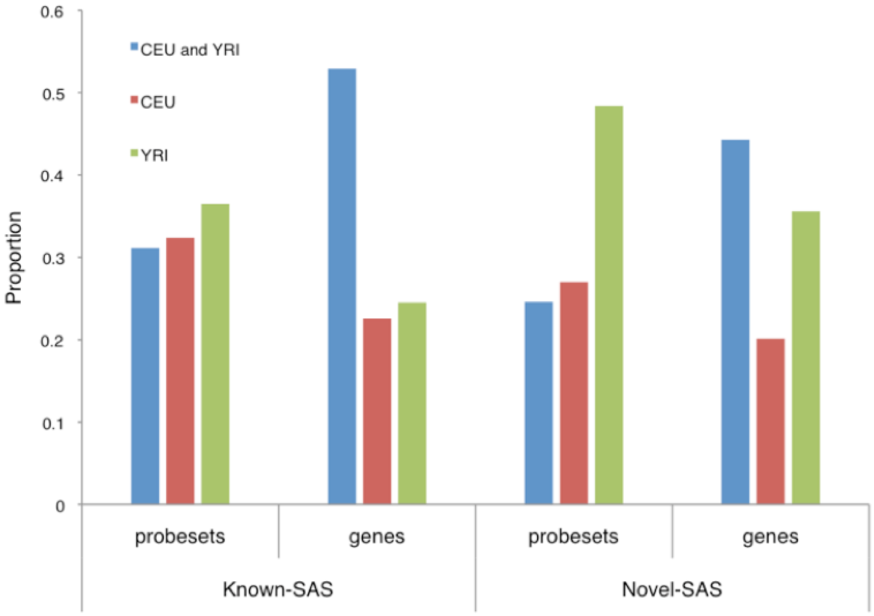
Probesets (ie. exons) passing the expression threshold and mapping to the positive and negative strands are shown in the top track, and are either positively correlated (green), negatively correlated (red), or not significantly correlated (grey) to antisense gene expression. Probeset tracks are labeled with “strand:probeset_id:correlation”.


Scale
5 kb
chr1:
159360000
159365000
AS-correlated probesets
+:2363391:0
+:2363394:0.43
+:2363396:-0.05
+:2363397:0.31
+:2363399:0.21
+:2363401:-0.14
+:2363402:0.03
+:2363404:-0.24
+:2363406:-0.12
+:2363407:-0.08
-:2440629:0.29
-:2440631:0.05
-:2440632:-0.22
-:2440635:0.13
-:2440636:0.16
+:2363412:-0.35
-:2440639:0.46
-:2440640:-0.22
-:2440641:0.2
-:2440643:-0.47
-:2440651:-0.2
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PFDN2
DEDD
NIT1
NIT1
NIT1
NIT1
DEDD
DEDD
DEDD
DEDPRO1
Ensembl Gene Predictions
ENST00000368010
ENST00000392186
ENST00000392190
ENST00000368009
ENST00000368008
ENST00000368007
ENST00000392187
ENST00000368006
ENST00000392188
ENST00000368005
AceView Gene Models With Alt-Splicing
PFDN2.aApr07
DEDD.bApr07
PFDN2.dApr07
DEDD.dApr07
PFDN2.cApr07
DEDD.oApr07
NIT1.cApr07
DEDD.gApr07
NIT1.kApr07
DEDD.kApr07
NIT1.iApr07
NIT1.hApr07
DEDD.iApr07
NIT1.fApr07
NIT1.gApr07
DEDD.cApr07
NIT1.lApr07
DEDD.aApr07
NIT1.jApr07
DEDD.fApr07
NIT1.oApr07
DEDD.eApr07
NIT1.dApr07
DEDD.nApr07-unspliced
NIT1.gApr07
NIT1.bApr07
foyweebo.aApr07
NIT1.aApr07
NIT1.eApr07
DEDD.mApr07
NIT1.mApr07
NIT1.nApr07
NIT1.pApr07
DEDD.pApr07
DEDD.hApr07
DEDD.gApr07
DEDD.jApr07
DEDD.lApr07
Affymetrix Exon Array 1.0: Probesets
Affy Exon Probes
Repeating Elements by RepeatMasker
RepeatMasker

Supplementary Figure 4. YRI individuals have a greater proportion of novel SAS genes with antisense-correlated splicing events.

The fraction **(a)** and number **(b)** of probesets (i.e. exons) with antisense-correlated splicing indexes, and the fraction of genes that these probesets map to, is shown for the individual CEU and YRI datasets and for both populations.

a

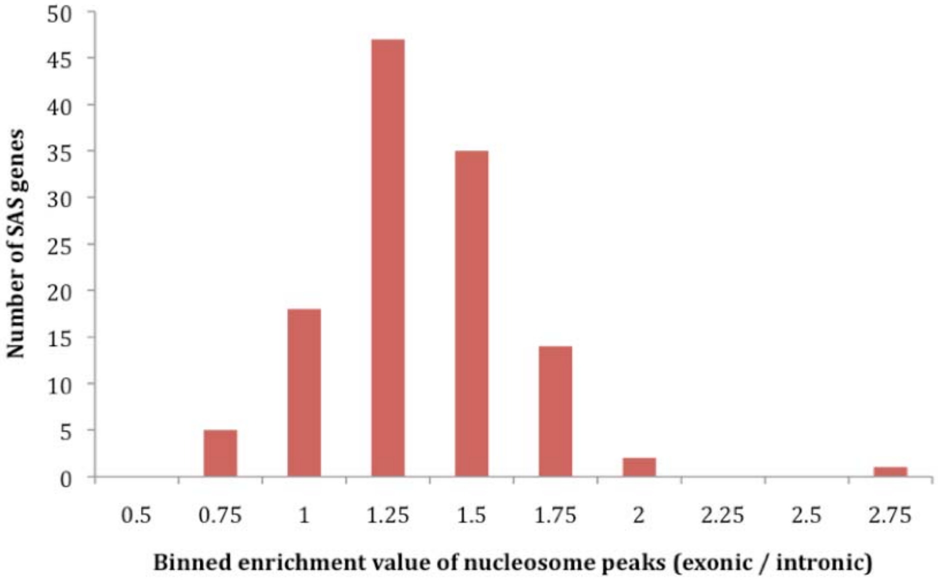


b

	Known SAS		**Novel SAS**	
	probesets	**genes**	**probesets**	**genes**
CEU and YRI	151	82	162	66
CEU	157	35	178	30
YRI	177	38	319	53
Total	485	155	659	149

Supplementary Figure 5. Nucleosomes are enriched in SAS gene exons relative to introns.

Nucleosome peak enrichment was calculated using a ratio of exonic vs intronic peaks. The majority of genes have a significant enrichment of exonic nucleosomes in introns (ratio > 1).



Supplementary Table 1. Antisense-correlation test results.

The correlation between the splicing index of each exon (ie. probeset) and the expression of the corresponding known antisense gene or novel antisense constructs was assessed at every expressed locus in LCL cells. Correlation p-values were corrected for multiple testing using the Bonferroni method. Correlations were also assessed between the gene expression levels of known SAS gene partners. Exons (or genes) whose splice index was positively or negatively correlated to antisense gene or construct expression were independently summarized.

	Knows SAS (splicing)	**Known SAS (gene expression)**	**Novel SAS (splicing)**
Total exons (probesets)	2,995	129	4,167
Total antisense genes / constructs	258	129	215
Significant exons (genes)	720 (191)	88	823 (168)
Significant exons with negative correlation (genes)	373 (165)	3	435 (136)
Significant exons with positive correlation (genes)	347 (155)	85	387 (137)

Supplementary Table 2. Structural differences between known and novel SAS genes and non SAS genes.

P-values for all pairwise comparisons between known SAS, novel SAS, and non SAS genes, as described in **Fig. 1b** (Welch two-sample t-test).

	Known SAS **vs** **Novel SAS**	**Known SAS** **vs** **Non SAS**	**Novel SAS** **vs** **Non SAS**
Gene length (kb)	2.4×10^{-7}	1.6×10^{-148}	1.0×10^{-285}
Transcript length (kb)	8.0×10^{-18}	5.1×10^{-40}	3.0×10^{-147}
Transcripts per gene	6.1×10^{-1}	3.2×10^{-84}	4.2×10^{-111}
Introns per gene	5.8×10^{-3}	5.8×10^{-78}	9.7×10^{-147}

Supplementary Table 3. Functional annotation analysis.

Functional annotations for **(a)** known SAS, **(b)** novel SAS genes, and **(c)** non SAS genes shows enriched Gene Ontology terms and UniProt Keywords.

a

Category	Term	Gene Count	Fold Enrichment	p-value
4,792 Known SAS genes				
GO Biological Process	cellular protein metabolic process	601	12.9	1.70E-04
	regulation of small GTPase mediated signal transduction	87	1.9	2.10E-03
	protein modification process	379	8.1	3.30E-03
	RNA metabolic process	254	5.4	8.60E-03
	intracellular signaling cascade	328	7	9.50E-03
GO Cellular Component	cytoplasm	1797	38.5	4.40E-12
	cytoplasmic part	1198	25.7	2.80E-05
	intracellular membrane-bounded organelle	1874	40.1	7.60E-05
	intracellular organelle	2086	44.7	9.10E-05
	Golgi apparatus	249	5.3	2.30E-04
	DNA-directed RNA polymerase II, holoenzyme	36	0.8	1.50E-03
GO Molecular Function	Ras guanyl-nucleotide exchange factor activity	38	0.8	7.20E-03
	protein kinase activity	174	3.7	7.60E-03
	adenyl ribonucleotide binding	386	8.3	6.70E-03
UniProt Keywo`rds	alternative splicing	2037	43.6	3.70E-44
	phosphoprotein	1846	39.5	6.60E-19
	cytoplasm	875	18.7	3.70E-09
	coiled coil	551	11.8	1.20E-07
	guanine-nucleotide releasing factor	51	1.1	1.70E-04
	tpr repeat	60	1.3	1.30E-03
	golgi apparatus	172	3.7	1.90E-03
	atp-binding	348	7.5	5.40E-03

b

Category	Term	Gene Count	Fold Enrichment	p-value
7,648 Novel SAS genes				
GO Biological Process	negative regulation of gene expression	271	3.5	5.70E-06
	negative regulation of cellular metabolic process	370	4.8	3.90E-06
	negative regulation of transcription	248	3.2	5.60E-06
	negative regulation of macromolecule metabolic process	373	4.9	8.90E-06
	negative regulation of biosynthetic process	299	3.9	9.20E-06
	negative regulation of macromolecule biosynthetic process	287	3.8	7.80E-06
	negative regulation of cellular biosynthetic process	293	3.8	8.30E-06
	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	269	3.5	1.50E-05
	negative regulation of nitrogen compound metabolic process	271	3.5	2.40E-05
	regulation of protein metabolic process	283	3.7	2.60E-05
	cellular protein metabolic process	1070	14	4.00E-05
	negative regulation of RNA metabolic process	196	2.6	5.40E-05
	negative regulation of transcription, DNA-dependent	193	2.5	5.50E-05
	regulation of cellular protein metabolic process	245	3.2	2.00E-04
	neurogenesis	302	3.9	2.50E-04
	regulation of signal transduction	425	5.6	2.80E-04
	positive regulation of cell differentiation	129	1.7	4.00E-04
	generation of neurons	281	3.7	5.00E-04
	regulation of cell development	117	1.5	4.80E-04
	regulation of neurogenesis	95	1.2	3.70E-03
	regulation of cellular carbohydrate metabolic process	29	0.4	6.10E-03
	negative regulation of programmed cell death	184	2.4	6.30E-03
	embryonic limb morphogenesis	55	0.7	6.50E-03
	embryonic appendage morphogenesis	55	0.7	6.50E-03
	regulation of programmed cell death	385	5	6.60E-03
	limb morphogenesis	61	0.8	6.50E-03
	negative regulation of cell death	184	2.4	6.50E-03
	regulation of apoptosis	381	5	6.80E-03
	posttranscriptional regulation of gene expression	115	1.5	6.60E-03
	positive regulation of cellular carbohydrate metabolic process	18	0.2	7.20E-03
	positive regulation of carbohydrate metabolic process	18	0.2	7.20E-03
	positive regulation of RNA metabolic process	238	3.1	7.10E-03
	positive regulation of transcription, DNA-dependent	236	3.1	7.40E-03
	regulation of carbohydrate metabolic process	29	0.4	8.70E-03
	negative regulation of apoptosis	180	2.4	8.90E-03
	regulation of neuron differentiation	77	1	8.70E-03
	organ morphogenesis	274	3.6	8.50E-03
	positive regulation of cellular metabolic process	412	5.4	9.30E-03
	protein modification process	658	8.6	9.60E-03

Category	Term	Gene Count	Fold Enrichment	p-value
7,648 Novel SAS genes				
GO Cellular Component	cytoplasm	3273	42.8	3.50E-29
	cytoplasmic part	2226	29.1	1.30E-19
	intracellular organelle	3860	50.5	3.90E-19
	intracellular membrane-bounded organelle	3465	45.3	9.90E-19
	cytosol	692	9	1.10E-18
	intracellular organelle part	1885	24.6	5.00E-11
	nuclear part	845	11	3.30E-07
	nucleus	2196	28.7	4.10E-07
	intracellular organelle lumen	811	10.6	2.20E-05
	nuclear lumen	665	8.7	1.10E-04
	organelle envelope	305	4	1.40E-04
	nuclear envelope	115	1.5	2.50E-04
	Golgi apparatus	411	5.4	5.20E-04
	nucleoplasm	415	5.4	5.40E-04
	intracellular non-membrane-bounded organelle	1131	14.8	1.70E-03
	Golgi apparatus part	152	2	2.10E-03
	nuclear membrane	46	0.6	6.40E-03
	nuclear body	91	1.2	9.90E-03
GO Molecular Function	adenyl ribonucleotide binding	707	9.2	7.50E-08

Category	Term	Gene Count	Fold Enrichment	p-value
7,648 Novel SAS genes				
UniProt Keywords	phosphoprotein	3445	45	5.40E-94
	acetylation	1339	17.5	1.80E-44
	alternative splicing	3319	43.4	1.20E-44
	cytoplasm	1512	19.8	8.60E-20
	nucleotide-binding	806	10.5	7.00E-16
	atp-binding	641	8.4	3.50E-13
	nucleus	1830	23.9	7.70E-11
	metal-binding	1296	16.9	1.10E-09
	chromosomal rearrangement	158	2.1	4.00E-08
	ubl conjugation	297	3.9	4.40E-08
	activator	266	3.5	7.20E-08
	kinase	336	4.4	3.40E-07
	Transcription	901	11.8	3.50E-06
	transcription regulation	880	11.5	6.60E-06
	zinc	944	12.3	9.50E-06
	zinc-finger	752	9.8	1.50E-05
	repressor	217	2.8	2.20E-05
	rna-binding	262	3.4	2.50E-05
	host-virus interaction	150	2	2.90E-05
	protein biosynthesis	105	1.4	4.80E-05
	neurogenesis	85	1.1	5.80E-05
	isopeptide bond	163	2.1	8.70E-05
	cytoskeleton	299	3.9	9.80E-05
	transferase	611	8	1.20E-04
	sh3 domain	113	1.5	1.60E-04
	transport	720	9.4	2.20E-04
	Proto-oncogene	121	1.6	2.80E-04
	disease mutation	686	9	3.30E-04
	ATP	123	1.6	3.90E-04
	magnesium	211	2.8	7.70E-04
	ligase	152	2	7.80E-04
	coiled coil	853	11.2	7.50E-04
	golgi apparatus	272	3.6	8.80E-04
	endoplasmic reticulum	324	4.2	8.70E-04
	developmental protein	351	4.6	9.10E-04
	Endocytosis	57	0.7	1.10E-03
	calmodulin-binding	68	0.9	1.40E-03
	ubl conjugation pathway	236	3.1	2.20E-03
	mrna processing	129	1.7	3.30E-03
	Apoptosis	180	2.4	4.60E-03
	ribosome	44	0.6	5.40E-03
	ribonucleoprotein	136	1.8	5.50E-03
	phosphotransferase	102	1.3	5.70E-03
	mrna splicing	105	1.4	7.30E-03
	tyrosine-protein kinase	61	0.8	8.90E-03

c

Category	**Term**	**Gene Count**	**Fold Enrichment**	**p-value**
7,137 Non SAS genes				
GO Biological Process	defense response to bacterium	75	1.1	2.70E-11
	chemotaxis	82	1.1	1.30E-04
GO Molecular Function	chemokine activity	35	0.5	1.10E-06
	chemokine receptor binding	36	0.5	1.50E-06
	serine-type endopeptidase inhibitor activity	52	0.7	1.30E-04
UniProt Keywords	Secreted	809	11.3	1.30E-43
	signal	1359	19	1.80E-36
	Antimicrobial	57	0.8	1.70E-13
	antibiotic	55	0.8	2.30E-13
	cytokine	107	1.5	2.20E-11
	defensin	36	0.5	6.60E-11
	disulfide bond	1101	15.4	1.10E-10
	hormone	59	0.8	3.40E-09
	protease inhibitor	64	0.9	8.20E-07
	cleavage on pair of basic residues	132	1.8	1.00E-06
	chemotaxis	45	0.6	2.70E-05
	inflammatory response	46	0.6	1.10E-04
	Lectin	80	1.1	1.60E-04
	Serine protease inhibitor	46	0.6	3.90E-04
	Intermediate filament	45	0.6	4.20E-04
	immune response	104	1.5	4.00E-04
	plasma	50	0.7	1.40E-03
	neuropeptide	24	0.3	2.10E-03
	inflammation	19	0.3	2.20E-03
	Monooxygenase	43	0.6	2.30E-03
	amidation	28	0.4	2.40E-03
	fungicide	11	0.2	3.60E-03

Supplementary Table 4. **Concordance of SAS genes with alternative splicing across species.**

For twelve metazoan organisms, a significant enrichment of known SAS genes (**a**) and novel antisense transcription (**b**; as measured by the presence of antisense ESTs with known orientation) was observed in genes with multiple rather than single transcripts. **(a)** For each species, the proportion of genes with multiple or single isoforms that are known SAS genes is tabulated along with the enrichment of SAS genes in the multiple transcript genes, and the p-value of that enrichment (Student's t-Test). Similarly **(b)**, the average expression of ESTs with known orientation mapping antisense to genes with multiple or single transcripts is enumerated along with the enrichment and significance for each species.

a

	Known SAS (proportion)			
Species	**Genes with multiple isoforms**	**Genes with single isoforms**	**Enrichment**	**P value**
human	0.22	0.16	1.37	3.48E-25
fugu	0.04	0.02	1.80	9.64E-14
mouse	0.20	0.12	1.69	2.25E-66
chimp	0.10	0.09	1.14	2.61E-03
rhesus	0.12	0.10	1.16	3.35E-04
rat	0.11	0.09	1.22	8.97E-06
ciona	0.14	0.11	1.36	5.15E-09
drosophila	0.29	0.25	1.16	2.51E-06
xenopus	0.14	0.05	3.16	1.24E-64
chicken	0.12	0.12	1.06	2.41E-01
nematode	0.16	0.09	1.84	1.05E-33
zebrafish	0.11	0.05	1.95	2.12E-20

b

		Spliced antisense ESTs (proportion)			
Species	**Nr ESTs**	**Genes with multiple isoforms**	**Genes with single isoforms**	**Enrichment**	**P value**
Human	473344	0.8	0.5	1.67	1.00E-307
frog	220668	0.5	0.5	1.00	4.00E-01
zebrafish	142692	0.5	0.4	1.34	5.26E-41
mouse	127259	0.6	0.3	1.81	2.34E-273
sea squirt	106988	0.3	0.2	1.35	3.60E-20
rat	101393	0.5	0.3	1.70	1.23E-164
nematode	98089	0.9	0.5	1.78	1.00E-307
fly	26609	0.7	0.3	2.06	1.75E-189
chicken	14219	0.3	0.2	2.06	5.60E-80
puffer fish	1443	0.0	0.0	0.85	2.78E-01
rhesus	569	0.0	0.0	1.59	6.45E-02
chimp*	66	0.0	0.0	0.91	8.95E-01

* Although there were over 1.4 milion ESTs available for chimp, the spliced EST table downloaded from UCSC was nearly unpopulated, and very likely under-represents the true number of spliced ESTs for chimp. When considering all chimp ESTs, there was a significant enrichment (1.68) of antisense ESTs in the multiple versus single transcript gene categories (p = 1.98 x 10^{-77}). These numbers are comparable to the results from all human ESTs (data not shown).

Supplementary Table 5. Known SAS genes of comparable length to non-SAS genes have more annotated isoforms.

Bin range (%)	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**80-90**
T-test pvalue (Nr. isoforms)	0.43	0.35	0.062	0.011	0.0086	0.0014
Mean (Non-SAS isoforms)	2.7	2.63	2.52	2.45	2.38	2.4
Mean (Known SAS isoforms)	2.81	2.53	2.75	2.73	2.68	2.67
Nr genes in bin (Known SAS)	307	320	343	371	374	635
Nr genes in category (Non-SAS)	2190	2750	3214	3592	3789	3629

Supplementary References

Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, Agirre E, Plass M, Eyras E, Elela SA et al. 2009. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat Struct Mol Biol* **16**(7): 717-724.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al. 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research* **37**(Database issue): D885.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4**(9): R60.61-R60.11.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. M2. *J Mol Biol* **313**(4): 903-919.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols* **4**(1): 44-57.

Hubbell E, Liu W-M, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18**(12): 1585-1592.

Michael A, Catherine AB, Judith AB, David B, Heather B, Cherry JM, Allan PD, Kara D, Selina SD, Janan TE et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1): 25.

Nag A, Narsinh K, Kazerouninia A, Martinson HG. 2006. The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. *RNA* **12**(8): 1534-1544.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**(5): 887-898.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology* **16**(9): 990-995.