compared with chromosomal duplication content (R^2 = 0.16). The correlation was due to intrachromosomal duplications (fig. S5; R^2 = 0.20; P = 0.04; F test) and was absent for interchromosomal duplications (R^2 = 0.002). The three most gene-rich chromosomes showed high levels of duplication, and the seven most gene-poor chromosomes were among the least duplicated chromosomes.

To determine what role recent segmental duplications have played in current gene evolution, we characterized the gene content in our filtered set of duplicated genomic sequence. We analyzed a highly curated set of 13,351 mRNAs assigned to the human genome assembly (RefSeq, www.ncbi.nlm.nih.gov/LocusLink/refseq.html). We partitioned exons from each gene into a unique or duplicated sequence on the basis of their map position (>90% sequence identity). We identified a total of 7777 exons as being transcribed from recently duplicated sequence, corresponding to 6.1% of all RefSeq exons (128,467). This is slightly greater than the genomic representation of segmental duplication (5.2%), which confirms that gene-poor regions have not been preferentially duplicated. In many cases, a complete complement of exons was not duplicated. These incomplete duplicated genes were often found adjacent to other duplicated cassettes that originated from elsewhere in the genome. By comparing our data with human expressed sequence tag databases, we found evidence for "chimeric" or fusion transcripts that emerged from the physical juxtaposition of incomplete segmental duplications. Although the mechanism for recent segmental duplications is not understood, the existing data suggest the process may play a role in exon shuffling associated with expanding protein diversity. A complete list of all genes with one more exons within duplicated genomic sequence is available (*8*).

To further assess whether specific kinds of genes or biological processes have been preferentially duplicated, we compared all RefSeq mRNAs on the basis of their INTERPRO protein domain classification (Table 2) (table S7) (*23*). In this analysis, we considered a gene duplicated only if all its exons were contained within a duplicated genomic region. Our analysis suggests a nonrandom distribution of segmental duplications within the proteome. Genes associated with immunity and defense (natural killer receptors, defensins, interferons, serine proteases, cytokines), membrane surface interactions (galectins, HLA, lipocalins, carcinoembryonic antigens), drug detoxification (cytochrome P450), and growth/development (somatotropins, chorionic gonadotropins, pregnancy-specific glycoproteins) were particularly enriched. It should be emphasized that our gene analysis is restricted to genomic segments that show ≥90% sequence identity. On the basis of neutral expectation of divergence, this corresponds to duplications that have emerged over the last ~40 million years of human evolution (*24*). Gene duplication followed by functional specialization has long been considered a major evolutionary force for gene innovation (*25*). Therefore, these genes embedded within recent genomic duplications may be considered excellent candidates for adaptations specific to primate evolution.

References and Notes

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* **11**, 1005 (2001).
3. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
4. S. Ohno, U. Wolf, N. Atkin, *Hereditas* **59**, 169 (1968).
5. E. E. Eichler, *Trends Genet.* **17**, 661 (2001).
6. P. Stankiewicz, J. R. Lupski, *Trends Genet.* **18**, 74 (2002).
7. E. E. Eichler, *Genome Res.* **11**, 653 (2001).
8. See supporting data on *Science* Online.
9. V. E. Cheung *et al.*, *Nature* **409**, 953 (2001).
10. The sequence and underlying test statistics for all duplicated regions of the genome are available at http://humanparalogy.cwru.edu/SDD. WGAC comparisons of the human genome assembly (UCSC, August freeze, 2001; http://genome.ucsc.edu) were done as described (*2*). The WSSD-filtered set of WGAC duplications can be interactively searched (http://humanparalogy.cwru.edu/SDD). This includes extracted sequence files, the actual alignments, the location of the alignments within the assembly, and whole-chromosomal views comparing WGAC and WSSD duplication patterns. An updated WSSD based on the analysis of 39,298 clones from April 2002, detecting an additional 36 Mb of duplicated sequence, is also available.
11. T. H. Shaikh *et al.*, *Hum. Mol. Genet.* **9**, 489 (2000).
12. J. A. Bailey *et al.*, *Am. J. Hum. Genet.* **70**, 83 (2002).
13. L. Edelmann, R. K. Pandita, B. E. Morrow, *Am. J. Hum. Genet.* **64**, 1076 (1999).
14. R. Mazzarella, D. Schlessinger, *Genome Res.* **8**, 1007 (1998).
15. J. R. Lupski, *Trends Genet.* **14**, 417 (1998).
16. S. T. Sherry *et al.*, *Nucleic Acids Res.* **29**, 308 (2001).
17. K. Chen *et al.*, *Nature Genet.* **17**, 154 (1997).
18. S. L. Christian, J. A. Fantes, S. K. Mewborn, B. Huang, D. H. Ledbetter, *Hum. Mol. Genet.* **8**, 1025 (1999).
19. D. E. Jenne *et al.*, *Am. J. Hum. Genet.* **69**, 516 (2001).
20. T. Kuroda-Kawaguchi *et al.*, *Nature Genet.* **29**, 279 (2001).
21. H. C. Mefford, B. J. Trask, *Nature Rev. Genet.* **3**, 91 (2002).
22. J. Guy *et al.*, *Hum. Mol. Genet.* **9**, 2029 (2000).
23. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
24. W. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
25. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).

26. We thank L. Christ, M. Eichler, and U. Neuss for technical assistance, and H. Willard, J. Nadeau, T. Hassold, D. Locke, and J. Horvath for helpful comments. Supported by NIH grants GM58815 and HG002318 and U.S. Department of Energy grant ER62862 (E.E.E.), NIH Career Development Program in Genomic Epidemiology of Cancer (CA094816) and Medical Scientist Training Grant (J.A.B.), the W. M. Keck Foundation, and the Charles B. Wang Foundation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/297/5583/1003/DC1
Materials and Methods
Tables S1 to S7
Figs. S1 to S5

20 March 2002; accepted 7 June 2002

Predictive Identification of Exonic Splicing Enhancers in Human Genes

William G. Fairbrother,[1,2]* Ru-Fang Yeh,[1]* Phillip A. Sharp,[1,2] Christopher B. Burge[1]†

Specific short oligonucleotide sequences that enhance pre-mRNA splicing when present in exons, termed exonic splicing enhancers (ESEs), play important roles in constitutive and alternative splicing. A computational method, RESCUE-ESE, was developed that predicts which sequences have ESE activity by statistical analysis of exon-intron and splice site composition. When large data sets of human gene sequences were used, this method identified 10 predicted ESE motifs. Representatives of all 10 motifs were found to display enhancer activity in vivo, whereas point mutants of these sequences exhibited sharply reduced activity. The motifs identified enable prediction of the splicing phenotypes of exonic mutations in human genes.

Human genes are generally transcribed as much longer precursors, typically tens of kilobases in length, from which large introns must be precisely removed and flanking exons precisely ligated to create the mRNA that will direct protein synthesis. Sequences around the splice junctions—the 5′ and 3′ splice sites (5′ss and 3′ss)—are clearly important for splice site recognition. However, these signals appear to contain only about half of the information required for exon and intron recognition in human transcripts (*1*). The sequence or structure context in the vi-

[1]Department of Biology, [2]Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: cburge@mit.edu

cinity of the 5′ss and 3′ss motifs is known to play an important role in splice site recognition (*2*–*4*). ESE sequences, which enhance splicing at nearby sites (*5*), are an important component of this context.

Exonic enhancers have been identified through the analysis of disease alleles (*6*), by site-directed mutagenesis of minigene constructs, and by protocols based on SELEX (Systematic Evolution of Ligands by EXponential enrichment) to identify sequences with enhancer activity from a pool of random sequences (*7*–*11*). These methods initially characterized ESEs as purine-rich sequences, but additional classes of AC-rich motifs and pyrimidine-rich motifs have since emerged (*7*, *10*).

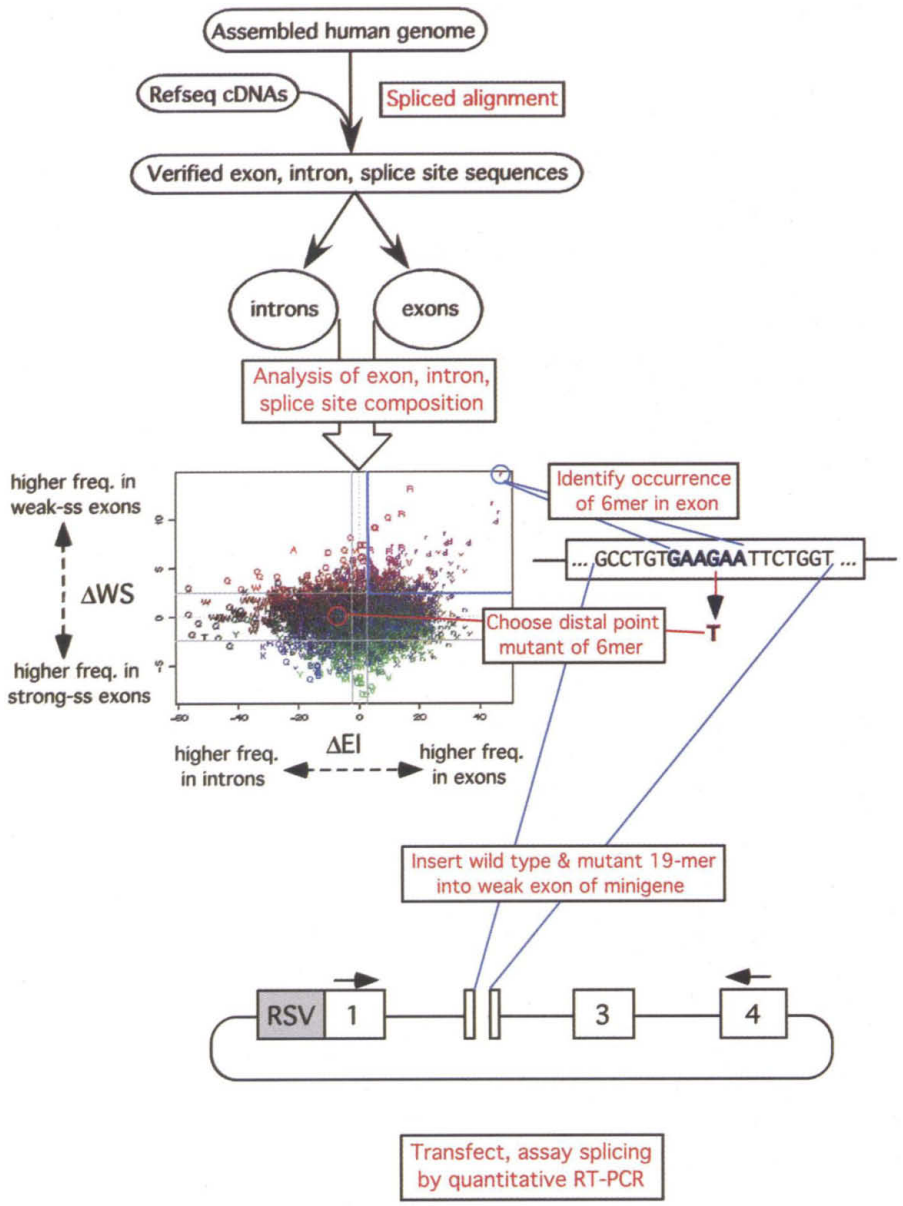


Fig. 1. Schematic of RESCUE-ESE approach. Exon-intron structures of human genes are derived by spliced alignment of cDNAs to the assembled genomic sequence, and splice sites are scored as described (*17*). Values of ΔEI (scaled difference in frequency between exons and introns) and ΔWS (scaled difference in frequency between weak and strong exons) are calculated as described for each of the 4096 possible hexanucleotides (*17*). Each hexamer is then represented by a colored letter at the point (ΔEI, ΔWS) in the scatterplot. The letters are chosen to reflect the base composition of the hexamer according to IUPAC nomenclature (e.g., hexamers containing only A and G are represented by the letter "r"). Hexamers containing homonucleotide runs of three or more bases (e.g., AAA) are represented by capital letters, all other hexamers by lowercase letters. Each letter is colored proportional to the relative content of A (red), C (green), G (blue), and T (black) of the hexamer. Hexamers (6mers) satisfying ΔEI > 2.5 and ΔWS > 2.5 (upper right portion of first quadrant) are predicted to have ESE activity. As a test of ESE activity, a 19-base "extended exemplar" sequence containing the hexamer in its natural context in a weak exon is chosen and inserted into the SXN splicing reporter construct as indicated. SXN is a β-globin–derived minigene with deleted translation start codon. A point mutant predicted to disrupt ESE activity is also chosen, generally the single-base mutant that is farthest to the left and below the predicted ESE hexamer in the scatterplot. Transient transfection of the reporter construct followed by quantitative RT-PCR with flanking primers is used to assay inclusion of the test exon for the candidate ESE and its mutant.

Our strategy for identifying human ESE sequences was to first develop a statistical/computational method to predict the ESE activity of oligonucleotide sequence motifs, to apply this method to large data sets of human genomic sequences, and then to test representatives of each predicted motif by means of an in vivo splicing assay. At the heart of this approach is a sequence analysis method that we call RESCUE (Relative Enhancer and Silencer Classification by Unanimous Enrichment). RESCUE identifies the set of oligonucleotide motifs that enhance or repress a particular biochemical process; it consists of four steps: (i) Identify two or more statistical "attributes" that should be manifested by sequences that enhance (or, alternatively, repress) the biochemical activity of interest. (ii) Use a statistical power calculation to determine an oligonucleotide "word" size *k* appropriate for the amount of data available. Then represent all possible oligonucleotides of size *k* by points in a multidimensional space, the axes of which represent the attributes chosen in the previous step. (iii) Define a region in this space corresponding to "unanimous enrichment" (i.e., significantly high values of all of the chosen attributes) and identify clusters of similar sequences that fall in this region. (iv) Align the sequences in each cluster to produce motifs, and test representative sequence(s) from each motif and appropriate point mutants with the use of a suitable functional assay.

A large body of work suggests that ESEs are located in the general vicinity of splice sites (*12*). Unlike transcriptional enhancers, ESEs function in a strongly position-dependent manner, enhancing splicing when present downstream of a 3′ss and/or upstream of a 5′ss (*13*), but often repressing splicing when present in intronic locations (*14*, *15*). These observations suggest that, as one attribute, ESE sequences should be strongly selected for in constitutively spliced exons and generally avoided in intronic sequences near splice sites.

Moreover, ESEs can compensate for the presence of "weak" (nonconsensus) 5′ or 3′ splice signals in exons, and strengthening of the splice sites of an enhancer-dependent exon generally eliminates enhancer dependence (*16*). Therefore, we conjecture that exons with nonconsensus splice sites ("weak exons") are under much stronger selective

pressure to retain ESEs than are exons with consensus splice sites ("strong exons"), resulting in a significantly higher frequency of ESEs in weak exons than in strong exons.

Available full-length cDNA sequences were aligned to the assembled human genome by means of the spliced alignment algorithm that is part of the "Genoa" gene annotation script (*17*). Reliable full-length alignments were obtained with this approach for 4817 human genes containing 31,463 introns and 28,933 internal exons. Position-specific log-odds score matrices were then used to score the 5′ss and 3′ss of these exons, and the distributions of 5′ss and 3′ss scores were used to partition exons into categories on the basis of the strength of their splice sites: "weak 5′ exons" (bottom 25% of 5′ss scores), "strong 5′ exons" (top 25% of 5′ss scores), with "weak 3′ exons" and "strong 3′ exons" defined analogously.

Application of the RESCUE-ESE method to this set of human genes is illustrated in Fig. 1. A power calculation dictated the use of a word size of six nucleotides, which is comparable in size to the binding sites of many known RNA binding factors (*17*). In step two, each of the 4096 oligonucleotides of length six was assigned two scores: ΔEI, the scaled difference between the frequency of occurrence of the hexamer in exons and the frequency of occurrence near splice sites in introns (scaled in standard deviation units); and Δ5WS, the scaled difference between the frequency of occurrence of the hexamer in weak 5′ exons and its frequency in strong 5′ exons (SD units), with Δ3WS defined analogously for weak 3′ exons versus strong 3′ exons. Each hexamer was then represented by a point in the plane with coordinates (ΔEI, Δ5WS) for identification of sequences that enhance 5′ss recognition (5′ESEs) (Fig. 2A). Alternatively, each hexamer was represented by the point (ΔEI, Δ3WS) for identification of sequences that enhance 3′ss recognition (3′ESEs) (Fig. 2B). A statistical significance threshold of 2.5 standard deviations above the mean (corresponding to a P value of ~0.01) was then applied to each axis independently; that is, any hexamer for which both ΔEI > 2.5 and Δ5WS > 2.5 is predicted to be a 5′ESE, and any hexamer with both ΔEI > 2.5 and Δ3WS > 2.5 is predicted to be a 3′ESE (hexamers in the upper right portion of the first quadrant in the scatterplots). The requirement that each hexamer exceed thresholds in two separate dimensions, both with $P \sim 0.01$, represents essentially a Bonferroni-type correction for multiple comparisons: Because 4096 different tests are being performed, the combined P value is set to $\sim(0.01)^2 = 10^{-4}$, giving an expectation of less than one false positive hexamer.

These criteria identified 103 different hexamers as candidate 5′ESEs and 198 hexamers as candidate 3′ESEs. These two sets overlap fairly extensively, with 63 of the 103 predicted 5′ESEs also contained in the set of predicted 3′ESEs, which suggests that many enhancers may be capable of acting at both splice sites [e.g., (*13*)]. The total number of hexamers predicted to display either 5′ or 3′ ESE activity was 238 out of the 4096 possible hexamers, about 6% of the total, consistent with the notion that ESEs are quite common.

In step three of the RESCUE procedure, predicted 5′ESE and 3′ESE hexamers were clustered on the basis of sequence similarity, and the hexamers in each cluster were multiply aligned using CLUSTALW (*18*) to identify candidate enhancer motifs (fig. S3) (*17*). This procedure yielded a total of five 5′ESE motifs (Fig. 2A) and eight 3′ESE motifs (Fig. 2B). Three of the five 5′ESE motifs—5A, 5B, and 5C—are significantly similar to 3′ESE motifs 3G, 3A, and 3D, respectively, so the three pairs 5A/3G, 5B/3A, and 5C/3D were considered to represent just three distinct classes, each comprising the union of the pair of similar hexamer clusters (*17*). The total number of distinct candidate enhancer motifs identified by RESCUE-ESE was therefore 10.

In the final step of the RESCUE procedure, representatives of these candidate enhancer motifs were tested for ESE activity in a splicing reporter construct. For each cluster of predicted ESEs, a representative hexamer was chosen—referred to as the "exemplar" of the class. To place each exemplar hexamer in its natural context, we screened our human spliced gene database for an occurrence of each exemplar hexamer in a weak 5′ exon (bottom 10% of 5′ss scores) or weak 3′ exon (bottom 10% of 3′ss scores), as appropriate. A slightly longer region of sequence centered on the exemplar—referred to as the "extended exemplar"—was then chosen from this exon and inserted into the reporter construct described below. The extended exemplar

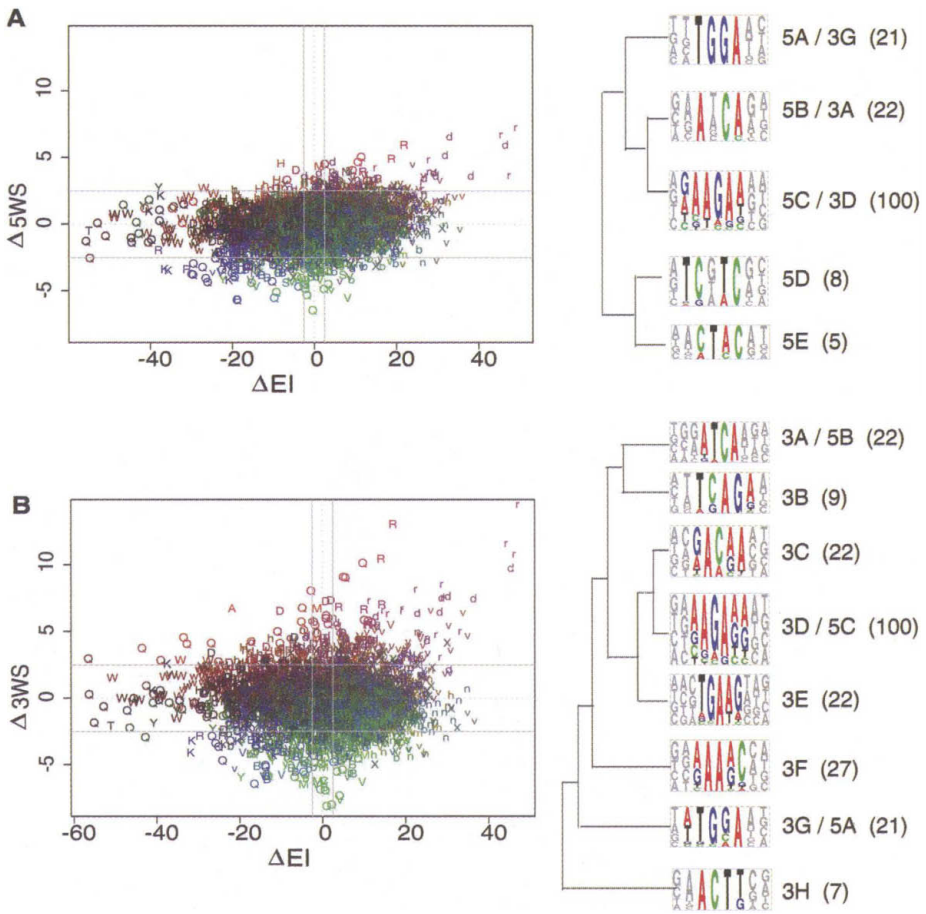


Fig. 2. RESCUE-ESE prediction of 5′ and 3′ ESEs in human genes. (**A**) Scatterplot for prediction of 5′ESE activity. Hexamers are represented by colored letters as described in Fig. 1. Simplified dendrogram shows clustering of 5′ESE hexamers (total of 103 hexamers with ΔEI > 2.5 and Δ5WS > 2.5) into five clusters of four or more hexamers. (**B**) Scatterplot for prediction of 3′ESE activity. Simplified dendrogram shows clustering of 3′ESE hexamers (total of 198 hexamers with ΔEI > 2.5 and Δ3WS > 2.5) into eight clusters of four or more hexamers. Complete dendrograms of all hexamers are shown in fig. S3. The aligned sequences in each cluster are represented as Pictograms (http://genes.mit.edu/pictogram.html). Cluster labels (e.g., 3B, 5A/3G) are listed to the right of each Pictogram, with the total number of hexamers in the cluster indicated in parentheses. Clustering and alignment were performed as described (*17*).

sequences comprise the 19-base region extending from six bases 5′ of the exemplar hexamer to seven bases 3′ of the exemplar hexamer (Fig. 1).

The splicing enhancer activity of each extended exemplar sequence was then assessed by measuring its ability to "rescue" splicing of exon 2 of the reporter construct, pSXN (*7*). SXN exon 2 is only 32 bases long, including the 19-base insert. Previously, this exon was observed to be predominantly skipped for most random insert sequences tested. This failure to be included is reversed when the exon is lengthened, when a splicing enhancer is present, or when the 5′ss, the branch point, or the polypyrimidine tract is improved (fig. S1) (*19*–*21*). Because strengthening either the 5′ss or 3′ss consensus sequence of SXN exon 2 or inserting an ESE causes exon inclusion, we reasoned that this exon would be a suitable reporter system for testing the activity of candidate 5′ESEs as well as candidate 3′ESEs.

It was of particular interest to assess the ability of the RESCUE approach to predict ESE-disrupting mutations. Therefore, for each exemplar hexamer, a single-base mutant was chosen that was predicted to lack enhancer activity [i.e., did not fall in the extreme upper right ("unanimous enrichment") region of the scatterplot]. Typically, the single-point mutant farthest "southwest of" (to the left and below) the exemplar in the scatterplot was chosen. A "mutant" extended exemplar sequence containing just this single base change was then generated for each extended exemplar and inserted into the same cloning site in the SXN minigene. Constructs containing the extended exemplars and mutants were transiently transfected into HeLa cells, and the splicing phenotype was assayed by quantitative reverse-transcription polymerase chain reaction (RT-PCR) (see fig. S2 for protocol and quantitation curves).

An initial set of experiments evaluated the robustness of the approach with respect to differences in the local context of the exemplar hexamer. For this purpose we focused on a representative hexamer, GAAGAA, chosen from the large purine-rich 5C/3D cluster of predicted enhancers (Fig. 2). The consensus sequences for these clusters and the chosen exemplar are similar to the classical "GARGAR" enhancer (R represents either purine nucleotide, A or G). Occurrences of GAAGAA were identified in three exons with weak splice sites, generating extended exemplar sequences GAAGAA.1, GAAGAA.2, and GAAGAA.3, which lack appreciable similarity other than the shared hexamer GAAGAA (Fig. 3). All three extended exemplars conferred high levels of inclusion on the test exon, ranging from ~50% for GAAGAA.3 to ~70% for GAAGAA.1 (see fig. S4 for representative gels). All three of these extended exemplars contained additional purine-rich hexamers overlapping the central GAAGAA hexamer, which were also predicted to have enhancer activity by RESCUE-ESE (indicated by the vertical blue bars in Fig. 3). Next, the mutation G4>T was introduced into each extended exemplar [i.e., each central GAAGAA hexamer was mutated to GAATAA, a hexamer that falls far "southwest" of GAAGAA in the scatterplots and is predicted to lack ESE activity (see Fig. 4, motif 5C/3D)]. This mutation also disrupts many or all (for GAAGAA.3) of the overlapping RESCUE-ESE hexamers. As predicted, this mutation produced sharply reduced levels of inclusion in each of the three contexts, ranging from ~5% to ~30% of the wild-type level (Fig. 3). Taken together, these data suggest that different occurrences of the same exemplar tend to be qualitatively similar in their ability to enhance splicing and in their response to specific point mutations, but that the precise level of ESE activity depends on local sequence context. Another mutation predicted to disrupt ESE activity of GAAGAA, A2>T (Fig. 3, M2), also gave reduced levels of exon inclusion in the context of GAAGAA.3. On the other hand, the mutation A5>C (Fig. 3, M3) is predicted to preserve ESE activity because it converts GAAGAA to GAAGCA, another predicted ESE hexamer, and this mutation slightly increases exon inclusion in the context of GAAGAA.3 (Fig. 3). These data anecdotally suggest that RESCUE-ESE can accurately predict which mutations will disrupt the enhancing activity of an ESE; some evidence for this conclusion is discussed below.

To assess the degree to which different exemplar hexamers from the same cluster would have similar ESE activity, we chose a quite different exemplar, AGAAAC, from the same 5C/3D cluster as GAAGAA. The extended exemplar AGAAAC.1 also displayed ESE activity in the range observed for the different extended exemplars of GAAGAA (Fig. 3). However, the mutation G2>T, predicted to disrupt the activity of AGAAAC, gave only a moderate (~27%) reduction in exon inclusion, from ~75% to ~55%. This remaining ESE activity might be attributable

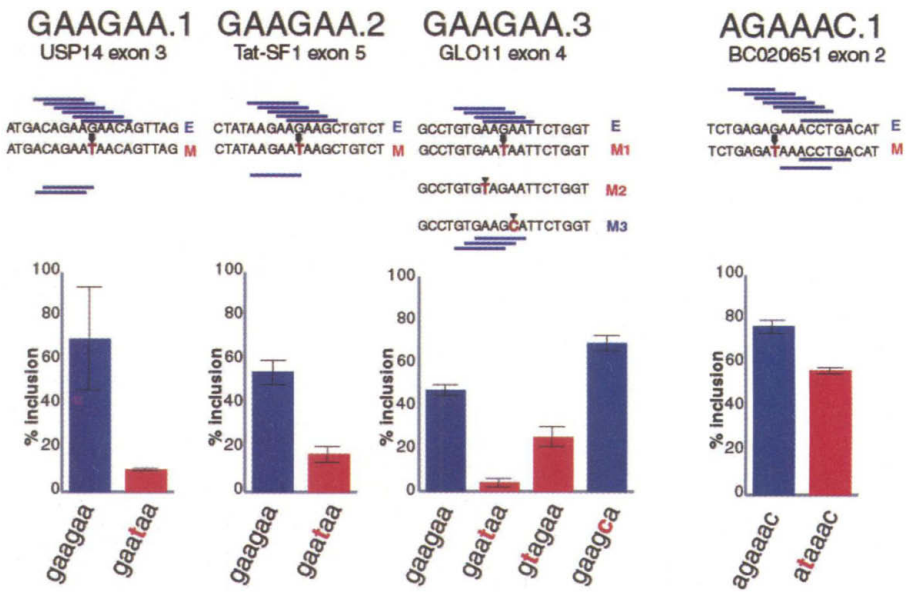


Fig. 3. Analysis of ESE activity for predicted enhancers of class 5C/3D. Upper panel: Extended exemplar sequences for three occurrences of the GAAGAA exemplar and one occurrence of the AGAAAC exemplar. All extended exemplars derive from arbitrarily selected occurrences of the exemplar in human exons with weak splice sites, as described in the text. Gene name and exon number are listed above each sequence. GenBank accession numbers for the mRNAs are as follows: XM_046769 (GAAGAA.1), XM_010365 (GAAGAA.2), AF212232 (GAAGAA.3), and BC020651 (AGAAAC.1). Predicted ESE hexamers in each extended exemplar are indicated by blue bars above the sequence. Point mutations introduced into these sequences are shown in red, with predicted ESE hexamers in the mutant sequence shown by blue bars below the sequence. Each mutant is labeled by a red M if the mutation is predicted to disrupt ESE activity, or by a blue M if the mutant sequence is predicted to retain ESE activity. Total RNA extracted from HeLa cells was amplified by RT-PCR after transient transfection with the SXN reporter containing the indicated insert. Radiolabeled products were analyzed by polyacrylamide gel electrophoresis and visualized using a phosphorimager. (Representative autoradiographs are shown in fig. S4.) Bottom panel: Percent inclusion for each construct was calculated as the ratio of the intensity of the upper band (including exon 2) to the sum of the intensities of the upper and lower bands. All transfections were performed at least twice. The height of the colored bar indicates the average of all measurements; horizontal black lines indicate the minimum and maximum inclusion values observed.

to the retention of two predicted ESE hexamers in the mutated AGAAAC.1 sequence (Fig. 3), although this was not tested. This example underscores the difficulty in interpreting the results of mutations in sequences containing additional predicted ESE hexamers. To rigorously test the predictions of the RESCUE-ESE method, we used sequences specifically chosen to contain exactly one predicted ESE hexamer (or zero, in the case of mutant sequences) in all other experiments reported here.

One exemplar and a corresponding extended exemplar were chosen from each of the 10 motifs for testing in the reporter construct. Only 19-nucleotide oligomers containing a single RESCUE-ESE–predicted hexamer in the middle were considered (table S1). Although this restriction might result in a bias toward selection of weaker enhancers, it was considered essential in order to avoid complications in interpreting the splicing phenotypes of sequences containing overlapping or adjacent predicted ESEs. For each motif, a single-base mutant predicted to disrupt the ESE activity of the exemplar hexamer was chosen as described above, introduced into the extended exemplar sequence, and cloned into the reporter construct. The 10 predicted enhancer and mutant constructs were transiently transfected and assayed for splicing as before (Fig. 4).

All 10 of the predicted enhancers (blue bars) displayed ESE activity in the reporter system, ranging from weakly enhancing (~20% inclusion for 5B/3A and 3B) to strongly enhancing (~60 to 80% inclusion, for 5D and 3E). In addition, for 9 of 10 classes of enhancer tested, the predicted ESE sequence gave a significantly higher level of inclusion than the mutant (blue bar higher than red bar), motif 3F being the only exception (see fig. S5 for representative gels). These results demonstrate the effectiveness of RESCUE-ESE for prediction of the effects of single base changes on ESE activity. The different point mutant sequences exhibited varying levels of inclusion. Mutant 3F gave comparable inclusion to wild-type 3F enhancer. Three other mutants, 3C, 3E, and 3H, gave about two-thirds the level of inclusion of the wild-type sequence, indicating that ESE activity had been only partially impaired. On the other hand, the remaining six mutants all had 10 to 50% of the wild-type level of inclusion. In absolute terms, these six mutants had inclusion levels below 20% and often less than 10%, comparable to that seen by others for typical random inserts in this context (*7*).

In a large set of human exons, slightly more than 10% of all the hexanucleotides were found to match RESCUE-ESE hexamers (*22*), often in overlapping clumps as in Fig. 3, suggesting that ESEs are very common in human genes. Counting each overlapping clump as a single enhancer, we found an average of 5.2 predicted enhancers per exon, with most exons containing between three and seven ESEs (20th and 80th percentiles, respectively). The hexamers in each cluster typically occurred more frequently in exons than in introns by a factor of 1.5 to 2 and more frequently in weak exons than in strong exons by a factor of 1.3 to 1.4. The average frequencies of hexamers in each of the 10 RESCUE-ESE motif clusters were comparable or slightly lower in a database of more than 2000 alternative (skipped) exons than in our database of constitutively spliced exons (*22*); this finding suggests that the motifs we have identified are involved in recognition of both constitutively and alternatively spliced exons.

Some sequences that display ESE activity were missed by the RESCUE method in its current form. For example, mutant 3F and one of the three predicted "neutral" sequences tested (fig. S5) displayed enhancer activity but did not contain any RESCUE-predicted ESE hexamers. Analysis of the three predicted neutral sequences—19-base segments that lack predicted ESE hexamers chosen from exons with weak splice sites—suggests a possible modification of the cutoffs used in the RESCUE-ESE protocol. Specifically, it was found that neutral sequence N3 contained a hexamer that was close to the cutoff for ESEs: The hexamer CTACGC had $\Delta EI = 16.9$ ($\gg 2.5$) and $\Delta 5WS = 2.2$, just below the cutoff. By contrast, no hexamer in neutral sequence N1 or N2 had both $\Delta EI > 2.5$ and $\Delta 5WS$ or $\Delta 3WS > 1.5$, and no hexamer in any of the neutral sequences had $\Delta 5WS$ or $\Delta 3WS > 2.5$. These and other data (*22*) suggest that altering the cutoffs used in RESCUE-ESE, perhaps by increasing the ΔEI cutoff while simultaneously reducing the ΔWS cutoff to 1.5 or 2, might result in improved detection of ESEs.

A database of published mutationally characterized natural ESE sequences was constructed, and these sequences were searched for occurrences of the hexamers in each cluster (tables S2 to S4). Five of the RESCUE-ESE clusters (5C/3D, 5E, 3C, 3E, and 3F) resemble

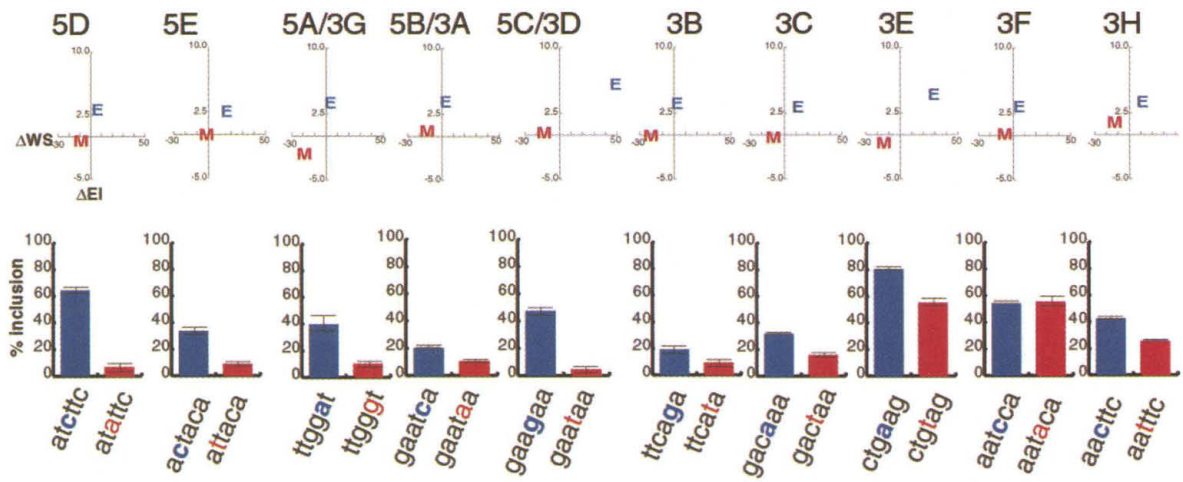
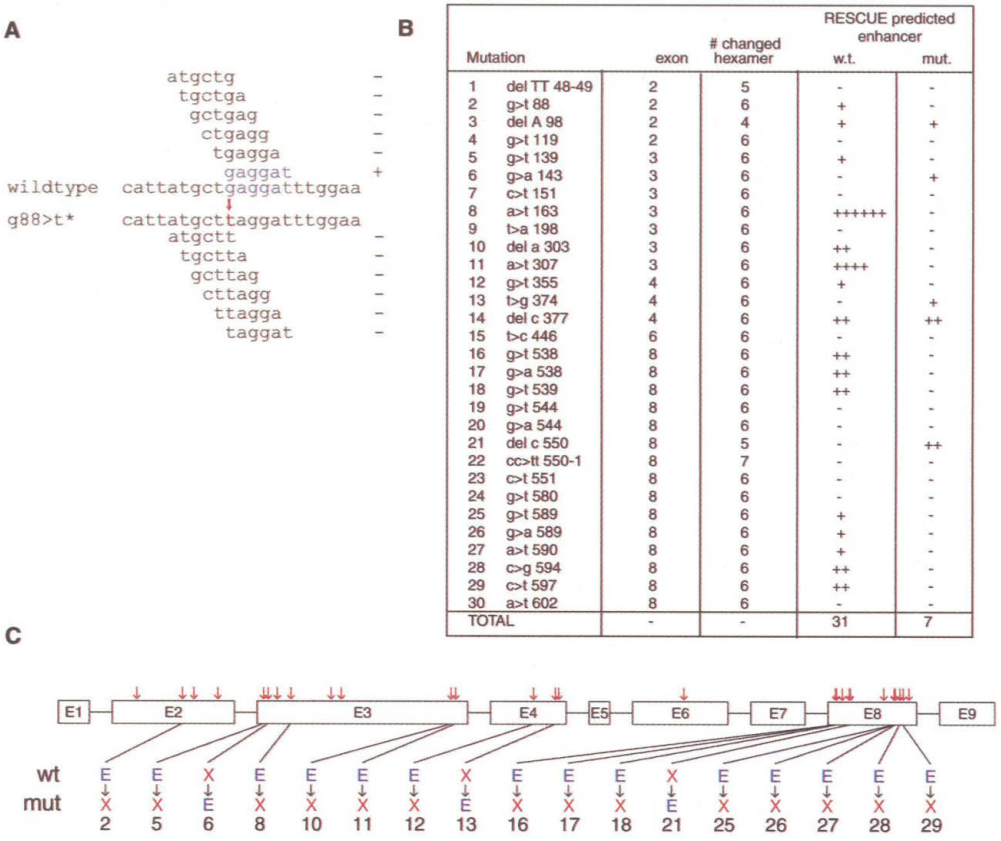


Fig. 4. Analysis of ESE activity for 10 classes of predicted enhancers. HeLa cells were transfected with SXN splicing reporter construct containing inserts representing all 10 classes of predicted ESEs and point mutants of these sequences. The extended exemplar sequences used are listed in table S1. Upper panel: schematic representing ΔEI and ΔWS values for each tested exemplar hexamer (blue E) and point mutant hexamer (red M) from Fig. 2A or 2B, as appropriate. The label of the predicted ESE cluster from Fig. 2 is indicated above. Lower panel: Percent inclusion for each construct, calculated as in Fig. 3. (Representative autoradiographs are shown in fig. S5.) All transfections were performed at least twice. The height of the colored bar (blue for predicted ESE, red for mutant predicted to disrupt ESE activity) indicates the average of all measurements; horizontal black lines indicate the minimum and maximum inclusion values observed. The predicted ESE hexamer and point mutant sequence are shown below the blue and red bars, respectively, with the mutated base shown in the corresponding color.



Mutation		exon	# changed hexamer	RESCUE predicted enhancer w.t.	mut.
1	del TT 48-49	2	5	-	-
2	g>t 88	2	6	+	-
3	del A 98	2	4	+	+
4	g>t 119	2	6	-	-
5	g>t 139	3	6	+	-
6	g>a 143	3	6	-	+
7	c>t 151	3	6	-	-
8	a>t 163	3	6	++++++	-
9	t>a 198	3	6	-	-
10	del a 303	3	6	++	-
11	a>t 307	3	6	++++	-
12	g>t 355	4	6	+	-
13	t>g 374	4	6	-	+
14	del c 377	4	6	++	++
15	t>c 446	6	6	-	-
16	g>t 538	8	6	++	-
17	g>a 538	8	6	++	-
18	g>t 539	8	6	++	-
19	g>t 544	8	6	-	-
20	g>a 544	8	6	-	-
21	del c 550	8	5	-	++
22	cc>tt 550-1	8	7	-	-
23	c>t 551	8	6	-	-
24	g>t 580	8	6	-	-
25	g>t 589	8	6	+	-
26	g>a 589	8	6	+	-
27	a>t 590	8	6	+	-
28	c>g 594	8	6	++	-
29	c>t 597	8	6	++	-
30	a>t 602	8	6	-	-
TOTAL		-	-	31	7

Fig. 5. Correlation between predicted ESEs and exon skipping mutations in human *HPRT* gene. (**A**) Exon skipping mutations were analyzed in terms of the set of hexamers affected; mutation 2 (G88>T) is shown here as an example. All hexamers affected by the mutation are shown, with matches to RESCUE-predicted ESEs shown in blue and represented by a plus sign. The location of the mutation is indicated by a red arrow. (**B**) Summary of human *HPRT* gene mutations known to cause exon skipping [from (*23, 24*)]. Base changes that occurred within five nucleotides of a splice junction were excluded, as they may alter the splice site signals. The first column lists the nature of the mutation (del = deletion, X>Y = substitution of base Y for base X), with coordinates listed relative to the translation start site of the *HPRT* cDNA (GenBank accession number NM_000194). The last two columns list the number of predicted ESE hexamers in the affected region of the wild-type and mutant sequences, respectively. (**C**) Locations of all mutations listed in (B) are indicated relative to the exon-intron structure of the *HPRT* gene by red arrows. Exon sizes are to scale; intron sizes are not. Mutations that alter RESCUE-predicted ESEs are shown below the exon-intron schematic, numbered according to (B). Mutations labeled E→X disrupt predicted ESE hexamers; those labeled X→E create predicted ESEs.

natural ESEs; the other five (5A/3G, 5B/3A, 5D, 3B, and 3H) are not similar to known, mutationally defined ESEs. Many SELEX motifs have not been confirmed by mutational analysis and so are not treated as "known" enhancers here. On the other hand, some functional SELEX and all binding SELEX results identify trans-factor interactions, which are often not known for natural ESEs. Databases of functional SELEX and binding SELEX consensus motifs were also constructed by surveying the literature (tables S5 and S6, respectively) and were compared to RESCUE-ESE hexamers. All "complete" matches spanning either the entire hexamer and/or the entire SELEX motif are listed in table S1 (e.g., the hexamer GAAGAA would be counted as a match to GARGAR, YGARGAR, or ARGAR, but not to YARGAR). There were matches to some of the clusters that lacked similarity to known natural ESEs, as well as several matches to the purine-rich cluster 5C/3D. The consensus sequence for cluster 5D, TC[GT]TC, is a particularly good match for the functional SELEX motif UCUUC and also matches fairly well to positions 2 to 6 of the SRp20-binding SELEX motif YWCUUCAU (Y = C or T; W = A or T). The cluster 3H consensus (ACTT) also matches fairly well to two functional SELEX motifs and to another binding SELEX motif for SRp20 (table S2). The match between cluster 3C, consensus GACRA, and the 9G8 binding SELEX motif AGACKACGAY (K = G or T) is also reasonable. These similarities suggest plausible and testable trans-factor interactions.

To evaluate the quality of RESCUE predictions in the context of a natural gene locus, we considered the published set of exon mutations that cause exon skipping in the *HPRT* gene, mutation of which is associated with Lesch-Nyhan syndrome (*23, 24*). The *HPRT* cDNA was considered as a set of overlapping hexanucleotides, each beginning one base 3′ of the beginning of the previous hexamer. In this representation, a point mutation disrupts six overlapping hexamers in the wild-type sequence, creating six new hexamers that are point mutants of the wild-type hexamers. Similarly, a one-base deletion at a given position disrupts six wild-type hexamers and creates five new ones in their place. Of the 30 *HPRT* mutations that alter hexamer composition and cause exon skipping, 16 positions coincide with at least one predicted enhancer motif (Fig. 5). Overall, a total of 31 predicted ESE hexamers are present in positions overlapping the mutated positions in the wild-type sequence, whereas only seven predicted ESE hexamers are present in the mutant sequences [$P < 0.01$ (*17*)]. For 14 of these cases, all predicted ESE activity is lost in the mutant sequence, as compared to only three positions where mutations create predicted ESEs not present in the wild-type sequence. This ratio (14/3 = ~4.7) of predicted ESE-down mutations to predicted ESE-up mutations is higher than that reported recently (43/21 = ~2.0) for an approach using weight matrices based on functional SELEX results [see table 1 of (*25*)]. Thus, RESCUE-ESE should prove useful as a predictive tool for analyzing the splicing phenotypes of exonic mutations or polymorphisms associated with human disease. The general RESCUE approach could also be applied to identify enhancers or silencers of other processes such as polyadenylation or transcription.

References and Notes

1. L. P. Lim, C. B. Burge, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11193 (2001).
2. D. L. Black, *RNA* **1**, 763 (1995).
3. K. K. Nelson, M. R. Green, *Genes Dev.* **2**, 319 (1988).
4. R. Reed, T. Maniatis, *Cell* **46**, 681 (1986).
5. B. J. Blencowe, *Trends Biochem. Sci.* **25**, 106 (2000).
6. L. Cartegni, S. L. Chew, A. R. Krainer, *Nature Rev. Genet.* **3**, 285 (2002).
7. L. R. Coulter, M. A. Landree, T. A. Cooper, *Mol. Cell. Biol.* **17**, 2143 (1997).
8. H. X. Liu, M. Zhang, A. R. Krainer, *Genes Dev.* **12**, 1998 (1998).
9. H. X. Liu, S. L. Chew, L. Cartegni, M. Q. Zhang, A. R. Krainer, *Mol. Cell. Biol.* **20**, 1063 (2000).

10. T. D. Schaal, T. Maniatis, *Mol. Cell. Biol.* **19**, 1705 (1999).
11. H. Tian, R. Kole, *Mol. Cell. Biol.* **15**, 6291 (1995).
12. S. M. Berget, *J. Biol. Chem.* **270**, 2411 (1995).
13. C. F. Bourgeois, M. Popielarz, G. Hildwein, J. Stevenin, *Mol. Cell. Biol.* **19**, 7347 (1999).
14. A. Kanopka, O. Muhlemann, G. Akusjarvi, *Nature* **381**, 535 (1996).
15. L. M. McNally, M. T. McNally, *Mol. Cell. Biol.* **18**, 3103 (1998).
16. B. R. Graveley, *RNA* **6**, 1197 (2000).
17. Supporting data are on *Science* Online.
18. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
19. D. L. Black, *Genes Dev.* **5**, 389 (1991).
20. Z. Dominski, R. Kole, *Mol. Cell. Biol.* **11**, 6075 (1991).
21. ———, *Mol. Cell. Biol.* **12**, 2108 (1992).
22. R.-F. Yeh, C. B. Burge, data not shown.
23. M. Tu, W. Tong, R. Perkins, C. R. Valentine, *Mutat. Res.* **432**, 15 (2000).
24. C. R. Valentine, *Mutat. Res.* **411**, 87 (1998).
25. H. X. Liu, L. Cartegni, M. Q. Zhang, A. R. Krainer, *Nature Genet.* **27**, 55 (2001).
26. We thank B. Blencowe, L. Chasin, L. Lim, D. Lipman, and D. Riordan for helpful comments on the manuscript; H. Cargill for help with the figures; T. Cooper for generously providing us with the SXN minigene construct; and anonymous reviewers for helpful suggestions. Supported by a Functional Genomics Innovation Award from the Burroughs Wellcome Fund (C.B.B., P.A.S.) and by NIH grant 1 R01 HG02439-01 (C.B.B.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1073774/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S6
References

9 May 2002; accepted 2 July 2002
Published online 11 July 2002;
10.1126/science.1073774
Include this information when citing this paper.

Microbial Reefs in the Black Sea Fueled by Anaerobic Oxidation of Methane

Walter Michaelis,[1]* Richard Seifert,[1] Katja Nauhaus,[2] Tina Treude,[2] Volker Thiel,[1] Martin Blumenberg,[1] Katrin Knittel,[2] Armin Gieseke,[2] Katharina Peterknecht,[1] Thomas Pape,[1] Antje Boetius,[3] Rudolf Amann,[2] Bo Barker Jørgensen,[2] Friedrich Widdel,[2] Jörn Peckmann,[4] Nikolai V. Pimenov,[5] Maksim B. Gulin[6]

Massive microbial mats covering up to 4-meter-high carbonate buildups prosper at methane seeps in anoxic waters of the northwestern Black Sea shelf. Strong ^{13}C depletions indicate an incorporation of methane carbon into carbonates, bulk biomass, and specific lipids. The mats mainly consist of densely aggregated archaea (phylogenetic ANME-1 cluster) and sulfate-reducing bacteria (*Desulfosarcina/Desulfococcus* group). If incubated in vitro, these mats perform anaerobic oxidation of methane coupled to sulfate reduction. Obviously, anaerobic microbial consortia can generate both carbonate precipitation and substantial biomass accumulation, which has implications for our understanding of carbon cycling during earlier periods of Earth's history.

Until recently, it was believed that only aerobic bacteria, depending on oxygen as an electron acceptor, build up substantial biomass from methane carbon in natural habitats (*1*). Because biogenic methane is strongly depleted in ^{13}C, a worldwide negative excursion in the isotopic signature of organic matter around 2.7 Ga (1 Ga = 10^9 years) ago was taken as an argument for methanotrophy and, consequently, an early oxygenation of the Earth's atmosphere (*2*). However, recent investigations have shown the existence of methane-consuming associations of archaea and sulfate-reducing bacteria (SRB) in anoxic marine sediments (*3*, *4*).

Microorganisms capable of anaerobic growth on methane have not been cultivated so far, and the biochemical pathway of the anaerobic oxidation of methane (AOM) remains speculative. Analyses of depth profiles and radiotracer studies in marine sediments (*5*, *6*) as well as molecular (*7–10*) and petrographic studies (*11*) argue for AOM (*12*) as a crucial process that channels ^{13}C-depleted methane carbon into carbonate and microbial biomass. Through AOM mediated by consortia of archaea and SRB, methane is oxidized with equimolar amounts of sulfate, yielding carbonate and sulfide, respectively (*13*). Generation of alkalinity favors the precipitation of methane-derived bicarbonate according to the following net reaction:

$$CH_4 + SO_4^{2-} + Ca^{2+} \rightarrow CaCO_3 + H_2S + H_2O$$

Here, we provide evidence that vast amounts of microbial biomass may accumulate in an anoxic marine environment because of the use of methane as an electron donor for sulfate reduction (SR) and as an organic carbon source for cell synthesis.

In the northwestern Black Sea, hundreds of active gas seeps occur along the shelf edge west of the Crimea peninsula at water depths between 35 and 800 m (*14*). At some of the shallow Crimean seeps, microbial mats were found associated with isotopically light carbonates. Aspects of the microbiology, sedimentology, mineralogy, and selected biomarker properties of these deposits were recently described (*11*, *15–17*). We explored the seeps on the lower Crimean shelf with the manned submersible *JAGO* from aboard the Russian R/V *Professor Logachev*. During dives to a seep area at 44°46′N, 31°60′E, we discovered a reef consisting of up to 4-m-high and 1-m-wide microbial structures projecting into permanently anoxic bottom water at a depth around 230 m (Fig. 1A). These buildups are formed by up to 10-cm-thick microbial mats that are internally stabilized by carbonate precipitates (Fig. 1A).

From holes in these structures, streams of gas bubbles emanate into the water column (Fig. 1A). The gas contains about 95% methane [see supporting online material (SOM)]. In cross section, the outside of the soft mat has a dark gray to black color (Fig. 1B). Inside the structure, most of the mat is pink to brownish. The interior rigid parts are porous carbonates (aragonite and calcite with up to 14% $MgCO_3$). Much of the structures consists of interconnected, irregularly distributed cavities and channels filled with seawater and gases. Apparently, the cavernous structure of these precipitates enables methane and sulfate to be transported and distributed throughout the massive mats. Smaller microbial structures and nodules from nearby areas were of the same morphology, with compact mat enclosing calcified parts and cavities. Obviously, the microorganisms do not grow on preformed carbonates but induce and shape their formation. Stable carbon isotope analyses of the carbonates yielded $\delta^{13}C$ values ranging from −25.5 to −32.2 per mil (‰) [for methods, see (*11*)]. Compared with the $\delta^{13}C$ values of dissolved inorganic carbon in the Black Sea water column from +0.8‰ at surface to −6.3‰ at depth (*18*), these values indicate that a major portion of the carbonate originates from the oxidation of

[1]Institute of Biogeochemistry and Marine Chemistry, University of Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany. [2]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany. [3]Alfred Wegener Institute for Polar and Marine Research, 27515 Bremerhaven, and International University Bremen, 28725 Bremen, Germany. [4]Geowissenschaftliches Zentrum, University of Göttingen, Goldschmidtstrasse 3, 37077 Göttingen, Germany. [5]Institute of Microbiology, Russian Academy of Sciences, pr. 60-letiya Oktyabrya 7, k. 2, Moscow, 117811, Russia. [6]Institute of Biology of Southern Seas, National Academy of Sciences of Ukraine, pr. Nakhimova 2, Sevastopol, Ukraine.

*To whom correspondence should be addressed. E-mail: michaelis@geowiss.uni-hamburg.de