UNDERSTANDING THE FUNCTIONALITY OF TRANSCRIPT DIVERSITY

EOGHAN HARRINGTON

Inauguraldissertation

zur
Erlangung der Würde eines
Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

August 2007

Heidelberg, Germany

ABSTRACT

Recent years have seen a huge increase in the amount of genomic DNA being sequenced from a wide variety of organisms, giving us an unprecedented insight into the molecular diversity seen in nature. As a result a host of methods have been developed, both experimental and computational, to understand the functional significance of such diversity and how it relates to organismal and environmental complexity. In this thesis I use comparative approaches to explore two areas of molecular biology where there is evidence for large amounts of transcript diversity. Firstly, I explore the unprecedented view of microbial sequence diversity offered by metagenomic sequencing projects, using sequence similarity and adapted genomic context methods to quantify the amount of functional novelty in these samples. Secondly, I look at the transcript diversity generated by alternative splicing. I develop methods to detect and visualise alternative splicing events and apply these to the detection of conserved alternative splicing events.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Published

1. Eoghan D Harrington, Stephanie Boue, Juan Valcarcel, Jens G Reich, and Peer Bork. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet*, 36(9):916917, September 2004 (Harrington et al., 2004).

2. Evgeny M Zdobnov, Monica Campillos, Eoghan D Harrington, David Torrents, and Peer Bork. Protein coding potent al of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res*, 33 (3):946954, 2005 (Zdobnov et al., 2005).

3. Francesca D Ciccarelli, Christian von Mering, Mikita Suyama, Eoghan D Harrington, Elisa Izaurralde, and Peer Bork. Complex genomic rearrangements lead to novel primate gene function. *Genome Res*, 15(3):343351, Mar 2005 (Ciccarelli et al., 2005).

4. Mikita Suyama, Eoghan D Harrington, Peer Bork, and David Torrents. Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. *PLoS Comput Biol*, 2(6):e76, Jun 2006 (Suyama et al., 2006).

5. Jeroen Raes, Eoghan D Harrington, Amoolya Hardev Singh, and Peer Bork. Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol*, 17(3):362369, Jun 2007 (Raes et al., 2007a).

6. Eoghan D Harrington, Amoolya Singh, Tobias Doerks, Christian von Mering, Lars Jensen, Jeroen Raes, and Peer Bork. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *PNAS*, August, 2007 (Harrington et al., 2007).

Submitted

1. Gautier Koscielny, V Le Texier, Eleanor Whitfield, Vasudev Kumanduri, Francesco Nardone, Chellappa Gopalakrishnan, Jean-Jack Riethoven, Christine Fallsehr, Magnus von Knebel Doeberitz, Oliver Hofmann, Winston Hide, Eoghan Harrington, Peer Bork, Stephanie Boue, Eduardo Eyras, Mireya Plass, Fabrice Lopez, William Ritchie, Virginie Moucadel, Daniel Gautheret. ASTD: the Alternative Splicing and Transcript Diversity Database. Submitted 2007.

ACKNOWLEDGMENTS

I would firstly like to thank Peer for giving me the opportunity to come to EMBL and for his patient mentorship over the years.

I would also like to acknowledge the input of members of my thesis advisory committe: Eileen Furlong, Reinhardt Schneider and Mihaela Zavlon. I am also grateful to Mihaela and Walter Keller for supervising my thesis at the University of Basel.

Over the course of my studies I have received invaluable guidance from Stéphanie Boué, Chris Creevey, Seán Hooper, Lars Jensen, Evengenia Kriventseva, Ivica Letunić, Christian von Mering, Brian Naughton, Jeroen Raes, Devin Scannell, Amoolya Singh, Mikita Suyama, David Torrents and Evengy Zdobnov. I would also like to thank all the members of the Bork group, especially Yan Yuan for keeping everything running. I would also like to thank the members of the ASTD consortium for their feedback on my work on alternative splicing.

I would also like to thank those who carried out the important task of helping me to forget about work: Alessia, Alex, Ambra, Anan, Andreia, Barry, Dan, Erwan, Gaëlle, Gráinne, Jan, Jeroen, Jessica, Joël, Jop, Juliette, Kate, Katrien, Laurent, Lorenz, Lukas, Mathilde, Matthieu, Meikel, Mikko, Sascha, Seán, Silvia, Sofia, Stéphanie, Steve, Thore, Warre.

Most of all I would like to thank my family –Mary, Donal, Dara and Cathal– for supporting me over the drawn-out student years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 TRANSCRIPT DIVERSITY AND FUNCTIONAL COMPLEXITY

1.1 INTRODUCTION

Anything found to be true of E. coli must also be true of elephants.

— Jaques Monod, 1954

Monod's famous phrase sums up his belief that the mechanisms responsible for functional complexity are fundamentally the same for all organisms, from simple unicellular prokaryotes to elaborate multicellular eukaryotes. This statement was made following Monod's discovery, along with Francois Jacob, of the *lac* operon, the regulatory module responsible for the transport and metabolism of lactose in *E. coli* (Jacob and Monod, 1961). With only a few components this module provides simple regulatory logic, the operon is activated in response to the presence of lactose, but only if glucose is absent. In the decades following this discovery increasingly complex functional modules have been characterised in a range of organisms, from the module that switches between lytic and lysogenic states of the bacteriophage lambda (Herskowitz and Hagen, 1980), to modules responsible for complex interaction with the environment such as bacterial chemotaxis (Baker et al., 2006), and even modules with complex spatial and temporal features such as developmental patterning (Reeves et al., 2006).

The understanding of progressively more complicated functional modules has been facilitated by advances in technology, allowing us to identify the components of these modules and the functional interactions between them. For example advances in DNA sequencing (Shendure et al., 2004) have made the sequencing of whole genomes cheaper and faster, providing the basis for a complete list of genes, transcripts and proteins for these functional modules. In parallel, advances in technologies such as oligonucleotide microarrays, high-throughput complex affinity purification and mass spectroscopy have allowed us to pick apart the regulatory interactions between these components. However, in contrast to genome sequencing, these technologies have only been applied to a handful of organisms and even then only to a fraction of the genes within. For instance, it is estimated that only 25-31% of human proteins are covered by predicted or experimentally determined structures (Xie and Bourne, 2005), and only 10% of the human interactome has been observed (Hart et al., 2006). The result is that for many organisms genes can be identified, but there is little experimental evidence describing the complexity with which they function together. In such cases where there is a large disparity between the amounts of experimental and genomic data, comparative approaches can be used in a variety of ways to infer both the functions of genes and their interactions (von Mering et al., 2003b). The goal of this thesis is to

apply comparative methods to two different contexts where transcript diversity is high and direct experimental evidence is low.

The study of microbes was on of the first areas to benefit from breakthroughs in DNA sequencing technologies. Since the sequencing of *Haemophilus influenzae* in 1995 by Fleischmann et al. hundreds of microbial genomes have been sequenced (Fleischmann et al., 1995). The wealth of molecular diversity uncovered by these sequencing projects has overturned many preconceptions and provided the basis for insights in many disparate fields (Fraser-Liggett, 2005). However, this diversity is likely to be a tiny fraction of the total. Historically genome sequencing was an expensive process, meaning that microbes of medical or industrial importance were sequenced first, with 40% of baterical genome sequences belonging to human pathogens (Fraser-Liggett, 2005). However this bias pales in comparison to the effect that the inability to culture microbes has had on our view of the molecular diversity of the microbial world. Traditional sequencing methods required large amounts of starting material to create libraries, meaning that only species that could be cultured in laboratory conditions were sequenced (Tringe and Rubin, 2005). Given that it is estimated that only 1% of all prokaryotic species can be cultured (Torsvik and Øvreås, 2002), it seems that our view of the microbial world is limited. Indeed of the 52 bacterial phyla identified by 16S rRNA sequences, only half are represented by cultured species (Riesenfeld et al., 2004). In the past few years, aided by the increasing speed and decreasing cost of DNA sequences, it has become possible to sequence naturally occurring microbial populations to a level where partial assembly is possible, giving us an unprecedented view of prokaryotic sequence diversity. In Chapter 2 I explore this diversity, assessing the level of functional novelty available in these datasets and adapting gene context methods to assign function to completely novel genes.

Monod's opening quote asserts that despite the obvious differences in organismal complexity between prokaryotes and eukaryotes, the same molecular processes are at work in both. This assertion has been largely borne out by the decades of research that have followed, however it has left researchers struggling to determine which of the differences at the molecular level are responsible for differences in organismal complexity. Part of the problem is due to the difficulty in quantifying biological complexity (Adami, 2002), however intuitively it should be some combination of the number of components in the system and the structure and dynamics of the interactions between them. In this sense alternative splicing, the mechanism by which the same primary transcript can yield different mature forms, could represent an important mechanism in the generation of biological complexity as it both increases the number of components and provides an extra regulatory step in gene expression. The first study to assess the importance of the first aspect, the ability to expand the transcriptome, found that it didn't seem to be related to organismal complexity (Brett et al., 2002). While this finding remains controversial (Kim et al., 2004; Harrington et al., 2004; Kim et al., 2007), none of the subsequent studies have looked in gene-level detail at the conservation of alternative splicing. In Chapter 3 I present a tool to that detects and visualises alternative transcription events and use it to detect conserved alternative splicing events.

2 MICROBIAL TRANSCRIPT DIVERSITY

2.1 INTRODUCTION

[1] Recent years have seen an explosion in the amount of shotgun sequence data gathered from diverse natural environments. Since 2004, almost 2 billion base pairs resulting from published large-scale metagenomics sequencing projects have been deposited (as of January of 2007 (Tyson et al., 2004; Venter et al., 2004; Hallam et al., 2004; Tringe et al., 2005; DeLong et al., 2006; Gill et al., 2006; Martín et al., 2006; Turnbaugh et al., 2006)), eclipsing the entire 764 Mbp of previously sequenced microbial genomes (Pruitt et al., 2007) (Figure 2.1). However, true potential for these data lies not only in their sheer volume but also the novel view it gives of microbial communities. Part of this novelty is due to the fact that the sequences produced by these projects are unbiased with respect to culturability, providing an insight into the estimated 99% of species that cannot be sequenced by traditional methods (Torsvik and Øvreås, 2002). However the most exciting insights from this data come from the novel views they give of the structure and functional complexity of microbial communities. For instance, by comparing the gut microbiomes of obese and lean mice, Turnbaugh et al. identified metabolic pathways overrepresented in the obese mouse microbiome that increased the potential for energy harvest from the diet (Turnbaugh et al., 2006). However the true potential of such studies depends on the correct functional annotation of the metagenomic ORFs. In this chapter I will assess the level of functional annotation possible for metagenomic ORFs using traditional sequence similarity methods and newly-adapted gene conext methods.

Currently, the first step in characterizing an unknown sequence involves comparing it to sequences or protein domains of known function in public databases, usually using BLAST (Altschul et al., 1990) or other homology search tools (Bork and Koonin, 1998). By applying BLAST-based annotation methods to the *Eschericia coli K12* genomes, functions can typically be assigned to approximately 80% of the gene products (Raes et al., 2007a) (Figure 2.2). However, these similarity-based methods work best in organisms like *Eschericia coli K12*, where there are many genome sequences available for relatively closely related, well-characterized species. At the lower end of the scale lie Archaebacteria, where there are few full genome sequences and relatively little experimental data, as shown by the fact that less than 40% of the genes of *Aeropyrum pernix* can be characterized by homology-based methods (Figure 2.2). At the other end of the scale lie the symbionts and pathogens with their vastly reduced genomes, such as *Wigglesworthia glossinidia*, with over 90% of genes functionally characterized by homology. For the average fully sequenced bacterial genome, however,

1 Material from this chapter has appeared previously in Harrington, Singh, Doerks, Letunic, von Mering, Jensen, Raes, and Bork (2007) and Raes, Harrington, Singh, and Bork (2007a)

homology-based methods can provide a broad functional characterization for ~73% of genes (Figure 2.2).

Such homology-based methods are subject to several limitations, the most obvious being that they can only assign function to an ORF if it displays significant homology to a previously characterized gene. Moreover, these predictions are susceptible to database propagation errors, which have been estimated to affect 13% of sequences (Brenner, 1999). To complement homology-based function prediction, particularly in prokaryotes, additional information from genomic neighborhood (Dandekar et al., 1998; Overbeek et al., 1999), phylogenetic profiles (Pellegrini et al., 1999), gene co-expression (Marcotte et al., 1999), and gene fusion (Marcotte et al., 1999; Enright et al., 1999) has been utilized and combined (Marcotte et al., 1999; von Mering et al., 2005). These data provide evidence for functional interactions between genes, giving biochemical context and even allowing the characterization of genes for which homology-based methods fail. When these data are added to the homology-based annotation described above, the proportion of genes in the average prokayote that can be functionally characterized rises to almost 85%. As yet, however, only the exploitation of genomic neighborhood (including gene fusions) is feasible in the context of metagenomic shotgun data.

In the first large-scale shotgun metagenomics projects from four diverse and complex environments (tropical surface water from the Sargasso Sea near Bermuda (Venter et al., 2004) , farm soil from Minnesota (Tringe et al., 2005), an acidophilic biofilm from an iron ore mine in northern California (Tyson et al., 2004), and three samples from "whale fall" carcasses on the deep Pacific and Antarctic ocean floor (Tringe et al., 2005)), functions have been predicted based on sequence similarity for only 27% to 48% of the 1.4 million genes in the different samples (Table A.2). This implies that for the majority of proteins in the environment, functions remain unknown and no attempt has yet been made to discover novel functionality. Furthermore, for each project different methods, parameters and even definitions of function were used, which are often not easily accessible to the community, making a comparison of the different samples difficult. To be able to comprehensively predict functions from various metagenomics samples and to get a consistent overview of function in different environments, we developed a sensitive prediction protocol that complements BLAST- and domain-based function predictions with newly developed and adapted gene neighborhood methods. Applying this protocol to the samples revealed a considerable predictive power, indicating that function can be inferred for most of the genes on earth; yet the majority of functions appear to reside in numerous rare, small protein families that remain largely unexplored.

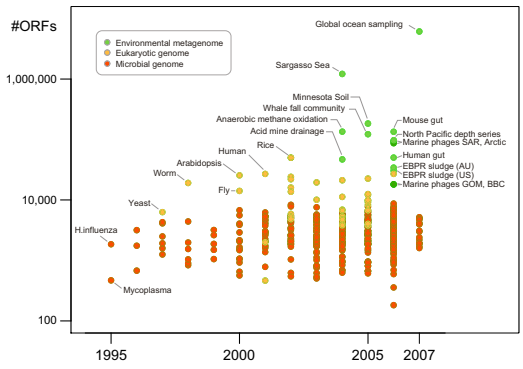


Figure 2.1. Number of ORFs generated by genome sequencing projects (red: bacteria, orange: eukaryotic) and metagenomics projects (light green: microbial, dark green: viral). Data were taken from the GOLD database

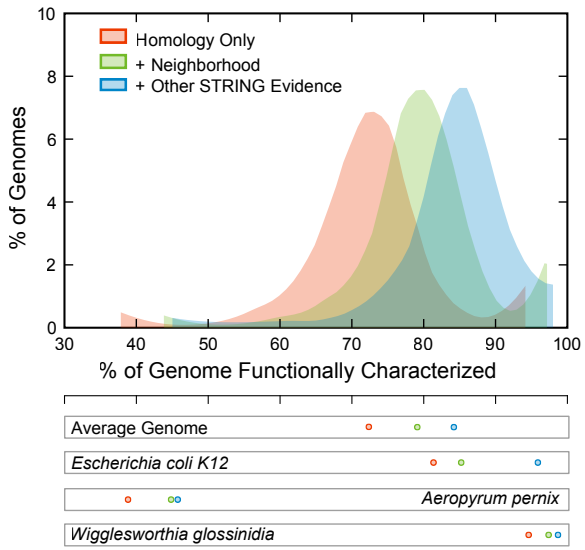


Figure 2.2. Assessment of novelty in fully sequenced genomes by computational methods. Our knowledge of function space is unevenly spread across the tree of life. The 338 prokaryotic genomes in the STRING database (version 7) were classified according to the proportion of proteins for which some inference of function is possible using three different criteria. Using simple homology, we considered functional inference possible for a protein if it can be mapped to a KEGG pathway, a characterized COG or UniRef90 cluster. We then added neighborhood evidence with a score greater than 0.7 from the STRING database to infer function for those proteins in the same neighborhood as those characterized by homology. Similarly, we added all combined evidence from STRING to infer function for the remaining proteins.

2.2 RESULTS AND DISCUSSION

2.2.1 *An operational definition of protein function.*

Biological function is a fuzzy term summarizing a complex concept applicable to different spatial scales (Bork and Koonin, 1998; Bork and Serrano, 2005). At the molecular and cellular level, an operational framework with clearly defined terms and thresholds is therefore required when attempting to quantify protein function. To infer specific function from existing database annotations using homology, we require similarity to an environmental ORF exceeding 60 bits, corresponding roughly to an e-value of 10^{-8} in Uniref90 searches(Tringe et al., 2005). This level of sequence similarity is rather strict in terms of homology identification, but without further analysis may be insufficient to distinguish between paralogs and orthologs, thus not capturing all functional features such as enzyme substrate specificity. It is, however, sufficient to capture basic functionality. To assess the sensitivity of our method to different values of this threshold, analyses were also carried out at 40- and 80-bit cutoffs. The results of these analyses, which show minor difference to those produced with a 60-bit cutoff, described in Section 2.3.

We used a hierarchical classification scheme, favoring manual annotation, to divide environmental ORFs and, for comparison, 124 prokaryotic proteomes into four categories based on the level of functional annotation possible: *(i)* those with strong similarity to, or in the genomic neighborhood of, a gene with specific functional annotation; *(ii)* those with strong similarity to genes with non-specific functional information, weak but significant similarity to genes with any functional annotation, or in the genomic neighborhood of either of these; *(iii)* those with strong similarity to, or in the genomic neighborhood of, a gene of unknown function; *(iv)* those with neither similarity to sequences in annotated databases nor significant genomic neighborhood (Figure 2.3).

We used sequence similarity to infer functional information from the KEGG (Kanehisa et al., 2004), COG (Tatusov et al., 2003), UniRef90 (Wu et al., 2006), SMART (Letunic et al., 2006) and Pfam (Bateman et al., 2004) databases (see Methods for parameter choices, benchmarks and definitions of functional annotation). We utilized gene neighborhood evidence from the STRING database (von Mering et al., 2005) and adapted existing gene neighborhood function prediction methods, based on intergenic distance and evolutionary conservation, for use in fragmented shotgun metagenomics data. First, we exploited the fact that intergenic distances tend to be shorter between genes of the same operon than between operons (Salgado et al., 2000). Although several operon prediction methods have been introduced that are based solely on intergenic distances (Price et al., 2005; Salgado et al., 2000; Okuda et al., 2006; Yan and Moult, 2006), they are either species-specific, trained with experimentally verified transcript information (Salgado et al., 2000), and/or require the context of a complete genome. Here we calibrated directly on each sample to establish the likelihood of being functionally associated given a positional distance within a read. Second, we utilized the fact that neighboring ORFs are more likely to be functionally associated if they are conserved over long evolutionary

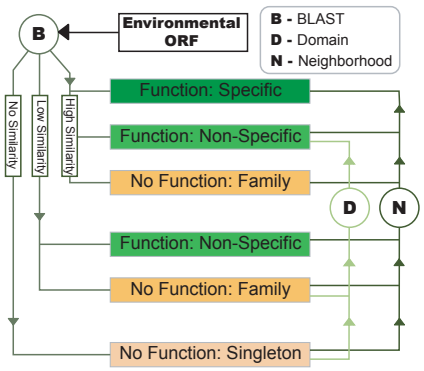


Figure 2.3. Using homology to genes in the KEGG, COG and UniRef90 databases, ORFs were divided into four categories based on the level of functional annotation possible: *(i)* specific functional annotation: ORFs similar to genes with specific functional information; *(ii)* non-specific functional annotation: ORFs similar to genes that have been characterized at a general level or low similarity; *(iii)* no functional annotation but member of an existing family: ORFs with homologs in one of the databases but no functional information (e.g. 'conserved hypothetical'); *(iv)* singletons: ORFs that have no significant similarity to known sequences. ORFs containing domains from the SMART and Pfam A databases were upgraded to having non-specific annotation where applicable. Finally genomic neighborhood methods were used to infer functional links between ORFs and upgrade the functional annotation accordingly.

distances (Dandekar et al., 1998; Overbeek et al., 1999; Korbel et al., 2004). We recorded multiple occurrences of neighboring genes, measured the sequence similarity of the respective neighborhoods to each other and derived a metric based on evolutionary distance. We then combined these measures for intergenic and evolutionary distance to predict functional relationships between genes in the metagenomic data (see Methods).

2.2.2 *Consistent functional characterization of ORFs in four environmental datasets.*

By combining homology searches and neighborhood methods, we were able to infer specific functional information for 76% of the 1.4 million predicted environmental ORFs and a more general level of functional information for a further 7% (dark and light green segments respectively of the outermost ring in Figure 2.4). Using sequence similarity alone, a specific function can be inferred for almost two-thirds (65%) of the ORFs, and a general function for another 13% (inner circle Figure 2.4). Neighborhood-based methods provide functional information for 30% of the ORFs (green segments in middle ring Figure 2.4), complementing similarity-based molecular characterizations with functional interactions. They also provide functional information for almost a quarter

of the ORFs (75,448) where homology-based methods fail. This 30% of neighborhood-based predictions is considerably lower than the 56% achieved when the same methods are applied to the 124 prokaryotic genomes. However, only 47% of the ORFs in the metagenomic datasets have a neighbor in the same transcription direction, as compared to 88% in completely sequenced genomes (Table A.3), which implies that the predictive power of neighborhood methods is comparable in genomes and metagenomes. Indeed, the combined methods perform almost equally well in metagenomes (83% functional characterization) as in fully sequenced genomes (86%). Moreover, the metagenomic ORFs that cannot be characterized by similarity are significantly shorter than those that can (Figure A.32). Some of these may be fragmented ORFs that are too short to assign significant similarity; others may have resulted from erroneous ORF predictions. The latter would imply that the true fraction of gene products for which functions can be predicted is even higher. In either case the quality of predictions should improve in the future as sequence coverage is likely to increase in metagenomics projects allowing more reads to be assembled into longer contigs.

In the original reports of the metagenomics datasets, specific functions were assigned to 27% of the predicted gene products (Tyson et al., 2004; Venter et al., 2004; Tringe and Rubin, 2005), indicating marked differences in the function prediction protocols caused by various technical issues such as the stringency of BLAST cutoffs, the choice of functional databases, and variations in gene calling (a detailed comparison is presented in Table A.2). Since our benchmarks and manual confirmations of parameter settings show a negligible false-positive rate (see Methods), we believe that the near doubling in functional assignments is not caused by a looser function definition or more spurious assignments, but is due to better utilization of existing functional information. The latter uncovers marked trends such as over-representation at the gene, family, or pathway level in line with earlier studies (Tringe et al., 2005) (Table A.7). For example, we find that bacterial chemotaxis, flagellar assembly, and type III secretion genes are 3-fold more frequent in the genomes than the metagenomes (dominated by the surface sea water dataset), perhaps due to the futility of bacterial motility in strong ocean currents. On the other hand, genes involved in amino acid metabolism, as well as in the biosynthesis of nucleotides, carbohydrates, and lipids are significantly under-represented in the genomes as compared to the metagenomes, perhaps due to the bias towards sequencing obligate pathogens, which tend to acquire these compounds from their hosts.

2.2.3 *Comparison of environmental samples.*

Among the four environments, the fraction of functional assignments differs considerably as it does between organisms (Figure 2.4, Figure 2.10, Figure 2.9). In the surface sea water, specific functions are inferable for 82% of ORFs (dark green sections in Figure 2.4); the corresponding fraction in whale fall is 66%, and in soil only 53%. These differences can be partially attributed to inherent differences in the sequence data: for example, the individual read length of the sea water data is longer than in soil (818bp vs. 673bp after quality filtering (Venter

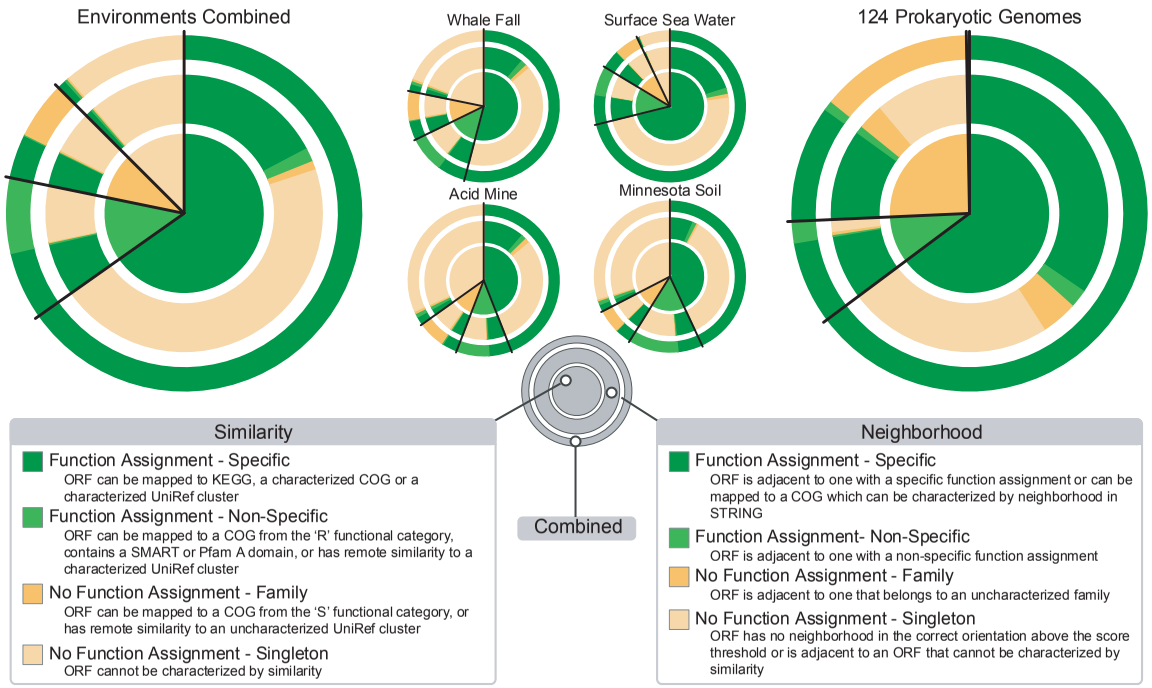


Figure 2.4. Many proteins can be functionally characterized in both datasets. The degree of functional characterization for four metagenomic datasets is shown on the left, and 124 prokaryotic genomes on the right. The inner pie chart represents the level of functional characterization possible using the homology-based approach. The middle ring shows the level of functional characterization possible using neighborhood methods. The outer ring summarises the combined level of characterization possible. Surprisingly, it implies that most metagenomic ORFs (83% of the data) can be functionally characterized, similar to the level possible in fully sequenced genomes.

et al., 2004; Tringe and Rubin, 2005)) and 60% of the sea water reads can be assembled into longer contigs compared to less than 1% in soil (Raes et al., 2007b). Also, environments have been previously characterized to different degrees, and for some environments complete genome sequences are available that closely resemble those from the environment (e.g. SAR11 as a frequent ocean bacterium (Giovannoni et al., 2005)). This not only means more gene context in a certain environment, but also more BLAST assignments for short fragmented ORFs and hence more reliable gene predictions. Finally, a major fraction of the acid mine sample is comprised of Archaea, which are generally less functionally characterized than bacteria, thus lowering our functional understanding of the sample. Nevertheless, we believe that most differences between the environments are caused by multiple effects linked to genuine diversity in phylogeny and lifestyle. For example, genomes of species in the sea water samples are smaller than in soil, with a higher fraction of essential, well-characterized genes (Raes et al., 2007b), but they also evolve faster (von Mering et al., 2007) which should make homology searches less sensitive. Farm soil might supply the most stressors to microbial life due to its high population density, microhabitats, physical and systemic perturbations (e.g. temperature, nutrient availability, and pH) (Torsvik and Øvreås, 2002), leading to a broad repertoire of stress-response phenotypes with hitherto uncharacterized functions. Similarly, the unusual ecological niche created by a deep-sea whale carcass, with its extreme conditions of darkness, cold, and high pressure, lead to highly specialized microbial adaptations such as barotolerance and temperature-induced lipid fluidity (Yayanos, 1995) that do not resemble those in other environments or genomes.

2.2.4 *Predicting functional novelty: in depth analysis of two neighborhood-based findings.*

Whereas homology-based methods require additional analysis to identify novel functions (e.g. via novel subgroups in a characterized sequence family), neighborhood methods can directly provide novel functional associations. Novelty can be obtained either by *(i)* seeing unexpected functional coupling of known genes or *(ii)* assigning unknown genes to known processes. The first is evident in the fact that there are as many as 5,851 pairs of neighboring COGs unique to metagenomes, even though these COGs occur individually in the 124 prokaryotic genomes, implying many novel functional interactions. These frequently include enzymes involved in amino acid biosynthesis with novel links to numerous protein degradation and regulatory proteins, probably reflecting the different nutritional constraints (Table A.8). The second can be seen in the 75,448 ORFs (5% of the total) that are solely characterized by neighborhood. Here we provide detailed functional annotation for two families: a previously uncharacterized gene family associated with a well-known pathway (heme biosynthesis) and a new transcription factor that potentially regulates the coupling of two opposing processes (fatty acid biosynthesis and degradation). These and other functional predictions, including novel annotations for nearly half a million proteins, are available online (http://www.bork.embl.de/Docu/harrington).

Neighborhood information can help characterize a gene family if members of that gene family occur next to different genes belonging to the same pathway in different species. Using such a query, we discovered members of a large uncharacterized gene family (COG1981), with several hundred ORFs in the surface sea water and whale fall samples, adjacent to various enzymes from the well-studied heme biosynthesis pathway (Figure 2.5a). Heme feeds into the synthesis of both cytochromes and chlorophyll and thus plays a key role in enzymatic reactions, energy production, and metabolic regulation (Michal, 1999). In addition, it functions as a prosthetic group to proteins involved in bacterial stress response, oxidative damage, and virulence (Frankenberg et al., 2003). Sequence analysis of the uncharacterized family reveals that it comprises hydrophobic, putative membrane-associated proteins that are unlikely to have enzymatic functions. They might thus be implicated as scaffolding proteins in tethering the pathway to the membrane and/or enabling sufficient substrate fluxes.

Whereas the heme-associated gene family had previously been observed in fully sequenced genomes, another family of 20 members was found exclusively in the surface sea water samples using our clustering procedure (see Methods). Even though no homology could be found using our automated methods, detailed analysis revealed weak but significant similarity to a family of helix-turn-helix (HTH)-transcription factors. An examination of its neighboring genes implies that this family is found in a variety of species, the most closely related being Actinobacteria. As the genes are on various contigs with differing gene orders, we could assign it to an entire operon that additionally contains three downstream genes consistently occurring in the same orientation. The first downstream gene of unknown function (NOG05011) has been observed in completely sequenced genomes; in depth sequence and secondary structure analyses suggest an enzymatic function (data not shown). The second and third genes of this potential operon (COG1024, COG1960) catalyze successive steps of the beta-oxidation of fatty acids (usually involved in degradation) (Yang et al., 1991; Michal, 1999). Interestingly, this invariant operon, apparently controlled by the newly predicted transcriptional regulator, frequently occurs downstream of various genes involved in fatty acids biosynthesis (Figure 2.5b). Thus, context-based methods predict a coupling between fatty acid degradation and biosynthesis, whereby the novel gene might provide the regulation of this link. It is intriguing to speculate that this coupling of two antagonistic processes is an adaptation to repeatedly changing environmental conditions. For instance, strongly regulated circadian rhythms are followed by several marine bacteria (Lakin-Thomas and Brody, 2004). These bacteria actively migrate to different depths in a periodic fashion to balance the efficient usage of light for energy against the danger of DNA-damage (Alexandre et al., 2004; Bebout and Garcia-Pichel, 1995). Energy storage during the light-dependent phase by biosynthesis of fatty acid and energy release in the light-independent phase could thus be a regulated switch during locomotion from light to dark and vice versa.

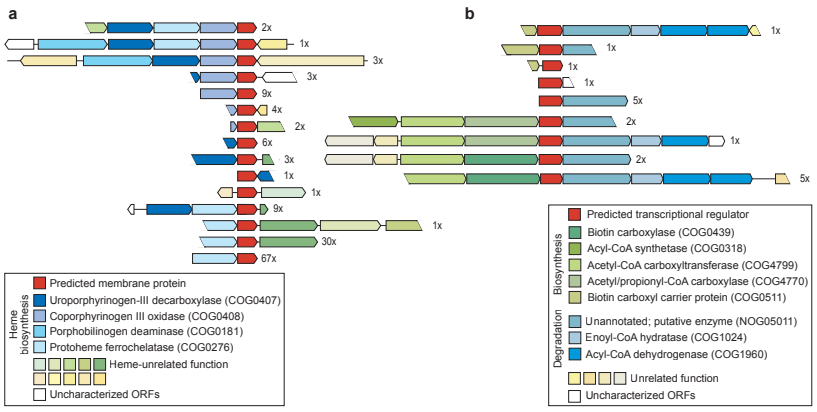


Figure 2.5. Prediction of function in previously uncharacterized gene families using genomic neighborhood. Whereas homology-based approaches quantify the known functions, neighborhood approaches reveal functional novelty, even in conjunction with well-known processes. (a) A putative transmembrane protein belonging to an uncharacterized COG (COG1981 shown in red) that consistently co-occurs with members of the well-characterized heme biosynthesis pathway (colored blue). The putative membrane-associated protein occurs on 174 distinct contigs in the surface sea water and whale fall datasets that can be grouped into at least 15 unique operon arrangements, strongly suggesting a role in this process. (b) A predicted putative regulator, shown in red, that links fatty acid biosynthesis (upstream, colored green) with fatty acid degradation (downstream, colored blue), a functional link not seen in fully sequenced genomes. The regulator appears on 20 distinct contigs in the sea water, of which there are at least five unique operon arrangements.

2.3 MATERIALS AND METHODS

2.3.1 *Sequence data*

We analyzed published microbial shotgun sequence data from four environmental samples, totaling 1,438,944 genes: 1,086,400 genes from tropical surface water from the Sargasso Sea (Venter et al., 2004), 183,586 genes from farm soil from Minnesota (Tringe et al., 2005), 122,146 genes from isolated whale fall carcasses (Tringe et al., 2005), and 46,862 genes from an acidophilic biofilm from an iron ore mine (Tyson et al., 2004) . In parallel, we analyzed 344,619 genes from 124 prokaryotic genomes from the STRING database (von Mering et al., 2005) (Table A.9).

2.3.2 *Function prediction using sequence similarity.*

Each dataset was BLASTed against itself and each of the other datasets. To functionally characterize the data we BLASTed each dataset against proteins from the STRING database (v6) and the UniRef90 database (downloaded 29 March 2006). The parameters used for each search are '*-p blastp -M BLOSUM62 -G 11 -E 1 -z 10000000 -Y 10000000 -v 300 -b 300*'. To assess the sensitivity of our method to different cutoffs we carried out all analyses using 40, 60 and 80 bit score cutoffs, which correspond to e-values of approximately 10^{-1}, 10^{-8} and 10^{-14} in a BLAST against the UniRef90 database with the above alignment parameters (except -z and -Y). To map functionally characterized domains to metagenomic ORFs, we scanned the HMMprofile signatures from Pfam (Bateman et al., 2004) and SMART (Letunic et al., 2006) against the metagenomic sequences using HMMER (http://hmmer.wustl.edu/) software and applied the corresponding family-specific cutoffs.

To be able to intergrate functional information based on similarity to UniRef90 clusters, we first had to divide the UniRef90 database into characterized and uncharacterized clusters. Clusters names matching the regular expression

```
1 (hypothetical)|(unknown)|(unassigned)|(unclassified)|(
      undetermined)|(uncharacteri[zs]ed)|
  (putative)|(predicted)|(probable)|(cluster related to UPI
      .+?;.+similar)
```

were classified as functionally uncharacterized and the remaining clusters were considered characterized. On this basis, 55% (1,086,355) of the UniRef90 clusters were considered functionally characterized. It would be extremely difficult to develop a regular expression that can detect all functionally uninformative annotation. We therefore took a random sample of 200 clusters and checked manually our functional classification. From this we estimate that approximately 4% of clusters are incorrectly classified as characterized (false positives) versus 14% that are incorrectly classified as uncharacterized (false negatives). In theory, any ORF that hits a characterized cluster could be considered characterized; however, due to false positive and negative rates of the classification method and error propagation in automatically annotated databases (Brenner, 1999), we used a threshold to limit the effect of spurious annotations. ORFs were considered characterized if more than

20% of the UniRef90 clusters they hit are characterized (see Figure A.31). To make the results comparable between the prokaryotic genomes and the environmental datasets, we removed self-hits from the results of the BLAST between the prokaryotic genomes and UniRef90 by excluding all 100% identical hits, unless the target cluster was composed of sequences from more than one species.

ORFs were assigned to KEGG pathways and COGs using the method described by Tringe et al. using a 60 bit cutoff (Tringe et al., 2005). For the 124 prokaryotic genomes, the KEGG and COG assignments from the STRING database were used. ORFs were also compared against the UniRef90 database, divided into functionally characterized and uncharacterized clusters (see Supp. Info), and annotated with domains from the SMART and Pfam databases. These annotations were combined in a hierarchical manner, favoring manually annotated databases, placing each ORF into one of the above categories. By definition any ORF that mapped to KEGG was considered to have a specific function assigned. Of the remaining ORFs those that mapped to a COG were considered to have a specific function assigned with the exception of those in functional classes 'R' and 'S' which were considered to have non-specific and no function assigned respectively. The remaining ORFs were considered to have specific functional annotation if they had strong similarity (>60 bits) to functionally characterized UniRef90 clusters, non-specific functional annotation if they contain a domain from the SMART or Pfam A database or have remote homology (>40 bits) to functionally characterized UniRef90 clusters. All other ORFs were considered to have no function assigned, those with similarity to uncharacterized UniRef90 clusters were considered to be part of a family and the rest singletons. This was repeated with cutoffs of 40 and 80 bits (the cutoff for remote homology remaining 40 bits). As seen can be seen from Figure 2.6, varying this cutoff doesn't greatly affect the overal number of ORFs that have some functional information, but does affect the balance between those with specific and non-specific functional annotation.

Any attempt to automatically provide functional annotation for a large dataset is prone to a range of potential errors (Iliopoulos et al., 2003). To test the sensitivity of our homology-based classification method to such errors, we took a random sample of 100 ORFs and carried out a detailed manual analysis, based on which we estimate that the overall false positive rate is 5% and the false negative rate is 18%.

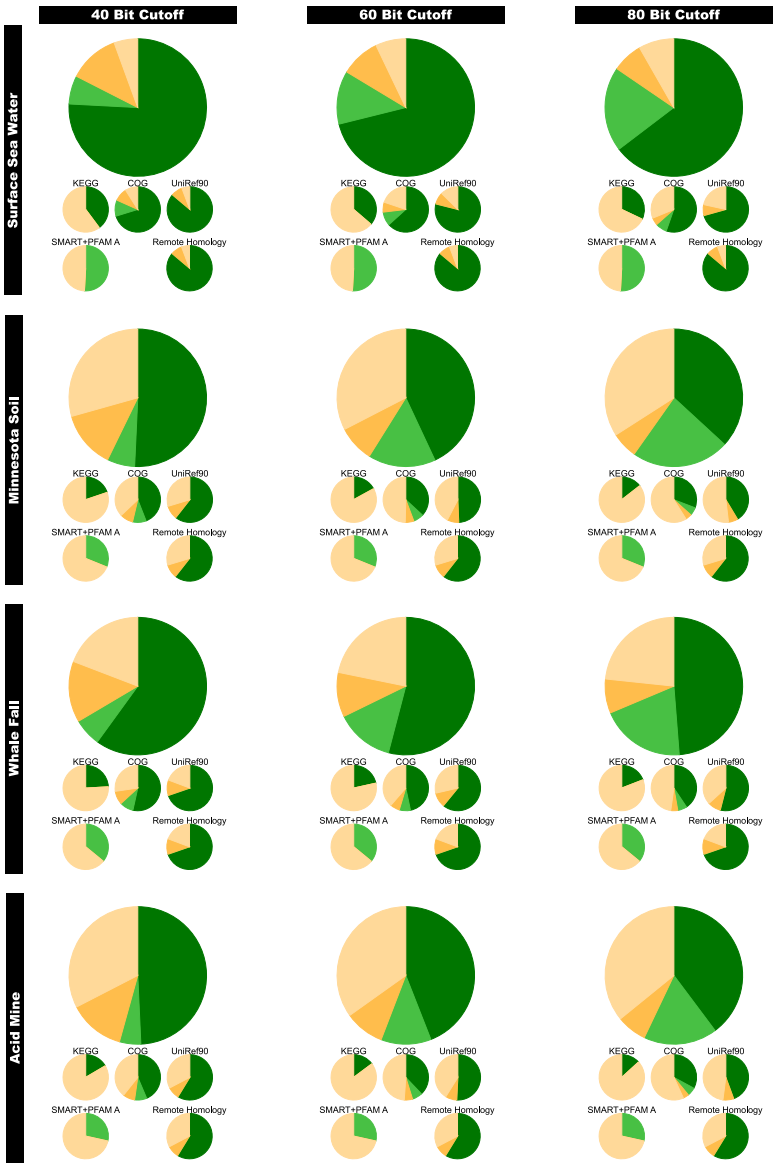


Figure 2.6. Similarity-based functional annotation of 4 metagenomic datasets at 3 different bitscore cutoffs. The smaller pie charts show the amount of functional characterization possible using each of the sources of functional annotation individually while the large pie chart shows the combination of these acording to the procedure described in the methods. Note that the bitscore cutoff only applies to the COG, KEGG and UniRef90 mappings, and remote homology is the same as the UniRef mapping with a 40 bit cutoff

2.3.3 *Function prediction using genomic neighborhood.*

Using the contig positions of the ORFs in each dataset, we constructed a list of pairwise neighborhoods. For this analysis we only considered codirectionally transcribed genes. The difficulty involved in predicting translation initiation sites has led to the prediction of a large number of overlapping genes (Suzek et al., 2001) in both the fully sequenced genomes and the metagenomic data. Some of these genes are in the same phase and therefore likely to be artifacts of the gene prediction process; however, there are also many ORFs with long overlaps. While some of these may represent real overlaps, manual inspection revealed that many are likely to be mispredictions. To reduce the effect that these might have on our analysis, where two genes overlapped by more than 100nt or overlapped in the same phase, we removed the shorter gene from the analysis. The 124 prokaryotic genomes used in this analysis (Table A.9) were chosen to have relatively few large overlaps.

To investigate the conservation of neighborhoods, we constructed a graph for each set of homologous neighborhoods for the metagenomic datasets at each of the three bitscore cutoffs (40, 60 and 80) and for the 124 prokaryotic genomes at a single 60-bit cutoff. An edge was placed between two neighborhoods if there were BLAST hits $\geqslant$ the cutoff between both pairs of genes. This graph was then used to construct clusters of neighborhoods representing a conserved gene pair. To measure the level of conservation of a given gene pair, we adapted a method developed to weight sequences for multiple sequence alignment (Gerstein et al., 1994). For each neighborhood cluster, a distance matrix was constructed where the distance between two neighborhoods was calculated as $1 -$ the average identity between the genes in each neighborhood. This matrix was then used to construct a UPGMA tree using the biopython treecluster algorithm, and then subjected to the algorithm described in *Gerstein et al.* to produce a series of weights for each neighborhood in the cluster. The evolutionary distance for this cluster was taken to be the sum of the unnormalized weights. This score has the property that it will be low for small clusters of closely related sequences and large for clusters with distantly related sequences. This data is plotted on the y-axis of rows A,B and C of Figure 2.7, Figure A.34, Figure A.33, Figure A.35 and Figure 2.8.

For each of the metagenomics datasets at each bitscore cutoff (40, 60, 80) and each individual prokaryotic genome (60 bit cutoff), we constructed a benchmark dataset of the neighborhoods where both members have a KEGG mapping. Using these neighborhoods, we constructed a two-dimensional histogram, the first dimension being intergenic distance (nucleotides) and the second evolutionary distance (conservation score described above). For each bin in this histogram, we measured the fraction of neighborhoods that map to the same KEGG pathway, which can be interpreted as p, the probability that a pair of genes are functionally related. It is possible that the difficulties in predicting genes in metagenomic datasets can lead to split genes that could cause our method to overestimate the value of p. Therefore we removed neighborhoods where both genes map to the same COG. This data is shown in row B of Figure 2.7, Figure A.34, Figure A.33, Figure A.35 and Figure 2.8. We also applied this method to individual

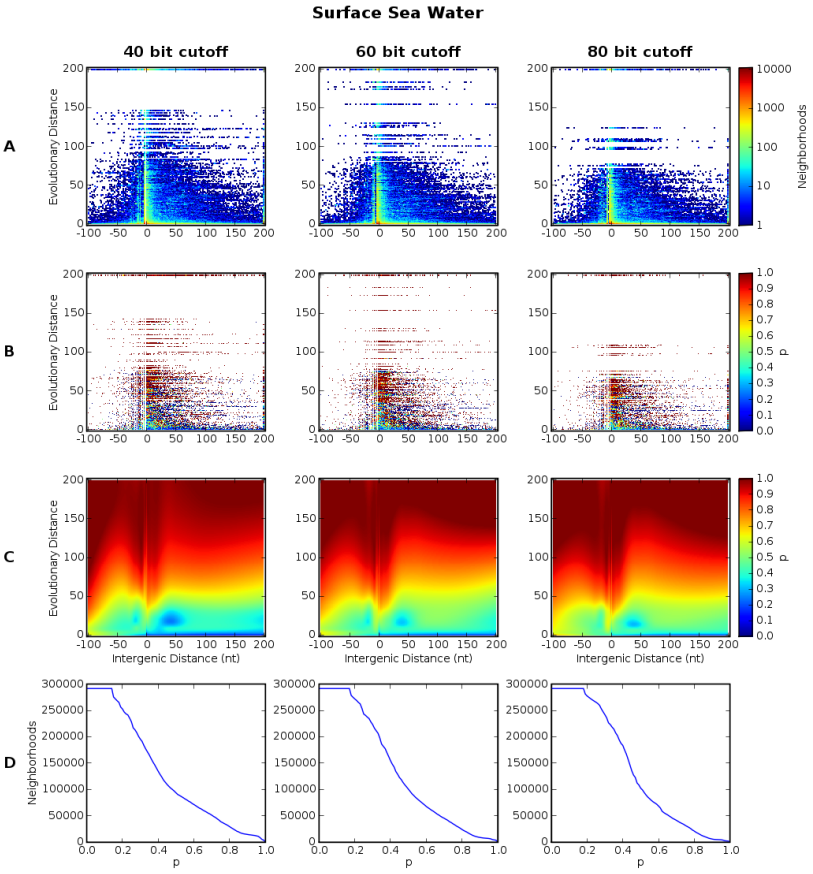


Figure 2.7. Neighborhood method applied to Surface Sea Water data at 3 different bitscore cutoffs. Each column shows the method applied at a different bitscore cutoff, affecting the detection of conserved neighborhoods and the stringency of the KEGG mapping used for the benchmark dataset. Row A shows a 2-dimensional histogram of the all the codirectionally transcribed neighborhoods in the dataset, binned on the x-axis by intergenic distance and on the y-axis by evolutionary distance (see Supp Info for full description). Row B shows the benchmark data, at each intergenic and evolutionary distance p (the proportion of neighborhoods where both genes are functionally related) is shown. Row C shows the interpolation of the data in row B. Row D shows the proportion of neighborhoods with p greater than the cutoff on the x-axis using the predictions from the interpolation in row C. The same plots for the other environments are shown in Figure A.34, Figure A.33, Figure A.35 andFigure 2.8.

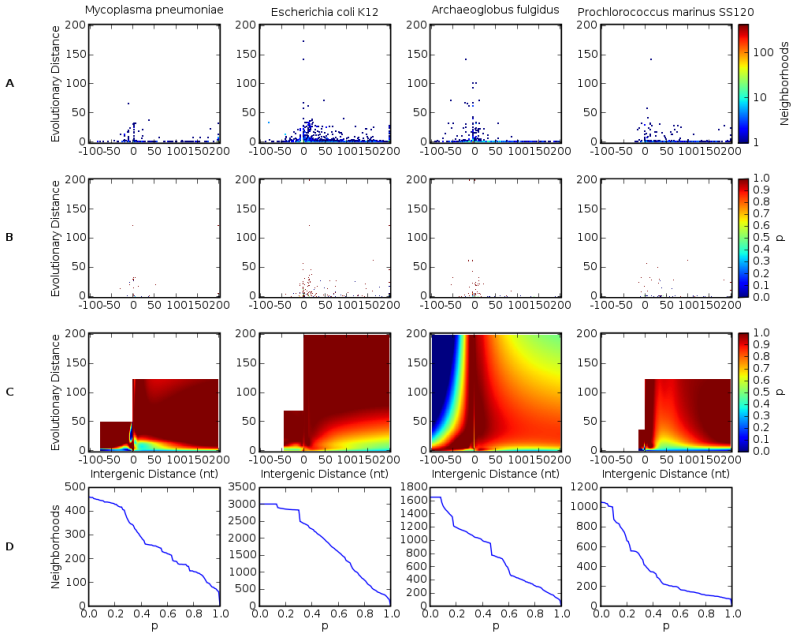


Figure 2.8. Neighborhood method applied to four different prokaryotic species.Row A shows a 2-dimensional histogram of the all the codirectionally transcribed neighborhoods in the dataset, binned on the x-axis by intergenic distance and on the y-axis by evolutionary distance (see Supp Info for full description). Row B shows the benchmark data, at each intergenic and evolutionary distance p (the proportion of neighborhoods where both genes are functionally related) is shown. Row C shows the interpolation of the data in row B. Row D shows the proportion of neighborhoods with p greater than the cutoff on the x-axis using the predictions from the interpolation in row C. Note that for clarity the axes limits are the same for all graphs, however due to the different genome architecture and levels of neighborhood conservation available for individual species the benchmark data may not extend over the full range, causing the blocked appearance of the interpolation in row C. The different genome architectures influence the relationship between intergenic and evoltionary distance and p

organisms (Figure 2.8,Figure 2.9 and Table A.5) to assess the effect of species-specific genome architectures on the method. It is clear that the relationship between intergenic and evolutionary distance and p is highly species-specific.

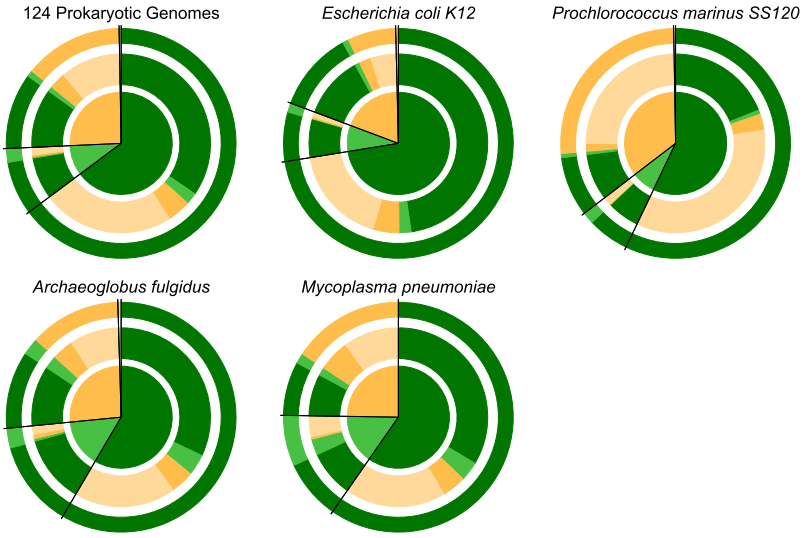


Figure 2.9. Results of the homology and neighborhood methods applied to four representative prokaryotic species

Next, we used the relationship between intergenic and evolutionary distance and p determined for the benchmark set to predict functional relationships for all neighborhoods. Given the sparse nature of the data, it was necessary to first interpolate the relationship over the range of values for intergenic and evolutionary distance. Since we expect different evolutionary pressures to be acting on negatively overlapping genes, we interpolated positive and negatively overlapping neighborhoods separately. A weighted 2-dimensional loess interpolation was carried out using the interp.loess function of the tgp package in R. Due to the sparsity of the data, we first log transformed both the evolutionary and intergenic distances before performing the interpolation. Each point was weighted by the number of neighborhoods contributing to that data point. Grid lengths of 1000 and 500 we used for the positive and negative overlaps respectively. A span parameter of 0.5 was chosen after considering a range of values. The vast majority of p values exceed the random expectation (16%, the probability that a random pair of genes map to the same KEGG pathway). To ensure that we were dealing with high quality predictions, however, we only considered a pair of genes to be functionally linked if the p value was greater than 0.4 (in a previous study (von Mering et al., 2003a) this was found to have an accuracy approaching 70% at the level of functional modules). In addition to utilising the neighborhood data availible within the metagenomic datasets we also integrated information from the STRING database. Genes that map to orthologous groups with no or

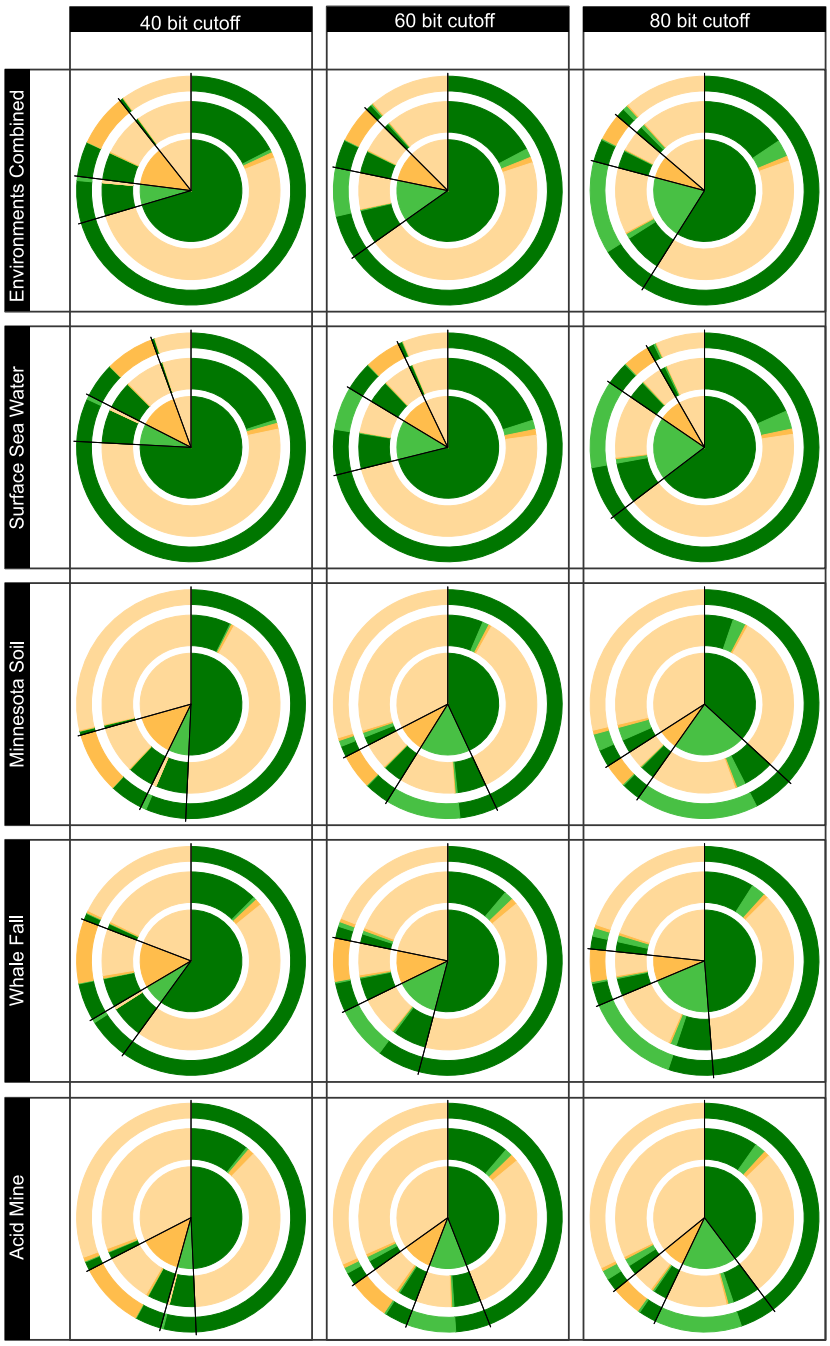


Figure 2.10. A comparison of the homology and neighborhood methods applied to the metagenomic datasets across 3 different bitscore cutoffs. For more a detailed look at the effect of the bitscore cutoff on homology-based methods see Figure 2.6 and for neighborhood methods see Figure 2.7,Figure A.34, Figure A.33 and Figure A.35

non-specific functional annotation were upgraded if that orthologous group was linked to a functionally characterized orthologous group by a significant neighborhood score (>2) in the STRING database.

2.3.4 *Identification of over/under-represented KEGG maps*

To identify biological processes that are significantly over- or under-represented in the environmental samples relative to the fully sequenced prokaryotic genomes, we counted the number of proteins from each of these to sets that could be assigned to each KEGG map. For a given map, the statistical significance of over- or under-representation was assessed using a two-sided Fisher's exact test, and the resulting p-values were corrected for multiple testing by applying the Bonferroni correction. For the maps that display a statistically significant skew, the absolute difference was summarized by calculating the fraction of proteins from each set that was assigned to the KEGG map in question. The most significant maps are displayed in Table A.7.

2.3.5 *Gene family analysis.*

We grouped genes from all four environmental datasets into 206,217 gene families by first constructing a single-linkage graph of an all-against-all BLAST (60 bit cutoff), with nodes representing proteins, and edges representing BLAST hits between proteins weighted by BLAST bitscores. This graph was then clustered using Markov Chain Linkage clustering with an inflation value of 1.1 (van Dongen, 2000; Enright et al., 2002)(Table A.6).

2.4 OUTLOOK

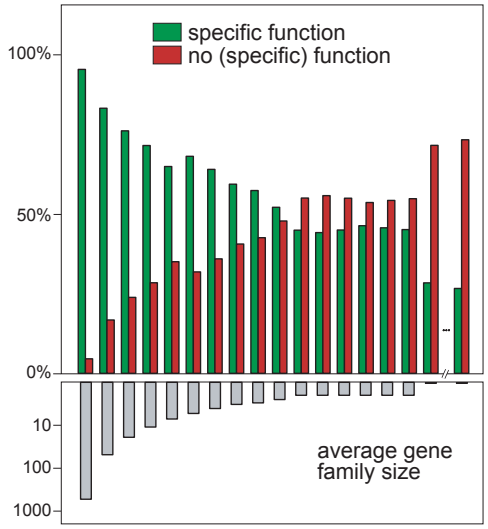


Figure 2.11. Dependence of functional characterization on family size.Colored bars in this histogram of gene families binned by size represent the proportion of families with specific functional annotation (if >20% of the members were classified as such; green) and no specific annotation (a combination of non-specific and no functional annotation; red). Grey bars indicate average gene family size in that bin. Only two out of 174,124 bins containing singletons are shown for clarity. Most large gene families have a known function while many small families remain uncharacterized.

As more environments are explored, we expect that core protein functions (for example, translational machinery) will be seen repeatedly, and will dominate every sample. Novel, rare, and perhaps environment-specific functions, on the other hand, might not be classifiable because they are not yet captured by the experimental studies that underlie most current knowledge about biological function. To reconcile our gene-centric view of the data with a function-based one, we performed an all-against-all similarity search of all predicted ORFs in all four environments, clustered the results into gene families and recorded their functional status according to our operational definition (see Figure 2.11 and Methods). We find that specific functional knowledge is indeed heavily skewed towards large families: functionally characterized families make up 89% of the largest families (200 or more members), while uncharacterized ones make up 72% of the smallest families (three or less members). Thus, although most of the proteins in the environmental samples can be functionally characterized because they belong to well-studied large gene families, numerous distinct, rare functions remain to be identified. As these are likely to be adaptations to specific environmental constraints, they should have the potential for exploitation in biotechnology and medicine. Of all the families (including singletons),

functions can be assigned for only 32%, but this fraction contains 85% of all the proteins studied here. If singletons are disregarded, the fraction of characterizable proteins in the complex environments studied increases further, from 72% to 79%. Although these remain qualitative assignments of low resolution (i.e. substrate specificity or cellular roles are often not specified), even general molecular classifications such as 'dehydrogenase' imply some basic functional understanding and more than a quarter of these are further complemented by associations to other genes predicted by the neighborhood method.

Despite this remarkably high coverage, our functional knowledge about the proteins on earth can be further increased by deeper sequencing that generates longer assemblies and less fragmented ORFs. This should improve gene predictions and reduce the number of uncharacterized singletons that are skewed towards short ORFs. Moreover, longer contigs would allow the application of indirect neighborhood methods (that is, operon membership) increasing the functional context available for each gene. This context can be further increased by using methods to place these contigs into phylogenetic bins, which can give some clues to the partioning of functions among organisms. Such methods, albeit applied to a simpler system than the metagenomic samples described here, uncovered the metabolic interactions underlying the symbiosis between the gutless worm *Olavius algarvensis* and its four bacterial endosymbionts (Woyke et al., 2006).

This huge potential in functionally characterizing the vast majority of proteins in current and upcoming complex samples calls for strategies to capture functional novelty, for example by experimental procedures that enrich in those many small and rare families of unknown functions, analogous to normalizations of EST libraries introduced in the early '90s (Venter et al., 2004). Coupled with systematic biochemical screens, a census of the repertoire of protein functions on earth (at least at the low level of resolution currently used in sequence annotation) might thus be feasible in the very near future.

3 EUKARYOTIC TRANSCRIPT DIVERSITY

Over the various scales of the biological sciences, from the study of single molecules to whole ecosystems, the unifying theme is the understanding of biological complexity. Despite this unity of purpose, it has proved very difficult to connect the biological complexity seen at these different scales due, in part, to the difficulty in defining and measuring biological complexity (Adami, 2002). Physical definitions of complexity tend to emphasise the dynamic aspects of complex systems, defining complex behaviour as somewhere between periodic and random. Biological definitions, on the other hand, have so far tended to focus on the structure of a system, simply put, complex systems have more components and more interactions between them.

These differing approaches to complexity are as much due to the availability of data as to any properties intrinsic to the systems themselves. The traditional approach to studying a complex system in molecular biology was to decompose it into its constituent components, study each individually and finally combine the results into a coherent model. This approach has had some notable successes, such as the lambda phage(Herskowitz and Hagen, 1980), however tended to stress the importance of the components (usually genes) at the expense of the interactions. The limitations of this approach became apparent with the publication of the human genome, with some expressing surprise at the low gene count in humans relative to *Drosophila melanogaster* and *Caenorhabditis elegans* (Claverie, 2001).

One of the explanations offered was that gene products and not genes themselves were the important determinants of biological complexity. Therefore alternative splicing, the mechanism by which a single gene can generate multiple products, was proposed to be an important contributor to the complexity of eukaryotes. This was an attractive proposition in light of the discovery of the extraordinary transcript diversity of the *Dscam* gene in *Drosophila*, which encodes over 38,000 different isoforms (Schmucker et al., 2000). However, the first study to assess this proposition at a global level found no major difference between the levels of alternative splicing between organisms of different complexities (Brett et al., 2002). Although this finding has been disputed (discussed below), there doesn't seem to be a simple relationship between the amount of alternative splicing and organismal complexity, suggesting that the total amount of alternative splicing isn't the major determinant of complexity. Perhaps to look for such a simple realtionship is to repeat the mistake of emphasising the components at expense of the dynamic interactions between them (Lareau et al., 2004). Indeed there is growing evidence that the importance of the transcript diversity generated by alternative splicing can only be understood in the context of the regulatory potential it provides.

In addition to its role in creating and regulating transcript diversity over the lifetime of an organism, there is a growing appreciation of its role in facilitating the evolution of biological complexity (Brett et al.,

2002; Kan et al., 2002; Modrek and Lee, 2003). By providing a nearly neutral path to the evolution of novel biological functions, alternative splicing is thought to play a similar role to gene duplication in the evolution of complexity (Kopelman et al., 2005). In fact it is now thought that such neutrally evolving characteristics might be behind much of the biological complexity we see in eukaryotes (Lynch and Richardson, 2002).

In Section 3.1 I will review the contribution of alternative splicing to the complexity of an organism both in terms of the transcript diversity it generates and the potential for regulatory complexity it provides. I will also look at alternative splicing in an evolutionary context, assessing its impact on the evolution of functional complexity. In Section 3.2 I will present a tool I have developed for the detection and visualisation of alternative splicing and in Section 3.3 I apply this tool to examine the conservation of alternative splicing across metazoans.

3.1 THE CONTRIBUTION OF ALTERNATIVE SPLICING TO BIOLOGICAL COMPLEXITY

3.1.1 *Alternative Splicing and Regulatory Complexity*

One of the earliest puzzles for relatively new field of gene expression in the 1970's was the fact that mRNAs in the nucleus of vertebrates were much longer than their counterparts in the cytoplasm. This was resolved when the sequence of the cytoplasmic mRNAs were compared to the corresponding genomic sequence, revealing that parts of the sequence, later called introns, had been removed (Berget et al., 1977; Chow et al., 1977; Sharp, 2005). Subsequently it was found that the process responsible, called splicing, could remove different introns from the transcript, allowing a single gene to encode multiple products. A summary of the basic patterns of alternative splicing is given in Figure 3.12. These may be combined into higher order patterns such as mutually exclusive exons, where only one of a set of neighbouring skipped exons is included in a transcript.

The splicing reaction is remarkable for the accuracy with which it determines the correct splice sites, even though they can be transcribed several hours apart and separated by hundreds of kilobases. It is even more remarkable that such a mechanism can maintain enough flexibility to allow splicing at alternative sites (Query and Konarska, 2006). The importance of maintaining splicing regulation is evident from the high proportion of hereditary diseases that are caused by mutations near splice sites (Krawczak et al., 1992; López-Bigas et al., 2005) and from a recent study showing that overexpression of the splicing factor SF2/ASF can lead to oncogenesis (Karni et al., 2007). The macromolecular complex responsible for maintaining fidelity and regulating alternative splicing is called the spliceosome. In humans it is composed of approximately 200 different proteins including both core components, responsible for the biochemical reactions of intron excision, and regulatory factors which maintain fidelity and mediate alternative splicing (Jurica and Moore, 2003; Nilsen, 2003).

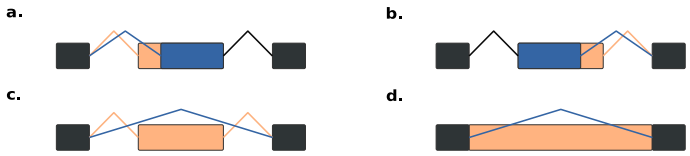


Figure 3.12. Classification of alternative splicing events. While alternative splicing in essence is the skipping of either a 5′or 3′ splice site, it more useful to classify alternative splicing events based on the effect they have on the exonic and intronic sequences that make up the mature transcript. In each case the constitutive exons are coloured grey, the exon and introns in the longer of the two isoforms is colored orange, and the shorter one blue: a) alternative 3′ splice site, where an exon contains two different 3′ splice site, only one of which is used, b) alternative 5′ splice site, c) skipped exon, where both the 3′ and 5′ splice sites of an exon are skipped, causing it to be removed from the transcript during the splicing reaction (sometimes referred to as a cassette or cryptic exon), and d) a retained intron, where both the splice sites of an intron are skipped, causing it to remain in the mature transcript.

The Splicing Reaction

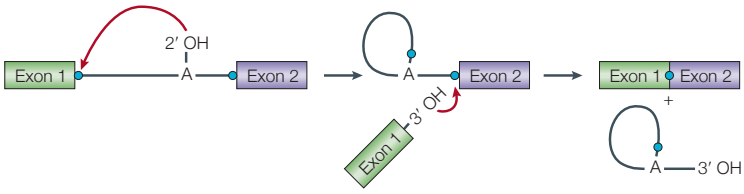


Figure 3.13. Intron removal is achieved by two *trans*-esterification reactions. This figure was taken from (Patel and Steitz, 2003)

The basic biochemical mechanism responsible for the splicing of introns, shown in Figure 3.13, is achieved by two *trans*-esterification reactions involving three sequence elements in the transcript: the 5′ splice site (5′ss), the branch point sequence (BPS) and the 3′ splice site (3′ss). In the first *trans*-esterification reaction the phosphodiester bond at the 5′ss is cleaved by a nucleophilic attack by the 2′ hydroxyl group of a conserved adenine in the BPS. In the second reaction the resulting 3′ hydroxyl group of the upstream exon attacks the 3′ss, ligating the two exons and releasing the lariat intron. It is possible to carry out this type of reaction using only ribozymes, indeed the group I and II introns found in some prokaryotes and eukaryotes are capable of catalysing their own excision. However, for the majority of eukaryotic introns, their removal is catalysed and regulated by the spliceosome.

Many eukaryotes utilize two different types of spliceosome, both of which consist of five small nuclear ribonucleoprotein particles (snRNPs) and a host of protein factors. The major spliceosome, found in all

eukaryotes (Collins and Penny, 2005) and responsible for splicing over 99% of all introns in higher eukaryotes (Sheth et al., 2006), contains the snRNPs U1, U2, U4, U5 and U6. On the other hand, the minor spliceosome is found in some but not all eukaryotes and contains the U5, U11, U12, U4atac and U6atac snRNPs. The four snRNPs that differ between the two spliceosomes are functionally analogous, and the presence of shared components between the two spliceosomes suggests an ancient divergence (Burge et al., 1998). This was further supported by the discovery of minor spliceosomal components in eukaryotes as diverse as protists and fungi (Russell et al., 2006). As the major spliceosome is dependent on U2 for its activity, the introns it splices are called U2 introns. Similarly the targets of the minor spliceosome are called U12 introns. Both types of introns may be distinguished by their differing concensus sequences at the 5′ and 3′ splice sites. Additionally U2 introns contain another conserved sequence element between the BPS and 3′ss called the polypyrimidine tract (PPT). Further discussion will be limited to the major spliceosome unless stated otherwise.

The catalytic core of the major spliceosome is composed of the five snRNPs described, each of which consists of an snRNA associated with members of the Sm and Lsm protein families (Barbosa-Morais et al., 2006). The assembly of the core components of the major spliceosome is depicted in Figure 3.14. It begins with the binding of U1 snRNP to the 5′ss, U2 auxiliary factor (U2AF) to the PPT and the 3′ss and splicing factor 1 (SF1) to the BPS. The U2 snRNA then forms a duplex with with the BPS to form the pre-spliceosomal complex (Smith and Valcárcel, 2000; Patel and Steitz, 2003) allowing the activation of the branchpoint adenosine residue and its subsequent nucleophilic attack of the 5′ss. Next a complex containing the U5 snRNP and the base paired U4-U6 snRNPs joins the pre-spliceosome, which undergoes a conformational change to form the mature spliceosome. This conformational change brings the 5′ and 3′ splice sites into juxtaposition forming the catalytic core for the second reaction. After the *trans*-esterification reactions are complete the intron lariat is released and most of the spliceosomal components dissociate and are reused. Some of the spliceosomal components remain on the transcript in an exon junction complex (EJC) which forms a link between splicing and the nonsense mediated decay (NMD) pathway (Lejeune and Maquat, 2005).

Regulation by interaction of cis and trans-acting factors

In the eukaryotic species with spliceosomal introns but without much alternative splicing, such as *Saccharomyces cerevisiae*, this description captures most of the basic features of the splicing reaction. However, in organisms with alternative splicing the picture is far more complex with numerous factors affecting the choice of splice site (Matlin et al., 2005). These factors, acting in both *cis* and *trans*, often work antagonistically and a delicate balance in their ratios and activities determines which splice site is chosen. This balance is struck within the spliceosome, affecting where it binds to the nascent transcript and possibly even the kinetics of the subsequent reactions (Ares, 2007). In practice there is no clear division between the factors involved in constitutive splicing and those involved in alternative splicing, however some sequence elements

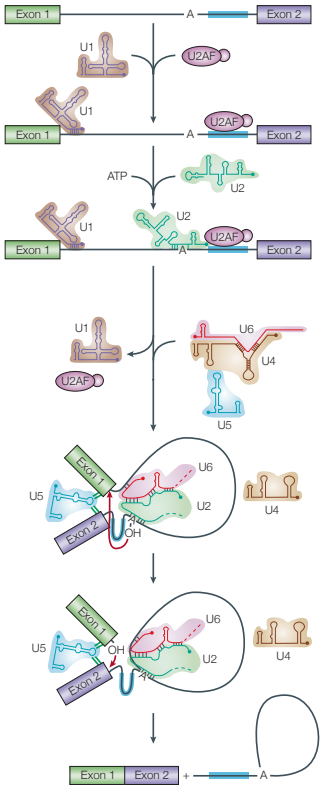


Figure 3.14. Removal of U2 introns by the major spliceosome. A detailed description can be found in the text. This figure was taken from (Patel and Steitz, 2003)

and protein families have been demonstrated to have important roles in the regulation of alternative splicing (Fairbrother et al., 2002; Jurica and Moore, 2003).

The *cis*-acting factors that mediate alternative splicing are classified according to their position within the transcript and the effect that they have on splicing: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs) (Figure 3.15) (Black, 2003; Matlin et al., 2005). It should be noted however that the same sequence element can have different effects on splicing depending on its position within the transcript (Goren et al., 2006; Ule et al., 2006). Clues to the potential mechanisms behind these functions can be gleaned from their distribution in the genome. A systematic study to identify the elements involved in constitutive splicing showed that there is a higher density of silencer elements in introns and pseudoexons (intronic sequences that contain both 5′ and 3′ splice sites, but are never incorporated into a full transcript), than real exons (Zhang and Chasin, 2004). Similarly there is a higher density of enhancer elements in exons than in introns, with no difference between the densities in real and pseudo exons. It is believed that the ratio between the density of enhancers and silencers may play a

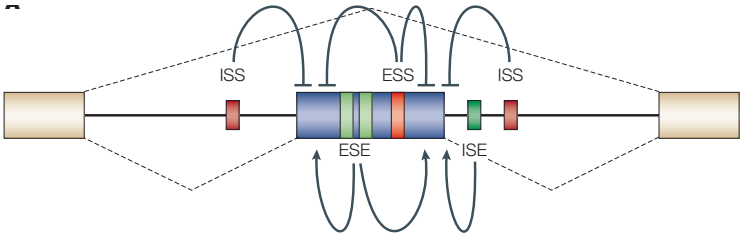


Figure 3.15. Splicing enhancers and silencers maybe located both in introns and exons, indeed the same sequence element can function as both and enhancer and silencer depending on the location (Ule et al., 2006). Figure taken from (Matlin et al., 2005)

role in regulating alternative splicing, for instance it has been shown that ratio between silencer and enhancer density is much higher in alternatively spliced exons compared to constitutive ones (Zhang and Chasin, 2004; Wang et al., 2004). Similarly silencer elements are more frequently found in the intervening sequence between two alternative 5′ or 3′ splice sites (Wang et al., 2006; Yeo et al., 2007).

The identities and densities of these *cis* elements along the transcript form a 'splicing code' which is interpreted by different components of the splicesome, influencing the choice of alternative splice sites (Matlin et al., 2005). The specificity of this code is achieved by regulating the levels and activities of the *trans*-acting factors that recognise the this code. Many ESEs are recognised by members of the serine-rich (SR) family of proteins, while silencers are thought to mostly interact with members of the heterogenous nuclear ribonucleoprotein (hnRNP) protein family (Stadler et al., 2006). There are nine families of SR proteins in metazoans, each consisting of an N-terminal RNA recognition motifs, responsible for interaction with the transcript, and a C-terminal RS domain made up of repeated serine-arginine dipeptides, which mediate interactions with both proteins and RNA (Smith and Valcárcel, 2000; Matlin et al., 2005; Barbosa-Morais et al., 2006). Although SR proteins are thought to have multiple functions in the spliceosome, recent results suggest that the RS domain enhances splicing by promoting the formation of double-stranded RNA between snRNAs and suboptimal splice sites (Izquierdo and Valcárcel, 2006). The levels of SR proteins are partly controlled by alternative splicing coupled to NMD (discussed below) (Lareau et al., 2007), while their activity and localization are modulated by reversible phosporylation (Matlin et al., 2005). The hnRNP proteins, on the other hand, are defined by their association to unspliced mRNA precursors (hnRNAs) and can be classified into 13 diverse families. Recently it was shown that one of the hnRNP proteins, PTBP1 (hnRNP I), is repressed by a micro-RNA (miRNA) in neuronal tissue (Makeyev et al., 2007). This in turn releases NMD-mediated repression of its paralog PTBP2, which in turn regulates neuron-specific splice patterns.

In addition to these major groups there are a host of other *trans*-acting factors that are involved in developmental and tissue-specific alternative splicing. One such splicing regulator is the *Sex-lethal* (*Sxl*) gene, which functions as the master switch in the *Drosophila* sex de-

termination pathway (Black, 2003; Penalva and Sánchez, 2003). This switch consists of a positive feedback loop, regulated at the level of alternative splicing, that is active throughout the life of females but not males. The constitutive transcript isoform of Sxl contains an exon with a PTC and is thus subject to NMD, however the Sxl protein can prevent the inclusion of this exon thereby creating a positive feedback loop. This loop is initiated in female embryos by embryonic transcription from an alternative promoter coupled to skipping of the PTC exon. The resulting levels of *Sxl* protein are sufficient to maintain skipping of the PTC exon when transcription initiation shifts to the promoter used in both male and female adult flies. This skipping is mediated by multiple Sxl-binding sites flanking the PTC exon, which promote cooperative binding of the Sxl along the exon preventing recognition of the exon by the spliceosome. The female-specific expression of *Sxl* provides the basis for a cascade of regulatory events, many of which operate at the level of splicing, to direct sex-specific morphology and behaviour. One of the direct targets of *Sxl* is the *Transformer* (*Tra*) gene which itself encodes a member of the SR family of splicing regulators. The action of Sxl on Tra is more typical of splicing repressor than the mechanism used in its autoregulation. In this case, Sxl has a binding site in the 3′ splice site of a *Tra* exon. In females binding of Sxl to this element prevents splicing at this site, thereby activating an alternative 3′ splice site. This results in the skipping of a termination codon, thus allowing production of an active protein. One of the most dramatic examples of the phenotypic consequences of misregulation of alternative splicing comes from a target of *Tra*, the *fruitless* (*fru*) gene. When male-specific isoforms of this gene are expressed in females it is sufficient to produce male behaviour, such as courtship towards other females (Demir and Dickson, 2005) and aggression (Vrontou et al., 2006).

The switch-like nature of the *Sxl* pathway and it persistence for the lifetime of the organism make it one of the more simple systems regulated by alternative splicing characterized so far. Often alternative splicing is regulated in response to a physiological stimulus (Stamm, 2002) or in a tissue-specific manner (Smith and Valcárcel, 2000), both of which are evident in recent studies of alternative splicing in neurons. Tissue-specific regulation of an entire functional module has been demonstrated in neurons using a large-scale microarray approach (Ule et al., 2005). It was found that the alternative splicing of 40 genes with synaptic functions is under the control of the Nova proteins, a pair splicing regulators expressed only in the central nervous system. These regulators can either promote or inhibit exon inclusion depending on where they bind to the transcript, even exhibiting opposite effects on different exons within the same gene (Ule et al., 2006). An additional level of complexity is added to this system by the fact that many neurotransmitter receptors and ion channels are also regulated by alternative splicing in response to neuron excitation (Lipscombe, 2005). One such channel is the NMDA receptor 1 (NMDA-R1), the splicing of which is not only under the tissue-specific regulation of the Nova proteins, but is also alternatively spliced in response neuron depolarisation (Ares, 2007). This regulation is achieved by an as yet unknown pathway, yet seems to rely on at least two different splicing regulators (An and Grabowski, 2007; Lee et al., 2007). Given that both NMDA-R1 and Nova-2 have

been implicated in long-term potentiation (LTP) (Huang et al., 2005), a physiological change behind memory and learning, it is possible that regulated alternative splicing may contribute to these higher brain functions (Ares, 2007).

Regulation by RNA secondary structure

While most of the research into the regulation of alternative splicing has so far focused on the interactions between the transcript and regulatory proteins, there is a growing appreciation for the role of interactions between RNA elements within the transcript (reviewed in Buratti and Baralle (2004)). For instance, it has been shown that RNA secondary structure influences the binding of several splicing regulators, including members of the SR protein family. Moreover, it has been shown that some silencer and enhancer elements are incorporated in secondary structures, the stability of which can determine the activity of the element. Comparative studies have determined that the level of constraint acting on exons, especially alternatively spliced exons, is greater than can be explained by a combination of amino acid conservation, silencer/enhancer density and codon usage bias, suggesting that secondary structure might have a general role to play in the regulation of splicing (Xing and Lee, 2006). Recently, two striking mechanisms of regulation by RNA secondary structure were discovered in *Drosophila melanogaster* and *Neurospora crassa*.

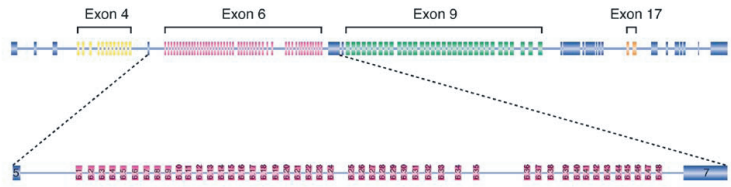


Figure 3.16. *DSCAM* contains four clusters of mutually exclusive exons. Exon clusters 4,6 and 9 each contribute to different immunoglobulin domains whereas exon cluster 17 codes for two alternative transmembrane domains. Taken from (Graveley, 2005).

The current record holder for the gene with the highest level of transcript diversity is the *DSCAM* gene in *Drosophila melanogaster* with approximately 38,000 isoforms, more than double the number of genes in the *Drosophila* genome. This diversity is essential for the complex neuronal patterning connecting adult fly bristles to the central nervous system (Chen et al., 2006), possibly through the different homophilic binding specificities of the isoforms (Wojtowicz et al., 2004). Even a two-fold reduction in the number of isoforms leads to defects in patterning. Perhaps the most remarkable aspect *DSCAM* transcript diversity is that it is also important in innate immunity. *DSCAM* is also expressed the *Drosophila* hemolymph, where it functions as a pattern recognition receptor (PRR) in the phagocytosis of microbes (Watson et al., 2005). Studies on the *Anopheles gambiae* homolog *AgDscam* showed that different isoforms are produced in response to different bacteria (Dong

et al., 2006). The recent evidence for an adaptive immune response in insects (Pham et al., 2007; Sadd and Schmid-Hempel, 2006) lead Sadd and Schmid-Hempel to suggest that *DSCAM* diversity might be responsible. Given that such a memory would have to be maintained epigenetically it is tempting to speculate that a mechanism analogous to the *Sxl* switch is at work.

The functional complexity of *DSCAM* is dependent on the complex alternative splicing that it undergoes, whereby mutually exclusive splicing of 95 alternative exons influences three of the ten immunoglobulin domains in the protein as well as a transmembrane domain (Graveley, 2005). These mutually exclusive events occur in four clusters where a single exon is incorporated from clusters of 12 (exon 4), 48 (exon 6), 33 (exon 9) and 2 (exon 17) variable exons (Figure 3.16). Using a comparative approach, Graveley found conserved sequence elements in the introns upstream of the exons in exon cluster 6 (Graveley, 2005). In the intron between the constitutive exon 5 and the first alternative exon in cluster 6, a 66 nucleotide element, called the docking sequence, was found to be highly conserved across 10 species of *Drosophila*. In the intron upstream of each of the other exons in cluster 6 a sequence complementary to part of the docking sequence, called the selector sequence, was found. Each selector sequences matches a different, yet overlapping, part of the docking site, meaning that only one selector element can interact with the docking element at a time. This suggests that alternative splicing of exon 6, and the maintenance of mutual exclusivity, is dependent on the formation of alternative RNA secondary structures. A similar, but mechanistically different, secondary strucutre is thought to be behind the alternative splicing of exon cluster 4 (Kreahling and Graveley, 2005).

Another example of regulation by secondary structure, albeit far less complex than *DSCAM*, was recently discovered in the fungus *Neurospora crassa* (Cheah et al., 2007). It had previously been shown that some eukaryotic genes contained conserved elements similar to bacterial aptamers that bind thiamine pyrophosphate (TPP) (Blencowe and Khanna, 2007). An aptamer is an RNA domain that can be used to sense the levels of small molecules such as metabolites. These aptamers are often found in riboswitches, where binding of the small molecule to the aptamer results in a conformational change, thus influencing the transcription or translation of the gene containing the riboswitch (Winkler and Breaker, 2005). These riboswitches can be used in feedback loops, where the gene regulated by the riboswitch is responsible for the synthesis of the metabolite. Cheah *et al.* found that two of the three genes in *Neurospora crassa* that contained these putative TPP aptamers were involved in the synthesis of thiamine (the other has no known function), and therefore could possibly be part of such a feedback loop. Moreover all of these aptamers were located in the introns of these genes, which upon addition of TPP underwent changes in their splicing patterns. Subsequent analysis of the *NMT1* gene revealed that, at low levels of TPP, the aptamer could act to block the recognition of an alternative 5′ splice site by the spliceosome producing a functional enzyme. However, at high levels, the conformational change induced by binding to TPP prevented the aptamer from blocking the recognition of the alternative 5′ splice site, resulting in a non-functional enzyme

(Figure 3.17). The intriguing aspect of this system is its simplicity, it doesn't require any regulatory proteins other than the spliceosome and yet can directly sense the levels of the metabolite. When we consider the widespread nature of riboswitches in bacteria (Winkler and Breaker, 2005), the likely bacterial origin of eukaryotic introns, and the fact that the spliceosome was most likely present in the eukaryotic ancestor (Collins and Penny, 2005), then it is possible that such systems may have provided some of the adaptive impetus behind the rapid expansion of introns during early eukaryotic evolution (Koonin, 2006).

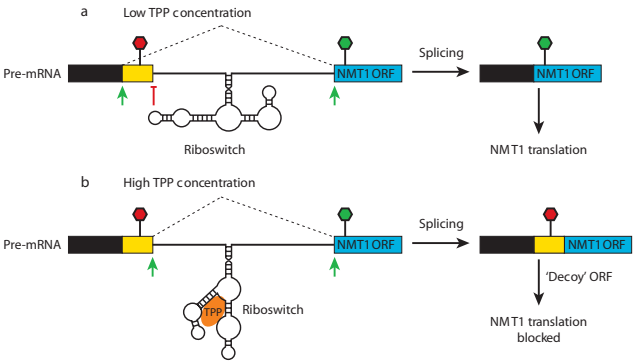


Figure 3.17. A riboswitch regulates alternative splicing in *Neurospora crassa*. The riboswitch is part of a feedback mechanism where the levels of the metabolite thiamine pyrophosphate determine (TPP) the levels of functional *NMT1* protein produced. *a* at low levels of TPP a distal splice site is repressed by the riboswitch, producing a shorter mRNA product which codes for a functional *NMT1* product. *b* at higher levels of TPP the aptamer of the riboswitch binds to the metabolite, releasing repression on the distal splice site. The resulting product, while longer at the level of mRNA, produces a non-functional truncated product. Taken from (Blencowe and Khanna, 2007)

Interaction with other regulatory mechanisms

The previous two sections showed how various developmental, tissue-specific and physiological signals can directly influence the action of the splicesome to bring about changes in alternative splicing. However if we consider the splicing reaction in the greater context of gene expression, then we see that it is at the center of a complex network of regulatory modules involving transcription, transcript processing, export from the nucleus and transcript degradation (Figure 3.18) (Maniatis and Tasic, 2002; Moore, 2005). Some of these regulatory events, such as alternative promoter usage and alternative polyadenylation, act in concert with alternative splicing to produce a combinatorial increase in the potential transcript diversity. In addition to this, the interconnected nature of this network allows regulatory signals from many different modules to be integrated by the spliceosome, which in turn can provide input to many other modules.

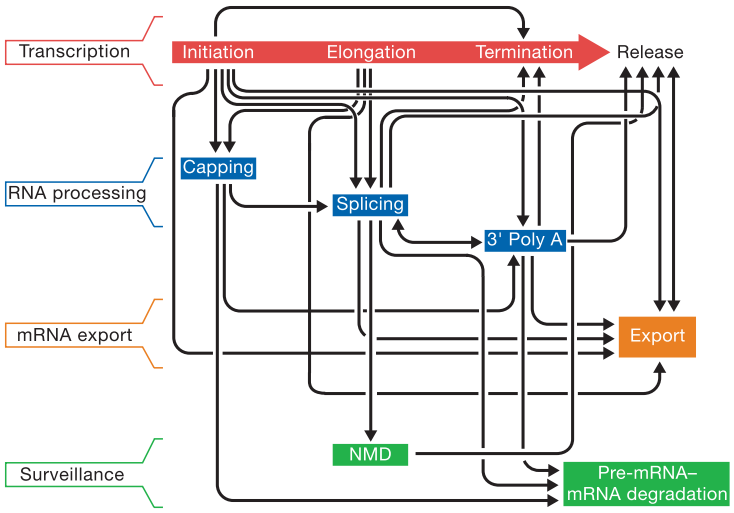


Figure 3.18. The splicing reaction is central to the regulation of gene expression. Taken from (Maniatis and Reed, 2002)

One of the earliest steps in gene expression that can influence the regulation of alternative splicing is transcription itself. There are two possible models for the regulatory link between transcription and alternative splicing: *(i)* the kinetic model and *(ii)* the recruitment model. Under the kinetic model of regulation, a slow rate of transcript elongation can promote the recognition of an upstream weak splice site over a downstream strong strong splice site (Figure 3.19) (Kornblihtt, 2005, 2006). This model is supported by several strands of evidence (reviewed in Kornblihtt (2005)), including a possible role for chromatin remodelling factors in the regulation of alternative splicing (Batsché et al., 2006). In this case it was shown that overexpression of the ATPase subunit of the SWI/SNF chromatin-remodelling factor resulted in a higher level of inclusion of skipped exons in certain genes. This activity was linked to an accumulation of RNA polymerase II (Pol II) at inter-

nal sites within the gene, suggesting that the pausing of transcription favours the inclusion of alternative exons in these genes. This finding is especially interesting in light of the recent discovery that another chromatin-associated protein, DEK, has been shown to regulate splicing at a later step by increasing the fidelity of 3′ splice site recognition (Soares et al., 2006). These studies suggest that perhaps there is a more extensive coupling between chromatin regulation and the downstream steps in gene expression than previously thought (Kress and Guthrie, 2006). The second model for the regulation of splicing by transcription is called the recruitment model (Kornblihtt, 2005; Lynch, 2006). Under this model, the promoter of a gene affects the recruitment of factors to the spliceosome via interactions with the C-terminal domain (CTD) of Pol II, which in turn regulate the choice of splice site. While this model has less support than the kinetic model (Lynch, 2006), there is convincing evidence for its existence. Recent work in the Kornblihtt lab has shown that the CTD can influence the inclusion of an exon in a mechanism that is not dependent on elongation rate (de la Mata and Kornblihtt, 2006). Therefore it seems that both mechanisms may be at work, however for the moment it is difficult to gauge the global levels of each. The large-scale production of chromatin immunoprecipitation (ChIP) data by projects such as ENCODE (Consortium, 2007) may soon allow us to assess the importance of the kinetic model on a global scale (Kornblihtt, 2006), and advances in methods of detection of protein-protein interactions may provide some insight into the interactions responsible for the recruitment model.

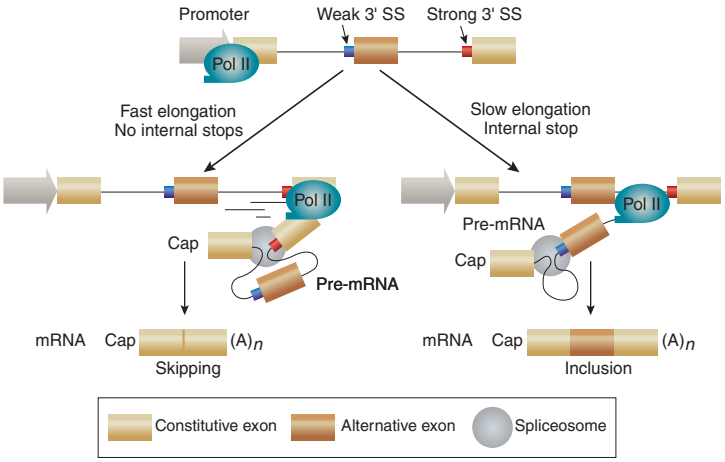


Figure 3.19. The kinetic model of splicing regulation by transcription. In the case where the 3′ splice site of a cassette exon (blue) is weaker than the downstream 3′ splice site (red), low transcription elongation rates will promote inclusion while faster rates will result in skipping. Taken from (Kornblihtt, 2006)

As well as being coupled to upstream events in gene expression, there is evidence of significant coupling to downstream events, especially to RNA surveillance pathways. One example comes from recent

study showed the importance of miRNA-mediated repression of a splicing factor in establishing neuronal-specific splicing patterns (Makeyev et al., 2007). However the most important coupling between alternative splicing and RNA surveillance may involve the NMD pathway. An early study found that 45% of genes that undergo alternative splicing produce at least one isoform that contains a premature termination codon (PTC) and therefore is likely to be degraded by NMD (Lewis et al., 2003). While microarray experiments have shown that coupling between alternative splicing and NMD is unlikely to occur on such a large scale (Pan et al., 2006), recent studies have shown that when it does occur it can have important functional consequences (Lareau et al., 2004). It seems as if the feedback loop utilised by *Sxl*, whereby the splicing regulator regulates its own levels by producing an isoform targeted to the NMD pathway, may be a general regulatory circuit for splicing regulators. Lareau et al. showed that each member of the SR family in humans was subject to NMD, reducing expression between 4 and 40-fold, and in some cases, the SR protein itself is responsible for regulating this event (Lareau et al., 2007). The importance of this mode of regulation was underscored by the fact that the exons containing the PTCs were associated with ultraconserved elements, long stretches of genomic sequence highly conserved between human and mouse. Moreover, the PTC containing exons were in different positions in different genes, suggesting that the same regulatory mechanism arose multiple times, indicative of an optimal regulatory mechanism (Conant and Wagner, 2003). This is further supported by the fact that SR proteins in plants are also regulated in such a manner (Kalyna et al., 2006). In addition the ultraconserved elements in SR proteins, a further six are associated with hnRNP proteins, highlighting the functional importance of alternative splicing regulation (Lareau et al., 2007).

3.1.2 *Alternative Splicing and Transcriptome Complexity*

It is clear from the preceding examples that the regulation of alternative splicing and the resulting transcript diversity can have very important consequences for organismal complexity. However it is not clear how representative these examples are of alternative splicing in general. For example, the mechanism of exon exclusion behind the huge combinatorial transcript diversity of *DSCAM* is probably not present in vertebrates (Graveley, 2005). Therefore to understand the global importance of alternative splicing we must consider both the prevalence of alternative splicing and the proportion of it that is likely to contribute to functional complexity.

The prevalence of alternative splicing

Before the advent of high-throughput sequencing technologies, alternative splicing was very much considered the exception to the one gene, one protein rule. Early estimates suggested that on 5% of human genes were alternatively spliced (Sharp, 1994), however as technology progressed, these estimates began to increase dramatically (Boue et al., 2003). The first systematic estimates were made by aligning ESTs against full-length cDNAs, which suggested that 38% of human genes were

alternatively spliced (Brett et al., 2000). The publication of the human genome (Lander et al., 2001; Venter et al., 2001) allowed the development of a more sensitive method to detect alternative splicing based on the spliced alignment of ESTs and cDNAs against the genome. The first such study concluded that, with the level of EST coverage available at the time, 41% of human genes are alternatively spliced (Modrek et al., 2001). As subsequent study used a similar method to look at the effect of EST coverage on this estimate (Kan et al., 2002). In contrast to the previous estimates they found that 68% of genes were alternatively spliced. However, when a stringent cut-off was applied, requiring that an alternative splicing event is seen in more than 5% of transcripts, this figure dropped to 17–28% of genes. A more telling statistic is that there is evidence for alternative splicing in 99% of genes with more than 700 ESTs, a proportion that drops to 47% when isoforms that occur in at least 1% of transcripts are considered.

These figures serve to highlight the effect that limited coverage of EST libraries can have on the detection of alternative splicing. Not only do ESTs represent a limited sample of the transcriptome, but they are also biased towards the areas of study popular for that given organism. For instance, human EST data is heavily biased towards sex-specific tissues and neuronal tissue (Figure 3.20). On top of this a large proportion of EST libraries are created from cancerous tissue (Figure 3.20), where at least in some cases, deregulation of splicing may be a symptom or causative agent in the pathology (Roy et al., 2005; Karni et al., 2007). A further complication arises from the fact that at least some of the different isoforms produced by a gene may not be due to regulated alternative splicing, but rather allele-specific transcript isoforms (Nembaware et al., 2004). These biases and the limited coverage can be partly alleviated by using more systematic approaches such as microarrays (Srinivasan et al., 2005). The first such study looked at the splice junctions of a set of genes in 52 human tissues, and used the validation rate and EST coverage to estimate that 74% of human genes are alternatively spliced (Johnson et al., 2003). Therefore it seems as if the majority of human multi-exon genes are alternatively spliced.

Similar genome-wide methods of detection of alternative splicing have been applied to a number of model organisms, however for the majority of them the levels of EST coverage are far below those of human (Figure 3.21). Given that there is no expression ontology similar to EVOC (Kelso et al., 2003) for these organisms it is difficult to assess potential biases within these datasets. For instance, some EST libraries may come from whole-body extracts in which case tissue-specific isoforms from larger tissues might be overrepresented. The organism with the highest EST coverage after human is the mouse and seems to have a similar level of alternative splicing to that of human. However, mouse has many more full-length cDNA sequences, allowing the delineation of full transcript isoforms (Zavolan et al., 2003). The first analysis of these data showed that 41% of mouse genes were subject to alternative splicing, a figure which increased to 60% when coverage was accounted for. A more recent analysis on a larger dataset detected alternative splicing for about half of the genes studied, but didn't account for coverage (Carninci et al., 2005). A microarray study in *Drosophila melanogaster*

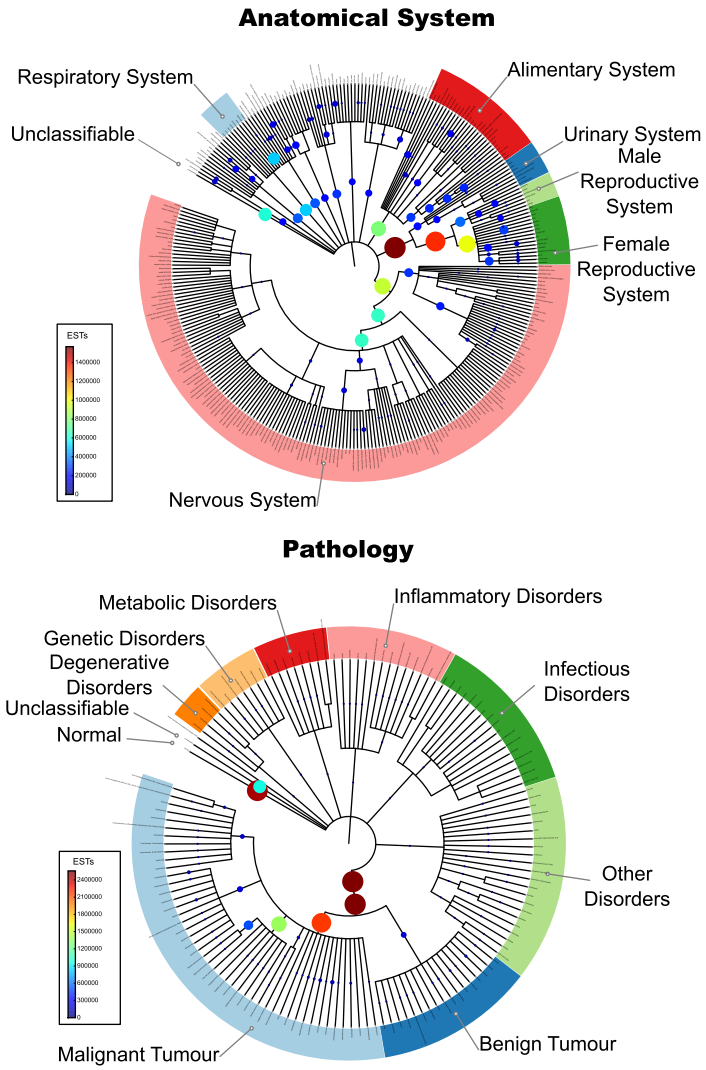


Figure 3.20. Distribtion of ESTs among the EVOC anatomical and pathological terms. The EVOC ontologies (Kelso et al., 2003) were visualised using iTOL and the cumulative distribution of ESTs in the tree is shown as circles along the branches of the tree. The area of each circle is proportional to the number of ESTs that can be mapped to that ontology term and its descendants, a minimum area is set so that terms with very few ESTs can be seen. The number of ESTs that can be mapped to each term is also represented by the color of each circle according to the scale at the bottom right of each tree.

used both exon and exon-junction to detect alternative splicing in a subset of genes (Stolc et al., 2004). They found that 53% of genes undergo exon skipping, however by comparing to EST data they estimated a false negative rate of 46%, suggesting that the true amount of alternative splicing is much higher. It is difficult to compare this study, which looked at expression across developmental stages, to that of Johnson *et al.* who looked at differential expression across tissues. The estimates for plants are much lower, approximately 20% in *Arapdidopsis thaliana* and *Oryza sativa*, however no estimate on the effect of coverage was made (Wang and Brendel, 2006). There are a handful of intron retention events in *Saccharomyces cerevisiae*, that are not regulated in the same way as metazoans, therefore it was thought that alternative splicing is mostly absent in fungi. However this assumption was overturned by the sequencing of the *Cryptococcus neoformans* genome, where it was found that 4% of genes are alternatively spliced. This was reinforced by the discovery of several introns in *Neurospora crassa* that undergo alternative splicing (Cheah et al., 2007) (described above). It was also noted that the only intron identified so far in the early branching eukaryote *Giardia lamblia* was removed in only a proportion of transcripts (Nixon et al., 2002; Johnson, 2002). Taken together these results suggest that alternative splicing might have already been present in some form in the last common ancestor of eukaryotes, and subsequently lost in some lineages.

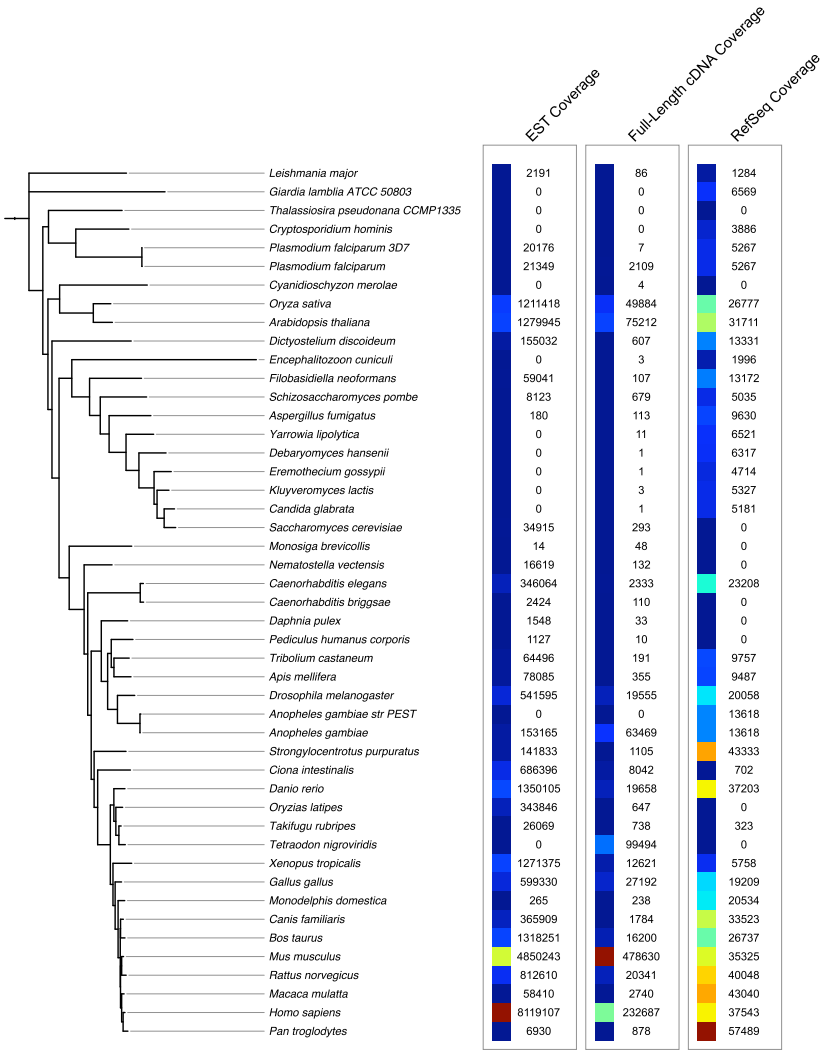


Figure 3.21. Coverage of eukaryotic species by EST, cDNA and gene prediction data

The functional impact of alternative splicing

Both the prevalence and levels of alternative splicing observed by these genome-wide studies were somewhat surprising, as traditional gene-oriented studies, with some notable exceptions, did not hint at such levels of diversity. One possible explanation is that a large amount of alternative transcripts are merely biological noise and have no functional impact on the organism (Kan et al., 2002). The first studies to test this were based on the assumption that if alternative splicing makes a significant contribution to the functional complexity of an organism, then we should expect to see higher amounts of alternative splicing in more complex organisms.

Testing this is far from straightforward for several reasons: *(i)* there is no quantitative measure of organismal complexity (Adami, 2002), *(ii)* the EST coverage for different organisms varies widely (Figure 3.21), *(iii)* different organisms will have a different distribution of developmental states, tissues and pathologies sampled by ESTs, *(iv)* the coverage of allele-specific splicing depends on the diversity of genetic backgrounds sampled by EST libraries, which will not be the same across species. To address the first limitation researchers tend to use an intuitive measure of biological complexity, limiting comparisons to broad categories such as vertebrates and invertebrates. The second can be addressed by using resampling to normalise the number of ESTs used to detect alternative splicing across species (Brett et al., 2002). The final two drawbacks cannot be resolved currently, although ESTs from human cancerous tissues can be removed using expression ontologies such as EVOC (Kelso et al., 2003).

The first study to look at the association between alternative splicing and organismal complexity used alignments of ESTs against cDNAs to detect alternative splicing (Brett et al., 2002). They found that when the ratio of ESTs to cDNAs was normalised across organisms, the proportion of genes that were alternatively spliced in each organism was comparable. This prompted several follow-up studies, the first of which concluded that there was more alternative splicing in vertebrates (Kim et al., 2004), but was found to be methodologically flawed (Harrington et al., 2004). Nagasaki et al. used alignments of full-length cDNAs against the genomes of 6 organisms and carried out a similar subsampling method as *Brett et al.* (Nagasaki et al., 2005). Although the results of this analysis are remarkably similar to those of *Brett et al.*, including a higher proportion of alternative splicing in *Drosophila melanogaster* than mouse at similar levels of coverage, they conclude that there is indeed a link between the level of alternative splicing and organismal complexity. The most recent study used EST and cDNA alignments against the genome to detect alternative splicing in 8 organisms (Kim et al., 2007). They employed a different resampling strategy, whereby genes with more that 10 ESTs were selected from each organism and then alternative splicing was repeatedly detected for each of these genes using only 10 ESTs each time.The results of this analysis are at odds with the previous studies and shows a higher proportion of alternatively spliced genes in vertebrates than invertebrates. This may be a result of the sampling strategy, where only the most highly expressed genes in an organism with poor coverage (those with >10 ESTs)

are compared to a group of genes containing both highly and lowly expressed genes of an organism with high coverage. At any rate the results don't conclusively support a link between the proportion of genes that are alternatively spliced and organismal complexity as it seems that the smallest difference between vertebrates and invertebrates (the proportion of alternatively spliced genes in mouse is 2-fold higher than *Drosophila*) isn't much greater than the greatest difference within vertebrates (1.3-fold higher in chicken than mouse).

The different methods of detecting alternative splicing, datasets and resampling strategies used in these studies make it difficult to resolve the disagreements between them. However the biggest hurdle to reaching a resolution is the fact that these studies aim to find a correlation between alternative splicing and an intuitive measure of complexity. Even if such a correlation could be found its explanatory power would be minimal, as the existence of an alternative splicing event does not necessarily mean that it is functional. Among the alternative splicing events detected above there are three sub-populations under different evolutionary pressures: *(i)* those under negative selection and therefore likely to be conserved between species, *(ii)* those under positive selection, which differ between species but show little allelic variation within a species, and *(iii)* real alternative events that have little or no functional impact on the organism and are therefore evolving neutrally (Khaitovich et al., 2006a). Without a neutral model of alternative splicing evolution and the difficulty in associating allelic variation to alternative splice events (Nembaware et al., 2004) it is difficult to distinguish between the latter two classes. However, for one class of alternative splicing events, alternative 5′ or 3′ splice sites that lead to small length variations, it has been shown that the choice of alternative splice site is influenced by the relative strengths of the alternative splice sites (Chern et al., 2006). The stochastic nature of this alternative splicing, along with the lack of evolutionary constraint, suggests that these events belong to the neutrally evolving class. However much of the focus, and indeed success has been in detecting alternative splicing events from the first class. Strictly speaking an alternative splicing event can be considered conserved when it is seen at the level of the transcriptome in both species. However, due to the relatively low coverage of the transcriptome in many species, transcriptomic evidence of the event in one species and genomic conservation of the exons and introns necessary for the alternative splicing event in the other species is often considered sufficient evidence for conservation of the alternative splicing event itself. This difference in stringency, along with the variety of methods and datasets used again make it difficult to compare results between studies (Lareau et al., 2004).

Most of the estimates of conservation have so far been based on comparisons of human and mouse and the first was carried out by Kan et al. who looked at the presence of human alternative exon junctions in mouse ESTs (Kan et al., 2002), and found that only 7% of alternative junctions were seen to be conserved. However, they reasoned that the this low number could be due to low EST coverage, when they only considered isoforms with sufficently high EST coverage in mouse this figure rose to 42%. A similar study by (Thanaraj et al., 2003) found that 15% of alternative junctions were conserved which

rose to 61% upon extrapolation using a model of transcript coverage (Thanaraj et al., 2003). Several groups also looked at the conservation of skipped exons, allowing them to address both the genomic and transcriptomic conservation of these events. Resch et al. found that that only 5% of human skipped exons were genomically conserved in mouse and of these 22% were seen in the mouse transcriptome (Resch et al., 2004b). This yields an overall conservation rate for skipped exons of 1%, however it doesn't take into the limited transcript coverage into account. In contrast, Sorek et al. used a much smaller dataset and found genomic conservation for 25% of human skipped exons (Sorek et al., 2004). A more recent study by Yeo et al. found that 5% of human skips that were genomically conserved in mouse were also found in the transcriptome (Yeo et al., 2005). They developed an algorithm to predict the likelihood of an exon to be a conserved skipped exon, and based on these likelihoods and the validation rate of this algorithm using RT-PCR, they estimated that only 11% of human skipped exons that are conserved in the mouse genome are likely to conserved at the level of the transcriptome. It's hard to definitively pin down why these estimates differ so much. It may be due the different measures of alternative splicing used (alternative junctions vs. skipped exons). Another possibility is that the choice of data can bias the result, for instance it has been shown that a skipped exon is far more likely to be genomically conserved if it is the major isoform than a minor (Modrek and Lee, 2003), so by only using well supported events Kan et al. were likely to bias their analysis toward this set.

Given that most of these studies have only considered human and mouse it's hard to know if the conclusions of these studies are applicable to other species. Wang and Brendel looked at the conservation of alternative splicing between *Arabidopsis* and rice and found that 24% of alternatively spliced genes in *Arabidopsis* had the same type of alternative event in rice (Wang and Brendel, 2006), however not all of these were necessarily conserved at the resolution of individual events. Malko *et al.* took a novel approach to detecting the genomic conservation of alternative events over a larger range of evolutionary distances (Malko et al., 2006). They utilised a spliced-alignment program to create alignments protein sequences from one species against the genome of another species, which were then used to detect conserved alternative splicing between *Drospohila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. They found that 75-80% of alternative segments of a protein were conserved between *melanogaster* and *pseudoobscura*, while only 45% were conserved between *melanogaster* and *Anopheles*. These figures should probably be interpreted with caution as they do not take into account EST evidence and there are technical limitations in doing protein against genome alignments across such large evolutionary distances. Rukov et al. took a smaller-scale approach to look in detail at the conservation of 21 alternative splicing events between *Caenorhabditis elegans* and *Caenorhabditis briggsae*, a comparable divergence time to that of human and mouse. Surprisingly they found a high level of conservation of not only the events themselves (93% at the level of transcript isoforms), but also the expression profiles of these isoforms (Rukov et al., 2007).

While it's hard to conclude the exact level of conservation of alterna-

tive splicing, it is generally accepted that it is considerably lower than seen for whole genes. This is consistent with comparative studies of gene expression which show that there is reduced constraint and possibly even positive selection acting on the regulation of gene expression (Khaitovich et al., 2006a,b). If there is indeed a low level of conservation and assuming that there isn't a large proportion of alternative events under positive selection, then the majority have little or no functional impact. Based on this assumption, by comparing conserved to non-conserved alternative splicing events, it should be possible delineate the properties of functional alternative splicing events.

The first property of functional alternative splicing uncovered by this approach is that it is associated with higher inclusion levels (Modrek and Lee, 2003; Pan et al., 2004), that is the major isoform (usually defined as being in present more than 66% of transcripts from a gene) is more likely to be conserved than a minor form. This is most likely due to the fact that a large proportion of nonconserved alternatively spliced exons in human are recent additions to the genome, and probably haven't acquired strong enough splicing signals to be included above a low level in transcripts. Other analyses have shown that conserved alternative exons and their flanking introns have a higher level of constraint, as measured by sequence conservation (Sugnet et al., 2004) and SNP density (Yeo et al., 2005). This constraint mostly affects the synonymous substitution rate (K_s) and is most likely due to the density of regulatory elements required for proper splicing, including splicing factor binding sites and possibly elements that regulate RNA secondary structure (Xing and Lee, 2006). While minor form exons are less likely to be conserved, those that are conserved are subject to higher levels of constraint. For instance, minor form exons that are conserved between human and chimpanzee have a 25% lower K_s than their major form counterparts, whereas those conserved between human and mouse have a four-fold lower K_s (Xing and Lee, 2005a). This may be partly explained by tissue-specific regulation, where the minor isoform becomes the major form in a single tissue, requiring a higher density of regulatory elements around that exon (Xing and Lee, 2005b). Another general trend among conserved alternatively spliced exons is that they tend to have relatively minor effects on the overall domain composition of proteins, either removing a domain in a modular fashion or affecting the functional residues of the domain (Kriventseva et al., 2003; Yeo et al., 2005). One manifestation of this pressure is the tendancy of skipped exons to be of a length divisible by 3, thus preserving the reading frame of the transcript (Resch et al., 2004a; Magen and Ast, 2005). As with the level of nucleotide conservation, this tendancy is especially marked in conserved minor form exons (Resch et al., 2004a). A similar trend was noted for conserved alternative 3′ and 5′ splice sites (Koren et al., 2007).

3.1.3 *Alternative Splicing and the Evolution of Complexity*

The apparent low rate of conservation of alternative splicing has lead some to suggest that its major contribution to biological complexity is not the direct contribution it makes to the regulatory and transcript complexity of an organism, but rather its role in facilitating the evolu-

tion of complexity (Kirschner and Gerhart, 1998; Brett et al., 2002; Kan et al., 2002; Modrek and Lee, 2003). In this sense it can be considered analogous to gene duplication as a source of functional novelty. However, unlike alternative splicing, where an EST or cDNA is required to confirm the existence of an isoform, gene duplicates can be detected directly from genomic sequences. For this reason, and the fact that the importance of gene duplication was recognised far earlier (Taylor and Raes, 2004), the contribution of gene duplication to the evolution of functional complexity is far better characterized than that of alternative splicing (Hurles, 2004). Moreover, the increasing number of genome sequences has allowed detailed exploration of both the mechanisms by which gene duplicates are generated, and the subsequent evolutionary pressures acting on them (Lynch and Katju, 2004). Given the paucity of such data for alternative splicing it is instructive to use the model of gene duplication to infer the possible roles for alternative splicing in the evolution of complexity.

The birth of new genes and isoforms

The first factor to understand in the evolution of gene duplicates is the mechanism by which they are generated. Gene duplicates can be generated in one of three ways: *(i)* unequal crossing over, *(ii)* replication errors or *(iii)* retrotransposition (Hurles, 2004). In the first case recombination occurs between paralogous sequences rather than homologous sequences, resulting in the tandem duplication of partial or even multiple genes. Other duplications are not associated with paralogous sequences are a thought to be the result of chromosomal breakages during replication. The final class of duplication event, and in human the most common (Suyama et al., 2006), is through the retrotransposition of a transcript into the genome, resulting in an intronless copy of the gene at a different location in the genome.

In contrast, the birth of an alternative transcript isoform is the product of several overalapping mutational events broadly divided into those that affect the structure of the gene and those that affect the *cis*-acting elements responsible for splicing. The maximum number of transcript isoforms possible for a given gene is set by its intron-exon structure. Therefore the loss and gain of introns and exons is an important process in the evolution of novel transcript isoforms. The role of exon gain in the birth of new transcript isoforms has already been highlighted by several studies (Letunic et al., 2002; Modrek and Lee, 2003; Lev-Maor et al., 2003; Crayton et al., 2006; Zhang and Chasin, 2006). Exon birth can be further subdivided into those resulting from exon duplication and those that are derived from non-coding intronic sequences (Kondrashov and Koonin, 2003). Exon duplication can automatically lead to mutually exclusive transcript isoforms if the duplicated exon is flanked by different intron types (U2 and U12) (Letunic et al., 2002). Similarly the huge transcript diversity of the *DSCAM* gene among insects is the result of many exon duplications. On the other hand, exons that are derived from intronic sequences are thought to be the major source of novel alternative isoforms in mammals (Modrek and Lee, 2003; Lev-Maor et al., 2003; Zhang and Chasin, 2006). Often these new exons are associated with repetitive elements, for instance Sorek et al. estimated

that 5% of human alternatively spliced exons contain *Alu* elements (Sorek et al., 2002). Intron birth, on the other hand, may not be a major force in the recent evolution of novel transcript isoforms. For instance, it seems as there has been little or no intron loss or gain during the evolution of mammals and the pattern for insects and fungi is dominated by intron loss (Carmel et al., 2007b) (Figure 3.22). However, there have been several episodes of rampant intron gain during eukaryotic evolution such as the emergence of animals. These episodes, which may have been due the reduced efficiency of selection associated with population bottlenecks, have been associated with increases in complexity (Lynch and Conery, 2003; Lynch, 2007). It is unclear what role, if any, these additional introns had in this increase of complexity, however it is possible that the resulting increase in the potential for alternative splicing may have contributed.

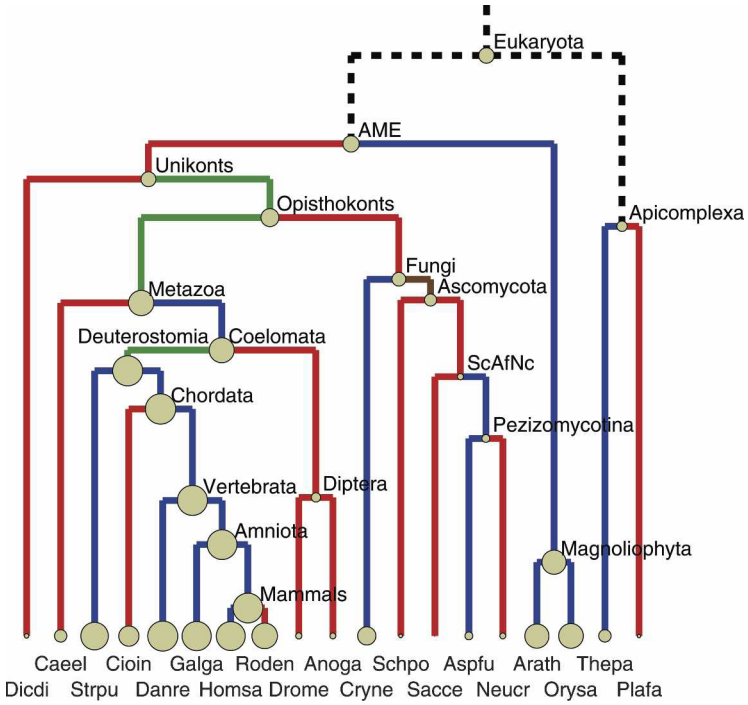


Figure 3.22. Distribution of intron gain and loss rates over the phylogenetic tree of eukaryotes. The size of the circles are proportional to the intron density (inferred for internal nodes) and the branches coloured according to whether the evolution of introns was dominated by gain (green), loss (red) or a balance between loss and gain. Figure taken from (Carmel et al., 2007b)

In addition to, and often in combination with, these relatively large mutational events, novel transcript isoforms may be generated by substitutions and indels in the *cis*-acting elements required for the regulation of alternative splicing. These mutations can act to weaken constitutive splicing signals and strengthen alternative signals, thereby converting a constitutive exon to an alternative one (Izquierdo and Valcárcel, 2006).

They can also act in the opposite direction to activate cryptic splice sites within introns to create whole new exons or alternative 3′ and 5′ splice sites (Koren et al., 2007). For example the high number of *Alu* elements associated with alternatively spliced exons in human is due to the presence of sequence motifs that resemble splice sites within the elements themselves (Lev-Maor et al., 2003). The role of such mutations in creating novel isoforms is likely to differ between organisms, for example it is known that there is considerably more redundancy in splice site sequences in human than in yeast (Ast, 2004), meaning that the probability of creating a novel splice site by random mutation is higher in humans. Similarly it seems that the sequence space occupied by splicing regulatory elements is larger than previously thought, with relatively frequent interconversion between different ESEs (Stadler et al., 2006). This functional redundancy may have been facilitated by the expansion of the SR family in metazoans (Barbosa-Morais et al., 2006).

The fate of novel genes and isoforms

The fate of the newly duplicated gene pair falls into one of three different categories: *(i)* non-functionalisation, *(ii)* sub-functionalisation and *(iii)* neo-functionalisation. In the first case, which is the most frequent, one of the duplicate pairs accumulates a deleterious mutation such as a premature stop codon or silencing mutation within the promoter, abolishing its function and thus creating a pseudogene. The second case occurs when a multifunctional gene is duplicated and the functions are partitioned out among each of the duplicates (Force et al., 1999). In the third scenario, originally proposed by Ohno (Ohno, 1970), one member of the duplicate pair acquires a completely novel function, while the other retains the ancestral function.

The exact role of selection in this process is still under debate (Lynch and Katju, 2004). In the original model by Ohno the duplication event creates exact copies of the gene, one of which retains the ancestral function thus relaxing the constraint on the other gene. This frees the new copy to gradually accumulate point mutations in a neutral manner until it acquires a novel function or is silenced. A variation on this model is that mutational events associated with the duplication event, such as exon shuffling and gene truncation, are responsible for the novel functionality (Katju and Lynch, 2006). In this case there is no period of functional redundancy, and the novel gene is immediately exposed to natural selection. Whereas both of these models stress the role of random mutation in creating new gene functions, the adaptive-conflict model proposed by Piatigorsky and Wistow and Hughes suggest that positive selection might be the major determinant (Piatigorsky and Wistow, 1991; Hughes, 1994). Under this model, the ancestral gene carries out two or more distinct functions, but, due to the resulting pleiotropy, carries out some of these functions suboptimally. After duplication the pleiotropic effects of mutation are reduced, allowing positive selection to partition and optimise the different functions among the gene duplicates.

Novel transcript isoforms share the same fate as gene duplicates but the evolutionary trajectories they follow are likely to be very different. For instance, due to the fact that many novel forms are due to

the exonisation of an intronic sequence, the product of a novel transcript isoform is already likely to be functionally distinct from the pre-existing isoforms of that gene. Therefore the initial period of functional redundancy between gene duplicates, responsible for the near neutral evolution of some gene duplicates, is not likely to be present for most novel alternative isoforms. On the other hand, many of these novel alternative isoforms are expressed at very low levels and are thus likely to be shielded from selective constraint (Modrek and Lee, 2003). It is also possible that the level of selection acting on a novel isoform may also be reduced by the NMD pathway, which prevents the translation of PTC-containing isoforms, although it seems as if few novel isoforms are targeted to this pathway (Pan et al., 2006). This path of near neutral evolution may not be present for alternative isoforms generated by exon duplication. Due to the fact that these exons already contain splicing signals they are likely to be included at a high level and therefore immediately subject to selection. However, in the case where the duplicated exon is spliced mutually exclusively with the ancestral exon, the novel isoform will initially be functionally redundant, relaxing the constraint on one of the two exons.

The differences between the evolutionary trajectories offered by gene duplication and alternative splicing are highlighted by the fact that they are inversely correlated (Kopelman et al., 2005; Su et al., 2006; Talavera et al., 2007). It turns out that the more prone to duplication a gene is, as measured by family size, the less likely it is to to be alternatively spliced. One possible explanation for this might be the adaptive-conflict model. In this case a gene obtains an extra function though the evolution of a novel alternative isoform, this results in pleiotropy, which is relieved by gene duplication followed by sub-functionalisation. Perhaps the most striking example of the pleiotropy associated with alternative splicing comes from the INK4a/ARF locus in mammals, which encodes two overlapping reading frames (Szklarczyk et al., 2007). Each of the proteins generated from this locus regulate different tumour suppression pathways and therefore mutations in either of these reading frames can cause cancer. This overlap, along with a high intrinsic mutation rate, makes this one of the most frequently mutated loci in cancers. While this locus may represent a pre-duplication state, there are some examples in nature where a pair of gene duplicates in one species are represented by alternative isoforms in other species (Pacheco et al., 2004). However a study by Talavera et al. suggests that that gene duplicates are not functionally equivalent to alternative isoforms and therefore not likely to be the result of sub-functionalisation (Talavera et al., 2007). The explanation offered was that genes that are alternatively spliced are more sensitive to dosage effects and therefore less likely to be duplicated (Talavera et al., 2007).

3.2 DETECTING AND VISUALISING ALTERNATIVE SPLICING

3.2.1 *Introduction*

The typical detection protocol for alternative splicing involves *(i)* the alignment of ESTs and cDNAs to the genomic sequences and *(ii)* the

detection of alternative splicing events from these alignments. Despite the similar approach taken by most analyses, they often implement their own alignment and detection algorithms, making it difficult to compare between analyses. Moreover, in most cases the results of the analysis but not software that implements it is available, limiting analysis to the organisms used in the original study. Therefore, in order to be able to study alternative splicing in a range of organisms I created a piece of software to detect and visualise alternative splicing called *Sircah*.

There were several goals in the design of *Sircah*. Firstly, to ensure that it can be widely used it was developed in the Python scripting language which can be run on all common operating systems. Secondly, instead of tying it to any particular alignment algorithm it takes as input a standard alignment description format, allowing it to be used with any spliced alignment tool. Thirdly, many splicing analyses involve comparing alternative splicing events seen in different subsets of data, either to look for tissue-specific events or to test the effect of EST coverage on the number of events detected. Therefore, *Sircah* was designed to be flexible enough to allow the analysis of arbitrary subsets of data. Finally, the analysis of alternative splicing often involves large amounts of data, therefore *Sircah* was designed to minimise redundancy and provide compact visualisations of complex data.

3.2.2 *Program Overview*

Input

Sircah takes as input transcript models in the GFF3 format allowing the user the flexibility to choose the sources of evidence for the use in detecting alternative transcription. Such transcript models may come from the gene prediction pipelines of genome databases or from spliced alignments of ESTs or proteins against the genome. Within the GFF3 file the user may also specify the completeness of the transcript model used and may provide a set of tags, which can later be used to analyse subsets of data (Figure 3.23).

Detection Alternative Splicing and Transcription Events

Sircah uses a splice graph data model as first proposed by Heber et al. to represent the transcripts models in a non-redundant form (Heber et al., 2002). The nodes of the directed graph are exons, the edges introns and transcripts are represented as subpaths of the graph. Additionally overlapping exons are clustered into superexons. A series of rules are then applied to the data and based on the topology of the splice graph and the membership of the superexons the following alternative events can be classified: alternative initiation exons, alternative termination exons, exons with alternative 3′ and/or 5′ splice sites, retained introns and skipped exons (Figure 3.24).

Visualisation of Alternative Splicing and Transcription Events

The splice graph and the transcript models used to construct it can be visualised in a variety of ways to show: *(i)* the alternative events detectable and the transcript models used, *(ii)* the different events

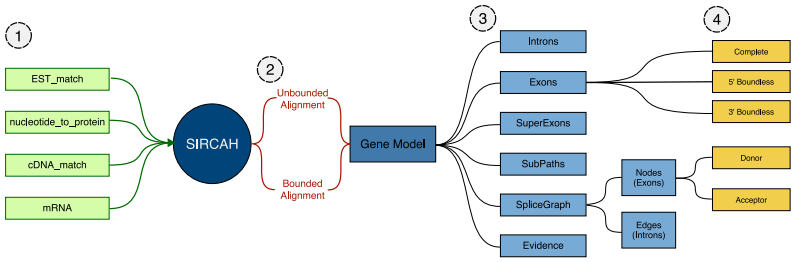


Figure 3.23. Sircah data models. *(1)* Transcript models are provided to *Sircah* in GFF3 format. *(2) Sircah* treats these alignments as either unbounded or bounded depending on whether they come from full length transcript models or not. *(3)* The transcript models are then used to create a gene model consisting of exons, introns, superexons (clusters of overlapping exons), a splice graph, subpaths (transcripts) and evidence (a mapping of the input data to subpaths). *(4)* Based on whether they're enclosed by introns or from a full-length transcript exons are further classified as boundless or complete, and based on their position within the splice graph they're classified as donors or acceptors.

detectable in subsets of the data and *(iii)* the coverage of introns and exons by transcript models (see Figure 3.25). The visualisation is created in the SVG format, allowing the creation of publication quality images. In addition the ids of all the elements in the SVG file are directly mappable to the ids used in the data objects, allowing the user to alter graphic after it's generated and even create interactive graphics using javascript.

Analysis of Subsets of Evidence

One of the most powerful features of *Sircah* is the ability to create splice graphs based on a subset of the total data. This allows the user to compare alternative transcription events under different conditions. For example by tagging the EST alignments with the EVOC expression ontology(Kelso et al., 2003) one can examine the tissue distribution of alternative transcription events (Figure 3.25c.).

Data Serialisation

In order to be able to carry out such analyses it is important to be able to save and reload the data models described above. To facilitate this *Sircah* can serialise its data objects to either an XML file or to a relational database using the SQLAlchemy python module.

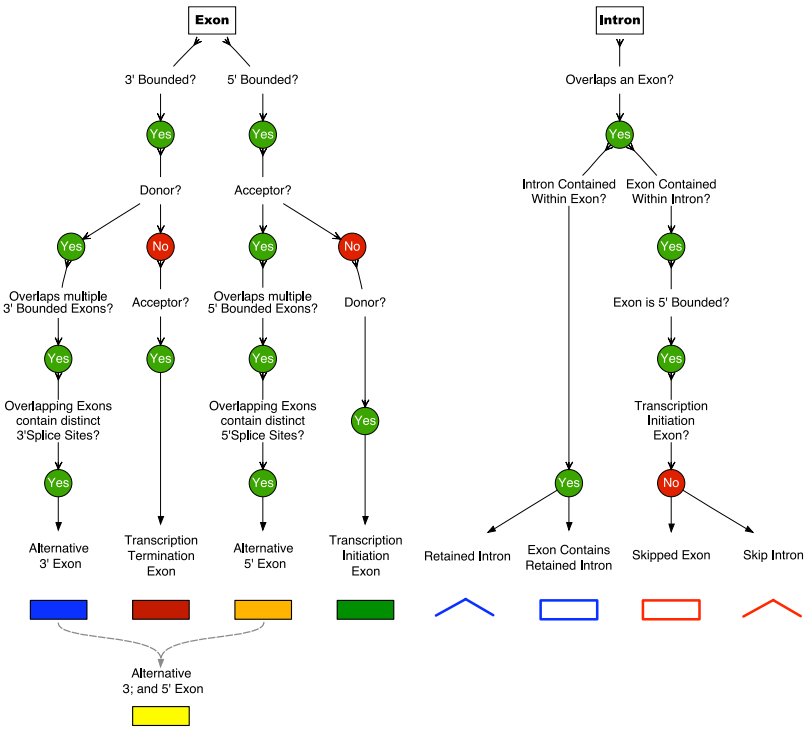


Figure 3.24. Rules used to detect alternative splicing. Based on the data model described in Figure 3.23 alternative initiation exons, alternative termination exons, exons with alternative 3′ and/or 5′ splice sites, retained introns and skipped exons are classified.

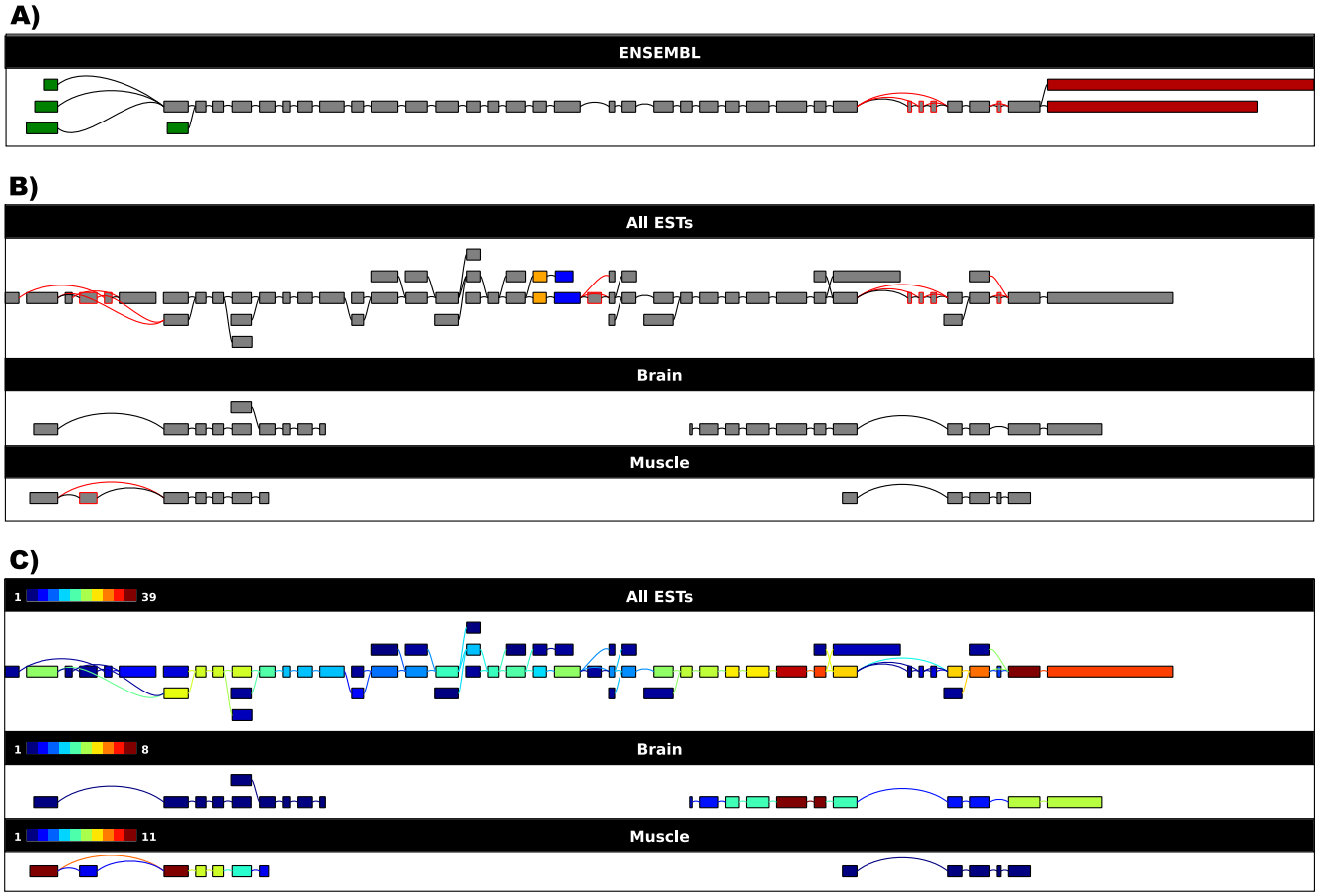


Figure 3.25. Sircah visualisations of the myosin 6 gene. (A) The splice graph created from the ENSEMBL transcripts, the *Sircah* visualisation includes the option of trimming long introns to make visualisation easier. The alternative events are coloured as in Figure 3.24. *(B)* The splice graph for the same gene, this time created using only ESTs. The bottom panels show the splice graphs from the subset of ESTs that are classified as brain or muscle in the EVOC ontology (Kelso et al., 2003). (C) The same plot as in *B* but this time coloured according to EST coverage rather than alternative event.

3.3 SEARCHING FOR CONSERVED ALTERNATIVE SPLICING EVENTS

3.3.1 *Introduction*

As discussed above, the majority of evolutionary studies of alternative splicing have focused on the conservation of alternative splice events in mammals, mainly focusing on human and mouse (Modrek and Lee, 2003; Thanaraj et al., 2003; Sorek et al., 2004, 2006; Yeo et al., 2005). However, in order to get a general understanding of the origins and subsequent evolution of alternative splice events it is important to look at not only a range of organisms but also over a broad range of evolutionary timescales. This approach can potentially reveal the effect of each organism's evolutionary history on the conservation of alternative splicing. For instance, there is a much higher rate of conservation of alternative splicing events between *Caenorhabditis elegans* and *Caenorhabditis briggsae* (Rukov et al., 2007) than between mouse and human (Yeo et al., 2005), even though their divergence times are similiar. One possible explanation could be due to the fact that the effective population size of nematodes is an order of magnitude higher than that of vertebrates (Lynch and Conery, 2003), leading to a higher efficiency of purifying selection. Therefore the higher conservation rate seen in nematodes could be due to the more efficient removal by selection of mildly deleterious alternative splice events. On the other hand, a more mechanistic explanation is based on the fact that complement of spliceosomal regulatory proteins in vertebrates is higher than in nematodes (Barbosa-Morais et al., 2006). This may have in turn have expanded the proportion of sequence space that can act as splicing regulatory elements, facilitating the evolution of novel isoforms. However, as long as these explanations are based on such a small number of species, and in the case of nematodes such a small number of genes, they will remain speculative.

There have been two studies so far that have looked at the conservation of alternative splicing over longer evolutionary distances. The first was by Resch et al. who looked at the conservation of alternative splicing events in human, mouse, rat, zebrafish and *Drosophila* (Resch et al., 2004a). They found a relatively high number of skipped exons in human that were also skipped in *Drosophila* (~300), however their method of detecting conservation could not distinguish between true orthologous alternative splice events and those that have evolved by convergent evolution (Copley, 2004). They found that the more conserved an alternative splicing event was, the more likely it was to preserve reading frame, however didn't look into detail at the conserved events. Another study by Malko et al. looked at the conservation of alternative splicing between *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae* and found a relatively low rate of conservation (Malko et al., 2006). However, they only looked at the genomic conservation of the gene segments involved in alternative splicing and didn't look at EST or cDNA evidence for confirmation, therefore these results should be treated with caution (discussed below).

3.3.2 *Methods*

A dataset for the detection of conserved alternative splicing

In order to determine the evolutionary forces affecting alternative splicing we selected eight organisms with a broad phylogenetic spread and a range of divergence times: *Apis mellifera*, *Drosophila melanogaster*, *Anopheles gambiae*, *Ciona intestinalis*, *Tetraodon nigroviridis*, *Gallus gallus*, *Mus musculus* and *Homo sapiens*. For each organism we downloaded the protein sequences from the Ensembl website (Hubbard et al., 2007) with the exception of *Apis mellifera* which was obtained from the honey bee genome consortium (Elsik et al., 2007) (sources of gene predictions are listed in Table 3.1). Orthologous groups were constructed according to the method described in (Zdobnov and Bork, 2007). From these a core set of groups was selected that maintained loose one-to-one orthology across all eight organisms, allowing for losses or gains in individual lineages. The number of genes in the core orthologous groups is shown in Table 3.1.

Organism	Source of Gene Predictions	Number of Orthologs
Apis mellifera	GLEAN3	2402
Drosophila melanogaster	Flybase	2440
Anopheles gambiae	Ensembl	2457
Ciona intestinalis	Ensembl	1778
Tetraodon nigroviridis	Genoscope	2758
Gallus gallus	Ensembl	2290
Mus musculus	Ensembl	2587
Homo sapiens	Ensembl	2440

Table 3.1. Orthologs used for the detection of conserved alternative splicing.

Detecting events present at the level of the transcriptome

Our starting set of alternative splicing events are those represented in the transcript isoforms predicted for each of the eight organisms, however not all of these alternative events have been seen in ESTs or cDNAs. Most gene prediction algorithms integrate information from the alignment of ESTs, cDNAs and proteins against the genome. In organisms with low EST/cDNA coverage the many of the predicted isoforms are based on the alignment of protein isoforms from other species. Therefore for some of the predicted isoforms in our data set there is no direct EST or cDNA evidence (see Figure 3.26). A significant number of alternatively spliced exons in human show conservation at the level of the genome, but not the transcriptome (Pan et al., 2005; Yeo et al., 2005). To ensure that we are dealing with genuinely conserved alternative splicing events we limit our analysis to those that are supported by either ESTs of cDNAs.

The dbEST database was downloaded from the NCBI and the relevant ESTs extracted by using the NCBI taxonomic id of each of the eight

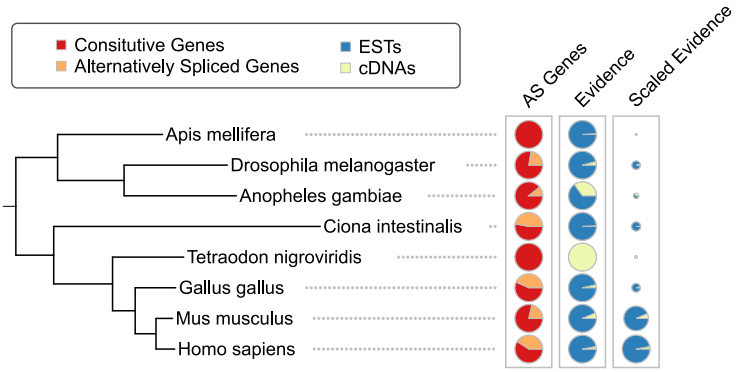


Figure 3.26. Data used for detection of conserved events. The first column shows the proportion of genes among the core orthologous groups that are predicted by the methods listed in Table 3.1 to have multiple protein splice forms. The second column shows the relative proportions ESTs and cDNAs that overlap these genes in each species. The final column contains the same data as the second, however this time the area of the pie charts is proportional to the total number of ESTs and cDNAs available for each organism.

species. Similarly full-length cDNAs for each organism were retrieved from NCBI's Entrez database using the query:

```
txid<ncbi taxonomic id>[Organism] AND (srcdb_genbank[prop
    ] OR srcdb_embl[prop] OR srcdb_ddbj[prop] NOT
    gbdiv_est[prop]) AND biomol_mrna[prop]
```

replacing the text between the angle brackets with the organism's taxonomic id. These were then mapped to the genomic sequence of the respective species using the GMAP spliced alignment program (Wu and Watanabe, 2005) and for each an alignment score based on the BLAT score was calculated (matching bases - (mismatching bases + number of non-intronic indels)). A filter was then applied to remove alignments that were: short (<100 nucleotides), low quality (less than 96% identity) or ambiguous (a score separation of less than 10 between the best and second best hits). Only alignments that overlapped the core set of orthologs were kept. Then the proteins predicted for each gene were aligned against the genomic locus using the Exonerate protein2genome model (Slater and Birney, 2005).

The EST, cDNA and protein alignments were then used as input to the *Sircah* program. From the resulting gene model, alternative splicing was detected using two subsets of the total data: *(i)* the protein against genome alignments only, and *(ii)* ESTs and cDNAs only. An alternative splicing event present in the first set (protein alignments) was considered to be present at the level of the transcriptome if the exon involved had the same alternative splicing classification in the second subset.

Detecting conserved alternative splicing using cross-species spliced alignments

In order to be able to detect conservation we must be able identify orthologous alternatively spliced regions of the gene. The most direct method of achieving this is to align orthologous genomic sequences against each other and identifying orthologous alternatively spliced exons. However this method is only suitable over relatively short evolutionary distances and may be sensitive to changes in gene structure. An alternative method is to use a spliced-alignment algorithm to align the proteins from one species against the genomic sequence of another (Malko et al., 2006). This has the added advantage that conservation can be detected over much longer evolutionary distances, and can even tolerate changes in gene structure (Malko et al., 2006).

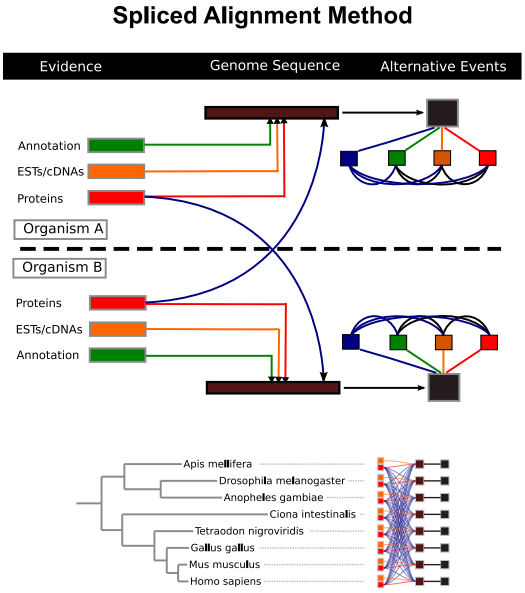


Figure 3.27. Spliced alignment method of detecting conserved alternative splicing. Conserved alternative splicing is detected by creating a spliced alignment between the proteins of one species and the orthologous genomic sequence in another species. The alternative splicing events detected with these alignments are compared with those detected by EST and cDNA alignments to remove those with no support. This is carried out in an all-against-all manner

To detect the conservation of alternative splicing in a given orthologous group and in a given organism (called the target), the genomic locus surrounding the gene was first extracted. Then the protein isoforms from one of the other seven organisms (called the query) were aligned against it using the protein2genome model of Exonerate (Slater and Birney, 2005). To allow for changes in protein length due to insertions and deletions, multiple alignments were permitted for a single protein as long as they didn't overlap in either genomic or protein coordinates. These alignments were then combined with the EST and

cDNA alignments described above and were used as input to the *Sircah* program. The alternative splicing events predicted by the cross-species alignments were then checked for EST/cDNA support as above, and if supported was considered a putative conserved alternative splicing events. In the same way putative conserved alternative splicing events were detected in the remaining six query organisms. This procedure was repeated for the remaining target organisms, such that an all-against-all comparison was done (Figure 3.27). Species that either multiple or no copy of a gene from this orthologous group were skipped.

Many of the conserved events predicted by this method, especially at longer distances, had no EST or cDNA support and manual inspection revealed that they were artifacts of the alignment process rather than true alternative splicing events. Given the difficulty that the spliced alignment program had over longer evolutionary distances we developed a complementary method of detecting conserved events using multiple sequence alignments.

Detecting conserved alternative splicing using multiple sequence alignments

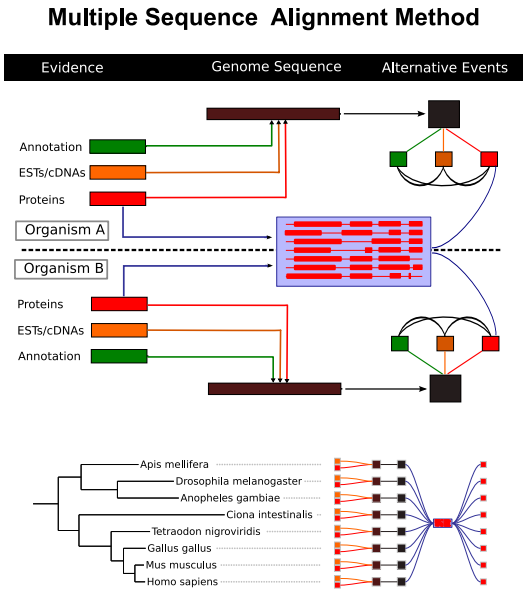


Figure 3.28. Multiple sequence alignment method of detecting conserved alternative splicing. Alternative splicing is detected using intraspecific protein against genome alignments and confirmed using EST and cDNA alignments. Then a multiple sequence alignment of all proteins from an orthologous group is then constructed and the genomic coordinates of the alternative events are first translated to protein coordinates and then multiple sequence alignment coordinates.

Another approach to detecting conserved alternative splicing events is use a multiple sequence alignment to look at the level of the protein for orthologous alternatively spliced segments (Figure 3.28). This

method has been used successfully to look at the evolution of introns over evolutionary distances greater than those under consideration here (Carmel et al., 2007a). Firstly, for each orthologous group a multiple sequence alignment was produced using Muscle (Edgar, 2004) (again excluding species that were not single-copy for this orthologous group). Then EST, cDNA and intraspecific protein against genome alignments were used to detect EST- and cDNA-supported alternative splicing as above. In order to map the these alternative splicing events, which are represented in genomic coordinates, to the multiple sequence alignment, we first need to translate from genomic coordinates to protein coordinates. This is possible using the CIGAR string produced by the Exonerate program, which details both the genomic and protein coordinates of the aligned blocks produced. The alternative splicing events, now represented in protein coordinates, are then translated into the coordinates of the multiple sequence alignment (Figure 3.29). Clusters of overlapping alternative splicing events were then created and clusters containing events from more than one organism represent putative conserved events.

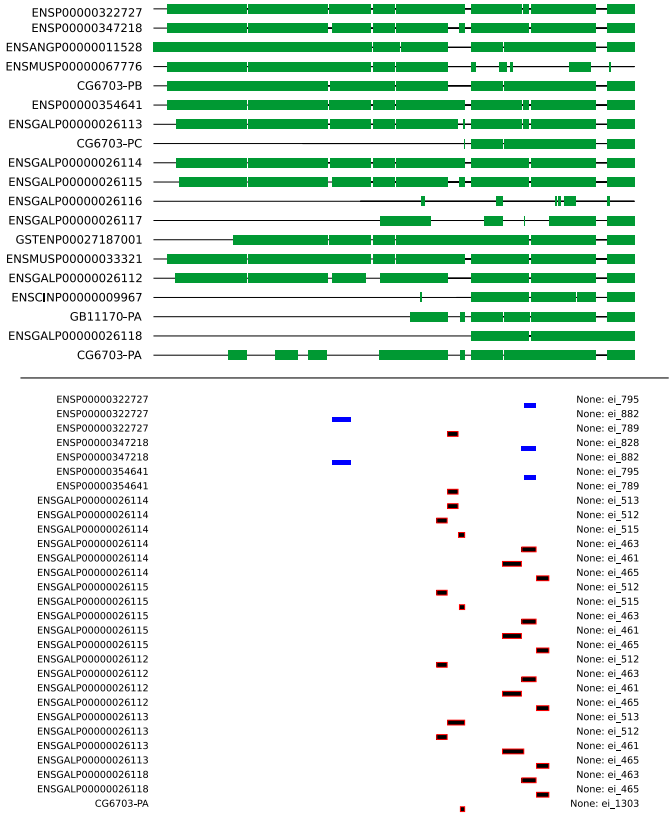


Figure 3.29. Alternative splicing events represented in multiple sequence alignment coordinates. The top panel shows a representation of the multiple sequence alignment, green blocks represent aligned positions in the protein. The bottom panel shows the EST- and cDNA-supported alternative splicing events coloured according to the same scheme in Figure 3.24. These events are then placed into clusters based on their overlap within the multiple sequence alignment coordinates, and clusters with events from more than one species represent putative conserved alternative splicing events.

3.3.3 *Results*

The results of the methods described above have yet to be analysed systematically, however we have an example that shows that such methods can detect alternative splicing events conserved between *Homo sapiens* and *Drosophila melanogaster* (Figure 3.30). Using the multiple sequence alignment method described above a conserved skipping event was detected with EST support in *Homo sapiens*, *Gallus gallus* and *Drosophila melanogaster*. The gene involved is a member of the membrane-associated guanylate kinase (MAGUK) family of proteins and is called *CASK* in vertebrates and either *CAKI* or *CAMGUK* in *Drosophila*. CASK is distinguished from the other members of the MAGUK family by the presence of an N-terminal Ca^{2+}/calmodulin-dependent kinase domain. The gene has been implicated in a range of neurological processes including neurotransmitter release, neural development and even the regulation of gene expression (Hsueh, 2006). In *Drosophila* the knockout of the gene leads to impaired movement (Martin and Ollo, 1996) and flight (Zordan et al., 2005), most likely due to spontaneous neurotransmitter release (Zordan et al., 2005). In contrast, the knockout of *CASK* in mouse is lethal, with the mice dying the first day after birth, however the effects on neuronal activity are more subtle (Atasoy et al., 2007). In addition *CASK* was recently identified as one of the targets of the *Nova* splicing regulators (Ule et al., 2005), which are responsible for regulation a host of alternative splicing events in synapse-related genes (Ule et al., 2006). It is not immediately clear how this conserved event might influence the activity of CASK as it located after the PDZ domain and immediately upstream of the SH3 domain. It's possible is that the skipped exon encodes a short linear motif (Neduva and Russell, 2005) or functions as a spacer between the two domains. However the conservation of this event over such a distance suggests that alternative splicing may have been important in the evolution of the nervous system. This is supported by the finding that another target of the *Nova* proteins, the *slo* gene, has experienced convergent evolution of exon skipping isoforms (Copley, 2004).

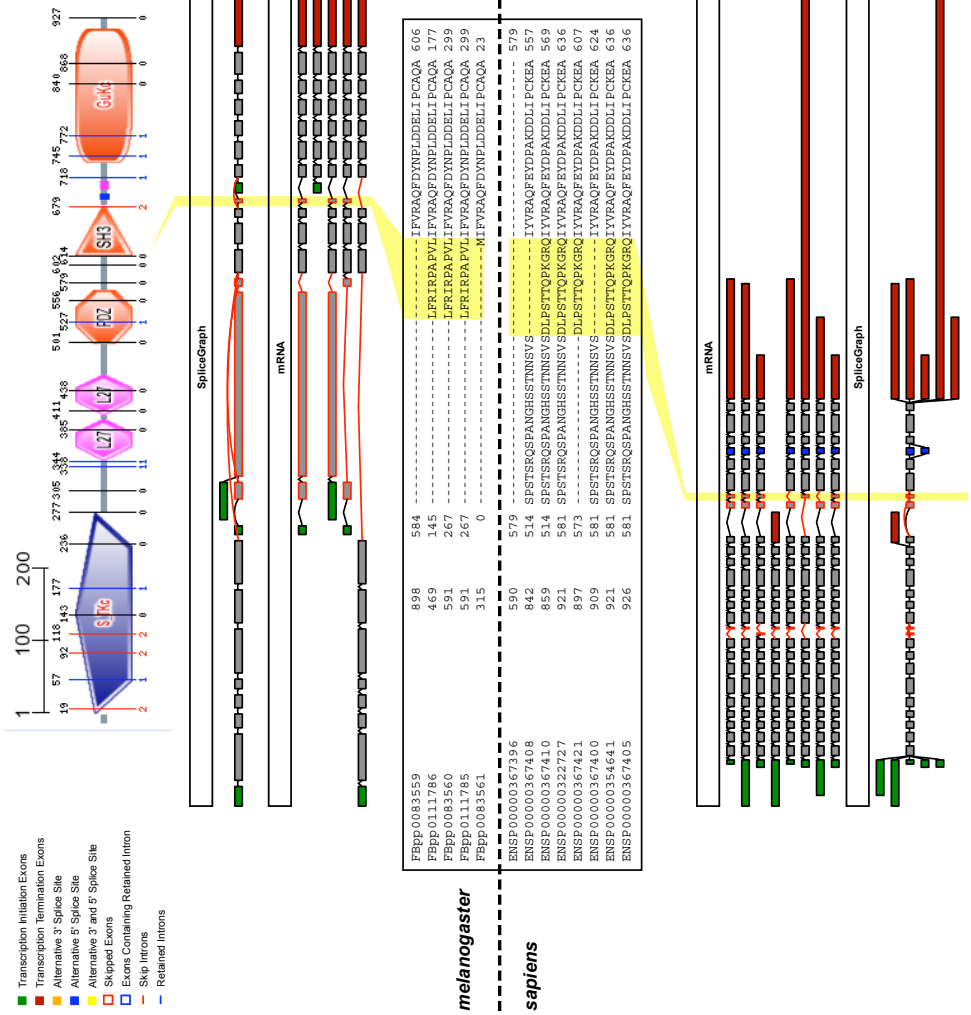


Figure 3.30. An exon skipping event conserved between human and fly. The central box in the figure show the multiple sequence alignment surrounding the conserved skipped exon, which is highlighted in yellow. This exon is also highlighted in the *Sircah* visualisations of the flybase (top) and Ensembl (bottom) transcripts. EST and cDNA alignments are not shown for clarity, however the exon skipping events in both species are supported by ESTs and/or cDNAs. At the top of the figure the domain structure of the longest fly isoform is shown. The conserved skipping event is located directly before the SH3 domain.

3.4 OUTLOOK

While the case story described above demonstrates the potential of being able to detect conserved alternative splicing events, there are still some hurdles to be overcome before we can make general statements about the evolution of alternative splicing. The most obvious one is the limiting effect of EST and cDNA coverage on our ability to detect alternative splice events, which in turn affects our ability to detect conserved events. We can account for this to some extent by using the frequency of an event among the ESTs that cover the region in question in one species to estimate the likelihood seeing the same event given the number of ESTs that cover the orthologous region. This is based on the assumption that the inclusion level of an alternative splicing event is also conserved, which has been shown to be true in some cases (Rukov et al., 2007) but might not be true for all. Another limitation is that we can only detect the conservation of alternative splicing events that are represented in proteins predicted by genome annotation pipelines, meaning that we can really only estimate conservation for a subset of all alternative splicing events.

Recent advances in sequencing technology have the potential to not only overcome these hurdles, but also to transform the study of alternative splicing. It was suggested recently by Khaitovich et al. that the future of gene expression profiling may lie in the high-throughput sequencing of full-length cDNA libraries rather than the hybridisation-based technologies used today (Khaitovich et al., 2006a). The value of this approach has already been demonstrated by the FANTOM consortium, which provided an unprecedented view of the mouse transcriptome (Carninci et al., 2005). If faster and cheaper sequencing allowed this approach to be carried out a large scale, then many of the limitations on the global analysis of alternative splicing would disappear. Firstly, sequence coverage would reach saturation, with the rate of discovery of novel isoforms dropping rapidly. Moreover, unlike microarray technologies which have to be developed anew for each species, such methods could easily be applied to any species, greatly facilitating the detection of conserved isoforms. Secondly, cDNA sequencing allows the detection of alternative splicing from the full transcript structure which avoids the "multi-assembly" problem associated with the transcript fragments represented by ESTs (Xing et al., 2004), allowing the direct prediction of protein sequences for coding cDNAs. Having the full transcript sequence also facilitates the detection of correlations between the alternative splicing events in parts of the transcript, which in turn can reveal potential regulatory components (Zavolan and van Nimwegen, 2006). Finally, this sequencing approach has the potential to allow the quantification of the absolute number of transcript molecules present (Khaitovich et al., 2006a), which is a vital step towards gaining a systems-level understanding of functional modules regulated by alternative splicing, such as the one regulated by the *Nova* genes (Ule et al., 2006). One of the challenges in dealing with this potential flood of data will be to cope with the large levels of redundancy in the cDNAs. By using the splice graph data structure *Sircah* can greatly reduce this redundancy while maintaining the original transcript data, and the corresponding visualisations can produce a compact representation of

a large number of transcripts.

While the focus on this part of the thesis has been on detecting conserved alternative splicing events, in reality this only represents a subset of all the functionally important alternative splicing events. Given the low conservation rate of alternative splicing between human and mouse it is not unreasonable to expect a population of alternative splicing events to be evolving under positive selection. However, in order to detect such selection one would need to develop a neutral model for the evolution of alternative splicing (Khaitovich et al., 2006a), which given complex nature of mutations that affect the evolution of alternative splicing would be difficult to derive. However, if the full-length cDNA sequencing approach described above is applied to a sample from a genotyped individual it will be possible to associate single nucleotide polymorphisms (SNPs) with the production of particular alternative splicing isoforms (Nembaware et al., 2004; Kwan et al., 2007; Hull et al., 2007). By combining such information with areas of the genome identified as being under positive selection, it may be possible to identify adaptive alternative splicing events.

BIBLIOGRAPHY

Christoph Adami. What is complexity? *Bioessays*, 24(12):1085–1094, Dec 2002. doi: 10.1002/bies.10192. URL http://dx.doi.org/10.1002/bies.10192. (Cited on pages 2, 25, and 43.)

Gladys Alexandre, Suzanne Greer-Phillips, and Igor B Zhulin. Ecological role of energy taxis in microorganisms. *FEMS Microbiol Rev*, 28(1):113–126, Feb 2004. doi: 10.1016/j.femsre.2003.10.003. URL http://dx.doi.org/10.1016/j.femsre.2003.10.003. (Cited on page 11.)

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990. doi: 10.1006/jmbi.1990.9999. URL http://dx.doi.org/10.1006/jmbi.1990.9999. (Cited on page 3.)

Ping An and Paula J Grabowski. Exon silencing by uagg motifs in response to neuronal excitation. *PLoS Biol*, 5(2):e36, Feb 2007. doi: 10.1371/journal.pbio.0050036. URL http://dx.doi.org/10.1371/journal.pbio.0050036. (Cited on page 31.)

Manuel Ares. Sing the genome electric: excited cells adjust their splicing. *PLoS Biol*, 5(2):e55, Feb 2007. doi: 10.1371/journal.pbio.0050055. URL http://dx.doi.org/10.1371/journal.pbio.0050055. (Cited on pages 28, 31, and 32.)

Gil Ast. How did alternative splicing evolve? *Nat Rev Genet*, 5(10):773–782, Oct 2004. doi: 10.1038/nrg1451. URL http://dx.doi.org/10.1038/nrg1451. (Cited on page 49.)

Deniz Atasoy, Susanne Schoch, Angela Ho, Krisztina A Nadasy, Xinran Liu, Weiqi Zhang, Konark Mukherjee, Elena D Nosyreva, Rafael Fernandez-Chacon, Markus Missler, Ege T Kavalali, and Thomas C Südhof. Deletion of cask in mice is lethal and impairs synaptic function. *Proc Natl Acad Sci U S A*, 104(7):2525–2530, Feb 2007. doi: 10.1073/pnas.0611003104. URL http://dx.doi.org/10.1073/pnas.0611003104. (Cited on page 62.)

Melinda D Baker, Peter M Wolanin, and Jeffry B Stock. Signal transduction in bacterial chemotaxis. *Bioessays*, 28(1):9–22, Jan 2006. doi: 10.1002/bies.20343. URL http://dx.doi.org/10.1002/bies.20343. (Cited on page 1.)

Nuno L Barbosa-Morais, Maria Carmo-Fonseca, and Samuel Aparício. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res*, 16(1):66–77, Jan 2006. doi: 10.1101/gr.3936206. URL http://dx.doi.org/10.1101/gr.3936206. (Cited on pages 28, 30, 49, and 55.)

Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L L Sonnhammer, David J Studholme, Corin Yeats,

and Sean R Eddy. The pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–D141, Jan 2004. doi: 10.1093/nar/gkh121. URL http://dx.doi.org/10.1093/nar/gkh121. (Cited on pages 6 and 13.)

Eric Batsché, Moshe Yaniv, and Christian Muchardt. The human swi/snf subunit brm is a regulator of alternative splicing. *Nat Struct Mol Biol*, 13(1):22–29, Jan 2006. doi: 10.1038/nsmb1030. URL http://dx.doi.org/10.1038/nsmb1030. (Cited on page 36.)

B. M. Bebout and F. Garcia-Pichel. Uv b-induced vertical migrations of cyanobacteria in a microbial mat. *Appl Environ Microbiol*, 61(12):4215–4222, Dec 1995. (Cited on page 11.)

S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mrna. *Proc Natl Acad Sci U S A*, 74(8):3171–3175, Aug 1977. (Cited on page 26.)

Douglas L Black. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336, 2003. doi: 10.1146/annurev.biochem.72.121801.161720. URL http://dx.doi.org/10.1146/annurev.biochem.72.121801.161720. (Cited on pages 29 and 31.)

Benjamin J Blencowe and May Khanna. Molecular biology: Rna in control. *Nature*, 447(7143):391–393, May 2007. doi: 10.1038/447391a. URL http://dx.doi.org/10.1038/447391a. (Cited on pages 33 and 35.)

P. Bork and E. V. Koonin. Predicting functions from protein sequences–where are the bottlenecks? *Nat Genet*, 18(4):313–318, Apr 1998. doi: 10.1038/ng0498-313. URL http://dx.doi.org/10.1038/ng0498-313. (Cited on pages 3 and 6.)

Peer Bork and Luis Serrano. Towards cellular systems in 4d. *Cell*, 121(4):507–509, May 2005. doi: 10.1016/j.cell.2005.05.001. URL http://dx.doi.org/10.1016/j.cell.2005.05.001. (Cited on page 6.)

Stephanie Boue, Ivica Letunic, and Peer Bork. Alternative splicing and evolution. *Bioessays*, 25(11):1031–1034, Nov 2003. doi: 10.1002/bies.10371. URL http://dx.doi.org/10.1002/bies.10371. (Cited on page 38.)

S. E. Brenner. Errors in genome annotation. *Trends Genet*, 15(4):132–133, Apr 1999. (Cited on pages 4 and 13.)

D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbrück, S. Krueger, J. Reich, and P. Bork. Est comparison indicates 38alternative splice forms. *FEBS Lett*, 474(1):83–86, May 2000. (Cited on page 39.)

David Brett, Heike Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, Jan 2002. doi: 10.1038/ng803. URL http://dx.doi.org/10.1038/ng803. (Cited on pages 2, 25, 43, and 47.)

Emanuele Buratti and Francisco E Baralle. Influence of rna secondary structure on the pre-mrna splicing process. *Mol Cell Biol*, 24(24):10505–10514, Dec 2004. doi: 10.1128/MCB.24.24.10505-10514.2004.

URL http://dx.doi.org/10.1128/MCB.24.24.10505-10514.2004. (Cited on page 32.)

C. B. Burge, R. A. Padgett, and P. A. Sharp. Evolutionary fates and origins of u12-type introns. *Mol Cell*, 2(6):773–785, Dec 1998. (Cited on page 28.)

Liran Carmel, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin. Evolutionarily conserved genes preferentially accumulate introns. *Genome Res*, 17(7):1045–1050, Jul 2007a. doi: 10.1101/gr.5978207. URL http://dx.doi.org/10.1101/gr.5978207. (Cited on page 60.)

Liran Carmel, Yuri I Wolf, Igor B Rogozin, and Eugene V Koonin. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res*, 17(7):1034–1044, Jul 2007b. doi: 10.1101/gr.6438607. URL http://dx.doi.org/10.1101/gr.6438607. (Cited on page 48.)

P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R R Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P T Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A M Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, F. A. N. T. O. M. Consortium, R. I.

K. E. N. Genome Exploration Research Group, and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, Sep 2005. (Cited on pages 39 and 64.)

Ming T Cheah, Andreas Wachter, Narasimhan Sudarsan, and Ronald R Breaker. Control of alternative rna splicing and gene expression by eukaryotic riboswitches. *Nature*, Apr 2007. doi: 10.1038/nature05769. URL http://dx.doi.org/10.1038/nature05769. (Cited on pages 33 and 41.)

Brian E Chen, Masahiro Kondo, Amélie Garnier, Fiona L Watson, Roland Püettmann-Holgado, David R Lamar, and Dietmar Schmucker. The molecular diversity of dscam is functionally required for neuronal wiring specificity in drosophila. *Cell*, 125 (3):607–620, May 2006. doi: 10.1016/j.cell.2006.03.034. URL http://dx.doi.org/10.1016/j.cell.2006.03.034. (Cited on page 32.)

Tzu-Ming Chern, Erik van Nimwegen, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, and Mihaela Zavolan. A simple physical model predicts small exon length variations. *PLoS Genet*, 2(4):e45, Apr 2006. doi: 10.1371/journal.pgen.0020045. URL http://dx.doi.org/10.1371/journal.pgen.0020045. (Cited on page 44.)

L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger rna. *Cell*, 12(1):1–8, Sep 1977. (Cited on page 26.)

Francesca D Ciccarelli, Christian von Mering, Mikita Suyama, Eoghan D Harrington, Elisa Izaurralde, and Peer Bork. Complex genomic rearrangements lead to novel primate gene function. *Genome Res*, 15 (3):343–351, Mar 2005. doi: 10.1101/gr.3266405. URL http://dx.doi.org/10.1101/gr.3266405. (Cited on page v.)

J. M. Claverie. Gene number. what if there are only 30,000 human genes? *Science*, 291(5507):1255–1257, Feb 2001. (Cited on page 25.)

Lesley Collins and David Penny. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol*, 22(4):1053–1066, Apr 2005. doi: 10.1093/molbev/msi091. URL http://dx.doi.org/10.1093/molbev/msi091. (Cited on pages 28 and 34.)

Gavin C Conant and Andreas Wagner. Convergent evolution of gene circuits. *Nat Genet*, 34(3):264–266, Jul 2003. doi: 10.1038/ng1181. URL http://dx.doi.org/10.1038/ng1181. (Cited on page 38.)

E. N. C. O. D. E. Project Consortium. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007. (Cited on page 37.)

Richard R Copley. Evolutionary convergence of alternative splicing in ion channels. *Trends Genet*, 20(4):171–176, Apr 2004. (Cited on pages 55 and 62.)

Mack E Crayton, Bradford C Powell, Todd J Vision, and Morgan C Giddings. Tracking the evolution of alternatively spliced exons within the dscam family. *BMC Evol Biol*, 6:16, 2006. doi: 10.1186/1471-2148-6-16. URL http://dx.doi.org/10.1186/1471-2148-6-16. (Cited on page 47.)

T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–328, Sep 1998. (Cited on pages 4 and 7.)

Manuel de la Mata and Alberto R Kornblihtt. Rna polymerase ii c-terminal domain mediates regulation of alternative splicing by srp20. *Nat Struct Mol Biol*, 13(11):973–980, Nov 2006. doi: 10.1038/nsmb1155. URL http://dx.doi.org/10.1038/nsmb1155. (Cited on page 37.)

Edward F DeLong, Christina M Preston, Tracy Mincer, Virginia Rich, Steven J Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matthew B Sullivan, Robert Edwards, Beltran Rodriguez Brito, Sallie W Chisholm, and David M Karl. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311 (5760):496–503, Jan 2006. doi: ence.1120250. URL http://dx.doi.org/ence.1120250. (Cited on page 3.)

Ebru Demir and Barry J Dickson. fruitless splicing specifies male courtship behavior in drosophila. *Cell*, 121(5):785–794, Jun 2005. doi: 10.1016/j.cell.2005.04.027. URL http://dx.doi.org/10.1016/j.cell.2005.04.027. (Cited on page 31.)

Yuemei Dong, Harry E Taylor, and George Dimopoulos. Agdscam, a hypervariable immunoglobulin domain-containing receptor of the anopheles gambiae innate immune system. *PLoS Biol*, 4(7):e229, Jun 2006. doi: 10.1371/journal.pbio.0040229. URL http://dx.doi.org/10.1371/journal.pbio.0040229. (Cited on page 32.)

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004. doi: 10.1093/nar/gkh340. URL http://dx.doi.org/10.1093/nar/gkh340. (Cited on page 60.)

Christine G Elsik, Aaron J Mackey, Justin T Reese, Natalia V Milshina, David S Roos, and George M Weinstock. Creating a honey bee consensus gene set. *Genome Biol*, 8(1):R13, 2007. doi: 10.1186/gb-2007-8-1-r13. URL http://dx.doi.org/10.1186/gb-2007-8-1-r13. (Cited on page 56.)

A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, Nov 1999. doi: 10.1038/47056. URL http://dx.doi.org/10.1038/47056. (Cited on page 4.)

A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7): 1575–1584, Apr 2002. (Cited on page 21.)

William G Fairbrother, Ru-Fang Yeh, Phillip A Sharp, and Christopher B Burge. Predictive identification of exonic splicing enhancers in human

genes. *Science*, 297(5583):1007–1013, Aug 2002. doi: 10.1126/science.1073774. URL http://dx.doi.org/10.1126/science.1073774. (Cited on page 29.)

R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, Jul 1995. (Cited on page 2.)

A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, Apr 1999. (Cited on page 49.)

N. Frankenberg, J. Moser, and D. Jahn. Bacterial heme biosynthesis and its biotechnological application. *Appl Microbiol Biotechnol*, 63 (2):115–127, Dec 2003. doi: 10.1007/s00253-003-1432-2. URL http://dx.doi.org/10.1007/s00253-003-1432-2. (Cited on page 11.)

Claire M Fraser-Liggett. Insights on biology and evolution from microbial genome sequencing. *Genome Res*, 15(12):1603–1610, Dec 2005. doi: 10.1101/gr.3724205. URL http://dx.doi.org/10.1101/gr.3724205. (Cited on page 2.)

M. Gerstein, E. L. Sonnhammer, and C. Chothia. Volume changes in protein evolution. *J Mol Biol*, 236(4):1067–1078, Mar 1994. (Cited on page 16.)

Steven R Gill, Mihai Pop, Robert T Deboy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359, Jun 2006. doi: 10.1126/science.1124234. URL http://dx.doi.org/10.1126/science.1124234. (Cited on page 3.)

Stephen J Giovannoni, H. James Tripp, Scott Givan, Mircea Podar, Kevin L Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H Richardson, Michiel Noordewier, Michael S Rappé, Jay M Short, James C Carrington, and Eric J Mathur. Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245, Aug 2005. doi: 10.1126/science.1114057. URL http://dx.doi.org/10.1126/science.1114057. (Cited on page 10.)

Amir Goren, Oren Ram, Maayan Amit, Hadas Keren, Galit Lev-Maor, Ida Vig, Tal Pupko, and Gil Ast. Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. *Mol Cell*, 22(6):769–781, Jun 2006. doi: 10.1016/j.molcel.2006.05.008. URL http://dx.doi.org/10.1016/j.molcel.2006.05.008. (Cited on page 29.)

Brenton R Graveley. Mutually exclusive splicing of the insect dscam pre-mrna directed by competing intronic rna secondary structures. *Cell*, 123(1):65–73, Oct 2005. doi: 10.1016/j.cell.2005.07.028. URL http://dx.doi.org/10.1016/j.cell.2005.07.028. (Cited on pages 32, 33, and 38.)

Steven J Hallam, Nik Putnam, Christina M Preston, John C Detter, Daniel Rokhsar, Paul M Richardson, and Edward F DeLong. Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science*, 305(5689):1457–1462, Sep 2004. doi: 10.1126/science.1100025. URL http://dx.doi.org/10.1126/science.1100025. (Cited on page 3.)

E D Harrington, S Boue, Juan Valcarcel, Jens G Reich, and Peer Bork. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet*, 36(9):916–917, September 2004. ISSN 1061-4036. URL http://dx.doi.org/10.1038/ng0904-916. (Cited on pages v, 2, and 43.)

E. D. Harrington, A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, and P. Bork. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *PNAS*, page 0702636104, 2007. doi: 10.1073/pnas.0702636104. URL http://www.pnas.org/cgi/content/abstract/0702636104v1. (Cited on pages v and 3.)

G. Traver Hart, Arun K Ramani, and Edward M Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006. doi: 10.1186/gb-2006-7-11-120. URL http://dx.doi.org/10.1186/gb-2006-7-11-120. (Cited on page 1.)

Steffen Heber, Max Alekseyev, Sing-Hoi Sze, Haixu Tang, and Pavel A Pevzner. Splicing graphs and est assembly problem. *Bioinformatics*, 18 Suppl 1:S181–S188, 2002. (Cited on page 51.)

I. Herskowitz and D. Hagen. The lysis-lysogeny decision of phage lambda: explicit programming and responsiveness. *Annu Rev Genet*, 14:399–445, 1980. doi: 10.1146/annurev.ge.14.120180.002151. URL http://dx.doi.org/10.1146/annurev.ge.14.120180.002151. (Cited on pages 1 and 25.)

Yi-Ping Hsueh. The role of the maguk protein cask in neural development and synaptic function. *Curr Med Chem*, 13(16):1915–1927, 2006. (Cited on page 62.)

Cindy Shen Huang, Song-Hai Shi, Jernej Ule, Matteo Ruggiu, Laura A Barker, Robert B Darnell, Yuh Nung Jan, and Lily Yeh Jan. Common molecular pathways mediate long-term potentiation of synaptic excitation and slow synaptic inhibition. *Cell*, 123(1):105–118, Oct 2005. doi: 10.1016/j.cell.2005.07.033. URL http://dx.doi.org/10.1016/j.cell.2005.07.033. (Cited on page 32.)

T. J P Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617, Jan 2007. doi:

10.1093/nar/gkl996. URL http://dx.doi.org/10.1093/nar/gkl996. (Cited on page 56.)

A. L. Hughes. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*, 256(1346):119–124, May 1994. doi: 10.1098/rspb.1994.0058. URL http://dx.doi.org/10.1098/rspb.1994.0058. (Cited on page 49.)

Jeremy Hull, Susana Campino, Kate Rowlands, Man-Suen Chan, Richard R Copley, Martin S Taylor, Kirk Rockett, Gareth Elvidge, Brendan Keating, Julian Knight, and Dominic Kwiatkowski. Identification of common genetic variation that modulates alternative splicing. *PLoS Genet*, 3(6):e99, Jun 2007. doi: 10.1371/journal.pgen.0030099. URL http://dx.doi.org/10.1371/journal.pgen.0030099. (Cited on page 65.)

Matthew Hurles. Gene duplication: the genomic trade in spare parts. *PLoS Biol*, 2(7):E206, Jul 2004. doi: 10.1371/journal.pbio.0020206. URL http://dx.doi.org/10.1371/journal.pbio.0020206. (Cited on page 47.)

Ioannis Iliopoulos, Sophia Tsoka, Miguel A Andrade, Anton J Enright, Mark Carroll, Patrick Poullet, Vassilis Promponas, Theodore Liakopoulos, Giorgos Palaios, Claude Pasquier, Stavros Hamodrakas, Javier Tamames, Asutosh T Yagnik, Anna Tramontano, Damien Devos, Christian Blaschke, Alfonso Valencia, David Brett, David Martin, Christophe Leroy, Isidore Rigoutsos, Chris Sander, and Christos A Ouzounis. Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, 19(6):717–726, Apr 2003. (Cited on page 14.)

José-María Izquierdo and Juan Valcárcel. A simple principle to explain the evolution of pre-mrna splicing. *Genes Dev*, 20(13):1679–1684, Jul 2006. doi: 10.1101/gad.1449106. URL http://dx.doi.org/10.1101/gad.1449106. (Cited on pages 30 and 48.)

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961. (Cited on page 1.)

Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton, and Daniel D Shoemaker. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–2144, Dec 2003. doi: 10.1126/science.1090100. URL http://dx.doi.org/10.1126/science.1090100. (Cited on page 39.)

Patricia J Johnson. Spliceosomal introns in a deep-branching eukaryote: the splice of life. *Proc Natl Acad Sci U S A*, 99(6):3359–3361, Mar 2002. doi: 10.1073/pnas.072084199. URL http://dx.doi.org/10.1073/pnas.072084199. (Cited on page 41.)

Melissa S Jurica and Melissa J Moore. Pre-mrna splicing: awash in a sea of proteins. *Mol Cell*, 12(1):5–14, Jul 2003. (Cited on pages 26 and 29.)

Maria Kalyna, Sergiy Lopato, Viktor Voronin, and Andrea Barta. Evolutionary conservation and regulation of particular alternative splicing events in plant sr proteins. *Nucleic Acids Res*, 34(16):4395–4405, 2006. doi: 10.1093/nar/gkl570. URL http://dx.doi.org/10.1093/nar/gkl570. (Cited on page 38.)

Zhengyan Kan, David States, and Warren Gish. Selecting for functional alternative splices in ests. *Genome Res*, 12(12):1837–1845, Dec 2002. doi: 10.1101/gr.764102. URL http://dx.doi.org/10.1101/gr.764102. (Cited on pages 26, 39, 43, 44, 45, and 47.)

Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280, Jan 2004. doi: 10.1093/nar/gkh063. URL http://dx.doi.org/10.1093/nar/gkh063. (Cited on page 6.)

Rotem Karni, Elisa de Stanchina, Scott W Lowe, Rahul Sinha, David Mu, and Adrian R Krainer. The gene encoding the splicing factor sf2/asf is a proto-oncogene. *Nat Struct Mol Biol*, 14(3):185–193, Mar 2007. doi: 10.1038/nsmb1209. URL http://dx.doi.org/10.1038/nsmb1209. (Cited on pages 26 and 39.)

Vaishali Katju and Michael Lynch. On the formation of novel genes by duplication in the caenorhabditis elegans genome. *Mol Biol Evol*, 23(5):1056–1067, May 2006. doi: 10.1093/molbev/msj114. URL http://dx.doi.org/10.1093/molbev/msj114. (Cited on page 49.)

Janet Kelso, Johann Visagie, Gregory Theiler, Alan Christoffels, Soraya Bardien, Damian Smedley, Darren Otgaar, Gary Greyling, C. Victor Jongeneel, Mark I McCarthy, Tania Hide, and Winston Hide. evoc: a controlled vocabulary for unifying gene expression data. *Genome Res*, 13(6A):1222–1230, Jun 2003. doi: 10.1101/gr.985203. URL http://dx.doi.org/10.1101/gr.985203. (Cited on pages 39, 40, 43, 52, and 54.)

Philipp Khaitovich, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. Evolution of primate gene expression. *Nat Rev Genet*, 7(9):693–702, Sep 2006a. doi: 10.1038/nrg1940. URL http://dx.doi.org/10.1038/nrg1940. (Cited on pages 44, 46, 64, and 65.)

Philipp Khaitovich, Kun Tang, Henriette Franz, Janet Kelso, Ines Hellmann, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. Positive selection on gene expression in the human brain. *Curr Biol*, 16(10):R356–R358, May 2006b. doi: 10.1016/j.cub.2006.03.082. URL http://dx.doi.org/10.1016/j.cub.2006.03.082. (Cited on page 46.)

Eddo Kim, Alon Magen, and Gil Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, 35(1):125–131, 2007. doi: 10.1093/nar/gkl924. URL http://dx.doi.org/10.1093/nar/gkl924. (Cited on pages 2 and 43.)

Heebal Kim, Robert Klein, Jacek Majewski, and Jurg Ott. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet*, 36(9):915–6; author reply 916–7, Sep 2004. doi: 10.1038/ng0904-915. URL http://dx.doi.org/10.1038/ng0904-915. (Cited on pages 2 and 43.)

M. Kirschner and J. Gerhart. Evolvability. *Proc Natl Acad Sci U S A*, 95 (15):8420–8427, Jul 1998. (Cited on page 47.)

Fyodor A Kondrashov and Eugene V Koonin. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet*, 19(3):115–119, Mar 2003. (Cited on page 47.)

Eugene V Koonin. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct*, 1:22, 2006. doi: 10.1186/1745-6150-1-22. URL http://dx.doi.org/10.1186/1745-6150-1-22. (Cited on page 34.)

Naama M Kopelman, Doron Lancet, and Itai Yanai. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet*, 37(6):588–589, Jun 2005. doi: 10.1038/ng1575. URL http://dx.doi.org/10.1038/ng1575. (Cited on pages 26 and 50.)

Jan O Korbel, Lars J Jensen, Christian von Mering, and Peer Bork. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol*, 22(7):911–917, Jul 2004. doi: 10.1038/nbt988. URL http://dx.doi.org/10.1038/nbt988. (Cited on page 7.)

Eli Koren, Galit Lev-Maor, and Gil Ast. The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. *PLoS Comput Biol*, 3(5):e95, May 2007. doi: 10.1371/journal.pcbi.0030095. URL http://dx.doi.org/10.1371/journal.pcbi.0030095. (Cited on pages 46 and 49.)

Alberto R Kornblihtt. Promoter usage and alternative splicing. *Curr Opin Cell Biol*, 17(3):262–268, Jun 2005. doi: 10.1016/j.ceb.2005.04.014. URL http://dx.doi.org/10.1016/j.ceb.2005.04.014. (Cited on pages 36 and 37.)

Alberto R Kornblihtt. Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol*, 13(1):5–7, Jan 2006. doi: 10.1038/nsmb0106-5. URL http://dx.doi.org/10.1038/nsmb0106-5. (Cited on pages 36 and 37.)

M. Krawczak, J. Reiss, and D. N. Cooper. The mutational spectrum of single base-pair substitutions in mrna splice junctions of human genes: causes and consequences. *Hum Genet*, 90(1-2):41–54, 1992. (Cited on page 26.)

Jenny M Kreahling and Brenton R Graveley. The istem, a long-range rna secondary structure element required for efficient exon inclusion in the drosophila dscam pre-mrna. *Mol Cell Biol*, 25(23):10251–10260, Dec 2005. doi: 10.1128/MCB.25.23.10251-10260.2005. URL http://dx.doi.org/10.1128/MCB.25.23.10251-10260.2005. (Cited on page 33.)

Tracy L Kress and Christine Guthrie. Molecular biology. accurate rna siting and splicing gets help from a dek-hand. *Science*, 312 (5782):1886–1887, Jun 2006. doi: 10.1126/science.1130324. URL http://dx.doi.org/10.1126/science.1130324. (Cited on page 37.)

Evgenia V Kriventseva, Ina Koch, Rolf Apweiler, Martin Vingron, Peer Bork, Mikhail S Gelfand, and Shamil Sunyaev. Increase of functional

diversity by alternative splicing. *Trends Genet*, 19(3):124–128, Mar 2003. (Cited on page 46.)

Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, David Serre, Harry Zuzan, Tyson A Clark, Anthony Schweitzer, Michelle K Staples, Hui Wang, John E Blume, Thomas J Hudson, Rob Sladek, and Jacek Majewski. Heritability of alternative splicing in the human genome. *Genome Res*, 17(8):1210–1218, Aug 2007. doi: 10.1101/gr.6281007. URL http://dx.doi.org/10.1101/gr.6281007. (Cited on page 65.)

Patricia L Lakin-Thomas and Stuart Brody. Circadian rhythms in microorganisms: new complexities. *Annu Rev Microbiol*, 58:489–519, 2004. doi: 10.1146/annurev.micro.58.030603.123744. URL http://dx.doi.org/10.1146/annurev.micro.58.030603.123744. (Cited on page 11.)

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp,

W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921, Feb 2001. (Cited on page 39.)

Liana F Lareau, Richard E Green, Rajiv S Bhatnagar, and Steven E Brenner. The evolving roles of alternative splicing. *Curr Opin Struct Biol*, 14(3):273–282, Jun 2004. doi: 10.1016/j.sbi.2004.05.002. URL http://dx.doi.org/10.1016/j.sbi.2004.05.002. (Cited on pages 25, 38, and 44.)

Liana F Lareau, Maki Inada, Richard E Green, Jordan C Wengrod, and Steven E Brenner. Unproductive splicing of sr genes associated with highly conserved and ultraconserved dna elements. *Nature*, Mar 2007. doi: 10.1038/nature05676. URL http://dx.doi.org/10.1038/nature05676. (Cited on pages 30 and 38.)

Ji-Ann Lee, Yi Xing, David Nguyen, Jiuyong Xie, Christopher J Lee, and Douglas L Black. Depolarization and cam kinase iv modulate nmda receptor splicing through two essential rna elements. *PLoS Biol*, 5(2):e40, Feb 2007. doi: 10.1371/journal.pbio.0050040. URL http://dx.doi.org/10.1371/journal.pbio.0050040. (Cited on page 31.)

Fabrice Lejeune and Lynne E Maquat. Mechanistic links between nonsense-mediated mrna decay and pre-mrna splicing in mammalian cells. *Curr Opin Cell Biol*, 17(3):309–315, Jun 2005. doi: 10.1016/j.ceb.2005.03.002. URL http://dx.doi.org/10.1016/j.ceb.2005.03.002. (Cited on page 28.)

Ivica Letunic, Richard R Copley, and Peer Bork. Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet*, 11(13):1561–1567, Jun 2002. (Cited on page 47.)

Ivica Letunic, Richard R Copley, Birgit Pils, Stefan Pinkert, Jörg Schultz, and Peer Bork. Smart 5: domains in the context of genomes and networks. *Nucleic Acids Res*, 34(Database issue):D257–D260, Jan 2006. doi: 10.1093/nar/gkj079. URL http://dx.doi.org/10.1093/nar/gkj079. (Cited on pages 6 and 13.)

Galit Lev-Maor, Rotem Sorek, Noam Shomron, and Gil Ast. The birth of an alternatively spliced exon: 3′ splice-site selection in alu exons. *Science*, 300(5623):1288–1291, May 2003. doi: 10.1126/science.1082588. URL http://dx.doi.org/10.1126/science.1082588. (Cited on pages 47 and 49.)

Benjamin P Lewis, Richard E Green, and Steven E Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans. *Proc Natl Acad Sci U S A*, 100(1):

189–192, Jan 2003. doi: 10.1073/pnas.0136770100. URL http://dx.doi.org/10.1073/pnas.0136770100. (Cited on page 38.)

Diane Lipscombe. Neuronal proteins custom designed by alternative splicing. *Curr Opin Neurobiol*, 15(3):358–363, Jun 2005. doi: 10.1016/j.conb.2005.04.002. URL http://dx.doi.org/10.1016/j.conb.2005.04.002. (Cited on page 31.)

Kristen W Lynch. Cotranscriptional splicing regulation: it's not just about speed. *Nat Struct Mol Biol*, 13(11):952–953, Nov 2006. doi: 10.1038/nsmb1106-952. URL http://dx.doi.org/10.1038/nsmb1106-952. (Cited on page 37.)

Michael Lynch. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A*, 104 Suppl 1: 8597–8604, May 2007. doi: 10.1073/pnas.0702207104. URL http://dx.doi.org/10.1073/pnas.0702207104. (Cited on page 48.)

Michael Lynch and John S Conery. The origins of genome complexity. *Science*, 302(5649):1401–1404, Nov 2003. doi: 10.1126/science.1089370. URL http://dx.doi.org/10.1126/science.1089370. (Cited on pages 48 and 55.)

Michael Lynch and Vaishali Katju. The altered evolutionary trajectories of gene duplicates. *Trends Genet*, 20(11):544–549, Nov 2004. doi: 10.1016/j.tig.2004.09.001. URL http://dx.doi.org/10.1016/j.tig.2004.09.001. (Cited on pages 47 and 49.)

Michael Lynch and Aaron O Richardson. The evolution of spliceosomal introns. *Curr Opin Genet Dev*, 12(6):701–710, Dec 2002. (Cited on page 26.)

Núria López-Bigas, Benjamin Audit, Christos Ouzounis, Genís Parra, and Roderic Guigó. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett*, 579(9):1900–1903, Mar 2005. doi: 10.1016/j.febslet.2005.02.047. URL http://dx.doi.org/10.1016/j.febslet.2005.02.047. (Cited on page 26.)

Alon Magen and Gil Ast. The importance of being divisible by three in alternative splicing. *Nucleic Acids Res*, 33(17):5574–5582, 2005. doi: 10.1093/nar/gki858. URL http://dx.doi.org/10.1093/nar/gki858. (Cited on page 46.)

Eugene V Makeyev, Jiangwen Zhang, Monica A Carrasco, and Tom Maniatis. The microrna mir-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mrna splicing. *Mol Cell*, 27(3):435–448, Aug 2007. doi: 10.1016/j.molcel.2007.07.015. URL http://dx.doi.org/10.1016/j.molcel.2007.07.015. (Cited on pages 30 and 38.)

Dmitry B Malko, Vsevolod J Makeev, Andrey A Mironov, and Mikhail S Gelfand. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res*, 16(4):505–509, Apr 2006. doi: 10.1101/gr.4236606. URL http://dx.doi.org/10.1101/gr.4236606. (Cited on pages 45, 55, and 58.)

Tom Maniatis and Robin Reed. An extensive network of coupling among gene expression machines. *Nature*, 416(6880):499–506, Apr 2002. doi: .1038/416499a. URL http://dx.doi.org/.1038/416499a. (Cited on page 36.)

Tom Maniatis and Bosiljka Tasic. Alternative pre-mrna splicing and proteome expansion in metazoans. *Nature*, 418(6894):236–243, Jul 2002. doi: 10.1038/418236a. URL http://dx.doi.org/10.1038/418236a. (Cited on page 36.)

E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, Nov 1999. doi: 10.1038/47048. URL http://dx.doi.org/10.1038/47048. (Cited on page 4.)

J. R. Martin and R. Ollo. A new drosophila ca2+/calmodulin-dependent protein kinase (caki) is localized in the central nervous system and implicated in walking speed. *EMBO J*, 15(8):1865–1876, Apr 1996. (Cited on page 62.)

Héctor García Martín, Natalia Ivanova, Victor Kunin, Falk Warnecke, Kerrie W Barry, Alice C McHardy, Christine Yeates, Shaomei He, Asaf A Salamov, Ernest Szeto, Eileen Dalin, Nik H Putnam, Harris J Shapiro, Jasmyn L Pangilinan, Isidore Rigoutsos, Nikos C Kyrpides, Linda Louise Blackall, Katherine D McMahon, and Philip Hugenholtz. Metagenomic analysis of two enhanced biological phosphorus removal (ebpr) sludge communities. *Nat Biotechnol*, 24(10):1263–1269, Oct 2006. doi: 10.1038/nbt1247. URL http://dx.doi.org/10.1038/nbt1247. (Cited on page 3.)

Arianne J Matlin, Francis Clark, and Christopher W J Smith. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–398, May 2005. doi: 10.1038/nrm1645. URL http://dx.doi.org/10.1038/nrm1645. (Cited on pages 28, 29, and 30.)

Gerhard Michal, editor. *Biochemical pathways: an atlas of biochemistry and molecular biology*. Wiley, New York, 1999. (Cited on page 11.)

B. Modrek, A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*, 29(13):2850–2859, Jul 2001. (Cited on page 39.)

Barmak Modrek and Christopher J Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*, 34(2):177–180, Jun 2003. doi: 10.1038/ng1159. URL http://dx.doi.org/10.1038/ng1159. (Cited on pages 26, 45, 46, 47, 50, and 55.)

Melissa J Moore. From birth to death: the complex lives of eukaryotic mrnas. *Science*, 309(5740):1514–1518, Sep 2005. doi: 10.1126/science.1111443. URL http://dx.doi.org/10.1126/science.1111443. (Cited on page 36.)

Hideki Nagasaki, Masanori Arita, Tatsuya Nishizawa, Makiko Suwa, and Osamu Gotoh. Species-specific variation of alternative splicing

and transcriptional initiation in six eukaryotes. *Gene*, 364:53–62, Dec 2005. doi: 10.1016/j.gene.2005.07.027. URL http://dx.doi.org/10.1016/j.gene.2005.07.027. (Cited on page 43.)

Victor Neduva and Robert B Russell. Linear motifs: evolutionary interaction switches. *FEBS Lett*, 579(15):3342–3345, Jun 2005. doi: 10.1016/j.febslet.2005.04.005. URL http://dx.doi.org/10.1016/j.febslet.2005.04.005. (Cited on page 62.)

Victoria Nembaware, Kenneth H Wolfe, Fabiana Bettoni, Janet Kelso, and Cathal Seoighe. Allele-specific transcript isoforms in human. *FEBS Lett*, 577(1-2):233–238, Nov 2004. doi: 10.1016/j.febslet.2004.10.018. URL http://dx.doi.org/10.1016/j.febslet.2004.10.018. (Cited on pages 39, 44, and 65.)

Timothy W Nilsen. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25(12):1147–1149, Dec 2003. doi: 10.1002/bies.10394. URL http://dx.doi.org/10.1002/bies.10394. (Cited on page 26.)

Julie E J Nixon, Amy Wang, Hilary G Morrison, Andrew G McArthur, Mitchell L Sogin, Brendan J Loftus, and John Samuelson. A spliceosomal intron in giardia lamblia. *Proc Natl Acad Sci U S A*, 99(6):3701–3705, Mar 2002. doi: 10.1073/pnas.042700299. URL http://dx.doi.org/10.1073/pnas.042700299. (Cited on page 41.)

Susumu Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970. (Cited on page 49.)

Shujiro Okuda, Toshiaki Katayama, Shuichi Kawashima, Susumu Goto, and Minoru Kanehisa. Odb: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res*, 34 (Database issue):D358–D362, Jan 2006. doi: 10.1093/nar/gkj037. URL http://dx.doi.org/10.1093/nar/gkj037. (Cited on page 6.)

R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–2901, Mar 1999. (Cited on pages 4 and 7.)

Teresa R Pacheco, Anita Q Gomes, Nuno L Barbosa-Morais, Vladimir Benes, Wilhelm Ansorge, Matthew Wollerton, Christopher W Smith, Juan Valcárcel, and Maria Carmo-Fonseca. Diversity of vertebrate splicing factor u2af35: identification of alternatively spliced u2af1 mrnas. *J Biol Chem*, 279(26):27039–27049, Jun 2004. doi: 10.1074/jbc.M402136200. URL http://dx.doi.org/10.1074/jbc.M402136200. (Cited on page 50.)

Qun Pan, Ofer Shai, Christine Misquitta, Wen Zhang, Arneet L Saltzman, Naveed Mohammad, Tomas Babak, Henry Siu, Timothy R Hughes, Quaid D Morris, Brendan J Frey, and Benjamin J Blencowe. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell*, 16(6):929–941, Dec 2004. doi: 10.1016/j.molcel.2004.12.004. URL http://dx.doi.org/10.1016/j.molcel.2004.12.004. (Cited on page 46.)

Qun Pan, Malina A Bakowski, Quaid Morris, Wen Zhang, Brendan J Frey, Timothy R Hughes, and Benjamin J Blencowe. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet*, 21(2):73–77, Feb 2005. doi: 10.1016/j.tig.2004.12.004. URL http://dx.doi.org/10.1016/j.tig.2004.12.004. (Cited on page 56.)

Qun Pan, Arneet L Saltzman, Yoon Ki Kim, Christine Misquitta, Ofer Shai, Lynne E Maquat, Brendan J Frey, and Benjamin J Blencowe. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mrna decay to control gene expression. *Genes Dev*, 20(2): 153–158, Jan 2006. doi: 10.1101/gad.1382806. URL http://dx.doi.org/10.1101/gad.1382806. (Cited on pages 38 and 50.)

Abhijit A Patel and Joan A Steitz. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*, 4(12):960–970, Dec 2003. doi: 10.1038/nrm1259. URL http://dx.doi.org/10.1038/nrm1259. (Cited on pages 27, 28, and 29.)

M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, Apr 1999. (Cited on page 4.)

Luiz O F Penalva and Lucas Sánchez. Rna binding protein sex-lethal (sxl) and control of drosophila sex determination and dosage compensation. *Microbiol Mol Biol Rev*, 67(3):343–59, table of contents, Sep 2003. (Cited on page 31.)

Linh N Pham, Marc S Dionne, Mimi Shirasu-Hiza, and David S Schneider. A specific primed immune response in drosophila is dependent on phagocytes. *PLoS Pathog*, 3(3):e26, Mar 2007. doi: 10.1371/journal.ppat.0030026. URL http://dx.doi.org/10.1371/journal.ppat.0030026. (Cited on page 33.)

J. Piatigorsky and G. Wistow. The recruitment of crystallins: new functions precede gene duplication. *Science*, 252(5010):1078–1079, May 1991. (Cited on page 49.)

Morgan N Price, Katherine H Huang, Eric J Alm, and Adam P Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, 33(3):880–892, 2005. doi: 10.1093/nar/gki232. URL http://dx.doi.org/10.1093/nar/gki232. (Cited on page 6.)

Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–D65, Jan 2007. doi: 10.1093/nar/gkl842. URL http://dx.doi.org/10.1093/nar/gkl842. (Cited on page 3.)

Charles C Query and Maria M Konarska. Splicing fidelity revisited. *Nat Struct Mol Biol*, 13(6):472–474, Jun 2006. doi: 10.1038/nsmb0606-472. URL http://dx.doi.org/10.1038/nsmb0606-472. (Cited on page 26.)

Jeroen Raes, Eoghan Donal Harrington, Amoolya Hardev Singh, and Peer Bork. Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol*, 17(3):362–369, Jun 2007a. doi: 10.1016/j.sbi.2007.05.010. URL http://dx.doi.org/10.1016/j.sbi.2007.05.010. (Cited on pages v and 3.)

Jeroen Raes, Jan Korbel, Martin Lercher, Christian von Mering, and Peer Bork. Prediction of effective genome size in metagenomic samples. *Genome Biol*, 8(1):R10, Jan 2007b. doi: 10.1186/gb-2007-8-1-r10. URL http://dx.doi.org/10.1186/gb-2007-8-1-r10. (Cited on page 10.)

Gregory T Reeves, Cyrill B Muratov, Trudi Schüpbach, and Stanislav Y Shvartsman. Quantitative models of developmental pattern formation. *Dev Cell*, 11(3):289–300, Sep 2006. doi: 10.1016/j.devcel.2006.08.006. URL http://dx.doi.org/10.1016/j.devcel.2006.08.006. (Cited on page 1.)

Alissa Resch, Yi Xing, Alexander Alekseyenko, Barmak Modrek, and Christopher Lee. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res*, 32(4):1261–1269, 2004a. doi: 10.1093/nar/gkh284. URL http://dx.doi.org/10.1093/nar/gkh284. (Cited on pages 46 and 55.)

Alissa Resch, Yi Xing, Barmak Modrek, Michael Gorlick, Robert Riley, and Christopher Lee. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res*, 3(1):76–83, 2004b. (Cited on page 45.)

Christian S Riesenfeld, Patrick D Schloss, and Jo Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*, 38:525–552, 2004. doi: 10.1146/annurev.genet.38.072902.091216. URL http://dx.doi.org/10.1146/annurev.genet.38.072902.091216. (Cited on page 2.)

Meenakshi Roy, Qiang Xu, and Christopher Lee. Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms. *Nucleic Acids Res*, 33(16):5026–5033, 2005. doi: 10.1093/nar/gki792. URL http://dx.doi.org/10.1093/nar/gki792. (Cited on page 39.)

Jakob Lewin Rukov, Manuel Irimia, Søren Mørk, Viktor Karlovich Lund, Jeppe Vinther, and Peter Arctander. High qualitative and quantitative conservation of alternative splicing in caenorhabditis elegans and caenorhabditis briggsae. *Mol Biol Evol*, 24(4):909–917, Apr 2007. doi: 10.1093/molbev/msm023. URL http://dx.doi.org/10.1093/molbev/msm023. (Cited on pages 45, 55, and 64.)

Anthony G Russell, J. Michael Charette, David F Spencer, and Michael W Gray. An early evolutionary origin for the minor spliceosome. *Nature*, 443(7113):863–866, Oct 2006. doi: 10.1038/nature05228. URL http://dx.doi.org/10.1038/nature05228. (Cited on page 28.)

Ben M Sadd and Paul Schmid-Hempel. Insect immunity shows specificity in protection upon secondary pathogen exposure. *Curr Biol*,

16(12):1206–1210, Jun 2006. doi: 10.1016/j.cub.2006.04.047. URL http://dx.doi.org/10.1016/j.cub.2006.04.047. (Cited on page 33.)

H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. Operons in escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A*, 97(12):6652–6657, Jun 2000. doi: 10.1073/pnas.110147297. URL http://dx.doi.org/10.1073/pnas.110147297. (Cited on page 6.)

D. Schmucker, J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky. Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6): 671–684, Jun 2000. (Cited on page 25.)

P. A. Sharp. Split genes and rna splicing. *Cell*, 77(6):805–815, Jun 1994. (Cited on page 38.)

Phillip A Sharp. The discovery of split genes and rna splicing. *Trends Biochem Sci*, 30(6):279–281, Jun 2005. doi: 10.1016/j.tibs.2005.04.002. URL http://dx.doi.org/10.1016/j.tibs.2005.04.002. (Cited on page 26.)

Jay Shendure, Robi D Mitra, Chris Varma, and George M Church. Advanced sequencing technologies: methods and goals. *Nat Rev Genet*, 5(5):335–344, May 2004. doi: 10.1038/nrg1325. URL http://dx.doi.org/10.1038/nrg1325. (Cited on page 1.)

Nihar Sheth, Xavier Roca, Michelle L Hastings, Ted Roeder, Adrian R Krainer, and Ravi Sachidanandam. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*, 34(14):3955–3967, 2006. doi: 10.1093/nar/gkl556. URL http://dx.doi.org/10.1093/nar/gkl556. (Cited on page 28.)

Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005. doi: 10.1186/1471-2105-6-31. URL http://dx.doi.org/10.1186/1471-2105-6-31. (Cited on pages 57 and 58.)

C. W. Smith and J. Valcárcel. Alternative pre-mrna splicing: the logic of combinatorial control. *Trends Biochem Sci*, 25(8):381–388, Aug 2000. (Cited on pages 28, 30, and 31.)

Luis Miguel Mendes Soares, Katia Zanier, Cameron Mackereth, Michael Sattler, and Juan Valcárcel. Intron removal requires proofreading of u2af/3' splice site recognition by dek. *Science*, 312(5782):1961–1965, Jun 2006. doi: 10.1126/science.1128659. URL http://dx.doi.org/10.1126/science.1128659. (Cited on page 37.)

Rotem Sorek, Gil Ast, and Dan Graur. Alu-containing exons are alternatively spliced. *Genome Res*, 12(7):1060–1067, Jul 2002. doi: 10.1101/gr.229302. URL http://dx.doi.org/10.1101/gr.229302. (Cited on pages 47 and 48.)

Rotem Sorek, Ron Shamir, and Gil Ast. How prevalent is functional alternative splicing in the human genome? *Trends Genet*, 20(2):68–71, Feb 2004. (Cited on pages 45 and 55.)

Rotem Sorek, Gideon Dror, and Ron Shamir. Assessing the number of ancestral alternatively spliced exons in the human genome. *BMC Genomics*, 7:273, 2006. doi: 10.1186/1471-2164-7-273. URL http://dx.doi.org/10.1186/1471-2164-7-273. (Cited on page 55.)

Karpagam Srinivasan, Lily Shiue, Justin D Hayes, Ross Centers, Sean Fitzwater, Rebecca Loewen, Lillian R Edmondson, Jessica Bryant, Michael Smith, Claire Rommelfanger, Valerie Welch, Tyson A Clark, Charles W Sugnet, Kenneth J Howe, Yael Mandel-Gutfreund, and Manuel Ares. Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, 37(4):345–359, Dec 2005. doi: .09.007. URL http://dx.doi.org/.09.007. (Cited on page 39.)

Michael B Stadler, Noam Shomron, Gene W Yeo, Aniket Schneider, Xinshu Xiao, and Christopher B Burge. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet*, 2(11):e191, Nov 2006. doi: 10.1371/journal.pgen.0020191. URL http://dx.doi.org/10.1371/journal.pgen.0020191. (Cited on pages 30 and 49.)

Stefan Stamm. Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum Mol Genet*, 11(20):2409–2416, Oct 2002. (Cited on page 31.)

Viktor Stolc, Zareen Gauhar, Christopher Mason, Gabor Halasz, Marinus F van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Emilio Barbano, Harmen J Bussemaker, and Kevin P White. A gene expression map for the euchromatic genome of drosophila melanogaster. *Science*, 306(5696):655–660, Oct 2004. doi: 10.1126/science.1101312. URL http://dx.doi.org/10.1126/science.1101312. (Cited on page 41.)

Zhixi Su, Jianmin Wang, Jun Yu, Xiaoqiu Huang, and Xun Gu. Evolution of alternative splicing after gene duplication. *Genome Res*, 16(2):182–189, Feb 2006. doi: 10.1101/gr.4197006. URL http://dx.doi.org/10.1101/gr.4197006. (Cited on page 50.)

C. W. Sugnet, W. J. Kent, M. Ares, and D. Haussler. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput*, pages 66–77, 2004. (Cited on page 46.)

Mikita Suyama, Eoghan Harrington, Peer Bork, and David Torrents. Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. *PLoS Comput Biol*, 2(6):e76, Jun 2006. doi: al.pcbi.0020076. URL http://dx.doi.org/al.pcbi.0020076. (Cited on pages v and 47.)

B. E. Suzek, M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg. A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, 17(12):1123–1130, Dec 2001. (Cited on page 16.)

Radek Szklarczyk, Jaap Heringa, Sergei Kosakovsky Pond, and Anton Nekrutenko. Rapid asymmetric evolution of a dual-coding tumor suppressor ink4a/arf locus contradicts its function. *Proc Natl Acad Sci U S A*, Jul 2007. doi: 10.1073/pnas.0703238104. URL http://dx.doi.org/10.1073/pnas.0703238104. (Cited on page 50.)

David Talavera, Christine Vogel, Modesto Orozco, Sarah A Teichmann, and Xavier de la Cruz. The (in)dependence of alternative splicing and gene duplication. *PLoS Comput Biol*, 3(3):e33, Mar 2007. doi: 10.1371/journal.pcbi.0030033. URL http://dx.doi.org/10.1371/journal.pcbi.0030033. (Cited on page 50.)

Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B. Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003. doi: 10.1186/1471-2105-4-41. URL http://dx.doi.org/10.1186/1471-2105-4-41. (Cited on page 6.)

John S Taylor and Jeroen Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*, 38:615–643, 2004. doi: 10.1146/annurev.genet.38.072902.092831. URL http://dx.doi.org/10.1146/annurev.genet.38.072902.092831. (Cited on page 47.)

T. A. Thanaraj, Francis Clark, and Juha Muilu. Conservation of human alternative splice events in mouse. *Nucleic Acids Res*, 31(10):2544–2552, May 2003. (Cited on pages 44, 45, and 55.)

Vigdis Torsvik and Lise Øvreås. Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol*, 5(3):240–245, Jun 2002. (Cited on pages 2, 3, and 10.)

Susannah Green Tringe and Edward M Rubin. Metagenomics: Dna sequencing of environmental samples. *Nat Rev Genet*, 6(11):805–814, Nov 2005. doi: 10.1038/nrg1709. URL http://dx.doi.org/10.1038/nrg1709. (Cited on pages 2, 8, and 10.)

Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, Peer Bork, Philip Hugenholtz, and Edward M Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, Apr 2005. doi: 10.1126/science.1107851. URL http://dx.doi.org/10.1126/science.1107851. (Cited on pages 3, 4, 6, 8, 13, and 14.)

Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1031, Dec 2006. doi: 10.1038/nature05414. URL http://dx.doi.org/10.1038/nature05414. (Cited on page 3.)

Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, Mar 2004. doi: 10.1038/nature02340. URL http://dx.doi.org/10.1038/nature02340. AMD. (Cited on pages 3, 4, 8, and 13.)

Jernej Ule, Aljaz Ule, Joanna Spencer, Alan Williams, Jing-Shan Hu, Melissa Cline, Hui Wang, Tyson Clark, Claire Fraser, Matteo Ruggiu, Barry R Zeeberg, David Kane, John N Weinstein, John Blume, and Robert B Darnell. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet*, 37(8):844–852, Aug 2005. doi: 10.1038/ng1610. URL http://dx.doi.org/10.1038/ng1610. (Cited on pages 31 and 62.)

Jernej Ule, Giovanni Stefani, Aldo Mele, Matteo Ruggiu, Xuning Wang, Bahar Taneri, Terry Gaasterland, Benjamin J Blencowe, and Robert B Darnell. An rna map predicting nova-dependent splicing regulation. *Nature*, 444(7119):580–586, Nov 2006. doi: 10.1038/nature05304. URL http://dx.doi.org/10.1038/nature05304. (Cited on pages 29, 30, 31, 62, and 64.)

Stijn van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000. URL http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm. (Cited on page 21.)

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz,

B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291 (5507):1304–1351, Feb 2001. doi: 10.1126/science.1058040. URL http://dx.doi.org/10.1126/science.1058040. (Cited on page 39.)

J. Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667): 66–74, Apr 2004. doi: 3857. URL http://dx.doi.org/3857. (Cited on pages 3, 4, 8, 13, and 23.)

C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815): 1126–1130, Feb 2007. doi: 10.1126/science.1133420. URL http://dx.doi.org/10.1126/science.1133420. (Cited on page 10.)

Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–261, Jan 2003a. (Cited on page 19.)

Christian von Mering, Evgeny M Zdobnov, Sophia Tsoka, Francesca D Ciccarelli, Jose B Pereira-Leal, Christos A Ouzounis, and Peer Bork. Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci U S A*, 100(26):15428–15433, Dec 2003b. doi: 10.1073/pnas.2136809100. URL http://dx.doi.org/10.1073/pnas.2136809100. (Cited on page 1.)

Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–D437, Jan 2005. doi: 10.1093/nar/gki005. URL http://dx.doi.org/10.1093/nar/gki005. (Cited on pages 4, 6, and 13.)

Eleftheria Vrontou, Steven P Nilsen, Ebru Demir, Edward A Kravitz, and Barry J Dickson. fruitless regulates aggression and dominance in drosophila. *Nat Neurosci*, 9(12):1469–1471, Dec 2006. doi: 10.1038/nn1809. URL http://dx.doi.org/10.1038/nn1809. (Cited on page 31.)

Bing-Bing Wang and Volker Brendel. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci U S A*, 103(18): 7175–7180, May 2006. doi: 10.1073/pnas.0602039103. URL http://dx.doi.org/10.1073/pnas.0602039103. (Cited on pages 41 and 45.)

Zefeng Wang, Michael E Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–845, Dec 2004. doi: 10.1016/j.cell.2004.11.010. URL http://dx.doi.org/10.1016/j.cell.2004.11.010. (Cited on page 30.)

Zefeng Wang, Xinshu Xiao, Eric Van Nostrand, and Christopher B Burge. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell*, 23(1):61–70, Jul 2006. doi: 10.1016/j.molcel.2006.05.018. URL http://dx.doi.org/10.1016/j.molcel.2006.05.018. (Cited on page 30.)

Fiona L Watson, Roland Püttmann-Holgado, Franziska Thomas, David L Lamar, Michael Hughes, Masahiro Kondo, Vivienne I Rebel, and Dietmar Schmucker. Extensive diversity of ig-superfamily proteins in the immune system of insects. *Science*, 309(5742):1874–1878, Sep 2005. doi: 10.1126/science.1116887. URL http://dx.doi.org/10.1126/science.1116887. (Cited on page 32.)

Wade C Winkler and Ronald R Breaker. Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol*, 59:487–517, 2005. doi: 10.1146/annurev.micro.59.030804.121336. URL http://dx.doi.org/10.1146/annurev.micro.59.030804.121336. (Cited on pages 33 and 34.)

Woj M Wojtowicz, John J Flanagan, S. Sean Millard, S. Lawrence Zipursky, and James C Clemens. Alternative splicing of drosophila dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118(5):619–633, Sep 2004. doi: 10.1016/j.cell.2004.08.021. URL http://dx.doi.org/10.1016/j.cell.2004.08.021. (Cited on page 32.)

Tanja Woyke, Hanno Teeling, Natalia N Ivanova, Marcel Huntemann, Michael Richter, Frank Oliver Gloeckner, Dario Boffelli, Iain J Anderson, Kerrie W Barry, Harris J Shapiro, Ernest Szeto, Nikos C Kyrpides, Marc Mussmann, Rudolf Amann, Claudia Bergin, Caroline Ruehland, Edward M Rubin, and Nicole Dubilier. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950–955, Oct 2006. doi: 10.1038/nature05192. URL http://dx.doi.org/10.1038/nature05192. (Cited on page 23.)

Cathy H Wu, Rolf Apweiler, Amos Bairoch, Darren A Natale, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Raja Mazumder, Claire O'Donovan, Nicole Redaschi, and Baris Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–D191, Jan 2006. doi: 10.1093/nar/gkj161. URL http://dx.doi.org/10.1093/nar/gkj161. (Cited on page 6.)

Thomas D Wu and Colin K Watanabe. Gmap: a genomic mapping and alignment program for mrna and est sequences. *Bioinformatics*, 21 (9):1859–1875, May 2005. doi: 10.1093/bioinformatics/bti310. URL http://dx.doi.org/10.1093/bioinformatics/bti310. (Cited on page 57.)

Lei Xie and Philip E Bourne. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput Biol*, 1(3):e31, Aug 2005. doi: 10.1371/journal.pcbi.0010031. URL http://dx.doi.org/10.1371/journal.pcbi.0010031. (Cited on page 1.)

Yi Xing and Christopher Lee. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci U S A*, 102(38):13526–13531, Sep 2005a. doi: /pnas.0501213102. URL http://dx.doi.org//pnas.0501213102. (Cited on page 46.)

Yi Xing and Christopher Lee. Alternative splicing and rna selection pressure–evolutionary consequences for eukaryotic genomes. *Nat Rev Genet*, 7(7):499–509, Jul 2006. doi: 10.1038/nrg1896. URL http://dx.doi.org/10.1038/nrg1896. (Cited on pages 32 and 46.)

Yi Xing and Christopher J Lee. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet*, 1(3):e34, Sep 2005b. doi: 10.1371/journal.pgen.0010034. URL http://dx.doi.org/10.1371/journal.pgen.0010034. (Cited on page 46.)

Yi Xing, Alissa Resch, and Christopher Lee. The multiassembly problem: reconstructing multiple transcript isoforms from est fragment mixtures. *Genome Res*, 14(3):426–441, Mar 2004. doi: 10.1101/gr.1304504. URL http://dx.doi.org/10.1101/gr.1304504. (Cited on page 64.)

Yongpan Yan and John Moult. Detection of operons. *Proteins*, 64(3): 615–628, Aug 2006. doi: 10.1002/prot.21021. URL http://dx.doi.org/10.1002/prot.21021. (Cited on page 6.)

X. Y. Yang, H. Schulz, M. Elzinga, and S. Y. Yang. Nucleotide sequence of the promoter and fadb gene of the fadba operon and primary structure of the multifunctional fatty acid oxidation protein from escherichia coli. *Biochemistry*, 30(27):6788–6795, Jul 1991. (Cited on page 11.)

A. A. Yayanos. Microbiology to 10,500 meters in the deep sea. *Annu Rev Microbiol*, 49:777–805, 1995. doi: 10.1146/annurev.mi.49.100195.004021. URL http://dx.doi.org/10.1146/annurev.mi.49.100195.004021. (Cited on page 10.)

Gene W Yeo, Eric Van Nostrand, Dirk Holste, Tomaso Poggio, and Christopher B Burge. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A*, 102(8):2850–2855, Feb 2005. doi: 10.1073/pnas.0409742102. URL http://dx.doi.org/10.1073/pnas.0409742102. (Cited on pages 45, 46, 55, and 56.)

Gene W Yeo, Eric L Van Nostrand, and Tiffany Y Liang. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet*, 3(5):e85, May 2007. doi: 10.1371/journal.pgen.0030085. URL http://dx.doi.org/10.1371/journal.pgen.0030085. (Cited on page 30.)

Mihaela Zavolan and Erik van Nimwegen. The types and prevalence of alternative splice forms. *Curr Opin Struct Biol*, 16(3):362–367, Jun 2006. doi: 10.1016/j.sbi.2006.05.002. URL http://dx.doi.org/10.1016/j.sbi.2006.05.002. (Cited on page 64.)

Mihaela Zavolan, Shinji Kondo, Christian Schonbach, Jun Adachi, David A Hume, Yoshihide Hayashizaki, Terry Gaasterland, R. I. K. E. N. GER Group, and G. S. L. Members. Impact of alternative initiation, splicing, and termination on the diversity of the mrna transcripts encoded by the mouse transcriptome. *Genome Res*, 13(6B):1290–1300, Jun 2003. (Cited on page 39.)

Evgeny M Zdobnov and Peer Bork. Quantification of insect genome divergence. *Trends Genet*, 23(1):16–20, Jan 2007. doi: 10.1016/j.tig.2006.10.004. URL http://dx.doi.org/10.1016/j.tig.2006.10.004. (Cited on page 56.)

Evgeny M Zdobnov, Mónica Campillos, Eoghan D Harrington, David Torrents, and Peer Bork. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res*, 33(3):946–954, 2005. doi: 10.1093/nar/gki236. URL http://dx.doi.org/10.1093/nar/gki236. (Cited on page v.)

Xiang H-F Zhang and Lawrence A Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–1250, Jun 2004. doi: 10.1101/gad.1195304. URL http://dx.doi.org/10.1101/gad.1195304. (Cited on pages 29 and 30.)

Xiang H-F Zhang and Lawrence A Chasin. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci U S A*, 103(36):13427–13432, Sep 2006. doi: 10.1073/pnas.0603042103. URL http://dx.doi.org/10.1073/pnas.0603042103. (Cited on page 47.)

Mauro A Zordan, Michele Massironi, Maria Giovanna Ducato, Geertruy Te Kronnie, Rodolfo Costa, Carlo Reggiani, Carine Chagneau, Jean-René Martin, and Aram Megighian. Drosophila caki/cmg protein, a homolog of human cask, is essential for regulation of neurotransmitter vesicle release. *J Neurophysiol*, 94(2):1074–1083, Aug 2005. doi: 10.1152/jn.00954.2004. URL http://dx.doi.org/10.1152/jn.00954.2004. (Cited on page 62.)

Part I

APPENDIX

A

MICROBIAL TRANSCRIPT DIVERSITY: SUPPLEMENTARY DATA

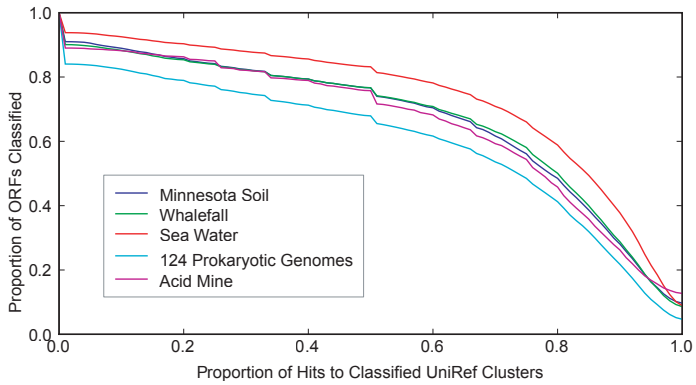


Figure A.31. Parameter exploration to decide threshold over which environmental ORFs can be considered characterized based on their hits against UniRef. The figure shows the proportion of ORFs considered 'characterized' based on the proportion of their hits in UniRef90 that are characterized. In theory, any metagenomic ORF that hits a characterized cluster could be considered characterized; however, due to false positive and negative rates of the classification method and error propagation in automatically annotated databases, we used a threshold to limit the effect of spurious annotations. ORFs were considered characterized if more than 20% of the UniRef90 clusters they hit are characterized. Other values of this parameter do not greatly affect the number of ORFs functionally characterized.

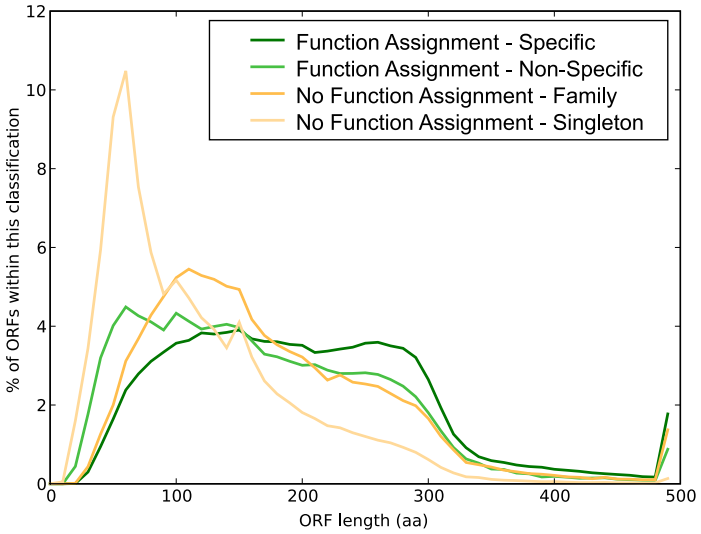


Figure A.32. Metagenomic ORFs with different functional characterizations have different length distributions. ORFs that cannot be characterized by similarity methods are significantly shorter than those that can.

					Gene calling				Functional annotation				
Pub. Yr.	**Environment (Location)**	**# ORFs (Mbp)**	**# Novel ORFs (%)**	**# COGs***	**Procedure**	**Sequence comparison algorithm**	**Parameters & cutoffs**	**Sequence DB searched**	**Procedure**	**Parameters**	**Functional DB Searched**	**Org.**	**Ref.**
2004	Acid Mine (California)	46,862 (76)	34,301 (73.2%)	1,824	FGENESB pipeline (Softberry Inc)	DBScan	1E-10, >100bp	nr	blastp against COG	manual refinement using ARTEMIS tool	COG, nr	UCB	
2004	Surface Sea Water (Sargasso Sea, samples 1-4)	1,001,987 (779)	649,608 (64.8%)	3,714	Evidence-based, using translation start & stop sites	tblastn tblastx	1E-03 1E-04	Genes: Bacterial portion of nraa rRNA: nraa	BLAST against TIGR blastn	>400p, 3 or more hits to a role category 1E-40 low-complexity	TIGR Role Category Sargasso (self-	Venter Inst.	
2005	Deep-Sea Whalefall (Pacific, Antarctic)	122,147 (75)	63,021 (51.6%)	3,332	FGENESB pipeline (Softberry Inc)	DBScan	Default parameters of software	nraa	blastp against COG, KEGG	filtering disabled, 60 bits cutoff equiv. to 1E-8	extCOG** v 6, KEGG	JGI	
2005	Farm Soil (Minnesota)	183,536 (100)	114,301 (62.3%)	3,394	FGENESB pipeline (Softberry Inc)	DBScan	Default parameters of software	nraa	blastp against COG, KEGG	low-complexity filtering disabled, 60 bits cutoff equiv. to 1E-8	extCOG** v 6, KEGG	JGI	
Total		1,354,532 1,030	861,231 (63.6%)										

* ACE/Chao1 estimates of community richness
** extCOG: extended COG database in STRING

Table A.2. Range of function prediction protocols in a sampling of metagenomics publications to date

		Genes in at least 1 Neighborhood		-> ->		-> <-		<-->	
Dataset	*Total Genes*	*Genes*	*% of total*	*Genes*	*% of total*	*Genes*	*% of total*	*Genes*	*% of total*
Whale Fall	122,147	76,456	62.59%	56,038	45.88%	13,860	11.35%	10,336	8.46%
Sea Water	1,086,400	681,651	62.74%	519,688	47.84%	111,774	10.29%	101,656	9.36%
Acid Mine	46,862	30,349	64.76%	21,434	45.74%	7,552	16.12%	7,994	17.06%
Minnesota Soil	183,536	111,154	60.56%	76,474	41.67%	21,274	11.59%	17,892	9.75%
Environments Combined	1,438,945	899,610	62.52%	673,634	46.81%	154,460	10.73%	137,878	9.58%
124 Prokaryotic Genomes	344,619	343,584	99.70%	303,968	88.20%	102,668	29.79%	102,597	29.77%

Table A.3. Neighborhood information available for each of the datasets analyzed.

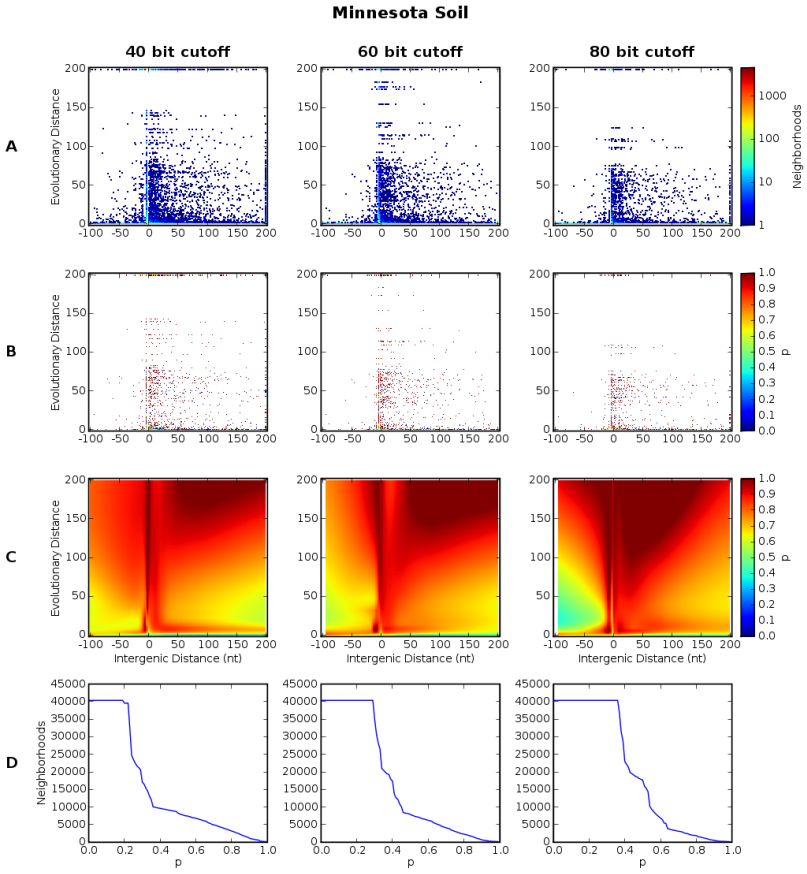


Figure A.33. Neighborhood method applied to Minnesota Soil data at 3 different bitscore cutoffs. Each column shows the method applied at a different bitscore cutoff, affecting the detection of conserved neighborhoods and the stringency of the KEGG mapping used for the benchmark dataset. Row A shows a 2-dimensional histogram of the all the codirectionally transcribed neighborhoods in the dataset, binned on the x-axis by intergenic distance and on the y-axis by evolutionary distance (see Supp Info for full description). Row B shows the benchmark data, at each intergenic and evolutionary distance p (the proportion of neighborhoods where both genes are functionally related) is shown. Row C shows the interpolation of the data in row B. Row D shows the proportion of neighborhoods with p greater than the cutoff on the x-axis using the predictions from the interpolation in row C.

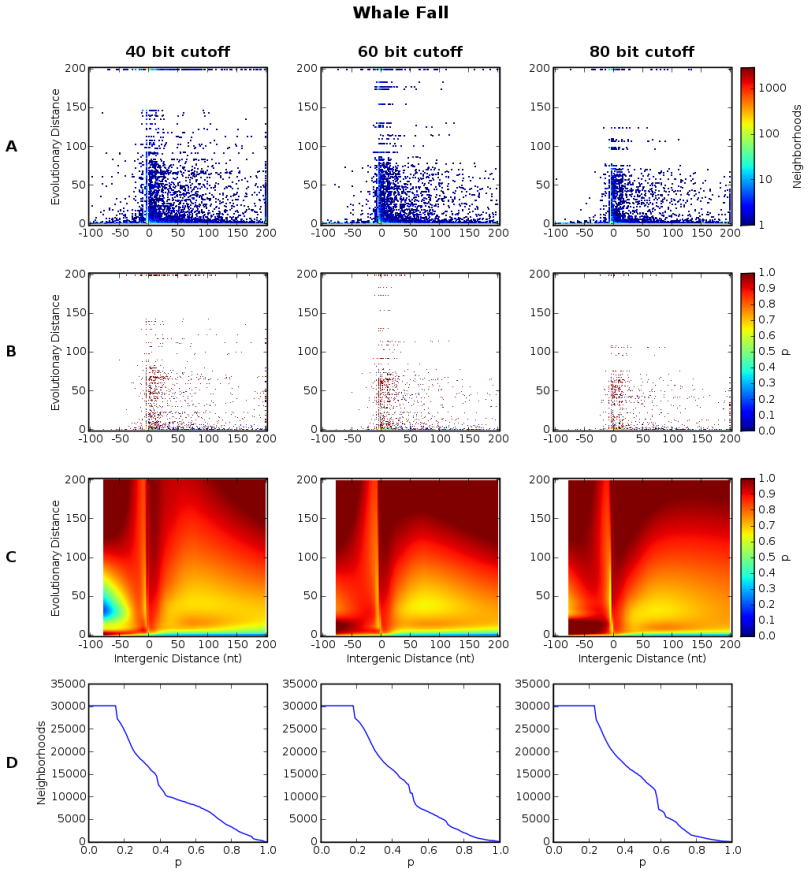


Figure A.34. Neighborhood method applied to Whale Fall data at 3 different bitscore cutoffs. Each column shows the method applied at a different bitscore cutoff, affecting the detection of conserved neighborhoods and the stringency of the KEGG mapping used for the benchmark dataset. Row A shows a 2-dimensional histogram of the all the codirectionally transcribed neighborhoods in the dataset, binned on the x-axis by intergenic distance and on the y-axis by evolutionary distance (see Supp Info for full description). Row B shows the benchmark data, at each intergenic and evolutionary distance p (the proportion of neighborhoods where both genes are functionally related) is shown. Row C shows the interpolation of the data in row B. Row D shows the proportion of neighborhoods with p greater than the cutoff on the x-axis using the predictions from the interpolation in row C.

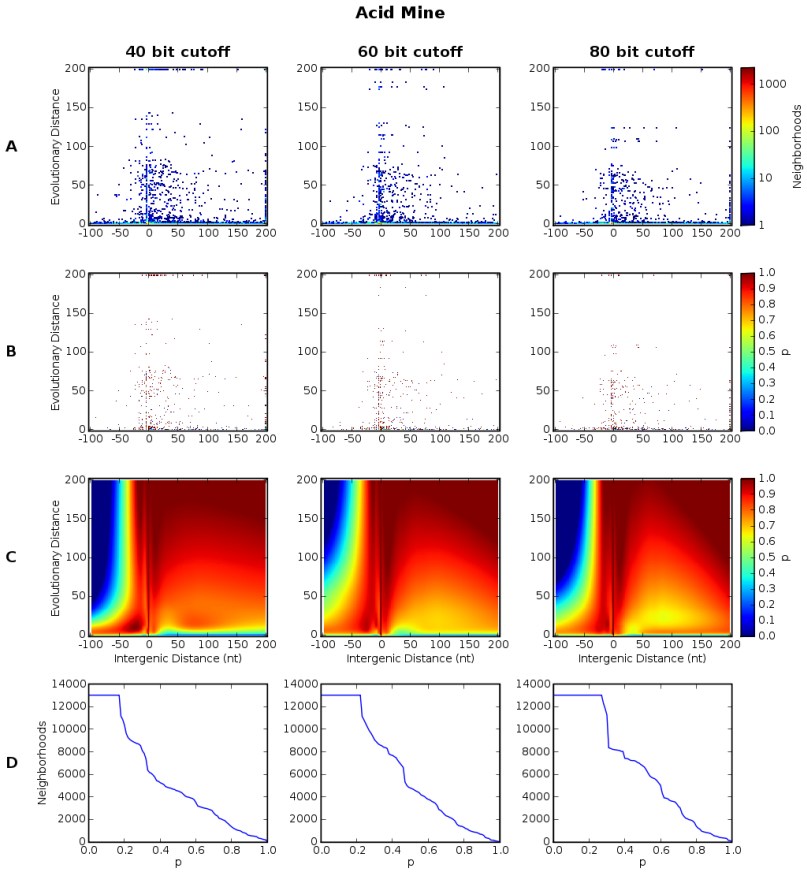
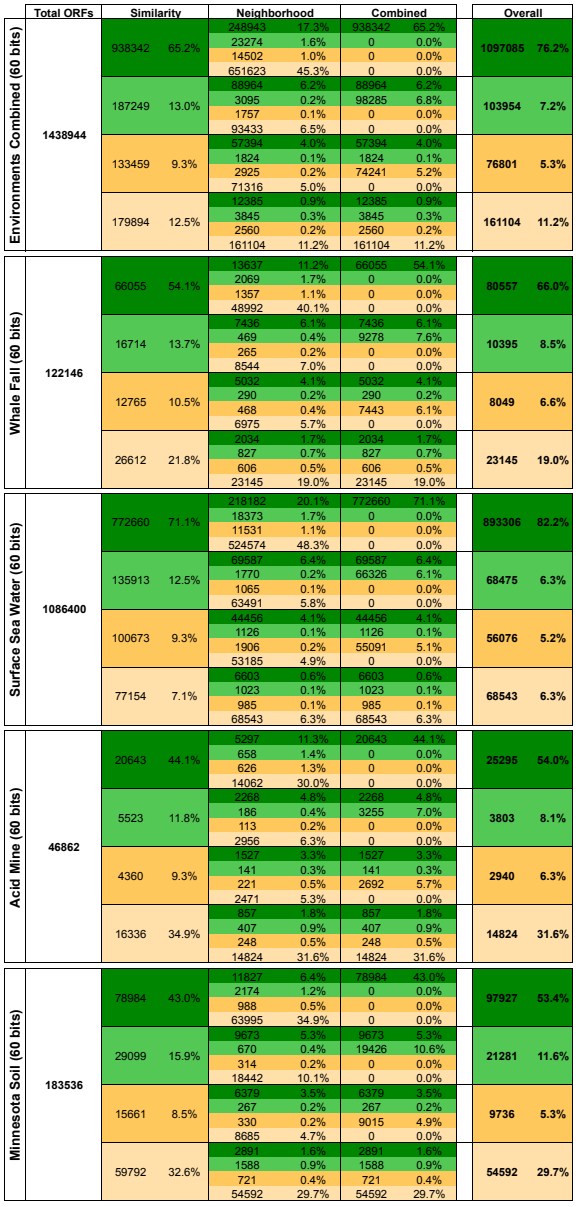


Figure A.35. Neighborhood method applied to Acid Mine data at 3 different bitscore cutoffs. Each column shows the method applied at a different bitscore cutoff, affecting the detection of conserved neighborhoods and the stringency of the KEGG mapping used for the benchmark dataset. Row A shows a 2-dimensional histogram of the all the codirectionally transcribed neighborhoods in the dataset, binned on the x-axis by intergenic distance and on the y-axis by evolutionary distance (see Supp Info for full description). Row B shows the benchmark data, at each intergenic and evolutionary distance p (the proportion of neighborhoods where both genes are functionally related) is shown. Row C shows the interpolation of the data in row B. Row D shows the proportion of neighborhoods with p greater than the cutoff on the x-axis using the predictions from the interpolation in row C.



	Total ORFs	Similarity		Neighborhood		Combined		Overall	
Environments Combined (60 bits)	1438944	938342	65.2%	248943	17.3%	938342	65.2%	1097085	76.2%
				23274	1.6%	0	0.0%		
				14502	1.0%	0	0.0%		
				651623	45.3%	0	0.0%		
		187249	13.0%	88964	6.2%	88964	6.2%	103954	7.2%
				3095	0.2%	98285	6.8%		
				1757	0.1%	0	0.0%		
				93433	6.5%	0	0.0%		
		133459	9.3%	57394	4.0%	57394	4.0%	76801	5.3%
				1824	0.1%	1824	0.1%		
				2925	0.2%	74241	5.2%		
				71316	5.0%	0	0.0%		
		179894	12.5%	12385	0.9%	12385	0.9%	161104	11.2%
				3845	0.3%	3845	0.3%		
				2560	0.2%	2560	0.2%		
				161104	11.2%	161104	11.2%		
Whale Fall (60 bits)	122146	66055	54.1%	13637	11.2%	66055	54.1%	80557	66.0%
				2069	1.7%	0	0.0%		
				1357	1.1%	0	0.0%		
				48992	40.1%	0	0.0%		
		16714	13.7%	7436	6.1%	7436	6.1%	10395	8.5%
				469	0.4%	9278	7.6%		
				265	0.2%	0	0.0%		
				8544	7.0%	0	0.0%		
		12765	10.5%	5032	4.1%	5032	4.1%	8049	6.6%
				290	0.2%	290	0.2%		
				468	0.4%	7443	6.1%		
				6975	5.7%	0	0.0%		
		26612	21.8%	2034	1.7%	2034	1.7%	23145	19.0%
				827	0.7%	827	0.7%		
				606	0.5%	606	0.5%		
				23145	19.0%	23145	19.0%		
Surface Sea Water (60 bits)	1086400	772660	71.1%	218182	20.1%	772660	71.1%	893306	82.2%
				18373	1.7%	0	0.0%		
				11531	1.1%	0	0.0%		
				524574	48.3%	0	0.0%		
		135913	12.5%	69587	6.4%	69587	6.4%	68475	6.3%
				1770	0.2%	66326	6.1%		
				1065	0.1%	0	0.0%		
				63491	5.8%	0	0.0%		
		100673	9.3%	44456	4.1%	44456	4.1%	56076	5.2%
				1126	0.1%	1126	0.1%		
				1906	0.2%	55091	5.1%		
				53185	4.9%	0	0.0%		
		77154	7.1%	6603	0.6%	6603	0.6%	68543	6.3%
				1023	0.1%	1023	0.1%		
				985	0.1%	985	0.1%		
				68543	6.3%	68543	6.3%		
Acid Mine (60 bits)	46862	20643	44.1%	5297	11.3%	20643	44.1%	25295	54.0%
				658	1.4%	0	0.0%		
				626	1.3%	0	0.0%		
				14062	30.0%	0	0.0%		
		5523	11.8%	2268	4.8%	2268	4.8%	3803	8.1%
				186	0.4%	3255	7.0%		
				113	0.2%	0	0.0%		
				2956	6.3%	0	0.0%		
		4360	9.3%	1527	3.3%	1527	3.3%	2940	6.3%
				141	0.3%	141	0.3%		
				221	0.5%	2692	5.7%		
				2471	5.3%	0	0.0%		
		16336	34.9%	857	1.8%	857	1.8%	14824	31.6%
				407	0.9%	407	0.9%		
				248	0.5%	248	0.5%		
				14824	31.6%	14824	31.6%		
Minnesota Soil (60 bits)	183536	78984	43.0%	11827	6.4%	78984	43.0%	97927	53.4%
				2174	1.2%	0	0.0%		
				988	0.5%	0	0.0%		
				63995	34.9%	0	0.0%		
		29099	15.9%	9673	5.3%	9673	5.3%	21281	11.6%
				670	0.4%	19426	10.6%		
				314	0.2%	0	0.0%		
				18442	10.1%	0	0.0%		
		15661	8.5%	6379	3.5%	6379	3.5%	9736	5.3%
				267	0.2%	267	0.2%		
				330	0.2%	9015	4.9%		
				8685	4.7%	0	0.0%		
		59792	32.6%	2891	1.6%	2891	1.6%	54592	29.7%
				1588	0.9%	1588	0.9%		
				721	0.4%	721	0.4%		
				54592	29.7%	54592	29.7%		

Table A.4. Metagenomic data in Figure 2.4 and Figure 2.10. Moving from top to bottom is equivalent to moving clockwise around the pie chart and moving from left to right across the table is equivalent to moving from the inner pie to outside ring in in Figure 2.10. A description of the colors used and the corresponding categories can be found in Figure 2.4.

	Total ORFs	Similarity		Neighborhood		Combined		Overall	
124 Prokaryotic Genomes	344619	223296	64.8%	119078	34.6%	223296	64.8%	286225	83.1%
				7337	2.1%	0	0.0%		
				14594	4.2%	0	0.0%		
				82287	23.9%	0	0.0%		
		32703	9.5%	26016	7.6%	26016	7.6%	9773	2.8%
				810	0.2%	6687	1.9%		
				1197	0.4%	0	0.0%		
				4680	1.4%	0	0.0%		
		87653	25.4%	36779	10.7%	36779	10.7%	47919	13.9%
				3057	0.9%	3057	0.9%		
				10382	3.0%	47817	13.9%		
				37435	10.9%	0	0.0%		
		967	0.3%	134	0.0%	134	0.0%	702	0.2%
				29	0.0%	29	0.0%		
				102	0.0%	102	0.0%		
				702	0.2%	702	0.2%		
Mycoplasma pneumoniae	687	410	59.7%	231	33.6%	410	59.7%	520	75.7%
				24	3.5%	0	0.0%		
				30	4.4%	0	0.0%		
				125	18.2%	0	0.0%		
		107	15.6%	58	8.4%	58	8.4%	59	8.6%
				20	2.9%	49	7.1%		
				3	0.4%	0	0.0%		
				26	3.8%	0	0.0%		
		170	24.8%	52	7.6%	52	7.6%	108	15.7%
				10	1.5%	10	1.5%		
				39	5.7%	108	15.7%		
				69	10.0%	0	0.0%		
		0	0.0%	0	0.0%	0	0.0%	0	0.0%
				0	0.0%	0	0.0%		
				0	0.0%	0	0.0%		
				0	0.0%	0	0.0%		
Escherichia coli K12	4305	3119	72.5%	2054	47.7%	3119	72.5%	3907	90.8%
				94	2.2%	0	0.0%		
				203	4.7%	0	0.0%		
				768	17.8%	0	0.0%		
		355	8.3%	299	7.0%	299	7.0%	94	2.2%
				5	0.1%	56	1.3%		
				12	0.3%	0	0.0%		
				39	0.9%	0	0.0%		
		815	18.9%	488	11.3%	488	11.3%	293	6.8%
				37	0.9%	37	0.9%		
				94	2.2%	292	6.8%		
				198	4.6%	0	0.0%		
		16	0.4%	3	0.1%	3	0.1%	11	0.3%
				1	0.0%	1	0.0%		
				1	0.0%	1	0.0%		
				11	0.3%	11	0.3%		
Archaeoglobus fulgidus	2396	1401	58.5%	768	32.1%	1401	58.5%	1957	81.7%
				92	3.8%	0	0.0%		
				102	4.3%	0	0.0%		
				439	18.3%	0	0.0%		
		359	15.0%	292	12.2%	292	12.2%	124	5.2%
				11	0.5%	67	2.8%		
				18	0.8%	0	0.0%		
				38	1.6%	0	0.0%		
		625	26.1%	261	10.9%	261	10.9%	310	12.9%
				57	2.4%	57	2.4%		
				93	3.9%	307	12.8%		
				214	8.9%	0	0.0%		
		11	0.5%	3	0.1%	3	0.1%	5	0.2%
				0	0.0%	0	0.0%		
				3	0.1%	3	0.1%		
				5	0.2%	5	0.2%		
Prochlorococcus marinus SS120	1881	1074	57.1%	356	18.9%	1074	57.1%	1342	71.4%
				15	0.8%	0	0.0%		
				55	2.9%	0	0.0%		
				648	34.5%	0	0.0%		
		141	7.5%	110	5.9%	110	5.9%	42	2.2%
				2	0.1%	31	1.7%		
				2	0.1%	0	0.0%		
				27	1.4%	0	0.0%		
		661	35.1%	158	8.4%	158	8.4%	492	26.2%
				11	0.6%	11	0.6%		
				27	1.4%	492	26.2%		
				465	24.7%	0	0.0%		
		5	0.3%	0	0.0%	0	0.0%	5	0.3%
				0	0.0%	0	0.0%		
				0	0.0%	0	0.0%		
				5	0.3%	5	0.3%		

Table A.5. The 124 prokaryotic species from the STRING database used in this analysis. Moving from top to bottom is equivalent to moving clockwise around the pie chart and moving from left to right across the table is equivalent to moving from the inner pie to outside ring in in Figure 2.10. A description of the colors used and the corresponding categories can be found in Figure 2.4

Bin	Mean Fam. Size	Fn. Characterized #	Fn. Characterized % of bin	Fn. Uncharacterized #	Fn. Uncharacterized % of bin	Total in Bin
1	517.36	1,909	95%	91	5%	2,000
2	47.96	1,665	83%	335	17%	2,000
3	18.88	1,523	76%	477	24%	2,000
4	10.90	1,431	72%	569	28%	2,000
5	7.10	1,299	65%	701	35%	2,000
6	5.28	1,363	68%	637	32%	2,000
7	4.05	1,281	64%	719	36%	2,000
8	3.26	1,188	59%	812	41%	2,000
9	3.00	1,148	57%	852	43%	2,000
10	2.51	1,043	52%	957	48%	2,000
11	2.00	899	45%	1,101	55%	2,000
12	2.00	884	44%	1,116	56%	2,000
13	2.00	900	45%	1,100	55%	2,000
14	2.00	927	46%	1,073	54%	2,000
15	2.00	914	46%	1,086	54%	2,000
16	2.00	903	45%	1,097	55%	2,000
17	1.05	568	28%	1,432	72%	2,000
Sample singleton bin		533	27%	1,467	73%	2,000
All singletons		47,394	27%	126,730	73%	174,124
Total		**66,715**	32%	**139,502**	68%	**206,217**

Table A.6. Data in Figure 2.11.

KEGG id	Description	Freq Meta	Freq String	p-value
2030	Bacterial chemotaxis	3.955E-03	1.028E-02	0.000E+00
2040	Flagellar assembly	7.637E-03	2.568E-02	0.000E+00
3010	Ribosome	5.441E-02	9.569E-02	0.000E+00
3070	Type III secretion system	2.154E-03	1.157E-02	0.000E+00
3080	Type IV secretion system	7.391E-04	3.904E-03	0.000E+00
3090	Type II secretion system	7.138E-03	1.750E-02	0.000E+00
860	Porphyrin and chlorophyll metabolism	1.907E-02	2.967E-02	8.228E-74
290	Valine, leucine and isoleucine biosynthesis	3.706E-02	2.348E-02	1.494E-71
260	Glycine, serine and threonine metabolism	5.683E-02	4.052E-02	6.741E-68
640	Propanoate metabolism	4.272E-02	2.914E-02	5.446E-62
380	Tryptophan metabolism	3.258E-02	2.088E-02	4.470E-60
780	Biotin metabolism	4.589E-03	9.323E-03	7.939E-57
512	O-Glycan biosynthesis	4.855E-03	1.044E-03	4.546E-44
533	Keratan sulfate biosyntheses	4.844E-03	1.044E-03	6.537E-44
602	Glycosphingolipid biosynthesis - neo-lactoseries	6.187E-03	2.059E-03	6.292E-40
604	Glycosphingolipid biosynthesis - ganglioseries	6.140E-03	2.030E-03	7.152E-40
30	Pentose phosphate pathway	2.067E-02	2.823E-02	2.144E-35
903	Limonene and pinene degradation	2.380E-02	1.609E-02	3.357E-35
71	Fatty acid metabolism	2.882E-02	2.035E-02	4.739E-35
195	Photosynthesis	1.384E-03	3.432E-03	3.694E-33
251	Glutamate metabolism	4.090E-02	3.114E-02	5.939E-33
790	Folate biosynthesis	1.190E-02	1.744E-02	2.363E-32
510	N-Glycan biosynthesis	8.634E-03	4.218E-03	4.657E-32
53	Ascorbate and aldarate metabolism	1.294E-02	7.650E-03	2.048E-30
120	Bile acid biosynthesis	2.078E-02	1.407E-02	2.313E-30
1052	Type I polyketide structures	2.010E-03	0.000E+00	2.551E-30
280	Valine, leucine and isoleucine degradation	3.329E-02	2.482E-02	2.649E-30
600	Sphingolipid metabolism	1.467E-02	9.065E-03	4.607E-30
4010	MAPK signaling pathway	1.862E-03	0.000E+00	4.811E-28
730	Thiamine metabolism	3.349E-03	6.134E-03	2.087E-27
310	Lysine degradation	3.426E-02	2.611E-02	3.464E-27
410	beta-Alanine metabolism	1.918E-02	1.318E-02	3.607E-26
500	Starch and sucrose metabolism	2.340E-02	3.016E-02	3.073E-25
603	Glycosphingolipid biosynthesis - globoseries	7.753E-03	4.032E-03	3.808E-25
3020	RNA polymerase	1.353E-02	8.679E-03	3.103E-24
40	Pentose and glucuronate interconversions	9.135E-03	1.323E-02	7.504E-23
20	Citrate cycle (TCA cycle)	3.275E-02	2.549E-02	2.027E-22
1051	Biosynthesis of ansamycins	1.489E-03	0.000E+00	2.562E-22
193	ATP synthesis	1.120E-02	1.554E-02	3.857E-21
3060	Protein export	1.851E-02	2.392E-02	2.633E-20
641	3-Chloroacrylic acid degradation	1.526E-03	1.001E-04	9.910E-20
51	Fructose and mannose metabolism	2.660E-02	3.280E-02	7.591E-19
970	Aminoacyl-tRNA biosynthesis	4.741E-02	3.949E-02	1.557E-18
1054	Nonribosomal peptide structures	1.158E-03	0.000E+00	3.213E-17
473	D-Alanine metabolism	2.257E-03	4.061E-03	6.060E-17
625	Tetrachloroethene degradation	5.132E-03	2.688E-03	3.402E-16
631	1,2-Dichloroethane degradation	2.259E-03	6.720E-04	6.538E-16
630	Glyoxylate and dicarboxylate metabolism	2.485E-02	1.949E-02	7.399E-16
720	Reductive carboxylate cycle (CO2 fixation)	2.923E-02	2.346E-02	1.231E-15
620	Pyruvate metabolism	5.180E-02	4.434E-02	5.271E-15
950	Alkaloid biosynthesis I	5.769E-03	3.274E-03	6.729E-15
540	Lipopolysaccharide biosynthesis	1.525E-02	1.127E-02	3.379E-14
523	Polyketide sugar unit biosynthesis	7.945E-03	5.105E-03	6.797E-14
632	Benzoate degradation via CoA ligation	3.191E-02	2.644E-02	7.740E-13
562	Inositol phosphate metabolism	5.924E-03	3.575E-03	9.599E-13
650	Butanoate metabolism	4.746E-02	4.085E-02	1.009E-12
230	Purine metabolism	8.719E-02	7.848E-02	1.830E-12
740	Riboflavin metabolism	8.028E-03	1.082E-02	4.236E-12
130	Ubiquinone biosynthesis	3.868E-02	3.290E-02	8.390E-12
660	C5-Branched dibasic acid metabolism	9.052E-03	6.306E-03	2.963E-11
520	Nucleotide sugars metabolism	2.141E-02	1.720E-02	3.621E-11
61	Fatty acid biosynthesis	2.062E-02	1.657E-02	1.058E-10
480	Glutathione metabolism	1.167E-02	8.636E-03	1.293E-10
791	Atrazine degradation	1.206E-03	2.860E-04	7.251E-10
271	Methionine metabolism	1.818E-02	1.454E-02	9.308E-10
770	Pantothenate and CoA biosynthesis	2.358E-02	1.953E-02	2.651E-09
450	Selenoamino acid metabolism	2.369E-02	1.965E-02	3.053E-09
240	Pyrimidine metabolism	7.160E-02	6.484E-02	7.310E-09
580	Phospholipid degradation	8.743E-04	1.702E-03	1.000E-08
220	Urea cycle and metabolism of amino groups	2.238E-02	1.859E-02	1.560E-08
31	Inositol metabolism	4.271E-03	2.660E-03	4.244E-08
62	Fatty acid elongation in mitochondria	1.161E-02	8.965E-03	6.304E-08
960	Alkaloid biosynthesis II	8.721E-05	5.005E-04	9.908E-08
52	Galactose metabolism	1.636E-02	1.945E-02	2.892E-07
312	beta-Lactam resistance	4.360E-06	1.716E-04	3.286E-07
563	Glycosylphosphatidylinositol(GPI)-anchor biosynthes	4.949E-04	0.000E+00	5.856E-07
710	Carbon fixation	2.803E-02	2.458E-02	1.803E-05
4110	Cell cycle	2.420E-04	0.000E+00	2.345E-05
601	Glycosphingolipid biosynthesis - lactoseries	6.519E-04	1.430E-04	2.730E-05
590	Arachidonic acid metabolism	6.584E-04	1.573E-04	4.886E-05
4120	Ubiquitin mediated proteolysis	2.289E-04	0.000E+00	5.496E-05
940	Stilbene, coumarine and lignin biosynthesis	5.621E-03	4.132E-03	5.619E-05
471	D-Glutamine and D-glutamate metabolism	3.789E-03	5.047E-03	7.600E-05
561	Glycerolipid metabolism	3.732E-02	4.099E-02	1.750E-04
430	Taurine and hypotaurine metabolism	2.871E-03	3.889E-03	4.518E-04
4210	Apoptosis	1.940E-04	0.000E+00	5.326E-04
300	Lysine biosynthesis	2.518E-02	2.803E-02	7.956E-04

Table A.7. KEGG maps over-represented in Environmental Datasets relative to fully sequenced genomes

Ocurrences	COG A	COG B	COG A Description	COG B Description
171	COG1077	COG0119	Actin-like ATPase involved in cell morphogenesis	Isopropylmalate/homocitrate/citramalate synthases
161	COG0015	COG0024	Adenylosuccinate lyase	Methionine aminopeptidase
153	COG0190	COG0762	5,10-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase	Predicted integral membrane protein
149	COG0072	COG0481	Phenylalanyl-tRNA synthetase beta subunit	Membrane GTPase LepA
144	COG0527	COG0821	Aspartokinases	Enzyme involved in the deoxyxylulose pathway of isoprenoid biosynthesis
143	COG2086	COG0404	Electron transfer flavoprotein, beta subunit	Glycine cleavage system T protein (aminomethyltransferase)
141	COG0141	COG0361	Histidinol dehydrogenase	Translation initiation factor 1 (IF-1)
139	COG0008	COG1974	Glutamyl- and glutaminyl-tRNA synthetases	SOS-response transcriptional repressors (RecA-mediated autopeptidases)
134	COG0289	COG0177	Dihydrodipicolinate reductase	Predicted EndoIII-related endonuclease
133	COG0192	COG0815	S-adenosylmethionine synthetase	Apolipoprotein N-acyltransferase
129	COG0119	COG0215	Isopropylmalate/homocitrate/citramalate	Cysteinyl-tRNA synthetase
128	COG0524	COG0177	Sugar kinases, ribokinase family	Predicted EndoIII-related endonuclease
126	COG0483	COG0254	Archaeal fructose-1,6-bisphosphatase and related enzymes of inositol monophosphatase	Ribosomal protein L31
124	COG0188	COG0629	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), A subunit	Single-stranded DNA-binding protein
123	COG0254	COG3820	Ribosomal protein L31	Uncharacterized protein conserved in bacteria
123	COG0192	COG0195	S-adenosylmethionine synthetase	Transcription elongation factor
122	COG0477	COG0168	Permeases of the major facilitator superfamily	Trk-type K+ transport systems, membrane components
121	COG0349	COG0794	Ribonuclease D	Predicted sugar phosphate isomerase involved in capsule formation
121	COG0206	COG2001	Cell division GTPase	Uncharacterized protein conserved in bacteria
120	COG1304	COG1304	L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid	L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid
120	COG0756	COG0476	dUTPase	Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis
119	COG0167	COG1384	Dihydroorotate dehydrogenase	Lysyl-tRNA synthetase (class I)
117	COG0668	COG0762	Small-conductance mechanosensitive	Predicted integral membrane protein
117	COG0499	COG0802	S-adenosylhomocysteine hydrolase	Predicted ATPase or kinase
116	COG0604	COG1741	NADPH:quinone reductase and related Zn-dependent oxidoreductases	Pirin-related protein
115	COG0462	COG1496	Phosphoribosylpyrophosphate synthetase	Uncharacterized conserved protein
115	COG0408	COG0782	Coproporphyrinogen III oxidase	Transcription elongation factor
114	COG0400	COG1403	Predicted esterase	Restriction endonuclease
114	COG0104	COG0351	Adenylosuccinate synthase	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
113	COG0180	COG0694	Tryptophanyl-tRNA synthetase	Thioredoxin-like proteins and domains
112	COG0225	COG2897	Peptide methionine sulfoxide reductase	Rhodanese-related sulfurtransferase
112	COG0408	COG0219	Coproporphyrinogen III oxidase	Predicted rRNA methylase (SpoU class)
112	COG2009	COG0136	Succinate dehydrogenase/fumarate reductase, cytochrome b subunit	Aspartate-semialdehyde dehydrogenase
109	COG0119	COG0565	Isopropylmalate/homocitrate/citramalate synthases	rRNA methylase
108	COG0793	COG0799	Periplasmic protease	Uncharacterized homolog of plant lojap
108	COG0717	COG0766	Deoxycytidine deaminase	UDP-N-acetylglucosamine enolpyruvyl
107	COG1304	COG0436	L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid	Aspartate/tyrosine/aromatic aminotransferase
105	COG0343	COG0652	Queuine/archaeosine tRNA-ribosyltransferase	Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family
104	COG0289	COG0484	Dihydrodipicolinate reductase	DnaJ-class molecular chaperone with C-terminal Zn finger domain
102	COG0492	COG0625	Thioredoxin reductase	Glutathione S-transferase
101	COG0686	COG0161	Alanine dehydrogenase	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
101	COG0006	COG3473	Xaa-Pro aminopeptidase	Maleate cis-trans isomerase
100	COG0621	COG0735	2-methylthioadenine synthetase	Fe2+/Zn2+ uptake regulation proteins
100	COG1109	COG0351	Phosphomannomutase	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
98	COG0461	COG0854	Orotate phosphoribosyltransferase	Pyridoxal phosphate biosynthesis protein
97	COG0821	COG0216	Enzyme involved in the deoxyxylulose pathway of isoprenoid biosynthesis	Protein chain release factor A
96	COG0077	COG2812	Prephenate dehydratase	DNA polymerase III, gamma/tau subunits
95	COG0548	COG0706	Acetylglutamate kinase	Preprotein translocase subunit YidC
95	COG0479	COG1012	Succinate dehydrogenase/fumarate reductase, Fe-S protein subunit	NAD-dependent aldehyde dehydrogenases
94	COG1028	COG0861	Dehydrogenases with different specificities (related to short-chain alcohol	Membrane protein TerC, possibly involved in tellurium resistance
94	COG0284	COG0776	Orotidine-5-phosphate decarboxylase	Bacterial nucleoid DNA-binding protein
94	COG0134	COG1974	Indole-3-glycerol phosphate synthase	SOS-response transcriptional repressors (RecA-mediated autopeptidases)
93	COG0129	COG0525	Dihydroxyacid dehydratase/phosphogluconate dehydratase	Valyl-tRNA synthetase
92	COG0324	COG0265	tRNA delta(2)-isopentenylpyrophosphate transferase	Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain
91	COG0604	COG0302	NADPH:quinone reductase and related Zn-dependent oxidoreductases	GTP cyclohydrolase I
91	COG0128	COG1137	5-enolpyruvylshikimate-3-phosphate synthase	ABC-type (unclassified) transport system, ATPase component
91	COG0006	COG0210	Xaa-Pro aminopeptidase	Superfamily I DNA and RNA helicases
89	COG1003	COG2907	Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain	Predicted NAD/FAD-binding protein
89	COG2303	COG0861	Choline dehydrogenase and related flavoproteins	Membrane protein TerC, possibly involved in tellurium resistance
88	COG0382	COG1132	4-hydroxybenzoate polyprenyltransferase and related prenyltransferases	ABC-type multidrug transport system, ATPase and permease components

Table A.8. The 60 Most frequently occurring COG neighborhoods unique to metagenomic datasets

NCBI Taxid	Species Name
139	Borrelia burgdorferi
158	Treponema denticola
160	Treponema pallidum
197	Campylobacter jejuni
210	Helicobacter pylori 26695
305	Ralstonia solanacearum
340	Xanthomonas campestris
375	Bradyrhizobium japonicum
381	Rhizobium loti
382	Rhizobium meliloti
491	Neisseria meningitidis B
562	Escherichia coli K12
601	Salmonella typhi
602	Salmonella typhimurium
623	Shigella flexneri 2a 301
632	Yersinia pestis CO92
666	Vibrio cholerae
670	Vibrio parahaemolyticus
727	Haemophilus influenzae
747	Pasteurella multocida
781	Rickettsia conorii
782	Rickettsia prowazekii
813	Chlamydia trachomatis
882	Desulfovibrio vulgaris
959	Bdellovibrio bacteriovorus
1076	Rhodopseudomonas palustris
1097	Chlorobium tepidum
1148	Synechocystis sp. PCC6803
1219	Prochlorococcus marinus SS120
1282	Staphylococcus epidermidis
1299	Deinococcus radiodurans
1309	Streptococcus mutans
1313	Streptococcus pneumoniae TIGR4
1314	Streptococcus pyogenes M1
1360	Lactococcus lactis
1423	Bacillus subtilis
1488	Clostridium acetobutylicum
1502	Clostridium perfringens
1513	Clostridium tetani
1590	Lactobacillus plantarum
1639	Listeria monocytogenes EGD
1642	Listeria innocua
1717	Corynebacterium diphtheriae
1718	Corynebacterium glutamicum
1765	Mycobacterium bovis
1769	Mycobacterium leprae
1770	Mycobacterium paratuberculosis
1902	Streptomyces coelicolor
2096	Mycoplasma gallisepticum
2097	Mycoplasma genitalium
2104	Mycoplasma pneumoniae
2107	Mycoplasma pulmonis
2190	Methanococcus jannaschii
2214	Methanosarcina acetivorans
2234	Archaeoglobus fulgidus
198466	Streptococcus pyogenes MGAS315
203907	Blochmannia floridanus
216466	Streptococcus agalactiae V
216495	Streptococcus agalactiae III
217992	Escherichia coli O6
222523	Bacillus cereus ATCC 10987
229193	Yersinia pestis Medievalis

Taxid	Species Name
2261	Pyrococcus furiosus
2287	Sulfolobus solfataricus
2303	Thermoplasma acidophilum
2320	Methanopyrus kandleri
2336	Thermotoga maritima
2371	Xylella fastidiosa 9a5c
13773	Pyrobaculum aerophilum
28227	Mycoplasma penetrans
29292	Pyrococcus abyssi
29459	Brucella melitensis
32046	Synechococcus elongatus
33072	Gloeobacter violaceus
33959	Lactobacillus johnsonii
35554	Geobacter sulfurreducens
36870	Wigglesworthia brevipalpis
39152	Methanococcus maripaludis
44101	Mycoplasma mycoides
44275	Leptospira interrogans L1-130
50339	Thermoplasma volcanium
53953	Pyrococcus horikoshii
56636	Aeropyrum pernix
59919	Prochlorococcus marinus CCMP1378
63363	Aquifex aeolicus
65699	Neisseria meningitidis A
66077	Wolbachia sp. wMel
70863	Shewanella oneidensis
76856	Fusobacterium nucleatum
83331	Mycobacterium tuberculosis CDC1551
83334	Escherichia coli O157:H7
83557	Chlamydophila caviae
83560	Chlamydia muridarum
84588	Synechococcus sp. WH8102
85963	Helicobacter pylori J99
86665	Bacillus halodurans
92829	Xanthomonas axonopodis
98794	Buchnera aphidicola Sg
100379	Phytoplasma Onion yellows
103690	Nostoc sp. PCC 7120
111955	Sulfolobus tokodaii
115711	Chlamydia pneumoniae AR39
115713	Chlamydia pneumoniae CWL029
118099	Buchnera aphidicola APS
119072	Thermoanaerobacter tengcongensis
134821	Ureaplasma parvum
135842	Buchnera aphidicola Bp
155864	Escherichia coli EDL933
155892	Caulobacter crescentus
158879	Staphylococcus aureus N315
160232	Nanoarchaeum equitans
171101	Streptococcus pneumoniae R6
180835	Agrobacterium tumefaciens WashU
181661	Agrobacterium tumefaciens Cereon
182082	Chlamydophila pneumoniae TW183
182710	Oceanobacillus iheyensis
186103	Streptococcus pyogenes MGAS8232
187410	Yersinia pestis KIM
193567	Streptococcus pyogenes SSI-1
196600	Vibrio vulnificus YJ016
196620	Staphylococcus aureus MW2
196627	Corynebacterium glutamicum 13032
198094	Bacillus anthracis
262724	Thermus thermophilus

Table A.9. The 124 prokaryotic species from the STRING database used in this analysis