

The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes[§]

Kosuke Hashimoto,* Akiyasu C. Yoshizawa,* Shujiro Okuda,* Keiichi Kuma,† Susumu Goto,* and Minoru Kanehisa^{1,*§}

Bioinformatics Center,* Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan; National Institute of Informatics,† Chiyoda-ku, Tokyo 101-8430, Japan; and Human Genome Center,[§] Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639, Japan

Abstract The repertoire of biosynthetic enzymes found in an organism is an important clue for elucidating the chemical structural variations of various compounds. In the case of fatty acids, it is essential to examine key enzymes that are desaturases and elongases, whose combination determine the range of fatty acid structures. We systematically investigated 56 eukaryotic genomes to obtain 275 desaturase and 265 elongase homologs. Phylogenetic and motif analysis indicated that the desaturases consisted of four functionally distinct subfamilies and the elongases consisted of two subfamilies. From the combination of the subfamilies, we then predicted the ability to synthesize six types of fatty acids. Consequently, we found that the ranges of synthesizable fatty acids were often different even between closely related organisms. The reason is that, as well as diverging into subfamilies, the enzymes have functionally diverged within the individual subfamilies. Finally, we discuss how the adaptation to individual environments and the ability to synthesize specific metabolites provides some explanation for the diversity of enzyme functions. **■** This study provides an example of a potent strategy to bridge the gap from genomic knowledge to chemical knowledge.—Hashimoto, K., A. C. Yoshizawa, S. Okuda, K. Kuma, S. Goto, and M. Kanehisa. The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. *J. Lipid Res.* 2008. 49: 183–191.

Supplementary key words genome • bioinformatics • lipid • histidine box • motif analysis • phylogenetic analysis • evolution

Fatty acids, which are the essential components of biomembranes, have great structural and functional diversity. In particular, there is accumulated clinical evidence that PUFAs have beneficial effects for human health (1). Thus, the artificial production of fatty acids is the subject of intensive research, and commercial interest is increasing (2). Several studies have demonstrated the reconstitution of the biosynthesis pathway for PUFAs and

the synthesis of fatty acids in transgenic plants (3, 4) and transgenic microalgae (5). In addition, fatty acids and their related genes attract great attention as potential targets for antibiotics (6) as well as treatments for obesity and the metabolic syndrome (7). With this background, much effort has been expended on the cloning and molecular identification of genes encoding fatty acid desaturases and elongases, which are key enzymes for generating a variety of fatty acids, especially in eukaryotes (8, 9).

Fatty acid desaturases catalyze the introduction of a double bond into an acyl chain with strict regioselectivity and stereoselectivity. These enzymes can be classified into two phylogenetically unrelated groups: the membrane-bound fatty acid desaturases and the acyl-acyl carrier protein desaturases (10). The former is the dominant enzyme in desaturation and is ubiquitous in eukaryotes and bacteria. Their sequences are characterized by three histidine box motifs containing eight histidine residues (11). The latter is a plant enzyme, which specifically catalyzes the conversion of 18:0 to 18:1 (12). The enzymes that are responsible for the rate-limiting step of acyl chain extension also can be classified into two phylogenetically unrelated groups: the fatty acid elongases and the β -ketoacyl-CoA synthases (13). The former is the dominant enzyme in elongation and is widely distributed in eukaryotes, characterized by one histidine box. The latter is specific for saturated fatty acids or MUFAs, but not for PUFAs, and has only been identified to date in plants (14). Thus, fatty acid structures are generally determined by the combination of two types of enzymes: the membrane-bound fatty acid desaturases and the fatty acid elongases.

Abbreviations: HMM, hidden Markov model; KEGG, Kyoto Encyclopedia of Genes and Genomes; SCD, stearoyl-coenzyme A desaturase.

¹ To whom correspondence should be addressed.

e-mail: kanehisa@kuicr.kyoto-u.ac.jp

§ The online version of this article (available at <http://www.jlr.org>) contains supplementary data in the form of two tables and eight figures.

Manuscript received 23 August 2007 and in revised form 28 September 2007.

Published, JLR Papers in Press, October 7, 2007.

DOI 10.1194/jlr.M700377-JLR200

Copyright © 2008 by the American Society for Biochemistry and Molecular Biology, Inc.

This article is available online at <http://www.jlr.org>

Bioinformatics approaches to lipid research have recently begun using large amounts of mass spectrometry and microarray data (15, 16). Phylogenetic analysis of protein families related to fatty acids has also been performed (10, 17). However, to our knowledge, there is no report that describes the investigation of fatty acid structures based on the comprehensive analysis of the gene contents in the genomes. In fact, the prediction of fatty acid structures from genomic information is difficult because of the functional diversity of the key enzymes. For example, the membrane desaturases constitute a highly diversified family including at least 10 different types of regioselectivities, such as $\Delta 4$, $\Delta 5$, $\Delta 6$, $\Delta 8$, $\Delta 9$, $\Delta 10$, $\Delta 11$, $\Delta 12$, $\Delta 13$, and $\Delta 15$, some of whose sequences are significantly close to one another (10). Several reports indicated that the function of experimentally characterized desaturases did not correspond to the annotation from similarity searches (18). It is thus difficult to connect enzymes with fatty acid structures from an analysis based merely on the sequence similarity of individual enzymes.

The resources and the strategy to solve such problems are provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG) project, namely the integration of genomic information and chemical information (19). One successful example is that of functional glycomics, in which glycan structures are related to the repertoire of glycosyltransferases, which synthesize glycan chains in a stepwise manner with distinct substrate specificities (20, 21). Another example of the integrative analysis of genomic and chemical information is the prediction of polyketide and nonribosomal peptide structures, which are also complex natural compounds synthesized by distinct types of syntheses (22). We assume that the application of a similar strategy to fatty acids can allow us to understand how different lipid structures are found among organisms and what the meaning of the difference is. Taking the analysis from genomic information through biological components to phenotypes is a challenge in fatty acids, which have important interactions with an organism's environment.

In this study, we first investigated the diversity of membrane fatty acid desaturases and elongases. Although these enzymes have been classified in previous work, the classifications were not comprehensive and not consistent with each other. Our phylogenetic analysis indicated that desaturases are divided into four functional subfamilies and elongases are divided into two functional subfamilies. Each subfamily has a distinct motif, whose profiles can be used for functional assignments of desaturases and elongases in newly sequenced genomes. In the next step, we examined the ability of a set of organisms to synthesize fatty acids, especially six types of fatty acids widely distributed in nature, from the pathway viewpoint. Our analysis suggests that differences in the repertoires of enzymes as well as functional divergence in each subfamily underlie the fatty acid diversity among organisms. Adaptation to individual environments and the ability to synthesize specific metabolites may provide an explanation for the diversity of enzyme functions and subsequent fatty acid structures.

First, we comprehensively collected all sequences similar to known desaturases and elongases, including slightly or partially similar sequences, from a genomic data set using the PSI-BLAST program (23). We then manually discarded any false-positive hits by investigating the results of hierarchical clustering and multiple alignments according to a similar method described previously (17, 24). The following analysis was independently performed against the desaturase and elongase data sets.

Searching for similar sequences with PSI-BLAST

As PSI-BLAST targets, we used amino acid sequences from 56 complete or draft-quality whole eukaryotic genomes, including 21 animals, 20 fungi, 3 plants, and 12 protists. These data were derived from KEGG GENES and DGENES Release 41.0 (<http://www.genome.jp/kegg/genes.html>). As the query sequences for PSI-BLAST (blastp2.2.10) search, we used experimentally known desaturase and elongase sequences from a wide range of organisms, including *Homo sapiens* from animals, *Saccharomyces cerevisiae* from fungi, *Arabidopsis thaliana* from plants, *Trypanosoma brucei* from protists, and *Bacillus subtilis*, *Pseudomonas aeruginosa*, and *Mycobacterium tuberculosis* from bacteria (indicated by circles in supplementary Data I). By combining all of the PSI-BLAST results with E values < 0.01 into one file and removing duplicate hits, the initial data set was obtained.

Discarding false-positive sequences

To remove the false-positive sequences from the initial data set, a hierarchical clustering analysis was performed. First, the sequence similarity for each pair of whole sequences was calculated with the SSEARCH 3.4t06 program (25), which is an implementation of the Smith-Waterman algorithm (26). Next, we defined the distance between the sequences as (distance) = 1,000/(Smith-Waterman score). Then, using the distance, a hierarchical cluster was calculated with the complete linkage method of the R program package for statistical computing version 1.7.1 (27) and with the BioRuby library version 1.0 (<http://bioruby.org/>). The hierarchical cluster tree was separated into clusters with a proper threshold.

We manually checked all of the clusters and subsequently determined false-positive clusters using two criteria: literature information and motif information. If one or more proteins in a cluster were annotated as nondesaturase or nonelongase protein by the literature or database annotation, the cluster was discarded. Then, clusters that did not contain a specific motif were discarded, because all known desaturases and elongases conserved each histidine motif.

Phylogenetic analysis

We used MAFFT version 5.8 (28) to obtain all multiple alignments for phylogenetic and motif analysis. Although phylogenetic trees of the entire desaturases and elongases were calculated with the neighbor-joining method (29) using ClustalW version 1.83 (30), trees of the individual subfamilies were calculated with the Bayesian method using MrBayes version 3.1.2 (31). In the Bayesian method, Markov chain Monte Carlo analysis was performed with 20,000–500,000 generations and four independent chains. The Markov chain was sampled every 100 generations. Both the entire trees and the subfamily trees were reconstructed using conserved regions independent of the cytochrome *b₅* domain. For the display and manipulation of phylogenetic trees, we used a web-based tool, Interactive Tree Of Life (32).

Motif analysis

We used HMMER version 2.3.2 (<http://hmm.janelia.org/>) to build hidden Markov model (HMM) profiles in subfamilies and to search for subfamily motifs. The cytochrome *b₅* domain was searched with the Pfam profile PF00173 with E values < 0.05 (33). Graphical representations of the conservation patterns of consensus sequences were generated by WebLogo (34).

Microarray analysis

Our microarray data set was derived from expression data of many human tissues provided by the Genomics Institute of the Novartis Research Foundation (35). We found expression data of the genes corresponding to five desaturases and three elongases of the human enzymes obtained in this study. Expressed genes were determined by Affymetrix MAS5 Absent/Present calls.

RESULTS

Desaturases consist of four functionally distinct subfamilies

We obtained 275 desaturase homologs from 56 eukaryotic genomes through the sequence analysis (listed in supplementary Data I-1). The phylogenetic tree of desaturase sequences comprises four large branches each with a distinct function (see supplementary Data II-1). We defined the branches as the following subfamilies: *a*) First Desaturase, introducing the first double bond into the saturated acyl chain; *b*) Omega Desaturase, introducing a double bond between an existing double bond and the acyl end; *c*) Front-End Desaturase, introducing a double bond between an existing double bond and the carboxyl end; and *d*) Sphingolipid Desaturase, sphingolipid $\Delta 4$ desaturases. All subfamilies contain sequences belonging to animals, fungi, plants, and protists; they probably diverged early. To clarify the difference in functions between subfamilies, we discuss them below.

The predominant members in the First Desaturase subfamily were stearoyl-coenzyme A desaturases (SCDs), which generally introduce a double bond to the $\Delta 9$ position of palmitic acid (16:0) or stearic acid (18:0). All experimentally confirmed SCDs were detected and classified into this subfamily, including two in *H. sapiens* (36, 37) and four in *Mus musculus* (38).

The Omega Desaturase subfamily contained 13 known desaturases whose functions were $\Delta 12$ or $\Delta 15$ desaturases. The two functions did not form two distinct branches but fell in each of four branches representing the four kingdoms: animals, fungi, plants, and protists. Additional phylogenetic analysis of this subfamily indicated that these functions independently diverged in each lineage (see supplementary Data II-2). For example, in the animal kingdom, the $\Delta 12$ and $\Delta 15$ sequences of *Caenorhabditis elegans* diverged after *C. elegans* separated from other animal species. This is supported by high posterior probabilities. In the same way, the two functions diverged in the fungi kingdom after it separated from the others. Plants also obtained both $\Delta 12$ and $\Delta 15$ desaturases at different duplication points.

The Front-End Desaturase subfamily also comprised desaturases whose substrate is unsaturated acyl chains. The difference from the Omega Desaturase subfamily is the position of the double bond, which in this case is introduced between an existing double bond and the carboxyl end. This subfamily included $\Delta 4$, $\Delta 5$, $\Delta 6$, and bifunctional $\Delta 6$ /sphingolipid $\Delta 8$ desaturases. Similar to Omega Desaturases, additional phylogenetic analysis suggested with high probability that the $\Delta 5$ and $\Delta 6$ desaturases of nematodes, vertebrates, and others diverged separately in each lineage (see supplementary Data II-3).

The last group is the Sphingolipid Desaturase subfamily, whose sole function is the sphingolipid $\Delta 4$ desaturase. Previous research has already indicated that these sequences form a distinct subfamily (39), and our results supported this conclusion.

Elongases consist of two functionally distinct subfamilies

We obtained 265 elongase homologs from 56 eukaryotic genomes (listed in supplementary Data I-2). The phylogenetic tree of elongase sequences was roughly separated into two branches (see supplementary Data II-4). Each branch was defined as follows: *a*) S/MUFA Elongase, elongating a saturated fatty acid or a MUFA; or *b*) PUFA Elongase, elongating a polyunsaturated fatty acid.

The S/MUFA Elongase subfamily contained 11 experimentally known elongases. Six of them elongate saturated fatty acids. Four of them elongate both saturated fatty acids and MUFAs. There is one exception, whose function is to elongate 18:2 (40). A notable feature of this subfamily is the various specificities for the length of the acyl chain. For example, EVOLV6 in *M. musculus* elongates C12–16 (41), whereas *S. cerevisiae* has three different enzymes whose substrate specificities are C14–16, C14–24, and C14–26 (42, 43). In addition, recent research evaluated three elongases in *T. brucei* whose substrate specificities were found to be C4–10, C10–14, and C14–18 (44).

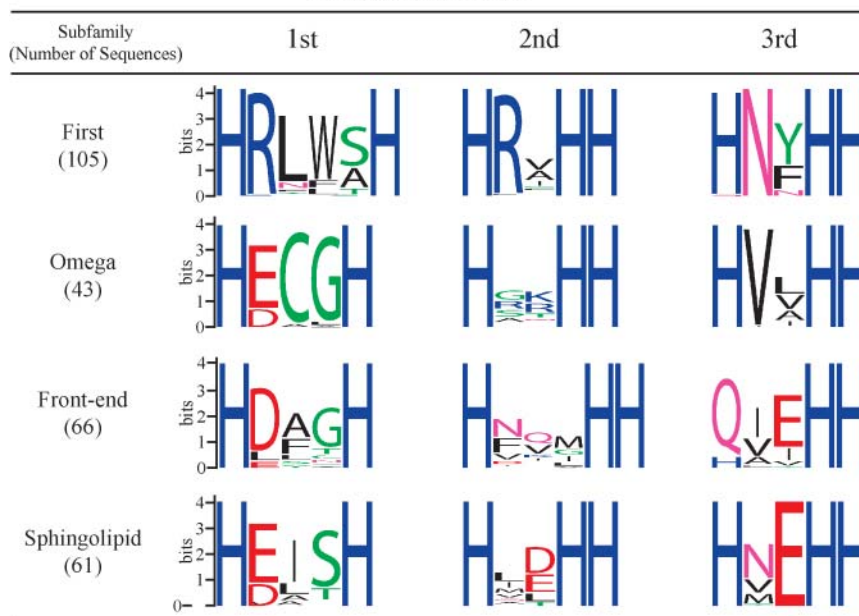
Sequences in the PUFA Elongase subfamily are composed of animals and protists. No sequences were detected in fungi and plants. Vertebrates have the most paralogs; for example, human and mouse have five paralogs, some of which have been experimentally confirmed to be involved in the elongation of PUFAs. Four of the seven sequences detected in protists have also proved to be elongases of PUFAs by a recent study (45). This subfamily also has one exception, which has been characterized as an elongase for short MUFAs in *Drosophila melanogaster* (46).

Several amino acids are clearly different in subfamilies

As described above, desaturases were divided into four subfamilies and elongases were divided into two subfamilies. In this section, first, we describe the difference in amino acids between subfamilies. Next, we show that HMM profiles constructed for each subfamily can classify test sequences into appropriate subfamilies.

Figure 1A shows sequence logos of three histidine boxes in desaturase subfamilies, which were clearly different even in conserved regions. The first histidine box including two

A Desaturase Motif



B Elongase Motif

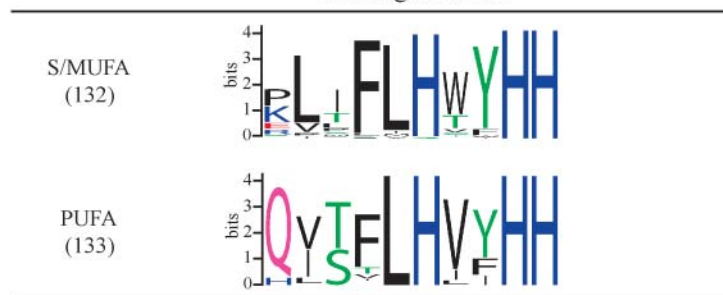


Fig. 1. Sequence logo representation of conserved histidine regions in each subfamily. The height of each amino acid symbol is proportional to its frequency of occurrence. A: Desaturase motif in four subfamilies. Desaturases possess three histidine boxes, shown as 1st, 2nd, and 3rd. Different amino acids are conserved between subfamilies; for example, there is a strongly conserved arginine in the 1st and 2nd histidine boxes and an asparagine in the 3rd histidine box exclusively in the First Desaturase subfamily. B: Elongase motif in two subfamilies. Elongases possess one histidine box, which is similar between subfamilies. However, several amino acids are different on the N-terminal side of the histidine box, such as lysine in the S/MUFA subfamily and glutamine in the PUFA subfamily.

histidines is located in the N-terminal region (Fig. 1A, 1st). There were three amino acids between the histidines in three subfamilies, whereas four amino acids existed between them in the First Desaturase subfamily. The second histidine box including three histidines is positioned ~30 amino acids downstream of the first one (Fig. 1A, 2nd). In this region, the number of amino acids between the histidines was different only in the Front-End Desaturase subfamily. In addition, a strongly conserved arginine was observed in the First Desaturase subfamily. The last histidine box is located in the C-terminal region far from the others (Fig. 1A, 3rd). As is often reported, in many sequences the first histidine changed to glutamine in the Front-End Desaturase subfamily. The second amino acid also has clear difference between subfamilies. It is strongly conserved as asparagine in First Desaturases and as valine

in Omega Desaturases. Figure 1B shows one histidine box and the surrounding region conserved in fatty acid elongases. Several different amino acids were found, such as leucine in the S/MUFA subfamily and glutamine in the PUFA subfamily.

HMM profiles were constructed using the length of ~200 amino acids for all of the sequences in each subfamily. To validate these profiles, we calculated the *P* values of test sequences against the profiles. The test sequences were 30 known desaturases and 17 elongases derived from a wide range of organisms, such as *Thalassiosira pseudonana*, *Spodoptera littoralis*, and *Mortierella alpina*. The sequences were not used in calculating the profiles, because our data set only consisted of nearly complete genomes. **Figure 2** summarizes the *P* values, indicating that the profiles can clearly classify desaturases into appropriate subfamilies,

Functions	Accession	Organisms	Fi	Om	Fr	Sp
Δ9	AAB92583	<i>T. ni</i>	■			
Δ9	AAL27034	<i>O. furnacalis</i>	■			
Δ9	CAB38177	<i>M. alpina</i>	■			
Δ9	CAB38178	<i>M. alpina</i>	■			
Δ9	CAC81988	<i>M. alpina</i>	■			
Δ9	AAQ74258	<i>S. littoralis</i>	■			
Δ9	AAM12238	<i>P. glauca</i>	■			
Δ11	AAQ74259	<i>S. littoralis</i>	■			
Δ11	AAD03775	<i>T. ni</i>	■			
Δ14	AAL35746	<i>O. furnacalis</i>	■			
Δ12	BAA81754	<i>M. alpina</i>		■		
Δ12	AAT58363	<i>R. oryzae</i>		■		
Δ12	BAB78716	<i>C. vulgaris</i>		■		
Δ15	BAD91495	<i>M. alpina</i>		■		
Δ15	BAB78717	<i>C. vulgaris</i>		■		
Δ12/15	ABK15557	<i>A. castellanii</i>		■		
Δ6	AF465281	<i>M. alpina</i>			■	
Δ6	AAQ10731	<i>A. leveillei</i>			■	
Δ6	CAA11033	<i>P. patens</i>			■	
Δ6	AAT85661	<i>M. polymorpha</i>			■	
Δ6	AAX14505	<i>T. pseudonana</i>			■	
Δ5	BAD95486	<i>M. alpina</i>			■	
Δ5	AAT85663	<i>M. polymorpha</i>			■	
Δ5	AAX14502	<i>T. pseudonana</i>			■	
Δ4	AAM09688	<i>T. sp.</i>			■	
Δ4	AAX14506	<i>T. pseudonana</i>			■	
Δ4	AAQ98793	<i>P. lutheri</i>			■	
Δ4	AAY15136	<i>P. salina</i>			■	
Δ8	ABF58684	<i>P. marinus</i>			■	
Δ8	AAD45877	<i>E. gracilis</i>			■	

Fi: First
Om: Omega
Fr: Front-end
Sp: Sphingolipid

■ ~ 1.0E-100
 ■ 1.0E-100 ~ 1.0E-50
 ■ 1.0E-50 ~ 1.0E-10
 ■ 1.0E-10 ~

Fig. 2. Classification of test sequences into subfamilies using hidden Markov model (HMM) profiles. Different tones of black indicate *P* values against the HMM profiles constructed in each subfamily. The profiles distinguish between subfamilies, namely Δ9, Δ11, and Δ14 into First Desaturase, Δ12 and Δ15 into Omega Desaturase, and Δ4, Δ5, Δ6, and Δ8 into Front-End Desaturase. Accession indicates GenBank accession numbers.

namely, Δ9 into First Desaturase, Δ12 and Δ15 into Omega Desaturase, and Δ4, Δ5, Δ6, and Δ8 into Front-End Desaturase. The profiles of elongases also distinguished test sequences into two elongase subfamilies, although not as clearly as the desaturase case (see supplementary Data II-5).

The repertoires of the subfamilies and predicted fatty acids are highly diverse

The phylogenetic and motif analyses indicated that desaturases and elongases consist of distinct subfamilies whose sequences and functions are different. Thus, we can determine a set of subfamilies in an organism from its genome sequence. The next important step is to predict fatty acids that each organism can synthesize from the set of subfamilies. The prediction was conducted by map-

ping the combination of enzymes onto the pathway of the fatty acid synthesis. We selected six types of unsaturated fatty acids as the prediction targets: oleic acid (18:1), linoleic acid (18:2), α-linolenic acid (18:3), arachidonic acid (20:4), eicosapentaenoic acid (EPA; 20:5), and docosahexaenoic acid (DHA; 22:6), which are widely distributed in the biological membrane. The repertoires of the subfamilies and predicted fatty acids in 56 eukaryotic genomes are summarized in supplementary Data II-6. The repertoires of desaturases and elongases in organisms show great diversity, which yields the diversity of fatty acids observed.

We present an overview of the prediction results using the simplified pathway illustrated in **Fig. 3A**. There are three reaction processes catalyzed by the distinct functional subfamilies in the pathway. First, through process 1 (**Fig. 3A**), oleic acid (18:1) is synthesized from stearic acid (18:0) by First Desaturase introducing the first double bond. We predicted that organisms that possessed First Desaturases could synthesize oleic acids.

Next, through process 2 (**Fig. 3A**), linoleic acid (18:2) and α-linolenic acid (18:3) are synthesized from oleic acid (18:1) by Δ12 and Δ15 desaturases. Because they were categorized in the Omega Desaturase subfamily, we determined that the process would be present in many fungi and plants that possess the subfamily, in contrast to vertebrates, which require such fatty acids as a diet to survive. These results agree with the fact that plasma membranes

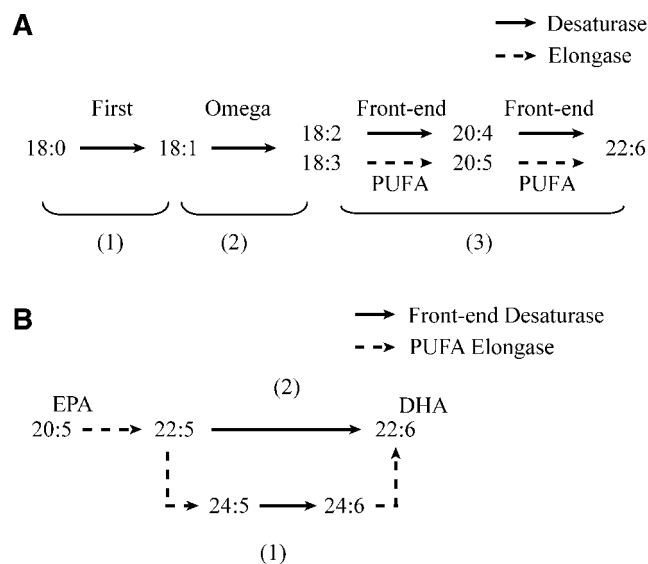


Fig. 3. A: Schematic pathway of unsaturated fatty acids with subfamily enzymes. 1: The pathway from stearic acid (18:0) to oleic acid (18:1), catalyzed by Δ9 desaturases belonging to the First Desaturase subfamily. 2: The pathway from oleic acid (18:1) to linoleic acid (18:2) and α-linolenic acid (18:3), catalyzed by Δ12 and Δ15 desaturases belonging to the Omega Desaturase subfamily. 3: The pathway from oleic acid (18:1) and linoleic acid (18:2) to docosahexaenoic acid (DHA; 22:6), catalyzed by Δ4, Δ5, and Δ6 desaturases belonging to the Front-End Desaturase subfamily and C18, C20, and C22 elongases belonging to the PUFA subfamily. B: The two distinct pathways from eicosapentaenoic acid (EPA) to DHA. 1: The Sprecher pathway found in mammals. 2: The pathway using the Δ4 desaturase found in microalgae.

in such organisms are abundant in linoleic acid (18:2) and α -linolenic acid (18:3) (47, 48).

Process 3 (Fig. 3A) is a complicated step involving two distinct pathways and plural subfamilies (details mentioned in Discussion). Front-End Desaturases and PUFA Elongases dominate the process and cooperatively synthesize arachidonic acid (20:4), EPA (20:5), and DHA (22:6). Therefore, we determined that organisms can perform the conversion if they possess both of the subfamilies. As a result, sea urchin and trypanosomatids had this pathway as well as vertebrates.

DISCUSSION

Functional divergence of desaturases and elongases in two phases

We identified >450 new putative genes encoding desaturases or elongases from various eukaryotic genomes. For example, 13 genes belonging to five different subfamilies were detected in *Strongylocentrotus purpuratus*, whose

repertoire is similar to that of vertebrates (see supplementary Data II-6). In addition, novel desaturases and elongases conserved in three *Plasmodium* species were identified. Our phylogenetic and motif analyses indicate that the functions of desaturases and elongases have diverged via two phases. In the first phase, these proteins diverged into the subfamilies introduced earlier. Subsequently, they have diverged into different specificities, $\Delta 12$ and $\Delta 15$ in the Omega Desaturase subfamilies and $\Delta 4$, $\Delta 5$, $\Delta 6$, and $\Delta 8$ in the Front-End Desaturase subfamilies, within the individual subfamilies in the second phase. We discuss the history of functional diversification of desaturases and elongases through the two phases.

In the first phase, desaturases and elongases diverged into four and two subfamilies, which had different ranges of substrate specificities. They can still be distinguished based on sequence similarities, because different amino acids are conserved (Fig. 1). For example, arginine residues in the first and second histidine boxes are conserved in the sequences of >95% of First Desaturases, whereas the other subfamilies possess few arginines in this region.

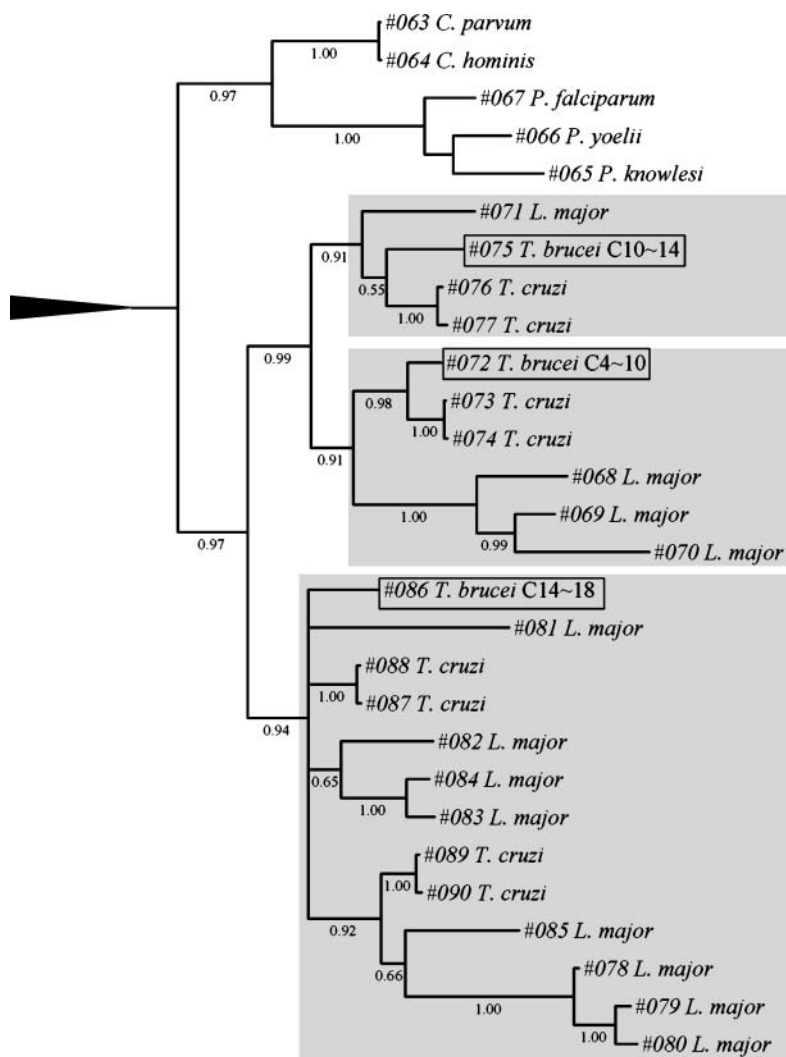


Fig. 4. Partial phylogenetic tree of the S/MUFA Elongase subfamily. This tree is the protist branch of the phylogenetic tree of the S/MUFA Elongase subfamily. The full tree is shown in supplementary Data II-8. Numbers below the branches indicate posterior probability values. Sequences of trypanosomatids are separated into three branches, each of which contains different substrate specificities.

Such specific residues could relate to the differences in functions. To elucidate the functions of such residues, further experiments, such as to determine the three-dimensional structures, are required.

In the second phase, we believe that a variety of substrate specificities or regioselectivities diverged within the individual subfamilies. Recent reports also show the independent divergence of $\Delta 12$ and $\Delta 15$ in several organisms, such as *Mortierella alpina* and *Saprolegnia diclina* (49, 50). It should be noted that the first diversification occurred in the old common ancestor, whereas the second occurred independently in each lineage. The second diversification likely causes the difference in fatty acids between even closely related organisms. In other words, the first phase restricted the range of functions and then, in the second phase, functions diverged within their ranges. However, some exceptional functions beyond the ranges of subfamilies were detected in elongases, suggesting that the functional constraints of the subfamilies are not always effective. In particular, elongases seem to be more flexible about substrate specificities.

Two pathways converting EPA to DHA are the consequences of independent divergence in the second phase. Figure 3B, pathway 1 shows the Sprecher pathway, which contains four steps: two consecutive elongation steps from EPA to 24:5, the $\Delta 6$ desaturation to 24:6, and the subsequent β -oxidation from 24:6 to 22:6 (51). The C20 and C22 elongases that convert EPA to 24:5 are key enzymes in this pathway, because the latter two steps are catalyzed by reused enzymes that also serve in other pathways. In our phylogenetic analysis (see supplementary Data II-7), it was found that many animals had multiple copies of elongases, as for the PUFA Elongase subfamily. In particular, vertebrates have five distinct branches, some of which include elongases characterized as C18, C20, and C22 elongases (52, 53). Another pathway for DHA, identified in lower eukaryotes, is simpler, consisting of two steps: an elongation from EPA to DPA, and the $\Delta 4$ desaturation to DHA (Fig. 3B, pathway 2) (54). This pathway was only detected in three trypanosomatids. We conclude that DHA can be synthesized by vertebrates and trypanosomatids, each of which acquired different enzymes for the extension of the pathway to DHA in the second phase diversification.

Fatty acids as an adaptation to environments and precursors of metabolites

Expansions and contractions of pathways caused by the diversity of desaturases and elongases lead to a considerable diversity in fatty acids among organisms. What roles do such a variety of fatty acids play in each organism? With respect to the adaptation to individual environments, one interesting example is *T. brucei*, a human parasite that causes sleeping sickness. When this parasite invades the human blood from the tsetse fly, it rapidly replaces fatty acids in the plasma membrane with myristic acid to evade the host's immune response (55). A recent report identified three elongases for a novel type of fatty acid synthesis, each of which has different conversion ranges, such as C4–10, C10–14, and C14–18 (44). This remark-

able finding suggested that the parasite can readily produce stage-specific fatty acids by regulating the expression of elongases. Our phylogenetic analysis of S/MUFA Elongase indicates that these three elongases fall into a single branch comprising exclusively trypanosomatid sequences apart from other protists. The branch then separates into three subbranches with different substrate specificities (Fig. 4; see supplementary Data II-8). Hence, these genes have probably been acquired in relatively recent duplication events and have subsequently mutated into different substrate specificities. Other trypanosomatids, *Trypanosoma cruzi* and *Leishmania major*, also have more S/MUFA elongases, which have arisen in a similar manner to *T. brucei*. They are intracellular parasites unlike *T. brucei* and use stage-specific fatty acids in their reproductive cycles [details of the lipid biology of trypanosomatids were reviewed recently (56)]. Hence, such paralogs are likely to have specificities corresponding to their life stages. Their characterization will lead to improved understanding of the lipid biology of parasites.

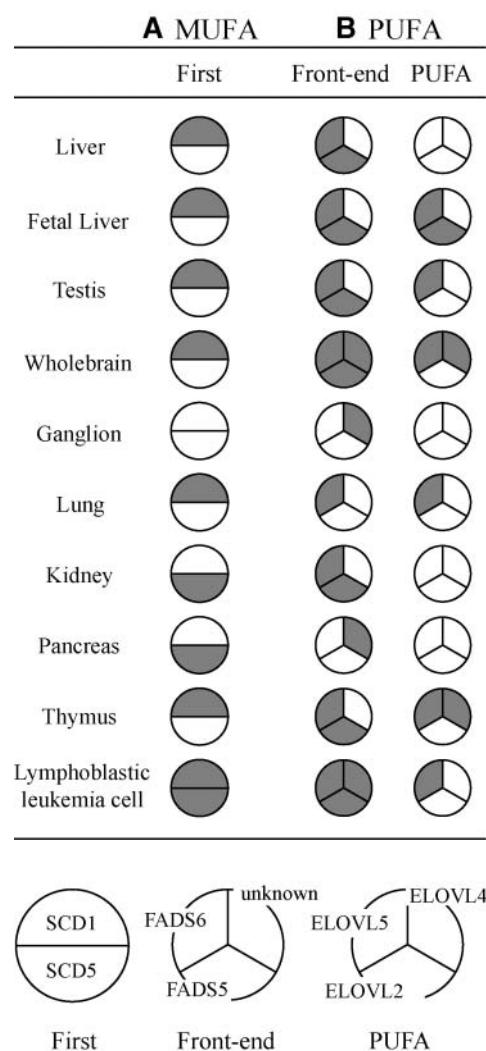



Fig. 5. Distribution of MUFA (A) and PUFA (B) expression in different human tissues. Gene names in three subfamilies are described at the bottom. Expressed genes are shaded in each tissue.

Another important role of fatty acids is as precursors of various metabolites. In mammals, metabolites of both MUFAs and PUFAs are used as signaling molecules (57, 58). Arachidonic acid is abundantly stored within the cell membrane and is required as a substrate for eicosanoid synthesis (59). EPA (20:5) is the precursor of prostaglandins, whereas DHA (22:6) is essential for nervous system maintenance and development (60). Meanwhile, insects use fatty acids as a precursor of pheromones (61). A typical example is moth sex pheromone, derived from fatty acids with desaturations and alterations of the chain length. Pheromone structures have a great diversity even among closely related species (62) as a result of desaturases with a variety of stereoselectivities and regioselectivities. This suggests that the diversification of the desaturases and the subsequent pheromone structures occurred under a birth-and-death process with strong selection pressure (63, 64). The number of paralogs in insects is apparently larger than in other groups, especially in the First Desaturase and PUFA Elongase subfamilies. This suggests that the genes of these subfamilies in insect genomes evolved under the specific circumstance of pheromone diversification pressure.

From genomic information to chemical structures

Mapping microarray data to pathways is another potential use of our results (21). The human pathway for the biosynthesis of unsaturated fatty acids is divided in two, MUFA and PUFA, because of the lack of the Omega Desaturase subfamily. MUFAs are synthesized by SCDs, including two paralogs, which are expressed in a tissue-specific manner (Fig. 5A). For example, SCD2 is expressed exclusively in pancreas and kidney. Previous reports also suggested that SCD1 was expressed in liver, muscle, and other tissues (65) and SCD2 was abundantly expressed in adult brain and pancreas (37). For PUFA biosynthesis, both Front-End Desaturases and PUFA Elongases are required. Their expressions were also different across tissues, as expected (Fig. 5B). A previous study indicated that the proportion of PUFAs including three or more double bonds increased in patients with acute lymphoblastic leukemia (66). These results suggest that genes encoding desaturases and elongases are strictly regulated according to tissues or cell types to control the composition of fatty acids in membranes. Similar analysis can be performed with all of the complete genomes. Furthermore, analysis of other enzymes would enable us to predict whole lipid structures, including head groups.

In this study, we comprehensively detected the key proteins for fatty acid synthesis from genome sequences. The combination of the proteins elucidated differences in the repertoires of major fatty acids in organisms. This analysis can be readily applied to newly sequenced genomes using our HMM profiles. In addition, our results can be combined with experimental data for further analysis. Biosynthesis pathways that integrate genetic information, such as desaturases and elongases, and a variety of fatty acid structures will be provided by KEGG. 

The authors thank Alex Gutteridge for critical reading of the manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science, and Technology of Japan and the Japan Science and Technology Agency. K.H. was supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

REFERENCES

1. Simopoulos, A. P. 2002. Omega-3 fatty acids in inflammation and autoimmune diseases. *J. Am. Coll. Nutr.* **21**: 495–505.
2. Warude, D., K. Joshi, and A. Harsulkar. 2006. Polyunsaturated fatty acids: biotechnology. *Crit. Rev. Biotechnol.* **26**: 83–93.
3. Qi, B., T. Fraser, S. Mugford, G. Dobson, O. Sayanova, J. Butler, J. A. Napier, A. K. Stobart, and C. M. Lazarus. 2004. Production of very long chain polyunsaturated omega-3 and omega-6 fatty acids in plants. *Nat. Biotechnol.* **22**: 739–745.
4. Graham, I. A., T. Larson, and J. A. Napier. 2007. Rational metabolic engineering of transgenic plants for biosynthesis of omega-3 polyunsaturates. *Curr. Opin. Biotechnol.* **18**: 142–147.
5. Sijtsma, L., and M. E. de Waaf. 2004. Biotechnological production and applications of the omega-3 polyunsaturated fatty acid docosahexaenoic acid. *Appl. Microbiol. Biotechnol.* **64**: 146–153.
6. Wang, J., S. M. Soisson, K. Young, W. Shoop, S. Kodali, A. Galgoci, R. Painter, G. Parthasarathy, Y. S. Tang, R. Cummings, et al. 2006. Platensimycin is a selective FabF inhibitor with potent antibiotic properties. *Nature.* **441**: 358–361.
7. Dobrzyn, A., and J. M. Ntambi. 2005. Stearoyl-CoA desaturase as a new drug target for obesity treatment. *Obes. Rev.* **6**: 169–174.
8. Nakamura, M. T., and T. Y. Nara. 2004. Structure, function, and dietary regulation of delta6, delta5, and delta9 desaturases. *Annu. Rev. Nutr.* **24**: 345–376.
9. Leonard, A. E., S. L. Pereira, H. Sprecher, and Y. S. Huang. 2004. Elongation of long-chain fatty acids. *Prog. Lipid Res.* **43**: 36–54.
10. Sperling, P., P. Ternes, T. K. Zank, and E. Heinz. 2003. The evolution of desaturases. *Prostaglandins Leukot. Essent. Fatty Acids.* **68**: 73–95.
11. Shanklin, J., E. Whittle, and B. G. Fox. 1994. Eight histidine residues are catalytically essential in a membrane-associated iron enzyme, stearoyl-CoA desaturase, and are conserved in alkane hydroxylase and xylene monooxygenase. *Biochemistry.* **33**: 12787–12794.
12. Kachroo, A., J. Shanklin, E. Whittle, L. Lapchyk, D. Hildebrand, and P. Kachroo. 2007. The Arabidopsis stearoyl-acyl carrier protein-desaturase family and the contribution of leaf isoforms to oleic acid synthesis. *Plant Mol. Biol.* **63**: 257–271.
13. Zank, T. K., U. Zahringer, C. Beckmann, G. Pohnert, W. Boland, H. Holtorf, R. Reski, J. Lerchl, and E. Heinz. 2002. Cloning and functional characterisation of an enzyme involved in the elongation of Delta6-polyunsaturated fatty acids from the moss *Physcomitrella patens*. *Plant J.* **31**: 255–268.
14. Jakobsson, A., R. Westerberg, and A. Jakobsson. 2006. Fatty acid elongases in mammals: their regulation and roles in metabolism. *Prog. Lipid Res.* **45**: 237–249.
15. Watson, A. D. 2006. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Lipidomics: a global approach to lipid analysis in biological systems. *J. Lipid Res.* **47**: 2101–2111.
16. Yetukuri, L., M. Katajamaa, G. Medina-Gomez, T. Seppanen-Laakso, A. Vidal-Puig, and M. Oresic. 2007. Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Syst Biol.* **1**: 12.
17. Hashimoto, K., A. C. Yoshizawa, K. Saito, T. Yamada, and M. Kanehisa. 2006. The repertoire of desaturases for unsaturated fatty acid synthesis in 397 genomes. *Genome Inform.* **17**: 173–183.
18. Tripodi, K. E., L. V. Buttiglieri, S. G. Altabe, and A. D. Uttaro. 2006. Functional characterization of front-end desaturases from trypanosomatids depicts the first polyunsaturated fatty acid biosynthetic pathway from a parasitic protozoan. *FEBS J.* **273**: 271–280.
19. Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**: D354–D357.

20. Hashimoto, K., S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki, and M. Kanehisa. 2006. KEGG as a glycome informatics resource. *Glycobiology*. **16**: 63R–70R.
21. Kawano, S., K. Hashimoto, T. Miyama, S. Goto, and M. Kanehisa. 2005. Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics*. **21**: 3976–3982.
22. Minowa, Y., M. Araki, and M. Kanehisa. 2007. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* **368**: 1500–1517.
23. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
24. Yoshizawa, A. C., S. Kawashima, S. Okuda, M. Fujita, M. Itoh, Y. Moriya, M. Hattori, and M. Kanehisa. 2006. Extracting sequence motifs and the phylogenetic features of SNARE-dependent membrane traffic. *Traffic*. **7**: 1104–1118.
25. Pearson, W. R. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*. **11**: 635–650.
26. Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**: 195–197.
27. R_Development_Core_Team. 2003. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
28. Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**: 511–518.
29. Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
30. Ayar, A. 2000. The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment. *Methods Mol. Biol.* **132**: 221–241.
31. Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. **19**: 1572–1574.
32. Letunic, I., and P. Bork. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. **23**: 127–128.
33. Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, et al. 2004. The Pfam protein families database. *Nucleic Acids Res.* **32**: D138–D141.
34. Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res.* **14**: 1188–1190.
35. Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*. **101**: 6062–6067.
36. Mugnai, G., and V. Boddi. 1977. Stearoyl-CoA desaturase in mitochondrial membrane fractions. *Ital. J. Biochem.* **26**: 245–253.
37. Wang, J., L. Yu, R. E. Schmidt, C. Su, X. Huang, K. Gould, and G. Cao. 2005. Characterization of HSCD5, a novel human stearyl-CoA desaturase unique to primates. *Biochem. Biophys. Res. Commun.* **332**: 735–742.
38. Miyazaki, M., S. M. Bruggink, and J. M. Ntambi. 2006. Identification of mouse palmitoyl-coenzyme A Delta9-desaturase. *J. Lipid Res.* **47**: 700–704.
39. Ternes, P., S. Franke, U. Zahringer, P. Sperling, and E. Heinz. 2002. Identification and characterization of a sphingolipid delta 4-desaturase family. *J. Biol. Chem.* **277**: 25512–25518.
40. Beaudoin, F., L. V. Michaelson, S. J. Hey, M. J. Lewis, P. R. Shewry, O. Sayanova, and J. A. Napier. 2000. Heterologous reconstitution in yeast of the polyunsaturated fatty acid biosynthetic pathway. *Proc. Natl. Acad. Sci. USA*. **97**: 6421–6426.
41. Matsuzaka, T., H. Shimano, N. Yahagi, T. Yoshikawa, M. Amemiya-Kudo, A. H. Hasty, H. Okazaki, Y. Tamura, Y. Iizuka, K. Ohashi, et al. 2002. Cloning and characterization of a mammalian fatty acyl-CoA elongase as a lipogenic enzyme regulated by SREBPs. *J. Lipid Res.* **43**: 911–920.
42. Toke, D. A., and C. E. Martin. 1996. Isolation and characterization of a gene affecting fatty acid elongation in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **271**: 18413–18422.
43. Oh, C. S., D. A. Toke, S. Mandala, and C. E. Martin. 1997. ELO2 and ELO3, homologues of the *Saccharomyces cerevisiae* ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. *J. Biol. Chem.* **272**: 17376–17384.
44. Lee, S. H., J. L. Stephens, K. S. Paul, and P. T. Englund. 2006. Fatty acid synthesis by elongases in trypanosomes. *Cell*. **126**: 691–699.
45. Livore, V. I., K. E. Tripodi, and A. D. Uttaro. 2007. Elongation of polyunsaturated fatty acids in trypanosomatids. *FEBS J.* **274**: 264–274.
46. Chertemps, T., L. Duportets, C. Labeur, and C. Wicker-Thomas. 2005. A new elongase selectively expressed in *Drosophila* male reproductive system. *Biochem. Biophys. Res. Commun.* **333**: 1066–1072.
47. Dobson, G., D. W. Griffiths, H. V. Davies, and J. W. McNicol. 2004. Comparison of fatty acid and polar lipid contents of tubers from two potato species, *Solanum tuberosum* and *Solanum phureja*. *J. Agric. Food Chem.* **52**: 6306–6314.
48. Karine, P., A. Paul, G. Andre, and J. T. Russell. 2006. Fatty acid composition of lipids from mushrooms belonging to the family Boletaceae. *Mycol. Res.* **110**: 1179–1183.
49. Damude, H. G., H. Zhang, L. Farrall, K. G. Ripp, J. F. Tomb, D. Hollerbach, and N. S. Yadav. 2006. Identification of bifunctional delta12/omega3 fatty acid desaturases for improving the ratio of omega3 to omega6 fatty acids in microbes and plants. *Proc. Natl. Acad. Sci. USA*. **103**: 9446–9451.
50. Zhang, S., E. Sakuradani, K. Ito, and S. Shimizu. 2007. Identification of a novel bifunctional delta12/delta15 fatty acid desaturase from a basidiomycete, *Coprinus cinereus* TD#822-2. *FEBS Lett.* **581**: 315–319.
51. Voss, A., M. Reinhardt, S. Sankarappa, and H. Sprecher. 1991. The metabolism of 7,10,13,16,19-docosapentaenoic acid to 4,7,10,13,16,19-docosahexaenoic acid in rat liver is independent of a 4-desaturase. *J. Biol. Chem.* **266**: 19995–20000.
52. Agaba, M., D. R. Tocher, C. A. Dickson, J. R. Dick, and A. J. Teale. 2004. Zebrafish cDNA encoding multifunctional fatty acid elongase involved in production of eicosapentaenoic (20:5n-3) and docosahexaenoic (22:6n-3) acids. *Mar. Biotechnol.* **6**: 251–261.
53. Lagali, P. S., J. Liu, R. Ambasudhan, L. E. Kakuk, S. L. Bernstein, G. M. Seigel, P. W. Wong, and R. Ayyagari. 2003. Evolutionarily conserved ELOVL4 gene expression in the vertebrate retina. *Invest. Ophthalmol. Vis. Sci.* **44**: 2841–2850.
54. Qiu, X., H. Hong, and S. L. MacKenzie. 2001. Identification of a Delta 4 fatty acid desaturase from *Thraustochytrium* sp. involved in the biosynthesis of docosahexaenoic acid by heterologous expression in *Saccharomyces cerevisiae* and *Brassica juncea*. *J. Biol. Chem.* **276**: 31561–31566.
55. Ferguson, M. A., and G. A. Cross. 1984. Myristylation of the membrane form of a *Trypanosoma brucei* variant surface glycoprotein. *J. Biol. Chem.* **259**: 3011–3015.
56. Hee Lee, S., J. L. Stephens, and P. T. Englund. 2007. A fatty-acid synthesis mechanism specialized for parasitism. *Nat. Rev. Microbiol.* **5**: 287–297.
57. Dobrzyn, A., and J. M. Ntambi. 2005. The role of stearyl-CoA desaturase in the control of metabolism. *Prostaglandins Leukot. Essent. Fatty Acids*. **73**: 35–41.
58. Jump, D. B. 2002. The biochemistry of n-3 polyunsaturated fatty acids. *J. Biol. Chem.* **277**: 8755–8758.
59. Sprecher, H. 1981. Biochemistry of essential fatty acids. *Prog. Lipid Res.* **20**: 13–22.
60. Marszalek, J. R., and H. F. Lodish. 2005. Docosahexaenoic acid, fatty acid-interacting proteins, and neuronal function: breastmilk and fish are good for you. *Annu. Rev. Cell Dev. Biol.* **21**: 633–657.
61. Tillman, J. A., S. J. Seybold, R. A. Jurenka, and G. J. Blomquist. 1999. Insect pheromones—an overview of biosynthesis and endocrine regulation. *Insect Biochem. Mol. Biol.* **29**: 481–514.
62. Roelofs, W. L., W. Liu, G. Hao, H. Jiao, A. P. Rooney, and C. E. Linn, Jr. 2002. Evolution of moth sex pheromones via ancestral genes. *Proc. Natl. Acad. Sci. USA*. **99**: 13621–13626.
63. Roelofs, W. L., and A. P. Rooney. 2003. Molecular genetics and evolution of pheromone biosynthesis in Lepidoptera. *Proc. Natl. Acad. Sci. USA*. **100**: 9179–9184.
64. Knipple, D. C., C. L. Rosenfield, R. Nielsen, K. M. You, and S. E. Jeong. 2002. Evolution of the integral membrane desaturase gene family in moths and flies. *Genetics*. **162**: 1737–1752.
65. Hulver, M. W., J. R. Berggren, M. J. Carper, M. Miyazaki, J. M. Ntambi, E. P. Hoffman, J. P. Thyfault, R. Stevens, G. L. Dohm, J. A. Houmard, et al. 2005. Elevated stearyl-CoA desaturase-1 expression in skeletal muscle contributes to abnormal fatty acid partitioning in obese humans. *Cell Metab.* **2**: 251–261.
66. Agatha, G., R. Hafer, and F. Zintl. 2001. Fatty acid composition of lymphocyte membrane phospholipids in children with acute leukemia. *Cancer Lett.* **173**: 139–144.