CURRENT OPINION
www.current-opinion.com

Missing genes in metabolic pathways: a comparative genomics approach

Andrei Osterman* and Ross Overbeek

The new techniques of genome context analysis — chromosomal gene clustering, protein fusions, occurrence profiles and shared regulatory sites — infer functional coupling between genes. In combination with metabolic reconstructions, these techniques can dramatically accelerate the pace of gene discovery.

Addresses
*Integrated Genomics, Inc., 2201 W. Campbell Park Drive, Chicago, IL 60612, USA
e-mail: andrei@integratedgenomics.com

Current Opinion in Chemical Biology 2003, **7**:238–251

This review comes from a themed section on Biocatalysis and biotransformation
Edited by Tadhg Begley and Ming-Daw Tsai

1367-5931/03/$ – see front matter

DOI 10.1016/S1367-5931(03)00027-9

Abbreviations

ACP	acyl carrier protein
CoA	coenzyme A
DMAPP	dimethylallyl diphosphate
DOXP	deoxyxylulose phosphate
DPCK	dephospho-CoA kinase
FAD	flavin adenine dinucleotide
FMN	flavin mononucleotide
IPP	isopentenyl diphosphate
PPAT	phosphopantetheine adenylyltransferase
SFA	saturated fatty acid
UFA	unsaturated fatty acids

Introduction

Comparative analysis of a large and growing number of diverse sequenced genomes is revolutionizing the pace of gene discovery. Consider the question: 'What is the most likely function of this gene?'. The most effective approach to answering such a question is based on projection of experimentally established functions of proteins from one species to another on the basis of homology, as revealed by sequence similarity. A set of powerful tools (such as BLAST and FastA) and public archives (such as GenBank and Swiss-Prot) are available to support such projection, as well as a significant body of literature (including recently published books [1,2••]).

Although the overall success of similarity-based tools has been remarkable, they fail to determine functions for many genes, and produce imprecise (and even incorrect) annotations for many others. These genes with no assigned function encode 20–60% of the proteins in most genomes, large or small, creating a well known *hypothetical proteins problem*. Ultimately, functional characterization of most of these hypothetical proteins will require advances in experimental biology; however, the emerging techniques of comparative genomics can dramatically reduce the efforts that will be required and have already increased the productivity of existing experimental technologies. Combining multiple new techniques in comparative genomics is often referred to as *genome context analysis*; it is the focus of many recent reviews and original research papers (some of them are listed in Table 1). A common theme of these efforts is the integration of various types of genomic evidence, such as clustering of genes on the chromosome [3], protein fusion events [4,5], occurrence profiles or signatures [6] and shared regulatory sites [7,8] to infer functional coupling for proteins participating in related cellular processes (e.g. enzymes involved in the same metabolic pathway). Application of these techniques for the analysis of all genes in a specific genome often produces valuable inferences [9••,10,11], which provide insight into a possible functional context but usually fall short of suggesting testable functional assignments, unless projected over a detailed reconstruction of relevant metabolic (or other cellular) pathways.

A metabolic reconstruction [12] is an attempt to develop a detailed overview of an organism's metabolism from an analysis of genomic sequence. This capability is a direct outgrowth of genomic sequencing and annotation efforts; a somewhat oversimplified summary of the technology would be that it supports inference of pathways on the basis of the presence or absence of relevant genes. Combining inferred pathways into hierarchical blocks produces metabolic charts specific for a particular organism and connected to individual genes [13,14••,15–19]. Metabolic reconstructions can reveal new aspects of metabolism in well-studied organisms (from *Escherichia coli* to humans), predict the metabolic potential of physiologically uncharacterized organisms, set the stage for network modeling [20], and support pathway re-engineering and the development of new therapies.

Since reconstruction technology is primarily focused on which components (e.g. metabolic enzymes) are actually present and which should be present but cannot be identified, it provides a rather specific and precise notion of what is actually missing [21]. This sets the stage for questions of the form, 'Which gene is most likely to play this given role?'. This question, which we define here as the *missing genes problem* is closely related to the

Table 1

Search for missing genes: major steps and techniques.

Milestones	Techniques	Fundamental concepts and observations	References	
			Background and implementation*	Applications†
I. Revealing missing genes				
Pathway reconstruction and projection of recognized orthologs across multiple diverse genomes				
List of relevant components (enzymes, transporters) in a functional context	Knowledge of metabolism	Template pathways: main routes and alternatives	[24]	
List of sequenced genes (groups of orthologs) connected to relevant functions	Homology-based searches	Sequence similarities: putative orthologs	[29,54]	
List of missing genes within a set of genomes	Metabolic reconstruction	Inferred pathways and functional systems	[12,14••,25,26]	
II. Identification and ranking of candidate genes				
Accumulation of genomic evidence of functional coupling and prioritization of candidate genes				
List of primary suspect candidate genes implicated by genomic evidence	Chromosomal clustering	Operons	[3,9••,30••,32••,37]	[28,62,63•,89•,92•,95••]
	Fusion events	Proteins with multiple functional domains	[4,5,41]	[69•,91]
	Occurrence profiles	Design commitments	[6,9••,42,43••]	[76,96]
	Shared regulatory sites	Regulons	[8,45–47]	[64•,87]
Prioritized list of candidates for further experimental verification	Long-range similarities and conserved motifs	Folds, superfamilies, ligand binding signatures	[51–53,55]	
	Biochemical and genetic data	Gene/protein features (phenotype, size, charge, localization, etc.)	[23]	
	Post-genomic data: microarrays, proteomics, gene knockouts	Co-expression profiles, physical interactions, gene essentiality	[56–59]	
III. Experimental verification				
New functional assignment for a protein family	Protein overexpression, purification, assays	Functional activity *in vitro*		
	Gene amplification, deletion, complementation	Functional activity *in vivo*		

*Many references and genomic resources described therein cover multiple techniques and they are relevant for more than one step in this analysis, nevertheless in this table they are cited only once. †Representative examples where one of the four techniques of genome context analysis provided the key evidence for a specific functional prediction followed by experimental verification.

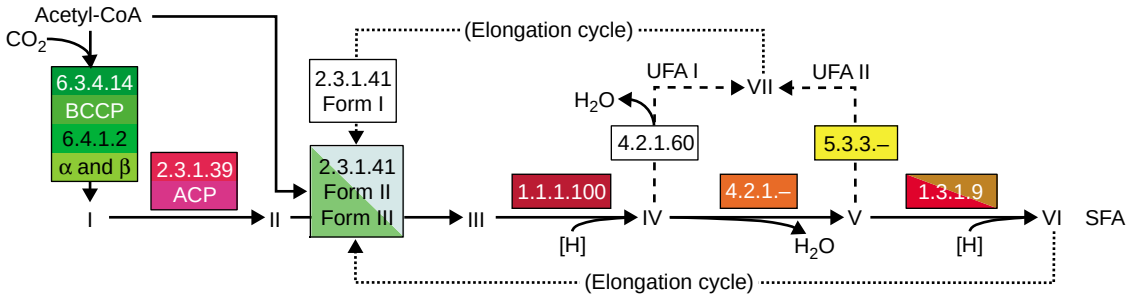
hypothetical proteins problem mentioned above — in both cases, one is attempting to connect functional roles to genes that have not yet been characterized.

Numerous instances of the hypothetical proteins problem are revealed with each sequenced genome. By contrast, just formulating a missing gene problem is dependent on the quality of pathway inference and a whole-genome metabolic reconstruction. At the same time, almost all experimental biologists are aware of one or more missing genes related to their immediate field of research. This specific and detailed knowledge, which is dispersed throughout the research community, is extremely difficult to integrate and encode for convenient computing. Therefore, with the current wealth of genomic information and sophisticated tools for comparative analysis, 'bench researchers' are in a much more favorable position to reveal numerous missing gene cases and generate reasonable predictions, let alone experimental verification, than their colleagues behind computer screens.

One major goal of this brief overview is to encourage experimental biologists and biochemists to use comparative genomics to search for missing genes involved with pathways and functional systems of their research interests. Since nothing can be more encouraging than a successful example, we decided to illustrate various aspects of contemporary techniques of genome context analysis using a set of representative examples. We limited our choice of examples to those published in the past two years, where functional predictions related to missing metabolic genes (predominantly enzymes) were made mostly by inferred functional coupling (rather than by similarity searches) and were immediately followed by experimental verification. We leave out a formal discussion and comparison of the various techniques

Figure 1

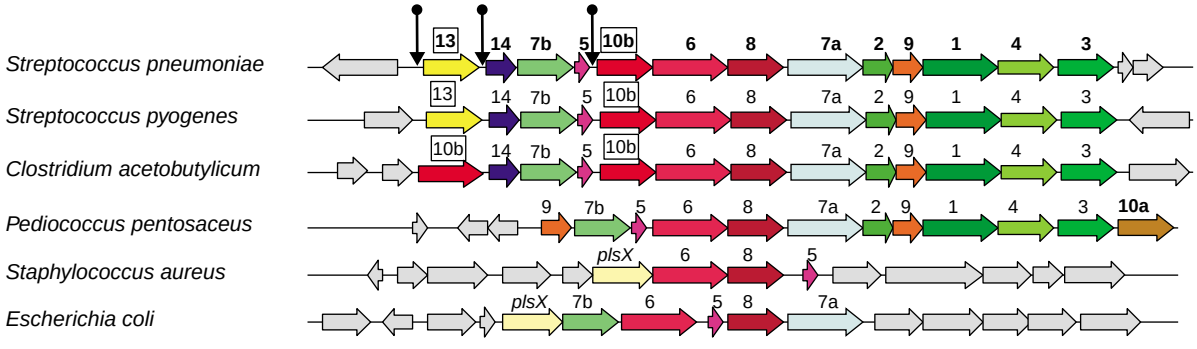
(a)



(b)

No	Enzymes and proteins directly involved in bacterial (type II) fatty acid biosynthesis	Gene name	EC number	*E. coli*	*S. aureus*	*P. pentosaceus*	*C. acetobutilicum*	*S. pyogenes*	*S. pneumoniae*
	Acetyl-CoA carboxylase								
1	Biotin carboxylase	***accC***	6.3.4.14	+	+	+	+	+	+
2	Biotin carboxyl carrier protein	***accB***		+	+	+	+	+	+
	Carboxyl-transferase		6.4.1.2						
3	Subunit α	***accA***		+	+	+	+	+	+
4	Subunit β	***accD***		+	+	+	+	+	+
	SFA biosynthesis								
5	ACP	***acpP***		+	+	+	+	+	+
6	Malonyl-CoA:ACP transacylase	***fabD***	2.3.1.39	+	+	+	+	+	+
	3-Ketoacyl-ACP-synthase		2.3.1.41						
7a	Form II	***fabF***		+	+	+	+	+	+
7b	Form III	***fabH***		+	+	+	+	+	+
8	3-Ketoacyl-ACP-reductase	***fabG***	1.1.1.100	+	+	+	+	+	+
9	3-Hydroxyacyl-ACP-dehydratase	***fabZ***	4.2.1.–	+	+	+	+	+	+
	Enoyl-ACP reductase		1.3.1.9						
10a	Form I	***fabI***		+	+	+	–	–	–
10b	Form II	***fabK****		–	–	+	+	+	+
	UFA biosynthesis								
11	**(UFA I)** 3-Ketoacyl-ACP-synthase; Form I	***fabB***	2.3.1.41	+	–	–	–	–	–
12	**(UFA I)** 3-Hydroxydecanoyl-ACP-dehydratase	***fabA***	4.2.1.60	+	–	–	–	–	–
13	**(UFA II)** *trans*-2,*cis* -3-Decenoyl-ACP isomerase	***fabM****	5.3.3.–	–	–	–	–	+	+

(c)



Current Opinion in Chemical Biology

and implementations, and we refer the reader to an excellent series of reviews and original research papers published on this subject in 2000 and 2002 (see Table 1).

Search for missing genes: the approach

The major steps and techniques used in a typical missing gene study are briefly described below and listed in Table I, where they are split in three phases: (I) building a case, (II) evidence accumulation and analysis, and (III) experimental verification. In reality, researchers often have pre-existing knowledge of a particular missing gene case in a target organism. Nevertheless, going through the first steps will help to strengthen such a case by checking for possible inconsistencies in sequencing data, annotations and pathway interpretation.

From a practical perspective, one may distinguish two categories of missing genes, which (for the lack of better terms) we will refer to as *globally missing* (for functions without any representative sequenced genes from any organism), and *locally missing* (for functions previously connected to one sequenced form of a gene in one group of species, but expected to exist in an alternative form in another group of species). Massive genome sequencing and comparative analysis has revealed an unexpectedly high frequency of non-orthologous gene displacements [22], which probably account for the majority of locally missing gene cases. In some cases, these alternative forms share the same fold and/or conserved motifs, implying an extremely divergent evolution, whereas in many other cases, no sequence/structure similarity is observed, suggesting that the same function could be 'invented' independently more than once.

Phase I: revealing missing genes

The determination that a specific enzyme is missing is made by compiling evidence supporting the existence of a specific pathway within an organism, identifying the specific genes that encode functions of the pathway, and then focusing on specific functions that cannot be connected to genes.

Step 1: establishing functional context

The search for one of these missing genes begins by computing a 'functional context', which usually amounts to the other enzymes that participate in the same pathway or variants of the pathway. To support this analysis, one uses traditional sources of biochemical information [23] (including books, such as [24]), supplemented by available public and commercial web-resources and databases (such as the electronic Biochemical Pathways Chart available from the ExPaSy server at www.expasy.org, KEGG [14••], ERGO [25] and PGDB [26]).

Step 2: gene inventory

Once the set of closely related functions has been determined, one builds a table showing which of these functions is present or absent within a diverse set of model organisms. The table contains a row for each of the enzymatic functions, and a column for each organism. Each cell contains genes believed to be instances of the functional role in a specific organism, inferred by homology analysis (limitations of homology-based functional annotations have been discussed, for example, see [27]). The construction of such a table has been described [15,28], and is illustrated here in Figure 1b. Available whole-genome annotations, as well as collections of protein families (such as clusters of orthologous groups (COGs) at NCBI [29]) are perfect starting points for this analysis. This table is the raw data for beginning to understand which organisms have variants of the pathway, which do not, and where the situation remains ambiguous.

Step 3: metabolic reconstruction

Once the gene inventory has been composed, the next step is to formulate an assessment of exactly what variants of the

(Figure 1 Legend) Missing genes in fatty acid biosynthesis and chromosomal clustering. **(a)** Pathway diagram. Simplified representation of major enzymatic steps in fatty acid biosynthesis. Before entering the cycle malonyl-CoA (**I**) is produced from acetyl-CoA, and malonyl residue is transferred to ACP to form malonyl-ACP (**II**). The first step of the SFA cycle is a condensation with another molecule of acetyl-CoA affording β-ketoacyl-ACP (**III**). This undergoes consecutive reduction to β-hydroxyacyl-ACP (**IV**), dehydration to *trans*-2-enoyl-ACP (**V**), and another reduction to acyl-ACP (**VI**) to enter the next elongation cycle (dotted arrow). Two alternative branching-out pathways of unsaturated fatty acid biosynthesis as known in *E. coli* (UFA I), and proposed for *S. pneumoniae* (UFA II) [63•] are shown by dashed arrows. Both proceed by an isomerization step to produce *cys*-3-enoyl-ACP (**VII**), which enters further elongation as shown by a dotted line. Structural formula of all intermediates in this and other figures are provided in the Supplementary Material (URL: http://www.integratedgenomics.com/online_material/osterman/index.html; Table S1). Enzymes are indicated by standard enzyme classification (EC) numbers explained in panel (b). The shading reflects correspondence to specific genes, as in (b,c). **(b)** Metabolic reconstruction. A list of major enzymes and protein components of bacterial FAS II. Gene names are as in *E. coli*, except for *fabK* and *fabM* (marked by an asterisk), recently discovered in *S. pneumoniae* and related species. Presence or absence of corresponding orthologous genes in a given genome is marked by '+' or '−' respectively. Numbers and colors are the same as in panel (c). (A third form of enoyl-ACP reductase gene (*fabL*, previously *ygaA*) recently identified in *B. subtilis* [94] is not present in any of the selected genomes, and it is not shown in this panel.) **(c)** Chromosomal clustering. The alignment of chromosomal regions 'pinned' around one of the FAS II genes (*fabG*) in *S. pneumoniae* and related species. Clustering of orthologous FASII-related genes (with corresponding colors to (b)) provided key evidence for the identification of two novel enzymes (*missing genes*) involved with SFA and UFA II pathways: *fabK* (11b) and *fabM* (13), respectively. Additionally, a putative UFA type II transcription regulator (14) and a protein of unknown function related to lipid biosynthesis (*plsX*) are outlined. Other genes that are not conserved in this neighborhood, and do not directly participate in fatty acid biosynthesis are colored gray. Note that a gene arrangement in *P. pentosaceus* is very similar to *S. pneumoniase*, *S. pyogenes* and *C. acetobutylicum*, with a most notable 'disappearance' of *fabK* (11b) in the middle of the cluster compensated by the 'appearance' of *fabI* (11a) at the end of the cluster. Multiple instances of a predicted regulatory site with a consensus sequence ac**TTTGA**twa**TCAAA**gt, are indicated in *S. pneumoniae* operon by arrows.

pathway are present in what organisms, in the process identifying which genes actually remain missing. There are numerous factors that complicate this analysis, including those related to non-committed enzymes (existing in multiple pathways) and enzymes with broad specificities.

Phase II: identification and ranking of candidate genes

Various techniques of genome context analysis are used to infer functional coupling and produce an initial list of candidate genes for a sought functional role. We briefly list the major techniques of missing gene analysis and the most relevant publications. We refer the reader to a more detailed (although still very sketchy) description of the approach and selected examples in the Supplementary Materials (URL: http://www.integratedgenomics.com/online_material/osterman/index.html).

Technique 1: clustering on the chromosome

Genes from the same pathway tend to cluster on prokaryotic chromosomes. This can be exploited to infer 'functional coupling' between genes [3]. Genome-scanning tools are used to look for cases in which it appears that multiple genes orthologous to members of the gene inventory occur in close proximity. [30••,31•,32••]. Background and application of this technique for enhancement of genome annotations are discussed in detail in several recent research papers and overviews [33•,34–37] (Figure 1).

Technique 2: protein fusion events

This technique involves searches for a pair of genes from one genome that appear to be fused into a single gene within another genome, providing further evidence of potential functional coupling. Since its introduction [4,5], the protein fusion approach has been implemented and successfully applied for genome-wide hypothetical protein analysis, mostly in combination with other techniques [38,39,40•,41] (Figure 2).

Technique 3: occurrence profiles

This approach [6] (often referred to as 'phylogenetic profiling') brings a truly independent type of genomic evidence. In a simplified form, the underlying assumption is that two proteins from the same cellular pathway are expected to either both occur or both not occur in any specific organism. The high-throughput version of this technique, implemented by various groups [9••,42,43••], generates instances of potential functional coupling for a pair of proteins on the basis of their occurrence profiles. Some users may find a simplified version of this technique more efficient for missing gene analysis (Figure 3b). Its application for the identification of uncharacterized bacterial photosynthetic proteins was recently described [44].

Technique 4: shared regulatory sites

This technique focuses on identification of so-called *regulons* (ensembles of genes subject to coordinated expression). Co-regulation of a pair of genes provides evidence that these genes may be functionally coupled. Recent publications describe new and improved algorithms to identify shared regulatory sites and putative regulons [45–47]. Attempts to apply this technique for gene discovery are at an early stage, and we are aware of only a limited number of functional predictions for previously uncharacterized proteins on the basis of shared regulatory sites. In a recent series of publications, a significant number of specific functional predictions were based on analysis of extremely conserved regulatory signals associated with genes involved in the biosynthesis of some vitamins [48,49,50••] (Figure 4).

Ranking candidate genes and additional types of evidence

The functional-context-based techniques described above produce partially overlapping conjectures that can be further prioritized based on strength and consistency of evidence. Among other techniques broadly used in gene discovery and also very helpful for additional candidate ranking are the methods revealing and analyzing putative folds [51–53], long-range sequence similarities [54] and conserved motifs [55]. An integration of the vast amounts of experimental data generated by post-genomic techniques, such as expression microarrays, protein–protein interaction analysis [56,57], and less established whole-genome conditional gene essentiality studies [58,59], provide us with an additional source of functional links for gene discovery.

Phase III: experimental verification

In most cases, the number of highly ranked gene candidates is very limited and they can be quickly challenged by traditional experimental techniques of experimental biology.

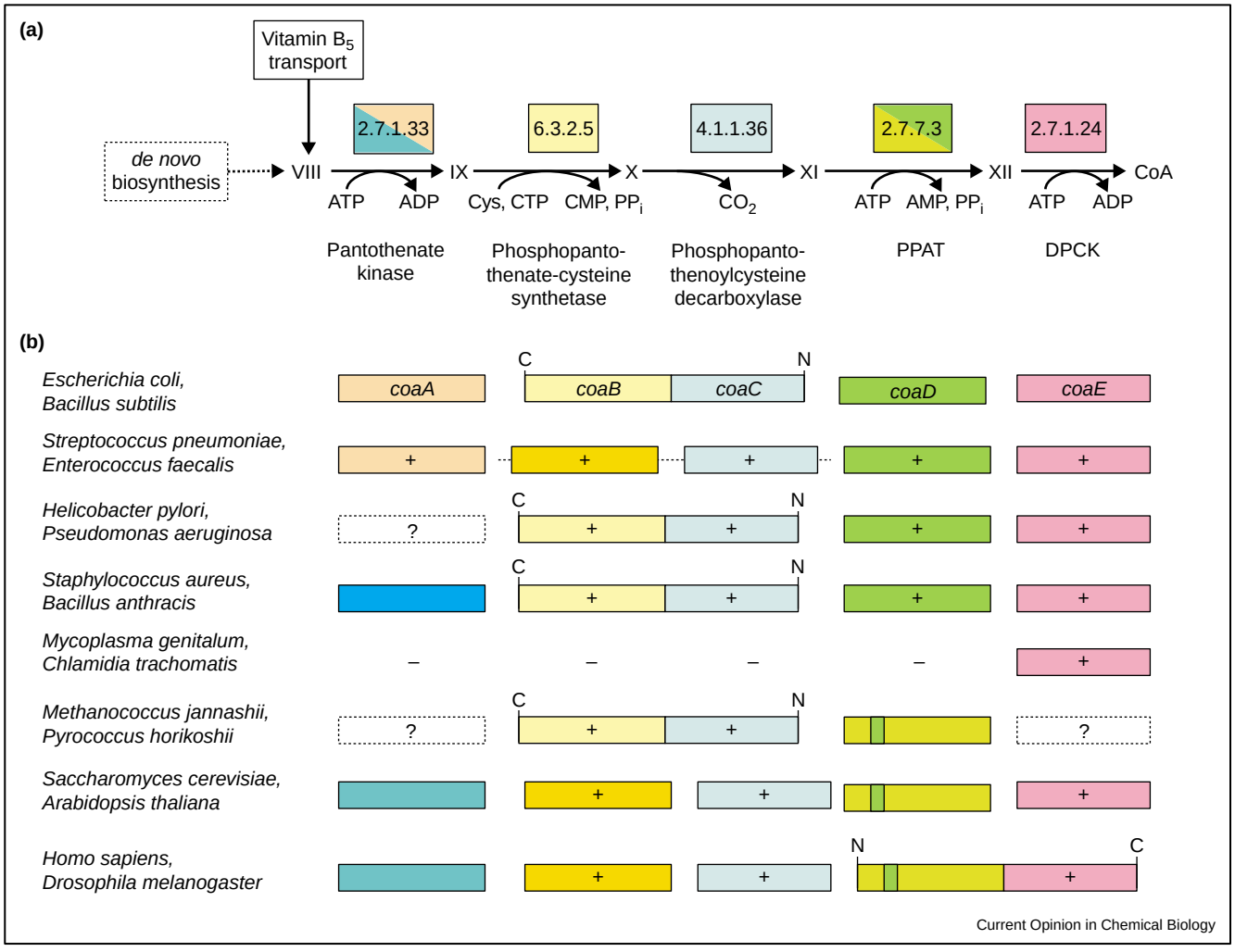
Missing genes in metabolic pathways: case studies

The following examples were selected from recent publications to illustrate applications of the four major techniques of genome context analysis. All of these examples contain functional predictions related to the most important metabolic pathways in the central machinery of life, followed by direct experimental verifications. We have found it impossible to adequately condense all of the important details of these examples. Therefore we only briefly introduce them here and provide a more expanded discussion in the Supplementary Materials at URL: http://www.integratedgenomics.com/online_material/osterman/index.html.

Fatty acid biosynthesis in *Streptococcus pneumoniae*: chromosomal clustering

Biosynthesis of fatty acids in bacteria (for a simplified diagram see Figure 1a) is a rich source of anti-infective drug targets [60,61]. Almost all of the essential components of fatty acid synthase complex producing saturated fatty acids (SFAs) can be projected by sequence similarity

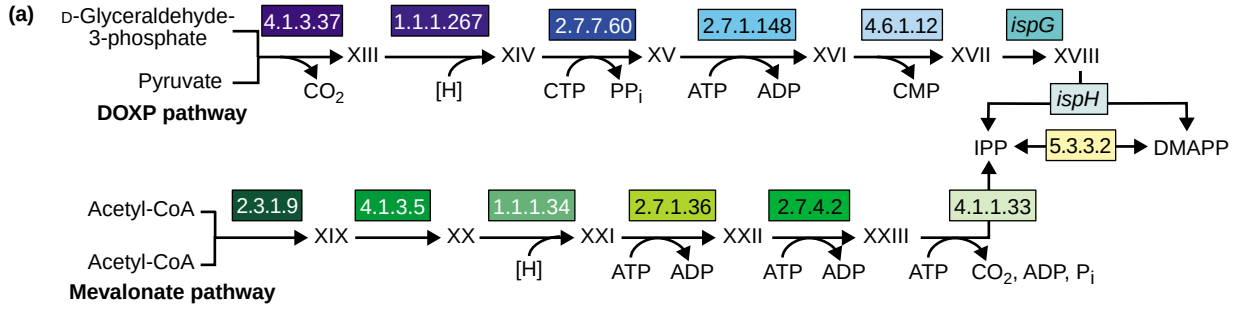
Figure 2



Missing genes in CoA biosynthesis and protein fusion events. **(a)** Pathway diagram. Five-step universal CoA biosynthetic pathway and enzymes involved therein (based on [66]). Pantothenate (**VIII**) produced *de novo* or salvaged from the medium is phosphorylated to produce 4′-phosphopantothenate (**IX**), which undergoes condensation with Cys affording 4′-phosphopantothenoylcysteine (**X**), and decarboxylation to 4′-phosphopantetheine (**XI**). Adenylyltransferase reaction yields dephosphocoenzyme A (**XII**), which gets further phosphorylated to the final form of CoA cofactor. **(b)** Domain arrangement. Orthologs of *E. coli* CoA biosynthetic enzymes in representative bacteria, archaea and eukarya are shown by boxes marked '+'. *Missing genes* (expected but unidentified) are indicated by uncolored boxes marked with '?', whereas those absent due to pathway truncations (as in *Mycoplasma* and *Chlamydia* spp.) are indicated by '–'. The eukaryotic form of pantothenate kinase belongs to a distinct structural class (marked by distinct color). In *S. aureus* (as well as in *B. anthracis*) a distant homolog (darker color) of the eukaryotic pantothenate kinase replaces a typical bacterial enzyme. Enzymes for the second and the third steps form a fusion protein in archaea and most bacteria (domain arrangement is indicated by positions of N- and C-termini), except for *Streptococci* and *Enterococci*, where a pair of monofunctional genes form a tight operon *coaB–coaC*. Eukaryotes also contain two monofunctional proteins, and one of them (corresponding to *coaB*) is significantly more divergent (darker color). Eukaryotic PPAT shows no sequence similarity with its bacterial counterpart (*coaD* gene) beyond an NTP-binding motif (as indicated by a light green stripe) and predicted common Rossman fold. In humans, this enzyme forms a fusion protein with a C-terminal domain clearly homologous to bacterial dephosphoCoA-kinase. This fusion event provided the major clue for functional prediction. Archaeal PPAT (closely related to the eukarytotic form) was independently identified by the research group at Virginia Polytech (R White, personal communication). The last enzymatic step in CoA biosynthesis appears to be a missing gene in all archaea.

from *E. coli* to other bacteria (see Figure 1b). However, orthologs of the *fabI* gene, encoding enoyl-ACP-reductase (a proven target for such drugs as izoniaside and triclosan; ACP = acyl carrier protein), are not found in *Streptococcus pneumoniae* and a group of related species, producing a case of a *locally missing gene*. The key evidence for the identification of a novel bacterial enoyl-ACP reductase (gene *fabK*) was provided by gene clustering on the chromosome (see Figure 1c). The prediction was verified by enzymatic characterization of the corresponding recombinant protein *in vitro*, and by genetic complementation of a *fabI* mutant of *E. coli* [62].

Figure 3



(b)

Step	Enzymes involved in isoprenoid biosynthesis	EC number	*Escherichia coli K12*	*Bacillus cereus*	*Xylella fastidiosa*	*Fusobacterium nucleatum*	*Brucella melitensis*	*Borrelia burgdorferi*	*Staphylococcus aureus*	*Streptococcus pneumoniae*	*Lactococcus lactis*	Co-occurrence score[1]
	DOXP pathway											
1	1-Deoxy-D-xylulose 5-phosphate synthase[2]	4.1.3.37	*dxs*	+	+	+	+	–	–	–	+	<0.7
2	1-Deoxy-D-xylulose 5-phosphate reductoisomerase	1.1.1.267	*ispC*	+	+	+	+	–	–	–	–	1.00
3	4-Diphosphocytidyl-2C-methyl-D-erythritol synthase	2.7.7.60	*ispD*	+	+	+	?	–	–	–	–	0.96
4	4-Diphosphocytidyl-2C-methyl-D-erythritol kinase[3]	2.7.1.148	*ispE*	+	+	+	+	–	+[3]	–	–	<0.7
5	2C-Methyl-D-erythritol2,4-cyclodiphosphate synthase	4.6.1.12	*ispF*	+	+	+	+	–	–	–	–	0.96
6	1-Hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate synthase[4]		*ispG*	+	+	+	+	–	–	–	–	1.00
7	1-Hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate reductase[4]		*ispH*	+	+	+	+	–	–	–	–	1.00
8	Isopentenyl-diphosphate delta-isomerase[5]	5.3.3.2	*(idi)*	+	–	–	–	(+)	(+)	(+)	(+)	<0.7
	Mevalonate pathway											
1	Acetyl-CoA-acetyltransferase[2]	2.3.1.9	*(atoB)*	(+)	–	(+)	(+)	+	+	+	+	
2	Hydroxymethylglutaryl-CoA synthase	4.1.3.5	–	–	–	–	–	+	+	+	+	
3	Hydroxymethylglutaryl-CoA reductase	1.1.1.34	–	–	–	–	–	+	+	+	+	
4	Mevalonate kinase	2.7.1.36	–	–	–	–	–	+	+	+	+	
5	Phosphomevalonate kinase	2.7.4.2	–	–	–	–	–	+	+	+	+	
6	Diphosphomevalonate decarboxylase	4.1.1.33	–	–	–	–	–	+	+	+	+	
7	Isopentenyl-diphosphate delta-isomerase	5.3.3.2	*(idi)*	(+)	–	–	–	+	+	+	+	

'In-group' representative | 'Out-group' representative

Current Opinion in Chemical Biology

Missing genes in isoprenoid biosynthesis and occurrence profiling. **(a)** Pathway diagram. Simplified representation of major enzymatic steps in the two alternative pathways of isoprenoid biosynthesis. In the DOXP-pathway, formation of 1-deoxy-D-xylulose 5-phosphate (**XIII**) is followed by NADPH-dependent reduction to 2C-methyl-D-erythritol 4-phosphate (**XIV**). The next intermediate, 4-diphosphocytidyl-2C-methyl-D-erythritol (**XV**), is produced by cytidyl-transferase reaction followed by phosphorylation to 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate (**XVI**), and CMP elimination/cyclization producing 2C-Methyl-D-erythritol 2,4-cyclodiphosphate (**XVII**). The final intermediate, 1-hydroxy-2-methyl-2(*E*)-butenyl 4-diphosphate (**XVII**) is converted to a mixture of the major isoprenoid building blocks IPP and DMAPP by the action of a single enzyme. In the alternative mevalonate pathway, the first intermediate, acetoacetyl-CoA (**XIX**), is converted to hydroxymethylglutaryl-CoA (**XX**), and then to mevalonate (**XXI**). The latter undergoes two consecutive phosphorylation steps to phosphomevalonate (**XXII**) and diphosphomevalonate (**XXIII**), followed by decarboxylation to IPP, which is further isomerized to DMAPP. IPP isomerase (EC 5.3.3.2), the only common enzyme in these two pathways, is optional for the DOXP pathway but indispensable for the mevalonate pathway. Merger of the two pathways in this diagram is not a pure abstraction, as both occur in some bacteria such as *Listeria monocytogenes* and *Mycobacterium marinum* (as well as in different compartments of plant cells). **(b)** Metabolic reconstruction and occurrence profiles. List of relevant enzymes and occurrence of corresponding genes within a set of representative bacterial genomes from the 'in-group' (DOXP pathway-dependent) and 'out-group' (mevalonate pathway-dependent). Presence or absence of putative orthologs is indicated by '+' and '−'. No orthologs of *ispD* are found in the completely sequenced genome of *B. melitensis*

S. pneumoniae (and many other species) also lack orthologs of *fabA* and *fabB* genes, which are involved with unsaturated fatty acid biosynthesis (UFA I) in *E. coli*. The same extended chromosomal cluster enabled the prediction and verification of a novel pathway (UFA II) in *S. pneumoniae* [63•], including a novel *trans-2,cis*-3-decenoyl-ACP isomerase (*fabM* gene, number 13 in Figure 1c)). Multiple instances of putative regulatory site (acTTTGAtwaTCAAAgt), and a predicted transcription regulator (HTH protein) are located within this cluster (Figure 1c), strengthening the functional prediction and suggesting a regulatory mechanism for UFA II. (This consensus, present in upstream regions of relevant genes, was independently derived from the analysis of six streptococcal and enterococcal genomes (M. Gelfand, Integrated Genomics, Moscow, Russia, unpublished data); however, the functional relevance of this observation was unclear until the identification of fabM, which led to elucidation of UFAII pathway and regulons.)

The identity of another missing gene related to fatty acid metabolism, encoding acyl coenzyme A dehydrogenase (gene *fadE*), was recently established [64•]. A previously uncharacterized *E. coli* gene, *yafH*, was implicated by the analysis of shared regulatory sites [65] and microarray expression data, and verified by direct genetic experiments [64•].

Human coenzyme A biosynthesis: protein fusions

Biosynthesis of coenzyme A (CoA) from pantothenate (vitamin B_5) by a universal five-step pathway, is schematically illustrated in Figure 2a (for a recent review see [66]). Bacterial genes encoding all of the enzymes in this pathway (*coaA* through *coaE*, see Figure 2b) were identified and characterized in *E. coli*. Until recently, only one of the human enzymes in this pathway (pantothenate kinase, structurally unrelated to the bacterial enzyme [67]) was connected to a particular gene. Similarity-based projection from bacterial genes allowed identification of human genes encoding all of the remaining enzymes, except phosphopantetheine adenylyltransferase (PPAT). This enzyme represented a typical case of a locally missing gene, and the key evidence for its elucidation was provided by a protein fusion (Figure 2b). On the basis of early biochemical data [68], a cDNA encoding a multidomain human protein with a C-terminal domain homologous to bacterial dephospho-CoA kinase (DPCK) was identified, and both predicted activities (DPCK and PPAT) were verified by enzymatic characterization of the purified recombinant protein [69•]. Two more research groups simultaneously reported identification and verification of the same human PPAT/DPCK gene [70,71], illustrating the impact of comparative genomics on modern gene discovery.

The current picture of the CoA biosynthetic pathway reveals a pronounced conservation of its enzymatic components across taxons (Figure 2b). At the same time, significant variations are observed at the level of individual enzymes, including non-orthologous gene displacements, domain fusions and what are likely to be lateral gene transfer events.

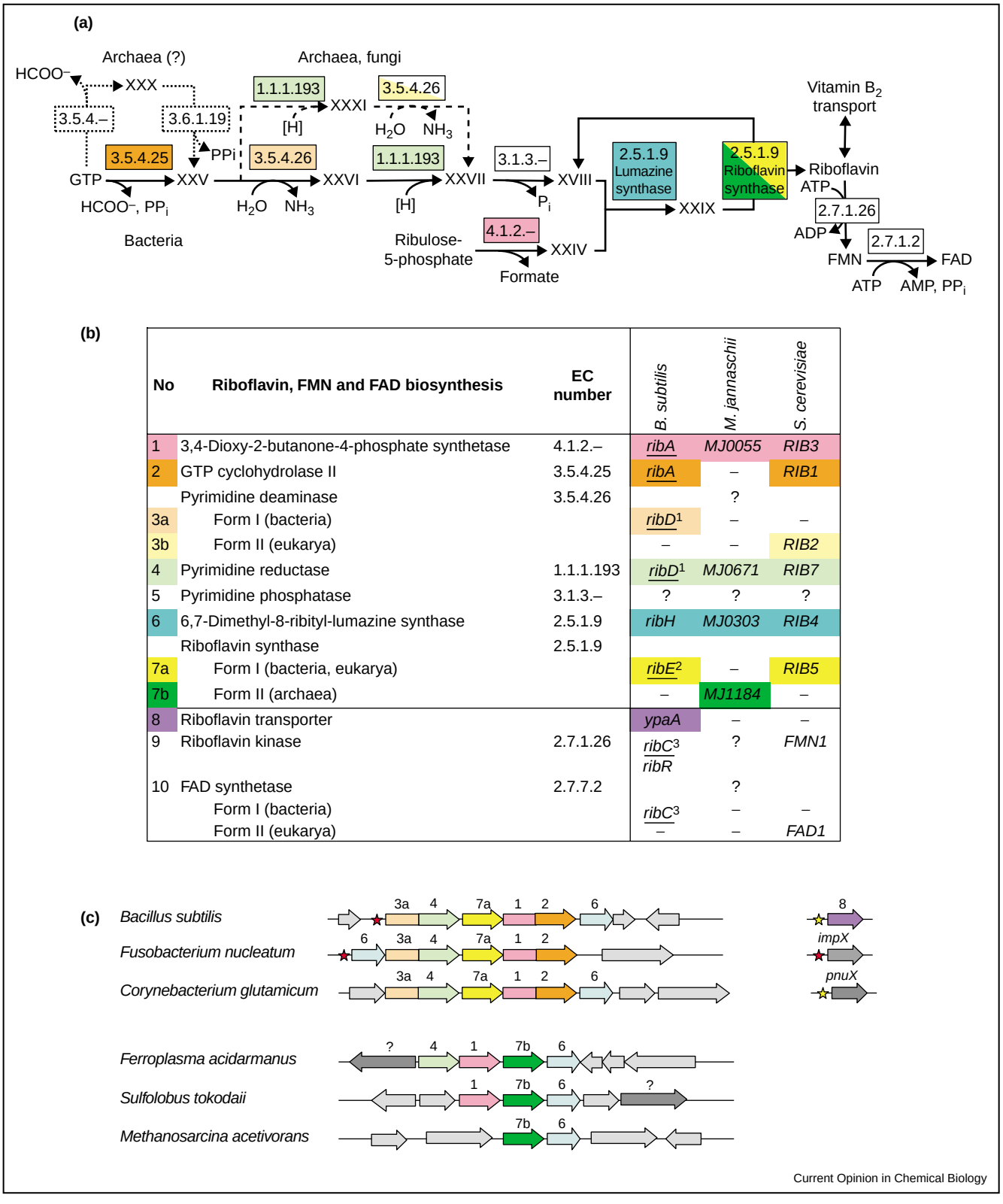
Nonmevalonate (deoxyxylulose phosphate) isoprenoid biosynthesis: occurrence profiles

Major terpenoid building blocks, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are produced by two different biosynthetic routes: in some species by the so-called *mevalonate* pathway, and in others by the *non-mevalonate* or deoxyxylulose phosphate (DOXP) pathway (Figure 3a). Historically, the mevalonate pathway and its enzymes have been thoroughly studied in eukaryotes. Some bacteria also use the mevalonate pathway, and all of the corresponding genes were identified on the basis of homology with eukaryotic counterparts [72,73]. Reconstruction of the mevalonate pathway in archaea, including conjectures for some locally missing genes, was recently described [74•]. The alternative DOXP pathway, characteristic of most bacteria, was not recognized until very recently (for a review see [75]), and some aspects of it remained obscure until last year.

The DOXP pathway provides a striking example of using occurrence profiles for missing gene analysis. In the original study, two uncharacterized *E. coli* genes (*gcpE* and *lytB*, now renamed to *ispG* and *ispH*) were implicated by their co-occurrence with DOXP genes known at that time (Figure 3b), and experimental evidence was provided for one of them (*lytB*) [76]. Later genetic experiments unambiguously confirmed *gcpE* and *lytB* participation in the last steps of the DOXP pathway, and experimental studies published within the past year have clarified corresponding reactions and enzymatic functions [77•,78,79•].

(Figure 3 Legend Continued) (as well as in *Desulfitobacterium halfniense* and *Mezorizobium loti*), suggesting another case of a locally missing gene (marked by '?'). [1]Co-occurrence scores for each protein were computed as the total number of genomes in the 'in-group' (maximum of 28) containing a homolog of a given protein minus the number of genomes in the 'out-group' (maximum of 10) containing such a homolog (using a FastA P-score cut-off 10^{-5}), normalized by a highest possible score (of 28). [2]Both enzymes participating at the first step of each pathway are not 'committed' to isoprenoid production, and their occurrence profiles deviate significantly. [3]Close homologs of 4-diphosphocytidyl-2C-methyl-D-erythritol kinase occur in several genomes of 'out-group', such as *S. aureus*. [4]Genes coding for the last two steps of DOXP pathway, *ispG(gcpE)* and *ispH(lytB)*, were originally implicated with this pathway on the basis of occurrence profiling [76]. [5]Orthologs of IPP isomerase are present in both groups of genomes. This activity is optional for the DOXP pathway but is absolutely required in the mevalonate pathway. [6]Mevalonate kinase and phosphomevalonate kinase in *Streptococci* (and in archaea) are closely related by sequence and are often located next to each other on the chromosome, causing annotation errors in many archives.

Figure 4



No	Riboflavin, FMN and FAD biosynthesis	EC number	*B. subtilis*	*M. jannaschii*	*S. cerevisiae*
1	3,4-Dioxy-2-butanone-4-phosphate synthetase	4.1.2.–	*ribA*	*MJ0055*	*RIB3*
2	GTP cyclohydrolase II	3.5.4.25	*ribA*	–	*RIB1*
	Pyrimidine deaminase	3.5.4.26		?	
3a	Form I (bacteria)		*ribD*[1]	–	–
3b	Form II (eukarya)		–	–	*RIB2*
4	Pyrimidine reductase	1.1.1.193	*ribD*[1]	*MJ0671*	*RIB7*
5	Pyrimidine phosphatase	3.1.3.–	?	?	?
6	6,7-Dimethyl-8-ribityl-lumazine synthase	2.5.1.9	*ribH*	*MJ0303*	*RIB4*
	Riboflavin synthase	2.5.1.9			
7a	Form I (bacteria, eukarya)		*ribE*[2]	–	*RIB5*
7b	Form II (archaea)		–	*MJ1184*	–
8	Riboflavin transporter		*ypaA*	–	–
9	Riboflavin kinase	2.7.1.26	*ribC*[3] *ribR*	?	*FMN1*
10	FAD synthetase	2.7.7.2		?	
	Form I (bacteria)		*ribC*[3]	–	–
	Form II (eukarya)		–	–	*FAD1*

Missing genes in riboflavin biosynthesis and conserved regulatory sites. **(a)** Pathway diagram. Simplified diagram of riboflavin (vitamin B_2) biosynthesis and conversion to FMN and FAD cofactors [80]. One of the committed precursors L-3,4-dihydroxy-2-butanone 4-phosphate (**XXIV**) is produced in one step from ribulose-5-phosphate. Conversion of GTP to 2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5′-phosphate (**XXV**),

To evaluate the impact of the significant increase in the number and diversity of sequenced genomes we have reproduced this analysis using only bacterial genomes (as illustrated in Figure 2b). In addition to the components of the DOXP pathway, this analysis revealed a limited number of genes with high occurrence scores. A significant fraction of these genes is related to thiamin and NAD biosynthesis, possibly revealing some common metabolic design commitments.

Riboflavin biosynthesis: shared regulatory sites

Riboflavin (vitamin B_2) is an ultimate precursor in the biosynthesis of two redox cofactors: flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD). Many aspects of riboflavin/FMN/FAD biosynthesis (for a simplified diagram see Figure 4a) are largely conserved across all taxons (for a recent review see [80]). The most significant variations occur in archaea, where some homologous and non-homologous forms of previously known enzymes have already been characterized [81•,82,83], but several enzymatic steps are still associated with missing genes (see Figure 4b).

For a long time, regulatory elements and mechanisms in riboflavin biosynthesis remained completely obscure. A novel regulatory mechanism mediated directly by FMN [84] was proposed on the basis of early experimental work in *Bacillus subtilis*, and comparative cross-genome analysis of upstream regions adjacent to operons and individual genes of riboflavin biosynthesis. Direct experimental verification of this mechanism, which involves alternative secondary structure formation by a conserved regulatory element (termed RFN) was recently published [85,86]. A search for additional occurrences of RFN-like sequences enabled the prediction and experimental verification of a missing riboflavin transporter (*ypaA*) in *B. subtilis* [87]. An extended comparative analysis of riboflavin biosynthetic genes in a broad range of bacterial genomes implicated more proteins as alternative riboflavin transporters in other species (see Figure 4c) [50••].

Miscellaneous examples: additional techniques

We have illustrated the application of the major techniques of genome context analysis by the analysis of four representative examples. In the Supplementary Materials (URL: http://www.integratedgenomics.com/online_material/osterman/index.html) we provide more details related to these and additional examples, including various biosynthetic enzymes in archaea, which are especially rich with missing genes [88]. Interesting examples of missing gene analysis are related to NAD biosynthesis [89•], tRNA-modification [90], thymidine biosynthesis [33•,91,92•] and propionyl-CoA metabolism [93]. The latter example provides an illustration of using gene clustering on prokaryotic chromosome as a key evidence for elucidation of a missing methylmalonyl-CoA racemase gene in humans. As the authors of this study, we also believe that this approach will soon gain much more popularity.

Conclusions: missing genes and central machinery

It is possible to systematically search for missing genes that encode metabolic enzymes, using a variety of emerging techniques. The use of these techniques to guide experimental efforts is improving the productivity of the experimental analysis, and we believe that this trend will accelerate. We have sketched, in the briefest terms, some of the more useful techniques. The reader who takes the time required to read the cited references and analyze these early success stories will almost inevitably begin to understand the enthusiasm that is growing. The underlying bioinformatic algorithms are believed to increase in power as the square of the number of complete genomes available. If this tendency turns out to be accurate, the hundreds of genomes that will become available in the next two years will dramatically enhance techniques that are already impressive.

Among all of the contemporary techniques of genome context analysis, gene clustering on the chromosome

(Figure 4 Legend Continued) followed by deamination to 5-amino-6-ribosylamino-2,4(1H,3H)-pyrimidinedione 5′-phosphate (**XXVI**), reduction to 5-amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione 5′-phosphate (**XXVII**) and dephosphorylation, yields another precursor 5-amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione (**XXVIII**). Condensation of these two precursors yields 6,7-dimethyl-8-ribityl-lumazine (**XXIX**). Two molecules of **XXIX** produce one molecule of riboflavin, while regenerating one molecule of **XXVIII**. Riboflavin is converted to flavin cofactors FMN and FAD by consecutive phosphorylation and adenylyltransferase reactions. Universal enzymatic steps and those characteristic of bacteria are shown by solid arrows. In methanogenic archaea, conversion of GTP to **XXV** was hypothesized to proceed in two steps (dotted arrows) via 2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5-triphosphate intermediate (**XXX**) [81•]. Deamination and reduction reactions were shown to occur in opposite order in archaea and yeast (dashed arrows), via 2,5-diamino-6-ribitylamino-4(3H)-pyrimidinone 5-phosphate (**XXXI**) intermediate. **(b)** Metabolic reconstruction. Enzymatic components of the pathway and corresponding genes are shown for the representative bacterial, archaeal and eukaryotic genomes. Identical bacterial gene names (underlined) associated with distinct enzymatic steps reflect fusion of corresponding functional domains. Absence of corresponding orthologs is marked by '–'; missing genes (such as globally missing pyrimidine phosphatase) are indicated by '?'. **(c)** Chromosomal arrangement and RFN regulatory sites. The alignment of chromosomal regions 'pinned' around two non-homologous forms of riboflavin synthase (7a, 7b) in selected bacteria and archaea. Orthologous genes conserved within displayed chromosomal neighborhoods are outlined by matching colors and labeled by the same numbers as in (b). A conserved uncharacterized gene, a proposed candidate for a missing archaeal pyrimidine deaminase is marked by pattern and '?'. Instances of the conserved regulatory element (RFN) with two predicted types of regulation, at the level of transcription and translation, are marked by red and yellow stars, respectively. In many cases, RFN elements are adjacent to bacterial rib-operons, and also to isolated genes in distal chromosomal loci, such as proven (8, *ypaA*) and inferred (*impX*, *pnuX*) flavin transporters. [1]Previously *ribG*; [2]previously *ribB*; [3]*ribF* in *E. coli*.

provides the single most critical contribution to missing gene discovery. Notwithstanding emerging evidence of chromosomal gene clustering in simple eukaryotes, this technique is almost exclusively applicable for the comparative analysis of prokaryotic genomes. The same is largely true for the analysis of shared regulatory sites, and to some extent for occurrence profiling, which is critically dependent on the number and diversity of complete genomes with well-defined genes.

At the same time, large-scale sequencing and comparative analysis of multiple and diverse prokaryotic genomes provide growing evidence that for an overwhelming majority (>90% by our estimates) of eukaryotic metabolic enzymes (or more generally, any protein components involved in the *central machinery of life*) it is possible to find functional counterparts (homologous or analogous) in one or another subset of prokaryotes. We use the term *central machinery*, a very useful and intuitively clear albeit quite loosely defined concept, to represent a set of ~4000 enzymatic and other functional roles involved in all major biochemical and informational pathways. Any particular organism contains a limited sub-set of this central machinery: from ~300 to 3000 distinct functions, depending on the genome complexity and organism life-style. Of those functions, approximately 10% remain as *globally missing genes*. Another trend revealed by comparative genome analysis is a growing number of *locally missing genes*. Indeed, as we study more and more diverse genomes, it becomes clear that there must be far more cases of non-orthologous gene displacements than most researchers would have estimated. With the rapid availability of hundreds (and soon thousands) of genomes, supplemented by functional data arriving from numerous sources, we predict that the majority of these missing genes will be characterized in the next 5–10 years, and that this monumental effort will be accomplished largely by groups of experimentalists that make effective use of the guidance provided by genome comparative analysis.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Galperin MY, Koonin EV: **Chapter 15: comparative genome analysis**. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, edn 2. Edited by Baxevanis A, Ouellette F. New York: Wiley-Liss, Inc; 2001:359-392.

2. •• Koonin EV, Galperin MY: *Sequence - Evolution - Function. Computational Approaches in Comparative Genomics*. Boston: Kluwer Academic Publishers; 2002.
We highly recommend this book. It presents a broad perspective of both fundamental and practical aspects of comparative genomics. For the scope of this review, Chapter 5 is of the most relevance, presenting techniques of genome annotation, including basics of genome context analysis and functional predictions.

3. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.

4. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature* 1999, **402**:86-90.

5. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**:751-753.

6. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.

7. Manson McGuire A, Church GM: **Predicting regulons and their cis-regulatory motifs by comparative genomics**. *Nucleic Acids Res* 2000, **28**:4523-4530.

8. Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA: **Comparative analysis of regulatory patterns in bacterial genomes**. *Brief Bioinform* 2000, **1**:357-371.

9. •• Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context**. *Genome Res* 2001, **11**:356-372.
Conserved 'gene strings' were used to infer functional coupling and to produce functional predictions for ~90 clusters of orthologous groups including a probable archaeal equivalent of the eukaryotic exosome.

10. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S: **Computational identification of operons in microbial genomes**. *Genome Res* 2002, **12**:1221-1230.

11. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis**. *Nucleic Acids Res* 2002, **30**:482-496.

12. Selkov E, Maltsev N, Olsen GJ, Overbeek R, Whitman WB: **A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data**. *Gene* 1997, **197**:GC11-GC26.

13. Bono H, Ogata H, Goto S, Kanehisa M: **Reconstruction of amino acid biosynthesis pathways from the complete genome sequence**. *Genome Res* 1998, **8**:203-210.

14. •• Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet**. *Nucleic Acids Res* 2002, **30**:42-46.
This is an introduction to one of the most valuable public web-based resources. The core of this integration is a collection of pathways connected to individual compounds, reactions, functional roles (enzymes) and genes. Recent additions include microarray expression and protein–protein interactions data.

15. Dandekar T, Schuster S, Snel B, Huynen M, Bork P: **Pathway alignment: application to the comparative analysis of glycolytic enzymes**. *Biochem J* 1999, **343**:115-124.

16. Dandekar T, Sauerborn R: **Comparative genome analysis and pathway reconstruction**. *Pharmacogenomics* 2002, **3**:245-256.

17. Kapatral V, Anderson I, Ivanova N, Reznik G, Los T, Lykidis A, Bhattacharyya A, Bartman A, Gardner W, Grechkin G *et al.*: **Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586**. *J Bacteriol* 2002, **184**:2005-2018.

18. DelVecchio VG, Kapatral V, Redkar RJ, Patra G, Mujer C, Los T, Ivanova N, Anderson I, Bhattacharyya A, Lykidis A *et al.*: **The genome sequence of the facultative intracellular pathogen *Brucella melitensis***. *Proc Natl Acad Sci USA* 2002, **99**:443-448.

19. Bhattacharyya A, Stilwagen S, Reznik G, Feil H, Feil WS, Anderson I, Bernal A, D'Souza M, Ivanova N, Kapatral V *et al.*: **Draft sequencing and comparative genomics of *Xylella fastidiosa* strains reveal novel biological insights**. *Genome Res* 2002, **12**:1556-1563.

20. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BO: **Metabolic modeling of microbial strains in silico**. *Trends Biochem Sci* 2001, **26**:179-186.

21. Cordwell SJ: **Microbial genomes and 'missing' enzymes: redefining biochemical pathways**. *Arch Microbiol* 1999, **172**:269-279.

22. Galperin MY, Koonin EV: **Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes**. *Genetica* 1999, **106**:159-170.

23. McEntyre J, Lipman D: **PubMed: bridging the information gap**. *CMAJ* 2001, **164**:1317-1319.

24. Michal G: *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. John Wiley & Sons; 1998.

25. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E Jr, Liolios K, Joukov V, Kaznadzey D, Anderson I *et al.*: **The ERGO genome analysis and discovery system**. *Nucleic Acids Res* 2003, **31**:1-8.

26. Karp PD, Paley S, Romero P: **The pathway tools software**. *Bioinformatics* 2002, **18(Suppl 1)**:S225-S232.

27. Gerlt JA, Babbitt PC: **Can sequence determine function?** *Genome Biol* 2000, **1**:reviews0005.

28. Daugherty M, Vonstein V, Overbeek R, Osterman A: **Archaeal shikimate kinase, a new member of the GHMP-kinase family**. *J Bacteriol* 2001, **183**:292-300.

29. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes**. *Nucleic Acids Res* 2001, **29**:22-28.

30. •• Kolesov G, Mewes HW, Frishman D: **SNAPping up functionally related genes based on context information: a colinearity-free approach**. *J Mol Biol* 2001, **311**:639-656.
A modification of the chromosomal clustering analysis (termed similarity-neighborhood approach) is described, validated and illustrated with examples of application in genome annotation. The major distinction of this implementation is in removing any gene colinearity constraint from the original technique.

31. • Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV: **Connected gene neighborhoods in prokaryotic genomes**. *Nucleic Acids Res* 2002, **30**:2212-2223.
Clustering of genes on the chromosome was extended to larger gene neighborhoods (superoperons), revealing new functional links between proteins, including a hypothesized 'genomic hitchhiking' phenomenon possibly related to maintaining optimal expression levels for large ensembles of genes.

32. •• Snel B, Lehmann G, Bork P, Huynen MA: **STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene**. *Nucleic Acids Res* 2000, **28**:3442-3444.
This paper describes a convenient web-based implementation of the chromosomal gene-clustering technique.

33. • Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics**. *Nat Biotechnol* 2000, **18**:609-613.
One of the earliest comprehensive reviews summarizing principles and applications of major techniques in comparative genomics going beyond sequence comparison. The functional prediction of an alternative thymidilate synthase (Thy1) was experimentally confirmed later and cited in this review.

34. Lathe WC III, Snel B, Bork P: **Gene context conservation of a higher order than operons**. *Trends Biochem Sci* 2000, **25**:474-479.

35. Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes**. *Nucleic Acids Res* 2001, **29**:1216-1221.

36. Snel B, Bork P, Huynen MA: **The identification of functional modules from the genomic association of genes**. *Proc Natl Acad Sci USA* 2002, **99**:5890-5895.

37. Yanai I, Mellor JC, DeLisi C: **Identifying functional links between genes using conserved chromosomal proximity**. *Trends Genet* 2002, **18**:176-179.

38. Kyrpides NC, Ouzounis CA, Iliopoulos I, Vonstein V, Overbeek R: **Analysis of the Thermotoga maritima genome combining a variety of sequence similarity and genome context tools**. *Nucleic Acids Res* 2000, **28**:4573-4576.

39. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context**. *Curr Opin Struct Biol* 2000, **10**:366-370.

40. • Marcotte EM: **Computational genetics: finding protein function by nonhomology methods**. *Curr Opin Struct Biol* 2000, **10**:359-365.
Genome context analysis techniques ('domain fusion, conserved gene position and gene co-inheritance and co-expression') are reviewed in the context of their application in genome-wide predictions of functions for proteins without characterized homologs.

41. Yanai I, Derti A, DeLisi C: **Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes**. *Proc Natl Acad Sci USA* 2001, **98**:7940-7945.

42. Huynen M, Snel B, Lathe W iii, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences**. *Genome Res* 2000, **10**:1204-1210.

43. •• Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C: **Predictome: a database of putative functional links between proteins**. *Nucleic Acids Res* 2002, **30**:306-309.
This paper describes one of the first web-based public resources providing a collection of predicted functional links between proteins (from 44 genomes) computed by a combination of genome context analysis techniques and protein–protein interaction data. The relative impact and potential synergy between all of these techniques is quantitatively evaluated and illustrated using the TCA cycle.

44. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes**. *Science* 2002, **298**:1616-1620.

45. McGuire AM, Hughes JD, Church GM: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes**. *Genome Res* 2000, **10**:744-757.

46. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED: **Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics**. *Proc Natl Acad Sci USA* 2002, **99**:7323-7328.

47. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12**. *Nucleic Acids Res* 2001, **29**:72-74.

48. Rodionov DA, Mironov AA, Gelfand MS: **Conservation of the biotin regulon and the BirA regulatory signal in eubacteria and archaea**. *Genome Res* 2002, **12**:1507-1516.

49. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms**. *J Biol Chem* 2002, **277**:48949-48959.

50. •• Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation**. *Nucleic Acids Res* 2002, **30**:3141-3151.
An extensive cross-genome analysis of occurrences of a conserved regulatory element (RFN) involved with riboflavin biosynthesis allowed the authors to implicate several uncharacterized genes as potential riboflavin transporters. Similar results were obtained by this group by extending the approach to biotin- and thiamin-related conserved regulatory elements (as described in their publications cited in this review).

51. Pawlowski K, Rychlewski L, Zhang B, Godzik A: **Fold predictions for bacterial genomes**. *J Struct Biol* 2001, **134**:219-231.

52. Kinch LN, Grishin NV: **Evolution of protein structures and functions**. *Curr Opin Struct Biol* 2002, **12**:400-408.

53. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2002: refinements accommodate structural genomics**. *Nucleic Acids Res* 2002, **30**:264-267.

54. Bateman A, Haft DH: **HMM-based databases in InterPro**. *Brief Bioinform* 2002, **3**:236-245.

55. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002**. *Nucleic Acids Res* 2002, **30**:235-238.

56. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA *et al.*: **The Stanford microarray database**. *Nucleic Acids Res* 2001, **29**:152-155.

57. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins: 2001 update**. *Nucleic Acids Res* 2001, **29**:239-241.

58. Badarinarayana V, Estep PW III, Shendure J, Edwards J, Tavazoie S, Lam F, Church GM: **Selection analyses of insertional mutants using subgenic-resolution arrays**. *Nat Biotechnol* 2001, **19**:1060-1065.

59. Gerdes SY, Scholle MD, D'Souza M, Bernal A, Baev MV, Farrell M, Kurnasov OV, Daugherty MD, Mseeh F, Polanuyer BM *et al.*: **From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways**. *J Bacteriol* 2002, **184**:4555-4572.

60. Heath RJ, White SW, Rock CO: **Inhibitors of fatty acid synthesis as antimicrobial chemotherapeutics**. *Appl Microbiol Biotechnol* 2002, **58**:695-703.

61. Campbell JW, Cronan JE Jr: **Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery**. *Annu Rev Microbiol* 2001, **55**:305-332.

62. Heath RJ, Rock CO: **A triclosan-resistant bacterial enzyme**. *Nature* 2000, **406**:145-146.

63. • Marrakchi H, Choi KH, Rock CO: **A new mechanism for anaerobic unsaturated fatty acid formation in *Streptococcus pneumoniae***. *J Biol Chem* 2002, **277**:44809-44816.
Further exploration of the extended cluster of fatty acid biosynthetic genes in *S. pneumoniae* enabled the discovery of a 'missing pathway', including a new enzyme and regulatory mechanism.

64. • Campbell JW, Cronan JE Jr: **The enigmatic *Escherichia coli* fadE gene is yafH**. *J Bacteriol* 2002, **184**:3759-3764.
A missing gene for acyl-CoA dehydrogenase was elucidated on the basis of evidence from microarrays and the analysis of regulatory sites, and experimental verification.

65. Sadovskaya NS, Laikova ON, Mironov AA, Gelfand MS: **Study of regulation of long-chain fatty acid metabolism using computer analysis of complete bacterial genomes**. *Mol Biol* 2001, **35**:862-866.

66. Begley TP, Kinsland C, Strauss E: **The biosynthesis of coenzyme A in bacteria**. *Vitam Horm* 2001, **61**:157-171.

67. Ni X, Ma Y, Cheng H, Jiang M, Ying K, Xie Y, Mao Y: **Cloning and characterization of a novel human pantothenate kinase gene**. *Int J Biochem Cell Biol* 2002, **34**:109-115.

68. Worrall DM, Tubbs PK: **A bifunctional enzyme complex in coenzyme A biosynthesis: purification of pantetheine phosphate adenylyltransferase and dephospho-CoA kinase**. *Biochem J* 1983, **215**:153-157.

69. • Daugherty M, Polanuyer B, Farrell M, Scholle M, Lykidis A, de Crecy-Lagard V, Osterman A: **Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics**. *J Biol Chem* 2002, **277**:21431-21439.
This paper provides an illustration of projection of function from prokaryotes to eukaryotes. The key evidence for identification of a missing human gene was provided by a protein fusion event.

70. Zhyvoloup A, Nemazanyy I, Babich A, Panasyuk G, Pobigailo N, Vudmaska M, Naidenov V, Kukharenko O, Palchevskii S, Savinska L *et al.*: **Molecular cloning of CoA synthase. The missing link in CoA biosynthesis**. *J Biol Chem* 2002, **277**:22107-22110.

71. Aghajanian S, Worrall DM: **Identification and characterization of the gene encoding the human phosphopantetheine adenylyltransferase and dephospho-CoA kinase bifunctional enzyme (CoA synthase)**. *Biochem J* 2002, **365**:13-18.

72. Wilding EI, Brown JR, Bryant AP, Chalker AF, Holmes DJ, Ingraham KA, Iordanescu S, So CY, Rosenberg M, Gwynn MN: **Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in Gram-positive cocci**. *J Bacteriol* 2000, **182**:4319-4327.

73. Humbelin M, Thomas A, Lin J, Li J, Jore J, Berry A: **Genetics of isoprenoid biosynthesis in *Paracoccus zeaxanthinifaciens***. *Gene* 2002, **297**:129-139.

74. • Smit A, Mushegian A: **Biosynthesis of isoprenoids via mevalonate in archaea: the lost pathway**. *Genome Res* 2000, **10**:1468-1484.
Traditional and new techniques of comparative genome analysis were elegantly applied for metabolic reconstruction and identification of missing gene candidates in the archaeal version of the mevalonate pathway.

75. Eisenreich W, Rohdich F, Bacher A: **Deoxyxylulose phosphate pathway to terpenoids**. *Trends Plant Sci* 2001, **6**:78-84.

76. Cunningham FX Jr, Lafond TP, Gantt E: **Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis**. *J Bacteriol* 2000, **182**:5841-5848.

77. • Seemann M, Bui BT, Wolff M, Tritsch D, Campos N, Boronat A, Marquet A, Rohmer M: **Isoprenoid biosynthesis through the methylerythritol phosphate pathway: the (*E*)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase (GcpE) is a [4Fe-4S] protein**. *Angew Chem Int Ed Engl* 2002, **41**:4337-4339.
In vitro experiments provided the ultimate confirmation of the predicted enzymatic role of the *gcpE* gene product in the nonmevalonate pathway, and implicated Fe–S clusters and an unknown regeneration system in its mode of action.

78. Hecht S, Eisenreich W, Adam P, Amslinger S, Kis K, Bacher A, Arigoni D, Rohdich F: **Studies on the nonmevalonate pathway to terpenes: the role of the GcpE (IspG) protein**. *Proc Natl Acad Sci USA* 2001, **98**:14837-14842.

79. • Adam P, Hecht S, Eisenreich W, Kaiser J, Grawert T, Arigoni D, Bacher A, Rohdich F: **Biosynthesis of terpenes: studies on 1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate reductase**. *Proc Natl Acad Sci USA* 2002, **99**:12108-12113.
This paper provides experimental verification of the last step in the nonmevalonate pathway, which produces a mixture of both critical isoprenoids (IPP and DMAPP). The step is driven by the product of *lytB* gene assisted by several cofactors, and probably one or more unidentified proteins.

80. Bacher A, Eberhardt S, Eisenreich W, Fischer M, Herz S, Illarionov B, Kis K, Richter G: **Biosynthesis of riboflavin**. *Vitam Horm* 2001, **61**:1-49.

81. • Graupner M, Xu H, White RH: **The pyrimidine nucleotide reductase step in riboflavin and F(420) biosynthesis in archaea proceeds by the eukaryotic route to riboflavin**. *J Bacteriol* 2002, **184**:1952-1957.
A combination of comparative genomics and experimental techniques is applied to characterize missing genes in archaeal riboflavin biosynthesis.

82. Fischer M, Romisch W, Schiffmann S, Kelly M, Oschkinat H, Steinbacher S, Huber R, Eisenreich W, Richter G, Bacher A: **Biosynthesis of riboflavin in archaea studies on the mechanism of 3,4-dihydroxy-2-butanone-4-phosphate synthase of *Methanococcus jannaschii***. *J Biol Chem* 2002, **277**:41410-41416.

83. Kaiser J, Schramek N, Eberhardt S, Puttmer S, Schuster M, Bacher A: **Biosynthesis of vitamin B2**. *Eur J Biochem* 2002, **269**:5264-5270.

84. Gelfand MS, Mironov AA, Jomantas J, Kozlov YI, Perumov DA: **A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes**. *Trends Genet* 1999, **15**:439-442.

85. Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E: **Sensing small molecules by nascent RNA. A mechanism to control transcription in bacteria**. *Cell* 2002, **111**:747-756.

86. Winkler WC, Cohen-Chalamish S, Breaker RR: **An mRNA structure that controls gene expression by binding FMN**. *Proc Natl Acad Sci USA* 2002, **99**:15908-15913.

87. Kreneva RA, Gel'fand MS, Mironov AA, Iomantas Iu A, Kozlov Iu I, Mironov AS, Perumov DA: **Inactivation of the ypaA gene in *Bacillus subtilis*: analysis of the resulting phenotypic expression**. *Russian J Genet* 2000, **36**:972-974.

88. Graham DE, White RH: **Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics**. *Nat Prod Rep* 2002, **19**:133-147.

89. • Graham DE, Graupner M, Xu H, White RH: **Identification of coenzyme M biosynthetic 2-phosphosulfolactate phosphatase. A member of a new class of $Mg(^{2+})$-dependent acid phosphatases**. *Eur J Biochem* 2001, **268**:5176-5188.

Clustering of genes on the chromosome provided key evidence for identification and experimental verification of the novel gene (*comB*) in the biosynthesis of coenzyme M cofactor. The broad conservation of *comB* homologs beyond methanogenic species illustrates an evolutionary recruitment of 'old genes' into 'new pathways'.

90. Mehl RA, Kinsland C, Begley TP: **Identification of the *Escherichia coli* nicotinic acid mononucleotide adenylyltransferase gene**. *J Bacteriol* 2000, **182**:4372-4374.

91. Kurnasov OV, Polanuyer BM, Ananta S, Sloutsky R, Tam A, Gerdes SY, Osterman AL: **Ribosylnicotinamide kinase domain of nadr protein: identification and implications in NAD biosynthesis**. *J Bacteriol* 2002, **184**:6906-6917.

92. Bishop AC, Xu J, Johnson RC, Schimmel P, de Crecy-Lagard V:
• **Identification of the tRNA-dihydrouridine synthase family**. *J Biol Chem* 2002, **277**:25090-25095.
A synergistic application of several genome context analysis techniques allowed the authors to implicate, experimentally verify and characterize members of a new family of enzymes involved in critical tRNA modification pathways.

93. Kuhn P, Lesley SA, Mathews II, Canaves JM, Brinen LS, Dai X, Deacon AM, Elsliger MA, Eshaghi S, Floyd R *et al.*: **Crystal structure of thy1, a thymidylate synthase complementing protein from *Thermotoga maritima* at 2.25 Å resolution**. *Proteins* 2002, **49**:142-145.

94. Heath RJ, Su N, Murphy CK, Rock CO: **The enoyl-[acyl-carrier-protein] reductases FabI and FabL from *Bacillus subtilis***. *J Biol Chem* 2000, **275**:40128-40133.

95. Bobik TA, Rasche ME: **Identification of the human**
•• **methylmalonyl-CoA racemase gene based on the analysis of prokaryotic gene arrangements. Implications for decoding the human genome**. *J Biol Chem* 2001, **276**:37194-37198.
One of the first reports in which a eukaryotic missing gene (involved in propionyl-CoA metabolism) was characterized by projection from the clusters on prokaryotic chromosomes. The authors express a well-justified enthusiasm about the potential of such an approach in the functional analysis of human genome.

96. Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U: **An alternative flavin-dependent mechanism for thymidylate synthesis**. *Science* 2002, **297**:105-107.