Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence

Joakim Näsvall *et al.*

Science **338**, 384 (2012);
DOI: 10.1126/science.1226521

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence

Joakim Näsvall,[1] Lei Sun,[1] John R. Roth,[2] Dan I. Andersson[1]*

Gene duplications allow evolution of genes with new functions. Here, we describe the innovation-amplification-divergence (IAD) model in which the new function appears before duplication and functionally distinct new genes evolve under continuous selection. One example fitting this model is a preexisting parental gene in *Salmonella enterica* that has low levels of two distinct activities. This gene is amplified to a high copy number, and the amplified gene copies accumulate mutations that provide enzymatic specialization of different copies and faster growth. Selection maintains the initial amplification and beneficial mutant alleles but is relaxed for other less improved gene copies, allowing their loss. This rapid process, completed in fewer than 3000 generations, shows the efficacy of the IAD model and allows the study of gene evolution in real time.

The origin of new genetic functions poses a fundamental biological question. In many bacteria, most new genes are acquired by horizontal gene transfer (HGT) from related organisms (*1*), but new genes with novel functions can also evolve from extra copies of duplicated genes in both bacteria and eukaryotes (*2*). New genes can evolve from a redundant copy of a duplicated parental gene, by removing one copy from selection. This extra copy is then able to acquire one or more mutations providing a new beneficial function. At this point, natural selection would maintain the duplication and drive further evolution. However, this model requires that the extra gene be maintained without selection long enough for rare improvements to occur; because tandem duplications are generally very unstable and short-lived, they are unlikely to remain long enough to acquire mutations (*3*, *4*). Furthermore, duplications typically have fitness costs (*3*, *4*), and deleterious mutations outnumber beneficial mutations, making inactivation of the gene (pseudo-gene formation) the most likely fate for any redundant gene copies.

We propose the innovation-amplification-divergence (IAD) model (Fig. 1A), which allows the evolution of new genes to be completed under continuous selection that favors maintenance of the functional duplicate copies and divergence of the extra copy from the parental allele (*5*). The IAD model proposes that the ancestral gene has a weak secondary activity (innovation) (*6*, *7*), and when a change in conditions makes this activity useful, selection favors increased gene dosage (amplification), resulting in two or more copies of the parent allele. The increased copy number provides multiple targets for beneficial mutations and buffers any negative effects a new mutation may have on the original activity. During continuous growth under conditions that select for both the original activity and the new activity, beneficial mutations will accumulate (divergence) in the copies. Any improved copy can be further amplified, whereas less functional copies, including the parental gene, can be lost. Ultimately, this results in a gene duplication in which one gene copy encodes the parent activity and another copy provides an improved, new activity.

To experimentally test the IAD model, we examined a histidine biosynthetic enzyme (HisA), and through continuous selection we created, by duplication and divergence, a new gene that catalyzes a step in tryptophan synthesis. The original HisA and TrpF enzymes both catalyze isomerization of a phosphoribosyl compound, but each acts on different substrates in the biosynthesis of the amino acids histidine and tryptophan (Fig. 1B). HisA and TrpF enzyme activities are selectable by growth in minimal media lacking histidine and tryptophan. In addition, the enzymes are structurally related and evolved from a common ancestor (*8*). Furthermore, *Streptomyces* and *Mycobacteria* lack TrpF but instead have one enzyme, PriA, that is a HisA ortholog and catalyzes both reactions (*9*).

In a strain lacking *trpF*, we selected a spontaneous *hisA* mutant of *Salmonella enterica* that maintained its original function (HisA) but acquired a low level of TrpF activity sufficient to support slow growth on a medium lacking histidine and tryptophan, representing the innovation of the IAD model (see table S1 for

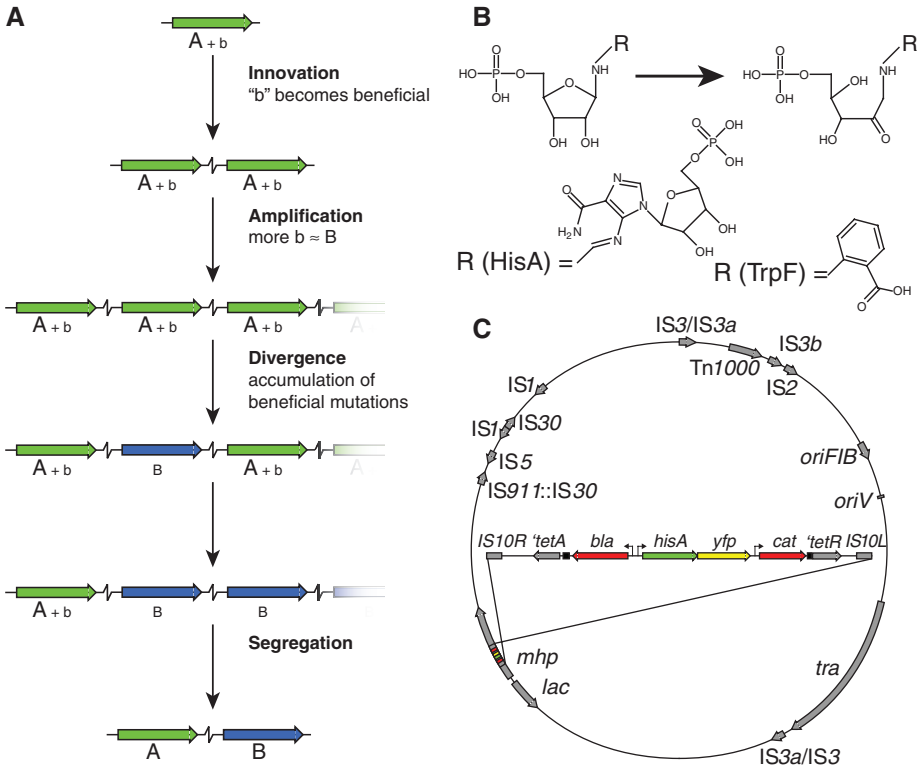

Fig. 1. (**A**) The IAD model. Innovation occurs when the ancestral gene (green) encodes a protein with the main function "A" and a minor activity "b." Amplification occurs when an environmental change makes the b activity beneficial and selection favors variants with increased b activity. Divergence may occur in any one of the amplified gene copies that acquires a beneficial mutation that increases "B" activity (blue gene copy). After a B mutation, selection for the amplified array is relaxed, and segregation occurs to leave alleles with original A activity and the evolved B activity. (**B**) The isomerization reaction catalyzed by HisA and TrpF. The respective substrates and products differ in which chemical group (R) is attached to the 1′-amino group of the phosphoribosylamine. (**C**) Structure of the T-*his* element (linear insert) and its location on F′128 (circle) with the relative genetic elements on the F′ as shown (transposons; IS elements; replication origins; and the *tra*, *lac*, and *mhp* operons).

[1]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. [2]Department of Biology, University of California, Davis, CA, USA.

*To whom correspondence should be addressed. E-mail: dan.andersson@imbim.uu.se

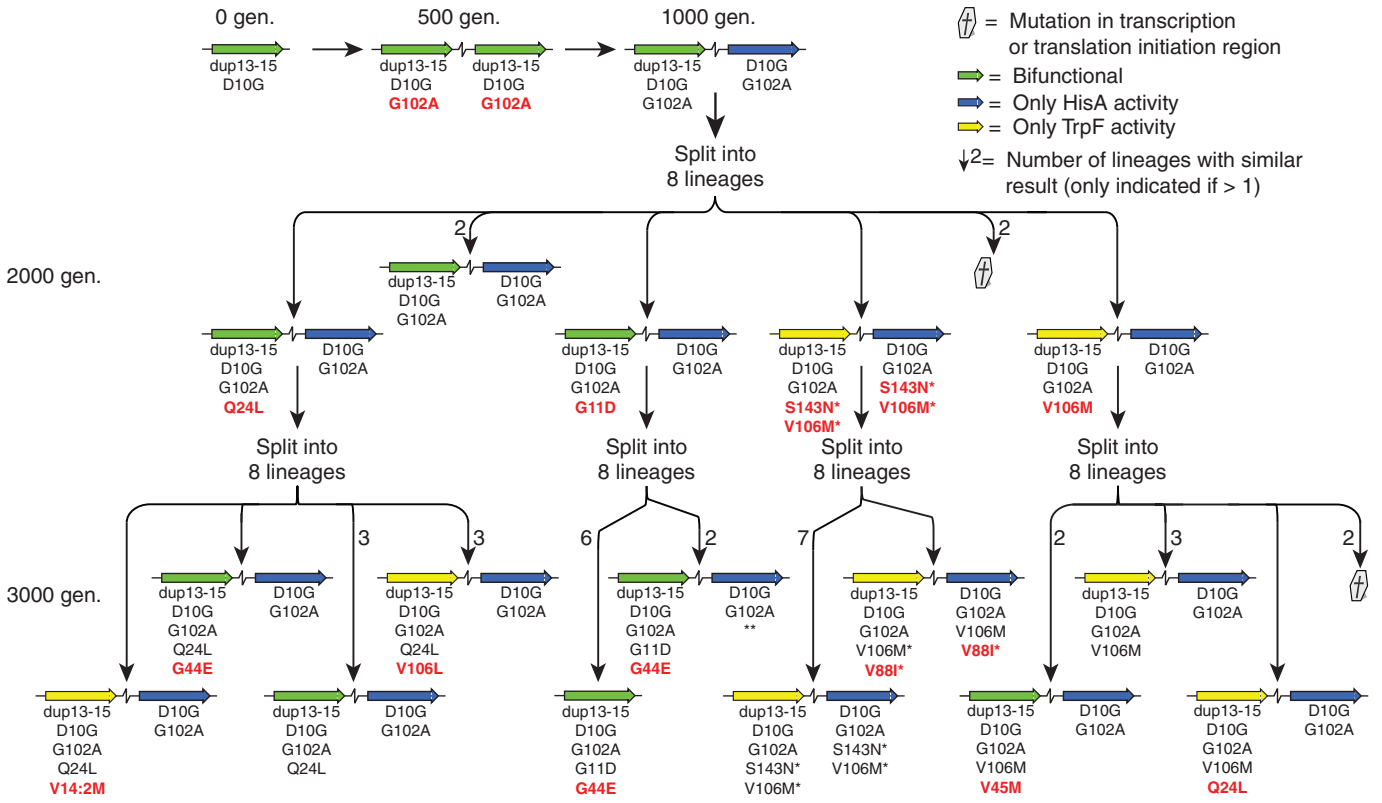

Fig. 2. Trajectory for 3000 generations of evolution of the bifunctional parental gene (dup13-15, D10G) during selection for improved TrpF and HisA activities from one main parental lineage to the numerous variants found in daughter lineages. Mutations verified by sequencing are shown below the gene symbols. Red text indicates the identification of a new mutation for that lineage after the indicated number of generations. Additional lineages are shown in fig. S1, A to C. Asterisks next to a mutation indicate the presence of more than one subpopulation, differing in which of the indicated mutations they contain. Two asterisks indicate that only a subpopulation of the cells in the culture contained the indicated gene copy. A, Ala; Q, Gln; L, Leu; S, Ser; N, Asn; V, Val; M, Met; E, Glu; I, Ile.

strains). Two mutations were required for this innovation: First, an internal duplication of codons 13 to 15 (dup13-15) gave a weak TrpF activity but led to a complete loss of HisA activity. A subsequent amino acid substitution [Asp^{10}→Gly^{10} (D10G)] restored some of the original HisA activity (*10*). We also isolated two other bifunctional derivatives of *hisA* that had acquired TrpF activity, but we will not discuss these mutants in this paper (fig. S1, A to C) (*10*).

We placed this bifunctional parental gene (dup13-15, D10G) under the control of a constitutive promoter that cotranscribed a yellow fluorescent protein (*yfp*) gene. We also placed the T-*his* operon in a transposition-inactive transposable element Tn*10d* Tet close to the *lac* operon on the low–copy number (about two copies per chromosome) (*11*) F'_{128} plasmid (Fig. 1C). Duplications and amplifications of this region are frequent and have low fitness cost (*3*), allowing experimental study of the process within a reasonable time frame. An F′ plasmid with the bifunctional gene inside T-*his* was introduced into a *S. enterica* strain with deleted *hisA* and *trpF* genes, dependent on the bifunctional gene for synthesis of both histidine and tryptophan. In the absence of both amino acids, the bifunctional gene supported a generation time

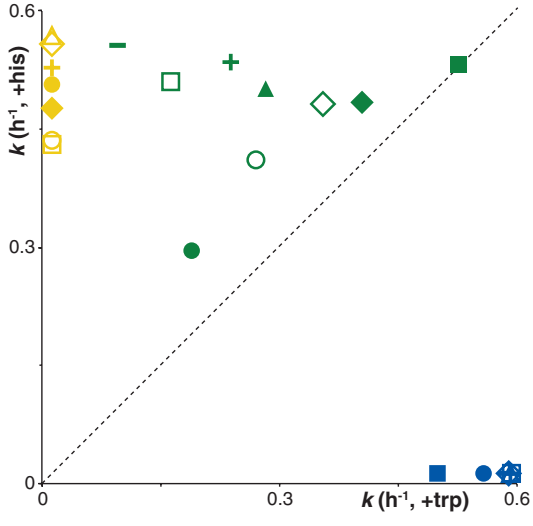

Fig. 3. Characteristics of 22 different evolved mutant gene variants. Each point represents the fitness of one specific mutant gene for its HisA activity on the *x* axis [assayed as growth rate (*k*) in minimal glycerol medium with added tryptophan] and TrpF activity on the *y* axis (assayed as growth rate in minimal glycerol medium with added histidine). Mutant genes fall into three main classes as indicated by the colors: Blue, HisA specialists [open diamond, *hisA*(wt); open triangle, D10G G102A; dash, D10G G102A V106M; cross, D10G; open square, D10G G102A S143N; solid circle, D10G R83C; solid square, D10G G102A V106M V88I]. Yellow, TrpF specialists (open triangle, dup13-15 D10G G102A Q24L V106L; open diamond, dup13-15 D10G G102A V106M V88I; cross, dup13-15 D10G G102A V106M Q24L; solid circle, dup13-15 D10G G102A Q24L V14:2M; solid diamond, dup13-15 D10G G102A V106M; open circle, dup13-15 D10G R83C; open square, dup13-15 D10G G81D). Green, generalist enzymes [solid circle, dup13-15 D10G (ancestral bifunctional gene); dash, dup13-15 D10G G102A V106M V45M; cross, dup13-15 D10G G102A Q24L G44E; solid square, dup13-15 D10G G102A G11D G44E; open square, dup13-15 D10G G102A Q24L; solid triangle, dup13-15 D10G G102A V88I; open diamond, dup13-15 D10G G102A S143N; solid diamond, dup13-15 D10G G102A G11D; open circle, dup13-15 D10G G102A].

of ~5.1 hours in minimal medium with doubling times of ~2.8 hours in the presence of tryptophan alone, ~2.6 hours in the presence of histidine alone, and ~1.5 hours in presence of both amino acids.

Several independent lineages of this strain evolved under continuous selection for improved growth and increased HisA and TrpF activities by serial passages in minimal glycerol medium lacking both amino acids (*10*). Within a few hundred generations, the growth rate increased from 5 hours per division to 1.9 to 2.5 hours, depending on the lineage. Associated with the increased growth rate, expression of the parental bifunctional gene (fig. S2 and table S2) increased stepwise (up to 20-fold) in most cultures due to amplifications of a region of the plasmid that includes the bifunctional parental gene and *yfp* (see fig. S3 and table S3 for structures of amplified units).

After evolution for up to 3000 generations, all lineages acquired mutations resulting in faster growth. In many of the lineages, different gene copies in the amplified array diverged by mutations that allowed enzymatic specialization (Fig. 2 and fig. S1, A to C). As predicted from the IAD model, we observed the appearance of a diverged gene copy with improved activity, relaxed selection for maintenance of the unimproved copies in the amplified array, loss of the unimproved copies, and, in some cases, reduction in the total gene copy number. (fig. S2 and table S2).

To test the HisA and TrpF activities of the evolved enzymes, 22 different genes from the evolved strain were individually cloned into the chromosomal *cobA* gene of a strain (lacking both the *hisA* and *trpF* genes) that had never been subjected to a histidine-tryptophan selection (*10*). This assured that the strain had only one copy of the gene to be tested, and no outside mutations contributed to activity. Strains with these single-copy evolved genes were tested for their ability to grow on a minimal medium with single amino acids (Trp or His), both, or neither. In every case, the evolved mutated gene increased the growth rate in the absence of either histidine or tryptophan or when both amino acids were absent. The evolved genes fell into three classes: (i) specialized genes with strongly improved HisA activity and loss of TrpF activity, (ii) specialized genes with strongly improved TrpF activity and loss of HisA activity, and (iii) generalist genes whose encoded enzyme showed a moderate increase in both activities (Fig. 3; Fig. 4, A to D; and table S4). In several cases, specialized mutant genes of both types (i) and (ii) were found in single bacterial clones, demonstrating that gene copies within a single amplified array had diverged to become specialized to perform either the HisA- or TrpF-specific reactions (Fig. 4, A to C). In other cases, the ancestral gene evolved into an individual gene with an improved level of both HisA and TrpF activities (Fig. 4D). In some strains an improved generalist enzyme evolved first and then duplicated with copies, subsequently diverging and becoming specialized (Fig. 4, B and C). Figure S4 shows the locations of the identified mutations on the HisA structure from *Thermotoga maritima*.

Thus, under suitable selective conditions, the IAD process rapidly generates genes with distinct enzymatic activities. In *Salmonella*, duplications of any particular gene form at a rate of roughly 10^{-5} per cell per division and reach a high steady-state frequency in the population, providing a reservoir of standing copy number variation upon which selection can act (*12*). Amplification to higher copy numbers occurs at 10^{-2} per cell per division (*3*), several orders of magnitude more frequent than point mutations. Thus, whenever a limiting gene product restricts cell growth, initial escape from

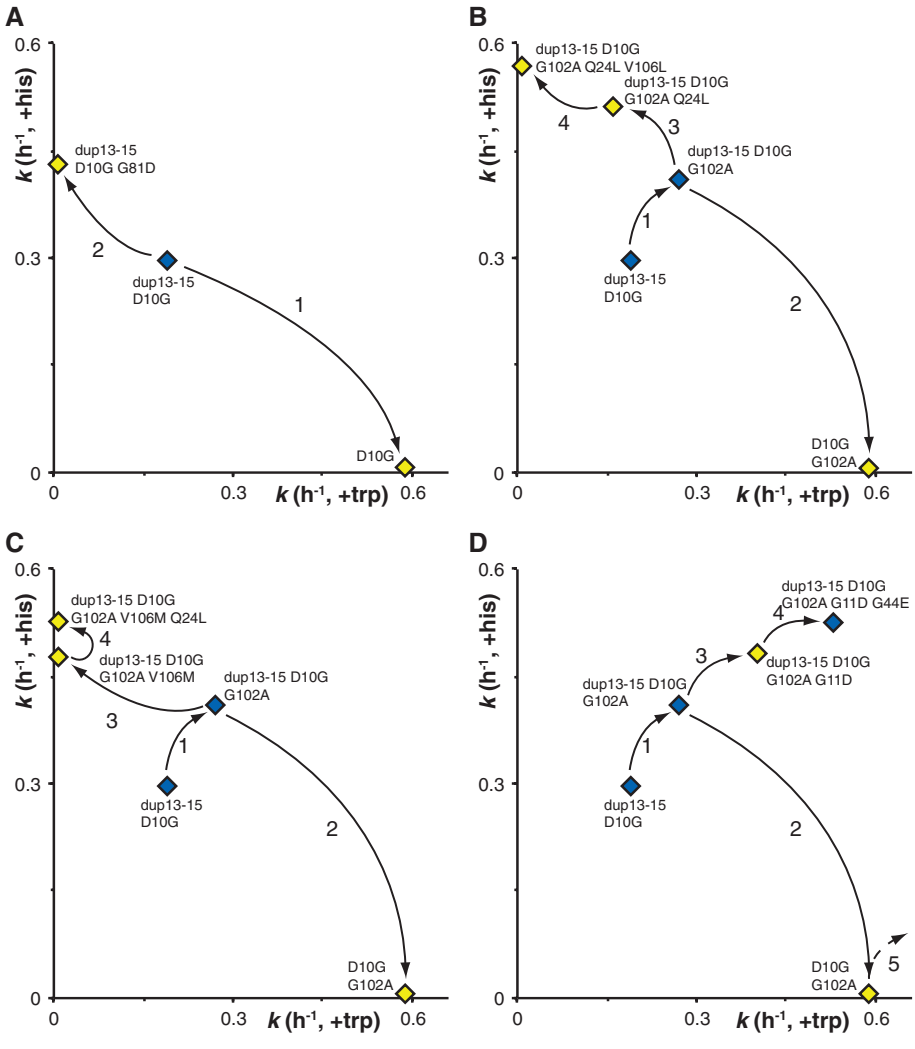

Fig. 4. Multiple evolutionary trajectories recovered through IAD. The *x* axis indicates the HisA activity (assayed as growth rate in minimal glycerol medium with added tryptophan); the *y* axis indicates the TrpF activity (assayed as growth rate in minimal glycerol medium with added histidine). (**A**) Evolution of specialist enzymes in which one activity is improved at the expense of the other. (**B** and **C**) Evolution of specialist enzymes after initial evolution of a generalist enzyme. (**D**) Evolution of a generalist enzyme with improvement of both activities. Arrows and numbers indicate the sequential order of appearance of the various mutations in the population. Yellow symbols denote gene variants that were always accompanied by another gene variant (generalist or with the complementary activity) in the same amplified array. Blue symbols denote gene variants that, at some point during the evolution, were the only variants found in the population.

this restriction may initially occur by duplication events and higher amplification, rather than rare point mutations that alter catalytic activity (*13*). The delayed appearance of point mutations suggests that the accumulation of a point mutation is the rate-limiting step in the IAD process.

Other sequence-based evidence supports the predictions of the IAD model, particularly in eukaryotes where new genes often evolve by amplification-divergence processes. For example, the evolution of a new gene may be accompanied by the appearance of paralogs and pseudogenes in the genome (*14*, *15*), new genes may show evidence of continuous selection (*16*, *17*), and new genes and pseudogenes may be tandemly clustered with the parent gene (*18*, *19*). On the contrary, in bacteria duplicate genes most commonly arise via HGT (*20*), but the IAD process could still generate new genes that can be distributed to other organisms by HGT. Conversely, horizontally acquired genes have a higher likelihood of possessing a new side-activity upon which selection and the IAD process could act, suggesting a potential coupling between the IAD process and HGT (*21*).

References and Notes

1. H. Ochman, J. G. Lawrence, E. A. Groisman, *Nature* **405**, 299 (2000).
2. S. Ohno, *Evolution by Gene Duplication* (Springer, New York, 1970).
3. A. B. Reams, E. Kofoid, M. Savageau, J. R. Roth, *Genetics* **184**, 1077 (2010).
4. M. E. Pettersson, S. Sun, D. I. Andersson, O. G. Berg, *Genetica* **135**, 309 (2009).
5. U. Bergthorsson, D. I. Andersson, J. R. Roth, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17004 (2007).
6. S. D. Copley, *Curr. Opin. Chem. Biol.* **7**, 265 (2003).
7. O. Khersonsky, D. S. Tawfik, *Annu. Rev. Biochem.* **79**, 471 (2010).
8. M. Henn-Sax, B. Höcker, M. Wilmanns, R. Sterner, *Biol. Chem.* **382**, 1315 (2001).
9. F. Barona-Gómez, D. A. Hodgson, *EMBO Rep.* **4**, 296 (2003).
10. Materials and methods are available as supplementary materials on *Science* Online.
11. R. Frame, J. O. Bishop, *Biochem. J.* **121**, 93 (1971).
12. P. Anderson, J. Roth, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3113 (1981).
13. D. I. Andersson, D. Hughes, *Annu. Rev. Genet.* **43**, 167 (2009).
14. J. Bergelson, M. Kreitman, E. A. Stahl, D. Tian, *Science* **292**, 2281 (2001).
15. R. W. Michelmore, B. C. Meyers, *Genome Res.* **8**, 1113 (1998).
16. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
17. F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, *Genome Biol.* **3**, research0008 (2002).
18. G. P. Wagner, C. Amemiya, F. Ruddle, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14603 (2003).
19. M. Hoffmann *et al.*, *Proc. Biol. Sci.* **274**, 33 (2007).
20. T. J. Treangen, E. P. Rocha, *PLoS Genet.* **7**, e1001284 (2011).
21. S. D. Hooper, O. G. Berg, *Genome Biol.* **4**, R48 (2003).

Acknowledgments: This work was supported by a grant from the Swedish Research Council to D.I.A. and from the NIH (grant GM27068) to J.R.R. The raw Illumina sequencing data sets have been deposited in the National Center for Biotechnology Information Sequence Read Archive (www.ncbi.nlm.nih.gov/Traces/sra/) with accession numbers SRX180378, SRX180382, SRX180383, SRX180384, SRX180385, SRX180387, SRX180388, SRX180390, SRX180391, and SRX180392.

Supplementary Materials

www.sciencemag.org/cgi/content/full/338/6105/384/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S4
Tables S1 to S4
References (*22*–*34*)

25 June 2012; accepted 4 September 2012
10.1126/science.1226521

Metagenome Mining Reveals Polytheonamides as Posttranslationally Modified Ribosomal Peptides

Michael F. Freeman,[1]* Cristian Gurgui,[1]* Maximilian J. Helf,[1] Brandon I. Morinaka,[1] Agustinus R. Uria,[1] Neil J. Oldham,[2] Hans-Georg Sahl,[3] Shigeki Matsunaga,[4] Jörn Piel[1,5]†

It is held as a paradigm that ribosomally synthesized peptides and proteins contain only L-amino acids. We demonstrate a ribosomal origin of the marine sponge–derived polytheonamides, exceptionally potent, giant natural-product toxins. Isolation of the biosynthetic genes from the sponge metagenome revealed a bacterial gene architecture. Only six candidate enzymes were identified for 48 posttranslational modifications, including 18 epimerizations and 17 methylations of nonactivated carbon centers. Three enzymes were functionally validated, which showed that a radical *S*-adenosylmethionine enzyme is responsible for the unidirectional epimerization of multiple and different amino acids. Collectively, these complex alterations create toxins that function as unimolecular minimalistic ion channels with near-femtomolar activity. This study broadens the biosynthetic scope of ribosomal systems and creates new opportunities for peptide and protein bioengineering.

The marine sponge *Theonella swinhoei*, a composite organism containing numerous uncultivated bacterial symbionts, is a rich source of bioactive metabolites (*1*). Among these, polytheonamides A and B (Fig. 1A) are particularly noteworthy for their structural complexity (*2*). Of the 19 different amino acids that constitute these unusual 48-residue peptides, 13 are nonproteinogenic. The compounds were therefore assumed to be products of a nonribosomal peptide synthetase (NRPS)—a large multifunctional protein complex that can generate peptides with unusual residues (*3*, *4*). However, polytheonamides are larger than other known NRPS-synthesized secondary metabolites, and the size of an NRPS biosynthetic machinery required to assemble 48 residues prompted us to speculate whether, alternatively, a ribosomal pathway (*5*, *6*) could be involved. This would require a ribosomal pathway that could introduce multiple D-configured and C-methylated residues (*6*–*8*). To test the ribosomal hypothesis, a seminested polymerase chain reaction (PCR) protocol (fig. S1) was used with primers designed on the basis of a hypothetical precursor peptide consisting of proteinogenic L-configured amino acids (Fig. 1A). Sequencing revealed a succession of codons that precisely corresponded to an unprocessed polytheonamide precursor; this supports a ribosomal origin. To identify the surrounding DNA region, 920,000 clones of a library of *T. swinhoei* total DNA (*9*) were screened in a pool-dilution strategy (*10*), yielding a single cosmid pTSMAC1. The few other clones detected were repeatedly lost during isolation. To expand the upstream sequence, we amplified a 7-kilobase portion directly from the partially enriched pool by long-range PCR, using primers based on sequences of the cosmid vector and the pTSMAC1 insert. The authenticity of the amplified region was subsequently confirmed by repeated PCR and sequencing with metagenomic DNA.

The assembled DNA region contained 11 additional genes, clustered around the initially identified open reading frame (ORF) (Fig. 1B). Nine ORFs, which we termed *poy* genes (*poyA-I*), form an operon, as apparent from the short or often absent intergenic regions. This polycistronic architecture, as well as the presence of Shine-Dalgarno motifs and lack of detectable introns, suggests a bacterial endosymbiont as the origin of the cloned region. Beyond the gene cluster, the presence of an upstream prokaryotic *hicAB*-type toxin-antitoxin system, numerous genes and gene fragments resembling bacterial transposition elements, and two downstream genes encoding a polyketide synthase of as-yet-unknown function further support this hypothesis (table S1). The 3′ terminus of *poyA* consists of 48 codons that match a

[1]Kekulé Institute of Organic Chemistry and Biochemistry, University of Bonn, Gerhard-Domagk-Strasse 1, 53121 Bonn, Germany. [2]School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK. [3]Institute of Medical Microbiology, Immunology and Parasitology, University of Bonn, Meckenheimer Allee 168, 53115 Bonn, Germany. [4]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan. [5]Institute of Microbiology, Eidgenössische Technische Hochschule (ETH) Zurich, Wolfgang-Pauli-Strasse 10, 8093 Zurich, Switzerland.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: joern.piel@uni-bonn.de, jpiel@ethz.ch