Evolutionary Anthropology 15:207–211 (2006)

Reading the Palimpsests of Life

Some relatives bear only the faintest trace of their ancestor.

KENNETH M. WEISS AND KAZUHIKO KAWASAKI

Parchment and vellum provided durable writing surfaces for scholars in the early Middle Ages in Europe. In the years before suitable paper and the technical means to print multiple copies of a work were available, it was important that a scribe leave a lasting impression. Parchment and vellum were convenient, easier to work with than stone tablets, and yet still durable. But they were costly to make, so when it came time to write a new book, often the writing on an old work was removed and the new text was lettered in a neat monastical hand on the same surface.

In some instances the original writing was washed off, while in others it was physically scraped away. In the former case, faint traces of the original sometimes remain and can be restored with laborious and patient effort so that, centuries later, scholars can read through the superimposed writing to discern what had been written before. These rewritten pages, called palimpsests, are a source of intellectual, religious, and science history, providing important ancient texts that would otherwise not be available.

The most famous example is a manuscript copy of a paper by the third-century BC mathematician, Archimedes of Syracuse, one of history's most brilliant scientists. In his *Method of Mechanical Theorems*, Archimedes devised a method of infinitesimals that anticipated Isaac Newton's invention of integral calculus by nearly 2,000 years. He conceptualized geometrical problems in mechanical terms like levers and centers of gravity. For example, as in the faint figures restored in Figure 1B, he divided a geometric structure such as a parabola into essentially an infinite number of small subparts, the areas of which could be added together to approximate the total to answer some question about the structure.

No copy of Archimedes' original presentation of his method survives, so his ideas had to be inferred from their use in his known works. But, as it happens, a tenth-century monk made a parchment copy of *Method* along with other of Archimedes' works. In the twelfth-century, however, a monk named Joannes Myronas washed away the original writing and recycled the parchment as a prayer book. Fortunately, the erasure was not complete. Modern methods have restored much of the earlier manuscript. The most complete restoration so far has recently been achieved by a method called X-ray fluorescence, which detects iron molecules in the original scribe's ink. So we now know what Archimedes said.

A phenomenon in evolution somewhat resembles a palimpsest. What we see today evolved from ancestral states, often "writing over" the original state as organisms adopt new functions so that the original state no longer exists. An example has to do with what has made vertebrates what we are today. There is a good vertebrate fossil record largely because what makes fossils hard is also what made vertebrates hard. Bones and other mineralized tissues, initially forming an external skeleton and subsequently an internal skeleton, as well as dentition, are perhaps the most important distinguishing characteristic of our group of animals. In a research project to understand the genetic basis of these evolving traits, we had to confront a problem akin to revealing and reading a palimpsest.

HOW THE NEW POINTS TO THE OLD: CHASING GENE GENEALOGIES

Reconstructing genetic evolution must be done by inferring the past from comparative studies of present-day DNA sequences. To find evolutionarily related genes, we use the fact that genes arise as accidental duplicate copies of existing genes forming a "gene family" that diversifies in function as mutations accumulate in each of the genes over time. The new functions are typically related in some way, which constrains the mutational variation they can tolerate. As a result, we can use DNA sequence similarity to identify gene family members within and between species. A test sequence is compared to the entire global database of DNA sequences (Genbank, at www.ncbi.nlm.nih.gov/), using a program called BLAST, which searches the database for sequences that have high degrees of sequence similarity (sometimes casually referred to as homology) to the test sequence. The program allows for some misalignment, substitutions, deletions, or insertions in the sequences being compared. That tolerance is necessary because even

Ken Weiss is Evan Pugh Professor and Kazuhiko Kawasaki is Senior Research Scientist, both in the Department of Anthropology at Penn State University.

DOI 10.1002/evan.20120
Published online in Wiley InterScience (www.interscience.wiley.com).

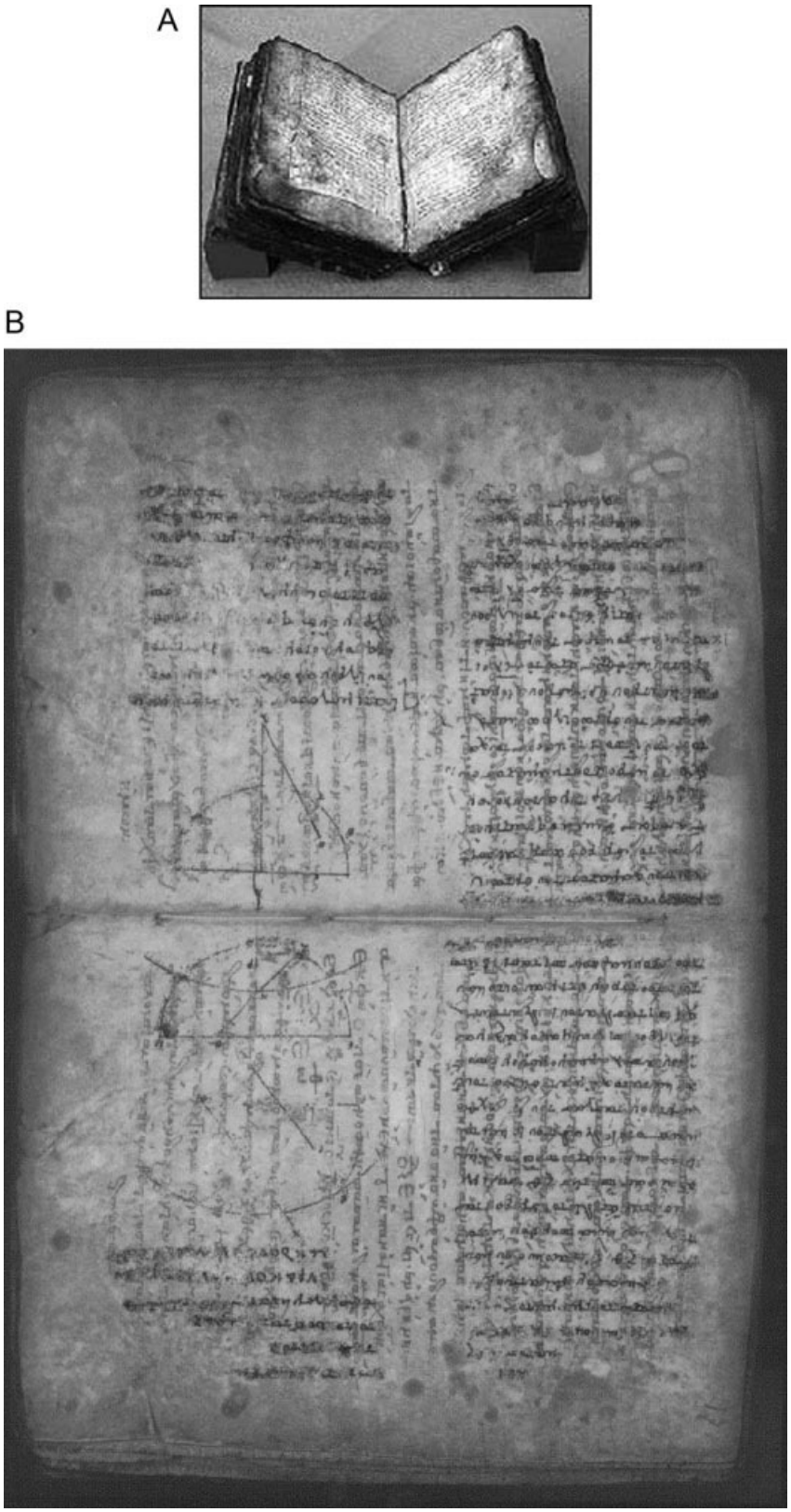

Figure 1. The Archimedes palimpsest, probably the most historically important and most widely accessible example. A. The book. B. A page of Archimedes Method. Source: The figures and text material on the Archimedes palimpsest are from many different public domain web sites, such as Wikipedia (http://en.wikipedia.org/wiki/Archimedes_palimpsest).

closely related genes are continually under threat of mutations, some of which persist over evolutionary time. Figure 2 shows an alignment of several homologous modern sequences in different species derived from the same ancestral gene, as well as a paralogous sequence (that is, a different gene family member) from the same species, for a given family of genes involved in vertebrate mineralization.

The figure shows the imperfect nucleotide correspondence that gives a sense of the kind of sequence relationships that can be found by alignments among related genes. To reinforce that point, a totally unrelated sequence is also shown. As expected, the unrelated sequence has little similarity to the other sequences.

A similar approach is then used to reconstruct the genealogy of the genes. Since mutations only accumulate slowly over time, we can compare the relative degrees of similarity among related genes to guess at their ancestral tree of origin. The most similar genes, and usually those most closely clustered together on the chromosome, are the most recently derived by duplication. Together, these approaches enable us to suggest which gene is the precursor of the entire family. The degree of difference among the genes indicates roughly how long ago their proliferation started.

VERTEBRATE ORIGINS, THE HARD WAY

We have been studying a family of genes, some of which are shown in Figure 2, that are primarily responsible for the mineralized tissues that have characterized vertebrate evolution. These genes, which we named the secretory calcium-binding phosphoprotein (SCPP) gene family, code for proteins that capture calcium, initially used in the mineralization of the hard tissues in early vertebrates: enamel, dentine, and bone. Subsequently, newer members of the SCPP family have taken on additional mineralization functions derived over evolutionary time from the original functions. Mammalian milk casein

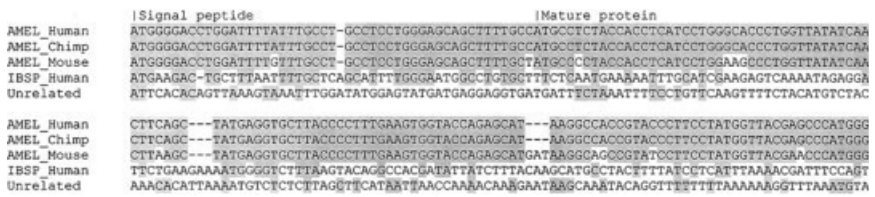

Figure 2. Homology the usual way. An alignment of corresponding sections of closely related amelogenin (AMEL, an enamel gene) sequences from human, chimpanzee, and mouse, plus a human dentine/bone matrix gene (IBSP), and an unrelated noncoding bit of human DNA sequence. In the region shown, human and chimpanzee AMEL sequences are identical, while mouse AMEL is slightly different. AMEL and IBSP show a weak sequence homology in the signal peptide region. Shading identifies shared nucleotides. Sequences from Genbank.

and salivary protein SCPP genes arose from an ancient enamel SCPP gene. Caseins enrich milk in calcium that helps infants grow teeth and bone; salivary proteins help maintain enamel integrity. In a similar way, bird eggshell is calcified with the help of an SCPP gene that originated from a dentine/bone mineralization gene.

These are interesting genes for a lot of reasons. With a single exception, they are located together on a single chromosome (in humans, chromosome 4) in two nearby gene clusters. The exception is *AMEL*, the gene for the enamel protein amelogenin, which is on the sex chromosomes in mammals (both X and Y in humans and some other mammals, or just X in others). The SCPP genes have stayed clustered together throughout their existence in vertebrate evolution, even as new individual genes arose by duplication or were destroyed by mutation. For reasons we don't yet know, that conserved clustering likely relates to the expression control of these genes.

The SCPP genes were very difficult to find because the nature of their function gives them great freedom to evolve. Although the amino acid sequences of the SCPP proteins have virtually no homology, they share common biochemical characteristics. They all have a signal peptide, a bit of the end of the proteins that is used to secrete them outside the cell, where they do their work. The SCPP genes also share a common overall organizational structure (Fig. 3). The most upstream region (the whole exon 1 and the first part of exon 2) is not translated into protein at all, while the second region (the middle part of exon 2) codes for the amino acids that make up the signal peptide (exons are protein-coding regions of genes). Following these regions and, depending on the gene and species, SCPP genes vary in their number of exons. However, the noncoding sequences (introns) that separate the exons splits the exons exactly between adjacent amino-acid specifying codons. The introns never disrupt a nucleotide-triplet codon, as shown in Figure 3A; these are called phase 0 exons. In all vertebrate genomes, including our own, only the SCPP genes possess all of these characteristics.

Unlike most proteins, the SCPPs are not neatly folded into sheets or helices, but are instead called disordered proteins. Their functional feature is not the shape of the protein, but, for example, the availability of the many negatively charged amino acids to attract and bind positively charged calcium ions in the extracellular matrix or fluids. This property does not require a *specific* arrangement of amino acids, which has left much of the sequence of the SCPP genes free to change so long as it still codes for enough suitably charged amino acids. This allows *phenogenetic drift* to occur, the evolutionary process by which a trait such as calcium binding is conserved by selection over long evolutionary times, while its underlying genetic basis diverges widely.

These SCPP features pose a problem for reconstructing the genetic basis of vertebrate mineralization. Because of low sequence similarity, it is difficult, if not impossible, to identify SCPP genes by traditional sequence comparison methods alone. The details making the features we see today have obscured the way that same feature was originally written into the genome. Until we realized this, genes like amelogenin seemed to have no relatives in the genome. That striking, if not incredible fact was clarified when we began searches based on gene structure rather than DNA sequence characteristics. Fortunately, within mammals the signal peptide part of SCPP genes is sufficiently conserved that it *can* be aligned by sequence searches. By itself, however, the signal peptide parts are too short, and their imperfect alignments might just have been the kind of coincidental similarity that is easy to find when searching the three billion nucleotides of vertebrate genomes. The additional information provided by the shared gene structure, including the chromosomal clustering, eventually allowed us to identify the genes and show that they were a single family of related genes (Fig. 3B).[1–4]

PEELING AWAY THE WRITING

Once the SCPP gene family was identified in mammals, the obvious thing was to go back phylogenetically, at least insofar as we can do that with DNA sequences from contemporary species as representatives of ancestral lineages, to see where these genes came from and how they evolved. But the deeper we went back in time in this comparative way, the harder the genes were to find. Because of phenogenetic drift and their particular family characteristics, in some deep lineages like fish only the faintest images of common origin remain. These have to be read by looking through the current sequences somewhat as scholars wash away layers of writing on palimpsests. Once we knew about the SCPP gene family and the nature of the proteins it codes for, this could be done.

A new gene that arises by gene duplication often stays on the same

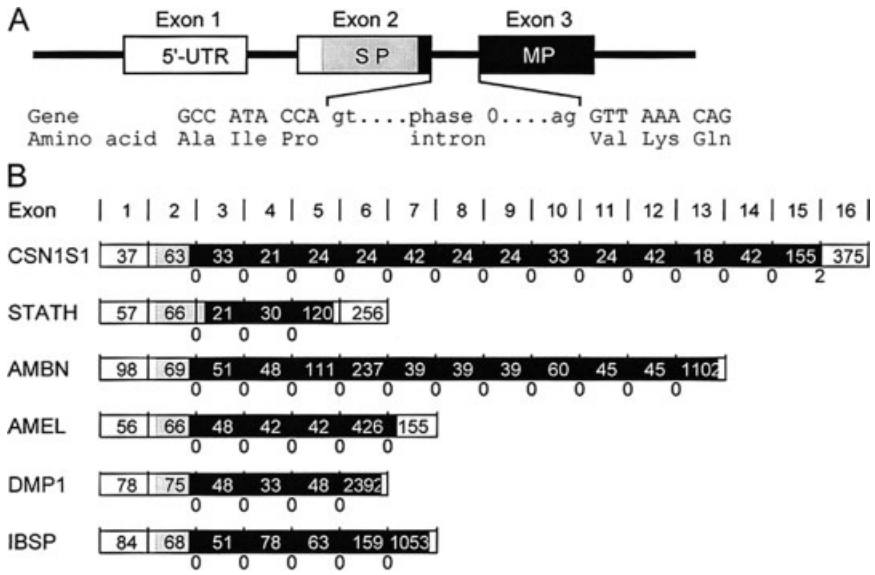

Figure 3. Conserved structure of the SCPP family of genes. A. The first three exons of typical SCPP genes. White, gray, and black regions respectively represent untranslated region (UTR), signal peptide (SP), and mature protein (MP). Introns located between two adjacent codons are called phase 0. B. Structure of some human SCPP genes (casein, CSN1S1; salivary protein, STATH; enamel protein, AMBN and AMEL; and dentine/bone protein, DMP1 and IBSP). The number in each exon represents length in nucleotides. Intron phases are shown between exons. Source: combined from the authors' work.

chromosome and adjacent to the original copy, at least for an evolutionary while. In both mammals and birds, a gene called *SPARCL1* is adjacent to the SCPP cluster. The SPARCL1 protein is a kind of superset of an SCPP gene, consisting of signal peptide followed by an upstream disordered domain, but then also a downstream part that is structurally well organized the way most proteins are. We proposed the evolutionary scenario in which *SPARCL1* was the "ancestral parent" of the other SCPP genes, which initially arose by a duplication event that only included the upstream *SPARCL1* regions. Subsequent SCPPs have been produced by duplications deriving from this initial SCPP gene. While the upstream disordered domain evolves rapidly, the downstream organized domain is very stable over long evolutionary times. So to go back into the deeper SCPP gene history, we used this stable domain to explore adjacent genes in the genome of a fish, the puffer fish, where we discovered many new SCPP genes. The puffer fish is an important model species because its genome is much shorter than those of most other vertebrates, including our own and those of other primates, its lineage having somehow excised many of the repeated sequence elements that litter other vertebrate genomes.

The puffer fish SCPP genes share the characteristic structure common to this gene family (Fig. 4B). Our specific interest is in dental evolution and, sure enough, the puffer SCPPs are expressed in its embryonic dental tissues. However, none of these genes exist in tetrapods or *vice versa*. They are expressed in somewhat different ways and have essentially no useful sequence homology with tetrapod SCPP genes (Fig. 4A). Outside the SCPP gene family, their characteristic structure is found only in the upstream region of the ancestral *SPARCL1* gene and a closely related *SPARC* gene.

SPARC is on a different chromosome from the SCPP genes and is even found in invertebrates. We think it is the ancestral gene of this whole vertebrate mineralization system, the "grandcestor" of the whole system. *SPARC* spawned *SPARCL1* some time after jawless fish like the lamprey, but before sharks, diverged from the common vertebrate lineage. Intriguingly, this is about the time when mineralized tissue first appeared in the body armor of Paleozoic jawless fish. Since then, various tetrapod SCPP genes have arisen and evolved new functions related to the evolution of egg-laying and lactating lineages in birds and mammal-like reptiles, respectively, while some SCPPs were secondarily lost by mutation in certain lineages. Thus, birds became edentulous in the Late Cretaceous and no longer have enamel genes.

Even more interesting is that, in addition to these derived functions, SCPP gene duplication has fashioned entirely new genes with similarly specialized functions. Independent proliferation of SCPP genes from a common *SPARCL1* ancestor in the tetrapod and teleost fish lineages has resulted, through phenogenetic drift, in entirely different, though SCPP-based, genetic bases for the mineralization of teeth.[3,4] Tetrapods and teleost fish both make teeth with an outer enamel-like layer over underlying dentine, but they do it in different ways using independently evolved SCPP genes.

NEW WAYS TO READ THE DISTANT PAST

Discerning genetic homology can sometimes be simple and straightforward. Even a computer can do it with only sequence data and no other information about the gene or its evolution. This is especially so when the taxa or genes are evolutionarily closely related and have not had too much time to accumulate differences, but it can even be done for distantly related genes when specific protein characteristics such as the folded shape of the protein are important to its function. But the SCPP genes show how adaptive natural selection can conserve homology of function, even when it involves different genes or only faintly retained DNA sequence homology.

The metaphor of a palimpsest is imperfect, but captures the general sense of what we see today in terms of the SCPP genes and the origin and evolution of vertebrates. We have to read the *logic* or relational

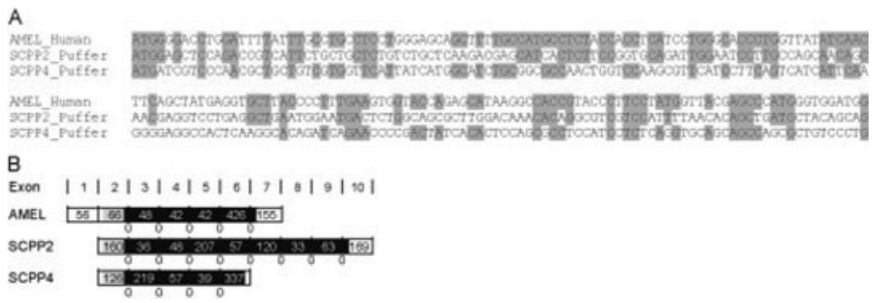

Figure 4. An SCPP palimpsest. A. Sequence alignment of human enamel protein (AMEL) and puffer fish enameloid protein (SCPP2, SCPP4) genes, showing no discernible sequence similarity. In fish, the surface tooth layer is called enameloid instead of enamel. B. Comparative structure of the same genes. Untranslated exon 1 is missing in these fish genes, but the overall structure is conserved. Source: combined from the authors' work.

principles of function, not just the writing on its genetic surface. Actually, the SCPP palimpsest was read by a combination of approaches. The tactic of looking for gene structure and the nature of protein rather than amino acid sequence helped place amelogenin in the SCPP gene family, conveniently close to several other enamel genes. Some of these genes had been identified from tooth germs, but not related evolutionarily to each other. The step-by-step peeling away of details to find the underlying message provided enough information for the more indirect challenge of finding the SCPP genes by working back to *SPARCL1* and forward from there to nearby SCPP genes in sharks and rays, whose SCPP genes probably arose totally independently from the cluster of their relatives in tetrapods.

Palimpsests provide a source of satisfying labor for scholars trying to understand life many centuries past. In a similar way, sometimes only patient work can discern from today's genomic writing what ancestral sequences were like. Other aspects of genetic as well as morphological or physiological evolution raise similar challenges. For example, protein coding is only one function in the genome. Gene regulation is another, in which nonprotein-coding sequence elements serve as physical binding sites for regulating proteins. Over time, these sequence elements can change so much that they are used by an entirely different regulatory protein even while what is being regulated does not change. Novel sequence-based functions with ancient histories continue to be discovered in our genome. What kinds of ghosts might such sequence patterns conceal? And how many other evolutionary palimpsests remain to be discovered in the Book of Life?

NOTES

I welcome comments on this column: kenweiss@psu.edu. I have a feedback and supplemental material page at http://www.anthro.psu.edu/weiss_lab/index.html. We thank Anne Buchanan and John Fleagle for critically reading this manuscript. Our work on the SCPP genes is supported by the National Science Foundation, grant BCS0343442, for which we are very grateful.

REFERENCES

1 Kawasaki K, Weiss KM. 2003. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. Proc Natl Acad Sci USA 100:4060–4065.

2 Kawasaki K, Suzuki T, Weiss KM. 2004. Genetic basis for the evolution of vertebrate mineralized tissue. Proc Natl Acad Sci USA 101:11356–11361.

3 Kawasaki K, Suzuki T, Weiss KM. 2005. Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. Proc Natl Acad Sci USA 102:18063–18068.

4 Kawasaki K, Weiss KM. 2006. Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family. J Exp Zool B Mol Dev Evol 306:295–316.