Darwinian alchemy: Human genes from noncoding DNA

Adam Siepel[1]

Department of Biological Statistics and Computational Biology, Cornell Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14853, USA

At least since the publication of Susumu Ohno's *Evolution by Gene Duplication* (Ohno 1970), the conventional wisdom has been that, in the emergence of novel genes, "natural selection merely modified, while redundancy created." In other words, new genes generally arise by the duplication of existing genes. While the notion that duplication plays a prominent role in the emergence of novel genes is perhaps most famously associated with Ohno, it actually traces back to the early days of the modern evolutionary synthesis (Bridges 1935; Muller 1936). Decades of modern sequence-based research have largely supported this general view (Graur and Li 2000). In recent years, the classic model of whole gene duplication and subsequent divergence has been enlarged to include phenomena such as exon shuffling, gene fusion and fission, retrotransposition, and lateral gene transfer (for review, see Long et al. 2003). Nevertheless, despite their additional complexity, these mechanisms remain essentially duplicative, in the sense that sequences encoding one or more protein-coding genes are copied, by one mechanism or another, and used as the starting point for a new gene sequence. (An exception is the exonization of noncoding transposable elements, such as *Alu*s, but this process tends to generate individual exons rather than entire genes;Makalowski et al. 1994; Nekrutenko and Li 2001.) By contrast, the origination of protein-coding genes de novo from nonrepetitive, noncoding DNA has been thought to occur only as an exceptionally rare event during evolution. Indeed, the emergence of complete, functional genes—with promoters, open reading frames (ORFs), and functional proteins—from "junk" DNA would seem highly improbable, almost like the elusive transmutation of lead into gold that was sought by medieval alchemists.

Over the past few years, this view has begun to change, with several reports of de novo gene origins in *Drosophila* and yeast (Levine et al. 2006; Begun et al. 2007; Chen et al. 2007; Cai et al. 2008). Zhou et al. (2008) have estimated that as many as ~12% of newly emerged genes in the *Drosophila melanogaster* subgroup may have arisen de novo from noncoding DNA, independently of transposable elements. Recently, Toll-Riera et al. (2009) identified 15 such genes in primates. Now, in this issue, Knowles and McLysaght (2009) demonstrate for the first time that human genes have arisen de novo from noncoding DNA since the divergence of the human and chimpanzee genomes. They identify and analyze three human genes that have no known homologs, in the human genome or any other, and do not appear to derive from transposable elements. Rather, these are cases in which mutation, natural selection, and/or neutral drift have evidently forged ORFs and functional promoters out of raw genomic DNA, like a blacksmith shaping a new tool from raw iron.

To identify these recent gene "births" in human, Knowles and McLysaght used a straightforward but rigorous approach. They began with a candidate set of several hundred human genes not annotated in the chimpanzee genome and winnowed these genes down to a high-confidence subset using a series of conservative bioinformatics filters. These filters eliminated candidate human genes that mapped to gaps in the chimp genome, that aligned to possible (unannotated) genes in orthologous locations in the chimp or macaque genomes, or that had annotated orthologs in any other species. In this way, a starting set of 644 human genes was reduced to just three genes. Several follow-up analyses then provided further support that these three genes represented de novo origins in recent human evolution.

Validating alleged gene births is a tricky business, because it requires showing not only that the new genes are functional but also that their evolutionary antecedents were nonfunctional. Knowles and McLysaght drew upon several lines of evidence in their efforts at validation. First, to establish that the human genes were probably functional, they considered evidence of both mRNA expression and protein expression. They showed that each gene was supported by at least one complete, spliced (human) cDNA sequence from GenBank and by at least one unique short (human) peptide from the PRIDE or PeptideAtlas proteomics databases, suggesting it was both transcribed and translated in human cells. Next, to establish that the genes most likely did not encode functional proteins in ancestral primates, they looked to orthologous sequences in the chimp and macaque genomes identified using syntenic alignments. In all three cases, they found multiple disabling mutations (such as absent start codons, premature stop codons, or frame-shifting indels) in the chimp and macaque orthologs. Moreover, each gene had at least one disabling mutation (supported by high-quality sequences) that was shared between its chimp and macaque orthologs, suggesting an absence of protein-coding function at least since the divergence of the Great Apes and Old World Monkeys (roughly 25 million years ago). These shared disablers were also present in the gorilla and gibbon genomes, and, for two out of three genes, in the orangutan genome. To help rule out the possibility that the presence/absence of a functional gene might be polymorphic in chimpanzees, Knowles and McLysaght resequenced the regions in question in another chimpanzee individual and verified that the disabling mutations were present. These experiments do not provide absolute proof that de novo gene origins occurred on the human lineage, but they strongly suggest that the three genes are transcribed and translated in humans, yet did not encode proteins in ancestral primates.

What properties, if any, do the three identified genes share? Not surprisingly, they all have short ORFs (121–163 amino acids) and lack introns in their coding regions, although they do (all three) have introns in their untranslated regions (UTRs). Most previously identified de novo genes have been short, with one or two exons (Toll-Riera et al. 2009). Interestingly, two out of the three genes in this case, and all three ORFs, fall within introns of genes on the opposite strand. (The third has a long 3′ UTR that

[1]**E-mail acs4@cornell.edu; fax (607) 255-4698.**
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.098376.109.

overlaps several exons and introns of another gene; see below.) As is typical of "orphan" genes, little is known about their functions. Two of the genes (encoding proteins called DNAH10OS and C22orf45) are completely uncharacterized. The third (encoding a protein called CLLU1) has been shown to be significantly upregulated in an aggressive form of chronic lymphocytic leukemia (Buhl et al. 2006) and subsequently has been analyzed in some detail at the level of mRNA expression (Buhl et al. 2009), but its function remains unknown. Their mRNA expression patterns are not distinctive (all are expressed in multiple tissues), nor are their patterns of within-species polymorphism (all are present in apparently functional form in three fully sequenced individuals and do not show significant evidence of positive selection). These last two features are notable (albeit based on somewhat sparse data) because novel genes in *Drosophila* have shown a strong tendency for testis-specific expression and evidence of positive selection (Levine et al. 2006; Begun et al. 2007; Chen et al. 2007). Retroposed genes in human are also strongly enriched for expression in the testis (Vinckenbosch et al. 2006).

These apparent de novo gene origins raise the question of how evolution by natural selection can produce functional genes from noncoding DNA. While a single gene is not as complex as a complete organ, such as an eye or even a feather, it still has a series of nontrivial requirements for functionality, for instance, an ORF, an encoded protein that serves some useful purpose, a promoter capable of initiating transcription, and presence in a region of open chromatin structure that permits transcription to occur. How could all of these pieces fall into place through the random processes of mutation, recombination, and neutral drift—or at least enough of these pieces to produce a protogene that was sufficiently useful for selection to take hold? One compelling solution for the general problem of the evolution of new features, called variously "preadaptation," "cooption," and "exaptation" (Gould and Vrba 1982) (and famously illustrated using the architectural metaphor of the "spandrel"; Gould and Lewontin 1979), is that complex new features can arise through alteration of pre-existing features—that evolution arrives at new forms by "tinkering" with forms that have previously evolved for other purposes (Jacob 1977). For example, bird feathers are believed to have evolved originally for temperature regulation, then to have been adapted for use in flight. Indeed, it is probably because of the principle of exaptation that most genes arise via gene duplication; there is no better starting point for a new gene than another gene. The fact that Knowles and McLysaght's novel genes overlap genes on the opposite strand hints at a more subtle form of tinkering. The overlapping genes might tend to make circumstances more favorable for transcription, by ensuring that the chromatin is open or by supplying *cis*-regulatory elements that promote transcription on both strands. They may also increase the likelihood that an ORF of nontrivial length occurs by chance, through CpG islands or elevated G + C content. Thus, while these new genes have not arisen directly from other genes, one might speculate that, in a sense, they "drafted" behind other genes on their journey to functionality. In other words, what was reused in the creation of these genes was not the actual protein-coding sequence but the general genomic context for protein-coding functionality. Interestingly, novel genes that emerge by retroposition have been shown to occur preferentially near other genes or within introns, suggesting the same type of reuse of genomic context (Vinckenbosch et al. 2006).

The study by Knowles and McLysaght (2009) does have some important limitations. First, any gene classified as "known" by Ensembl was assumed to be accurately annotated, even though some of these genes have scant support (one or two cDNA sequences with no other supporting evidence). Gene prediction, even with cDNAs, is an unsolved problem, and the catalogs of "known" genes have been found to contain significant numbers of spurious annotations (Clamp et al. 2007). Single-exon genes and genes that overlap other genes are especially difficult to predict correctly. Knowles and McLysaght's requirement of supporting peptides from proteomics experiments should help to alleviate this problem, but such data have their own limitations, for example, relating to uniqueness of peptides and sample contamination. Indeed, one of the three genes identified in the study, associated with the peptide C22orf45 and called ENSG00000204626 in Ensembl, appears dubious—it is supported by only a single spliced cDNA sequence (AK127211) and is predicted to have an intron within a long 3′ UTR, which is extremely rare in eukaryotic genes (Nagy and Maquat 1998). This gene is not present in the RefSeq, UCSC Genes, Vega, or CCDS gene sets, and it appears to have been recently removed from Ensembl. This gene does have two supporting peptides and may truly be functional, but more supporting evidence would be welcome.

In addition to the issue of false positive genes is the question of false negatives—that is, genes that were missing from the starting gene set or erroneously discarded. Because of the use of strict filters, only a relatively small subset of known genes (an estimated 4000) was ultimately considered by Knowles and McLysaght (2009). Assuming a total of 24,000 genes, the authors estimate that the total number of de novo gene births since the human/chimp divergence is about $(24{,}000/4{,}000) \times 3 = 18$. However, this estimate is very crude. It could be strongly biased by a nonrandom association between gene births and genes excluded by the filters—for example, an increased likelihood of gene births in duplicated or rearranged regions of the genome, which were excluded because of a requirement of conserved synteny with other primates. It also does not consider the possibility that significant numbers of genes may be absent from the current gene catalogs (Siepel et al. 2007) and that genes like the three that were identified—with short, single-exon ORFs, relatively weak cDNA support, and no known homologs—are especially likely to be missing. It seems fair to say that the number of recently emerged human protein-coding genes has not yet been estimated with any certainty.

Finally, the possibility that apparent gene births were actually functional in ancestral genomes and were lost independently in multiple lineages, although remote for these genes, cannot be completely discounted. Mutational hotspots could lead to non-negligible probabilities of parallel (homoplastic) disabling mutations. Indeed, Knowles and McLysaght observe a case in which an apparently enabling mutation in human (an ORF-creating deletion) has an exact parallel in orangutan. The same type of scenario could occur in the opposite direction, rendering multiple disabled descendant genes from a functional precursor. In addition, the low probability of any particular nonparsimonious scenario has to be weighed against the fact that hundreds of genes were tested, and only the cases in which the hypothesis of ancestral protein-coding function had low probability were selected. Proper modeling of mutational rate variation and the effects of multiple testing might show that the probabilities of these multiple-disablement scenarios are considerably larger than intuition would suggest.

While it is not the final word on de novo gene origins in human, Knowles and McLysaght's elegantly simple study is notable for several reasons. Along with other recent work (Toll-Riera

et al. 2009), it demonstrates convincingly that primate genomes contain true "orphan" genes, lacking known homologs in other species. This serves as an important reminder of the limits of sequence similarity (whether of orthologs or paralogs) in identifying and characterizing protein-coding genes in these genomes. Because the methods for identifying these genes so far have been quite conservative, it is possible that many more exist but have yet to be found. In addition, this study helps to shed light on the process by which evolution by natural selection can forge completely new functional elements from apparently nonfunctional DNA—the process by which molecular evolution turns lead into gold, as it were. These genes appear to be cases in which a few serendipitous mutations were sufficient to generate minimal ORFs and working promoters from noncoding sequences, perhaps aided by the presence of genes on the opposite strand. The genome is large and at any given time is likely to contain sequences that are at most a few mutational steps from minimal functional elements. One can imagine a process by which short, simple genes periodically arise de novo, then gradually become more complex over time, by obtaining longer coding regions, introns, alternative splice forms, and so on, through processes such as duplication, mobile element insertion, rearrangement, and point mutation—much as in the well-studied case of *hydra*, in *Drosophila* (Chen et al. 2007). Thus, the genes identified by Knowles and McLysaght (2009), together with similar genes in *Drosophila*, yeast, and other primates, can be thought of as missing links that help to demystify the alchemist's sorcery.

Acknowledgments

Support for this article was provided by early career awards from the David and Lucile Packard Foundation and the National Science Foundation (grant DBI0644111). I thank Dan Barbash for comments on the manuscript.

References

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* **176:** 1131–1137.

Bridges C. 1935. Salivary chromosome maps. *J Hered* **26:** 60–64.

Buhl AM, Jurlander J, Jorgensen FS, Ottesen AM, Cowland JB, Gjerdrum LM, Hansen BV, Leffers H. 2006. Identification of a gene on chromosome 12q22 uniquely overexpressed in chronic lymphocytic leukemia. *Blood* **107:** 2904–2911.

Buhl AM, Novotny GW, Josefsson P, Nielsen JE, Pedersen LB, Geisler C, Rassenti LZ, Kipps TJ, Jurlander J, Leffers H, et al. 2009. The CLLU1 expression level is a stable and inherent feature of the chronic lymphocytic leukemia clone. *Leukemia* **23:** 1182–1186.

Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179:** 487–496.

Chen ST, Cheng HC, Barbash DA, Yang HP. 2007. Evolution of *hydra*, a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster*. *PLoS Genet* **3:** e107. doi: 10.1371/journal.pgen.0030107.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci* **104:** 19428–19433.

Gould SJ, Lewontin RC. 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* **205:** 581–598.

Gould SJ, Vrba ES. 1982. Exaptation—a missing term in the science of form. *Paleobiology* **8:** 4–15.

Graur D, Li WH. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, MA.

Jacob F. 1977. Evolution and tinkering. *Science* **196:** 1161–1166.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* (this issue). doi: 10.1101/gr.095026.109.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci* **103:** 9935–9939.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat Rev Genet* **4:** 865–875.

Makalowski W, Mitchell GA, Labuda D. 1994. *Alu* sequences in the coding regions of mRNA: A source of protein variability. *Trends Genet* **10:** 188–193.

Muller HJ. 1936. Bar duplication. *Science* **83:** 528–530.

Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. *Trends Biochem Sci* **23:** 198–199.

Nekrutenko A, Li WH. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17:** 619–621.

Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, Heidelberg, Germany.

Siepel A, Diekhans M, Brejova B, Langton L, Stevens M, Comstock C, Davis C, Ewing B, Oommen S, Lau C, et al. 2007. Targeted discovery of novel human exons by comparative genomics. *Genome Res* **17:** 1763–1773.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of primate orphan genes: A comparative genomics approach. *Mol Biol Evol* **26:** 603–612.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci* **103:** 3220–3225.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W, et al. 2008. On the origin of new genes in *Drosophila*. *Genome Res* **18:** 1446–1455.