SCIENCE & SCILIFELAB PRIZE

Evolution of Vertebrate Transcriptional Regulator Binding

Comparative evolutionary analysis of transcriptional regulator binding across several vertebrate species reveals intricate regulatory genome evolution.

Dominic Schmidt

Vertebrates contain hundreds of different cell types that develop and maintain their phenotypic identity by a combination of genomic and epigenomic regulation. What are the regulatory mechanisms that enable one vertebrate genome to give rise to this magnificent diversity? And how are these mechanisms exploited over evolutionary time to allow for divergence and to give rise to new functions and ultimately species?

Science and SciLifeLab are pleased to present the essay by Dominic Schmidt, a 2013 second runner-up for the *Science* & SciLifeLab Prize for Young Scientists.

In 2001, the first nearly complete sequence of a vertebrate genome, the human genome, was published (*1*, *2*). Soon after, several other genomes of vertebrates—such as mouse, rat, dog, opossum, and chicken—were reported. The tremendous effort put into sequencing and assembling these genome sequences is a prerequisite to furthering our understanding of genetic information and its role in development, disease, and evolution. One of the first insights from comparative genomics was unexpected, namely, that the majority of human genes have a single identifiable ortholog in other vertebrate species (*3*, *4*). Because of the combination of our understanding of the genetic code and comparative genomic sequencing, we know that protein-coding sequences are under strong purifying selection and are, therefore, highly conserved between species (*3*). However, the vast majority of a vertebrate's genome does not code for proteins, and the evolution and function of those noncoding sequences is poorly understood. Some of the noncoding sequences in the human genome serve regulatory functions, and it was proposed decades ago that regulatory variation may explain many of the phenotypic differences that can be observed between closely related species given the few differences in their protein-coding sequence (*5*). Exactly how regulatory sequences evolve over evolutionary times remains to be understood and is of particular importance given the frequent involvement of regulatory changes in many human diseases.

A comparison between human and mouse showed that transcription factor–binding sites are considerably less conserved than protein-coding sequences (*6*–*9*). My Ph.D. thesis extended these prior analyses by comparing in vivo binding of the tissue-specific transcription factors CEBPA and HNF4A among human, mouse, dog, opossum, and chicken. Although tens of thousands of binding events are found in each individual species and the DNA binding preferences of the transcription factors are highly conserved, most binding is species-specific. For example, any two of the three placental mammals we analyzed shared 10 to 20% of the binding events, and this divergence increased further with greater evolutionary distance. Nonetheless, we found that functional target genes of these two factors were enriched for shared binding events. It is conceivable that binding events found in two species represent a core set of functional regions that are deeply conserved across multiple species. Thus, we tested whether there exists a subset of binding events that is shared among all five vertebrate species and, consequently, that must have been preserved for more than 300 million years. We found that only very few binding events were conserved across all five species and that they represent less than 0.3% of the total binding events found in humans.

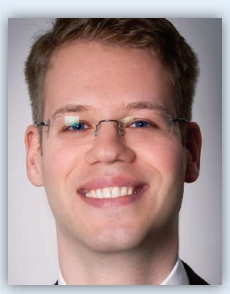
By comparing multiple species, we were further able to investigate the genetic mechanisms underlying the rapid gain and loss of binding events that we observed. The loss of the majority of binding events can be explained by disruption of the transcription factor's binding motif as a result of changes in the DNA sequence, whereas species-specific gains of binding events are frequently found in novel sequences that cannot be aligned with the other species (*10*).

Not all transcription factor binding seems to evolve in the same way as we observed for CEBPA and HNF4A. CCCTC-binding factor (CTCF) is an almost ubiquitously expressed DNA-binding protein that can divide transcriptional domains and appears to be involved in the three-dimensional organization of the genome (*11*, *12*). There have been somewhat conflicting reports suggesting that the binding events of CTCF are considerably more conserved between mammals, whereas (at the same time) they appear to have evolved in the mouse genome by means of rodent-specific retrotransposon expan-

L.E.K. Consulting, 40 Grosvenor Place, London, SW1X 7JL, UK. E-mail: dominic.schmidt@cantab.net

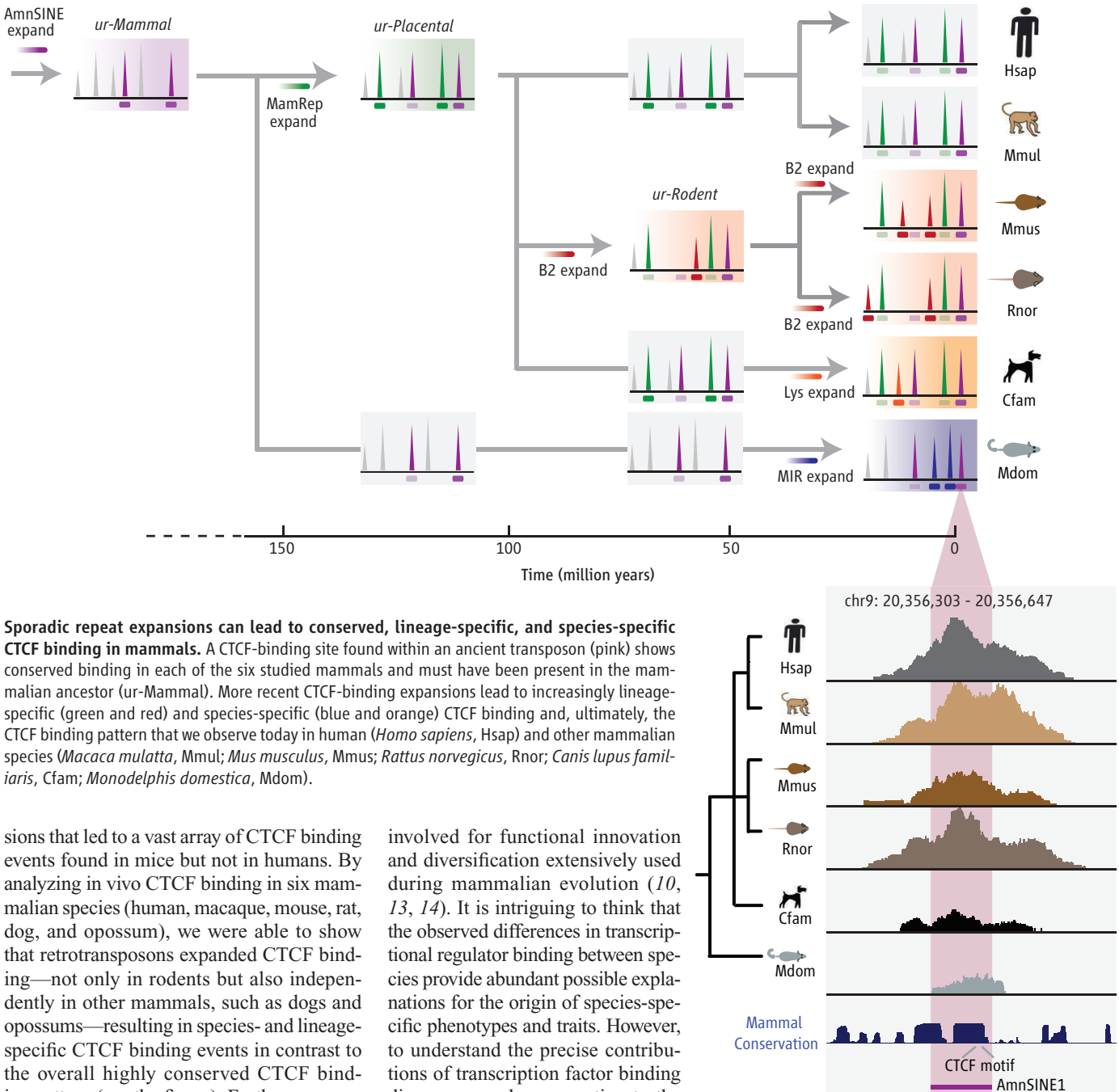
CREDIT: MARK WOOD

2013 Second Runner-Up



For his essay in the category of Genomics/Proteomics/Systems Biology, **Dominic Schmidt** is a second runner-up. Dr. Schmidt is a Strategy Consultant at L.E.K. Consulting in London where he works as a strategic adviser to the biopharma and life sciences industry. He received his Ph.D. in Oncology from the University of Cambridge where he combined experimental and computational approaches across multiple species to study how gene-regulation and genomes are evolving. Before getting his Ph.D., he received his German diploma degree in biochemistry at the Max Planck Institute for Molecular Genetics and the Free University of Berlin.

For the full text of all winning essays and further information, see the *Science* site at http://scim.ag/SciLifeLab.



Sporadic repeat expansions can lead to conserved, lineage-specific, and species-specific CTCF binding in mammals. A CTCF-binding site found within an ancient transposon (pink) shows conserved binding in each of the six studied mammals and must have been present in the mammalian ancestor (ur-Mammal). More recent CTCF-binding expansions lead to increasingly lineage-specific (green and red) and species-specific (blue and orange) CTCF binding and, ultimately, the CTCF binding pattern that we observe today in human (*Homo sapiens*, Hsap) and other mammalian species (*Macaca mulatta*, Mmul; *Mus musculus*, Mmus; *Rattus norvegicus*, Rnor; *Canis lupus familiaris*, Cfam; *Monodelphis domestica*, Mdom).

sions that led to a vast array of CTCF binding events found in mice but not in humans. By analyzing in vivo CTCF binding in six mammalian species (human, macaque, mouse, rat, dog, and opossum), we were able to show that retrotransposons expanded CTCF binding—not only in rodents but also independently in other mammals, such as dogs and opossums—resulting in species- and lineage-specific CTCF binding events in contrast to the overall highly conserved CTCF binding pattern (see the figure). Furthermore, we established that CTCF binding that has been conserved over millions of years is sometimes found within ancient, fossilized repeat elements outside protein-coding regions that are still shared between distinct mammalian lineages and are likely of critical importance for mammalian characteristics. This indicates that similar retrotransposon expansions that occurred millions of years ago might have resulted in the highly conserved CTCF binding pattern that we observe today (*13*).

Taken together, my thesis work produced insights into the evolution of transcription factor binding and some of the mechanisms involved for functional innovation and diversification extensively used during mammalian evolution (*10*, *13*, *14*). It is intriguing to think that the observed differences in transcriptional regulator binding between species provide abundant possible explanations for the origin of species-specific phenotypes and traits. However, to understand the precise contributions of transcription factor binding divergence and conservation to the organismal phenotypes of vertebrate species will require that we can read the regulatory code as easily as we read the genetic code. Further combined efforts of experimental and computational approaches across multiple cell types and species will be required for eventually deciphering the regulatory code.

References and Notes

1. E. S. Lander *et al.*; International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
3. R. H. Waterston *et al.*; Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (2002).
4. International Chicken Genome Sequencing Consortium, *Nature* **432**, 695 (2004).
5. M. C. King, A. C. Wilson, *Science* **188**, 107 (1975).
6. D. T. Odom *et al.*, *Nat. Genet.* **39**, 730 (2007).
7. Y.-H. Loh *et al.*, *Nat. Genet.* **38**, 431 (2006).
8. C. M. Conboy *et al.*, *PLOS ONE* **2**, e1061 (2007).
9. E. T. Dermitzakis, A. G. Clark, *Mol. Biol. Evol.* **19**, 1114 (2002).
10. D. Schmidt *et al.*, *Science* **328**, 1036 (2010).
11. K. L. Dunn, J. R. Davie, *Biochem. Cell Biol.* **81**, 161 (2003).
12. J. E. Phillips, V. G. Corces, *Cell* **137**, 1194 (2009).
13. D. Schmidt *et al.*, *Cell* **148**, 335 (2012).
14. D. Schmidt *et al.*, *Genome Res.* **20**, 578 (2010).

Acknowledgments: Supported by the European Research Council (to D. Odom), European Molecular Biology Organization Young Investigator Program (to D. Odom), Hutchinson Whampoa

(to D. Odom), Cancer Research UK, the University of Cambridge, the Wellcome Trust (to P. Flicek), and European Molecular Biology Laboratory (to P. Flicek).

10.1126/science.1247569

Evolution of Vertebrate Transcriptional Regulator Binding

Dominic Schmidt

Science **342** (6163), 1186.
DOI: 10.1126/science.1247569