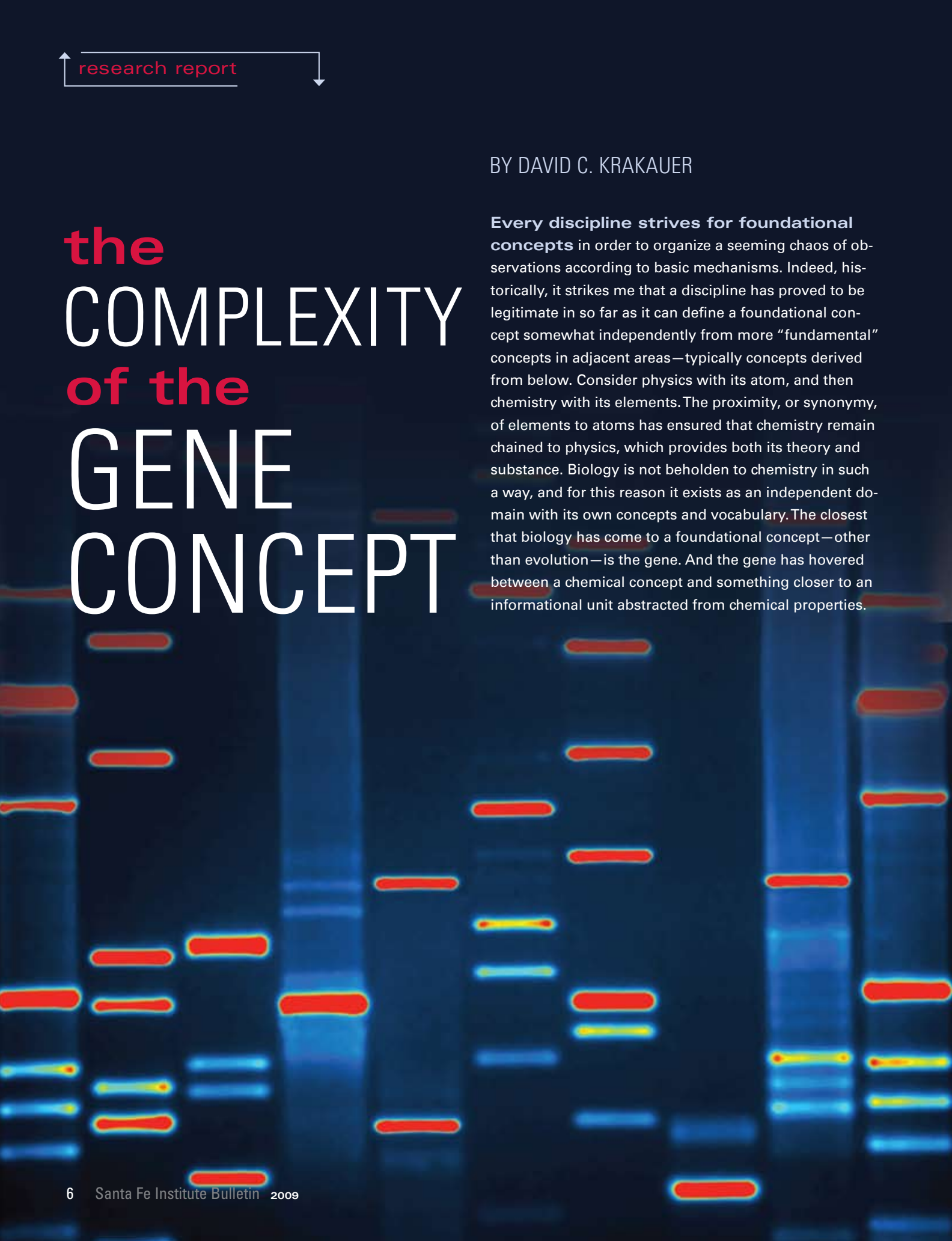
# the COMPLEXITY of the GENE CONCEPT

BY DAVID C. KRAKAUER

**Every discipline strives for foundational concepts** in order to organize a seeming chaos of observations according to basic mechanisms. Indeed, historically, it strikes me that a discipline has proved to be legitimate in so far as it can define a foundational concept somewhat independently from more "fundamental" concepts in adjacent areas—typically concepts derived from below. Consider physics with its atom, and then chemistry with its elements. The proximity, or synonymy, of elements to atoms has ensured that chemistry remain chained to physics, which provides both its theory and substance. Biology is not beholden to chemistry in such a way, and for this reason it exists as an independent domain with its own concepts and vocabulary. The closest that biology has come to a foundational concept—other than evolution—is the gene. And the gene has hovered between a chemical concept and something closer to an informational unit abstracted from chemical properties.



The foundational status of the gene concept is evident when we consider how it is used at all levels of biological organization—molecular biology, development, physiology, behavior, medicine, evolution, and even culture. In few of these cases is atomic chemistry the critical property, but rather some kind of discrete, regulatory unit with a heritable, causal influence. The two senses in which the gene is most commonly used are either as a memory molecule or as a determinant of the phenotype—anatomy, physiology, or behavior. From these two usages derives the gene's value in the study of inheritance, and in applied areas such as medicine. And these two qualities fuse when we consider how phenotypic traits are transmitted between generations and how these traits evolve.

In order to evaluate the current state of the gene concept, the Santa Fe Institute recently convened a workshop on the "Complexity of the Gene Concept." The meeting was engendered by the central role of the gene in organizing biological observations and theories, and the failure of the simple, chromosomal model of the gene—first proposed by Thomas Hunt Morgan—in the light of huge quantities of genetic data in digital form. The Morgan model describes the gene as beads on a string, each string a chromosome, and with each bead standing for a DNA nucleotide contributing a quantum of character. What emerged from the meeting were gene concepts grounded in a more compelling view of the relationship between a gene's structure and its function, which often included informational and computational principles.

For such a concept with such widespread influence and acceptance, the gene remains surprisingly slippery. At one level there is just the chemistry, and at another level, there are its effects—the color of the eyes, differential resistance to disease, and the ability to fly or swim. Relating phenotypes and their functions to the materiality of the gene is similar to the challenge of relating the mind to the brain or software to hardware. And these analogies provide a clue as to the nature of the gene. Rather like a computation which can be understood both from the perspective of transistors and of algorithms, the gene can be conceived both as coded information and materially. In striving to define the gene, these two varieties of meaning compete for scientific dominance.

The gene, thought about as a single stretch of contiguous DNA or RNA, transcribed and translated into a unique protein, with unambiguous expression and quantifiable selective value, has always been an ideal rather than a reality. Under most definitions the gene has been presumed to perform three functions: to serve as a unit of inheritance, a regulatory element in developmental dynamics, and an atomic unit of selection. In each case there is a mutable component, well behaved and easily identifiable, that survives cell division, can be turned on and off as a unit through suitable regulatory pathways, and contributes a quantum of fitness to an organism when expressed. The work—contribution to heritable, regulatory or selective variance—in each of these cases is presumed to be done by the gene, and so the gene occupies, understandably, a central position.

These many properties of the gene have contributed to a lively historical debate. Darwin, who was so clear on natural selection, became a little hazy when discussing the units of inheritance. Darwin's first problem was finding a suitable name. In a letter to his son George, at the time

PHOTO RESEARCHERS, INC.

Facing Page: Colorized images from human DNA; Above: Inbreeding has been common among royal families, causing genetic disorders. Married to her first cousin, Queen Victoria (1819–1901) carried the gene causing hemophilia, which passed to her children, and then on to the Spanish and Russian royal families.

LEFT: DNA AS ART, BY DNA11.COM

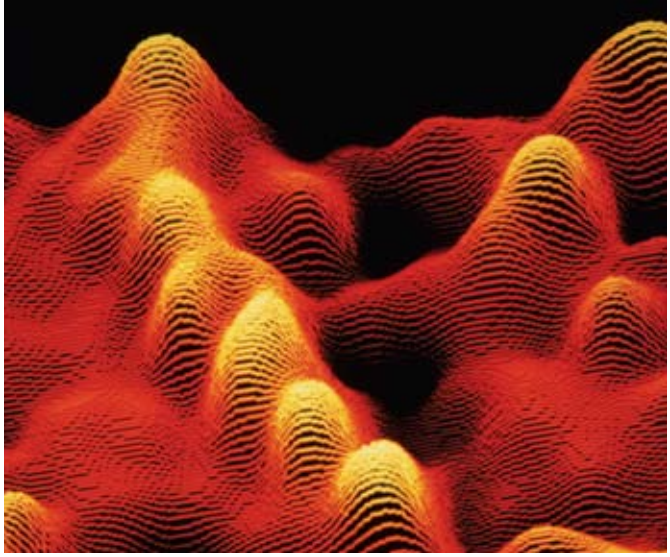


PHOTO RESEARCHERS, INC.

The Center for Human Genome Research created this color-enhanced image, a scanning tunneling electron micrograph (STEM) of a right-handed DNA duplex. The center's goal is to construct a physical map of DNA in 24 chromosomes.

studying mathematics at Cambridge, Darwin requested the advice of a classicist "who could suggest any Greek word expressing cell, and which could be united with genesis." Finally settling on the term pangenesis for his theory of inheritance, which suffered neglect for its complexity, Darwin wrote to the geologist Charles Lyell, "My fear has always been that pangenesis would be a still-born infant, over whom no one would rejoice or cry."

The SFI workshop, hosted by Institute researchers Peter Stadler, from the University of Leipzig, Sonja Prohaska from Arizona State University, Manfred Laubichler also from Arizona State University, and me, and supported by the McDonnell Foundation, sought to synthesize the growing body of somewhat contradictory data bearing on the gene concept. The idea was to bring together researchers for whom genetics is a critical consideration, but among whom the details of analysis vary enough to foster rather different operational definitions of the gene.

Representing bioinformatics, Stadler and Prohaska both expressed concern that annotation and taxonomic identification of genes is being hindered and obfuscated by the traditional, beads-on-a-string concept, which needs to be replaced. They suggest a DNA-based concept of distributed sequences understood in terms of context-dependent mappings onto RNA and protein. From a DNA-editing viewpoint, James Shapiro (Univ. of Chicago) called for the abolition of the gene concept based on its spurious unity and operational disutility. From RNA editing, Thomas Gingeras (Cold Spring Harbor Laboratory) recommended locating the gene concept at the level of the growing set

Darwin wrote to the geologist Charles Lyell, "My fear has always been that pangenesis would be a still-born infant, over whom no one would rejoice or cry."

of RNA transcripts where information is integrated. From the philosophy of biology, Richard Burian (Virginia Tech) was keen to ensure that a new gene concept could accommodate the elaborate roles of single gene sequences in multiple developmental contexts. Kenneth Weiss (Penn State), who has worked on the genetics of disease, emphasized the role that numerous, small mutations distributed over the genome play in defining traits and the value of operational-based definition of ordered sequences rather than genetic units. Douglas Erwin (SFI, Smithsonian) compared the proliferation of gene concepts to the zoo of species definitions and urged a practical approach so as to avoid discord and focus on critical developmental implications. From theoretical chemistry, Christian Forst (Univ. of Texas Southwestern Medical Center) was outspoken in dismissing the gene as an idea that has outlived its usefulness in an age of detailed, microscopic data.

All speakers agreed that a new gene concept needs to deal with the problem of distributed sequences, playing multiple roles in multiple contexts. If the regions of DNA sequence from which an RNA transcript is synthesized are distributed over the entire genome, or if multiple proteins all make use of the same sequence, the work is not performed purely by a sequence gene, but by the constructive processes capable of locating, transcribing, and concatenating all the relevant transcripts into a new sequence which behaves as if it were the traditional reference gene. This implies that most of the interesting dynamics take place in the coordination of the transcripts, and suggests that the gene might better be thought about in terms of input-output functions or mappings: That is, those functions that take as input, or arguments, a heterogeneous set of sequences typically in DNA form, and transform those inputs onto downstream RNA and protein targets that possess the functions that we formerly assigned to "the gene" as a contiguous DNA sequence.

Under this model, the new reference gene is partly a state-

ment about DNA-based memory, and partly a regulatory concept. This is because it suggests that the appropriate units of function, or modules, are those pathways capable of turning distributed DNA sequences into functional RNA transcripts. Mutations to these pathways are associated with modification of the phenotype, and the regulation of these pathways provides the raw material upon which gene regulation and natural selection then operates.

We might think about this modified gene concept in computational terms as some procedural element, or function, instantiated in sequences of code, contributing to one or more adaptive behaviors. The procedure or function describes the set of regulatory operations to be executed in some systematic fashion to generate a stable transcript. The code that furnishes the arguments for the function is the ordered collection of nucleotides stored in an enzyme-readable form distributed over the genome. And the final output of the procedure is the modification of phenotypic variability through contributions to cellular function. The gene is thereby a computational, or algorithmic element, exploiting underlying sequence structures, and is not merely a distributed structure itself.

Thus when we speak of selfish genes we are really speaking of a selfish function and its arguments, and not just an inert sequence of DNA or RNA base-pairs. And when we compare genomes among species, we ought to be comparing them at the level of these functions that are the true source of evolutionary variation, rather than at the level of the sequences which provide the combinatorial raw material for transcript production. One intriguing implication of this approach to the evolution of biological complexity is that it should provide a more satisfying metric than the current one, in which a simple gene-as-sequence has the disconcerting property of making primates and worms virtually indistinguishable at the genomic level, and severely reducing the resolution of cross-species comparison. ◀

*David Krakauer is an SFI professor.*

*The Parade of Memories* by Desmond Morris (20th C., British)

COURTESY OF DESMOND MORRIS