1 Overview

THE GENETIC MATERIAL?

Proteins have been associated with all biological processes since the inception of biochemistry in the 19th century, given their abundance, enzymatic properties and versatility. On the other hand, although identified in 1869, the functions of nucleic acids remained obscure until the 1940s. In the years leading up to the turn of the 20th century, DNA was found to be localized in chromosomes, which were shown to be the vehicles for genetic inheritance. In 1909, it was proposed that nucleic acids form simple tetramers containing each of the four component nucleotides, following which it was generally thought, because of their presumed repetitive structure, that nucleic acids have only peripheral functions.

Accordingly, proteins, which are also found in chromosomes, were regarded as the repository of genetic information for the first four decades of the 20th century, with DNA functioning as a scaffold. However, in 1944, DNA was demonstrated to be the 'transforming principle' in bacteria, although this finding was only widely accepted after bacteriophage infection 'pulse-chase' experiments in 1952, the elucidation of the structure of DNA in 1953 and the demonstration of its semi-conservative replication in 1958.

While having a nucleotide composition similar to DNA, RNA did not appear to play a role in the intergenerational transmission of genetic information, although RNA viruses were later found to exist. It was regarded for decades as an uninteresting metabolic molecule in bacteria, yeast and plants, and only conclusively shown to exist in animal cells in the 1930s.

Microbial genetics from the 1920s established that (some) genes encode proteins, but the mechanism by which this occurred was unknown. Gradually it dawned that RNA might be involved, inferred from histochemical, ultracentrifugation and spectroscopic studies in the 1940s that showed that RNA is present in cytoplasmic microsomes (ribosomes), which were becoming recognized as the sites of protein synthesis.

Meanwhile, theoretical biologists had declared, in the so-called Modern Synthesis reconciling Darwinian evolution with Mendelian genetics, that mutations are random, and that Lamarckian inheritance of experience does not occur. Moreover, the emphasis on lethal protein-coding mutations muddled the interpretation of genetic variation, with ongoing debates between the 'Mendelians' and the quantitative geneticists.

HALCYON DAYS

The abundant ribosomal RNAs (rRNAs) were identified in the mid-1950s, but a specific function for RNA was demonstrated only in 1958, when small RNAs were shown to act as 'adaptors' for the incorporation of amino acids into microsomal proteins, named 'transfer RNAs' (tRNAs). In 1961, the radioactive labeling of 'messenger RNAs' (mRNAs) finally identified the 'unstable' intermediate between genes and proteins, establishing the connection.

In the following decade, the triplet 'genetic code' for protein synthesis was deciphered. Analysis of the lactose (*lac*) operon of *Escherichia coli* cemented the conclusion that genes are synonymous with proteins. The regulation of gene activity by protein 'transcription factors' was established and assumed to hold not just in bacteria but also in developmentally complex organisms. All that remained to do, it seemed, was to flesh out the details.

WORLDS APART

It was obvious by that time that plants and animals are orders of magnitude more complex than bacteria and have different cellular and genetic features, including much greater internal compartmentalization and far larger genomes. It was later shown that eukaryotic cells arose by fusion of bacterial and archaeal cells and that developmentally complex organisms burst onto the scene in spectacular adaptive radiations, most likely following regulatory innovations required to orchestrate organized cell division and differentiation.

Studies using newer techniques in the 1960s and 1970s showed that that eukaryotic DNA is packaged in a repeating structure ('nucleosomes') comprised

DOI: 10.1201/9781003109242-1

of basic proteins called histones, and that chromatin is compacted and remodeled during development. It was found that histones are dynamically modified by methylation and acetylation, which suggested that histone modifications act as a regulatory mechanism. It was also shown that RNA is associated with chromatin and that very high molecular weight 'heterogeneous' RNAs are synthesized in the nucleus, predicted to be precursors of mRNAs, but the function of the remainder of these transcripts was mysterious.

STRANGE GENOMES, STRANGE GENETICS

The use of the fruit fly *Drosophila melanogaster* as a model genetic system from the 1910s enabled the mapping of genes along chromosomes by measuring recombination distances (co-inheritance frequencies), which established the view of genes as discrete, 'particulate' entities. Analysis of naturally occurring and radiation-induced mutations identified 'homeotic' loci that caused bizarre segmental transformations along with other encoding epigenetic 'modifiers' that exhibited strange interactions.

Odd genetic phenomena were also reported in plants. 'Rogue' non-Mendelian patterns of inheritance were observed in peas in 1915 and characterized in other species from the 1950s, termed 'paramutation', later understood to be a feature of transgenerational epigenetic inheritance. Mobile 'controlling elements' were identified in maize in the 1940s and shown to be due to the transposition of regulatory cassettes. In the mid-1960s, large fractions of the genomes of plants and animals were found to be comprised of 'repetitive sequences', most of which derive from transposable elements. It was also found that the repetitive sequences are differentially transcribed.

In 1969, these disparate molecular observations were integrated into a schema of gene regulation in embryonic development, which included the concepts of 'structural' (protein-coding) and 'integrator' genes (most likely) expressing regulatory RNAs recognized by cognate receptor sequences, connected into networks by repetitive sequences. Processed nuclear RNAs were posited in other models to be global regulators of gene expression, but the problem was the lack of detail about the actual information in genomes, which rendered these models, as reasonable as they were, speculative and largely overlooked.

THE AGE OF AQUARIUS

The problem of lack of detail began to be solved by the gene cloning revolution and the development of DNA sequencing in the 1970s. These technologies led an explosion in knowledge, and by the mid-1990s shotgun cloning and sequencing was being used to characterize the many mRNAs that had eluded identification by biochemical and genetic assays. A myriad of protein-coding genes was discovered in organisms from bacteria to humans, including those that regulate development, cell division, cell differentiation, cell signaling, trafficking pathways and immunological responses, among many others, as well as mutated versions in cancer. These advances, however, diverted attention from the broader questions of genome regulation and reinforced the concept of genes as protein-coding.

ALL THAT JUNK

By the 1970s, it was evident, however, that most sequences in the genomes of complex organisms are not protein-coding (Figure 1.1). The amount of cellular DNA was found to broadly increase with developmental complexity, but there were incongruities, termed the C-value enigma. Theoretical considerations of population genetics, the lethality of protein-coding mutations, the presence of large numbers of repetitive sequences and seemingly defective 'pseudogenes' all suggested that some, and perhaps most, multicellular organisms carry substantial loads of non-functional DNA.

The corollary of 'neutral' evolution of non-functional sequences was widely accepted, although there was debate between the 'near-neutralists' and 'adaptationists' concerning the signatures of protein-coding genes (and, later, regulatory sequences) underpinning quantitative trait variation. Nonetheless, there was growing consensus that much if not most of the DNA in plant and animal genomes must be junk and that the many repetitive sequences are 'selfish' genetic hobos.

The discovery in 1977 that eukaryotic genes are mosaics of short fragments of mRNA protein-coding and flanking regulatory sequences ('exons') interspersed with non-coding sequences ('introns') that are removed by post-transcriptional splicing explained heterogeneous nuclear RNA and was proffered as further evidence of junk. Introns were rationalized as the remnants of the prebiotic assembly of genes, which had been purged from microbial

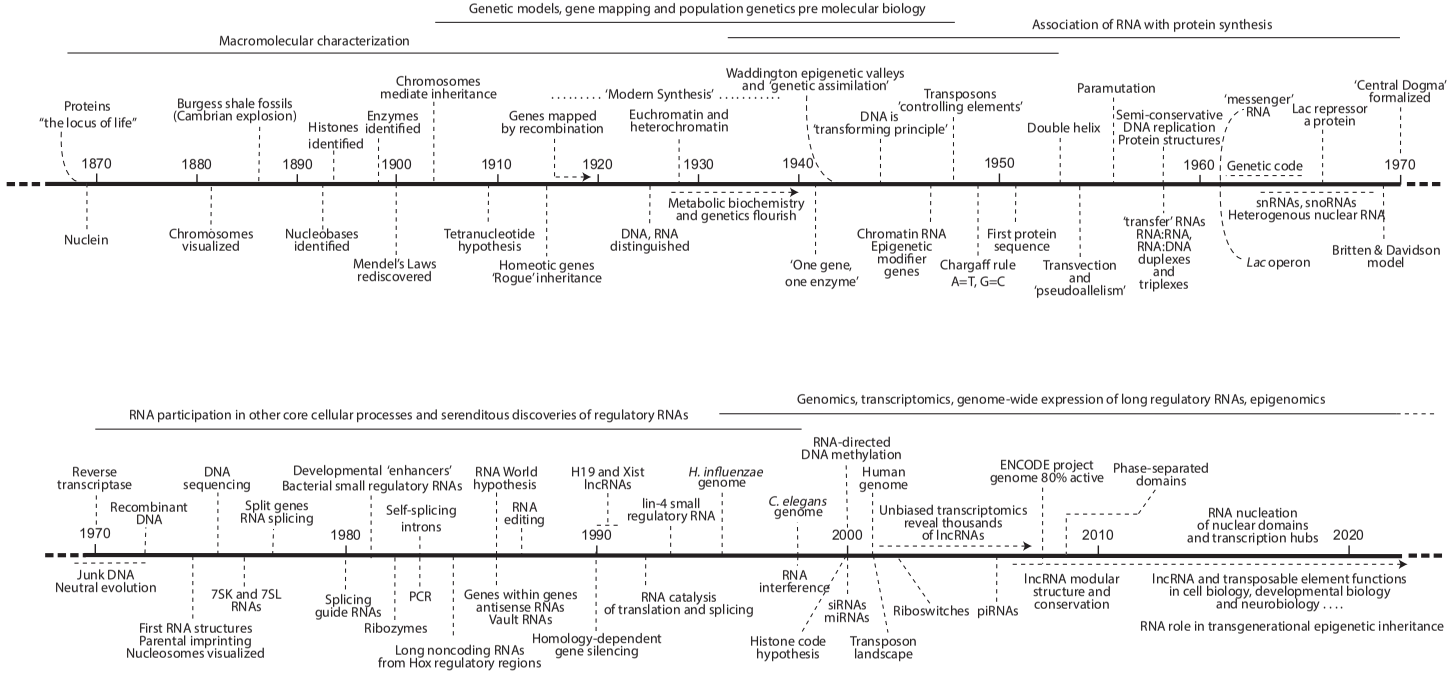

FIGURE 1.1 Historical overview.

genomes under selective pressure for rapid replication, even though the ancestors of complex organisms were also microbial. On the other hand, while small in unicellular eukaryotes, introns were found to increase in number and size with the developmental complexity of multicellular organisms, which suggested that these sequences had acquired important functions.

THE EXPANDING REPERTOIRE OF RNA

In parallel with the gene cloning revolution, the increasing sophistication of biochemical techniques identified relatively abundant RNA species beyond the canonical trio of tRNA, mRNA and rRNA. These included small nuclear RNAs (snRNAs) that guide splicing and other aspects of gene expression; small nucleolar RNAs (snoRNAs) that guide modifications of rRNAs, tRNAs and snRNAs; 7SK RNA, a negative regulator of transcription; 7SL RNA, an essential component of the 'signal recognition particle' that targets proteins to the endoplasmic reticulum and precursor of the ubiquitous Alu elements in the human genome; 'vault' RNAs of mysterious function but known to be involved in recycling of cellular components in lysosomes and neuronal synaptic plasticity; and rodent brain-specific transposon-derived RNAs that modulate behavior.

In the 1980s, RNAs were discovered to have self-splicing and cleavage activities, and that RNA catalyzes both translation and splicing, leading to the conclusion that RNA was the primordial molecule of life – the 'RNA World' hypothesis, whereby RNA subsequently outsourced its enzymatic functions to the more versatile proteins and its information functions to the more stable and easily replicable DNA. The early examples of the structural and functional capacities of RNA were, however, largely interpreted as relic infrastructural components rather than another dimension of molecular biology.

GLIMPSES OF A MODERN RNA WORLD

In the decades leading up to the turn of the century and shortly thereafter, as analytical sensitivity improved, many less abundant RNAs were identified. Small antisense RNAs ('riboregulators') and cis-acting RNA structures ('riboswitches') were found to control transcription and translation in bacteria, the latter by allosteric sensing of metabolites and environmental signals. Synthetic antisense oligonucleotides began to be used to artificially control gene expression in eukaryotic cells.

Overlapping 'antisense' transcription and 'nested' genes within genes were observed in animals and plants, hinting at intertwined genetic information and regulatory complexity. Differentially transcribed long 'untranslated' RNAs were reported to regulate ribosomal RNA transcription, and to be produced from the regulatory regions of homeotic and heat shock–induced genes in *Drosophila* and mammalian immunoglobulin class-switching, cancer-associated and parentally imprinted loci, among others.

Xist was identified as a long non-coding RNA that mediates female X-chromosome inactivation in mammals, and analogous RNAs mediating male X-chromosome activation were identified in *Drosophila*. 3' untranslated regions (3'UTRs) in mRNAs were found to be separately expressed and to transmit genetic information independently of their normally associated protein-coding sequences, and small RNAs antisense to 3′UTRs were found to control developmental timing in *C. elegans*. Although some speculated that these small and large RNAs may be the first examples of a more extensive RNA regulatory system in cell and developmental biology, they were generally regarded as oddities.

GENOME SEQUENCING AND TRANSPOSABLE ELEMENTS

By the mid-1990s, the extraordinary advances in DNA cloning, amplification and sequencing had made feasible the sequencing of whole genomes. The subsequent exponential growth of data led to progressively well-annotated genome databases and suites of computational tools for gene prediction, ortholog identification and the analysis of gene structure and expression. For the first time, the full complement of DNA sequence information in bacteria and archaea, protists, fungi, plants and animals began to be revealed, enabling comparative genomics to interrogate evolutionary relationships and functional indices at increasingly high resolution, including in complex microbial ecologies.

Prokaryote genomes were confirmed to be dominated by protein-coding genes, with phenotypic diversity achieved primarily by proteomic variation. On the other hand, animals differing by orders of magnitude in developmental complexity were unexpectedly found to have a similar number

and repertoire of protein-coding genes – only about 20,000 in both nematodes and mammals – the 'G-value enigma'.

By contrast, increased developmental complexity correlated with the extent of non-protein-coding DNA, reaching over 98% in humans and other mammals, indicating that the developmental sophistication of multicellular organisms is achieved by the expansion of regulatory information. Moreover, transposable element and retroviral-derived repetitive sequences began to be recognized as major drivers of phenotypic innovation in a wide range of plants and animals.

THE HUMAN GENOME

The first draft of the human genome sequence was published in 2001, notwithstanding the controversies that surrounded the project. The number of identified human protein-coding genes was far lower than expected by most in the field. Comparison with the mouse genome suggested that ~95% of the human genome is non-functional, based on the assumption that ancient transposon-derived sequences can be used to measure the rate of neutral evolution. On the other hand, analyses of genomic features such as transcription, sequence accessibility, DNA and histone modifications and transcription factor binding led to the conclusion that most of the human genome exhibits biochemical indices of function.

Human 'Mendelian' disorders were mapped and confirmed to be largely due to disabling mutations in protein-coding sequences. By contrast, genome-wide association studies showed that variations affecting complex traits and disorders reside mainly in non-coding regions of the genome, although an appreciable fraction of the known genetic contribution to these traits appeared unaccounted, suggesting other factors at play.

SMALL RNAs WITH MIGHTY FUNCTIONS

Genetic observations in the 1980s and 1990s indicated that RNA may play a general role in gene regulation, when it was reported that sense and antisense RNAs could modulate endogenous gene expression transcriptionally and post-transcriptionally, referred to as 'co-suppression', 'gene silencing' and (ultimately) 'RNA interference' (RNAi). The finding that introducing sense and antisense RNAs together resulted in strong systemic repression of target genes led to the dissection of the RNAi pathways, showing that double-stranded RNAs are processed to form 'small interfering RNAs' (siRNAs) that guide DNA methylation and cleavage of orthologous sequences in mRNAs.

At the turn of the 21st century, it was discovered that the RNAi pathway is used extensively to control gene expression during animal and plant development, via 'microRNAs' (miRNAs) derived from introns and other non-protein-coding transcripts. Related small RNAs, 'piRNAs', many produced from repetitive sequences, were found to be required for fertility, germ and stem cell development in animals. Other classes of small regulatory RNAs were found to be derived from tRNAs, rRNAs, snoRNAs, snRNAs, gene promoters and splice junctions, and small RNAs were shown to have many functions, including intergenerational and interspecies communication.

A similar pathway, termed 'CRISPR', was later found in bacteria to use RNA guides to target cleavage of bacteriophage genomes, manipulation of which has revolutionized genetic analysis and genetic engineering. The common feature of the RNAi and CRISPR pathways is that they use small RNAs to guide generic effector proteins to target cognate sequences in RNAs and DNAs, a highly efficient and flexible system of gene control.

LARGE RNAs WITH MANY FUNCTIONS

The high-throughput RNA profiling projects that followed the genome projects revealed the existence in both animals and plants of large numbers of low abundance long, often multi-exonic, RNAs that have little or no protein-coding potential. These 'long non-coding RNAs' (lncRNAs) were found to be expressed 'intergenically', intronically and antisense to or overlapping protein-coding genes, as well as from thousands of 'pseudogenes' and 3'UTRs. The data also showed that most of the genome of eukaryotes is transcribed in highly complex overlapping patterns, substantially from both DNA strands.

Although initially suspected to be noise, lncRNAs were found to be dynamically expressed during differentiation and development, mostly in cell-type specific patterns. LncRNAs were also found be associated with membrane-less cellular organelles, chromatin-modifying proteins and/or chromatin domains. While the genetic signatures of lncRNAs are, in the main, subtler than protein-coding genes, many have been shown to be involved in cancer and

developmental, autoimmune, neurodegenerative and neuropsychiatric disorders. Large numbers of lncRNAs – many of which are clade- or species-specific – were also discovered to have functions in cell fate determination and reprogramming, DNA damage repair, germ layer specification, hematopoietic, immunological and neuronal differentiation, retinal, skeletal, muscle and brain development, and memory and behavior, among many others.

THE EPIGENOME

It became increasingly evident during this period that the chromosomes of higher organisms are highly organized and epigenetically modified. Cytogenetic and molecular studies from the 1980s had shown the existence of chromosome territories, gene-rich and gene-poor regions and fine-scale 'topologically associated domains' with variable GC contents and non-random distributions of sequences derived from transposable elements. New genetic loci termed enhancers, hundreds of thousands of which exist in mammalian genomes, were identified and found to control plant and animal development by selective activation of protein-coding genes in their vicinity.

Nucleosomes were shown to contain canonical and specialist histones, some specific to mammalian germ and neuronal cells. The histones were found to be subject to a bewildering variety of post-translational modifications that are imposed, interpreted and erased by protein complexes that often have no intrinsic sequence specificity, including many essential for the developmental regulation of gene expression.

Histone modifications were shown to vary by gene expression and differentiation state. Exons were found to be preferentially located in nucleosomes, suggesting that epigenetic control of gene expression can be exon-specific. Vertebrate DNA was also found to be dynamically methylated during development, perturbed in cancer, and associated with gene repression. Little was known, however, of the pathways that determine the locus specificity of epigenetic modifications during development or in response to environmental influences.

THE PROGRAMMING OF DEVELOPMENT

The overarching question, rarely considered, is how much information is required to program development? The nematode worm has ~1,000 leaves in its developmental tree that are genetically hard-wired. Similarly, humans and other mammals must make trillions of divide or differentiate cell fate decisions with high accuracy, also hard-wired, as evidenced by the phenotypic congruency of monozygotic twins.

It had been widely assumed that Boolean combinatorics of transcription and other regulatory factors acting on cis-acting regulatory DNA sequences would suffice to direct developmental ontogeny, but this proposition was not rigorously justified theoretically, mathematically or mechanistically. By contrast, a decisional tree with N leaves requires an exponentially greater number of regulatory decisions, which is consistent with the quasi-quadratic increase in the number of regulatory genes with total gene number in bacteria. In all organisms, presumably, the proportion of the genome devoted to regulatory information increases with metabolic, developmental or cognitive complexity.

The fact that the genomes of plant and animals are transcribed in dynamic patterns during development and millions of different epigenetic marks are imposed at different positions in different cells across developmental stages suggests that RNA regulation has been enlisted as the most flexible and information efficient solution to the challenge of orchestrating multicellular ontogeny.

RNA RULES

Over the past two decades, RNA has been shown to regulate chromosome structure through interaction with transposon-derived sequences. DNA methylation, often differentially imposed at repetitive elements, had been known since the 1990s to be RNA-guided. Chromatin remodeling proteins, sometimes referred to as 'pioneer transcription factors', which have little or no sequence specificity and address different loci at different developmental stages, bind RNA. RNA-DNA hybrids and RNA-DNA-DNA triplexes were found to be common in eukaryotic chromatin. Histone-modifying proteins also have no intrinsic sequence specificity but some have been shown to associate with RNA 'promiscuously', i.e., bind to many different RNAs.

The largest class of sequence-specific transcription factors, containing zinc finger motifs, also

addresses target loci differentially and binds RNA as well as DNA, with many having higher affinity for RNA-DNA hybrids than for double-stranded DNA. Half of the C_2H_2 zinc finger proteins in the human genome contain KRAB domains, many primate-specific, which wire them into regulatory networks by binding cognate transposon-derived sequences.

Enhancers were found to express non-coding RNAs that are required for enhancer action, which involves chromatin 'looping' to form transcriptional hubs. Enhancers have all of the signatures of genes, except that they do not encode proteins. The number of mapped enhancers is approximately the same as the number of lncRNAs expressed from the human genome, which resolves the G-value enigma.

It was also discovered that most proteins involved in regulating gene expression in plants and animals, including transcription factors and histone modifiers, contain 'intrinsically disordered regions' (IDRs), the fraction of which increases with developmental complexity. IDRs interact with RNAs to form phase-separated condensates, which are widely deployed to organize subnuclear and cytoplasmic domains, including topologically associated transcriptional hubs in chromatin. RNA interaction with primitive proteins containing IDRs to form phase-separated domains may also comprise the third dimension of the ancestral protocell.

LncRNAs have a modular and highly alternatively spliced structure, with many domains derived from 'repetitive' elements. LncRNAs also act as scaffolds and guides for ribonucleoprotein complexes, a highly efficient and flexible system that, like RNAi and CRISPR, uses RNA signals to regulate and direct generic protein effectors to their sites of action to program development and adaptive radiation.

PLASTICITY

Over 170 different modifications of nucleotides have been identified in RNA, some important for the structure, function or stability of rRNAs, tRNAs, snRNAs and snoRNAs, as well as mRNAs and other non-coding RNAs. These modifications have also been found to be, at least in some cases, reversible and to modulate the structure-function relationships of RNAs to control processes as diverse as chromatin organization, stem cell differentiation, development, brain function, stress responses, mRNA stability and miRNA processing, among others. RNA modifications have been used to allow mRNA vaccines to evade the innate immune response.

RNA is also 'edited' by cytosine and adenosine deamination, to form uracil and inosine, respectively. Adenosine editing has expanded in vertebrate, mammalian and primate evolution, especially in the brain, and in humans occurs largely in Alu elements, which invaded the genome in three waves during primate evolution and occupy over 10% of the genome, with more than 1 million copies.

The APOBEC enzymes that deaminate cytosine to form thymine or uracil are vertebrate-specific, the first involved in somatic rearrangement and hypermutation of immunoglobulin domains. The ABOBECs have expanded under positive selection during mammalian and primate evolution, apparently to regulate transposable element and retroviral activity. Repetitive elements are mobilized in the brain, which is being shown to have many other unusual molecular dynamics associated with its ability to re-wire synaptic connections.

Transgenerational epigenetic inheritance (such as 'paramutation') was shown to involve small RNA signals and DNA methylation. Paramutation is associated with simple tandem sequence repeats (STRs), over 1 million of which are present in the human genome and are enriched in promoters of protein-coding genes and enhancers. STR variation has been linked with psychiatric disorders and cancer, as well as the modulation of physiological and neurological traits, suggesting that the extent of soft-wired inheritance of experience has been underestimated.

BEYOND THE JUNGLE OF DOGMAS

It seems that the nature of genetic information in complex organisms has been misunderstood since the inception of molecular biology, primarily because of the assumption that most genetic information is transacted by proteins. Other assumptions made during the formative years of genetics also appear to be incorrect, notably that mutations are random, and that epigenetic memory of experience is not inherited.

A transformation is taking place in the understanding of the role of RNA in evolution, inheritance, cell and developmental biology, brain function and disorders, ranging from basic science to a myriad of applications, including a new generation of RNA therapies.

Genomes contain biological software encompassing codes for components, self-assembly, differentiation and reproduction, supplemented by information transmitted by epigenetic memories. Not only has the data evolved, but also the data structures, implementation systems, evolutionary search algorithms and the interplay between hard- and soft-wired inheritance. Indeed, it is likely that evolution has learned how to learn, and that many primitive preconceptions will have to be reevaluated, with more surprises in store.

The details follow.