2001 and all that: a tale of a third science

Karola Stotz

Draft, do not cite without permission

Word count: w/o references and figures 9,620, total: 12,043

> "Last year (2001) will be identified in the history of biology by the publication of the first draft of the complete sequence of the human genome"(Collado-Vides and Hofestaedt 2002, vii) (Collado-Vides and Hofestaedt 2002, vii).
>
> "The more we lift the lid on the human genome, the more vulnerable to experience genes appear to be" (Ridley 2003).

1. Introduction

In '1953 and all that', one of the seminal papers on the relationship between classical and molecular genetics, Philip Kitcher argued that while molecular genetics has solved the major questions of replication, mutation and the action of genes these celebrated achievements do not fall into either of the traditional categories of theory reduction and explanatory extension (Kitcher 1984). The classical and molecular conceptions of the gene both remain valid. For the last 15 years C. Kenneth Waters has argued relentlessly against the antireductionist consensus that Kitcher helped create (Waters 1990, 1994, 2000, forthcoming a, forthcoming b). According to Waters the molecular gene concept "unifies our understanding of the molecular basis of a wide variety of phenomena, including the phenomena that classical genetics explains in terms of gene differences causing phenotypic differences" (Waters 1994, 163). He identifies the privileged role of the molecular gene in many biological explanations as that of an "actual difference maker" with "causal specificity" (Waters forthcoming b). I will argue that while Waters may have offered an accurate analysis of the role and status of the molecular gene concept during the classical period of molecular genetics from the 50s to the 70s, his account clearly downplays some of the major theoretical insights into genome structure or function revealed by contemporary molecular genetics and genomics, including surprising ways in which DNA performs its traditional gene-like functions, new un-gene-

like functions, and other cellular structures that may share some of DNA's cellular function. These revolutionary findings have propelled us into a new scientific era of 'postgenomic' biology. Its 'postgenomic gene' concept embodies the continuing project of understanding how genome structure supports genome function but with a deflationary picture of the gene as a structural unit and causal agent and a massively increased role for regulatory mechanisms including its environmental signaling pathways. The complex *cellular regulation* of genome expression forces us to distinguish a gene or protein's "molecular function" from its contingent "cellular function" (Marcotte 2002). It is one aim of systems biology to study the location, time and condition of these genes and proteins' expression, the networks of their interaction and the systemic context in which they operate. *The distributed control of genome expression, the extent to which it amplifies the literal coding sequence of the "reactive genome"*[1] *by providing additional sequence specificity to an underspecified DNA sequence, extends the range of "constitutive epigenesis"*[2] *all the way down to the molecular level of sequence determination.*

There is reason to believe that as much as 1953 marked the starting point of molecular genetics (Watson and Crick 1953a, 1953b), 2001 will come to signify the advent of postgenomic or systems biology (Lander et al. 2001; Venter et al. 2001). The Human Genome project, while sequencing the whole *genome*, highlighted molecular genetics' obsession of the last 50 years with identifying and annotating *protein coding* genes. In both a positive and a negative sense the final draft marked the climax of this old era, the end of the 'Century of the Gene' (Keller 2000), and the beginning of the post-genome era. On the positive side the sequence of the human genome, along with other sequenced genomes, have revolutionized molecular biology especially with respect to sequencing technology and its impact on fields like comparative genomics and molecular evolution. Biodiversity studies expanded enormously thanks to genomics, and when genome data made it clear that variation had hardly been tapped, it was genomics that supplied the technology and analyses that formed the basis of a radically new approach

[1] See Scott Gilbert's quote at the end of subsection 6.2
[2] For a more detailed description of constituent epigenesis see (Stotz 2006; Robert 2004).

(metagenomics) and perhaps ultimately a new vision about the interrelatedness of all human life (see discussions about the 'second human genome project', for example). On the negative side, in contrast to the widespread understanding of the goal of the human genome project as '*decoding'* the blueprint of life the reality turned out to be much more sobering: it merely *deciphered* the final draft of the code. Decoding may be eons away, and exactly for the reasons outlined in this paper. Despite our growing knowledge of the importance of non-coding RNA genes, both sequencing projects for practical reasons focused on annotating protein-coding genes, for which they came up with an almost "humiliating" small number of around 24,000, which is 1.5% of the entire genome (Ridley 2003). Detailed studies of partial sequences revealed the complexity involved in transcription patterns and showed us how little we know about the structure and function of genes and other genetic elements. Genomics assaulted genetic determinism, culminating in a much broader multilevel approach to molecular studies of life.

> "In sequencing the human genome, researchers have already climbed mountains and traveled a long and winding road. But we are only at the end of the beginning: ahead lies another mountain range that we will need to map out and explore as we seek to understand how all the parts revealed by the genome sequence work together to make life" (Stein 2004, 916).

So what is the postgenomic era about? Let us recall that most of the research in molecular genetics and medicine in the past two to three decades has been characterized by the efforts to identify the gene(s) responsible for a given biological function or disease. The completion of the sequencing of the human genome and at an accelerating pace of the genomes of many other organisms signifies the change of focus from the gathering and archiving of genomic data to its analysis and use in prediction and discovery. Comparing the human genome with its transcriptome reveals sequence information not encoded by the literal DNA code alone. Intra- and intercellular and even extra-organismal environmental signals impose instructional specificity on regulatory RNAs and proteins organized in expression mechanisms of mind-numbing complexity. The analysis of gene regulatory networks is exactly the sort of challenge that postgenomic tools such as bioinformatics can impact. This has undoubtedly changed the outlook of biological

research, marking the need to approach the study of living organisms with a different perspective. Rather than being satisfied with DNA sequence information, the focus has shifted towards how these sequences are used in a transient and flexible way through a network of transcriptional, co- and posttranscriptional, translational and posttranslational mechanisms of gene expression. While the molecular decades behind us were characterized by the attempt to decompose organisms into their smallest components, the postgenomic era with its systems-biological outlook marks new enterprises to reassemble these components to learn how they interact to form complex living system.

The paper will first introduce the central concept of *specificity*, the way it has changed from *conformational* specificity in the decades of classical genetics to *informational* specificity in the neo-classical era of molecular biology, and how this relates to Kenneth Waters' central thesis of *causal* specificity. This will provide the foundation for the remainder of the paper which argues in three steps (section 3 – 5) for a new, *distributed* specificity based on *combinatorial control* and how this is played out by the three 'genome' expression processes[3] that I have termed sequence *activation*, *selection* and *creation*. Section 6 will summarize this argument that I have based on a scientific survey of regulatory mechanisms of genome expression and the way they impose *sequence specificity* on an underdetermined DNA sequence. It will conclude with the introduction of a deflationary, 'postgenomic gene' concept that sets the stage for the introduction of a new postgenomic biology, outlined in section 7.

2. Sequence specificity and the co-linearity hypothesis

Just as 1953 didn't mark the end of genetic analysis used in classical genetics postgenomic biology doesn't reduce or otherwise substitute molecular genetics and genomics. The classical molecular gene concept was the product of a highly successful attempt to identify the physical basis of the 'instrumental gene'. That classical concept, however, is embedded in biological practice in ways that would be artificially and unhelpfully restricted by replacing it with the molecular concept. Marcel Weber

[3] This term does not presuppose the existence of a pre-defined boundary of any expressed sequence.

concludes in his insightful comparison of Mendelian and molecular analyses of *Drosophila* loci that "even though the classical gene concept had long been abandoned at the theoretical level, it continues to function in experimental practice up to the present" (Weber 2004, 223). What today is called 'genetical genomics' utilizes traditional genetic analysis to investigate differences in the level of expression of identical genes in different individuals, and to find the 'genes for' molecular phenotypes, such as the expression level of some transcript. This technique reveals a wide variety of distal regulatory regions, many of which are classical 'genes for' a particular phenotype but are not "nominal" genes in its molecular sense (Griffiths et al. forthcoming). The same goes with molecular genetics: as of Oct 13th 2006 the number of published genome sequences has reached 433, and counting (with more than 2000 running genome projects)[4]. The exercise of 'counting genes' continues in the post-genome era, if more critically reflected than before (Stein 2001; Snyder and Gerstein 2003; Kampa et al. 2004; Kapranov et al. 2005). For that reason Griffiths and Stotz have argued for the parallel existence of three gene concepts: the instrumental gene, the nominal gene, and the postgenomic gene (Griffiths and Stotz 2006). The instrumental gene has a critical role in the construction and interpretation of experiments in which the relationship between genotype and phenotype is explored via hybridization between organisms or directly between nucleic acid molecules. It also plays an important theoretical role in the foundations of disciplines such as quantitative and population genetics. The classical molecular gene concept, we argue, developed into two current concepts with quite different functions: Richard Burian has called the first the "nominal gene", which is grounded in well-defined sequences of nucleotides and a critical practical tool that allows stable communication between bioscientists in a wide range of fields (Burian 2004). This concept, however, does not embody major theoretical insights in genome structure or function. Most notions of 'gene' in this paper, if not otherwise specified, will refer to this concept. The second, postgenomic gene concept emerged out of the 'breakdown' of the classical molecular gene concept. We will encounter the genome expression mechanism responsible for this breakdown in section 3 to 5.

[4] GOLD (Genome OnLine Database): http://www.genomesonline.org/

Kenneth Waters' has recently made several attempts to repeat, clarify and justify a central thesis of his former analysis of the molecular gene concept. As I have argued elsewhere, his central claim is no longer suitable to capture our current knowledge of genome structure and function (Stotz 2006). Here I take issue with several of his most recent formulations of his genetic causation model phrased in terms of "causal specificity"[5]:

Thesis 1: "Only the *activated* DNA segments (the genes) are actual difference makers of RNA sequences" (Waters forthcoming b, my emphasis).

Thesis 2a: "The initial synthesis of RNA in prokaryotes and eukaryotes involves many causes, but only DNA is the *causally specific actual difference maker*" (Waters forthcoming b, my emphasis).

2b: "*Possible* exceptions involve cases of differential RNA splicing and editing. *If differential RNA splicing occurs within the same cell structure at the same time*, then differences in the linear sequences among *these* polypeptides … could be said to be caused by differences in splicing factors, rather than differences in DNA. It would still technically be true that different "split genes" were involved"[6] (Waters forthcoming a, my emphasis).

Thesis 3: "I will note that this qualifier does not need to be added for the case of genes for RNA or polypeptides in Prokaryotes or for the case of genes for unprocessed RNA in Eukaryotes" (Waters forthcoming a). "DNA is *the causally specific actual difference maker* with respect to the population of RNA molecules first synthesized in eukaryotic cells." (Waters forthcoming b, emphasis in original)

Before I can react to these theses in detail in section 4 and 5, I need to clarify their central concept of (causal) specificity.

Specificity, first for macromolecular structure and than later also for linear sequence, has been the touchstone for modern biology. It transformed our understanding of biological mechanism from a highly fluid and interactive process into an assembly of pieces each

[5] For argument's sake lets pretend that I accept his general model of causation in terms of actual difference makers.

[6] This move would depart from conventional molecular genetics, and it would mean that the pre mRNA and the RNA are specified by two different genes. Waters seems prepared to go a long way to withhold causal specificity from splicing agents.

with its own specific and restricted part to play (Greenspan 2001). The first half of the last century was characterized by the concept of chemical or *conformational* specificity, namely the ability of an enzyme's binding site to recognize the chemical structure of its specific ligands. The fewer substrates a protein can bind, the greater its specificity. Quantum mechanics provided the necessary insight to explain the idea of structural complementarity or a 'key and lock system' of recognition in terms of the stereospecificity of enzyme and substrate to form a certain number of weak hydrogen bonds. Molecular biology replaced this concept of specificity based on the idea of 'form' with a new concept of genetic specificity of nucleic acid based on the idea of 'information' encoded in the sequence of nucleotides. This new sequence or "colinearity hypothesis" Francis Crick laid down in the central dogma of molecular biology in 1958 and restated in *Nature* in 1970: "The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid" (Crick 1970, 561; 1958; Sarabhai et al. 1964).

Waters' thesis of causal specificity is basically restating Crick's dogmatic sequence or colinearity hypothesis in causal language and with a slight modification (namely of reluctantly accepting splicing agents to share this specificity in certain cases). In the light of the developments of the last 35 years (see below) this appears as an attempt to 'rescue' DNA as the (more or less) sole bearer of causal specificity in order to a) justify "why so much research attention in developmental biology is centered on DNA", and b) to "reveal the fallacy of causal parity arguments" (Waters forthcoming b). Al Hershey, who's experimental results finally convinced the majority that genes are made of nucleic acid instead of proteins, has once said: "Influential ideas are always simple. Since natural phenomena need not be simple, we master them, if at all, by formulating simple ideas and exploring their limitations" (cited in: Ptashne and Gann 2002, 59). So let us accept Waters' (and Crick's earlier) claims as necessarily simple ideas to master a complex reality and take my following arguments as exploring their limitations. Even if we restrict ourselves to the investigation of sequence specificity of gene products we see that the organism's molecular complexity is not specified by its limited number of protein coding

genes but by what it can do with its genome. I will prove this point with *detailed examples of how nucleotide sequences are activated, selected and created by causally specific regulatory mechanisms of genome expression*. My goal is not to understand the full complexity involved in the regulation of genome expression, much less so the biological mechanisms beyond the primary sequence of gene products. *This paper has the limited agenda of giving examples of sequence modifying processes to understand which agents other than genes carry causal or sequence specificity.*

3. Sequence is not destiny: Distributed causal specificity

The view of conformational and informational specificity as selective and exclusive has recently given way to a picture of both conformational and causal specificity as being highly distributed, modular and combinatorial. Through the expansive possibilities of combinatorial associations more sophisticated cellular functions can be achieved and fine-tuned by increasing the number of interactions any specifying agent can engage in by virtue of its intrinsic molecular function. Even in prokaryotes, and to a much larger extent in eukaryotes, the regulation of gene expression works by means of the *regulated recruitment* of trans-acting factors (proteins, RNAs and metabolites) into larger complexes and to cis-acting sequence modules, so that *the specificity of an enzyme, a sequence, transcription or splicing factor comes to depend on its proper recruitment and combinatorial interaction* (Ptashne and Gann 2002; Buchler et al. 2003). Hence a gene product is specified by a genomic template and the differential recruitment of agents of genome expression mechanisms that activate and alter the transcript specifically. This versatility of the genome by means of the combinatorial complexity of its regulation resolves the 'N-value' paradox (Claverie 2001; Harrison et al. 2002): The proportion of protein-coding sequences indeed seems to decline as a function of complexity, but the ratio of non-coding DNA rises, and so does the number of functional, regulatory roles played by non-coding DNA and RNA and other cellular factors that help to translate sequential information encoded in the genome into developmental complexity (Mattick 2004).

> "There is increasing awareness that multiple, often overlapping mechanisms exist for amplifying the repertoire of protein products specified through the mammalian genome. An expanding array of processing and targeting mechanisms is now emerging, each representing a potentially important restriction point in the regulation of eukaryotic gene expression, and each expanding the possibilities specified by the literal code of the genome. These co- and posttranscriptional regulatory events include capping, alternative splicing, differential polyadenylation, RNA editing, nuclear export, alternative decay and degradation pathways, as well as alterations in ribosomal loading or translation" (Davidson 2002).

Latest estimates place the number of alternatively spliced human genes to over 70%, with close to 100 genes with over 5,000 splice variants, and the Drosophila record of up to 38,000 temporally and spatially regulated splice variants derived from the Drosophila cell adhesion molecule gene (*DSCAM*), and (Kapranov et al. 2005; Celottoa and Graveley 2001; Graveley 2005; Rowen et al. 2002; Leipzig et al. 2004). In some mitochondria of higher plants a total of more than 1000 C-to-U changes are known to alter the total coding text of the entire RNA population, mostly within the first two positions of codons, hence changing the amino acid. The RNA editing of cellular RNAs of many eukaryotic organisms can result in up to 50% modified adenosine residues in a transcript (Gott and Emeson 2000). This form of editing is absolutely critical for normal brain function in humans and very prevalent in mammalian cells with a suspected 85% of all mRNA as a target (Athanasiadis et al. 2004). A recent study of the architecture of the human transcriptome paints "a picture of a highly overlapping, complex, and dynamic nature of the human transcriptome, where one base pair can be part of many transcripts emanating from both strands of the genome. The data further suggest that base pairs normally thought to contribute to transcripts from different genes can be joined together in a single RNA molecule" (Kapranov et al. 2005; see also: Kampa et al. 2004; Cheng et al. 2005). New metaphors to capture this emerging picture have been suggested: "the sociable gene" conveys the interactivity, fluidity and dynamics of the genomic system; to describe the control of genome transcription Lenny Moss has coined the term of "ad hoc committees" of molecules convened on the basis of the history of the cell and its

interaction with the environment (Turney 2005; Moss 2003). There is so much more to our genome than its officially annotated number of 'genes' suggests, and in 2003 the National Human Genome Research Institute (NHGRI) with launching "ENCODE" has recognized this. The ENCyclopedia Of DNA Elements Project aims to comprehensively identify all the functional elements in the human genome, including transcriptional and other regulatory elements, gene and exon variants, alternative promoters in tissue-specific gene expression, chromatin-mapping sites, and conserved non-coding elements. The factors that interactively regulate genomic expression are far from mere background conditions or supportive environments; rather they are on a par with coding information since they *co-specify* the linear sequence of the gene product together with the target DNA sequence. From this follows the radical thesis of 'molecular epigenesis': *Networks of genome regulation made up of cis-regulatory sequences, trans-acting factors and environmental signals causally specify the physical structure of a gene and the range of its products through the activation, the selective use, and, more radically, the creation of nucleotide sequence information* (Stotz 2006).

4. The flexible genome: sequence selection

Transcriptional regulation:

(Thesis 1): To restate Waters' first thesis, he singles out "activated" DNA as the causally specific agent responsible for the composition of a population of RNAs in a cell. The default position of eukaryotic DNA is inactivation, and Waters deliberately neglects and downplays all the processes that are involved to *activate* DNA as causal agents. Second, he forgets to clarify between which two states DNA should function as the actual *difference* maker: it could be the difference in the linear sequence between any two gene products, or the difference between two populations of RNAs in two cells of an organism. The second problem is what is commonly called the foremost 'problem of development': the differentiation of cells from a single cell in multicellular organisms. The peculiarity of the differentiated cells is that despite their immense differences they all share the *same* DNA (with some notable exception as immune cells). Hence the actual *difference* between two cells is not their DNA but activating agents such as specific transcription

factors and inducing signals that co-differ between two cells. They orchestrate the tissue-dependent and time-specific *activation* and sequence *selection* of a subset of 'genes' that translates into different cellular phenotypes. The phenotypic difference between two daughter cells could result from the expression of *different* genes (with different causal specificity) or the time-, tissue- and combination-dependent expression of *common* genes (with the same causal specificity). *Activation* of DNA is therefore a causally specifying mechanism by determining a particular RNA product to be there. In addition, since activation *selects* between different promoters, and is likely to influence co-transcriptional activities such as splicing and editing, activation is causally specifying the particular sequence of a RNA product from the same DNA sequence through sequence *selection* and *creation*.

(Thesis 2a): Waters main thesis states the exclusivity of DNA in providing causal sequence specificity: (with some notable exceptions) only DNA provides the linear sequence specificity of any gene product. So while we are agreeing in principle that DNA alone is not the sole source of sequence specificity, I believe my argument presents a *radical* shift in focus from (molecular) genetic to (systems biological) distributed sequence specificity. Against Waters' almost *exclusive* notion of specificity in this and the following section I set a picture of *distributed* causal specificity, where already *pre-selected and activated* DNA shares the stage with the RNA processing machineries of splicing, editing, modification and translational recoding that further *select*, *modify* and *newly create* DNA and RNA sequences. All gene expression mechanisms such as activation or inhibition, splicing, editing and other co- and posttranscriptional processes have in common the combinatorial interaction of multiple and variable *cis*-acting sequence modules both upstream, downstream and within coding sequences that through their primary sequence or secondary structure formation bind a large range of diverse *trans*-acting factors such as proteins and RNAs. These factors either need to be transported, recruited or induced by intra- or extra-cellular signals and by these means *function as mediators of environmental information to the genome*. Most *cis*- and *trans*-acting elements have in common that they are individually weak, variable, and present in multiple copies. As we now know, "there is little or no *constitutive* regulation in higher

organisms; i.e., the differentiated state of normal cells is unstable and the environment regulates gene expression" (Bissell 1981, 27; quoted in Bissell 2003). Because of the structure of their complex, modular, but weak promoter sequences gene expression in eukaryotes *always* requires the recruitment of a large transcriptional machinery of *trans*-acting factors to the *cis*-acting modules through activators (another kind of trans-acting factors that bind to very distant *cis*-acting sequences, the enhancers). The exact order and nature of their recruitment is still largely unknown, we know, however, that the full machinery comes in the form of separate complexes often made up by a large number of proteins (see Figure 1): the activator complex assembles at the enhancer (enhanceosome) to recruit the chromatin remodeling complex (to make the DNA accessible) and the TATA-binding proteins and associated factors, which bind to the TATA site of the promoter to recruit the transcription enzyme polymerase, specific transcription factors and transcription cofactors. *It is the specific recruitment of transcription factors to varying complexes by trans-acting factors (proteins, RNA and environmental factors) that imposes their specificity*.

Through the differential use of alternative promoters, transcription start sites and transcription termination sites activation can at the same time be the *pre-selection of the actual sequence of the gene*. The differential use of promoters can also specify splice site selection when alternative exons can come with their own promoter (Dorn et al. 2001; Tasic et al. 2002). Some of the components of the transcription machinery will subsequently move along with the polymerase during the transcriptional process and may interact with the capping, splicing, polyadenylation and editing machinery (Ptashne and Gann 2002; Davidson 2001). Cells constantly response to intra- and extracellular signals such as a hormonal or nutritional changes with a change in gene expression *mediated through the environment-specific use of regulatory elements* (Ptashne and Gann 2002). In other words *RNA and proteins relay environmental information to the genome*. A common mechanism of induction is the phosphorylation of transcriptional regulators that changes their conformational specificity. *Specificity is imposed by environmental induction of activators, differential recruitment and combinatorial control*.

Insert figure 1 about here

One hallmark of postgenomic biology is the recognition of an extension of the possibilities specified by the literal code of the genome through the extensive amplification of the repertoire of RNA and protein products. This change in focus is mirrored by the terms 'histone code' and 'cellular code'[7], coined for co-specifying the *activation*, *selection* and *modification* of DNA through transcriptional, co- and post-transcriptional processes. These mechanisms of causal specificity introduce "a new element of difficulty in understanding the transmission of information from the DNA code to the functioning organism" (Eisenberg et al. 2006). The following section highlights agents involved in the final *selection* of DNA sequences especially through alternative splicing. Together with the mechanisms of activation and pre-selection they present 'conservative' cases of shared causal specificity because leave intact, however split up, *the linear order* of the original sequence in the gene product.

Alternative Splicing:

In eukaryotes, the DNA sequence is transcribed into a pre-messenger RNA from which the final RNA transcript is processed by cutting out large non-coding sequences, called *introns*, and splicing together the remaining, mostly but not always coding sequences, called *exons*. Biologists speak of alternative *cis*-splicing when more than one mature mRNA transcript results from these processes through the cutting and splicing of alternative exons. This is a very prevalent mechanism in complex organisms that affects quantitative and qualitative control of gene expression and the generation of protein diversity from a single DNA precursor. *Agents other than the original coding sequence have to provide sufficien tspice-site specificity controlling this diversification.*

Pre-mRNA splicing is the process by which two successive transesterification reactions

[7] The histone code is the sum of all chromatin- and DNA-modifying chemical complexes that imprint DNA sequences by rendering them inaccessible to the transcriptional machinery. The cellular code is the combination and their potential of combinatorial interaction of all available transcription factors, functional RNAs and inducing agents in a cell at one point in time. The term 'code' refers to the combinatorics of all these factors as conveying the 'meaning' to the gene expression machinery.

cleave the upstream exon from the intron and ligate it to the downstream exon. This takes place on the spliceosome, a dynamic complex of small nuclear RNAs/proteins and extrinsic SR protein factors assembled on the juxtaposed 5' and 3' splice sites. The splice site sequences are generally small and weak and not sufficient to specify splicing. Splicing specificity is imposed by the assistance of additional *cis*-acting elements in either of the adjoining introns or the exon itself and by *trans*-acting factors binding to them. While exonic and intronic splicing enhancers (ESE and ISE) positively stimulate the spliceosome assembly at certain sites, exonic and intronic splicing silencers (ESS and ISS) block certain splicing choices (see figure 2) (Smith and Valcarcel 2000). In other words, the *availability of certain trans-acting factors and the differential and combinatorial binding of spliceosomal binding RNAs and proteins to splice sites and regulatory sequences (the 'cellular splice code') seems to be the major contributor to splicing specificity*.

Insert Figure 2 about here

Occasionally the formation of a double-stranded RNA secondary structure can add specificity, similar to certain editing mechanisms. A mechanism called "variable window binding" between sequences in an intron separating a canonical exon from its mutually exclusive downstream exons with their partially overlapping and complimentary sequences seems responsible for the *mutually exclusive* splicing of alternative downstream exons to a canonical upstream exon in the Drosophila *Dscam* gene, which contains four clusters of a large number of mutually exclusive alternative exons (Anastassiou et al. 2006; Graveley 2005). A recent discovery has shown the provision of splicing specificity by RNA modification via a small nucleolar (sno) RNA (Kishore and Stamm 2006). This and similar examples show the interdependency of splicing, RNA modification guided by snoRNA and RNA editing (Bachellerie et al. 2002; Flomen et al. 2004). Three major mechanisms are known that change the "cellular code" for splice site selection: the de novo synthesis of splicing proteins, the activation of these proteins through phosphorylation, and a change in localization of splicing regulatory proteins from the nucleus into the cytosol or into stress-induced nuclear bodies (Stamm 2002;

Shin and Manley 2004). "The combinatorial mechanism for the control of alternative splicing … could allow cells to adjust splicing outcome (and consequently which proteins they express) rapidly in response to intracellular or extracellular cues, as well as contributing to the generation of protein diversity" (Bradbury 2005). In other words, *the cellular context imposes splice site specificity*.

(Thesis 2b): Under certain, restrictive conditions Waters is willing to extend causal specificity to splicing and editing agents, namely when different splice variants exist in the same cell at the same time; this is not credited when each cell produces its own splice variants, which would render the regulatory machinery as background condition. For the argument's sake, I interpret Waters to reason as follows: *From an observer's viewpoint*, in certain cellular conditions a gene is *always* specifying a particular splice variant, hence it holds the causal specificity. However, *from the viewpoint of the DNA sequence or the entire cell*, the relevant splicing and editing mechanisms are the providers of sufficient sequence specificity for the right product. In reality, however, most cells just differ in their *ratios* of a particular splice variant: "for most alternatively spliced transcripts there is no 'default' or unregulated state; instead, the ratio of alternative splice forms observed for a given pre-mRNA results from a balance between positive and negative regulation" (Ladd and Cooper 2002, 3; e.g. Celottoa and Graveley 2001; Athanasiadis et al. 2004).

(Thesis 3): Waters names prokaryotic gene expression and the specification of pre-mRNA as the clearest case for his (preferred but limited) exclusive DNA causal specificity thesis. However, not even in prokaryotes or in the production of *preliminary* mRNAs in eukaryotic cells does the DNA sequence exclusively specify the products. We now know that there exist RNA editing and modifying mechanisms in bacteria, and that transcription in eukaryotes is being carried out by what has come to be known as the cotranscriptional machinery or mRNA assembly line. This means that there is indeed no time at which a fully sequenced pre mRNA exists in the cell.

The Cotranscriptional machinery

Although all mechanisms of DNA expression and regulation have their biochemical identity, all of them feature in an "extensive network of coupling among gene expression machines". It is now clear that alternative splicing does not represent a distinct and decoupled step but is tightly coupled to transcription, polyadenylation, RNA editing, RNA surveillance and transport. "Recent studies suggest that this task is facilitated by a combination of protein–RNA and protein–protein interactions within a 'mRNA factory' that comprises the elongating RNA polymerase and associated processing factors. This 'factory' undergoes dynamic changes in composition as it traverses a gene and provides the setting for regulatory interactions that couple processing to transcriptional elongation and termination" (Bentley 2005). Polymerase II and many other transcriptionnal proteins cooperate with the cotranscriptional processing factors. For instance, some SR proteins involved in the spliceosome have been known to react with transcription factors, while other proteins even exhibit a dual function as transcription and splicing regulator (Maniatis and Reed 2002; Bentley 2002). The cotranscriptional assembly of the spliceosome in this 'mRNA assembly line' suggests profound implications for the regulation of splice site choice. Splicing has also been implicated in downstream processes such as RNA transport, stability, translation, location (Black 2003, 323). In addition, important links between RNA editing and other co- and posttranscriptional events that regulate gene expression have been suggested (Davidson 2002). These co-transcriptional agents (*cis*-regulatory elements, *trans*-acting factors, intra- and extracellular signals) *in combinatorial interplay with each other share causal specificity with genomic coding sequences in the production of gene products through their involvement in sequence selection (e.g. splice-site specificity) and sequence creation (e.g. editing-site specificity).*

Other sequence selection mechanisms

Beside the 'normal' splice variants the genome produces a large variety of transcripts that are even harder to attribute to a single nominal gene.[8] Many transcripts contains exons from adjacent genes and even pseudogenes that are 'co-transcribed' to produce a *single*

[8] For a more extensive list of such cases see (Stotz 2006; Griffiths and Stotz 2006) and Griffiths and Stotz' "Representing Genes" website: www.representinggenes.org

pre-mRNA (Communi et al. 2001; Kapranov et al. 2005; Finta and Zaphiropoulos 2000b, 2002). Such cotranscription may be produced by the insufficient termination efficiency of polymerase II or by a process that in prokaryotes has been called antitermination. In the latter case a regulatory protein induces changes in the termination properties of the polymerase. Many pseudogenes are processed, and while often we don't know their function, in some cases their mRNA seem to exert a stabilizing effect on the transcript of their homologous, functional gene (Hirotsune et al. 2003). Alternative gene products may be derived from so-called 'overlapping genes' including transcripts from the antisense strand, or read in an alternative reading frame (Blumenthal et al. 2002; Coelho et al. 2002). Instead of receiving mutually exclusive alternative transcripts from the same DNA sequence, as is the case with all alternative splicing and many overlapping phenomena, multiple simultaneous transcripts can occur, as is the case of the parallel processing of functional non-coding RNAs (such as microRNAs and snoRNAs) from the intronic regions of the transcript. These RNAs may be involved in the regulation of the coding transcript of the same gene, but need not be. *In all of the above instances the selective use of nucleotide sequences through a range of transcriptional, co- and post-transcriptional mechanisms co-specify the linear sequence of the final product.*

5. The versatile genome: sequence creation

In the following 'radical' cases of sequence specificity the *linear sequence* of the final product is not mirrored by the DNA sequence but is extensively *scrambled*, *modified* or literally *created* through a variety of co- and post-transcriptional processes, which often are interdependent with mechanisms of sequence activation and selection. All of the following cases are even stronger counterarguments to Waters main thesis (2b) of exclusive DNA sequence specificity than any of the 'conservative' cases provided above.

Trans-splicing

Biologists speak of *trans*-splicing when a final mRNA transcript is processed from two or more independently transcribed pre-mRNAs.[9] These separate pre-mRNAs can be derived

[9] For examples and diagrams of trans-splicing cases see www.representinggenes.org

from different DNA sequences or from multiple copies of transcripts from the very same sequence. The latter case allows the inclusion of multiple copies of the same exons or to *scramble the original order of exons* in the final transcript (Finta and Zaphiropoulos 2000a). In other words, *trans-splicing is changing the linear order of the original DNA sequence in the gene product, and co-linearity is the hallmark of Crick's and Water's sequence specificity*. Alternative exons can feature their own promoter that specify their individual selection while their inclusion in the final transcript must involve *trans*-splicing (Pirrotta 2002). Different *trans*-splicing pathways exist in nuclei and organelles, where they mostly resemble their cis-splicing counter parts. Mechanisms for splicing in trans are *supported by splicing agents that seem to be split versions of their equivalent cis-acting agents* (Caudevilla et al. 2001). Split introns can assemble a split spliceosome complex to provide specificity for *trans*-splicing in the nucleus and for spliced leader *trans*-splicing in kinetoplastid, while split self-splicing group II introns support *trans*-splicing in plant organellar genomes (Rivier et al. 2001; Sturm and Campbell 1999; Malek and Knoop 1998; Wissinger et al. 1991). So while we still don't know the exact inducing agents for the specific recognition of the autonomous pre-mRNAs, finding related mechanisms to *cis*-splicing should not surprise; in genes with very long introns splicing specificity happens almost in *trans*. Hence as far as we know *similar agents and mechanisms to alternative cis-splicing provide specificity to trans-splicing*.

RNA editing

RNA editing is another and very prevalent[10] mechanism of sequence modification that can significantly diversify the transcriptome or proteome (the total complement of final transcripts or proteins in the cells of an organism). Whereas most other forms of co-transcriptional modification of mRNA (capping, polyadenilation and *cis*-splicing) can be said to retain the *correspondence* of coding sequence and gene product (even though certain coding and noncoding regions have been cut out), RNA editing disturbs this correspondence, in some cases to a very large extent. But while trans-splicing did so by

[10] See (Gott and Emeson 2000) for a very good overview.

scrambling the order of the primary DNA sequence, editing *changes* the primary sequence of mRNA during or after its transcription via the site-specific *insertion* or *deletion* or *substitution* of nucleotides (cytidine-to-uridine and adenosine-to-inosine deamination, uridine-to-cytidine transamination) (Gray 2003). This creation of 'cryptogenes' affects most kinds of RNA (mRNAs, tRNAs, rRNAs, and 7 SLRNA) and can potentially have radical effects on the final product. U insertion or C-to-U conversions can lead to the creation of new translation start and stop codons (e.g. trypanosomatid protozoa, plant organelles, humans), while U-to-C changes can remove them (eg. Plants). Editing events within coding sequences reach from widespread nucleotide insertions ('pan-editing' in kinetoplasts and Physarum mitochondria, where over 50% of the final mRNA can be the product of editing) to singular amino acid substitutions due to C-to-U, U-to-C, and A-to-I changes, enlarging the number of protein isoforms created from a transcript. Other consequences include frameshifting between alternative ORFs (paramyxoviruses), alterations in splice sites by A-to-I conversion in mammals, and other alterations within introns and 5' and 3' untranslated regions (UTRs) potentially affecting mRNA stability, transport, translatability, and processing. *This phenomenon provides a potential break in the central dogma according to which coding information must be template derived, and in many cases, as seen in the human brain, the editing-derived coding information is essential for the normal functioning of the organism.*

Just a few of the myriad of different editing mechanisms are explained below that focus on agents that provide editing specificity. The main specificity-providing agent involved in nucleotide insertion/deletion in kinetoplasts are *trans*-acting guideRNAs (gRNAs) that bind to a complementary 'anchor sequence' just downstream of the editing sites (see figure 4). This anchor duplex directs the endonucleolytic cleavage event that initiates the editing cycle just upstream of the duplex. gRNAs act as templates to encode insertion or deletions of uridines into the nascent mRNA transcript due to their incomplete complementarity to the nascent mRNA strand. Bulges in the gRNA indicate insertion sites for the mRNA, while bulges in the mRNA get deleted. gRNAs may fold into a secondary structure to bind *trans*-acting factors such as a ribonucleoprotein (gRNP)

complex. Highly diverse mechanisms are utilized to accomplish the same end in different organisms. Insertional and deletional editing can occur either cotranscriptionally or posttranscriptionally with varying editing pathways. In cases where editing is tightly connected to transcription RNA polymerase could be involved in the recognition of editing sites.

Insert Figure 3 about here

A-to-I editing of cellular RNAs of many eukaryotic organisms modifies the adenosine residues to inosine, which will be translated into guanine. While the total number of known genes is still rather small (e.g. the serotonin receptor 5-HT_{2C}), recent estimates speculate that one in a thousand nucleotides in human brains are edited. Editing requires a partially base-paired RNA foldback structure, often provided by a pair of inverted Alu repeat sequences (see figure 5). These base-pairing events, however, don't provide the template for editing events as in the case of gRNA but mainly provide a substrate for the editing enzyme ADAR (adenosine deaminase that acts on RNA) (ADAR). The basis for the observed editing selectivity is still poorly understood, but it seems that mainly A-to-C mismatches (leading to G-to-C matches) and A-to-U matches (leading to G-to-U mismatches) are targeted for editing, partially guided by sequence biases within neighboring bases (Athanasiadis et al. 2004). As said before, one function of editing can be to create alternative splice sites. *ADAR2,* for example, produces four different splice variants of the mammalian editing enzyme ADAR2. One particular splicing event that creates a short ADAR2 protein with *no editing activity* relies on the gene's own product, ADAR2, to edit this mRNA. In other words, ADAR2 controls its own level of expression through a negative feedback mechanism. When ADAR2 levels get too high, it edits its own RNA to shut down its expression (Rueter et al. 1999). *As is the case with splicing, editing specificity seems to be distributed between cis- and trans-acting factors as well cell-signaling factors inducing certain trans-acting agents.*

Insert Figure 4 about here

Another common mechanism able to disrupt the colinearity between DNA sequence and final product is the nonstandard *translational recoding* of mRNA. The three different ways through which the translational machinery is able to *recode* the message are frameshifting, programmed slippage or bypassing, and codon redefinition (Baranov et al. 2003). *The details of transcriptional activation, alternative splicing, trans-splicing, RNA editing, and translational recoding are meant to show that the specifying relationship between DNA and gene product is indirect, mediated and specifically intervened by other sequence specifying agents (table 1).*

Insert Table 1 here

6. By way of conclusion: The postgenomic 'gene' concept

There are several general conclusions to be drawn from the mechanisms of sequence activation, selection and creation outlined in the last two sections:

The Reactive Genome

1. The causal specificity for the linear sequence of a final gene product, including pre-mRNAs, is distributed between the local DNA sequence, *cis*-acting sequences, *trans*-acting regulators, environmental signaling factors, and the contingent history of the cell (the cellular code) (see Table1 for an overview). In certain extreme cases the guide RNAs (or their DNA sequences) together provide the template ('gene'?) for the final product rather than the nominal gene, but nobody would commonly call the guide rDNA the gene. Many if not most agents involved in the regulation of gene expression of higher organisms not only must work in interaction with other agents in order to achieve full specificity, which is imposed by regulated recruitment and combinatorial control. The modular organization of genes, cis-regulatory sequences and trans-acting factors into actively distinct subunits (DNA binding sites, protein-protein and protein-RNA recognition sites, and catalytically active sites) is actively supporting this distributed and combinatorial specificity. Many people have argued that greater complexity is achieved

not by the addition of exclusively specific agents but by increased regulation, interaction, integration and the combination of structural-analog with informational-digital specificity.

2. One might concede that while any *local* DNA sequence ('gene') may indeed only partially specify its product, the genomic sequence as a whole via *cis*-acting sequences and *trans*-acting genome products is indeed sufficient for full causal specificity. Hence sequence specificity and causal agency remains with the genome. Since higher eukaryotes have no default transcriptional activation or splicing pattern, since information is provided by *difference-making* RNA and protein factors which in turn need to be recruited or turned on by external factors. Certain RNAs and proteins undergo crucial changes in shape in response to signals, which render them active and impose their causal specificity (Ptashne and Gann 2002, 6-7). *By this means they relay difference-making environmental information to the genome*. While many genetic accounts describe the environment as merely permissive, in many cases of gene expression the environment actually provides instructional specificity, including for gene products.

> "Organisms have evolved [a reactive genome] to let environmental factors play major roles in phenotype determination. […] In instructive interactions, a signal from the inducer initiates new patterns of gene expression in responding cells. […] It is usually assumed that the developing organism's environment constitutes a necessary permissive set of factors, whereas the genome provides the specificity of the interaction. In phenotypic plasticity, however, the genome is permissive and the environment is instructive" (Gilbert 2003, 92).

3. Waters' focus on the specificity of coding sequences shares the bias of the last 50 years of genetic research in its focus on (protein) coding genes while it neglects our growing understanding that the complexity of higher organisms lies not in its number of genes but within the flexibility, versatility and reactivity of its whole genome. Complexity is not encoded in the literal sequence of coding genes but in the processes that can amplify this information. These regulatory mechanisms involve among other

agents a large number of different non-coding RNAs and non-coding DNA sequences with important binding or structural domains, and even transcriptional capacity, for the longest time dismissed as 'junk' (Levine and Tjian 2003; Buchler et al. 2003). "We continue to learn new ways in which nature has exploited the specificity of interactions between RNA and nucleotide sequences. We now know that RNA, after being transcribed from DNA, can feed back to direct modifications of the genome. These modifications can be inherited through cell divisions and influence development" (Kawasaki and Taira 2004). It may be that the "explosion in complexity in virtual all systems occurred as a result of advanced controls and embedded networking, most of which is invisible to the observer". Some people now believe that we are at the brink of a "digital revolution" of an RNA – controlled parallel system of regulatory control (Mattick 2004, 320).

From Pathways to Networks:

4. Waters' analysis is built on an outdated model of *pathway* analysis that is in danger of overstating the importance of single nodes and the linear sequence of events. "Feedback loops and back-up *pathways* have been invoked to account for these properties. […] A more flexible and fluid view of the relationships among these signaling and regulatory systems allows for the same net result *without invoking a predetermined mechanism for it*. The malleability and versatility of gene *networks* and their ability to find new solutions when constituents are changes, help to account for the properties of robustness, buffering and emergence" (Greenspan 2001, 386, my emphasis). This point restates some newer criticism against the central dogma not as necessarily literally wrong but as misdirecting research into dogmatic pathway analyses and away from systems thinking (Werner 2005).

5. The Central Dogma has been enormously influential in molecular biology in the last 50 years. Several have claimed to prove it wrong, reverse transcriptase and prions being the main contenders. The flow back from RNA, however, was never explicitly excluded and prions don't confer *sequence* specificity. The shown phenomena, however, should provide a serious blow to the central dogma according to which sequence information or

causal specificity should be template-derived. This dogma now turns out to give a very limited and in its exclusity wrong account of the origin of sequence information. We don't find ourselves with a *reverse flow, but with a range of alternative sources of sequence information*, *some of which are derived from secondary products and environmental factors.*

Parity between similar functional roles

6. Last but not least, Waters misunderstands the principle of 'causal parity', which "derives its name from Oyama's earlier call for 'parity of reasoning' when thinking about the roles of DNA elements and other developmental resources. She argued that if one of the above distinctions applies to some but not all DNA elements and also applies to some non-DNA influences in development, we should treat both the DNA and the non-DNA factors alike in the area of theory where the distinction is useful. In order to be able to follow this principle of parity it is essential not to build grand, metaphysical distinctions, like that between form and matter or information and matter, on top of the many empirical differences between the roles of DNA elements and the roles of other causal factors in development … DNA does play a distinctive set of roles in development, but it does not play just one role (partly because DNA elements are themselves so diverse) and the important roles of those various DNA elements are sometimes played by non-DNA factors in development" (Griffiths and Gray 2003, 421).

When distinguishing different causal processes (italized below) in an organism, there are always more than one agent fitting this causal role: *Sequence specificity* is held by DNA, but also by splicing and editing agents, as well as other regulatory mechanisms that are involved in modifying the primary sequence of RNA. *Enzymatic activity* has for the longest time been attributed to proteins alone but is, as we now know, regularly achieved by tertiary RNA structures (ribozymes). Protein transcription factors now have to share their fame with regulatory non-coding RNAs and inducing environmental factors such as lactose in the *regulation of genome expression*; and unrelated to this topic but important nonetheless to the very idea of parity: though by different means which can be

distinguished if useful, organisms *inherit* a good deal more beside their DNA. Alone at the molecular level there is the histone code, structural components of the cell such as organelles and membranes, maternal RNA and transcription factors. And at higher levels of organization we have whatever else is provided by the parental generation in form of an 'ontogenetic niche' for the zygote, fetus, and born individual in a providing environment.

Focusing on the cutting-edge of contemporary genomics can induce an extremely deflationary, postgenomic view of the gene. As Falk suggested already 20 years ago: "Today the gene is … neither discrete … nor continuous …, nor does it have a constant location …, nor a clearcut function …, not even constant sequences … nor definite borderlines" (Falk 1986, 169). Some molecular biologists, realizing that the concepts of 'gene' transcription or 'gene' expression may not suffice to capture the complex architecture of the transcriptome of many eukaryotes have proposed the more general term of "genome transcription" to allow for the incorporation of RNA transcripts that contain sequences outside the border of canonical genes. From this new perspective the classical molecular conception of genes seems like "statistical peaks within a wider pattern of genome expression" (Finta and Zaphiropoulos 2001, 160). Recent investigation of the complexities of the human transcriptome supports these views (Kapranov et al. 2005). If correct, these results have important implications for the definition of a gene, and for the relationship between genotype and phenotype (Griffiths and Stotz 2006, in press). In contemporary postgenomic bioscience genes are hard to define as a straightforwardly structural entities, but a purely functional definition as suggested by Snyder and Gerstein would also run counter to some of our longstanding practices (e.g. to allow for alternative splice forms without increasing the number of genes[11]) (Snyder and Gerstein 2003). Or we just throw overboard these old stereotypes and the long-held ideal that a gene definition must combine functional and structural criteria, and say that genes

[11] Celera emphasizes the importance of alternative splicing in their definition of a gene as "a locus of cotranscribed exons". Ensembl's GeneSweepstake Web page goes into the same direction with their definition of a gene as "a set of connected transcripts ... [which] share at least part of one exon in the genomic coordinates". Both definitions partition the genome into regions of connected exons defined by the transcription process. By these definitions *trans*-spliced, polycistronic transcripts are split into multiple genes.

are ways in which cells utilize available template resources in almost anyway they like to create biomolecules that are needed in a specific place at a specific time: genes are "things an organism can do with its genome" (Stotz et al. 2006). The deflationary postgenomic gene concept, together with distributed causal specificity provided by environmentally induced combinatorial control and differential recruitment, forces us to look beyond exclusive specificity, single gene searches, the increasingly restraining central dogma, and linear pathway analyses towards the new science of postgenomic and systems biology.

7. An outlook to the 21st century: Postgenomic biology

The entry into post-genome biology has been accompanied by many phrases like 'from sequence to biology' or 'to the center of biology' which suggest a move from reductionist molecular biology to systems biology (Stein 2001; Lander and Weinberg 2000). Metaphorically speaking, while the former has lost the forest (the whole network, the whole system) for keeping the trees (single genes/molecules/sequences/nucleotides), the latter tries to see the forest again with its focus on the integration of networks of genes, protein, gene regulatory functions, metabolic products and their interactions[12]. Since then systems biology developed into a new discipline with an increasing number of new research and training centers popping up worldwide. It aims at an in-depth understanding of living organisms at system-level firmly grounded firmly to the molecular level. Hence, their relationship is not one of a simple reduction: systems biology utilizes molecular biology as providing the raw data in need of integration and interpretation in light of the system as a whole. Postgenomic biology accepts molecular research as a legitimate investigative strategy but uses a system-level explanatory strategy.

> "Twentieth century biology triumphed because of its focus on intensive analysis of the individual components of complex biological systems. The 21st century discipline will focus increasingly on the study of entire biological systems, by attempting to understand how component parts collaborate to create a whole. For

[12] I am owing this metaphor to Sandra Mitchell.

> the first time in a century, reductionists have yielded ground to those trying to gain a holistic view of cells and tissues" (Lander and Weinberg 2000, 1781).

In other words, postgenomic biology has as one of its goals to reassemble the living organisms. But it also understands that the essence of systems biology lies not in computational power or high-throughput analysis – even though that is a big part of it - it is all about dynamics, the quantitative analysis of biological processes over time and space. Thus, systems biology seeks to explain biological phenomena not on a gene-by-gene basis but through the interaction of all the individual components in a cell or organism.

The postgenomic gene concept that centrally features in this new science owes its physical structure and biological function to time-and-tissue-dependent regulatory mechanisms of genome expression, with which it has to share causal specificity. These mechanisms themselves have their specificity imposed by combinatorial control, which is ultimately dependent on contingent factors. Many core regulatory processes are by default inactive, and when activated generally unstable unless stabilized by the contingent interaction with multiple other core processes. Most interactions are weak, transient, and reactive to the combinatorial influence of new participants. "By compartmentalizing different core processes and rendering their linkages to each other highly contingent, i.e. regulated by circumstance, core processes can evolve greater specializations" (Moss in press). Hence the focus of postgenomic biology is no longer on any single gene with its exclusive causal specificity but on the network of regulatory mechanisms of *genome* expression with distributed specificity, of which the intra- and extracellular environment is a central determining part. Although history reminds us that no victory is final, systems biology with its profound change in the culture and content of the life sciences seems to provide a harsh blow to the success of reductionism. It's successes are a) the nearing completion of a complete list of parts of eukaryotic cells in the form of multiple 'omic' enterprises and b) a reasonable complete diagram of their interaction not to far in the future. However:

> The cell is not hard wired, therefore a "wiring diagram" only provides, after much analysis, a combinatorically rich repertoire of circuit modules, particular subsets

of which are selected by particular environments. And because a cell's environment is in fugue, the problem of systems biology is understanding the rules of subset selection, and connecting recurrent functional modules to phenotype" (DeLisi 2004).

So it remains questionable that both ingredients alone give us the necessary and sufficient conditions for understanding the complexity of a living organism or the development of a particular trait, even if that trait is reduced to a single mRNA product.

Acknowledgments:

This work was supported by the NSF grants # 0217567 and # 0323496 (STS/SDESTdevisions), University of Pittsburgh and Indiana University. This paper would not have been written without Paul Griffiths, my collaborator for many years, who has read and commented on several drafts of the paper. I want to thank the Work in Progress Seminar at the University of Queensland, the Biostudies Reading and Discussion Group at Indiana University and the PhilDevo 2006 workshop for comments.

References:

Anastassiou, Dimitris, Hairuo Liu, and Vinay Varadan (2006), "Variable window binding for mutually exclusive alternative splicing", Genome Research 7 (1):R2.

Athanasiadis, Alekos, Alexander Rich, and Stefan Maas (2004), "Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome", PSoS Biology 2 (12):e391.

Bachellerie, Jean-Pierre, Jerome Cavaille, and Alexander Huettenhofer (2002), "The expanding snoRNA world", Biochimie 84:775–790.

Baranov, Pavel. V., Olga L. Gurvich, Andrew W. Hammer, Raymond F. Gesteland, and John F. Atkins (2003), "Recode 2003", Nucleic Acids Research 31 (1):87-89.

Bentley, David (2002), "The mRNA assembly line: transcription and processing machines in the same factory", Curr. Opin. Cell Biol. 14 (3):336-342.

Bentley, David L. (2005), "Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors." Curr Opin Cell Biol. 17 (3).

Bissell, Mina J. (1981), "The differentiated state of normal and malignant cells or how to define a "normal" cell in culture", Int Rev Cytol 70 (27-100).

——— (2003), "Tissue specificity: structural cues allow diverse phenotypes from a constant genotype", in Gerd B. Müller and Stuart A. Newman (eds.), Origination

of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology, Cambridge, MA: The MIT Press, 103-117.
Black, D. L. (2003), "Mechanisms of alternative pre-messenger RNA splicing", Annu. Rev. Biochem. 72:291–336.
Blumenthal, T., D. Evans, C. D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W. L. Chiu, K. Duke, M. Kiraly, and S. K. Kim (2002), "A global analysis of Caenorhabditis elegans operons", Nature 417 (6891):851-854.
Bradbury, Jane (2005), "Alternative mRNA Splicing: Control by Combination", PLoS Biology 3 (11):e369.
Buchler, Nicolas E., Ulrich Gerland, and Terence Hwa (2003), "On schemes of combinatorial transcription logic", Proc. Natl Acad. Sci. USA 100:5136–5141.
Burian, Richard M. (2004), "Molecular epigenesis, molecular pleiotropy, and molecular gene definitions", History and Philosophy of the Life Sciences 26 (1, Special issue on 'Genes, Genomes and Genetic Elements', ed. by Karola Stotz):59-80.
Caudevilla, C., C. Codony, D. Serra, G. Plasencia, R. Roman, A. Graessmann, G. Asins, M. Bach-Elias, and F. G. Hegardt (2001), "Localization of an exonic splicing enhancer responsible for mammalian natural trans-splicing", Nucleic Acids Research 29 (14):3108-3115.
Celottoa, Alicia M., and Brenton R. Graveley (2001), "Alternative Splicing of the Drosophila Dscam Pre-mRNA Is Both Temporally and Spatially Regulated", Genetics 159:599-608.
Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, and et al. 2005. (2005), "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution", Science http://www.sciencemag.org/cgi/content/abstract/1108625v2.
Claverie, J. M. (2001), "Gene number: what if there are only 30,000 human genes?" Science 291:1255–1257.
Coelho, Paulo S.R., Anthony C. Bryan, Anuj Kumar, Gerald S. Shadel, and Michael Snyder (2002), "A novel mitochondrial protein, TAR1p, is encoded on the antisense strand of the nuclear 25S rDNA", Genes and Development 16:2755 - 2760.
Collado-Vides, Julio, and Ralf Hofestaedt (2002), Gene Regulation and Metabolism: Postgenomic Computational Approaches. Cambridge, MA: The MIT Press.
Communi, Didier, Nathalie Suarez-Huerta, Danielle Dussossoy, Pierre Savi, and Jean-Marie Boeynaems (2001), "Cotranscription and intergenic splicing of human P2Y(11) SSF1 genes", Journal of Biological Chemistry 276 (19):16561-16566.
Crick, Francis H. C. (1958), "The Biological Replication of Macromolecules", Symp. Soc. Exp. Biol. XII:138-163.
——— (1970), "Central Dogma of Molecular Biology", Nature 227:561-563.
Davidson, Eric R (2001), Genomic Regulatory Systems: Development and Evolution. San Diego: Academic Press.
Davidson, Nicholas O. (2002), "The challenge of target sequence specificity in C_U RNA editing", J Clin Invest. 109 (3):291–294.
DeLisi, Charles (2004), "Systems Biology, the Second Time Around", Environmental Health Perspectives 112 (16):A2-A3.

Dorn, Rainer, Gunter Reuter, and Andrea Loewendorf (2001), "Transgene Analysis Proves mRNA Trans-Splicing at the Complex mod(mdg4) Locus in Drosophila", Proc. Natl. Acad. Sci. USA 98 (17):9724-9729.

Eisenberg, David, Edward M. Marcotte, Andrew D. McLachlan, and Matteo Pellegrini (2006), "Bioinformatic challenges for the next decade(s)", Phil. Trans. R. Soc. B doi:10.1098/rstb.2005.1797 (Published online).

Falk, Raphael (1986), "What is a gene?" Studies in the History and Philosophy of Science 17:133-173.

Finta, Csaba, and Peter G. Zaphiropoulos (2000a), "The human CYP2C locus: A prototype for intergenic and exon repetition splicing events", Genomics 63 (3):433-438.

——— (2000b), "The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons", Gene 260 (1-2):13-23.

——— (2001), "A statistical view of genome transcription", Journal of Molecular Evolution 53:160-162.

——— (2002), "Intergenic mRNA molecules resulting from trans-splicing", Journal of Biological Chemistry 277 (8):5882-5890.

Flomen, Rachel, Joanne Knight, Pak Sham, Robert Kerwin, and Andrew Makoff (2004), "Evidence that RNA editing modulates splice site selection in the 5-HT2C receptor gene", Nucleic Acids Research 32 (7):2113-2122.

Gilbert, Scott F. (2003), "The reactive genome", in Gerd B. Müller and Stuart A. Newman (eds.), Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology, Cambridge, MA: The MIT Press, 87-101.

Gott, Jonatha M., and Ronald B. Emeson (2000), "Functions and mechanisms of RNA editing", Annual Review of Genetics 34:499-531.

Graveley, Brenton R. (2005), "Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures", Cell 123:65-73.

Gray, M. W. (2003), "Diversity and evolution of mitochondrial RNA editing systems", IUBMB Life 55 (4-5):227-233.

Greenspan, Ralph J. (2001), "The flexible genome", Nature Reviews Genetic 2:383-387.

Griffiths, Paul E, and Russell D Gray (2003), "The developmental systems perspective: Organism-environment systems as units of evolution", in Katherine Preston and Massimo Pigliucci (eds.), The Evolutionary Biology of Complex Phenotypes, Oxford and New York: Oxford University Press, xxx-xxx.

Griffiths, Paul E., Joan Leach, Karola Stotz, and Polly Ambermoon (forthcoming), "Is there a problem with the public understanding of genetics?" Science, Technology and Human Values xx (xx).

Griffiths, Paul E., and Karola Stotz (2006), "Genes in the Postgenomic era", Theoretical Medicine and Bioethics 27 (6):xxx-xxx.

——— (in press), "Gene", in David Hull and Michael Ruse (eds.), Cambridge Companion for the Philosophy of Biology, Cambridge: Cambridge University Press.

Harrison, Paul M., Anuj Kumar, Ning Lang, Michael Snyder, and Mark Gerstein (2002), "A question of size: the eukaryotic proteome and the problems in defining it", Nucleic Acids Research 30 (5):1083–1090.

Hirotsune, Shinji, Noriyuki Yoshida, Amy Chen, Lisa Garrett, Fumihiro Sugiyama, Satoru Takahashi, Ken-ichi Yagami, Anthony Wynshaw-Boris, and Atsushi Yoshiki (2003), "An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene", Nature 423 (6935):91-96.

Kampa, Dione, Jill Cheng, Philipp Kapranov, Mark Yamanaka, Shane Brubaker, Simon Cawley, Jorg Drenkow, Antonio Piccolboni, Stefan Bekiranov, Gregg Helt, Hari Tammana, and Thomas R. Gingeras (2004), "Novel RNAs identified from an in-depth analysis of the transcriotome of human chromosomes 21 and 22", Genome Research 14:331-342.

Kapranov, Phillip, Jorg Drenkow, Jill Cheng, Jeffrey Long, Gregg Helt, Sujit Dike, and Thomas R. Gingeras (2005), "Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays", Genome Research 15:987-997.

Kawasaki, Hiroaki, and Kazunari Taira (2004), "Induction of DNA methylation and gene silencing by short interfering RNAs in human cells", Nature 431:211-217.

Keller, Evelyn Fox (2000), The Century of the Gene. Cambridge, Mass.: MIT Press.

Kishore, Shivendra, and Stefan Stamm (2006), "The snoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C", Science 311 (5758):230 - 232.

Kitcher, Philip (1984), "1953 and all that: A Tale of Two Sciences ", Philosophical Review 93:335-373.

Ladd, Andrea N., and Thomas A. Cooper (2002), "Finding signals that regulate alternative splicing in the post-genomic era", Genome Biology 3 (11):1-16.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, and et al. (2001), "Initial sequencing and analysis of the human genome", Nature 409:860–921.

Lander, Eric S., and Robert A. Weinberg (2000), "Journey to the centre of biology", Science 287 (5459):1777-1782.

Leipzig, Jeremy, Pavel Pevzner, and Steffen Heber (2004), "The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome", Nucleic Acids Research 32 (13): 3977–3983.

Levine, Michael, and Robert Tjian (2003), "Transcription regulation and animal diversity", Nature 424:147–151.

Malek, Olaf, and Volker Knoop (1998), "Trans-splicing group II introns in plant mitochondria: The complete set of cis-arranged homologs in ferns, fern allies, and a hornwort", Rna-a Publication of the Rna Society 4 (12):1599-1609.

Maniatis, Tom, and Robin Reed (2002), "An extensive network of coupling among gene expression machines", Nature 416:499–506.

Marcotte, Edward M. (2002), "Predicting protein function and networks on a genomewide scale", in Julio Collado-Vides and Ralf Hofestaedt (eds.), Gene Regulation and Metabolism: Postgenomic Computational Approaches, Cambridge, MA: The MIT Press.

Mattick, John S. (2004), "RNA regulation: a new genetics?" Nature Reviews Genetics 5 (4):316-323.
Moss, Lenny (2003), What Genes Can't Do. Cambridge, Mass.: MIT Press.
——— (in press), "Redundancy, Plasticity and Detachment: the Implications of Comparative Genomics for Evolutionary Thinking", Philosophy of Science 73 (5, PSA 2004 proceedings):xxx-xxx.
Pirrotta, Vincenzo (2002), "Trans-splicing in Drosophila", Bioessays 24 (11):988-991.
Ptashne, Mark, and Alexander Gann (2002), Genes and Signals. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Ridley, Matt (2003), Nature via Nurture: Genes, Experience, and What Makes us Human. New York: Harper Collins.
Rivier, C., M. Goldschmidt-Clermont, and J. D. Rochaix (2001), "Identification of an RNA-protein complex involved in chloroplast group II intron trans-splicing in Chlamydomonas reinhardtii", Embo Journal 20 (7):1765-1773.
Robert, Jason S. (2004), Embryology, Epigenesis and Evolution: Taking Development Seriously. Cambridge: Cambridge University Press.
Rowen, Lee, Janet Young, Brian Birditt, Amardeep Kaur, Anup Madan, Dana L. Philipps, Shizhen Qin, Patrick Minx, Richard K. Wilson, Leroy Hood, and Brenton R. Graveley (2002), "Analysis of the Human Neurexin Genes: Alternative Splicing and the Generation of Protein Diversity", Genomics 79 (4):587-598.
Rueter, S. M., T. R. Dawson, and R. B. Emeson (1999), "Regulation of alternative splicing by RNA editing", Nature 399:75-80.
Sarabhai, A.S., A.O.W. Stretton, S. Brenner, and A. Bolleh (1964), "The Colinearity Hypothesis", Nature 201:13-17.
Shin, Chanseok, and James L. Manley (2004), "Cell signalling and the control fo the pre-mRNA splicing", Nature Reviews Molecular Cell Biology 5:727-738.
Smith, C.W.J., and J. Valcarcel (2000), "Alternative pre-mRNA splicing: the logic of combinatorial control", Trends. Biochem. Sci. 25:381–388.
Snyder, Michael, and Mark Gerstein (2003), "Defining Genes in the Genomics Era", Science 300:258- 260.
Stamm, Stefan (2002), "Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome", Human Molecular Genetics 11 (20):2409-2416.
Stein, Lincoln D. (2001), "Genome Annotation: From Sequence to Biology", Nature Reviews Genetics 2 493-503.
——— (2004), "End of the beginning", Nature 431:915-917.
Stotz, Karola (2006), "With genes like that, who needs an environment? Postgenomics' argument for the ontogeny of information", Philosophy of Science 73 (5, PSA 2004 proceedings):(Preprint in PhiSci Archive).
Stotz, Karola, Adam Bostanci, and Paul E. Griffiths (2006), "Tracking the shift to 'post-genomics'", Community Genetics 9 (3):xx-xx.
Sturm, N. R., and D. K. Campbell (1999), "The role of intron structures in trans-splicing and Cap 4 formation for the Leishmania spliced leader RNA", Journal of Biological Chemistry 274 (27):19361-19367.

Tasic, Bosiljka , Christoph E. Nabholz, Kristin K. Baldwin, Youngwook Kim, Erroll H. Rueckert, Scott A. Ribick, Paula Cramer, Qiang Wu, Richard Axel, and Tom Maniatis (2002), "Promoter choice determines splice site selection in protocadherin alpha and -gamma pre-mRNA splicing", Molecular Cell 10 (1):21-33.

Turney, Jon (2005), "The Sociable Gene", EMBO Reports 6 (9):809-810.

Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, and et al. (2001), "The sequence of the human genome", Science 291:1304–1351.

Waters, C. Kenneth (1990), "Why the Antireductionist Consensus Won't Survive the Case of Classical Mendelian Genetics", in Arthur Fine, Micky Forbes and Linda Wessells (eds.), Proceedings of the Biennial Meeting of the Philosophy of Science Association: Philosophy of Science Association, 125-139.

——— (1994), "Genes made molecular", Philosophy of Science 61:163-185.

——— (2000), "Molecules Made Biological", Rev. Int. de Philosophie 4 (214):539- 564.

——— (forthcoming a), "A Pluralist Interpretation of Gene-centered Biology", in Stephen Kellert, Helen E. Longino and C. Kenneth Waters (eds.), Scientific Pluralism, Minneapolis: University of Minnesota Press.

——— (forthcoming b), "Causes that make a difference", Journal of Philosophy.

Watson, James D., and Francis H. C. Crick (1953a), "A structure for deoxyribose nucleic acid", Nature 171:737-738.

——— (1953b), "Genetical implications of the structure of deoxyribose nucleic acid", Nature 171:964-967.

Weber, Marcel (2004), Philosophy of Experimental Biology. Cambridge, New York: Cambridge University Press.

Werner, Eric (2005), "Genome semantics, in silico multicellular systems and the Central Dogma", FEBS Letters 579 (8):1779-1782.

Wissinger, B., W. Schuster, and A. Brennicke (1991), "Trans Splicing in Oenothera Mitochondria: nad1 mRNAs Are Edited in Exon and trans-Splicing GroupII Intron Sequences", Cell 65 (3):473-482.

Figure 1

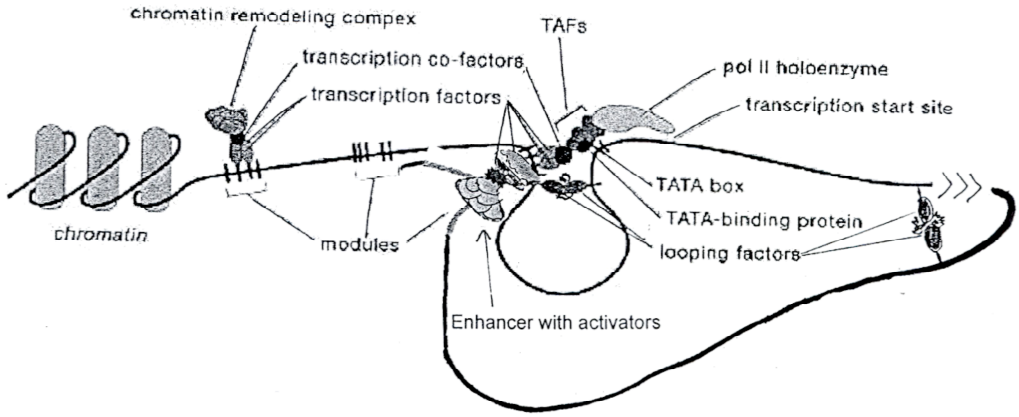

Figure 1: Programmatic schema of the diverse 3' and 5' cis-regulatory modules and its chromatin remodeling and diverse transcription complexes. Shown are some auxiliary looping factors that help bringing the complexes into contact, especially the mostly very distant enhancer with its activators, who is recruiting several of the other complexes.

Figure 2

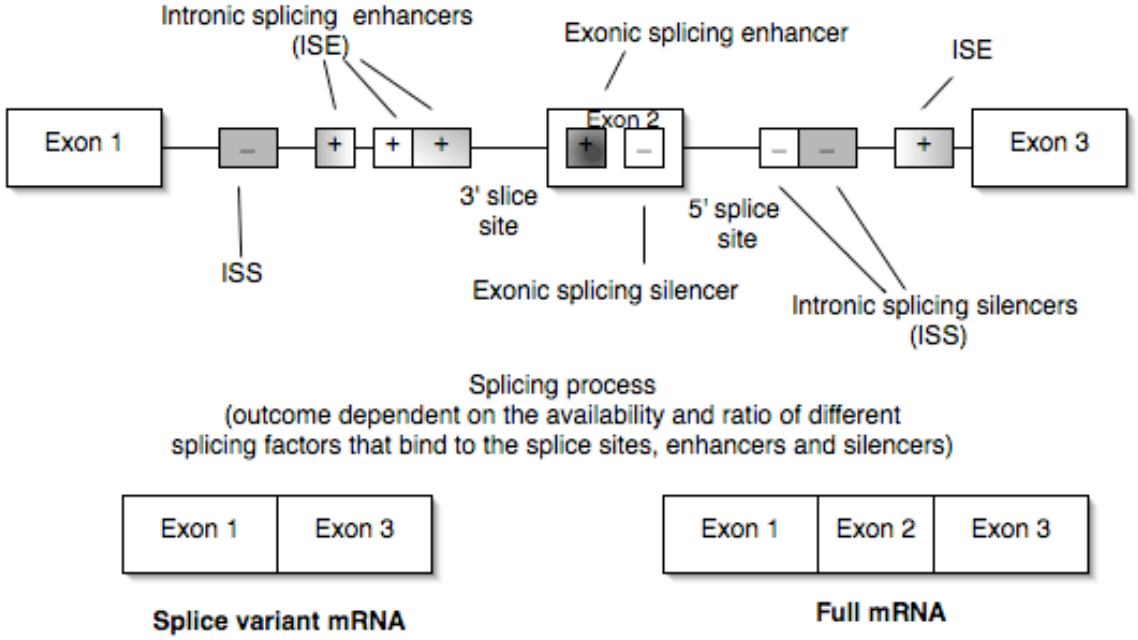

Figure 2: Programmatic schema of the distribution of cis-regulatory splicing modules for one exon. Beside the canonical (plus possible non-canonical) splice sites (here very simplified), there exist a range of enhancers and inhibitors within the exon and in the flanking introns (shaded boxes). Multiple copies of the same sequence (colored in the same shade) bind the same splicing factors, either serine/arginine-rich (SR) splicing proteins or heterogeneous nuclear ribonucleoproteins (hnRNPsFigure 3:

Figure 3

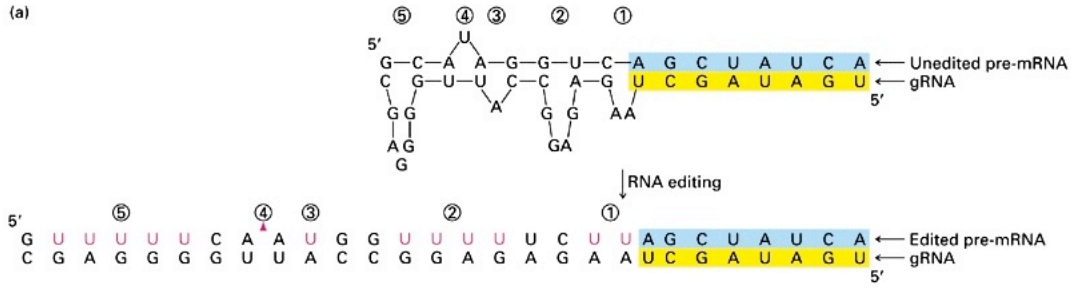

Figure 3: U-insertion and U-deletion in Kitanoplasts via guideRNA (gRNA). The gRNA forms binds to its 8-nucleotide-long complementary 3' anchor sequence in the pre-mRNA. The rest of the gRNA forms an incomplete double strand with the pre-mRNA, with bulges and loops in the gRNA presenting the editing targets for insertion, and in the mRNA for deletion. The edited mRNA matches the gRNA

Figure 4

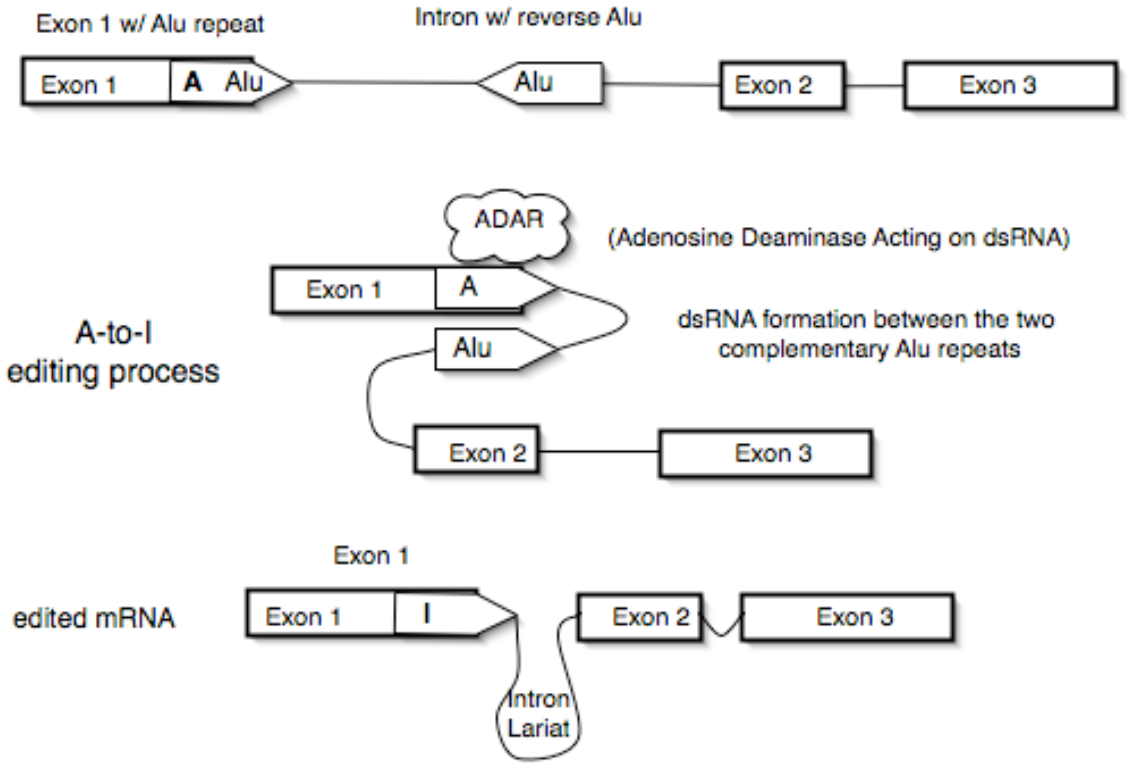

Figure 4: A-I nuclear RNA editing via exon-intron double stranded RNA formation (often by means of reverse Alu repeats of which many coding genes contain several copies. For simplicity only one editing site is shown, but the number is often much higher (around 20 editing sites in one dsRNA formation) and negatively correlated with the distance between the two Alu repeats.

Table 1

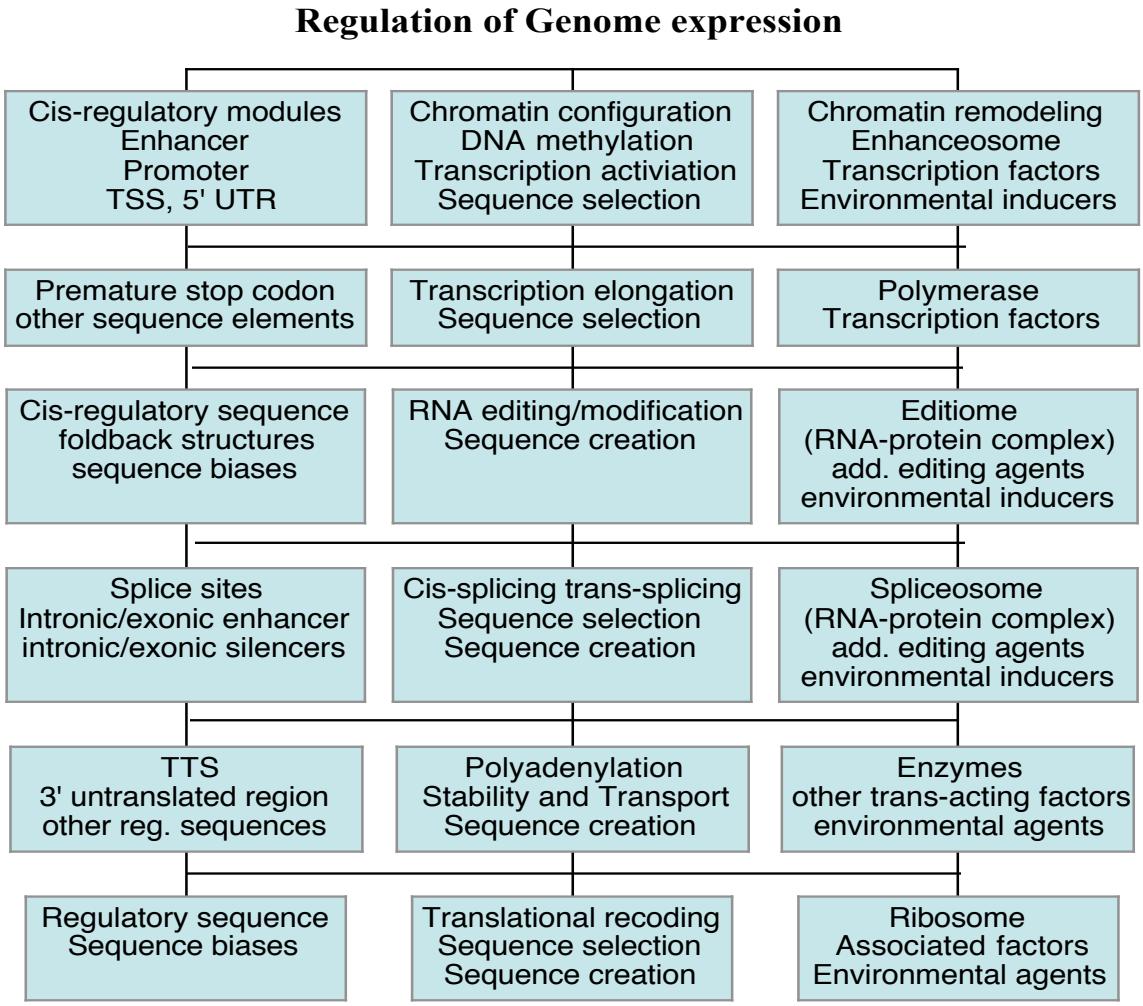

Table 1: Elements with sequence specificity in eukaryotic genome expression. The middle column the transcriptional and translational stages of nucleic acid from the packaged DNA over RNA to the amino acid sequence. The left side depicts the different cis-regulatory sequences involved at different stages of genome expression, while the right side shows the divers trans-acting factors such as transcription factors, splicing proteins and ncRNA and environmental inducers.