

Searching for Molecular Solutions – Cited Notes

CHAPTER 9

These Files contain details on all references to this ftp site within **Chapter 9** of *Searching for Molecular Solutions*. The page numbers of the book where the reference is made are shown in the Table below, the corresponding page number for this file, and the title of each relevant section.

Contents:

Book Reference Page Number	Page Number in this File	Section	
		No.	Title
336	2	29	RNAi Genomic Screens
338	8	30	Genomic Cycle Notes
339	24	31	Global Protein-Protein Interaction Screening

Subsection titles for Section 30	Page Number in this File
Generators of Genomic Diversity and Complexity	8
Biological compactness	14
The Largesse of Large T	17

Section 29: *RNAi Genomic Screens*

Cited on p. 336 of *Searching for Molecular Solutions*

This section provides some background information on functional genomic screens, with an emphasis on those employing RNAi.

Functional Genomic Screens

Functional screens on a genomic scale can be at the level of the transcriptome (the total transcribed genome), or the proteome[♥] (the total protein diversity), or involve interactions between them. These investigations can be performed directly on cells of interest with a battery of probes, or following specific perturbation of cellular systems aimed at screening for desired phenotypes. For example, high-density protein microarrays (especially antibody arrays)^{1,2} can be used to simultaneously measure the relative levels of a large set of cellular protein markers either directly, or before and after a specific chemical or biological treatment.

A treatment which perturbs cellular function is highly informative if it targets a specific endogenous gene product of interest, such that the resulting phenotype is directly linked to the alteration in the normal function of the target of interest. This has long been pursued by ablating target gene products, making them in excess, or modifying them in a variety of ways. The significant change in relatively recent times is that this process can be conducted on a genome-wide basis as a functional screen. The initiation of such

[♥]The proteome of an organism refers to the total complement of proteins, encoded by its genome, which it produces. For multicellular organisms where cells undergo differentiation into a variety of types, only a subset of the total proteome is expressed in any one differentiated cell. (Some proteins are ubiquitously expressed, and some are found only in specific cell lineages). Even if a study using an isolated cell type from a multicellular organism physically analyzes all proteins present, this represents only a specific subset of the whole proteome encoded by the organism's genome. This is understood in cases where 'proteomic' studies refer to only a single differentiated cell category.

screening as an artificial input perturbation is usually (but not always) at the transcription level, such that a modification occurs for that fraction of the transcriptome normally expressed in the cells of interest. The read-out, though, is usually at the level of the functional consequences of such a perturbation at the level of the protein phenotype, or the cell's proteome. In the context of searching for drug targets, in Chapter 9 of *Searching for Molecular Solutions* we referred to small inhibitory RNAs (RNAi), the deployment of which has become a major tool for genome-wide functional analyses. Libraries of RNAi species covering a large majority of known expressed genes have been prepared³, whose vast potential has inspired mushrooming applications since around 2003⁴⁻⁶. These can be delivered to cells in microwell or microarray formats⁷, either using direct uptake of double-stranded RNAs⁸, or using retroviral vectors for efficient transfection and transcription of short hairpin RNAs⁹. The latter vector-based strategy has the considerable advantage that it can be arranged such that stable transfectants are generated, and the transcription of the RNAi can be inducible³. (Inducibility is especially important when a target is an essential component of cellular survival).

How are cell phenotypes analyzed during a genomic functional screen by RNAi, such that the relevant RNAi molecules mediating the effect are identified, and thereby their target genes? If knock-down of a cellular gene by stably-expressed RNAi confers a positively-selectable phenotype, then cells with the relevant RNAi can be isolated and analyzed. One way to facilitate the identification of RNAi species out of large libraries is to equip the vector which encodes these RNA transcripts with a readily-identifiable and unique 'barcode'. A cell expressing the RNAi of interest is then always linked to the barcode through the corresponding vector, which itself can be conveniently screened for via hybridization-based microarray analysis³. The nature of the sought-after phenotype is important in the screening design. A positive phenotype (such as anchorage-independent cell growth) enables direct selection of cells bearing the specific member of

⁹ Artificial exploitation of the ancient system of RNA interference can deliver interfering RNAs in several ways, which converge at the level of short interfering RNAs (siRNA). These include direct transfection of siRNAs, cellular uptake of long double-stranded RNAs which are then cleaved and processed, or transcription of short hairpin RNAs from suitable transfected vectors (including viral vectors).

an RNAi library which ‘knocks down’ expression of genes which normally *prevent* such substrate-independent proliferation (as depicted in Fig. 9.Na below). On the other hand, if essential genes for cell growth are to be screened for, by definition their knock-down will ablate cell growth, or even promote cell death. When barcodes are present in integrated copies of the library vector, these can still identify RNAi species which negatively affect growth. Barcodes corresponding to such RNAi library members will at best not be amplified by cell proliferation, and may through cell death be removed from the total population. They can thus be identified by comparing barcode populations with and without induction of the RNAi library (Fig. 9.Na). RNAi screens on a genomic scale are also amenable to analysis by microscopic high content screening^{10,11} (referred to in Cited Notes for Chapter 8[♥]).

While the power of RNAi for functional genomic screening is not in dispute, there have been certain potential pitfalls noted. In particular, since the lengths of siRNAs which ultimately hybridize to target mRNAs (and mediate their destruction) are not long and absolute matching is not required for an effect, ‘off-target’ effects on non-target mRNAs have been commonly observed. These also have the potential for raising the false-positive rate in RNAi screens^{12,13}. One answer to this is to screen with multiple RNAi species against each target gene, but obviously this creates logistic headaches when taken to genomic scales. In mammalian (but not invertebrate) systems, RNAi duplexes cannot be too long or the interferon system is triggered, with undesirable consequences¹³. When transient RNAi systems are used (as with transfected pre-made RNA duplexes), the efficiency of knock-down of protein targets depends on the protein half-life^{14,15}. RNAi can only affect protein levels by preventing new synthesis, so if pre-formed proteins persist on a time scale comparable to the half-life of the RNAi effect itself, such target proteins may fail to show much diminution through the agency of RNAi, even if their corresponding mRNAs are efficiently destroyed.

[♥] See the file SMS–CitedNotes–Ch8/Section 21; from the same ftp site.

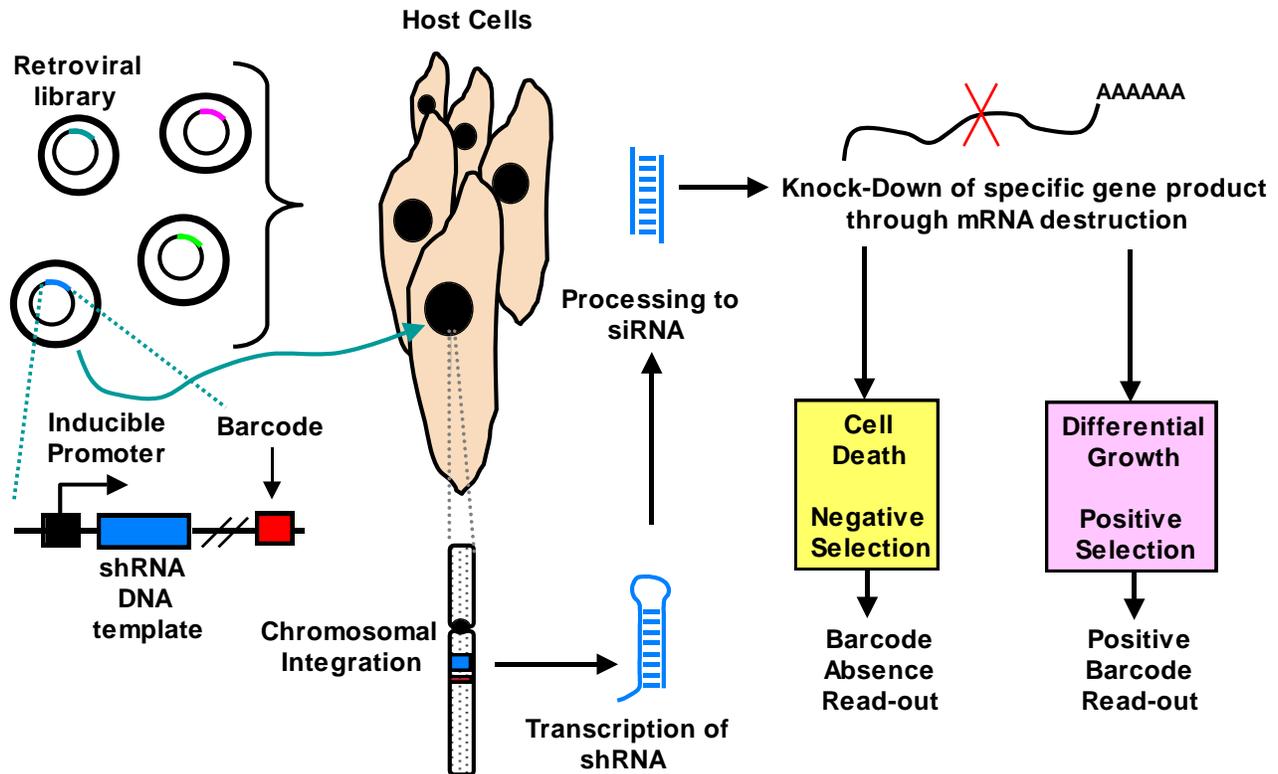


Fig. 9.Na

Functional screens with a retroviral short-hairpin RNA (shRNA) genome-wide library. The retroviral vector contains a library of templates for shRNAs against a genome-wide set of target mRNAs, and also unique 'barcode' sequences which accompany each specific shRNA. Upon infection of cells with the library, retroviral RNAs are reverse-transcribed and chromosomally integrated. Induction of transcription of the shRNAs results in their processing (by the endogenous RNAi machinery) into siRNAs which target specific mRNAs. A resulting phenotype conferring positive selection can be directly isolated, and barcodes 'read' to identify shRNAs and their corresponding targets¹⁶. If a negative selection occurs resulting in cell death or halting of growth, the relevant barcodes will be absent or highly under-represented (in 'dropout' mode) from within the final cell population. Such codes can be consequently identified by comparing sets with and without induction of shRNAs through hybridization-based microarray analyses^{17,18}. The barcode can incorporate a section of the shRNA itself¹⁸.

With these caveats on RNAi applications in mind, it is well worth noting that there are potential alternatives for genome-wide modulation of gene expression or gene product function. A long-standing approach, originating in prokaryotic systems but readily extendable to mammalian cells, is insertional mutagenesis by mobile genetic elements. The concept is simple enough: if cells of interest are exposed to an efficiently-translocating genetic element such that an average of one random 'hit' per genome is achieved, the resulting insertional library will provide a gamut of altered phenotypes through disruption of specific gene functions. Clearly, insertion of a sizable mobile DNA sequence into or adjacent to a host cell coding sequence is unlikely to be compatible with continued normal function. While intergenic insertions may be at least superficially phenotypically neutral, insertions which act as a mutagenic event for important genes can be scored as novel phenotypic alterations. Identification of cells with altered properties (from complex effects such as growth rates, to specific gene expression patterns) then provides a 'handle' on finding the genes conferring the relevant phenotype, through the insertional tag [▼]. By a number of methods ^{*}, sequences flanking specific insertions can be defined, in turn allowing identification of the gene whose expression is thereby disrupted.

For mammalian cells, retroviruses have been a convenient tool for insertional mutagenic scans, through their in-built facility for reverse transcription and genomic integration ^{19,20}. These have been used for many years for identification of genes initiating or promoting carcinogenesis ²¹, which has productively continued in more recent times ²². Yet most retroviral genomic insertions deviate strongly from true randomness ²³, and some alternatives have emerged in the form of transposons active in mammalian cells. A defective mobile element in fish was artificially restored to function and found to have a wide host range ^{24,25}, with useful applicability in many insertional mammalian

[▼] Note the difference here from some of the above-mentioned RNAi screens (as in Fig. 9.Na), where a retroviral vector is used to achieve single-copy genomic integration, but the phenotypic 'knock-down' is through a *trans*-effect (transcription and activation of the siRNA molecules towards target mRNAs). For insertional mutageneses (whether via retroviruses or other elements), disruption of expression is through the *cis*-effect of insertion itself.

^{*} One such option is 'inverse' PCR (see Cited Notes for Chapter 4 in the file SMS–CitedNotes-Ch4/Section 6; from the same ftp site).

mutagenesis screens^{23,26,27}. While this reconstructed transposon, dubbed '*Sleeping Beauty*', shows low transposability in some cellular backgrounds (including embryonal stem cells²⁸), alternatives have arisen to solve this problem. Another transposon from an insect (moth) source[▼], '*PiggyBac*'²⁹, has proven to be a useful transposable element in murine embryonal stem cells of a suitable genetic background³⁰. The theme of insertional mutagenesis is extended in a variation termed the RAGE approach (Random Activation of Gene Expression)³¹, where endogenous genes are activated if random insertions of a specially designed plasmid occur upstream of their coding sequences.

We might recall from previous chapters that both intramers and intrabodies (intracellular aptamer and antibody reagents) can target proteins directly, and could potentially be applied in array-based formats for proteome-wide screening. Alternatives also exist for genomic functional studies at the level of transcriptional initiation, by means of artificial transcription factors (specifically zinc finger proteins^{*}). Combinatorial shuffling of zinc finger building blocks can result in a huge range of DNA binding specificities, which theoretically can bind to regulatory sequences for all genes³². When such a library of binding specificities is linked with either a transcriptional activation or repression domain, it can be potentially used for selection of novel phenotypes on a genomic scale³². Zinc finger technology can also be married with mutagenesis through the development of zinc finger nucleases^{*}, where the DNA-interactive zinc fingers conferring binding specificity are linked with a bacterial restriction enzyme of a specific class.

▼ These instances of the utility of genetic elements from fish and moth sources emphasize the point that the biosphere as a whole is a potential source of novelty and benefit for biotechnology and research.

* For additional details, see the file SMS–CitedNotes–Ch4/Section 8B, from the same ftp site.

Section 30: **Genomic Cycle Notes**

Cited on p. 338 of *Searching for Molecular Solutions*

This section extends a theme raised in Fig. 9.7 of *Searching for Molecular Solutions*, where the function of the subsidiary 'omics (transcriptome, proteome, etc.) are depicted as serving the replication of the genome in a cycle. In part B of this figure, the 'unfolding' of a developing organism via its genome is depicted. The 'unfolding' of the genome during the development of a complex organism reveals many levels of complexity and diversification, far beyond that which the raw number of genes alone would seem to indicate. Some further details of this are provided here.

Generators of Genomic Diversity and Complexity

The realization of the information embodied in a DNA genome can be viewed as arising from an initial layer of RNA transcription and a subsequent layer of the attainment of function through proteins and RNA molecules assuming specific three-dimensional shapes through folding. These layers interact with each other through intricate and highly complex pathways such that the entire process is regulated and controlled in an ordered and efficient manner. Of course, many higher-level layers also become superimposed on the first layers, as increasingly complex order is generated, from multicellular organ systems up to the level of complex neurobiology and rational thought. Here we can consider some of the underlying processes which allow such complexity to arise from what appears at face value to be a relatively small number of linear strings of nucleotides assignable as genes.

Table 9.N1-A indicates some known mechanisms which drive the accumulation of diversity and complexity in organisms, from their underlying genomic information:

Table 9.N1-A Molecular Diversifiers			
Diversification Effect		Diversification Mechanism(s)	Diversification Principle For Protein Expression
Nucleic Acid level	Somatic DNA rearrangement	Genomic DNA segment shuffling	Coding Segment combinatorics
	Differential promoter usage	Control of expression with respect to specific exons	Exon combinatorics
	Differential RNA splicing	<i>Cis-splicing</i> : alternate exons / intron retention <i>Trans-splicing</i> : alternate coding sequences	Exon / intron / coding segment combinatorics
	Differential RNA polyadenylation	Alternative poly(A) site choice	Alternative terminal exons
	Somatic Mutation	Genomic DNA mutations	Mutational diversity (combinatorics at amino acid level)
	RNA editing	Changes to expressed RNA sequences	RNA coding sequence diversity (combinatorics at amino acid level)
	Differential tRNA use	Modulation of protein folding	Alternative protein folding pathways
Protein Level	Post-translational modifications	Enzyme-mediated covalent protein alterations	Transfer of a variety of groups to proteins post-expression; functional diversity
	Protein splicing	C-terminal autocatalytic domains; inteins (self-splicing), other trans-splicing events	Increase in functional diversity from one protein sequence
	Ligand binding	Conformational protein change after non-covalent ligand binding	Allosteric functional diversity
	Prion-like processes	Protein conformational changes	Increase in functional diversity from one protein sequence
	Proteolytic processing	(1) Modification of main protein (2) Release of functional peptides	Increase in functional diversity. Processed polypeptide may be dedicated source of peptides, or have secondary role as source of 'cryptic' functional peptides.
	'Moonlighting'	Protein functional or structural diversity	Increase in functional diversity. Includes some intrinsically unstructured proteins.

Most of the processes within Table 9.N1-A are alluded to within *Searching for Molecular Solutions*[▼]. The full extent of individual diversification mechanisms and their global importance are still being defined. For example, although RNA editing has been known for decades, its genome-wide impact in humans as a diversifier of the transcriptome is only just becoming appreciated^{33,34}. Continuing this theme into the area of regulation, some of the diversity of control mechanisms are listed in Table 9. N1-B below.

Table 9.N1-B System and Regulatory Controls			
Control Effect		Control Mechanism(s)	Control Principle For System
RNA mediators	General RNA interference	RNAi / miRNAs /antisense RNA	Control of transcription (RNAi, miRNA, antisense RNA) or translation (miRNA)
	Functional RNAs	RNAs with non-ribosomal roles in expression regulation	Natural aptamers
	Riboswitches	Ligand-based expression regulation at RNA level	Feedback control of expression on relevant RNAs through specific ligand interactions
	Natural ribozymes	RNA Processing	Intron excision, other splicing effects
Protein mediators; RNA Targets	RNA-binding proteins	Protein binding of RNA at specific sites / motifs	Control of transcription / stability
	RNA degradation / stability	Protein-based degradation systems; Nonsense-mediated decay system.	RNA turnover, quality control; regulation of expression through RNA half-life
	RNA nuclear / cytoplasmic traffic	Protein-based RNA transport to and from nucleus	Control of RNA function by regulation of cellular compartmentalization
Protein mediators; DNA Targets	Transcription Factors (TFs)	Binding to DNA control sequences and / or protein-protein interactions	Combinatorial control (multiple TFs); Control by post-translational modifications; Redox control

[▼] Protein splicing; inteins are discussed in the file SMS–Extras-Ch3/Section A2, from the same ftp site.

Table 9.N1-B System and Regulatory Controls, continued.

Control Effect		Control Mechanism(s)	Control Principle For System
Protein mediators; Protein or other Targets	Post-translational Modifications	Covalent modifications for information transmission (principally phosphorylation)	Signal / informational transduction
	Ligand Binding	Non-covalent ligand binding for information transmission	Regulation of expression or other processes
	Protein degradation / stability	Tagging for degradation pathways (ubiquitin, N-end rule, other)	Protein turnover control; expression control through half-life regulation; monitoring of defective or unfolded proteins
	Protein Transport	Protein sequences mediating transport into cellular compartments; post-translational modifications	Functional regulation through control of cellular compartmentalization

Direct and Indirect Information Transmission

The unfolding of genomic information can be broadly subdivided into effects which are direct or indirect. The former refers to functional RNAs or proteins which are directly 'read-off' from genomic sequence, by transcription alone for RNA molecules, or with the necessary subsequent translation processes for proteins. But this is only the beginning, since enzymes directly encoded by the genome then act on their specific substrates, with the resulting formation of products which inevitably arise from the expression of the genome, but are not directly encoded by it. If molecule **C** is not (or cannot be) directly encoded in a genome, but products **A** and **B** interact to form **C**, then for a genome to direct the production of **C**, it is only necessary to specify **A** and **B**. And **C** is then an indirect product of the outflow of genomic information. Indirect effects can also result from non-covalent (and often transient) interactions. If **A** and **B** transiently form **A•B**, where the complex has an altered function, then another indirect pathway to such a

function is achieved, through the ‘design’ of the two components. Of such effects and vastly more is the higher-level interactome born.

These hypothetical circumstances are simple enough, but in real biological situations the relevant inter-relationships determining a product or system are often much less obvious. Consider, for example, a subset of the glycome (itself the subset of the metabolome which includes simple and complex carbohydrates [♥]) which incorporates polysaccharides called heparans, and their sulfated derivatives. These sulfated heparans are very important as coreceptors during embryonic development ^{35,36}, and their patterns of sulfation modification appears to be diverse and specifically regulated (specific cell types will reproduce specific modification patterns ³⁷). The importance of heparans has led to the coining of the term ‘heparanome’ ^{*}, as a ‘non-templated’ process for information transmission ^{38,39}. For the glycome in general, carbohydrate-mediated information transmission has been referred to as ‘the sugar code’ ⁴⁰ or ‘glycocode’ ⁴¹. But is this somehow ‘beyond the genetic code’ and the primacy of the genome as the biological information repository? While the answer to this is ‘no’, the processes involved are far more subtle than the above **A**, **B**, **C** hypothetical.

Heparan sulfates, for example, are built by a tool-box of enzymes which both transfer sulfate groups (sulfyltransferases) and remove them (sulfatases) ³⁷. The latter may have an ‘editing’ function for heparan sulfates already on cell surfaces ³⁸. While the precise mechanisms for the ‘regulated diversity’ of heparan sulfation patterns are not defined, many indirect effects may come into play. As general examples, again with the above hypothetical players, consider some model scenarios :

Compound **L** is enzymatically produced by genomic products **U**, **V**, and **W**.
U, **V**, and **W** are only produced in certain cell differentiation lineages under the control of genomically-encoded transcription factor **Y**.

[♥] See Cited Notes for Chapter 8; in the file: SMS–CitedNotes-Ch8/Section 22 (Genomics and Chemogenomics); from the same ftp site.

^{*} Here we might recall the ‘polyomic’ proliferation also referred to in SMS–CitedNotes-Ch8/Section 22.

Compound **L** is secreted but only taken up other cells expressing genomically-encoded receptor **R**.

If **A** binds **L** as a ligand (forming **AL**), then **AL** has changed enzymatic specificity or **AL** activates the expression of a third relevant enzyme **D** or modulates the activity of **B** or(many other possibilities).

If concentration of **A** exceeds a certain threshold, **A** will block the effects of **B**.

A specific cofactor for **B** (genomically encoded factor **E**) modifies substrate recognition of **B** such that a different substrate modification pattern by **B•E** occurs (without changing the catalytic specificity). Cofactor **E** is expressed as an end result of a signaling pathway in cells triggered when receptor **R** binds ligand **L**.

...And we could go on and on. The challenge for real systems biology is to define the actual complex networks involved in real biological systems, including tracing the 'sugar code' back to its genomically-specified origins. But this is not to say that genomes cannot be the blind beneficiaries of inherent organizational properties of complex systems. As noted in Chapter 2 of *Searching for Molecular Solutions*, self-organizational phenomena ⁴² are an important aspect of biosystems. If a genome directly or indirectly specifies **A-Z**, and these 26 products mutually associate and initiate an on-going cascade of effects of increasing complexity, at least some of this might be attributable to 'spontaneous' self-organizational properties of specific system components (for example, **G**, **H**, **I** and **J**). If so, it would be advantageous for natural selection to 'find' components **G-J** as effective parts for the larger biosystem toolbox.

Another important instance of 'indirect' genomic specification is exemplified by the immune system (Chapter 3 of *Searching for Molecular Solutions*). While the genome 'sets up' the initial conditions for the immune system, by definition it cannot specify the output, since that is determined by environmental input (pathogens or other non-self molecules). This recalls the issue of defining 'self', where antibody idiotypes are obviously synthesized by organisms but not specified by their genomic coding sequences. But immune systems themselves are established through developmental

processes from a single zygote, and this recalls again the introduction to this topic by the ‘heparanome’, through its role in developmental regulation. Development itself can be viewed as the outcome of complex cascades of indirect effects, all ultimately emanating from the genome ♥.

An important aspect of genomic diversification and complexity generation is its incredible parsimony and compactness. Let us consider this in some detail, with a focus on some specific proteins as examples....

Biological compactness

An aspect of ‘natural design’ at both the levels of individual molecules and complex interactive molecular systems is the often stunning economy or compactness of such biological components, especially when considered in the context of the cell or organism as a whole. The above material in this Section has considered the means by which ~20,000 human genes can give rise to vastly more intricacies than the sheer number itself would imply. It could be proposed that a positive evolutionary selection exists towards maximizing the efficiency of the packing of genomic information, although again this is to be distinguished from genome size itself, which is also under the influence of parasitic DNA replicative elements *. A less efficiently packed arrangement might suffer from inferior global system energetics, or possibly tight coupling of genetic circuitry is a fundamental requirement for complex genomes and thus a basic constraint on the type of genomes which can encode multi-cellular organisms. On the other hand, at least some of the ‘efficiencies’ of genomic arrangements may stem from the nature of their evolutionary origins themselves. Duplications of promoter regions, for example, might result in neofunctionalizing divergence where a shared gene coding sequence is differentially expressed in different

♥ To qualify this slightly, input from the maternal genome may be needed to ‘kick-start’ the process, but it is still genomically-specified.

* This is also not to say that much of the non-coding regions of the genome has no function. As we have seen, there is increasing recognition of the regulatory roles of such genomic segments. In contrast to ‘selfish’ parasitic elements, such regions have been termed ‘polite’ DNAs ⁴³.

tissue sites during development. The mammalian *microphthalmia* (MITF) gene locus is a case in point for this type of arrangement. Its prototype gene (MITF-M) is a master regulator of melanocytic (pigmented) cells and melanocytic gene expression, and is expressed in a tightly regulated manner in this cell lineage^{44,45}. However, at least nine other MITF gene isoforms are known, each with a different promoter, first exon, and alternate transcript splicing⁴⁶. This kind of configuration, which in the case of MITF results in a minimum of nine different expressed isoforms from one gene[▼], is depicted in simplified and generalized form in Fig. 9.Nb below.

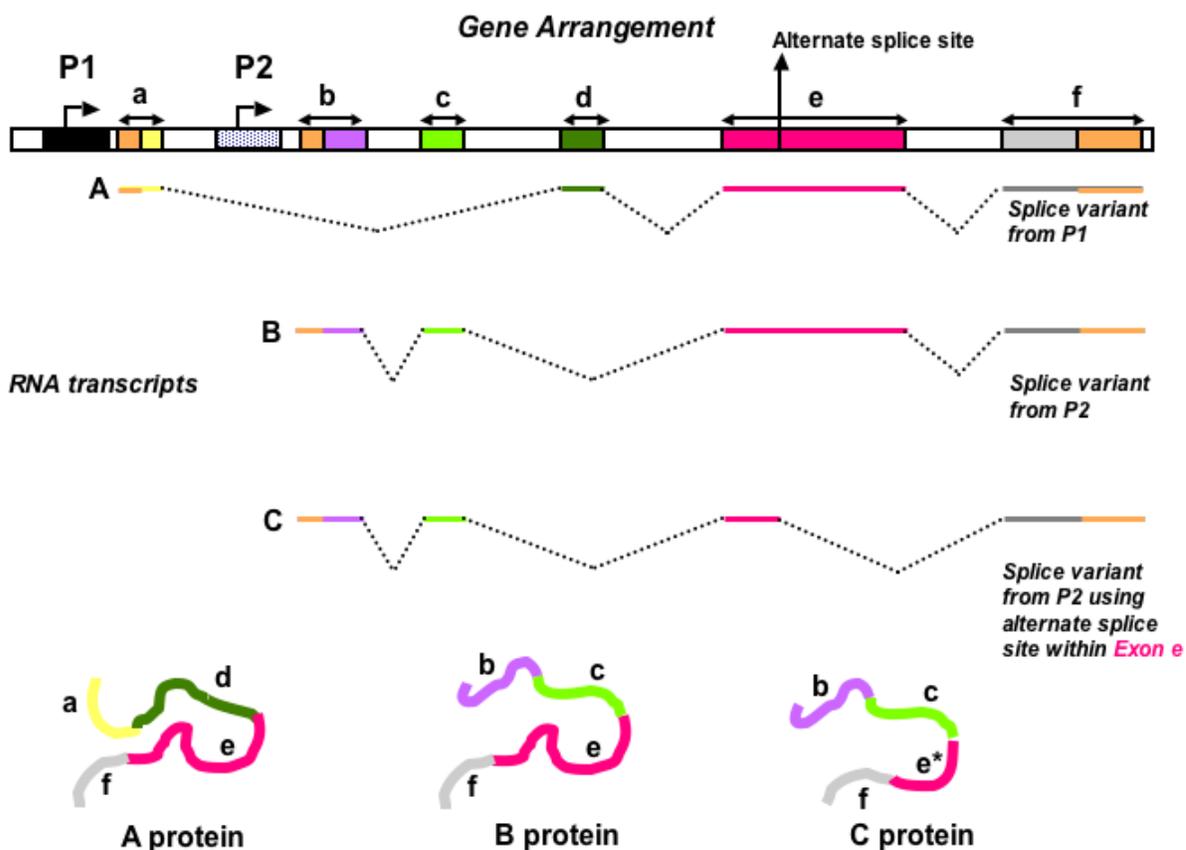


Fig. 9.Nb

▼ This is a minimum since each MITF isoform can also form alternate spliceforms with or without certain internal exons. Thus each protein isoform (directed by a separate promoter) itself can exist in at least two alternate forms.

Fig. 9. Nb. Gene arrangement with two promoters (P1 and P2), alternate splicing, and shared downstream coding regions. Upon transcription, gene exons at the DNA level are spliced into mature mRNA transcripts (involving splicing out of the intervening sequences of the primary transcript; RNA sequences co-colored as for corresponding genomic DNA segments). Each exonic sequence is denoted a-f as shown; untranslated 5' and 3' exon segments are shown in orange (that is, the segments of each exon which contain coding sequence are given distinct non-orange coloring). The transcript from promoter P1 here is spliced to contain exons a, d, e and f; from P2 the transcript contains b, c, e, and f. Dotted lines represent the regions of the primary transcript removed by splicing. With the simple rule that the first and last exons in a primary transcript must be included, and intervening exons can be skipped provided the original order is maintained, many splice variants are possible. For example, 8 possible variants can be obtained from the P1 promoter primary transcript if the b exon (the first transcribed from promoter P2) is excluded as a splice target. (a|c|d|e|f; a|c|d|f; a|c|e|f; a|c|f; a|d|e|f; a|d|f; a|e|f; a|f). But there are also precedents for exonic sequences possessing alternate splice sites which enable fusion of part of an exonic coding sequence with a downstream exon, as indicated with the alternative splice site within exon e. The resulting proteins (with shared C-terminal regions) expressed from the two promoters are also depicted at the bottom, with peptide segments color-matched with the corresponding exonic coding sequences (The truncated exon e is indicated as e*).

Biological functional compactness can certainly apply at the level of single molecules. Although the great majority of eukaryotic proteins are multi-domain^{47,48}, they vary in the degree that each of their residues participate in specific functions (the 'functional density' alluded to in the Evolving Proteins section of Chapter 2 of *Searching for Molecular Solutions*), whether at the level of protein-protein, protein-ligand, or catalytic interactions. As an example of how compactly diverse functions can be packed into a single protein, let's consider a protein of moderate size (81 kDa), the Large T antigen of SV40 virus. This protein has the power to transform cultured normal human cells into the unrestrained replicative state of cancer cells, and causes tumors in animals, and yet its study has been quite beneficial for the advancement of oncology, molecular and cellular biology, and the understanding of protein structure-function relationships.

The Largesse of Large T

The SV40 virus (simian virus-40) was discovered as a contaminant in ~30% of the Salk polio vaccine preparations used between 1955 and 1963, which had been prepared in monkey kidney cells⁴⁹. This was a very disturbing finding, given the subsequent demonstration of the ability of SV40 to transform human cells and generate tumors in rodents⁴⁹, but no associations between the vaccine usage and cancer rates have been made despite intensive studies[▼]⁵¹. The transforming property of SV40 was shown to be mediated by a viral protein termed Large T (T for tumor) antigen, in contrast to Small t antigen which shares a part of the Large T reading frame (Fig. 9.Nc; Table 9.N2)^{52,53}. Other animal viruses with transforming abilities have analogous proteins to Large T^{53,54}, but the SV40 protein is by far the best-studied. It is this large mine of information regarding Large T, which has by no means yet reached an end-point, which enables it to act as an excellent exemplar of how far protein functional packing may be taken.

▼ The issue of a direct association of SV40 with certain human tumors (that is, not linked to the vaccine in any way) is more controversial. Some studies have detected SV40 and/or Large T in significant percentages of some cancers (especially mesotheliomas and brain tumors⁵⁰), but this is not accepted as conclusive evidence⁵¹.

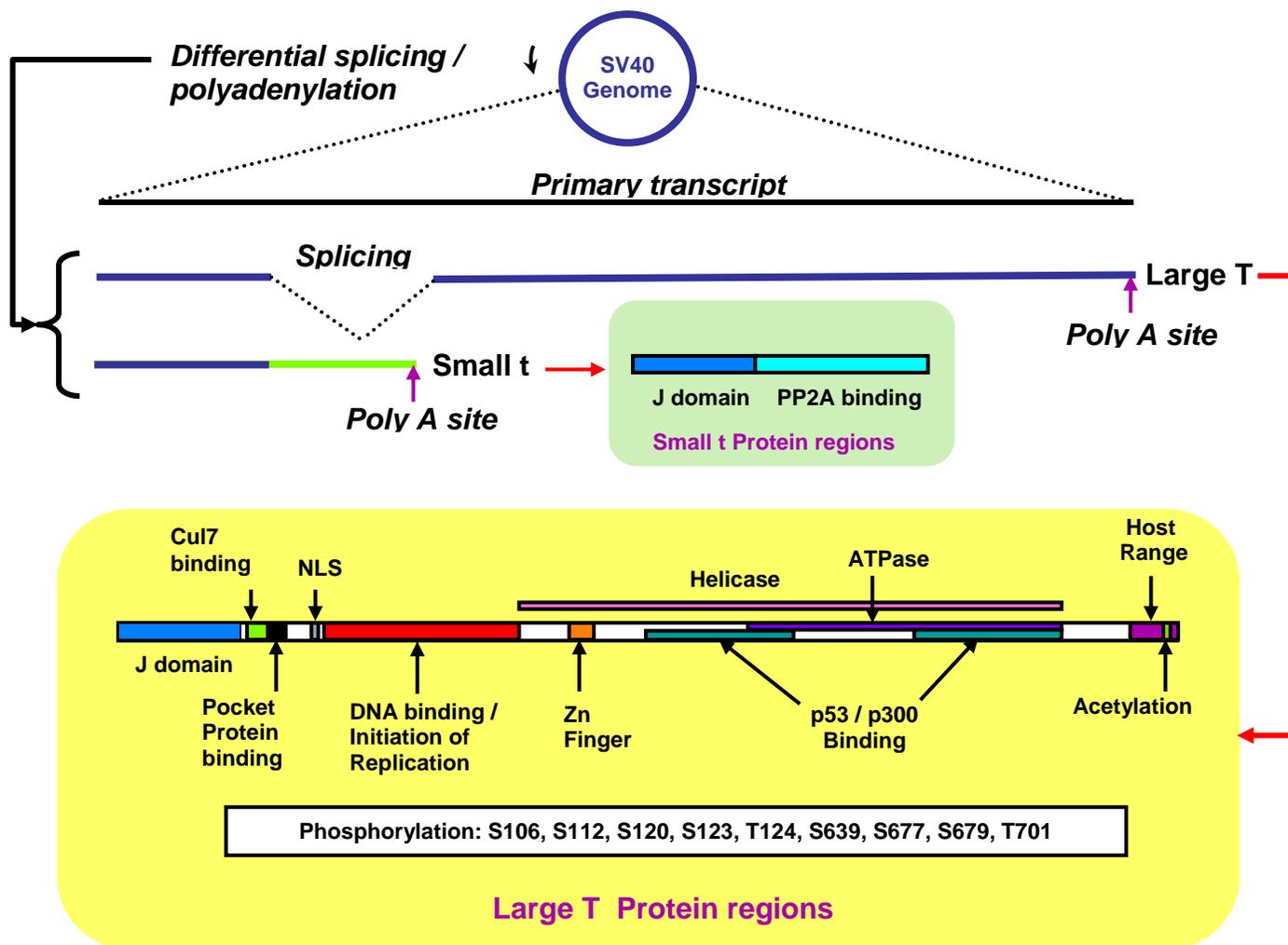


Fig. 9.Nc

Schematic depiction of the transcription and splicing of Large T vs. Small t, and the relative positions of their major functional features. For abbreviations and other designations, see footnotes to Table 9.N2 below.

Table 9.N2 SV40 Large T multi-functions		
Role / Effect	Specific Function	Insights
DNA replication of SV40	Helicase	Initiation of DNA replication Role of chaperones in eukaryotic DNA replication
	Single-stranded DNA binding	
	Specific DNA duplex binding	
	DNA polymerase α binding	
	ATPase	
	Molecular Chaperone	
Cellular transformation Anti-apoptosis	p53 Binding	Mechanism of cellular transformation Normal roles of tumor suppressor genes Decoy phosphodegron
	Pocket protein (pRb; p107; p130) binding	
	Transcriptional cofactor binding (p300; CBP)	
	Ubiquitin ligase pathways: Cul7 binding	
Cellular localization	Nuclear targeting signal	First demonstration of nuclear localization signals
Virion Assembly	J-domain chaperone function	(To be determined)
Host range	Virion assembly	(To be determined)
Protein multimers	Oligomerization	Protein Higher-order assembly; Zinc finger requirement
Other Functions / Post-translational modifications	Hsp90 binding Phosphorylation	Regulation of DNA binding
	Acetylation	p53-mediated; regulation of Large T activity?
Large T vs. small t	Small t: PP2A binding	Splicing
	Overlapping reading frame with Large T	Overlapping reading frames

Footnotes to Table 9.N2:

SV40 Large T functions. The overall functions are shown in the leftmost column, the specific subfunctions in the middle, and useful information derived from understanding these processes shown in the right-hand column.

Abbreviations and diagram key: Helicase; DNA unwinding activity required for commencing replication; ATPase: activity producing hydrolysis of adenosine triphosphate for energetics of replication; chaperone: a protein assisting other proteins to fold correctly (as noted in *Searching for Molecular Solutions*); p53: protein mediating protective cellular responses; Pocket proteins: a set of related proteins regulating the cell cycle, of which the retinoblastoma protein (pRb) was the first discovered; p300 and CBP: cofactors for transcription of many genes; ubiquitin: a small protein regulating protein turnover through its ligation to protein targets via ubiquitin ligases, one of which is Cul7; oligomerization: formation of multimeric forms of a single protein; PP2A; protein phosphatase 2A. *Derived from* ^{52,53,55-63}.

Although the analysis of Large T has proceeded since its discovery, with certain functions only recently discovered, even decades ago it was recognized that it was a remarkable protein: 'A Lot Packed into a Little' ⁶⁴ (see Fig. 9.Nc; Table 9.N2). Large T mediates the replication of the SV40 DNA genome and transformation of host cells, and assists with viral capsid (virion) assembly. It also acts as a substrate for post-translational modification, binds zinc ions (including in a zinc finger motif) determines host range for the virus, and targets to the nucleus. Large T has a hand, so to speak, in virtually the entire viral life-cycle.

Replication and transformation are complex operations which involve multiple sub-tasks, all performed by Large T (except where host proteins are co-opted and utilized). SV40 genome replication requires Large T for specific DNA binding at the viral origin of replication, chaperone activity (through the N-terminal J-domain), ATPase, and helicase ⁵³. The binding of host DNA polymerase α to Large T ⁶⁵ is also important for replication to proceed. Since it is an advantage for SV40 if its host cells are driven into active cycling, by means of Large T the virus targets key proteins involved in cell cycle regulation. These include the indicated 'pocket' proteins (especially the retinoblastoma protein [pRb]) ⁶⁶, and p53 ⁶⁷. The latter is a transcription factor which upon activation

can arrest the cell cycle or promote apoptosis (programmed cell death), and as such p53 has been considered a genomic 'guardian'⁶⁸ by preventing abnormal cell growth. The sequestration of pRb and p53 by Large T are then key steps in producing unrestrained cell division which can lead to tumorigenesis. However, other Large T functions are also linked with cell transformation, such as binding of ubiquitin ligase pathway proteins⁶³ and transcriptional cofactors⁶⁹.

At least some Large T functions (such as its ATPase) appear to be dependent on the protein undergoing oligomerization (i.e., forming multimeric, specifically hexameric, higher-order complexes⁶⁰). Structures of oligomerizing Large T domains have been determined, including the Large T helicase segment⁷⁰ and Large T complexed with p53⁷¹. The C-terminal region of Large T controls the host range activity of the virus (SV40 with mutations in the host-range region will not replicate in some types of monkey cells⁶²) and is also required for virion assembly, in conjunction with the N-terminal chaperone-like function^{53,56}.

Large T undergoes post-translational modifications in the form of multiple phosphorylations across the protein⁵² and acetylation in the C-terminal region, although the latter is apparently not required for the determination of host-range⁶². Finally, the Small t antigen, which shares the same N-terminal region as Large T and employs the Large T intron as coding sequence (Fig. 9.Nc; Table 9.N2), contains a binding region for protein phosphatase 2A, which serves as a regulator of Large T activity⁵². Protein phosphatase 2A regulates other cellular effects including adhesion⁷², and has properties of a tumor suppressor^{73,74}.

This broad-brush portrait of Large T is only a superficial sketch, which is far from exhaustive in terms of the complexities of its biology. It was necessary, though, to provide a short description of Large T's activities in order to convey the extent of the packing of function into its 708 amino acid residues. Even an overview of this knowledge portrays Large T as a beautifully efficient molecular machine, dedicated to the replication of SV40. And we may presume that selective pressures favored this kind of high-level protein multi-tasking.

But does this presumption contradict other findings? In Chapter 2 of *Searching for Molecular Solutions* (in the context of gene duplications) it was noted that sometimes selective pressures favor subfunctionalization of the expressed products of duplicated genes. The *splitting* of functions (formerly performed by a single protein) into two can thus be a favorable outcome in some circumstances. If combining multiple functions in a single protein forces a sub-optimal compromise for each, then separating them into discrete molecules may be a beneficial arrangement. What does this have to say about the fitness of the polyfunctional Large T? Firstly, the nature of the specific functions in question must be significant, as some tasks are very likely to be more amenable to modularity (that is, the ability to operate as a discrete protein domain and resist perturbation from the remainder of the protein) than others [▼]. The functions performed by Large T may fall into this category. Secondly, in the case of viruses such as SV40, relegation of multiple functions to a single protein of moderate size may confer an overall viral fitness benefit (for example, by allowing maintenance of small genome size) which outweigh slight suboptimality in one or more of the protein's tasks.

But this argument is countered by the knowledge that other transforming viruses are known where equivalent functions to some of those observed in Large T are in fact split into separate molecules. Human adenoviruses express discrete proteins E1a (binding pRb) and E1b (binding p53) which are important for cellular transformation ^{77,78}, and thus there is no obvious reason why fusing these activities into a single protein should provide an advantage. (Although to further complicate matters, these adenoviral proteins themselves have multiple additional functional activities ⁷⁹⁻⁸¹). Where specific protein domains encapsulate defined activities, simply stringing them together with appropriate intervening flexible linker sequences (either by natural evolution or artificially) may allow them to be included in a contiguous polypeptide. This does account for some compartmentalization of Large T function, as with the DNA-binding domain and the helicase domain ⁷⁰, but not all. As shown in Fig. 9.Nc; Table 9.N2, some

[▼] Many DNA-binding domains and transcriptional activation domains are examples of modular functions, where they can be 'chopped and changed' to alter the function of a protein. This capacity is the basis of the yeast two-hybrid system ^{75,76} and its derivatives, which are powerful means for detecting protein-protein and protein-ligand interactions (see the file SMS-CitedNotes-Ch4/Section 9; from the same ftp site).

of Large T's functions are embedded in other functional regions (especially at the C-terminus), so this protein is not a simple concatenation of discrete functional domains. At least one part of Large T (an N-terminal segment encompassing the pRb-binding region) shows limited homology to a host protein (Pur \square ⁸²), but the evolutionary history of Large T has yet to be accounted for.

While it may not always be a favored evolutionary arrangement to turn a single protein into a veritable Swiss army knife, it may indeed be useful in some circumstances to artificially engineer a protein towards this kind of ideal of compactness. Thanks to Large T, and other multi-functional proteins, we are well aware that this functional packing is possible.

Section 31: ***Global Protein-Protein Interaction Screens***

Cited on p. 339 of *Searching for Molecular Solutions*

Global Screens for Protein-Protein Interactions and Other Aspects of the Interactome

A great many global screens for the protein-based interactome have used the two-hybrid system and related methods, as considered in a section of Cited Notes for Chapter 4 [▼]. As also noted in the same section, mass spectrometry has emerged as an alternative for global analyses of protein-protein and protein-ligand interactions ^{83,84}. By its nature, this technique discriminates samples based on their mass-to-charge ratios following ionization, and can be used to identify enzymatically-digested proteins by their corresponding peptide signatures. Using affinity-based purifications of tagged proteins, mass spectrometry has been used to define the yeast protein interactome ⁸⁵.

Beyond protein identification, a key advantage of mass spectrometry is the potential to characterize a wide range of metabolites. In this respect, mass spectrometric analyses are an advancing tool for studies of the lipidome ^{86,87} and glycome ^{88,89}. A plethora of protein modifications are also subject to characterization by this versatile method. Given the profound importance of phosphorylation for the transmission of biological signaling information, it should come as no surprise to note that mass spectrometry is of fundamental importance for phosphoproteomic analyses ⁹⁰. The glycoproteome ^{91,92} and lipoproteome ^{93,94} are also amenable to characterization by this general approach. Mass spectrometry has been applied towards studying conjugates of proteins with ubiquitin or related peptides ^{95,96}, and markers of inflammation and oxidation ^{97,98}. Histone modifications (especially methylation and acetylation) have a fundamental role in gene regulation and expression, and these too have been addressed by means of mass spectrometric technologies ^{99,100}. A particular point of

[▼] See the file SMS-CitedNotes-Ch4/Section 9; from the same ftp site.

interest with respect to the latter area, and one that is likely to have widespread ramifications, is the use of mass spectrometry to analyze combinatorial histone modification patterns^{101,102}. Where specific patterns of post-translational modifications with a limited set of functional groups have informational significance, this method promises to have a major impact in defining the relevant codes, and how they are utilized. And regulatory systems such as the 'histone code' by definition have a global impact in higher-order eukaryotic biosystems.

References:

1. Angenendt, P. Progress in protein and antibody microarray technology. *Drug Discov Today* **10**, 503-11 (2005).
2. Merkel, J. S., Michaud, G. A., Salcius, M., Schweitzer, B. & Predki, P. F. Functional protein microarrays: just how functional are they? *Curr Opin Biotechnol* **16**, 447-52 (2005).
3. Fewell, G. D. & Schmitt, K. Vector-based RNAi approaches for stable, inducible and genome-wide screens. *Drug Discov Today* **11**, 975-82 (2006).
4. Paddison, P. J. & Hannon, G. J. siRNAs and shRNAs: skeleton keys to the human genome. *Curr Opin Mol Ther* **5**, 217-24 (2003).
5. Cullen, L. M. & Arndt, G. M. Genome-wide screening for gene function using RNAi in mammalian cells. *Immunol Cell Biol* **83**, 217-23 (2005).
6. Paddison, P. J. RNA interference in mammalian cell systems. *Curr Top Microbiol Immunol* **320**, 1-19 (2008).
7. Castel, D., Pitaval, A., Debily, M. A. & Gidrol, X. Cell microarrays in drug discovery. *Drug Discov Today* **11**, 616-22 (2006).
8. Wheeler, D. B. et al. RNAi living-cell microarrays for loss-of-function screens in *Drosophila melanogaster* cells. *Nat Methods* **1**, 127-32 (2004).
9. Bailey, S. N., Ali, S. M., Carpenter, A. E., Higgins, C. O. & Sabatini, D. M. Microarrays of lentiviruses for gene function screens in immortalized and primary cells. *Nat Methods* **3**, 117-22 (2006).
10. Rausch, O. High content cellular screening. *Curr Opin Chem Biol* **10**, 316-20 (2006).
11. Haney, S. A., LaPan, P., Pan, J. & Zhang, J. High-content screening moves to the front of the line. *Drug Discov Today* **11**, 889-94 (2006).
12. Moffat, J., Reiling, J. H. & Sabatini, D. M. Off-target effects associated with long dsRNAs in *Drosophila* RNAi screens. *Trends Pharmacol Sci* **28**, 149-51 (2007).
13. Svoboda, P. Off-targeting and other non-specific effects of RNAi experiments in mammalian cells. *Curr Opin Mol Ther* **9**, 248-57 (2007).
14. Wei, Q., Marchler, G., Edington, K., Karsch-Mizrachi, I. & Paterson, B. M. RNA interference demonstrates a role for nautilus in the myogenic conversion of Schneider cells by daughterless. *Dev Biol* **228**, 239-55 (2000).
15. Choi, I. et al. Choice of the adequate detection time for the accurate evaluation of the efficiency of siRNA-induced gene silencing. *J Biotechnol* **120**, 251-61 (2005).
16. Westbrook, T. F. et al. A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121**, 837-48 (2005).
17. Ngo, V. N. et al. A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441**, 106-10 (2006).

18. Schlabach, M. R. et al. Cancer proliferation gene discovery through functional genomics. *Science* **319**, 620-4 (2008).
19. Uren, A. G., Kool, J., Berns, A. & van Lohuizen, M. Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24**, 7656-72 (2005).
20. Baum, C., Kustikova, O., Modlich, U., Li, Z. & Fehse, B. Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. *Hum Gene Ther* **17**, 253-63 (2006).
21. Haupt, Y., Alexander, W. S., Barri, G., Klinken, S. P. & Adams, J. M. Novel zinc finger gene implicated as myc collaborator by retrovirally accelerated lymphomagenesis in E mu-myc transgenic mice. *Cell* **65**, 753-63 (1991).
22. Uren, A. G. et al. Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133**, 727-41 (2008).
23. Collier, L. S. & Largaespada, D. A. Transposons for cancer gene discovery: Sleeping Beauty and beyond. *Genome Biol* **8 Suppl 1**, S15 (2007).
24. Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-10 (1997).
25. Izsvak, Z., Ivics, Z. & Plasterk, R. H. Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol* **302**, 93-102 (2000).
26. Collier, L. S. & Largaespada, D. A. Hopping around the tumor genome: transposons for cancer gene discovery. *Cancer Res* **65**, 9607-10 (2005).
27. Dupuy, A. J., Jenkins, N. A. & Copeland, N. G. Sleeping beauty: a novel cancer gene discovery tool. *Hum Mol Genet* **15 Spec No 1**, R75-9 (2006).
28. Luo, G., Ivics, Z., Izsvak, Z. & Bradley, A. Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* **95**, 10769-73 (1998).
29. Ding, S. et al. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-83 (2005).
30. Wang, W., Bradley, A. & Huang, Y. A piggyBac transposon-based genome-wide library of insertionally mutated Bln-deficient murine ES cells. *Genome Res* **19**, 667-73 (2009).
31. Harrington, J. J. et al. Creation of genome-wide protein expression libraries using random activation of gene expression. *Nat Biotechnol* **19**, 440-5 (2001).
32. Beltran, A., Liu, Y., Parikh, S., Temple, B. & Blancafort, P. Interrogating genomes with combinatorial artificial transcription factor libraries: asking zinc finger questions. *Assay Drug Dev Technol* **4**, 317-31 (2006).
33. Barak, M. et al. Evidence for large diversity in the human transcriptome created by Alu RNA editing. *Nucleic Acids Res* (2009).
34. Li, J. B. et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210-3 (2009).
35. Gorski, B. & Stringer, S. E. Tinkering with heparan sulfate sulfation to steer development. *Trends Cell Biol* **17**, 173-7 (2007).

36. Baldwin, R. J. et al. A developmentally regulated heparan sulfate epitope defines a subpopulation with increased blood potential during mesodermal differentiation. *Stem Cells* **26**, 3108-18 (2008).
37. Lindahl, U., Kusche-Gullberg, M. & Kjellen, L. Regulated diversity of heparan sulfate. *J Biol Chem* **273**, 24979-82 (1998).
38. Lamanna, W. C. et al. The heparanome--the enigma of encoding and decoding heparan sulfate sulfation. *J Biotechnol* **129**, 290-307 (2007).
39. Ori, A., Wilkinson, M. C. & Fernig, D. G. The heparanome and regulation of cell function: structures, functions and challenges. *Front Biosci* **13**, 4309-38 (2008).
40. Gabius, H. J. Biological information transfer beyond the genetic code: the sugar code. *Naturwissenschaften* **87**, 108-21 (2000).
41. Pilobello, K. T. & Mahal, L. K. Deciphering the glycode: the complexity and analytical challenge of glycomics. *Curr Opin Chem Biol* **11**, 300-5 (2007).
42. Kauffman, S. *The Origins of Order. Self-Organization and Selection in Evolution* (Oxford University Press, 1993).
43. Zuckerkandl, E. Polite DNA: functional density and functional compatibility in genomes. *J Mol Evol* **24**, 12-27 (1986).
44. Tachibana, M. MITF: a stream flowing for pigment cells. *Pigment Cell Res* **13**, 230-40 (2000).
45. Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol Med* **12**, 406-14 (2006).
46. Hershey, C. L. & Fisher, D. E. Genomic analysis of the Microphthalmia locus and identification of the MITF-J/Mitf-J isoform. *Gene* **347**, 73-82 (2005).
47. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300**, 1701-3 (2003).
48. Orengo, C. A. & Thornton, J. M. Protein families and their evolution-a structural perspective. *Annu Rev Biochem* **74**, 867-900 (2005).
49. Butel, J. S. & Lednicky, J. A. Cell and molecular biology of simian virus 40: implications for human infections and disease. *J Natl Cancer Inst* **91**, 119-34 (1999).
50. Vilchez, R. A. & Butel, J. S. Emergent human pathogen simian virus 40 and its role in cancer. *Clin Microbiol Rev* **17**, 495-508, table of contents (2004).
51. Poulin, D. L. & DeCaprio, J. A. Is there a role for SV40 in human cancer? *J Clin Oncol* **24**, 4356-65 (2006).
52. Fanning, E. Simian virus 40 large T antigen: the puzzle, the pieces, and the emerging picture. *J Virol* **66**, 1289-93 (1992).
53. Sullivan, C. S. & Pipas, J. M. T antigens of simian virus 40: molecular chaperones for viral replication and tumorigenesis. *Microbiol Mol Biol Rev* **66**, 179-202 (2002).
54. Pipas, J. M. Common and unique features of T antigens encoded by the polyomavirus group. *J Virol* **66**, 3979-85 (1992).
55. Loeber, G. et al. The zinc finger region of simian virus 40 large T antigen is needed for hexamer assembly and origin melting. *J Virol* **65**, 3167-74 (1991).

56. Spence, S. L. & Pipas, J. M. Simian virus 40 large T antigen host range domain functions in virion assembly. *J Virol* **68**, 4227-40 (1994).
57. Miyata, Y. & Yahara, I. p53-independent association between SV40 large T antigen and the major cytosolic heat shock protein, HSP90. *Oncogene* **19**, 1477-84 (2000).
58. Ali, S. H. & DeCaprio, J. A. Cellular transformation by SV40 large T antigen: interaction with host proteins. *Semin Cancer Biol* **11**, 15-23 (2001).
59. Ali, S. H., Kasper, J. S., Arai, T. & DeCaprio, J. A. Cul7/p185/p193 binding to simian virus 40 large T antigen has a role in cellular transformation. *J Virol* **78**, 2749-57 (2004).
60. Gai, D. et al. Insights into the oligomeric states, conformational changes, and helicase activities of SV40 large tumor antigen. *J Biol Chem* **279**, 38952-9 (2004).
61. Ahuja, D., Saenz-Robles, M. T. & Pipas, J. M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24**, 7729-45 (2005).
62. Poulin, D. L. & DeCaprio, J. A. The carboxyl-terminal domain of large T antigen rescues SV40 host range activity in trans independent of acetylation. *Virology* **349**, 212-21 (2006).
63. Welcker, M. & Clurman, B. E. The SV40 large T antigen contains a decoy phosphodegron that mediates its interactions with Fbw7/hCdc4. *J Biol Chem* **280**, 7654-8 (2005).
64. Livingston, D. M. & Bradley, M. K. The simian virus 40 large T antigen. A lot packed into a little. *Mol Biol Med* **4**, 63-80 (1987).
65. Gannon, J. V. & Lane, D. P. p53 and DNA polymerase alpha compete for binding to SV40 T antigen. *Nature* **329**, 456-8 (1987).
66. DeCaprio, J. A. et al. SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. *Cell* **54**, 275-83 (1988).
67. Kierstead, T. D. & Tevethia, M. J. Association of p53 binding and immortalization of primary C57BL/6 mouse embryo fibroblasts by using simian virus 40 T-antigen mutants bearing internal overlapping deletion mutations. *J Virol* **67**, 1817-29 (1993).
68. Lane, D. P. Cancer. p53, guardian of the genome. *Nature* **358**, 15-6 (1992).
69. Eckner, R. et al. Association of p300 and CBP with simian virus 40 large T antigen. *Mol Cell Biol* **16**, 3454-64 (1996).
70. Li, D. et al. Structure of the replicative helicase of the oncoprotein SV40 large tumour antigen. *Nature* **423**, 512-8 (2003).
71. Lilyestrom, W., Klein, M. G., Zhang, R., Joachimiak, A. & Chen, X. S. Crystal structure of SV40 large T-antigen bound to p53: interplay between a viral oncoprotein and a cellular tumor suppressor. *Genes Dev* **20**, 2373-82 (2006).
72. Sontag, J. M. & Sontag, E. Regulation of cell adhesion by PP2A and SV40 small tumor antigen: an important link to cell transformation. *Cell Mol Life Sci* **63**, 2979-91 (2006).
73. Arroyo, J. D. & Hahn, W. C. Involvement of PP2A in viral and cellular transformation. *Oncogene* **24**, 7746-55 (2005).
74. Janssens, V., Goris, J. & Van Hoof, C. PP2A: the expected tumor suppressor. *Curr Opin Genet Dev* **15**, 34-41 (2005).

75. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6 (1989).
76. Chien, C. T., Bartel, P. L., Sternglanz, R. & Fields, S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A* **88**, 9578-82 (1991).
77. Endter, C. & Dobner, T. Cell transformation by human adenoviruses. *Curr Top Microbiol Immunol* **273**, 163-214 (2004).
78. Berk, A. J. Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus. *Oncogene* **24**, 7673-85 (2005).
79. Sang, N. & Giordano, A. Extreme N terminus of E1A oncoprotein specifically associates with a new set of cellular proteins. *J Cell Physiol* **170**, 182-91 (1997).
80. Sang, N., Caro, J. & Giordano, A. Adenoviral E1A: everlasting tool, versatile applications, continuous contributions and new hypotheses. *Front Biosci* **7**, d407-13 (2002).
81. Lavia, P., Mileo, A. M., Giordano, A. & Paggi, M. G. Emerging roles of DNA tumor viruses in cell proliferation: new insights into genomic instability. *Oncogene* **22**, 6508-16 (2003).
82. Gallia, G. L., Johnson, E. M. & Khalili, K. Puralpha: a multifunctional single-stranded DNA- and RNA-binding protein. *Nucleic Acids Res* **28**, 3197-205 (2000).
83. Smith, R. D. Advanced mass spectrometric methods for the rapid and quantitative characterization of proteomes. *Comp Funct Genomics* **3**, 143-50 (2002).
84. Wingren, C., James, P. & Borrebaeck, C. A. Strategy for surveying the proteome using affinity proteomics and mass spectrometry. *Proteomics* **9**, 1511-7 (2009).
85. Krogan, N. J. et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637-43 (2006).
86. Han, X. & Gross, R. W. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. *J Lipid Res* **44**, 1071-9 (2003).
87. Ejsing, C. S. et al. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. *Proc Natl Acad Sci U S A* **106**, 2136-41 (2009).
88. Haslam, S. M., North, S. J. & Dell, A. Mass spectrometric analysis of N- and O-glycosylation of tissues and cells. *Curr Opin Struct Biol* **16**, 584-91 (2006).
89. Satomaa, T. et al. The N-glycome of human embryonic stem cells. *BMC Cell Biol* **10**, 42 (2009).
90. Lemeer, S. & Heck, A. J. The phosphoproteomics data explosion. *Curr Opin Chem Biol* (2009).
91. North, S. J., Hitchen, P. G., Haslam, S. M. & Dell, A. Mass spectrometry in the analysis of N-linked and O-linked glycans. *Curr Opin Struct Biol* (2009).
92. Zhao, J., Patwa, T. H., Pal, M., Qiu, W. & Lubman, D. M. Analysis of protein glycosylation and phosphorylation using liquid phase separation, protein microarray technology, and mass spectrometry. *Methods Mol Biol* **492**, 321-51 (2009).
93. Wan, J., Roth, A. F., Bailey, A. O. & Davis, N. G. Palmitoylated proteins: purification and identification. *Nat Protoc* **2**, 1573-84 (2007).

94. Hoofnagle, A. F. & Heinecke, J. W. Lipoproteomics: using mass spectrometry-based proteomics to explore the assembly, structures, and functions of lipoproteins. *J Lipid Res* (2009).
95. Jeram, S. M., Srikumar, T., Pedrioli, P. G. & Raught, B. Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites. *Proteomics* **9**, 922-34 (2009).
96. Wohlschlegel, J. A. Identification of SUMO-conjugated proteins and their SUMO attachment sites using proteomic mass spectrometry. *Methods Mol Biol* **497**, 33-49 (2009).
97. Bigelow, D. J. & Qian, W. J. Quantitative proteome mapping of nitrotyrosines. *Methods Enzymol* **440**, 191-205 (2008).
98. Vivekanadan-Giri, A., Wang, J. H., Byun, J. & Pennathur, S. Mass spectrometric quantification of amino acid oxidation products identifies oxidative mechanisms of diabetic end-organ damage. *Rev Endocr Metab Disord* **9**, 275-87 (2008).
99. Rathert, P., Dhayalan, A., Ma, H. & Jeltsch, A. Specificity of protein lysine methyltransferases and methods for detection of lysine methylation of non-histone proteins. *Mol Biosyst* **4**, 1186-90 (2008).
100. Basu, A. et al. Proteome-wide prediction of acetylation substrates. *Proc Natl Acad Sci U S A* (2009).
101. Jiang, L. et al. Global assessment of combinatorial post-translational modification of core histones in yeast using contemporary mass spectrometry. LYS4 trimethylation correlates with degree of acetylation on the same H3 tail. *J Biol Chem* **282**, 27923-34 (2007).
102. Young, N. L. et al. High-throughput characterization of combinatorial histone codes. *Mol Cell Proteomics* (2009).