Review

Cell PRESS

Genome-wide approaches for cancer gene discovery

Paul M. Lizardi[1,2], Matteo Forloni[1] and Narendra Wajapeyee[1]

[1] Department of Pathology, Yale University School of Medicine, New Haven, CT 06520-8023, USA
[2] Yale Computational Biology and Bioinformatics, New Haven, CT, 06520-8023, USA

One of the central aims of cancer research is to identify and characterize cancer-causing alterations in cancer genomes. In recent years, unprecedented advances in genome-wide sequencing, functional genomics technologies for RNA interference screens and methods for evaluating three-dimensional chromatin organization *in vivo* have resulted in important discoveries regarding human cancer. The cancer-causing genes identified from these new genome-wide technologies have also provided opportunities for effective and personalized cancer therapy. In this review, we describe some of the most recent technologies for cancer gene discovery. We also provide specific examples in which these technologies have proven remarkably successful in uncovering important cancer-causing alterations.

Introduction

Although cancer research has generated an impressive body of knowledge about this highly diverse class of diseases, our understanding of neoplasia at the systems level remains in its infancy. Impressive advances in structural biology and protein interactomes are leading to elucidation of the three-dimensional structures of key protein molecules with important roles in cancer and identification of the logic of protein interaction and modification networks. However, the field of genomics remains far from the lofty goal of defining chromatin structural logic and dynamics specific to cancer and their relation to the abnormal transcriptional events that define disrupted regulatory states in cancer genomes. In this context, elucidation of cancer genomics represents a daunting challenge, since cancer genomes are not static but instead exist in dynamic and flexible evolutionary trajectories, characterized by mutation and structural rearrangements made even richer by lineage-specific epigenetic variegation and functional state mosaicism. Thus, a systems approach to cancer functional genomics requires dramatic improvements in the molecular analysis of tumor genomes and transcriptomes. Fortunately, the ongoing revolution in DNA sequencing and chromatin analysis technologies are already providing the scientific community with an arsenal of powerful tools to meet this challenge.

Here we outline and briefly review several examples of high-throughput methodological approaches that have emerged over the last few years and that hold promise for rapid progress towards a systems view of genomics and transcriptomics in cancer cells (Figure 1). Most of these methodologies are already in use for cancer genome analyses and have led to the identification of multiple cancer-causing genes, genetic alterations and deregulated pathways. The availability of a reference human genome implied that DNA sequencing could become the main tool for the exploration of cancer genomes. High-throughput ChIP-seq (see Glossary) and RNA-seq techniques are in principle capable of documenting most chromatin modification

Glossary

ChIP sequencing (ChIP-seq): high-throughput methodology that combines ChIP with massively parallel sequencing. Use of the latter approach for analysis of DNA segments yields greater and deeper ChIP-seq coverage in a shorter time compared to the traditional ChIP technique.
Chromatin immunoprecipitation (ChIP): technique for immunoprecipitation-based enrichment of a specific protein crosslinked to genomic DNA. ChIP has been adapted to perform genome-wide analyses of the occupancy of DNA-binding proteins and has led to a better understanding of proteins that regulate gene expression by binding to DNA.
Chromosome conformation capture (3C): methodology used to analyze the spatial organization of chromosomes. This facilitates evaluation of the existence of long-range chromatin interactions. 3C has been adapted for genome-wide surveys and has led to the development of technologies such as 4C, 5C, Hi-C and ChIP-PET.
Dropout RNAi screen: in contrast to positive selection screens, dropout RNAi screens identify loss of proliferation or survival as an end point. For example, an RNAi screen that identifies essential housekeeping genes in a given cell type would require a negative selection approach.
In vivo RNAi screens: most biological processes occur within the context of an organism and are the outcome of interactions between multiple cell types, so significant effort has recently been devoted to applying *in vitro* RNAi screening techniques to whole organisms to achieve results that might be relevant in an *in vivo* context. Set-up of RNAi screens *in vivo* is relatively less cumbersome in planarian worms and *C. elegans* than in a mouse model.
Massive parallel sequencing: powerful cell-free sequencing method that can read millions of bases of DNA in a few hours. This methodology is faster, more accurate and significantly less expensive compared to traditional methods such as capillary electrophoresis.
Mosaic mouse models: murine models in which genetically altered stem cells are retrotransplanted into recipient mice to produce clusters of mutated cells surrounded by normal counterparts. Development of a mosaic mouse model is much faster and less expensive compared to traditional knockout generation methodologies. The overall benefit of these mosaic models lies in their suitability for testing cooperation between multiple genetic lesions.
Positive-selection RNAi screen: screen in which cells containing a specific shRNA that endows them with a growth advantage survive and proliferate in a given assay. An example is a screen that identifies genes for which knockdown rescues the growth inhibitory effects of a tumor suppressor.
RNA sequencing (RNA-seq): powerful methodology used to sequence cDNA generated from cellular mRNA to gain insight into the complete transcriptome of a given cell. RNA-seq provides much deeper coverage compared to microarray-based methodologies and can yield up to single-base resolution. It is thus an invaluable tool for determining the cancer-specific transcriptome.
Synthetic lethality RNAi screen: synthetic lethality refers to a situation in which a genetic defect leads to cell lethality only in combination with a second genetic change, whereas neither of the genetic changes alone is lethal.

Corresponding authors: Lizardi, P.M. (paul.lizardi@yale.edu); Wajapeyee, N. (narendra.wajapeyee@yale.edu).

0167-7799/$ – see front matter doi:10.1016/j.tibtech.2011.06.003 Trends in Biotechnology xx (2011) 1–11

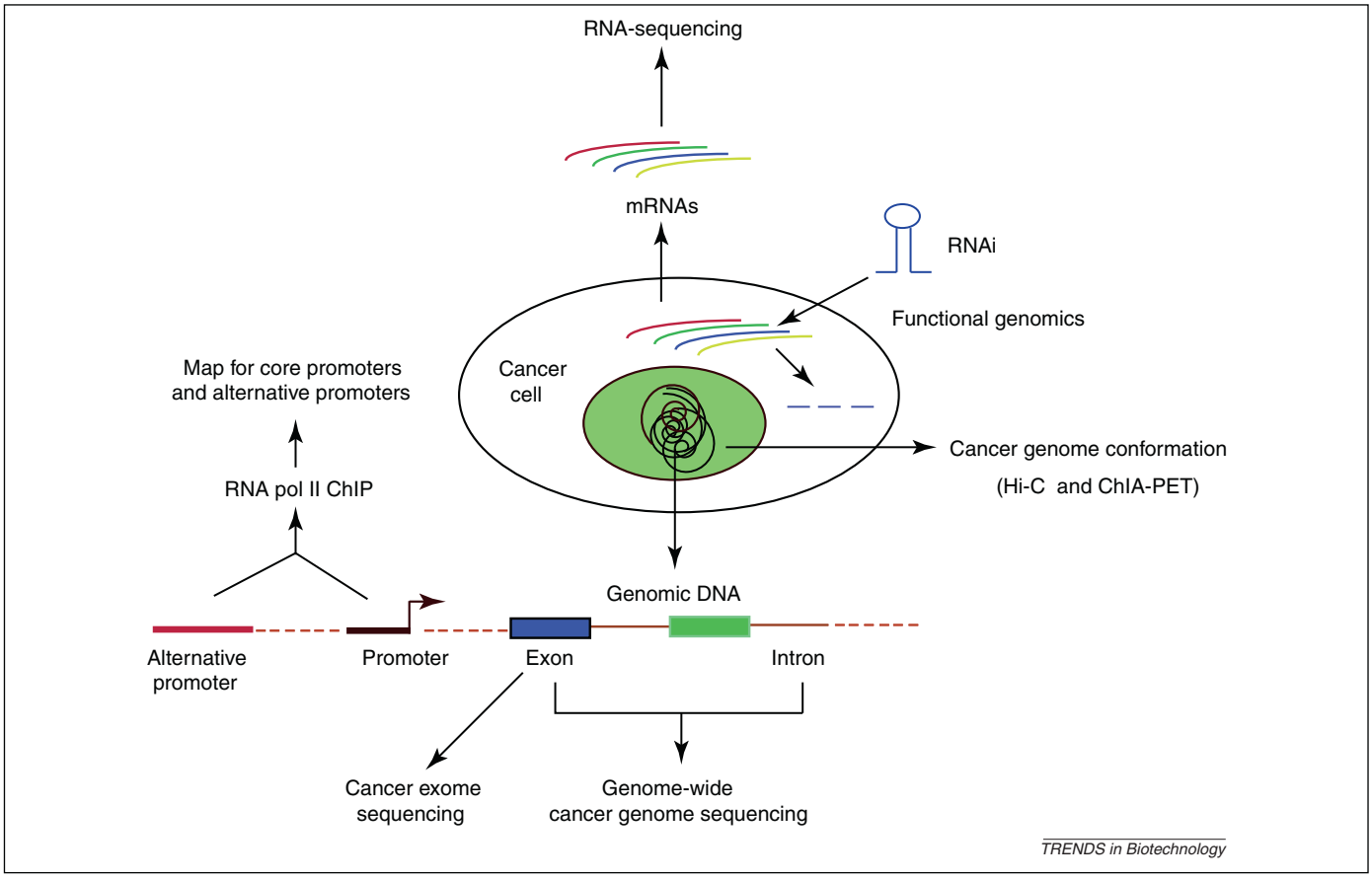

Figure 1. Schematic of the technologies for cancer genome analyses and functional characterization. The figure shows different experimental approaches that can be used to gain genome-wide insight into changes associated with cancer cells. Methodologies such as RNAi help in both identifying new cancer genes and functionally validating them as important regulator of tumorigenesis.

events and virtually all of the transcriptional events in a tumor cell. Many of the ongoing cancer genome projects incorporate the methods that we describe in this review (Box 1). Our aim is to provide the reader with a basic understanding of these contemporary, cutting-edge methodologies. In Table 1 we summarize the methods described in this review and provide some guidelines that will help researchers to identify a particular technology described here that might be suitable for their research problem.

High-throughput sequencing of cancer genomes and their value for understanding cancer

During tumorigenesis normal cells undergo complex genetic and epigenetic changes to become cancerous [1–3].

Box 1. Some noteworthy initiatives for cancer genome analyses and characterization

Cancer Genome Anatomy Project (CGAP) (http://cgap.nci.nih.gov/cgap.html)

The aim of the National Cancer Institute (NCI) CGAP is to determine gene expression profiles for normal, precancerous and cancerous cells to facilitate improved detection, diagnosis and treatment for all patients. The resources generated by CGAP are available to a broad research community. Interconnected modules provide access to all CGAP data, bioinformatics analysis tools and biological resources, so that users can find *in silico* answers to biological questions.

Cancer Genome Characterization Initiative (CGCI) (http://cgap.nci.nih.gov/cgci.html)

The CGCI assesses the utility of new genomics technologies used in strategically characterizing a subset of genomic changes involved in different tumors. Research groups involved with the CGCI make all of their data available through a publicly accessible database. CGCI incorporates genomic characterization methods including exome and transcriptome analysis using second-generation sequencing technologies to facilitate a better understanding the underlying genetic changes that lead to cancer.

The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/)

The NCI and the National Human Genome Research Institute (NHGRI) launched the TCGA project to create a comprehensive atlas of genomic changes involved in more than 20 common types of cancer. These analyses will be carried out over the next 5 years. This large-scale, high-throughput effort is being carried out by a network of more than 100 researchers at different organizations across the USA. The long-term aims of TCGA, similar to those of other cancer genome analyses initiatives, are to improve the technologies for cancer diagnosis and to develop new methods for treating cancers.

International Cancer Genome Consortium (ICGC) (http://www.icgc.org/)

The main ICGC goal is to perform comprehensive analysis on 50 different tumor types or subtypes to gain insights into genomic, transcriptomic and epigenomic changes associated with these tumor types. Some 22 countries and 120 research groups are participating in ICGC studies. The data generated through this initiative will be freely available to researchers.

TIBTEC-911; No. of Pages 11

Table 1. Genome-wide methods for cancer gene discovery

Method	Applications	Requirements	General considerations for use
Genome-wide deep sequencing	Suitable for comprehensive human cancer profiling applications, such as identification of mutations, SNPs, deletions, amplifications, copy number aberrations (CNAs) and gene fusion events	Large data storage space and specialized bioinformatics support	For specific needs, less resource-intensive alternative approaches can be explored. For example, SNP arrays can provide information on SNPs, CNAs, deletions and amplifications in the cancer genome. Compared to a genome-wide deep sequencing approach, SNP arrays are less expensive and require only limited bioinformatics support
Exome sequencing	Suitable for targeted sequencing of coding regions of the genome. Similar to genome-wide deep sequencing, exome sequencing can identify mutations, SNPs, deletions, amplification, and CNAs	Specific methods to enrich the target sequences (Box 2, Table 2) Large data storage space and specialized bioinformatics support Less resource intensive compared to whole-genome sequencing, only ~1% of the entire genome needs to be sequenced	If information is only required for coding regions, this might be a method of choice. However, if only a small number of genes are being analyzed (e.g. 50 genes) in multiple cancer samples, these can still be deep-sequenced after targeted PCR amplification and pooling of PCR products. This approach might substantially reduce the time required for sample preparation. This will also be the most cost-effective approach in terms of sample preparation, deep sequencing and data analyses
Genome-wide RNAi screens	Suitable for functional analysis of any gene through a loss-of-function method. An RNAi methodology can be adapted for positive selection, negative selection and synthetic lethality screens. In recent years, mouse model-based *in vivo* RNAi screens have also been performed (Box 3)	siRNA-based or shRNA-based high content screens might require robotics and automation Limited robotics required for pooling of shRNA screens Dropout, synthetic lethality and *in vivo* screens may require deep-sequencing-based approaches or barcode microarrays to deconvolute the results	In several instances for which preliminary results point to a specific pathway(s) of interest, small-scale RNAi screens targeting a limited number of genes can be performed with limited resources to address the biological problem
ORFome screens	Suitable for evaluating gain-of-function events in cancer cells identified by genome-wide or targeted sequencing approaches (Box 4)	Might require specific cDNA library construction depending on the tumor type analyzed and the biological question posed	ORFome screens are as powerful as siRNA and shRNA screens. However, they should be carefully designed to avoid loss of evaluation of rare oncogenic changes (Box 4)
ChIP-seq	Can be used for global DNA–protein interaction analyses Specifically useful for identifying unbiased transcription factor binding sites and alternative promoters and studying histone modifications	Good-quality antibodies for ChIP, large data storage space and strong bioinformatics support	ChIP-on-ChIP, which requires limited bioinformatics support and resources, can be first explored as an alternative to ChIP-seq. With advances in array technologies, high-density arrays that provide almost genome-wide coverage are now available and might suffice in many cases
RNA-seq	Useful for studying gene transcription regulation or RNA processing Particularly suitable for identifying rare splicing events and alternative splicing and global changes in RNA editing	Specific pipeline for data analysis, limited to transcribed sequences Might require special library preparation protocols when both mRNAs and small RNAs need to be sequenced	Although very useful for identifying splicing events and rare transcriptomics changes, should not be used an as alternative for gene expression arrays
ChIA-PET	ChIP-based 3D conformation analysis Useful for long-range interaction studies	Deep sequencing and a specialized pipeline for sequence analysis and specific antibodies suitable for ChIP	If there are indications that particular genomic loci are involved in long-range interaction, a more directed analysis such as 3C might suffice
Hi-C	Advanced version of chromosome conformation capture (3C) Can detect global chromatin conformation in any given cell type Unlike ChIA-PET, Hi-C is not antibody-based and is therefore completely unbiased	Specific bioinformatics analysis pipeline	If there are indications that particular genomic loci are involved in long-range interaction, a more directed analysis such as 3C might suffice
CAGE	Can be used to map transcription start sites induced by a transcription factor in the same cancer samples	Strong bioinformatics support	Should not be used as a replacement for targeted strategies such as rapid amplification of cDNA ends (5′-RACE) when only a few genes are being analyzed

Early studies to evaluate the role of cancer-causing genes used hypothesis-driven approaches for candidate genes. These approaches relied on sequencing the gene of interest first, which led to identification of mutations or mutation hotspots within a gene. This was followed by characterization of an identified mutation regarding its functional impact on the protein of interest, eventually implicating both the mutation and the functional alteration in tumorigenesis. Successes of these candidate-based approaches are well represented by the identification of mutations in the tumor suppressor genes *p53* and *PTEN* [4–6]. Although targeted gene sequencing approaches are straightforward and inexpensive, they are biased and have low throughput. Furthermore, owing to the complex nature of human cancers, these candidate-based strategies fail to reveal the complete landscape of all the genomic changes that occur in cancer. Therefore, genome-wide sequencing methods have become the methods of choice for cancer genome analyses [7,8].

In recent years, several new studies have taken advantage of high-throughput deep-sequencing methodologies for human cancer profiling [8–15]. Massively parallel DNA sequencing methods provide opportunities to carry out genome-wide screening for point mutations, copy-number variation and rearrangements on a single platform. For example, Chiang *et al.* used high-throughput sequencing for high-resolution mapping of copy-number alterations. Identification of genome regions with copy number variation is a powerful method for revealing cancer-causing genes. This study analyzed three matched pairs of normal and cancer cell lines. The authors showed that a collection of ~14 million aligned sequence reads from human cell lines has comparable power to detect events as the current generation of DNA microarrays [16]. However, massive parallel sequencing showed over a two-fold higher precision for localizing breakpoints, typically, to within a ~1 kb region [16].

Genome-wide deep sequencing approaches and many other approaches that we describe later in this review are extremely effective in their discovery power and are applicable to any human cancer. However, they require specialized support for bioinformatics analyses of genomic data. Some sequencing platforms such as the Illumina genome analyzer have their own pipeline that provide semi-processed data, which then require further analyses on custom bioinformatics platforms. The results can then be used to map the sequence reads to genomic regions to gain insight into genetic changes associated with cancer cells. Therefore, before applying high-throughput sequencing approaches, it is of utmost importance to first establish bioinformatics methods for analysis of large amounts of sequencing data. A good way to become familiar with high-throughput sequencing data sets is to download some of the data sets published in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and to practice analysis of these data sets. This can yield a preliminary understanding of the type of analysis required for such research projects. Some of the genomic data analysis issues pertaining to cancer genomes are discussed in a recent review by Chin *et al.* [17]. Table 2 compares different sequencing platforms and genome-wide and targeted sequencing approaches.

Exome sequencing to identify cancer-causing genes

The term exome refers to the complete set of regions of the genome corresponding to mature mRNA, most of which encodes proteins. The human genome contains ~180,000 exons that translates into ~30 Mb of DNA sequences, which in total represent ~1% of the total genomic DNA content. Most of the genes implicated in human cancers are protein-coding genes, so exome sequencing is expected to reveal important cancer-associated changes. Because exome sequencing requires sequencing of only ~1% of the genome, it is therefore less expensive compared to sequencing of the whole cancer genome. However, exome sequencing poses a technical challenge because the genomic DNA sequence corresponding to exons must first be enriched before sequencing. A brief account of the target enrichment methods that can be used in combination with deep sequencing technologies is presented in Table 2 and Box 2. These methods have also been discussed by others [18–20].

Many groups are now applying targeted exome sequencing to detect cancer-associated mutations [13,21,22]. A good example of this approach for cancer gene discovery is a recent study in which exome sequencing was performed on eight ovarian clear-cell carcinoma (OCCC) cancer samples [13]. The study identified two previously unreported ovarian cancer genes: tumor suppressor *ARID1B*, a known chromatin modifier, and oncogene *PPP2R1A* [13]. The results of this genome-wide exome

Box 2. Methods for target enrichment for exome sequencing

PCR-based methods

PCR-based techniques are by far the simplest and most straightforward for target enrichment. A multiplex PCR can be used for large-scale target enrichment and subsequent sequencing. RainDance Technologies (http://www.raindancetechnologies.com) has adapted a PCR-based method for target enrichment in preparation for massively parallel sequencing.

Molecular inversion probes

Molecular inversion probes (MIPs) are single-stranded DNA sequences and are complementary to the target regions. MIPs have linker that can be used for PCR amplification. For exon capture applications, MIPs can be used in combination with DNA polymerase-based gap filling, thereby covering the intermediate regions for complete coverage of the target sequences.

Microarray-based methods

Microarray-based target enrichment methods depend on the nucleic acid hybridization principle. Solid-phase methods for target enrichment use DNA probes on a microarray platform. To perform solid-phase capture, the fragmented genome is first ligated to an adapter sequence, with subsequent hybridization and then elution. This leads to selective enrichment for the region of choice. The enriched genomic region can then be subjected to PCR amplification and used for deep sequencing of the enriched sequences.

In-solution methods

In-solution target enrichment methods, similar to solid-phase microarray-based methods, depend on the nucleic acid hybridization principle. However, in-solution target enrichment provides an opportunity to increase throughput, because it can enrich a greater number of samples compared to microarray-based methods in the same amount of time. This method is also less resource-intensive and provides cost-effective solutions for deep sequencing methodologies.

sequencing were confirmed by analyzing 42 additional OCCC samples; *ARID1A* and *PPP2R1A* were mutated in 57% and 7%, respectively, of the samples [13].

This and other similar studies have shown that, similar to genome-wide sequencing, exome sequencing can reveal novel features of cancer genome and could further our understanding of the role of protein-coding genes in tumorigenesis and tumor progression [13,21].

Genome-wide RNAi and cDNA library screens for functional genomic analyses in cancer

The discovery of RNA interference (RNAi) in *Caenorhabditis elegans* and its adaptation to mammalian cells revolutionized the way in which biological functions of a gene can be explored [23–26]. In 2001, two groups developed retroviral vector-based RNAi libraries that are capable of targeting several thousand human genes [27,28]. Using shRNA libraries, these groups demonstrated that unbiased genome-wide RNAi screens can be performed in mammalian cells and that RNAi screens constitute an extremely powerful functional genomics method for cancer gene discovery and validation [27,28]. Since then, RNAi libraries with several new features have been developed, including siRNA-based and lentiviral vector-based shRNA libraries [29,30]. Many viral vector-based libraries are also barcoded, so RNAi screens can be performed using pools of several thousand shRNAs. The results can later be deconvoluted to identify shRNAs using either barcode microarrays or deep-sequencing technologies, such as the Illumina genome analyzer [30–32]. Several shRNA and siRNA libraries are now available (Table 3). These libraries facilitated several genome-wide RNAi screens that have revealed new cancer genes and pathways [33–37]. Box 3 provides more details regarding RNAi screens and discusses two classes of RNAi screens in this section [38–40]. The first class is referred to as *in vivo* RNAi screens and can be performed using a mouse model of human tumorigenesis. Using such an RNAi screen, Zender *et al.* combined cancer genomics data with RNAi and a mosaic mouse model of hepatocellular carcinoma and identified tumor suppressor genes [38]. First, based on comparative genome hybridization (CGH) arrays of hepatocellular carcinomas, the authors shortlisted genes that were deleted in human samples of hepatocellular carcinoma. Following gene identification, pools of shRNAs targeting the genes identified from the CGH arrays were generated. These shRNA pools were then used to evaluate the tumor suppressive function of the targeted genes in a mouse model of hepatocellular carcinoma. This approach revealed 13 tumor suppressor genes, 12 of which were previously not implicated in cancer. This study is very important because it provides an *in vivo* platform for integrating cancer genome data with RNAi screens and mouse models of cancer. In the future, these types of screens will be extremely useful for functional characterization of genetic information obtained in several ongoing large-scale cancer genome analysis initiatives.

The second class of RNAi screens targets cancer cells with activating RAS mutations [39,40]. These screens are based on the principle of synthetic lethality and aim to develop new ways to target cancer by exploiting non-oncogenic addictions in cancer cells [41–43]. In the first such screen, Luo et al. used a genome-wide RNAi approach for isogenic DLD-1 cells with either oncogenic or wild-type *KRAS*. The cells were infected with six pools of ~13 000 shRNAs per pool in triplicate. The authors then analyzed the change in relative abundance of each shRNA by microarray hybridization to identify and compare the lethality signature of mutant and wild-type *KRAS* DLD-1 cells. This led to identification of shRNAs that were specifically synthetic lethal to RAS mutant cells. The authors also observed enrichment of shRNAs targeting genes with mitotic functions, which led to identification of a pharmacologically tractable pathway involving polo-like kinase 1 (PLK1).

In a synthetic lethality screen for *KRAS* mutant cancers, Scholl *et al.* used the RNAi Consortium shRNA library [40]. They used 5024 shRNAs targeting 1011 human genes, including the majority of known and putative protein kinase genes and a selection of protein phosphatase genes and known cancer-related genes [40]. The screen was performed in eight human cancer cell lines representing five different tumor types, as well as fibroblasts and immortalized human mammary epithelial cells (HMECs) [40]. This RNAi screen revealed that STK33 is required for the survival and proliferation of *KRAS* mutant cancer cells, which was then functionally validated.

The results of these studies showcase the power of unbiased genome-wide RNAi screens in identifying and validating cancer genes and revealing important genetic relationships among genes in cancer cells.

Complementary to RNAi screens, cDNA library-based screens (also known as ORFome screens) can also be performed to identify gain-of-function events in cancer. A brief description of ORFome screens and considerations when performing such screens are presented in Box 4. A good example of this type of screen is an interesting study carried out by Boehm *et al.* [44]. The authors performed an

Box 3. RNAi screens

siRNA- versus shRNA-based screens

Genome-wide RNAi approaches provide opportunities to perform unbiased loss-of-function screens. In general, RNAi screens use either small interfering RNAs (siRNAs) or short hairpin-based RNAs (shRNAs). siRNA-based screens are particularly suitable in cases for which desired phenotypes or the outcome of an assay can be observed in a short time frame without a requirement for several rounds of cell division. Examples of these RNAi screen types include identification of apoptosis regulators and transcription regulators. However, for some experiments that require several rounds of cell division before a phenotype can be observed, stable knockdown using shRNAs might be required. Examples of such RNAi screens include identification of genes required for cellular transformation.

Endpoint choices: survival versus high-content screens

RNAi screens can be set up with survival as the readout (examples include positive selection and dropout screens). Depending on the research question, a high-content RNAi screen might be more appropriate. High-content RNAi screens are usually set up in multiwell plates in which cells are monitored for morphological, molecular or transcriptional changes using fluorescence- or luminescence-based assays. Therefore, high-content screens should be performed using siRNAs targeting one gene per well in almost all cases. It should be noted that at a genome-wide scale these screens can be quite expensive and require substantial automation for set-up and end-point monitoring.

TIBTEC-911; No. of Pages 11

Table 2. Comparison of whole-genome sequencing and targeted sequencing approaches

General features	Whole-genome sequencing			Targeted sequencing		
Sequence requirements	None			Prior knowledge of the target regions		
Coverage	Entire genome			Selected genomic regions of interest		
Sequencing cost	~ $5000–10 000			For exome sequencing cost is around 15% of that of sequencing the whole genome		
Advantages and disadvantages	Shorter sample preparation time and complete full-genome coverage Higher sequencing cost			Lower sequencing costs and shorter sequencing time Higher sample preparation costs and longer time		
Platform	454 Roche	Illumina	SOLiD	In situ exon capture	MIP-based exon capture	PCR-based
System	Pyrosequencing	Solid-phase amplification	Sequencing-by-ligation chemistry	Array-based target capture	In-solution capture by molecular inversion probes	PCR-based target amplification
Throughput	100 Mb	1–1.5 Mb	1–4.5 Mb	On average 2–3 Mb; up to 34 Mb each array with the HD2 NimbleGen array[b]	>55,000 loci in a single assay by multiplex MIP	Hundreds to thousands of genomic loci in a single tube with the RainStorm platform[b]
Cost	~$5000–6000	~$5000–6000	~$5000–10 000	Medium	<10 samples, high; >100 samples, low	High
Applications	*De novo* genome sequencing	RIP-seq, ChIP-seq, transcriptome sequencing, expression profiling	Resequencing, expression profiling, structural rearrangement analysis with paired-end, large-insert libraries	Useful for a large number of genomic targets but low number of samples	Useful for a large number of genomic targets and many samples	Useful for small- or medium-scale studies
Run time[a]	8 h	3 or more days	3 or more days			
Read length	200–300 nt	30–40 nt	30–40 nt			
Multiplex level[c]				10^5–10^6	10^4–10^5	10^2–10^3
Sensitivity				~98.6%	>98%	~95%
Accuracy	~0.2% error rate	<1.5% error rate	~0.2% error rate			
Advantages and disadvantages				High level of multiplexing Limited resolution and specificity and high DNA input requirement (10–15 μg)	Direct sequencing without the need for shotgun library construction, high specificity and low DNA input requirement (200 ng) Low capture uniformity	Potentially compatible with any next-generation sequencing platform Individual oligonucleotide synthesis and large numbers of amplification reactions

[a]Time required for single run.

[b]RainStorm platform (http://www.raindancetechnologies.com); HD2 NimbleGen array (Roche, http://www.nimblegen.com/products/).

[c]Number of probes used for target enrichment in each assay.

TIBTEC-911; No. of Pages 11

Table 3. List of human and mouse shRNA/siRNA libraries available for genome-wide RNAi screens

Library	Supplier/Institute	Genes Targeted	Total shRNA or siRNA	shRNA or siRNA/gene	Targeted Organisms	Vector	Type of Library	Features
pSM2 Retroviral Library	Thermo Scientific	28,000 mouse genes/ 28,500 human genes	~61,000/ 81,500	~3	Human and mouse	pSM2	Retroviral	miRNA-based design; Puromycin selection marker
GIPZ Lentiviral Library	Thermo Scientific	''Entire mouse genome''/ ''Entire human genome''	62,000/ 62,000	~2	Human and mouse	pGIPZ	Lentiviral	RNA pol II promoter; Turbo GFP; Puromycin selection marker; Can infect non-dividing cells
TRIPZ Inducible shRNA Library	Thermo Scientific	~16,000 human annotated genes/ ~15,950 mouse annotated genes	~159,000	~4-5	Human and mouse	pLKO.1	Lentiviral	Human U6 promoter; Inducible; Puromycin; Can infect non-dividing cells
MISSION shRNA Library	Sigma-Aldrich	~16,000 human annotated genes/ ~15,950 mouse annotated genes	~159,000	~4-5	Human and mouse	pLKO.1	Lentiviral	Human U6 promoter; Puromycin; Can infect non-dividing cells
NKI Library	NKI	~8,000 human genes/ 15,000 mouse genes	24,000/ 30,000	~3 ~2	Human and mouse	pRSC	Retroviral	RNA Pol III promoter; Puromycin selection
GeneNet shRNA Library	System Biosciences	39,000 mouse genes/ 47,400 human genes	150,000/ 200,000	4	Human and mouse	HIV and FIV-based	Lentiviral	Fluorescent proteins such as GFP etc.; Puromycin selection
siGenome SMARTpool siRNA Library	Thermo Scientific (Dharmacon)	18,236 human genes		~4	Human		siRNA-based	Four mRNA regions tareted at once to reduce false negatives; Guaranteed 75% silencing
Silencer siRNA Libraries	Ambion	12,585 human genes/ 11,134 mouse genes	37,755/ 33,402	~3	Human and mouse		siRNA-based	Can be used in low concentration (>30nM); Off-target, polymorphic, and antiviral inducting regions have been eliminated

Box 4. Gain-of-function screens using cDNA libraries

Gain-of-function screens using cDNA libraries (also called ORFome screens) are complementary to RNAi screens and can be used to assess gain-of-function phenotypes in cancers. However, in contrast to siRNA/shRNA libraries, there is no standard cDNA library that can be used to address all types of gain-of-function questions. For example, a standard HeLa mammalian cell cDNA expression library cannot be used to assess gain-of-function phenotypes in lung adenocarcinoma with a specific gain of function mutation or with a transforming oncogene resulting from a fusion such as the *EML4-ALK* fusion gene.

Although ORFome screens pose more challenges in terms of assessing gain-of-function changes for a given tumor, they have discovery potential as powerful as that of RNAi screens once set up. One of the most important considerations when setting up ORFome screen is to first choose an appropriate tissue to generate a cDNA library. Next, the cDNA library should be normalized to ensure that high- and low-expression transcripts are represented equally in the library to prevent loss of the latter owing to under-representation. Finally, researchers trying to perform ORFome should realize that, depending on the size or abundance of a transcript, they could end up missing some gain-of-function events. However, shRNA screens also have the same drawback in that they might not be saturating because of a lack of shRNAs against a gene or an inability to cause gene knockdown.

ORFome screen to identify kinases that can substitute for myristylated-AKT (myr-AKT) in transforming cells in cooperation with constitutively active MEK (MEK^{DD}). They cloned 354 kinases and kinases-related open reading frames (ORFs) using myristyl and Flag tags [44]. After cloning and confirming the activity of these kinases, the authors performed a screen with pools of 10–12 unique ORFs to identify activated kinases that could substitute for myr-AKT and induce transformation in cooperation with activated MEK1. The pools that permitted anchorage-independent growth were then deconvoluted by testing the ORFs individually. Using this approach, the authors identified four kinases that can cooperate with activated MEK1 to cause anchorage-independent growth, one of which was IKKε. Further analysis revealed that IKKε is amplified and overexpressed in breast cancer. Furthermore, it is required for survival of breast cancer cells, because knockdown of IKKε leads to apoptosis induction. This study provides a framework for performing integrative genomics analyses to identify gain-of-function events [44]. Eventually, in combination with RNAi-based approaches, this methodology should enable researchers to assess the functional consequences of cancer cell-associated changes identified in genome-wide analyses.

Non-canonical cancer genes emerging from alternative promoters and alternative splicing

Promoters are DNA elements that facilitate recruitment of RNA polymerase, which leads to establishment of a specific transcriptional state [45]. Human genes can have one or multiple promoters, many of which can act as alternative promoters [46,47]. Alternative promoters not only provide opportunities for differential gene expression, but can also influence the relative amounts of alternative transcripts produced, as well as their stability [46,47]. Genome-wide promoter analyses studies have indicated that over 50% of human genes have at least one alternative promoter [48,49]. Interestingly, disease-associated genes are more likely to be associated with more alternative promoters and tend to be differentially expressed [50,51]. Furthermore, a recent study has indicated that cancer-related genes have on average two alternative promoters, compared to 1.5 for other human genes [50], which indicates that cancer cells might rely in part on the use of alternative promoters for tumorigenesis and even for maintenance of the tumorigenic state.

A recent study of Hodgkin's lymphoma revealed that loss of the transcriptional repressor CBFA2T3 leads to reactivation of a long terminal repeat (LTR) of mammalian apparent LTR retrotransposon (MaLR) family member THE1B, which then functions as an alternative promoter. THE1B then leads to transcriptional upregulation of protooncogene colony-stimulating factor 1 receptor (*CSF1R*) in Hodgkin's lymphoma cells, which is required for their survival [52]. This study highlights the importance of analyzing the cancer transcriptome to identify the emergence of cancer cell-specific alternative promoters and evaluate their role in cancer.

Several methodologies have been developed in the last few years to identify promoters at genome-wide scale and to quantify their specific usage [53–57]. Here we describe one of these methods and evaluate its ability to identify alternative promoters and the possibility of using it for cancer-specific alternative promoter discovery. Sun et al. used RNA pol II ChIP followed by ChIP-seq for five different mouse tissues: brain, liver, lung, spleen and kidney [54]. Their analysis revealed 38 639 promoters, 12 270 of which were novel promoters, that included both protein-coding and non-coding genes. Furthermore, by identifying the RNA pol II-bound promoter(s) of each annotated gene in a given tissue, the authors observed that 37% of the protein-coding genes use alternative promoters. This approach provides an opportunity to probe cancer genomes and compare them to normal cell genomes to understand how cancer specific-promoters might drive tumorigenesis.

Similar to alternative promoters, misregulation of pre-mRNA splicing can lead to oncogenic changes in otherwise normal genes [58]. One of the best examples of alternative splicing leading to conversion of a non-tumorigenic protein to an oncogenic protein was investigated for splicing of the pyruvate kinase gene [59,60]. Surprisingly, cancer cells exclusively express the embryonic splice isoform M2 of pyruvate kinase [59,61,62]. Switching from the M1 isoform to the M2 isoform in cancer cells results in increased anaerobic respiration and reduced oxidative phosphorylation, thereby contributing to tumorigenesis [59]. This study highlights the importance of identifying tumor-specific gene isoforms, of understanding their role in tumorigenesis and of exploiting them for developing targeted cancer therapies.

Several methods have been used to identify alternative splicing events at a genome-wide scale [63–65]. One approach that has yielded remarkable results is RNA-seq [63,64]. A recent study used RNA-seq to obtain a snapshot of the human transcriptome and revealed 25% more transcripts compared to microarray technology [64]. The authors used the RNA-seq data set to elucidate alternative splicing and were able to identify 95% of the splicing events expected in their data set. The study identified

4096 previously unknown splice junctions in 3106 genes that were unique to one cell type. Within a given cell type, using junction reads, alternative splicing was identified in 30% of the expected genes and exon skipping was found to be largely over-represented. Most importantly, the authors were able to decipher complex patterns of alternative splicing. For instance, using *EIF4G1*, which encodes eukaryotic translation initiation factor 4 γ1, the authors identified 12 alternative splicing junctions in B cells, five of which had not been identified in previous studies. Therefore, RNA-seq should be the preferred method for cancer genome transcriptomics, especially for identifying previously undocumented alternative splicing patterns in cancer cells, with the possibility of targeting cancer-cell-specific isoforms for cancer treatment.

Beyond the conventional cancer gene concept: abnormal cancer-related chromatin looping

The cancer genome is very dynamic. During tumorigenesis and cancer progression, the constantly evolving cancer genome drastically affects the transcription profile through both local chromatin changes and major alterations in long-range genomic interactions [66,67].

Several studies have indicated that long-range chromatin interactions might be important for regulating the expression of oncogenes and imprinting of cancer-related genes [66]. Evidence of long-range interactions and their effects on oncogene *MYC* expression came from a study that analyzed the role of an inherited 8q24 cancer risk variant, rs6983267, which is significantly associated with increased cancer risk in many malignancies, including colon cancer. The rs6983267 variant is intriguing because of its location in a gene desert region. The authors noted that the nearest gene is located ~335 kb telomeric from this region and identified it as the proto-oncogene *MYC*. This study demonstrated that the 8q24 variant region shows all the features that define an enhancer. Using chromosome conformation capture (3C) technology, the authors demonstrated that long-range interaction of this 8q24 variant enhancer with the *MYC* locus contributes to MYC overexpression leading to increased risk of colorectal cancer [66,68].

These observations were very timely because two new methods were developed in 2009 that can be applied to identify long-range interactions on a genome-wide scale [69,70]. Both these methods are adaptation of 3C technology for high-throughput genome-wide long-range interaction analysis [68]. In the first study, the authors developed a method called Hi-C, which can probe the three-dimensional architecture of the human genome using proximity-based ligation followed by massive parallel sequencing. Using Hi-C, the authors constructed spatial proximity maps of the human genome at a resolution of 1 Mb [69]. In the second study, the authors developed a method called chromatin interaction analysis by paired end sequencing (ChIA-PET) [70]. ChIA-PET incorporates ChIP-based enrichment, chromatin proximity ligation, paired end tags and high-throughput sequencing. Using ChIA-PET, the authors developed a human chromatin interactome of estrogen receptor (ER)-α binding. The data were corroborated by trimethylation of lysine 4 (H3K4me3) ChIP, RNA pol II ChIP and gene expression arrays [70]. In summary, the results indicate that long-range interactions play an important role in transcription regulation by ER-α [70].

Both the Hi-C and ChIA-PET methods provide researchers with opportunities to understand and evaluate cancer-specific changes that occur in the context of the three-dimensional cancer genome. This information could be used to elucidate cancer-specific transcriptional regulation and its role in tumorigenesis.

Integration of high-throughput chromatin structure and transcription start-site mappings

All the methodologies we have described are capable of providing large data sets of genome-wide information regarding cancer genome structure, chromatin modification states, gene transcriptional outputs or gene functions. Interestingly, more powerful insights can be generated via integration of multiple data sets using the genomic coordinates of each observation as a common frame of reference.

For example, an RNAi screen might reveal a specific transcription factor that acts as a key oncogene and is overexpressed in the development of a certain type of cancer. ChIP data obtained using an antibody specific for the same transcription factor can be used to identify the occupancy of hundreds of different genomic loci by this transcription factor, whereas a separate set of ChIP experiments can reveal multiple histone modifications associated with loci bound by the transcription factor. In addition, a cap analysis of gene expression (CAGE) experiment can be performed to identify all transcription start sites (TSSs) induced by the transcription factor in the same cancer samples. CAGE is a high-throughput method for analysis of the nucleotide sequence at the capped 5′ terminus of RNA molecules [71–73]. Capped molecules are selected from total RNA using affinity capture and the resulting material is used to make complementary DNA, with subsequent ligation of a specially designed DNA tag sequence. The tagged cDNA is then processed to generate concatenated DNA molecules containing the first 20 bases present at the 5′ terminus of each of the original capped RNA molecules. Sequencing of concatenated DNA facilitates mapping of the first 20 bases of each RNA to the corresponding location in the human genome. The exact positions of thousands of TSSs are thus defined in a single CAGE experiment. The power of this approach lies in its ability to identify changes in transcriptional profiles that involve both genes and regulatory RNA loci, such as microRNAs, antisense RNAs, long noncoding RNAs and other RNA regulatory molecules that might be specific to a cancer cell. By integrating chromatin-based data and CAGE-TSS mapping data, we might begin to reveal informative patterns of transcriptional regulatory logic that could extend our current gene-centric view of molecular abnormalities in cancer biology [74].

Conclusions

Human cancer poses an increasing healthcare challenge owing to the aging population and has dramatic socioeconomic implications [75]. However, the high prevalence and complexity of cancer have accelerated the development of powerful research tools for furthering our understanding

all complex human diseases [76,77]. Our review highlights a new generation of technologies that are rapidly being developed and applied as tools in ambitious cancer functional genomics initiatives and even to combat all human diseases.

Although these new genomics methodologies are advancing our understanding of cancer functional genomics, they are also creating new challenges for intelligent integration of the large amount of information obtained. We are only starting to place this information in the context of a systems view of human cancer to elucidate regulatory logic at a deeper level, which should facilitate the development of novel cancer treatment strategies. In the future, cancer research will benefit from conceptual integration using structural genomics, proteomics, and metabolomics approaches. An integrative systems biology view will be crucial for advances in cancer prevention and treatment, the ultimate goal of all cancer research efforts.

Acknowledgements

We apologize to those whose work could not be cited owing to space limitations. N.W. and P.M.L. are members of Yale Cancer Center. N.W. is a Sidney Kimmel Scholar for Cancer Research and is supported by Yale Department of Pathology Start-up funds, a Yale Liver Center Pilot Grant (NIDDK P30-34989) and an AACR Career Development Award for Pediatric Cancer Research (10-20-03-WAJA). P.M.L. acknowledges support from grants 1 R21 CA116079-01 and 5 R01GM080242-03 from the National Institutes of Health. We thank Job Dekker for suggestions and Alex Liu for help with manuscript preparation.

References

1 Kinzler, K.W. and Vogelstein, B. (1996) Lessons from hereditary colorectal cancer. *Cell* 87, 159–170
2 Jones, P.A. and Baylin, S.B. (2007) The epigenomics of cancer. *Cell* 128, 683–692
3 Ting, A.H. *et al.* (2006) The cancer epigenome – components and functional correlates. *Genes Dev.* 20, 3215–3231
4 Takahashi, T. *et al.* (1989) p53: a frequent target for genetic abnormalities in lung cancer. *Science* 246, 491–494
5 Hollstein, M. *et al.* (1991) p53 mutations in human cancers. *Science* 253, 49–53
6 Li, J. *et al.* (1997) *PTEN*, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* 275, 1943–1947
7 Metzker, M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11, 31–46
8 Ley, T.J. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66–72
9 Mardis, E.R. *et al.* (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* 361, 1058–1066
10 Pleasance, E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191–196
11 Taylor, B.S. *et al.* (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell* 18, 11–22
12 Lee, W. *et al.* (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 465, 473–477
13 Jones, S. *et al.* (2010) Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* 330, 228–231
14 The Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068
15 Verhaak, R.G. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1, EGFR,* and *NF1*. *Cancer Cell* 17, 98–110
16 Chiang, D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103
17 Chin, L. *et al.* (2011) Making sense of cancer genomic data. *Genes Dev.* 25, 534–555
18 Mamanova, L. *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118
19 Turner, E.H. *et al.* (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* 6, 315–316
20 Hodges, E. *et al.* (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527
21 Jones, S. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321, 1801–1806
22 Harbour, J.W. *et al.* (2010) Frequent mutation of *BAP1* in metastasizing uveal melanomas. *Science* 330, 1410–1413
23 Fire, A. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391, 806–811
24 Elbashir, S.M. *et al.* (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411, 494–498
25 Hammond, S.M. *et al.* (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nat. Rev. Genet.* 2, 110–119
26 Chang, K. *et al.* (2006) Lessons from Nature: microRNA-based shRNA libraries. *Nat. Methods* 3, 707–714
27 Paddison, P.J. *et al.* (2004) A resource for large-scale RNA-interference-based screens in mammals. *Nature* 428, 427–431
28 Berns, K. *et al.* (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 428, 431–437
29 Silva, J.M. *et al.* (2005) Second-generation shRNA libraries covering the mouse and human genomes. *Nat. Genet.* 37, 1281–1288
30 Bassik, M.C. *et al.* (2009) Rapid creation and quantitative monitoring of high coverage shRNA libraries. *Nat. Methods* 6, 443–445
31 Silva, J.M. *et al.* (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* 319, 617–620
32 Schlabach, M.R. *et al.* (2008) Cancer proliferation gene discovery through functional genomics. *Science* 319, 620–624
33 Gazin, C. *et al.* (2007) An elaborate pathway required for Ras-mediated epigenetic silencing. *Nature* 449, 1073–1077
34 Gobeil, S. *et al.* (2008) A genome-wide shRNA screen identifies *GAS1* as a novel melanoma metastasis suppressor gene. *Genes Dev.* 22, 2932–2940
35 Wajapeyee, N. *et al.* (2008) Oncogenic *BRAF* induces senescence and apoptosis through pathways mediated by the secreted protein IGFBP7. *Cell* 132, 363–374
36 Palakurthy, R.K. *et al.* (2009) Epigenetic silencing of the *RASSF1A* tumor suppressor gene through HOXB3-mediated induction of *DNMT3B* expression. *Mol. Cell* 36, 219–230
37 Sheng, Z. *et al.* (2010) A genome-wide RNA interference screen reveals an essential CREB3L2–ATF5–MCL1 survival pathway in malignant glioma with therapeutic implications. *Nat. Med.* 16, 671–677
38 Zender, L. *et al.* (2008) An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* 135, 852–864
39 Luo, J. *et al.* (2009) A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* 137, 835–848
40 Scholl, C. *et al.* (2009) Synthetic lethal interaction between oncogenic *KRAS* dependency and *STK33* suppression in human cancer cells. *Cell* 137, 821–834
41 Hartwell, L.H. *et al.* (1997) Integrating genetic approaches into the discovery of anticancer drugs. *Science* 278, 1064–1068
42 Kaelin, W.G., Jr (2005) The concept of synthetic lethality in the context of anticancer therapy. *Nat. Rev. Cancer* 5, 689–698
43 Luo, J. *et al.* (2009) Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* 136, 823–837
44 Boehm, J.S. *et al.* (2007) Integrative genomic approaches identify *IKBKE* as a breast cancer oncogene. *Cell* 129, 1065–1079
45 Maston, G.A. *et al.* (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 7, 29–59
46 Schibler, U. and Sierra, F. (1987) Alternative promoters in developmental gene expression. *Annu. Rev. Genet.* 21, 237–257
47 Ayoubi, T.A. and Van De Ven, W.J. (1996) Regulation of gene expression by alternative promoters. *FASEB J.* 10, 453–460
48 Kimura, K. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* 16, 55–65
49 Baek, D. *et al.* (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.* 17, 145–155
50 Davuluri, R.V. *et al.* (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.* 24, 167–177

51 Liu, S. (2010) Increasing alternative promoter repertories is positively associated with differential expression and disease susceptibility. *PLoS ONE* 5, e9482
52 Lamprecht, B. *et al.* (2010) Derepression of an endogenous long terminal repeat activates the *CSF1R* proto-oncogene in human lymphoma. *Nat. Med.* 16, 571–579
53 Singer, G.A. *et al.* (2008) Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. *BMC Genomics* 9, 349
54 Sun, H. *et al.* (2010) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res.* DOI: 10.1093/nar/gkq775
55 Barrera, L.O. *et al.* (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.* 18, 46–59
56 Sandelin, A. *et al.* (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* 8, 424–436
57 Carninci, P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635
58 David, C.J. and Manley, J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.* 24, 2343–2364
59 Christofk, H.R. *et al.* (2008) The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* 452, 230–233
60 Christofk, H.R. *et al.* (2008) Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* 452, 181–186
61 Mazurek, S. *et al.* (2005) Pyruvate kinase type M2 and its role in tumor growth and spreading. *Semin. Cancer Biol.* 15, 300–308
62 Dombrauckas, J.D. *et al.* (2005) Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis. *Biochemistry* 44, 9417–9429
63 Richard, H. *et al.* (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.* 38, e112
64 Sultan, M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960
65 Gupta, S. *et al.* (2004) Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics* 20, 2579–2585
66 Pomerantz, M.M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* 41, 882–884
67 Vu, T.H. *et al.* (2010) Loss of IGF2 imprinting is associated with abrogation of long-range intrachromosomal interactions in human cancer cells. *Hum. Mol. Genet.* 19, 901–919
68 Dekker, J. *et al.* (2002) Capturing chromosome conformation. *Science* 295, 1306–1311
69 Lieberman-Aiden, E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293
70 Fullwood, M.J. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64
71 Balwierz, P.J. *et al.* (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.* 10, R79
72 Kodzius, R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222
73 Shiraki, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15776–15781
74 Mattick, J.S. *et al.* (2010) A global view of genomic information – moving beyond the gene and the master regulator. *Trends Genet.* 26, 21–28
75 Greenberg, E.R. *et al.* (1988) Social and economic factors in the choice of lung cancer treatment. A population-based study in two rural states. *N. Engl. J. Med.* 318, 612–617
76 Xie, Y. and Minna, J.D. (2008) Predicting the future for people with lung cancer. *Nat. Med.* 14, 812–813
77 Meyerson, M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11, 685–696