Making sense of cancer genomic data

Lynda Chin,1,2,3 William C. Hahn,1,2 Gad Getz,2 and Matthew Meyerson1,2

1Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA; 2Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

High-throughput tools for nucleic acid characterization now provide the means to conduct comprehensive analyses of all somatic alterations in the cancer genomes. Both large-scale and focused efforts have identified new targets of translational potential. The deluge of information that emerges from these genome-scale investigations has stimulated a parallel development of new analytical frameworks and tools. The complexity of somatic genomic alterations in cancer genomes also requires the development of robust methods for the interrogation of the function of genes identified by these genomics efforts. Here we provide an overview of the current state of cancer genomics, appraise the current portals and tools for accessing and analyzing cancer genomic data, and discuss emerging approaches to exploring the functions of somatically altered genes in cancer.

The development of powerful and scalable methods to analyze nucleic acids has transformed biological inquiry and has the potential to alter the practice of medicine [Lander 1996; Collins et al. 2003]. The application of such technologies, together with powerful computational methods in human disease and animal model systems, has facilitated the study of both normal and disease-affected tissues in a manner previously not possible. Indeed, the connection between basic inquiry and potential clinical translation has never been more intimate.

This convergence is particularly evident in cancer, a complex multigenic disease characterized by a diversity of genetic and epigenetic alterations (Vogelstein and Kinzler 1993; Weir et al. 2004; Jones and Baylin 2007; Stratton et al. 2009). Early cancer genome analysis has already led to new targets for cancer therapy and new insights into the relationship of specific genetic mutations and clinical response, as well as new approaches useful for diagnosis and prognosis. These initial efforts have motivated large-scale coordinated cancer genomic efforts to obtain complete catalogs of the genomic alterations in specific cancer types [The Cancer Genome Atlas [TCGA], http://cancergenome.nih.gov; Hudson et al. 2010]. Moreover, the current pace of technological advances make it increasingly clear that the ability to perform prospective and comprehensive molecular profiling of tumors will become commonplace and enable genome-informed personalized cancer medicine.

However, the bottlenecks along this path are formidable and numerous. For one, these large-scale genome characterization efforts involve the generation and interpretation of data at an unprecedented scale, which has brought into sharp focus the need for improved information technology infrastructure and new computational tools to render the data suitable for meaningful analyses. Moreover, exploiting this information to develop new therapeutic strategies depends on further biological insights derived from understanding the functional consequences of such genomic alterations. Thus, it is also clear that new approaches that permit the efficient validation of genomic data are required as a first step in distinguishing mutations responsible for disease pathogenesis from other mutations that are the consequence of genomic instability; defining genes involved in cancer initiation, progression, or maintenance; and identifying the optimal ways to exploit this information therapeutically.

In this review, we provide an overview of the current state of cancer genomics, describe the types of data being generated and where they can be accessed, and discuss recent progress in developing tools, models, and methods for analysis of gene functions in cancer, which is the requisite next step in the translation of cancer genome information.

State of cancer genomics

Nearly all cancer genomes contain many nucleotide sequence changes compared with the germline of the cancer patient [Vogelstein and Kinzler 1993; Stratton et al. 2009]. These variations include the genomic alterations that cause or promote cancer, often referred to colloquially as “drivers,” as well as alterations present in the cancer genome but without obvious advantage to the cancerous cells when they occurred, referred to as “passengers” [Davies et al. 2005]. The major known somatic alterations in the cancer genome include nucleotide substitution mutations and small insertion/deletions (indels), copy number gains and losses, chromosomal rearrangements, and nucleic acids of foreign origin [e.g., oncogenic viruses] [Weir et al. 2004; Chin and Gray 2008; Stratton et al. 2009]. In addition, alterations in the epigenetic
mechanisms that regulate gene expression occur in most cancers; this subject has been covered elsewhere (Jones and Baylin 2002, 2007) and is not treated in detail here. All of these acquired changes occur in the setting of germline variations of copy number and nucleotide sequence, which may influence the rate of occurrence and/or the effects of somatic genetic alterations [Balmain et al. 2003]. Moreover, although somatic mutations occur in tumor cells, it is increasingly clear that the tumor microenvironment mediates important heterotypic signals between the tumor and stromal cells for growth and survival (Hanahan and Weinberg 2000). In this respect, global gene expression encompassing the full transcriptome—including coding messenger RNAs (mRNAs) (Schena et al. 1995) and noncoding microRNAs (miRNAs) (Lu et al. 2005)—of the complex tumor tissue reflects the panoply of somatic and epigenomic alterations, together with the state of cell differentiation of the tumor and the admixture of noncancerous cells. Hence, transcriptional profiling can define a unique gene expression signature for each tumor that may prove useful for classification and prognosis (Golub et al. 1999; Alizadeh et al. 2000; van’t Veer et al. 2002).

**Evolution of cancer genomics**

Over the past decade, technologies for detection of each of these types of alterations have been developed and applied to analyses of the cancer genomes. The initial studies focused on a single technology platform and/or type of genetic alteration. For example, high-resolution copy number profiling has led to the discovery of novel oncogenes in ovarian cancer [Nanjundan et al. 2007], melanoma [Garraway et al. 2005; Kim et al. 2006; Scott et al. 2009], lung carcinoma [Weir et al. 2007; Bass et al. 2009], and colon carcinoma [Firestein et al. 2008], and tumor suppressor genes in leukemias [Mullighan et al. 2007, 2008]. Similarly, the application of directed sequencing of specific classes of genes has identified novel genes involved in specific types of cancer (Davies et al. 2002, Lynch et al. 2004; Paez et al. 2004; Pao et al. 2004; Samuels et al. 2004; Stephens et al. 2004; Baxter et al. 2005; James et al. 2005; Kralovics et al. 2005; Levine et al. 2005; Zhao et al. 2005; Pollock et al. 2007; Chen et al. 2008; Dutt et al. 2008; George et al. 2008; Janoueix-Lerosey et al. 2008; Mosse et al. 2008]. These observations underscore the contribution of different types of somatic genome alterations in different subsets of cancer, and that comprehensive profiling of the cancer genome will require interrogation of different types of genome alterations in diverse cancer types and subtypes.

As technologies to perform comprehensive profiling of the cancer genome progressed, different technology platforms from microarrays to capillary sequencing were brought together on unified sample sets (The Cancer Genome Atlas Network 2008; S Jones et al. 2008; Parsons et al. 2008). For example, Velculescu and colleagues (S Jones et al. 2008) integrated sequencing with expression and copy number profiling to identify IDH1 mutations in glioblastomas (GBM). The Cancer Genome Atlas pilot project applied targeted sequencing, copy number, and expression profiling, in addition to epigenetic assessment to a large number of stringently qualified tumor samples to define core pathways of deregulation in GBM [The Cancer Genome Atlas Network 2008] and discover genomic and epigenomic definition of molecular subtypes [Noushmehr et al. 2010, Verhaak et al. 2010]. Indeed, global comprehensive analysis with complementary genome annotation tools in statistically powered high-quality sample cohorts is a key aspect of the current consensus standard for large-scale cancer genomics efforts under the International Cancer Genome Consortium [ICGC] (Hudson et al. 2010), now encompassing >20 projects from 14 countries [Table 1].

**Second-generation sequencing technologies**

During the past several decades, continuous improvements in genomic technology have led to a series of breakthroughs in our understanding of cancer genetics. The advent of second-generation sequencing technologies and their applications to cancer have already accelerated the pace of genome discovery, as summarized in recent reviews [Shendure and Ji 2008; Meyerson et al. 2010]. During the last 5 years, a variety of array-based methods have been developed, including picotiter plate pyrosequencing [Margulies et al. 2005; Wheeler et al. 2008], single-nucleotide fluorescent base extension with reversible terminators (Bentley et al. 2008), and ligation-based sequencing [Shendure et al. 2005; Drmanac et al. 2010]. All of these second-generation methods involve the amplification of individual DNA molecules on arrays or beads prior to massively parallel sequence generation.

The throughput limitation and cost of first-generation Sanger-based capillary sequencing technology had, until now, dictated two predominant study designs for cancer genome discovery: one in which large numbers of samples were analyzed but only a small number of genes were interrogated [Greenman et al. 2007; The Cancer Genome Atlas Network 2008; Dalgliesh et al. 2010; Kan et al. 2010], versus a second in which all coding genes were sequenced but in only a handful of discovery samples, followed by targeted sequencing of candidates in an extension cohort comprised of an independent set of samples [Sjoblom et al. 2006; Wood et al. 2007; S Jones et al. 2008; Parsons et al. 2008]. Second-generation sequencing technology enables the complete sequencing of entire genomes in a time- and cost-efficient manner. Today, a single sequencing run of an Illumina HiSeq 2000 sequencer can generate ~200 gigabases of sequence data in 8 d—an output that easily exceeds the annual sequencing production of a genome sequencing center a few years ago [http://www.genome.gov/10016911]. This astronomical increase in sequencing capacity, along with the rapid reduction in sequencing cost (which is faster than the doubling of semiconductor/computer capacity every 18 mo, known as Moore’s law) [Pettersson et al. 2009], has completely transformed cancer genome discovery science.

Second-generation sequencing offers several advantages over previous technologies. It has the power to
identify mutations in highly admixed samples by virtue of deep coverage (Thomas et al. 2006), overcoming a major limitation of Sanger-based capillary sequencing technology. Whereas previous technologies could query one modality of cancer genome alteration (mutation, copy number, or expression) at a time, second-generation sequencing analyses permit the identification of all such alterations simultaneously. For example, one can obtain high-resolution and accurate measurements of somatic copy number alterations (SCNAs) from whole-genome sequencing (Campbell et al. 2008; Chiang et al. 2009), and the same data can identify nucleotide substitutions. Furthermore, second-generation sequencing offers structural information never before available from other genomic platforms, thus enabling for the first time global assessment of chromosomal rearrangements in cancer (Campbell et al. 2008; Mardis et al. 2009; Stephens et al. 2009; Ding et al. 2010a; Pleasance et al. 2010a,b). Similar approaches can be applied to cDNA, also known as RNA-seq, which permits accurate digital measurements of gene expression across the whole transcriptome. Importantly, this latter approach provides the means to measure known, and discover novel, splice variants as well as aberrant transcripts generated by somatic structural genome rearrangements (Maher et al. 2009a,b; Berger et al. 2010; Palanisamy et al. 2010). This type of data will undoubtedly reveal new insights into the regulation of gene transcription and RNA processing.

In the near future, sequencing-based approaches will be applied to nearly all aspects of cancer genome characterization. For example, current TCGA projects involve the comprehensive sequencing of all protein-coding genes and transcripts by hybrid capture/whole-exome sequencing in hundreds of tumor- and germline-matched pairs,
complemented by deep sequencing of the whole genomes (WGS) [at >30-fold coverage] in 10% of the samples. Given the rate at which sequencing capacity is increasing and cost is shrinking, it is highly likely that deep-coverage WGS will soon be applied to the majority of the discovery samples. In parallel, efforts to use low-coverage sequencing to conduct structural and copy number analyses are likely to replace array-based technologies in the near future. Similar study designs are being adopted by ICGC projects.

Accessing cancer genomics data

Although the data generated from these large-scale cancer genome characterization efforts have been and will continue to be made publicly available, accessing and using these cancer genome data remains a major challenge. In this section, we attempt to provide a framework for nongenomic noncomputational cancer biologists to become familiar with what data are available and where to download or query each of the major genomic data types. Specifically, we describe briefly the technology platform[s] used to generate each data type, followed by common public sites where cancer genome data can be downloaded and summarized results can be queried. We also point out basic open source analytical tools or computational algorithms for manipulation and analysis of cancer genome data. However, it should be noted that the tools described here are illustrative examples, rather than a complete survey of sites, data sources, or analytical tools, as these are rapidly evolving.

Data structure and data access policies

Generally speaking, cancer genomic data can be divided into [1] raw, [2] processed or normalized, [3] interpreted, and [4] summarized categories based on the degree of computational modification and integration applied to the data. These categories are sometimes referred to as Level I–Level IV data. Raw, processed, and interpreted [Level I–III] data apply to individual samples, while summarized [Level IV] data refer to analyses across sample sets. For example, normalized or processed data represent data that have been assigned to a genome reference, such as alignment of sequences to reference genome or mapping of probes to chromosomal positions. For microarray-based platforms, normalization refers to combining multiple probes measuring a single genomic locus to a single value and transforming the measured intensities such that the values can be compared between experiments; examples of normalization steps include correction for background noise and total brightness. Interpreted data represent meaningful biological results extracted from each specimen, such as genome-wide copy number profile, where copy number breakpoints have been statistically defined, or gene expression profiles, where individual gene expression levels have been collated from multiple loci across the gene. Summarized [Level IV] data represent analysis of interpreted data across a cohort of samples, where statistical methodologies can be applied to define significant events or molecular subtypes. This category of analyzed data is often presented as the findings of a genomic study in a publication. Major sites where these data sets can be accessed are listed in Table 2 and are described in some detail below.

Although cancer genomic data from various large-scale projects—including the Cancer Genome Project [CGP] at Wellcome Trust Sanger Institute [http://www.sanger.ac.uk/genetics/CGP], TCGA [http://cancergenome.nih.gov/dataportal], and ICGC [http://dcc.icgc.org]—are publicly available, for protection of patient privacy, access is either open or controlled. Prior to second-generation sequencing, most raw data and some type of normalized data (e.g., single nucleotide polymorphism [SNP] profiles) are subjected to controlled-access restriction, while interpreted and summarized data are openly accessible. With the transition to second-generation sequencing data, it is likely that raw and processed data, and possibly some interpreted data, will fall under the “controlled-access” category, since the level of resolution may provide the means to identify specific patients. Controlled-access data are restricted to qualified researchers [with certification by host institution] with a specific proposal of data use that is deemed compliant with the project’s data access policy, typically requiring preapproval by the institutional review board of the requesting investigator. For TCGA, access to controlled data is obtained through dbGAP [http://www.ncbi.nlm.nih.gov/gap], for ICGC projects, access is obtained through its Data Access Compliance Office [http://www.icgc.org/daco]. In the case of the Sanger Institute’s Cancer Genome Project, genotyping and first-generation sequencing traces can be requested at its data archive [http://www.sanger.ac.uk/genetics/CGP/Archive], and its second-generation sequencing data must be obtained through the European Genome-Phenome Archive [EGA, http://www.ebi.ac.uk/ega]. At present, downloading raw data from these sources is a technically and logistically challenging task that requires significant network infrastructure to handle the size of the data files.

Nucleotide sequence mutations

Nucleotide substitutions and small insertions/deletions are common mechanisms for activating oncogenes and inactivating tumor suppressor genes. The initial development of methods to determine the nucleotide sequence of DNA in 1975 [Sanger and Coulson 1975] led to the discovery of cancer-specific somatic mutations in the \textit{RAS} gene family in the early 1980s [Parada et al. 1982; Shimizu et al. 1983; Santos et al. 1984; Bos et al. 1985] and, later, mutations in human tumor suppressor genes [Friend et al. 1986; Hahn et al. 1996]. Subsequently, the invention of automated sequencing instruments [Hunkapiller et al. 1991] led to the initial sequencing of the human genome [Lander et al. 2001; Venter et al. 2001], and then to systematic efforts to sequence gene families [Davies et al. 2005; Stephens et al. 2005; Greenman et al. 2007]. These latter efforts identified several new oncogene mutations that are targets for cancer therapy—most notably the
<table>
<thead>
<tr>
<th>Database</th>
<th>Link</th>
<th>Data type</th>
<th>Type of information</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICGC</td>
<td><a href="http://dcc.icgc.org/">http://dcc.icgc.org/</a></td>
<td>Levels I–IV</td>
<td>Copy number, rearrangement, expression, and mutation data</td>
<td>Open and controlled</td>
</tr>
<tr>
<td>TCGA</td>
<td><a href="http://cancergenome.nih.gov/dataportal">http://cancergenome.nih.gov/dataportal</a></td>
<td>Levels I–III</td>
<td>Copy number, expression [mRNA and miRNA], promoter methylation, and mutation sequencing</td>
<td>Open and controlled</td>
</tr>
<tr>
<td>COSMIC</td>
<td><a href="http://www.sanger.ac.uk/genetics/CGP/cosmic">http://www.sanger.ac.uk/genetics/CGP/cosmic</a></td>
<td>Levels III–IV</td>
<td>Somatic mutations and copy number alterations by gene: amino acid position, tumor type, literature references</td>
<td>Open</td>
</tr>
<tr>
<td>Cancer Gene Census</td>
<td><a href="http://www.sanger.ac.uk/genetics/CGP/Census">http://www.sanger.ac.uk/genetics/CGP/Census</a></td>
<td>Level IV</td>
<td>Annotation of mutated or genomically altered genes</td>
<td>Open</td>
</tr>
<tr>
<td>WTSI CGP</td>
<td><a href="http://www.sanger.ac.uk/genetics/CGP/Archive">http://www.sanger.ac.uk/genetics/CGP/Archive</a></td>
<td>Levels I–II</td>
<td>First-generation trace archive, SNP genotype profiles</td>
<td>Controlled</td>
</tr>
<tr>
<td>EGA</td>
<td><a href="http://www.ebi.ac.uk/ega">http://www.ebi.ac.uk/ega</a></td>
<td>Levels I–II</td>
<td>Second-generation sequencing BAM files generated by WTSI CGP</td>
<td>Controlled</td>
</tr>
<tr>
<td>Tumorscape</td>
<td><a href="http://www.broadinstitute.org/tumorscape">http://www.broadinstitute.org/tumorscape</a></td>
<td>Levels I–IV</td>
<td>Browsable, searchable cancer copy number viewer using SNP array data</td>
<td>Open</td>
</tr>
<tr>
<td>Oncomine</td>
<td><a href="http://www.oncomine.org">http://www.oncomine.org</a></td>
<td>Level IV</td>
<td>Gene expression and copy number data in readily searchable and comparable fashion</td>
<td>Password-protected</td>
</tr>
<tr>
<td>caArray</td>
<td><a href="http://caarray.nci.nih.gov">http://caarray.nci.nih.gov</a></td>
<td>Level I</td>
<td>Gene expression data</td>
<td>Password-protected</td>
</tr>
<tr>
<td>UCSC Cancer Genome</td>
<td><a href="https://genome-cancer.ucsc.edu">https://genome-cancer.ucsc.edu</a></td>
<td>Levels III–IV</td>
<td>Browsable viewer for cancer copy number and expression data</td>
<td>Open</td>
</tr>
<tr>
<td>The cBio Cancer</td>
<td><a href="http://cbioportal.org">http://cbioportal.org</a></td>
<td>Levels III–IV</td>
<td>Browsable and searchable viewer for cancer copy number and expression data</td>
<td>Open</td>
</tr>
<tr>
<td>Genomics Portal</td>
<td><a href="http://www.ncbi.nlm.nih.gov/omim">http://www.ncbi.nlm.nih.gov/omim</a></td>
<td></td>
<td>Inherited syndromes and causative genes for cancer and other diseases, with extensive literature review</td>
<td>Open</td>
</tr>
<tr>
<td>Mitelman</td>
<td><a href="http://cgap.nci.nih.gov/Chromosomes/Mitelman">http://cgap.nci.nih.gov/Chromosomes/Mitelman</a></td>
<td></td>
<td>Copy number alterations and translocations based on cytogenetic data</td>
<td>Open</td>
</tr>
</tbody>
</table>

*Level I* Raw; *Level II* normalized/processed; *Level III* interpreted; *Level IV* summarized.
BRAF and EGFR protein kinase genes and the PIK3CA phosphatidylinositol kinase gene (Davies et al. 2002; Lynch et al. 2004; Paez et al. 2004; Pao et al. 2004; Samuels et al. 2004)—leading to approved and in-development targeted therapeutics for cancers.

For sequencing data, the normalized data category represents sequencing reads that have been aligned to a specific version of the human reference genome. As the reference genome is refined and filled in with each new version, mapping data may change; thus, researchers should pay attention to and always note the specific reference genome build used in an analysis. Raw sequencing reads from both Sanger-based capillary sequencing and second-generation platforms are stored at NCBI Sequence Read Archive and dbGAP (see Table 2). Access to these sequencing data is restricted and requires data use approval by the appropriate data access committee. The interpreted category for mutation data includes the sequence variant calls, the types of sequence variants (such as synonymous versus nonsynonymous or missense versus indel), and the location of the nucleotide change in relation to annotated gene structure; e.g., intron versus exon and consequent amino acid changes. One aspect of this analysis includes an annotation of whether such sequence variants are reported in dbSNP, a database representing likely common SNPs. If not found in dbSNP and not observed in matched germline-derived sequences, such variants are generally considered somatic in nature. Putative variants should be verified (e.g., result reproduced in an independent assay using the same technology platform) or validated (e.g., result observed by an orthogonal method) by methods including genotyping. For TCGA projects, all verified, validated, or putative somatic mutations and associated descriptions discovered in a sample or a cohort of samples can be found in the .MAF file, which is available on the TCGA data portal [http://cancergenome.nih.gov/dataportal]. Similar data files can be found at the ICGC Data Coordination Center [http://dcc.icgc.org] under the Download Data page. New file formats will likely emerge in the near future to support the increasing applications of next-generation sequencing platforms for data generation.

A key downstream (Level IV) analysis of verified or validated mutations is the determination of significance, accounting for the background mutation rate and size as well as composition of a gene. Several methodologies have been developed for this purpose (Getz et al. 2007; Greenman et al. 2007). For example, in the MutSig algorithm, a P-value is calculated for each gene, testing the hypothesis that all of the observed mutations in that gene are a consequence of random background mutation processes, taking into account the list of bases that are successfully interrogated by sequencing (i.e., “covered”) and the list of observed somatic mutations, as well as the length and composition of the gene in addition to the background mutation rates in different sequence contexts. As in analyses of other genomic data, such calculations must then be corrected for multiple hypothesis testing (see below). Using these types of significance analyses, the majority of the somatic mutations found in cancer genomes is likely to represent passenger events and only a minority is likely drivers. For example, of the 453 validated nonsilent mutations in GBM scattered across 223 genes, only eight genes were considered having higher than background mutation frequency, suggestive of positive selection pressure (The Cancer Genome Atlas Network 2008). However, it is worth noting that, as for any statistical test, the lack of statistical significance by MutSig or similar analyses does not preclude true cancer relevance, as, relatively speaking, the number of samples having been adequately sequenced is still low. Moreover, computational algorithms for mutation calling and statistical frameworks for significance calculation are still being developed and refined.

Beyond statistical analyses, there are various theoretical and computational models designed to predict the likely functional consequences of specific nucleotide mutations, particularly for mutations in coding genes. These models are often based on the impact of specific amino acid substitutions on protein structure or known functional domains or evolutionarily conserved regions. For example, the PolyPhen [for polymorphism phenotyping] tool predicts the possible impact of an amino acid substitution on the structure and function of a protein using a variety of structural and chemical parameters in addition to evolutionary conservation (Sunyaev et al. 2001; Ramensky et al. 2002). Indeed, a recurring theme in analysis of genomic data is the leverage of evolutionary information. MutationAssessor [Reva et al. 2007] is a recently published algorithm for predicting potential functional impact of a sequence mutation based heavily on the assumption that if a highly conserved residue is changed to a different residue type, the change is presumed to have high functional impact on the function of the affected protein. By analyzing aligned sequence families of paralogous and orthologous proteins within the human genome and across many other species, this algorithm calculates the functional impact [FI] score for a mutation. Both of these tools are Web accessible [http://genetics.bwh.harvard.edu/pph, http://mutationassessor.org] and offer an intuitive and easy-to-use query interface as well as a batch processing feature. With MutationAssessor, in addition to the calculated FI score for each variance, users can inspect the placement of mutations in a multiple sequence alignment relative to amino acid residues that are conserved globally or in a specific subfamily, as well as observe the consequences of the residue change in an interactive three-dimensional protein structure. Other examples of prediction algorithms include SIFT, CanPredict, and CHASM [Ng and Henikoff 2001; Kaminker et al. 2007; Carter et al. 2009; Adzhubei et al. 2010]. In the end, beyond statistical analysis or the prediction of functional impact, the relevance of any mutational event to human cancer will require functional validation (see below).

The COSMIC (Catalog of Somatic Mutations in Cancer) site is the single most comprehensive source of curated analyzed somatic mutation data in cancers developed and maintained by the Cancer Genome Project at the Wellcome Trust Sanger Institute [Futreal et al. 2004; Forbes et al. 2008]. COSMIC is an open source, easily
SCNAs and structural rearrangements

Cancer genomes are highly disordered compared with normal genomes, with extensive changes in chromosome structure and copy number. The SCNAs found in cancer include whole-chromosome or regional alterations spanning part to whole arms of a chromosomes, as well as focal events involving one or a few genes. The development of array-based comparative genomic hybridization using bacterial artificial chromosomes [Hodgson et al. 2001; Cai et al. 2002], cDNA [Pollack et al. 1999], oligomers [O’Hagan et al. 2003; Brennan et al. 2004], and SNP arrays [Lindblad-Toh et al. 2000; Mei et al. 2001; Bignell et al. 2004; Zhao et al. 2004] has enabled systematic analyses of the cancer genome and defined many new recurrent SCNAs in cancer [Beroukhim et al. 2010]. To date, the major CNAs linked to cancer are somatic changes, although many germline copy number variations are found in human populations [Sebat et al. 2004; Redon et al. 2006]. Further work is needed to determine the role of germline copy number aberrations in cancer.

Most of the available genome-wide high-resolution copy number profiling data were generated on either Agilent or Affymetrix microarray platforms. Performance of these various platforms has been compared [Brennan et al. 2004; Lai et al. 2005, 2008; Willenbrock and Fridlyand 2005] and used comparatively on the same sample cohort by TCGA during its pilot phase (The Cancer Genome Atlas Network 2008). Raw data for copy number are probe-level signals, whereas processed data are the results of normalization, calculation of tumor-to-normal copy number ratio, and mapping to chromosomal positions. Interpreted data generally contain segmented copy number profiles where breakpoints along the chromosomes have been defined in each individual tumor using segmentation methods such as circular binary segmentation (CBS) [Venkatraman and Olshen 2007], GLAD [Hupe et al. 2004], and others [Picard et al. 2005; Wang et al. 2005; Day et al. 2007; Ben-Yaacov and Eldar 2008]. Recent methods for measuring and determining segments of absolute allele-specific copy number from SNP arrays provide a more accurate description of allelic gains and losses and loss of heterozygosity in cancer genomes [LaFramboise et al. 2005; Bengtsson et al. 2010; Greenman et al. 2010; Van Loo et al. 2010].

When a cohort of segmented copy number profiles is analyzed together, several methodological tools are available to define the Level IV summarized data, where significantly altered regions are defined [by specifying the boundaries of “peaks” and significance). GISTIC is a popular algorithm based on statistical considerations [Beroukhim et al. 2007], allowing users to define the most significant regions and peaks; GISTIC works most robustly in large sample cohorts. RAE uses a similar methodology [Taylor et al. 2008]. GTS [Wiedemeyer et al. 2008] provides independent measures of recurrence frequency and focality, the latter can be a strong indicator for relevance when sample size is small. For those with basic R programming skills, cghMCR and CNTools are two Bioconductor packages [http://www.r-project.org; Gentleman et al. 2004] available for copy number data analyses. The former provides a fast and platform-independent approach to identifying and visualizing altered regions, while the latter enables the conversion of segmented copy number profiles into a matrix structure to allow further downstream analyses. It should be noted that most of these algorithms are built using the prevalence and shape [e.g., focal and high amplitude vs. flat and broad] of a numerical copy number aberration to discriminate likely target[s] at the peaks from passengers. Although this framework has led to identification of new cancer genes that have been experimentally validated, next-generation sequencing is beginning to provide information on sequence-level structures underlying these numerical copy number aberrations. These emerging data will provide much finer details of structural rearrangements in addition to simple numerical changes such as duplication or amplification or deletion. Such new insights will likely lead to different or improved algorithms to identify candidate targets of these genomic alterations.

The major repositories for somatic copy number data in cancer include Gene Expression Omnibus (GEO) [Edgar et al. 2002; Barrett et al. 2009], Tumorscape [Beroukhim et al. 2010], TCGA [http://cancergenome.nih.gov], COSMIC [Forbes et al. 2008], and Oncomine (Rhodes et al. 2004, 2007) [Table 3]. The GEO site contains predominantly raw data, while COSMIC and Oncomine emphasize normalized and interpreted data. All three categories of data are available from the TCGA DCC as well as at Tumorscape. For systematic queries of analyzed summary data on CNAs in human cancers, Tumorscape is particularly useful, as it provides segmented data for >3000 high-resolution copy number profiles from SNP arrays in a format that can be visualized with the interactive Integrative Genome Viewer [http://www.broadinstitute.org/igv]. COSMIC also contains copy number data annotated on the gene level in addition to raw data download at the CGP archive, while Oncomine has recently begun to curate copy number profile data sets in addition to transcriptome data sets.

Before the advent of second-generation sequencing [Maher et al. 2009a,b; Berger et al. 2010; Palanisamy et al. 2010], conventional cytogenetic methodologies such as FISH were the primary means for identification of chromosomal translocations in cancers, predominantly...
in leukemia and lymphoma (Rabbitts 1994). One of the most useful compilations of cytogenetic alterations was established by Mitelman et al. (2007), which includes both translocations and CNAs in >50,000 samples. The recent identification of translocations in epithelial cancers suggests that the application of high-throughput sequencing will expand the number of such translocations (Tomlins et al. 2005), although the number and significance of such recurrent translocations in solid tumors remains to be delineated.

Expression analysis of cancer

Global gene expression profiles offer a global view to the transcriptome of a tumor, the signature of which has the potential to provide diagnostic (Golub et al. 1999; Alizadeh et al. 2000) or prognostic (van de Vijver et al. 2002, Paik et al. 2004) information. This approach has also been used to molecularly subclassify tumors that are currently assigned to one histopathological class. Gene expression is also one measure of functional consequence of a genomic alteration, thus potentially enabling the interpretation of inactivating genetic changes on the DNA level or epigenomic (methylation) on DNA promoters. Interpreted expression data typically represent gene-level data where multiple probes reporting on one annotated gene are collapsed [by taking either the median or mean values or best probe].

A major summarized category for expression data type is definition of molecular subtypes. The discovery of novel molecular subclasses is based primarily on differences in gene expression between groups within a cohort using various unsupervised classification methodologies, such as hierarchical clustering (Eisen et al. 1998), self-organizing maps (Tamayo et al. 1999), and nonnegative matrix factorization (Kim and Tidor 2003; Brunet et al. 2004). Visual verification or identification of clusters requires transforming the high-dimensional expression data [each sample is represented by the expression values of all genes] to two or three dimensions. This is often performed using principal component analysis (PCA) (Raychaudhuri et al. 2000) or multidimensional scaling (MDS) (Khan et al. 1998). However, it should be noted that, given the large number of available gene expression profiles and the number of independent measurements contained within each profile, one can produce class discriminations that may or may not reflect underlying biological differences. For example, batch effects, introduced by profiling samples on different days using different lots of reagents or at different sites, can introduce variations and confound such analyses. These considerations require reproducing the classification results in independent test cohorts of samples and using multiple hypothesis testing correction methods.

Once subtypes are defined, there are several user-friendly analytical algorithms that can be employed to interrogate these data sets. For example, one of the most commonly asked questions with gene expression data is: What gene expression difference exists between two biologically, clinically, or molecularly defined subgroups? Significance analysis of microarrays [SAM, http://www-stat.stanford.edu/~tibs/SAM; Tusher et al. 2001] is a commonly used tool to discern genes that are significantly different between two groups of samples. In GenePattern, ComparativeMarkerSelection (Gould et al. 2006) is another tool that can be used to define genes that are characteristic of a subgroup. Finally, one of the most commonly used sites for querying analyzed cancer gene expression data is Oncomine, which allows comparison of any two data sets from different cancers or normal tissues to determine the genes that are specifically expressed in the data set of interest (Rhodes et al. 2004, 2007).

Once differential expression gene lists or signatures for each subtype are generated, these lists can be interrogated for pathway activation using knowledge-based pathway analysis tools such as Ontologizer (Bauer et al. 2008), which looks for statistical enrichment of Gene Ontology terms, or Gorilla, a tool to visualize Gene Ontology terms that are enriched in gene lists ranked by a user-defined criterion [Eden et al. 2009]. Gene set enrichment analysis (GSEA) is an algorithm that allows one to assess whether the differentially expressed genes are enriched for particular gene sets even though each member of the gene set individually is not necessarily strongly differentially expressed. Gene sets can be defined in various ways, such as manual curation of pathways, based on genomic position or shared motifs, expression correlation, or by experimentally identifying signatures that represent a molecular event or phenotype [Subramanian et al. 2005], such as KRAS activation [Sweet-Cordero et al. 2005]. Gene signatures that can be used as input for GSEA can be downloaded from the Molecular Signatures Database [MSigDB, http://www.broadinstitute.org/gsea/msigdb/index.jsp]. Other gene set and pathway repositories include Pathway Commons [http://www.pathwaycommons.org], KEGG [Kanehisa et al. 2006], and others. Another commonly used tool for pathway and Gene Ontology enrichment is DAVID [http://david.abcc.ncifcrf.gov/home.jsp; Huang et al. 2009]. These types of analyses provide a framework for generating hypotheses for further testing.

Table 3. Sites with open source analytical tools for cancer genomics data

<table>
<thead>
<tr>
<th>Tools</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioconductor</td>
<td><a href="http://www.bioconductor.org">http://www.bioconductor.org</a></td>
</tr>
<tr>
<td>GenePattern</td>
<td><a href="http://www.broadinstitute.org/genePattern">http://www.broadinstitute.org/genePattern</a></td>
</tr>
<tr>
<td>UCSC Cancer Genome Browser</td>
<td><a href="https://genome-cancer.ucsc.edu">https://genome-cancer.ucsc.edu</a></td>
</tr>
<tr>
<td>Integrative Genomics Viewer (IGV)</td>
<td><a href="http://www.broadinstitute.org/igv">http://www.broadinstitute.org/igv</a></td>
</tr>
<tr>
<td>The Cancer Genomics Pathway Portal</td>
<td><a href="http://cbioportal.org">http://cbioportal.org</a></td>
</tr>
</tbody>
</table>
However, users should be reminded that annotation of pathways or definition of gene sets is biased by what is known and published. Genes that are widely studied will be linked to many different processes, while ones that have not been explored in depth would have few connections to others.

Second-generation sequencing data

It is clear that, in the very near future, all genome characterization data will be generated by second-generation sequencing technologies. Output from these sequencing platforms is in the form of short-sequence reads, ranging from 35 base pairs (bp) to 100 bp or longer. These raw sequencing reads and their mapped reads after alignments (to reference genome) are captured in a BAM file. Therefore, one can consider the BAM file as containing both raw and normalized [Level I–II] data; hence, BAM files are controlled access data. For TCGA, these data are stored at the Sequence Read Archive of NCBI [http://www.ncbi.nlm.nih.gov/Traces/sra] and accessed through dbGAP. Similar second-generation sequencing data from the Sanger CGP are accessed through EGA [http://www.ebi.ac.uk/ega]. The size of these data files (typically in tens to hundreds of gigabytes) and the complexity of manipulating such data make it challenging to access and analyze these data by noncomputational groups. A detailed discussion of the computational challenges, algorithms, and software for analyzing second-generation sequencing for cancer is provided in recent reviews (Ding et al. 2010b; Meyerson et al. 2010).

Viewers and tools

Some of the open source software tools and viewers are summarized in Table 3. Several user-friendly Web-based viewers provide intuitive environments with which to visualize and explore cancer genome data, although most do not allow fully interactive queries at the present time. The University of California at Santa Cruz (UCSC) Cancer Genome Browser (Zhu et al. 2009) is perhaps the most commonly accessed site, which allows users to view and query a variety of cancer data types in the context of the widely used UCSC Genome Browser [https://genome-cancer.soe.ucsc.edu]. The Cancer Genomics Pathway Portal [http://www.cbioportal.org], hosted at Memorial Sloan-Kettering Cancer Center (MSKCC), is a recently launched site aiming to provide direct visualization and queries of summarized results by cancer biologists with little to no bioinformatic expertise as well as download of large-scale cancer genomics data sets for bioinformatic power users. For example, its major feature, Oncoprints, provides an easy way to visualize distinct genomic alterations (e.g., somatic mutations, CNAs, and mRNA expression changes) of a gene or genes of interest across a set of tumor samples. Another tool is the Integrative Genome Viewer [IGV, http://www.broadinstitute.org/igv], a scalable and readily browsable interface for accessing any type of cancer genome data, including mutations, expression, and copy number, particularly suitable for second-generation sequencing data. IGV also allows browsing clinical annotations alongside the genomic data. In addition to local data, a user can load data from a server that contains many data sets, including open access TCGA data. IGV can be downloaded and launched from a desktop computer. Similarly, for viewing and manipulating cancer copy number data and gene expression data to generate heat map figures, the dChip software system [Li and Wong 2001] can be installed on a standard desktop or laptop computer.

To perform customized analyses beyond querying summarized results, many open source analytical tools requiring only basic programming and bioinformatics expertise are now available. For example, Bioconductor (Gentleman et al. 2004) is an open source site with many useful analytical packages, written in the R language [http://www.r-project.org], for the analysis and interpretation of cancer genomic data, including second-generation sequencing data. Another site is GenePattern (Reich et al. 2006) [http://www.broadinstitute.org/cancer/software/genepattern], which is a user-friendly Web-based interface for >125 different genomics analysis tools and pipelines for various types of data, including gene expression, copy number data, proteomic data, and others. GenePattern can also keep track of parameters and versions of tools, and thus achieves an important goal of enabling reproducible research. The Gene Ontology site [http://www.geneontology.org/GO.tools.microarray.shtml] also maintains a collection of open source analytical tools contributed by its consortial members for analyses of microarray-based expression data.

Computational considerations in cancer genome analysis

In addition to specific challenges inherent in analysis of each type of cancer genome data, several general considerations should be kept in mind when one analyzes, interprets, and uses cancer genomics data. These include (1) quality control (QC) of data, (2) the accurate estimation of signal and noise in large data sets, (3) reproducible approaches to complex genomic analyses, and (4) achieving sufficient power in the face of multiple hypothesis testing. We briefly touch on each of these issues, but recommend several books for a broader introduction to bioinformatics; these include ones that are focused on principles and applications (Xiong 2006; Pevsner 2009), as well as on computational methodologies (Jones and Pevzner 2004).

Before analyzing genomic data, either publicly available or locally produced, and generating hypotheses for further experimental follow-up, one has to ensure the data are of sufficient quality. Two key aspects contributing to raw data quality are biospecimen and technical execution. Criteria used for biospecimen inclusion and exclusion can influence data quality; for example, a tumor specimen with a high proportion of stromal contamination will reduce one's ability to detect somatic alterations in the tumor cells. On the data generation front, standardization of technical methods [i.e., standard operating procedures (SOPs)] and execution by highly trained
individuals can minimize experimental variation and ensure reproducibility and high data quality. Once data are generated, the normalization step (Level II) attempts to remove any experimental artifacts that can negatively impact the data quality; however, some level of batch effects often remains and, if not controlled, may lead to spurious findings. A simple approach for detecting batch effects and estimating their extent is to use standard supervised methods to search for differences between batches; e.g., use SAM, ComparativeMarkerSelection, or ANOVA to look for genes that are differentially expressed between batches.

Another potential problem with large genomic data sets beyond data generation is sample mismatching, hence, it is recommended that one double-checks that the data originated from the intended sample. This is particularly important when analyzing multiple data types, since a mix-up could happen in some but not other data types. A useful approach to screening for sample mismatching is to leverage our understanding of the data types. For example, one can [1] correlate copy number and expression data, expecting overall positive correlation; [2] correlate expression and promoter methylation data with the expectation of finding a negative correlation; and [3] detect a drop in expression levels of specific genes in samples harboring truncating mutations (nonsense and frameshift insertions or deletions). With second-generation sequencing for somatic alterations, a common QC check is to compare the SNP genotype profiles between array-based and sequencing data to ensure that there is a match between tumor and germline DNAs, as a mismatch will result in erroneous calls of somatic mutations.

Like any other type of experimental data acquisition and analysis, maximizing signal and minimizing noise is essential in cancer genomics. In particular, using noise filters to remove poor-quality samples helps to achieve the most accurate analysis (Kauffmann and Huber 2010). Platform-specific noise measures can be used to optimize data quality for each type of genomic measurement. With such measures, the analysis of independent genomic data sets has proven to be reproducible across multiple independent laboratories, for example, for microarray-based gene expression analysis of human cancers (Dobbin et al. 2005).

Another key element in any computational analysis is reproducibility, a concept that is familiar to experimentalists, but the framework to ensure reproducibility in computational analysis is still in development. Naturally, the reproducibility standards expected from in silico experiments are beyond what is possible from bench experiments, since the initial raw data are available and computational tools (at least deterministic ones) should always reproduce the same values for all measurements. For example, if 200 genes are reported to be differentially expressed between two tumor subtypes, one would expect to obtain the exact same list of 200 genes when reproducing the analysis using the same tool, parameters, and input data. To aid in reproducibility, one must have a clear record of all inputs, parameters, and data manipulations for tracking and debugging. Such information will permit one to identify, for example, an input error that results in mislabeling of subsequent samples. Therefore, some of the key elements in reproducible bioinformatics include associating each analysis with a freeze of not just the data, but also the analytic software (codes) and parameters [Mesirow 2010]. For example, recording the specific reference genome build used for a particular analysis is critical to one’s ability to reproduce a result using the same input data. Increasing use of tools that have automated version tracking capability will greatly enhance reproducibility.

Unlike traditional molecular biology experiments, where only relatively few measurements are generally made (<10) in any single experiment, the number of individual measurements in a cancer genomics study is in the thousands to millions. This scale requires not only a large number of samples to achieve statistical power, but also addressing the issue of multiple hypothesis testing. Statistical power is a concept discussed in general publications on biostatistics, but it is also specifically illustrated in the case of whole-exome sequencing (Getz et al. 2007; Hudson et al. 2010), where it was calculated that 500 tumor samples are needed to detect, with ~80% power, genes that are mutated in 3% of patients (assuming a typical background mutation rate). In other words, a small discovery cohort will have very little power to detect infrequently mutated genes. Therefore, a lack of somatic mutation in a gene of interest in a study of a small number of samples should be considered non-informative, rather than interpreted as the gene is not mutated in cancers.

The concept of correcting for multiple hypothesis testing is very important in rigorous data interpretation of cancer genomics data, as it addresses the issue of false discovery. When thousands to millions of measurements are queried in each sample, significant differences will be observed, but the likelihood that such observed differences would occur by chance is extremely high. For example, when searching for differentially expressed genes by comparing two transcriptomes of 20,000 coding genes, a typical P-value cutoff of 0.05 will mean that, regardless of biological relevance, 5% of 20,000 genes (1000 genes) will be identified as differentially expressed by chance alone. To compensate for this, a q-value or false discovery rate (FDR) [Benjamini and Hochberg 1995] is calculated to bound the expected fraction of false discoveries in the results. In the above example of comparing two transcriptomes, if a set of genes in the differentially expressed list has a calculated FDR value of 0.2 (a commonly used cutoff) or less, it implies that the expected fraction of false discoveries among the list of differentially expressed genes is, at most, 20%. The FDR approach has become widely used in cancer genome analyses, as controlling the FDR is a more liberal approach than the standard Bonferroni correction, which bounds the chance of having even a single false discovery.

In summary, when using publicly available cancer genomic data, one should pay attention to the design of the study that generated the data; when expertise is available,
it is recommended that one downloads the raw data to repeat normalization and other QC checks prior to analyses. Analysts performing analyses, whether using off-the-shelf tools or developing new algorithms, must be mindful of reproducibility. With respect to interpretation of results, one must keep in mind the statistical power of a particular study and correct for multiple hypothesis testing. Finally, the most stringent test to assess the significance of these conclusions is the ability to reproduce the findings in independent data sets.

**Integrative analyses of the cancer genomes**

Cancer genomic data remains incomplete due to limitations of the current experimental methods as well as the inherent complexity of its biology. One method to identify genes of particular interest involves integrating the outputs from different types of experiments. For example, finding that a gene is targeted for genomic deletion, inactivating mutations, promoter hypermethylation, alterations of miRNA expression, and/or transcriptional down-regulation in different tumor samples would collectively suggest that this gene is a candidate tumor suppressor gene, even if each type of genomic alteration may be infrequent.

By extension, one can interrogate several types of data derived from the same set of tumors for evidence of dysregulation in a functional complex or pathway based on known constituents. The recent work in clear cell renal cancer is an excellent example of integrative analysis of a functional complex [Dalgliesh et al. 2010]. Here, the finding of low-frequency (3%) mutations in several enzyme-coding genes (SETD2, UTX, and JARID1C), all implicated in modifying regulatory lysine residues on histone H3, was interpreted collectively as evidence pinpointing the deregulation of H3 histone modification as a likely cancer-promoting process in a significant proportion of clear cell RCC.

An example of pathway integration is the analysis of GBM by TCGA. In this study, hundreds of clinically annotated GBM with matched blood normal (white blood cells) as a reference were characterized for (1) somatic mutation in 600 known or candidate cancer genes, (2) global SCNA patterns, (3) mRNA and miRNA expression, and (4) DNA promoter methylation status. By analyzing mutations and SCNAS in the same samples, TCGA showed that nearly all GBM harbored alterations in some components of the receptor tyrosine kinase/PI3K/RAS, p53, and RB/cell cycle pathways [The Cancer Genome Atlas Network 2008], providing definitive genomic evidence of the deregulation of these core pathways as obligate events for this cancer. In addition, by integrating sequencing and copy number data with expression profile analyses, TCGA linked major molecular subtypes of GBM defined by transcriptional profiles to specific genotypes [Verhaak et al. 2010]. In addition, analysis of DNA promoter methylation defined a further subclass of GBM exhibiting a CpG island Methylator Phenotype linked to IDH1 mutation and better survival [Noushmehr et al. 2010].

**Cross-species comparative oncogenomics**

The structural complexity of human cancer genomes has motivated efforts to leverage different data sets involving complementary information to identify somatically altered genes that likely contribute to tumorigenesis. In addition to integrating information derived from the same samples, the comparison of genetic alterations found in murine cancer models with those found in human cancers provides another method to identify cancer drivers. The rationale for the use of model organisms rests on the view that truly important driver genes and their linked mechanisms will be evolutionarily conserved, while bystander events not linked to biological processes are less likely to be shared in these cross-species comparisons.

Several studies have successfully leveraged the mouse cancer genomics along with human cancer genome data to identify novel cancer genes [Kim et al. 2006; Zender et al. 2006]. For example, Zender et al. (2006) compared regions that were amplified in both murine and human liver cancer genomes and found that cIAP1 and Yap are coamplified. Similarly, Kim et al. (2006) investigated a focal amplification in a murine model of melanoma that acquired new metastatic capability and found NEDD9 as a metastasis gene that is the target of the large 6p23 regional gain observed in ~30% of human metastatic melanomas. In a more recent example using second-generation sequencing technology, a mouse mammary tumor was found to carry an internal deletion of exons of the Lrp1b gene, an event similar to internal deletions found in the corresponding human ortholog in ~4% of human cancer cell lines [Varela et al. 2010]. Unlike what is usually observed in human cancer genomes, most mouse cancer genomes in genetically engineered mouse (GEM) strains harbor relatively few structural and copy number aberrations. Although this species difference limits the number of events available for comparison across the species, it also means that the presence of a focal amplification or deletion in a mouse cancer genome reflects strong selective pressure, thus providing strong evolutionary evidence that implicates the syntenic event in humans as a likely driver event. Indeed, when the mouse genome is engineered to experience telomere dysfunction and consequent genome instability like most human cells, the resultant tumors acquired a human-like genome harboring complex rearrangement and alterations that are syntenic to loci altered in human cancers [Maser et al. 2007]. These examples provide a rationale for cross-species comparative oncogenomics as an efficient strategy for the annotation of human cancer genes.

A complementary way to use mouse cancer models to identify cancer genes involves the use of forward genetic screens. For example, Berns and colleagues [Uren et al. 2008] have created cohorts of mice containing retroviral insertions for identification of genes that cooperate in a specific genetic background to drive tumorigenesis. In the setting of Ink4a/Arf or p53 deletion, they have identified many candidate oncogenes and tumor suppressor genes that corresponded to known human cancer-associated
mutations. Similarly, Copeland and Jenkins (2010) have deployed the Sleeping Beauty transposon system to generate both activation and inactivation of genes in the murine genome, and have used this system to discover genes involved in colorectal and other cancers (Dupuy et al. 2009; Starr et al. 2009). In another recent study, genome-wide copy number profiles of hundreds of human cancer cell lines were compared with hundreds of common insertion sites isolated from >1000 mouse tumors induced with the murine leukemia virus (MuLV), showing a significant enrichment of human ortholog genes with known mutations in COSMIC and overrepresentation of human orthologs that are annotated as oncogenes in the Cancer Gene Census [http://www.sanger.ac.uk/genetics/CGP/Census, Futreal et al. 2004]. Taken together, these efforts thus far have demonstrated the power of cross-species comparisons as a powerful approach to identify cancer genes, and argue for systematic and comprehensive genomic characterization of appropriately engineered mouse models of human cancers.

Converting genomic data into biological knowledge

The current large-scale cancer genomics efforts will identify and enumerate the frequency of every genetic element of interest that is structurally altered in the cancer genome, including those impacting annotated genes, noncoding miRNA, or other conserved elements. While statistical significance based on frequency will be an important filter to cull bystanders from likely driver events, it is likely that mutations that occur at a lower frequency may also contribute to specific types of cancer. For example, mutations or translocations involving ALK occur in 4%–5% of non-small-cell lung cancers (NSCLCs) (Soda et al. 2007), but confer sensitivity to small-molecule ALK inhibitors (McDermott et al. 2008). In such cases, additional experimental evidence is necessary to identify such low-frequency events as contributors to cancer initiation or progression (Chin and Gray 2008). Indeed, since it is likely that multiple genetic alterations are required to program the behavior of any specific cancer, functional analyses will be required to complement genome annotation. For example, the recent demonstration that the mutation status of KRAS dictates the response of tumors that harbor mutant EGFR to treatment with EGFR inhibitors and the complex interplay between BRAF, CRAF, and RAS in response to selective BRAF inhibitor (Heidorn et al. 2010; Joseph et al. 2010; Poulilakos et al. 2010; Whittaker et al. 2010) confirm that further functional studies will be required to exploit knowledge of somatic mutations. In addition, genes that are not mutated in cancers may also contribute to the survival of cancers harboring other genetic alterations, such as the PARP1 gene, whose inactivation is synthetically lethal in breast and ovarian cancers that lack BRCA1 or BRCA2 (Fong et al. 2009, 2010). Thus, functional analyses of cancer genomes are needed to complement structural analyses of cancer genomes.

High-throughput evaluation of gene function in cancer

In addition to laying the foundation for analysis of the cancer genomes, the information provided by the human genome project has facilitated the development of reagents to perform comprehensive somatic cell genetics in mammalian cells through the systematic manipulation of gene expression. Specifically, libraries that permit the expression or suppression of the majority of human or murine genes now exist in several formats. Although these tools can be used to study nearly any aspect of biology, early studies focused on cancer phenotypes have not only provided a proof-of-principle demonstration of the utility of such systematic functional approaches, but have also begun to provide insights into novel aspects of cancer biology.

Expression-based studies Although cDNA libraries have been used for many years to identify genes whose overexpression confers specific phenotypes [Seed and Aruffo 1987; Lin et al. 1991; Wang et al. 1991], and played a key role in the discovery of some of the first oncogenes [Shih and Weinberg 1982], such screens have often been limited by the efficiency of gene transduction as well as the representation of genes present in such libraries. As such, most successful screens involved positive selection strategies. However, several vector systems now exist that permit high-efficiency transduction of cDNAs into a wide range of mammalian cells [Koh et al. 2002], and increasingly complete collections of human cDNAs or ORFs now permit more comprehensive gain-of-function screens. Probably the most dramatic recent example of such a gain-of-function screen is the discovery of the EML4-ALK translocation in NSCLC by a retroviral expression system (Soda et al. 2007).

Using a positive selection screening strategy, several groups have identified genes that bypass specific anti-proliferative signals. For example, building on work that identified the retinoblastoma and p53 signaling networks as key regulators of cell proliferation and survival, TBX2 was found as a gene amplified in breast cancer that permits cells to proliferate in Bmi1-deficient fibroblasts [Jacobs et al. 2000], DRIL1 was discovered as a gene that bypassed RAS-induced senescence [Peeper et al. 2002], and BCL6 was found to permit cell proliferation in the presence of active p19ARF/p53 signaling [Shvarts et al. 2002].

This approach has also been used to identify genes involved in other cancer-related phenotypes. For example, the prostate-derived ETS factor gene SPDEF was identified as a gene that permits immortalized mammary epithelial cells to invade and migrate [Gunawardane et al. 2005]. SPDEF was subsequently found to be overexpressed in both breast and prostate cancers, and to cooperate with receptor tyrosine kinase genes such as ERBB2 and CSF1R to drive cell transformation.

In each of these examples, the cDNA libraries used by these investigators were derived from cell lines by reverse transcription of mRNA. Although this approach has been used successfully, several limitations of this methodology are that each gene is not represented at equal frequency in the library, longer cDNAs are underrepresented, and the
full repertoire of splice variants is not present. With the development of large collections of ORFs (Table 4, Lamesch et al. 2007; Rolfs et al. 2008; Varjosalo et al. 2008), it is now possible to create expression libraries in which at least one splice version of every gene is present at equal numbers. For example, analysis of a relatively small cDNA library targeting 353 kinases identified IKBKE as a breast cancer oncogene that substitutes for AKT to permit cell transformation (Boehm et al. 2007). In addition to this example of intersecting hits from a genetic screen with genomic profiles of human cancers, another approach is to create customized libraries of candidate genes—e.g., a library of amplified genes or mutated genes—to assess which ones have oncogenic activities. This type of approach can now be expanded to hundreds, rather than tens, of genes that emerge from genomic efforts, so one can enlist many candidates into a gain-of-function genetic screen for functional activity.

**Loss-of-function approaches** Similar to the cDNA or ORF libraries used for gain-of-function approaches, libraries of RNAi reagents can be introduced into cells either stably or transiently. In mammalian cells, RNAi-mediated gene suppression can be induced by the introduction of chemically synthesized siRNAs or plasmids expressing RNA hairpins, known as shRNAs, which are processed to siRNAs by Dicer (Bernstein et al. 2001). Chemically synthesized siRNAs are available from many different commercial sources as individual reagents, as pools targeting specific genes, or as genome-scale libraries. In general, siRNAs are easily synthesized and highly effective in inducing gene knockdown. However, such oligonucleotide reagents are relatively expensive and can be used only for transient loss-of-function experiments.

Vector-based systems to express RNAi provide several advantages compared with siRNA. By creating viruses, these vectors permit long-term, stable expression of the RNAi construct and expand the range and type of cells into which such constructs can be introduced. Both academic and commercial groups have produced large libraries of shRNAs in a variety of expression vectors (Table 4, Brummelkamp and Bernards 2003; Paddison et al. 2004; Silva et al. 2005, 2008; Buchholz et al. 2006; Moffat et al. 2006; Luo et al. 2009). Each of these systems has unique features, including high-efficiency infections, ease of recombination-based cloning, and inducible expression. In addition, these vector-based systems are useful for both arrayed and pooled screening approaches.

Both siRNA and shRNA libraries have been used successfully in arrayed screens (Aza-Blanc et al. 2003; Kittler and Buchholz 2005; MacKeigan et al. 2005; Whitehurst et al. 2007). For many other cancer-related phenotypic assays—such as anchorage-independent colony formation, bypass of senescence, or tumor xenografts—long-term gene suppression is essential, requiring stable integration and expression of the RNAi vector. Recent work from several laboratories has shown that these approaches are tractable in human cells. For example, PITX1 was found as a negative regulator of RAS signaling (Kollischoten et al. 2005), REST1 has been identified a negative regulator of PI3K signaling (Westbrook et al. 2005), CDK8 has been identified as a regulator of β-catenin signaling in colon cancer (Firestein et al. 2008), SIK1 was found to be a negative regulator of anikosis and metastasis (Cheng et al. 2009), and CDK6 has been shown to be an oncogene in GBM (Wiedemeyer et al. 2010). Although such arrayed format screens require assays that are amenable to well-based miniaturization, this experimental design permits the use of high-content imaging to identify subtle or complex phenotypes such as changes in cell morphology [Moffat et al. 2006; TR Jones et al. 2008].

In addition, vector-based shRNA libraries can be used to interrogate gene function in a massively parallel manner by creating pools of shRNAs. The advantages of this approach are that such pooled screens permit the study of a larger number of genes with decreased cost and provide the possibility of using loss-of-function genetics in assays that cannot be performed in vitro. Several large-scale screens using pooled libraries have been performed (Brummekamp et al. 2006; Ngo et al. 2006; Luo et al. 2008; Schlabach et al. 2008), demonstrating that both positive and negative selection screens are possible using these formats. To facilitate the deconvolution of genes targeted by shRNAs in these screens, each of these groups has developed strategies to quantify the abundance of each shRNA at the beginning and end of each screen by

<table>
<thead>
<tr>
<th>ORF/cDNA collection</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFCI Center for Cancer Systems Biology</td>
<td><a href="http://ccsb.dfc.harvard.edu/web/www/ccsb">http://ccsb.dfc.harvard.edu/web/www/ccsb</a></td>
</tr>
<tr>
<td>Harvard Institute of Proteomics (HIP)</td>
<td><a href="http://www.hip.harvard.edu">http://www.hip.harvard.edu</a></td>
</tr>
<tr>
<td>NEDO (FLJ)</td>
<td><a href="http://www.kazusa.or.jp/NEDO">http://www.kazusa.or.jp/NEDO</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>shRNA collection</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hannon-Elledge</td>
<td><a href="http://hannonlab.cshl.edu/index.html">http://hannonlab.cshl.edu/index.html</a></td>
</tr>
<tr>
<td>MISSION esiRNA</td>
<td><a href="http://elledgelab.bwh.harvard.edu/index.html">http://elledgelab.bwh.harvard.edu/index.html</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.broadinstitute.org/rna/trc">http://www.broadinstitute.org/rna/trc</a></td>
</tr>
</tbody>
</table>
been implicated in particular tumors. For example, processing, show an increased susceptibility to cancer the endo-ribonuclease that is required for miRNA pro-
to malignant transformation. Indeed, mice lacking Dicer, established that dysregulation of miRNAs contributes the targets of most of these miRNAs remain undefined regu-
lously through induction of transcript degradation or through expression of their target genes, either primar-
ously or transformed cells. The optimal use of these models requires an understanding of their ideal applications and experimental limitations, and the most predictive results will come from complementary uses of multiple models.

Model systems

A wide spectrum of model systems exists that permits the investigation of the context necessary for cell transforma-
tion and progression. The commonly used model systems include large panels of genome-annotated hu-
man cancer cells, early passage primary cancer cells from patients, genetically engineered immortal primary hu-
man cells, and GEM models and their derivative primary or transformed cells. The optimal use of these models requires an understanding of their ideal applications and experimental limitations, and the most predictive results will come from complementary uses of multiple models.

Established cancer cell model systems

Established human cancer cell lines and primary human cancer cell cultures have been used extensively in functional valida-
tion assays due to their ease of manipulation and versa-
tility. Although such systems only partially model the more complex biological features of cancers in vivo, they have proven powerful in advancing the validation of novel cancer genes, defining signaling pathways, and establishing pharmacogenomic relationships. A major challenge to the optimal use of these established cancer cell line panels has been the lack of a complete atlas of the genetic alterations in these cells, as it is appreciated that the specific genotypes in these cell models will dictate or modify the response to any molecular (RNAi or over-
expression) or pharmacological perturbation. Using the same tools described above to characterize tumor ge-
nomes, several groups are engaged in characterizing large panels of cell lines, such as the Cancer Cell Line Project [http://www.sanger.ac.uk/genetics/CGP/CellLines], or the Cancer Cell Line Encyclopedia Project [http://www.
broadinstitute.org/ccle]. Additionally, when multiple in-
dependent cell lines with molecular diversity are used, the risk that the observed phenotypes are idiosyncratic to a particular cell line can be minimized. Although these cell line models are powerful in throughput, they do not capture all molecular subtypes of a particular tumor type, nor do they retain any interaction with stromal microen-
vironment, thus raising the possibility that certain gene functions will be audited or represent artifacts of in vitro biology. Therefore, re-enforcing data from other model systems will continue to be important to complement information derived from established cell lines.

Nontransformed genetically engineered cells

Primary human cells engineered with initiating events that are insufficient to achieve full transformation represent a powerful system to validate novel genes. In particular,
such experimental models permit the creation of isogenic cells with specific mutations found in specific tumor subtypes, enabling the investigation of the role of a candidate cancer driver in a stringently defined genetic context. Such models can complement the use of established human cancer cell lines to interrogate functions of a candidate gene through both suppression by RNAi in engineered with signature mutations encountered in engineered stem and progenitor cells from GEM models, termed stem transgenesis systems as hematopoietic stem and progenitor cells are readily isolated from bone marrow or fetal livers, concep- tually similar approaches can be applied to other organ systems, such as liver or brain, as stem or progenitor cells have been identified, isolated, and transplanted in these systems [Bachoo et al. 2002; Zender et al. 2006; Zindy et al. 2007].

**Approaches to functional validation**

To fully credential a candidate gene as a therapeutic target or a diagnostic biomarker requires extensive functional assays and downstream biological studies that are time- and labor-intensive [Chin and Gray 2008]. One approach to validation is to begin with assays that offer throughput rather than depth, and move only the validated ones onto lower-throughput labor-intensive assays that provide insight into specific biological aspects. However, as noted, every assay—whether it is low or high throughput—will yield biological false positives and false negatives due to the context-specific nature of gene function.

Context can relate to cellular, genetic, and microenvironmental factors. One well-known example is the opposing roles of TGFβ signaling in initiation versus progression [Massague 2008]; hence, it is possible that the role of a candidate cancer gene may be oncogenic or tumor-suppressive, depending on the specific cellular or developmental contexts. Alternatively, a candidate may require cooperation of another genetic event to manifest its oncogenicity. An example is NEDD9, a metastasis gene that resides on chromosome 6 that is frequently gained in melanoma [Kim et al. 2006]. NEDD9 exhibited robust invasion activity in a Modified Boyden Chamber assay only in melanocytes harboring RAS mutations, consistent with its focal amplification in RAS mutant melanoma cells that have acquired metastatic capacity in vivo [Kim et al. 2006]. Thus, the specific genetic and cellular back- ground is necessary to understand the context in which a candidate oncogene or tumor suppressor gene operates.

Biological false positives can also emerge as a direct consequence of the artificial nature of experimental models. For example, overexpression may induce phenotypes due to supraphysiologic levels of expression, or suppression of a gene may have a different phenotype in vivo. The combina- tion of both functional studies and information derived from the analysis of cancer genomes can help mitigate these concerns. In addition, clinicopathological validation using tumor tissue microarrays can be highly informative, offering added evidentiary support for cancer relevance by demonstrating the prevalence of dysregulation on DNA [by FISH] and protein [by immunohistochemistry or immu- nonfluorescence] levels in independent large cohorts of specific tumor types and of broad tumor spectrums.

Although we have at our disposal a series of assays that permit the assessment of a large number of candidates with high throughput, as well as ones that enable deeper interrogation with lower throughput, the existing repertoire of cancer-related assays remains incomplete. Even in assays that are well developed, cellular as well as genetic contexts are important in interpretation of the results. Hence, it is important that conclusions drawn from these functional assays performed in experimental model systems are validated in human cancer specimens.

**Conclusions and challenges**

Comprehensive characterization of human cancer genomes will soon generate comprehensive lists of genomic and epigenomic alterations present in diverse human cancers. The integration of these genome characterization efforts—both structural and functional—promises to provide the foundation for a complete understanding of the somatic alterations that program cancer initiation, maintenance, and progression. Although the majority of early efforts have focused on known cancer pathways as a means to validate these methodologies, the full application of these approaches will identify new pathways and
networks necessary for establishment and maintenance of the malignant state. Moreover, such efforts will provide a framework on which to develop new therapeutics and rational combination regimens.

Although the throughput and reproducibility of methods for analyzing cancer genomes has improved at a rapid rate, many challenges remain. For example, collecting accurate clinical information on tumor samples remains an important and difficult task, but one that is necessary to afford the interpretation of genomic findings in a larger context. However, most available samples are associated with incomplete annotation or lack appropriate consent to permit the linkage of clinical information. Further progress will require the development and expansion of an infrastructure to collect samples and associated data with appropriate consent. The challenges—scientific, operational, and legal—of this process should not be underestimated.

In addition, it is clear that future efforts will need to account for genetic heterogeneity within tumors. As technologies improve in the near future, one will be able to explore patterns of molecular heterogeneity on the single-cell level to determine how such heterogeneity affects tumor biology and the response to treatment.

Finally, as emphasized above, integrative analyses of comprehensive cancer genomics data will generate hypotheses [such as candidate targets] that require experimental testing and validation. Such experimental validation results will guide further refinements of these analytical tools and approaches. The release of data sets prior to publication will aid in these efforts, but it is clear that more work is necessary to make these data sets available in useful formats.

In summary, cancer is the consequence of accumulated somatic genomic and epigenomic alterations within the tumor cells, influenced by their heterotypic interactions with a tumor microenvironment. The ability to catalog the somatic genomic alterations in large numbers of cancers, together with systematic functional analyses, will provide a foundation that will facilitate efforts to understand how these alterations induce malignant transformation. Moreover, as these genome characterization efforts are applied prospectively in patients, these efforts will provide a framework for relating studies in experimental models to the corresponding human tumors.

Acknowledgments

We apologize to colleagues whose studies, algorithms, or portals are not cited in this review due to limitation in space. L.C., G.G., and M.M. are TCGA-funded investigators [NIH U24CA143845, U24CA144025, and U24CA143867]. L.C. and W.C.H. are CTD2 [Cancer Target Discovery and Development] network investigators [NIH RC2CA148268].

References


Erratum


Making sense of cancer genomic data
Lynda Chin, William C. Hahn, Gad Getz, and Matthew Meyerson

In the above-mentioned review, the authors regret that a reference [Vicent et al. 2010] was inadvertently omitted. On page 547, in the second paragraph, the statement should read as follows:

“For example, several groups have identified genes that, when suppressed, lead to cell death only in the context of cells that are dependent on oncogenic KRAS. Specifically, TBK1, STK33, PLK1, WT1, and SNAIL2 have been identified as genes that are required for the survival of cells dependent on KRAS [Barbie et al. 2009; Luo et al. 2009; Scholl et al. 2009; Vicent et al. 2010; Wang et al. 2010].”

In addition, the Reference section should include the following:
