Chapter 9

Cancer Genomics

Bioinformatics for Cancer Genomics

Katayoon Kasaian
Yvonne Y. Li
Steven J.M. Jones

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

Contents

Cancer Genomics. DOI: http://dx.doi.org/10.1016/B978-0-12-396967-5.00009-8

Key Concepts

- Advances in DNA sequencing techniques have transformed the field of cancer biology. Using next-generation sequencing (NGS) technologies, we are now able to examine cancers on the molecular level and identify somatic mutations that have accrued during tumorigenesis. Analysis of these alterations can lead to the identification of molecular pathway(s) that are driving the disease
- Large volumes of data are generated as part of even the smallest cancer genomic experiments. Bioinformatic methods have provided solutions for easy data storage and access as well as sharing of data among different laboratories regardless of their geographical locations
- Different types of data can be generated as part of a cancer genomic experiment. Among these are whole genome, exome, transcriptome and epigenome sequencing datasets. Each type of experiment has its unique strengths and limitations; the most suitable technique for a particular experiment will depend on the scientific question in mind
- Somatic mutations in the tumor can be identified using the aligned or assembled sequence reads. These aberrations could involve only one base pair, i.e. single nucleotide variant (SNV), or could cover larger areas of the genome. Examples of such include indels, copy number variants (CNVs) and large structural variants (SVs)
- In the genomic analysis of a cancer sample, the challenge is to discriminate between driver and passenger mutations in a typically large pool of somatic variations. Driver mutations are those that give the tumor the capability to grow and divide without control while the passengers are merely the byproduct of the unstable cancer genome. Computational tools have been developed to aid this process; however, functional analysis of putative driver mutations is still a necessity
- The integration of different data types and mutation calls, with clinical information, as well as other "omics" datasets, will provide a comprehensive understanding of an individual's cancer and the disease pathway(s)

INTRODUCTION

Bioinformatics is an interdisciplinary field of study where computer science, statistics and information technology are applied to the field of biology. Advances in molecular biology tools and techniques have provided biologists with an unprecedented opportunity to gather large amounts of data. Bioinformatics offers not only a way to manage, store and easily access this information but also a way to visualize and analyze efficiently the data, enabling us to draw biologically significant conclusions.

Since the completion of the Human Genome Project (HGP), there has been a revolution in genomic technologies, particularly DNA sequencing techniques. Next-generation sequencing (NGS) technologies have transformed the field of cancer genomics; a complete human genome can now be sequenced at a high depth of coverage for a fraction of the cost and time it would take only a few years ago. Unravelling the genomic abnormalities that lead to cancer, potential therapeutic targets and the mechanisms behind tumor response or resistance to a particular treatment modality is integral to the advancement of cancer medicine. Therefore, the ultimate goal of the cancer genomics field is to explore fully the potential of these sequencing technologies in characterizing different types of cancer on the molecular level, understanding the mechanism of the disease, identifying diagnostic, prognostic and predictive markers and, finally, translating this knowledge into patient-based therapies. Computational biology and bioinformatic techniques provide solutions for examining complete genetic material of a cancer sample for every type of mutation including SNVs, insertions and deletions (indels), CNVs as well as SVs. The utility of the vast parallel sequencing machines is not, however, limited to analyzing the genome. The epigenome, the transcriptome, the proteome and the metabolome of a cell can all be investigated through these high-throughput technologies (Figure 9.1).

Bioinformatics helps to understand complex biological systems by systematically analyzing large biological datasets and by providing the necessary techniques for integrating different data types. This enables us to derive a global view of the healthy state of a cell and to identify how these are altered in the disease state. In this chapter, we focus on the bioinformatic algorithms and software applied in cancer genomics, particularly those used in the analysis of next-generation sequencing datasets. The algorithms and software discussed in this chapter are only illustrative examples of the more widely used tools and techniques in the field rather than a comprehensive survey.

DATA TYPES IN CANCER GENOMICS

Whole Genome and Exome Sequence Data

Cancers arise due to mutations that provide the cell with a growth advantage. In sporadic or non-familial cases of

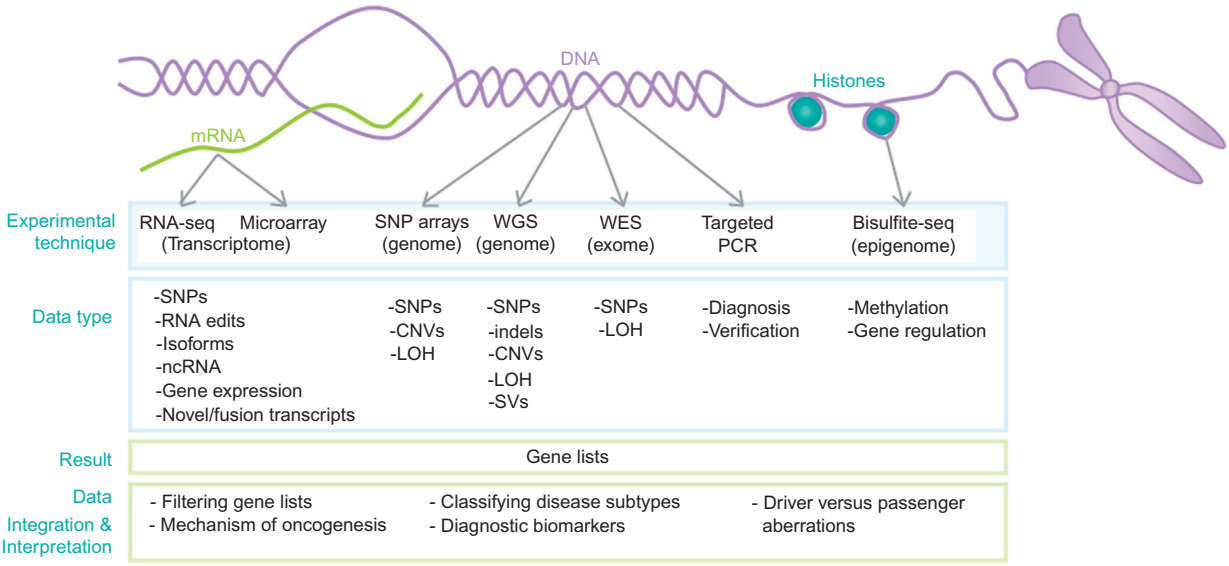

FIGURE 9.1 Through the application of high-throughput sequencing technologies, the genome, the epigenome and the transcriptome can be examined in great detail, providing a comprehensive picture of the state of health or any alterations leading to disease. Such experiments allow the identification of both small and large variations in individual samples.

cancer, these somatic events can be identified through the comparison of cancer and normal genomes of a patient. Whole genome shotgun sequencing provides the sequence of the complete DNA of a sample. These sequence data can be examined for the presence of various somatic alterations such as single nucleotide mutations, insertions and deletions, changes in copy number and large structural variations such as inversions, duplications, translocations and gene fusions. Although, whole genome sequencing provides the complete genetic landscape of a cell, most of the effort of the research community in cancer genomics has focused on identifying alterations in the coding regions of the genome. These are the mutations that change the sequence of the proteins and are thus more likely to alter protein structure and function and lead to cellular malfunctioning. This, along with the fact that whole genome sequencing is still not affordable enough to be carried out for individual patients in clinical settings or even in every research laboratory, has made whole exome sequencing an appealing alternative. Sequencing only the exons provides the information in the complete coding region of the genome at a high depth and for a lower cost than whole genome sequencing. Currently, the sequencing technologies provide such high sequence coverage that multiple exome libraries can be indexed, pooled and sequenced in a single experiment without losing any information while decreasing the cost even further. Whole exome sequence data can still unveil small mutations such as SNVs and indels. Recently, a few tools have been developed that promise the identification of regions of copy number loss or gain as well as coding structural variations from the exome capture data. However, most of the progress to date in finding somatic CNVs and SVs has been the result of whole genome sequencing experiments. This is due to change as more advanced algorithms and software become available.

Examining the cell's DNA provides a static view of the mutations that could potentially be disrupting protein functions. However, cells are dynamic entities, transcribing and translating the genetic information into protein products in accordance to their needs. Studying the dynamic profile of the cell through transcriptome sequencing or characterizing the protein collection of the cell can serve as a powerful tool for identifying disrupted pathways in a disease state.

Whole Transcriptome Sequence Data

It has long been known that there is a global change in the expression of genes in cancer cells compared with their normal counterparts. Some of these alterations, such as changes in the expression of oncogenes and tumor suppressors, will be drivers of the disease while others are the result of the malfunctioning cell and the fragile cancer genome. Using NGS technologies, the complete transcriptome of a cell can now be sequenced, providing a digital count of the expression of all genes. Through whole transcriptome sequencing, also referred to as RNA-seq, expressed mutations such as SNVs and indels are identified. *De novo* assembly of transcriptome data can also serve as a powerful tool for identifying events such as novel transcripts, skipped exons, retained introns or novel

TABLE 9.1 Advantages and Disadvantages of Different Data Types in Cancer Genomics

Data	Analysis Type	Advantages	Disadvantages
Whole genome shotgun sequencing	CNVs Indels SNVs SVs	Comprehensive interrogation of mutations	Costly No information on expression status
Whole exome capture sequencing	Indels SNVs	Cost efficient	Restricted to known annotations Detects only small coding mutations
Whole transcriptome shotgun sequencing	Expression Indels SNVs SVs	Cost efficient Digital gene expression Detects novel events	Detects only expressed alterations

splicing events. Differential expression analysis between malignant and adjacent normal tissues can shed light on the altered pathways in the disease and help in developing diagnostic and prognostic panels. However, such analysis in cancer genomics is hindered due to the typically limited access to neighboring matched normal tissue. Patient's blood usually serves as the normal sample and though it is a good reference for the tumor genome, the expression profile of the blood cells will be entirely different from that of a solid tumor, for example. Different data types in cancer genomics have varying strengths and limitations, generating as many different datasets as possible and their integration is the most promising solution in deciphering cancer signatures (Table 9.1).

Proteomic Data

Various high-throughput techniques such as protein microarrays and mass spectrometry have been developed for studying the complete collection of a cell's proteins, often referred to as the proteome. Proteomic analysis of a biological sample can unveil all the proteins present, their amount, specific post-translation modifications and all protein−protein interactions. Through such analyses of cancer and matched normal tissues or various cancer subtypes, one can identify diagnostic and prognostic biomarkers as well as novel drug targets. Our knowledge of the human proteome, however, has lagged behind the efforts such as the HGP which decoded the sequence of almost the entire genome. Understanding the structure and function of proteins is an important step in cancer genomics, leading to conclusions about the function of mutated proteins, whether they contribute to disease initiation and progression and how they can be targeted. The Human Proteome Project, launched in 2011, aims to identify the structure and function of at least one protein product of each protein-coding gene (http://www.hupo.org/research/hpp/) [1]. Such efforts combined with improvements in DNA sequencing technologies and integrative analyses tools will pave the way for delivering targeted therapies to patients.

Epigenomic Data

Next-generation sequencing technologies have also enabled the study of the epigenome, the transcriptional control of the cell. Mutations of several epigenetic enzymes are found in various cancers and thus there is increasing evidence that changes in the epigenome and the resultant alterations in the expressional profile of the cell could be the cause of many diseases including cancers. Examining the pattern of epigenetic marks associated with both the DNA and histone proteins throughout the whole genome of the cancer and matched normal tissue can provide a profound understanding of the changes leading to the disease state. Chromatin immunoprecipitation followed by sequencing (CHIP-seq) [2], with higher throughput and better sensitivity than CHIP-on-chip [3], provides a genome-wide view of specific DNA−protein interactions including histone modification marks. Profiling the methylation state of the genome is also now possible through techniques coupled with high-throughput sequencing [4]. These methods are divided into those which enrich for methyl-DNA [5−7], those which utilize methylation-dependent restriction enzymes [8,9] and the third category which is based on direct bisulfite conversion [10−16].

Data generation has arguably become the easiest and the most efficient step in studying a cancer genome. The challenge now is to analyze the sheer volume of data generated by the high-throughput technologies and to be able to integrate different types of mutational datasets such as SNVs, indels, CNVs, SVs, expression profiles and epigenetic alterations in order to draw a biologically correct

and meaningful conclusion about the underlying cause of the disease and how best to treat it.

DATA MANAGEMENT

A vast amount of information is gathered as part of cancer genomic initiatives such as The Cancer Genome Atlas (TCGA), the Pediatric Cancer Genome Project (PCGP), International Cancer Genome Consortium (ICGC) [17], Cancer Genome Project (CGP), Cancer Cell Line Project (http://www.sanger.ac.uk/genetics/CGP/CellLines/), Cancer Genome Characterization Initiative (CGCI), Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and the Cancer Cell Line Encyclopedia (CCLE) (http://www.broadinstitute.org/ccle/home) [18]. These large-scale initiatives are not restricted to studying diseases such as cancer but also include projects that examine the healthy population; examples include the HapMap [19], the 1000 Genomes Project [20], Personal Genome Project and the Human Genome Structural Variation Project [21]. In order to meet the collective goal of characterizing different cancers on the molecular level and identifying therapeutic targets, there is a need for efficient and rigorous management of the large and complex datasets generated mostly by the NGS technologies. Data collected as part of these studies include clinical data, raw genomic data as well as processed data such as short read alignment files and variant calls. An efficient data management strategy should allow easy data storage and access, as well as the ability to link different types of data. Cancer Genomic Hub (CGHub) (https://cghub.ucsc.edu/index.html), a national data center built by the University of California Santa Cruz (UCSC), is an example of such a centralized data storage system. It serves as the data repository center for TCGA and other related cancer genomic efforts; it is an automated resource that gives researchers access to sequence read alignment files as well as any available metadata.

In addition to raw and aligned data repositories, databases populated with derived information from large genomic studies are of considerable importance to the field. Great examples of these databases include dbSNP [22] where common variants in the general population identified through the HapMap [19], the 1000 Genomes Project [20] as well as other validated experiments are stored, and the COSMIC database [23] which stores curated somatic mutations identified in human cancers.

Tools such as NCBI's Entrez database browser [24] and BioMart Central Portal [25] provide a single point of access to several databases; this facilitates sharing as well as integration of data in the research community. Web services have also become indispensable tools in bioinformatics. Services provided by host institutes provide data and application access to scientists located anywhere in the world without having locally to install and maintain these resources.

The need for genomic data management is not restricted to large initiatives but also to smaller genome centers. Setting up automated pipelines for analyzing the data immediately after their production will ensure their efficient processing in a timely manner. Quality control and processing steps such as alignment of short reads to the reference genome, *de novo* assembly and variant calling would be part of such pipelines. These centers also benefit from efficient and robust data storage systems. Databases designed specifically for storage of cancer genomic variant calls will minimize redundancy, enable data integration with other databases and facilitate query of large data collections. As a result, such repositories can serve as powerful knowledge resources. They facilitate studies by which the profile of a specific point mutation, for instance, over thousands of sequenced cancer and normal libraries could provide the basis for a statistically based conclusion about its role in cancer and whether the affected gene could serve as a biomarker.

DATA ANALYSIS

Even more pronounced than the need for data storage in cancer genomics is the ever increasing need for powerful computational resources and efficient algorithms for data analysis and visualization.

Computational Resources

Currently, a single high-throughput DNA sequencing run using NGS technologies can produce 200 giga base pairs (Gbp) of sequence data. A tremendous amount of computational power is thus needed in every step of analysis, from the collection of raw data, to the alignment of reads to the 3 billion base pairs of the human reference genome, to variant calling and, finally, assembly. High-performance computing (HPC) facilities equipped with a cluster of high-speed computer nodes and multipetabyte storage systems can provide the hardware and the support needed for large-scale genomic projects. The HPC facility housed at the Michael Smith Genome Sciences Centre (http://www.bcgsc.ca/) with over 8000 cores/16 000 threads and 8.5 PB of storage system is an example of such an HPC system in Canada. Distributed and parallel computing, cloud computing and graphics processing unit (GPU) computing [26] are some of the examples of techniques and technologies that are moving to the spotlight as the genomics field is awaiting the arrival of single molecule sequencing technologies.

Analysis Algorithms and Tools

Sequence Alignment and Assembly

NGS technologies produce large numbers of short reads in a relatively short period of time. Application of these technologies in cancer genomics depends on the ability to reconstruct the complete genome from these reads with great accuracy in a time- and memory-efficient manner. Generally, two options exist: one is to align the reads to the reference genome and the other is to perform a *de novo* assembly.

Simply put, alignment refers to the task of finding the location in the complete genome where a sequence read was generated from. This is in essence a string-matching problem. There already exist various algorithms and software for solving this problem, all however facing a trade-off between accuracy and speed. Although the standard Smith–Waterman algorithm [27], widely used for the alignment of longer reads, provides the most optimal solution, it becomes computationally intractable when working with a large number of short sequence reads. As a result, a growing number of algorithms for the alignment of NGS reads to the human reference genome have been implemented. The most widely used techniques fall into two broad categories, hash table- and suffix tree-based algorithms. The former constructs a hash data structure for indexing the sequence data and fast look up of strings. These hash tables can be constructed from either the reference genome or the set of sequence reads. The unindexed set is subsequently used to scan the hash table for matches and to identify the approximate location of each read. Finally, dynamic programming algorithms such as Smith–Waterman and its improved equivalents are used to find the exact placement of each read. Some of the more widely used aligners in this category include MAQ [28], SOAP [29] and SHRiMP [30]. The second category of aligners uses a different set of data structures for indexing the sequence data. These generally rely on indexing the reference genome using suffix trees, suffix arrays or the more widely used Ferragina Manzini (FM)-index, a compressed data structure based on the Burrows–Wheeler Transform (BWT). Suffix trees, data structures, where all the suffixes of a string are stored, and suffix arrays, arrays of integers corresponding to the starting position of all possible substrings, are not efficient for indexing large search spaces such as the human reference genome. Ferragina and Manzini demonstrated that suffix arrays built from the BWT transformed strings are more efficient than arrays built from the non-transformed strings [31]. As a consequence, the FM-index is the data structure of choice in the majority of current short-read aligners. They allow for rapid substring search and are generally more memory efficient than suffix trees and arrays. Examples of such aligners include Bowtie [32], BWA [33] and SOAP2 [34].

The often ignored limitation of the current aligners is their intended use for the alignment of normal sequence reads only. These tools are not yet optimized for the alignment of reads generated from cancer samples which could potentially contain large insertions and deletions, as well as evidence of other structural variations such as duplications, translocations and gene fusions. In addition, as the sequencing technologies improve and the reads become longer, the majority of the tools for alignment of short reads to the reference will not be applicable anymore. There will be a need for specialized software to map optimally longer reads, perhaps containing indels or other structural variations, to the reference genome.

An alternative option to the alignment process is *de novo* assembly of sequence reads. Such an approach allows for the identification of highly diverged DNA regions in the sequenced sample compared with the reference genome. The techniques used in assembling longer reads produced by the early sequencing technologies, such as Sanger, generally involved finding areas of overlap between reads and extending those into longer contigs. Shorter reads and higher coverage produced by the NGS technologies, however, make such algorithms computationally inefficient, if not unfeasible. Currently, the more widely used assemblers make use of the de Bruijn graph data structure where all possible substrings of size k are stored in the nodes of the graph and each edge indicates an overlap of size $k-1$ between the two connecting nodes [35–38]. Traversing such a graph built from raw sequence reads will yield a collection of contigs representing the sample's sequence. *De novo* assembly techniques are not yet as computationally efficient as alignment of reads directly to the reference genome and, hence, not yet as widely used. Currently, the genomic analysis of a cancer and its matched normal involves separate alignment of each sample; this is followed by variant calling and the identification of novel somatic mutations in the tumor tissue. With advances in assembly algorithms as well as increases in read length and insert sizes of paired-end libraries, it is conceivable that *de novo* assembly of tumor and normal genomes will eliminate the need for alignment to a reference. As a result, this approach can provide more comprehensive insights into each individual's unique genomic landscape and pave the way for more personalized diagnosis and treatment options.

Discovery of Point Mutations

The alignment and/or assembly results are subsequently explored for the presence of any type of somatic mutation including single nucleotide variants. The majority of the

early SNV detection tools [39–41] rely on setting arbitrary thresholds for variables such as sequence coverage, read mapping quality, base quality and distance between mismatched bases in order to filter out technical noise and identify the positions that show true variability from the reference. These tools, however, are best suited for the analysis of normal samples and detection of germline variations where, for example, a heterozygote SNV would be expected to have variant allele frequency of 50% while, in a homozygote position, the variant base would be observed at 100% frequency. When analyzing tumor samples, contamination with adjacent normal tissue, the presence of multiple clonal populations within the tumor, as well as tumor aneuploidy can result in single nucleotide variants that are observed at any frequency. Probability model-based tools designed specifically for the detection of variants in cancer samples have been developed; these identify the most likely genotype at each position based on a probabilistic model for allelic distribution [42,43]. The dependence of all these tools on separate analysis of cancer and normal samples followed by their pair-wise subtraction has, however, deemed them as suboptimal in detecting somatic mutations. Recent developments in simultaneous analysis of matched sample pairs have resulted in more confident somatic mutation calls by calculating the likelihood of genotype differences between the two genomes, at all locations [39,44–46]. These algorithms allow for the detection of true somatic mutations which lack strong support in the tumor sequence data and distinguish them from false somatic events with weak support in the normal sequence data. The current state of cancer genomics requires the verification of computationally detected variant calls in their corresponding samples using orthogonal methods. In future, such verification may no longer be needed should advances in sequencing technologies and analysis tools lead to near optimal quality of reads and genotype calls.

Identification of Indels

Detecting small insertions and deletions (indels) from NGS short read products has proved more challenging than detecting single nucleotide variants. This is mainly attributed to the limitations of current aligners which, by default, allow a set number of small mismatches between a read and the reference, typically with no gaps, leading to misalignment or no alignment of reads spanning indels. Using split-read approaches, the Pindel software aims to detect large deletions and medium-size insertions from pair-end datasets [47]. Mapping short reads to repetitive regions and tandem repeats typically poses difficulties, leading to low sensitivity and specificity of the majority of indel detection tools. Gapped [33,48] and paired-end alignments [28,33,47] are methods that can improve detection sensitivity; parameters such as the number of reads supporting an indel, mapping and base qualities as well as presence or absence of homopolymer regions should be taken into account when estimating the true positive probabilities [39,40]. Dindel [49], the 1000 Genomes Project indel-caller [20], uses local realignment of reads to increase the accuracy of indel detection rate. Dindel accepts a list of potential indels and SNP calls as input, identifies all candidate haplotypes surrounding these sites and realigns reads to all the candidates in order to identify true events [49]. One limitation of Dindel, however, is its dependence on the sensitivity of the aligner that provides the initial list of potential insertion and deletions.

Indels, having the potential to alter or completely eliminate a protein's function, are the second most abundant type of variation in the human genome after SNVs [50]. As a result, there is a great need for development of robust probabilistic algorithms for detecting somatic indels from paired cancer and normal samples.

Structural Variation Detection

Structural alterations including large insertions and deletions, duplications, inversions, translocations and gene fusions have been associated with various cancer types [51]. Before the advent of NGS technologies, cytogenetics, karyotyping and fluorescent *in situ* hybridization, as well as array-bases techniques such as SNP arrays and array comparative genomic hybridization (CGH), were used in detecting large SVs. However, the emergence of next-generation sequencing technologies and the corresponding analysis tools has enabled the detection of various SVs including copy-neutral events and the corresponding break points at a much higher resolution and with greater accuracy.

Paired-end sequencing protocols, where the two ends of a single DNA molecule are read, allow the detection of SVs in the genomic data. Since the order and orientation of read pairs and the insert size distribution are known, any deviation from these expectations in the alignment might suggest a variation in the sample. Several tools have been developed which detect read pair anomalies and infer specific SVs [52–56]. However, we now know that the majority of structural variations are found in duplicated regions of the genome [21,57], regions that pose the most difficulty during the alignment process. As a result, alignment-based SV detection may result in many false positives while missing true events. An alternative to examining the alignment data for finding anomalies is to assemble the sequence reads *de novo* and compare the resultant contigs with the reference genome [58]. As the reads get longer, the assembly of individual genomes becomes more feasible and detection of SVs will have higher sensitivity and specificity.

Large deletions and amplifications, at times encompassing chromosome arms or whole chromosomes, lead to changes in number of gene copies and, in some cases, their expression levels. These structural variations are often collectively referred to as copy number variations. Variations in gene copy number can be detected using single-end as well as paired-end reads. Given the assumption that the whole genome is sampled uniformly and reads are generated with equal probability, depth of coverage can serve as a quantitative measure of copy number [59,60]. This assumption is not strictly correct, however. GC content, for instance, introduces bias during the sequencing experiment [61], while challenges such as alignment of short reads to repetitive regions of the reference genome leads to computational biases. Various techniques have been employed in identifying somatic CNVs by correcting these deviations from the expected distribution [62–64]. Several repositories containing somatic copy numbers from cancer datasets are publicly available [23,65–67]; these resources can facilitate the analysis of variation calls in one or multiple samples.

Expression Analysis

High-throughput sequencing of the complete transcriptome offers a few advantages over the more traditional means of expression analysis such as oligonucleotide microarray technologies. All expressed entities including novel transcripts, novel isoforms and non-human transcripts are sampled in these surveys of the whole transcriptome as opposed to microarray experiments which are restricted to known genes and annotations. Digital analysis of the transcriptome also increases both specificity and sensitivity; the high coverage that can be achieved through these experiments enables the identification of genes with even the lowest expression levels. Identifying differentially expressed genes or specific isoforms between malignant and normal states can reveal pathways which, when altered, might lead to tumorigenesis. Differential expression analysis can also identify subtypes of a disease and subsequently aid in finding diagnostic [68,69] and prognostic markers [70,71]. The discovery of novel subtypes of a disease can be accomplished through unsupervised clustering techniques such as hierarchical clustering of the expression data [72], use of self-organizing maps [73] or non-negative matrix factorization [74,75].

Expression analysis is not restricted to the cell's messenger RNA. Small non-coding transcripts such as miRNAs can also be subjected to high-throughput sequencing and analysis. Integration of protein-coding gene expression profiles with miRNA expression, promoter methylation and copy number variation data can provide clues to which genes are silenced and thus may function in tumor suppression, and which are overexpressed and might be acting as oncogenes.

Analytical Tools

In addition to computational infrastructure and software, programming languages and toolkits designed specifically for working with biological data are required for efficient analysis of genomic datasets. Analysis of variant calls, most likely stored in different file formats across a network, and their integration with externally available resources will require tools that allow easy data management and processing. Perl, Python, Java, C++ and R are some of the popular scripting and programming languages used in cancer genomics. Several toolkits designed specifically for biological data are developed for each of these languages; some of these include BioPerl (http://www.bioperl.org) [76], BioPython (http://biopython.org) [77], BioJava (http://biojava.org) [78], NCBI's C++ toolkit (http://ncbi.nih.gov/IEB/ToolBox/CPP_DOC/) and R BioConductor (http://bioconductor.org) [79] which provides powerful capabilities for statistical analysis of biological data. The majority of these resources are released under an Open Source license and thus can serve as building blocks for larger and more complex tools and software [80].

Visualizing the genomic data is an integral part of comprehensive investigation of cancer genomes. Tools such as genome browsers provide a graphical interface where the complete human genome as well as all available annotations of it can be explored. The widely used genome browsers in the field include the University of California Santa Cruz (http://genome.ucsc.edu) [81] and Ensembl browsers [82]. Various types of sequence data tracks can also be uploaded and visualized using tools such as Integrative Genomics Viewer (IGV) [83] and Circos [84]. The recently released cBio Cancer Genomics Portal [85] provides a platform for integrating and visualizing various types of cancer genomic data available through initiatives such as TCGA and ICGC; instances of the portal can also be installed locally (Table 9.2).

DATA INTERPRETATION

Over just a few years, the cancer genomics community has made great progress in developing algorithms and software for detecting various types of mutation from whole genome and transcriptome datasets. As improvements are made to sequencing technologies and detection tools, the most challenging task becomes mining the large and diverse mutational profiles that are generated in even a single patient experiment for mutations which contribute to disease initiation and progression.

TABLE 9.2 List of Bioinformatics Resources for Cancer Genomics

Description	Tools	URL
Programming languages & biology-specific modules	Perl, BioPerl	http://www.bioperl.org
	Python, BioPython	http://biopython.org
	Java, BioJava	http://biojava.org
	C++, NCBI's C++ toolkit	http://ncbi.nih.gov/IEB/ToolBox/CPP_DOC/
	R, BioConductor	http://bioconductor.org
Genome browsers & annotation resources	UCSC	http://genome.ucsc.edu/
	Ensembl	http://www.ensembl.org
	NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/projects/mapview/
Sequence data analysis toolkits	SAMtools	http://samtools.sourceforge.net/
	GATK	http://www.broadinstitute.org/gsa/wiki/
Population polymorphism data	dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
	1000 Genomes Project	http://www.1000genomes.org/home
	Database of Genomic Variants	http://projects.tcag.ca/variation/
Repositories of cancer gene & mutations	COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/
	Cancer Gene Census	http://www.sanger.ac.uk/genetics/CGP/Census/
	OMIM	http://www.ncbi.nlm.nih.gov/omim/
	Mitelman Database	http://cgap.nci.nih.gov/Chromosomes/Mitelman
Data from cancer genomic initiatives	TCGA	http://cancergenome.nih.gov/
	ICGC	http://www.icgc.org/
	PCGP	http://www.pediatriccancergenomeproject.org/site/
	Cancer Cell Line Project	http://www.sanger.ac.uk/genetics/CGP/CellLines/
	CGCI	http://cgap.nci.nih.gov/cgci.html
	CCLE	http://www.broadinstitute.org/ccle/home
Gene expression resources	GEO	http://www.ncbi.nlm.nih.gov/geo/
	ArrayExpress	http://www.ebi.ac.uk/arrayexpress/
	Oncomine	https://www.oncomine.org/resource/login.html
Viewers	IGV	http://www.broadinstitute.org/igv/
	Circos	http://circos.ca/
	Genome Browsers, e.g. UCSC	http://genome.ucsc.edu/
Pathway analysis tools & databases	KEGG	http://www.genome.jp/kegg/
	DAVID	http://david.abcc.ncifcrf.gov/
	IPA	http://www.ingenuity.com/
	Pathway Interaction Database	http://pid.nci.nih.gov/
	Reactome	http://www.reactome.org
	STRING	http://string-db.org/
Integrative data analysis	PARADIGM	http://sbenz.github.com/Paradigm/
	cBio Cancer Genomics Portal	http://www.cbioportal.org
Resources for therapeutic intervention	DrugBank	http://www.drugbank.ca/
	TTD	http://bidd.nus.edu.sg/group/ttd/
	PharmGKB	http://www.pharmgkb.org/

The list of putative cancer-related somatic mutations can be refined using publicly available reference databases. These include repositories where variations in the healthy population are curated such as dbSNP [22], Database of Genomic Variants [86] and 1000 Genomes Project datasets [20], as well as databases where known cancer genes and their mutations are stored. Examples of such include: the COSMIC (Catalogue of Somatic Mutations in Cancer) [23], an open source database containing somatic mutations and copy number alterations associated with cancers; OMIM (Online Mendelian Inheritance in Man), which collates information on familial cancer genes and susceptibility loci; Cancer Gene Census [87], which catalogs all genes shown to be causally implicated in cancers; and the Mitelman database of chromosome aberrations and gene fusions in cancer [88]. The common assumption when using these databases is that variations that are commonly found in the general population are less likely to contribute to diseases such as cancer, while genes recurrently mutated in various cancer types are potential oncogenes and tumor suppressors.

When interpreting genomic data, it is imperative to be aware of potential confounding factors. The presence of circulating tumor cells in a normal blood sample, normal cell contamination in a tumor sample or a heterogeneous tumor sample with various different subclonal populations can lead to false positive and false negative mutation calls. Bioinformatic algorithms have been developed to estimate and correct for the amount of contamination and to determine more sensitively the copy number variation (CNANorm) [89], SNVs (MutationSeq) [46], or loss of heterozygosity (APOLLOH) [90]. In some cases, manual inspection of tumor and normal sequence files using tools such as IGV [83] can help in eliminating false positive somatic calls. The next step following the computational discovery of candidate mutations typically entails verification of those events in their corresponding sample(s). This step identifies the variant calls that were falsely identified as somatic due to sequencing or computational errors. Mutation verification usually involves the amplification of the potential variant site in both cancer and matched normal tissues using polymerase chain reaction (PCR) techniques followed by Sanger or next-generation sequencing (Figure 9.2).

The number and profile of somatic mutations demonstrate great variability in different cancers [91]. A high number of mutations are seen in cancers such as melanomas [92] whereas some pediatric cancers exhibit a very low number of alterations [93]. Regardless, all somatic mutations can be categorized into “drivers”, changes that are responsible for disease pathogenesis and tumor evolution, or “passengers” which are simply the byproduct of the unstable cancer genome and provide no growth advantage to tumor cells [91]. Distinguishing these two types of entities is critical given that passenger mutations play no functional role in disease initiation, progression or maintenance and thus treatment(s) targeting them may prove ineffective. Although, gene(s) that are recurrently mutated in a large cohort of samples are easily distinguishable as drivers, the task becomes more difficult when examining a single or only a few samples. Computational techniques are developed that aim at distinguishing drivers and passengers *in silico* prior to the more labor-intensive and time-consuming procedures of functional validation in the wet lab. Since it is believed that mutations which result in changes in protein structure and function are more likely to act as drivers of cancer, the majority of the focus to date has been on mutations that affect protein-coding regions of the genome. A common computational strategy in examining the functional role of a somatic mutation is to determine its location with respect to functional domains and key amino acid residues in the protein product using resources such as UniProt. Software tools such as PolyPhen [94], MutationAssessor [95] and SIFT [96] use evolutionary conservation of gene sequences as well as homology to provide a likelihood score for the deleterious effect of a point mutation on protein structure and function. The general assumption that driver mutations providing growth advantage for the tumor must be under evolutionary positive selection while the passengers are less likely to be selected for has given rise to algorithms aiming to quantify the selection pressure on mutated genes in cancer samples. The algorithm developed by Greenman et al. [97] estimates the background rate of mutation in each gene using the set of silent mutations; any gene with more than the expected number of non-silent mutations might then be concluded to be under positive selection.

In addition to sequence-level analysis, genomic aberrations can be investigated on a structural level. This approach is limited to genes that have solved three-dimensional protein structures in the Protein Data Bank (PDB) [98] but can be extended to include genes with homology models, predicted protein structures of genes based on a solved structure of a similar-sequence gene. Repositories like SWISS-MODEL [99] have automatically built homology models for all known human genes where a solved structure of a similar gene exists. The protein and mutation can be visualized using molecular graphics software such as PyMOL (http://www.pymol.org/) or SwissPdb Viewer [100] and can provide a more realistic view of how the mutation might affect the catalytic site, nearby key residues, or protein−protein binding residues such as substrate/cofactor binding or dimerization sites. Conclusions based on structural analysis, however, are limited in that they do not take into consideration the

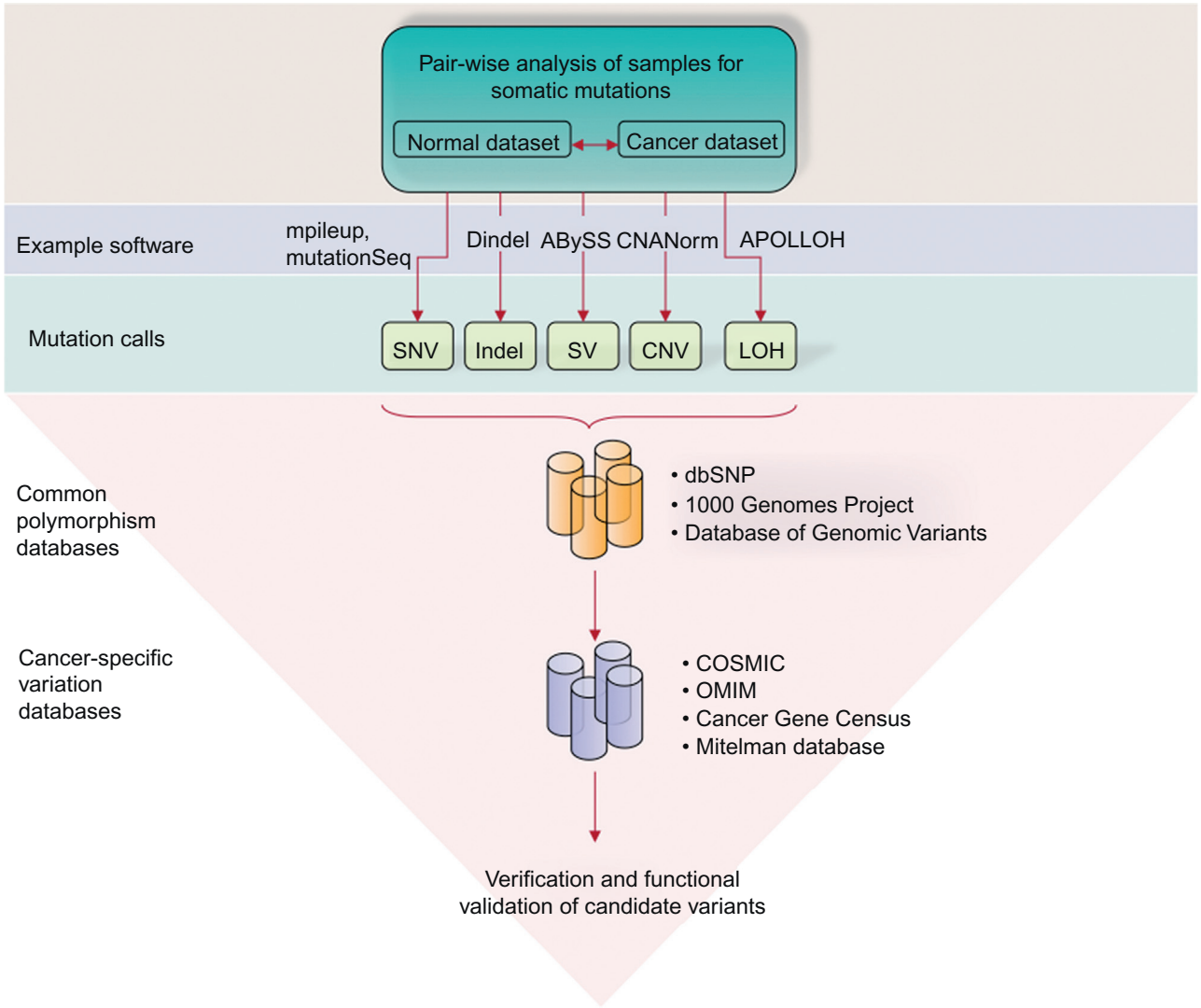

FIGURE 9.2 Filtering the identified somatic mutations in a cancer sample using publicly available databases of common genetic polymorphisms as well as known cancer-specific variants can narrow down the potentially long list of candidates to the most likely drivers.

actual protein conformation in the cell or residues that are key in retaining protein stability and/or play a role in protein−protein interactions.

Determining the deleterious effect(s) of mutations spanning more than one base pair is typically an easier task. Indels that occur in-frame add or remove at least one amino acid residue while out-of-frame indels can completely disrupt the sequence of the gene and hence the protein product. Inversions, duplications and translocations with break points in the middle of a gene can also alter the protein function. Changes in gene copy numbers and expression might be more difficult to interpret in whole genome and transcriptome studies. A single cancer sample could potentially have thousands of genes which are up- or downregulated compared with the normal tissue. Similar to SNPs, statistical approaches have been developed for identifying larger alterations such as CNVs that serve as disease drivers [101]. Integration of all mutational and expression data promises to provide a global view of the altered pathways and identify mutations that are indispensable to the cancer cell.

Although the majority of somatic mutations in any cancer sample fall outside the protein-coding regions, the scientific community for the most part has focused on examining the protein-coding changes. This is attributed not to a lack of interest but perhaps to a lack of comprehensive knowledge about the regulatory elements of the genome. Although it is more challenging to identify the role these non-coding mutations might be playing in cancer, integration of mutation data with known regulatory sites from resources such as the ENCODE data [102], TRANSFAC [103], JASPAR [104] and ORegAnno [105]

databases will be a first step in this direction. Mutations affecting 5′ and 3′ splice sites at the exon−intron junctions are another important category of non-protein-coding changes associated with cancer. A point mutation in these invariant dinucleotides could lead to loss of canonical splice sites while mutations in introns may create novel sites. Such alterations lead to aberrant protein products by retaining introns, skipping and/or producing new exons. The result of such aberrant splicing will be evident from the transcriptome data. Publicly available databases with known alternative splicing events [82], novel pathogenic exon boundaries [106], as well as resources integrating RNA splicing mutations and various diseases [107] are invaluable tools in analyzing cancer genomes.

Analysis of all the different data types in cancer genomics will generate one or more gene lists. These could catalog the differentially expressed genes, amplified or lost genes or genes with point mutations and indels, to name a few. They can vary in size from tens of mutations to thousands of amplified genes, to expression values of approximately 20 000 protein-coding genes. Bioinformatics analysis allows us to rank and select the most biologically interesting set of genes, cluster them into biologically significant subgroups, or select a subset for experimental verification or functional studies.

DATA INTEGRATION

Integration of all somatic mutation calls and expression data from a cancer sample plays an important role in creating a molecular pathway hypothesis of aberrations driving the tumor. The biology of a cell is complex and involves processes and controls of those processes on the genomic, epigenomic, transcriptomic and proteomic levels (Figure 9.3). Molecules in the cell are not isolated, but are part of a collective "system" of interacting parts [108], and aberrations in one molecule can perturb the whole system. Network construction and modeling allows the simulation of complex biological pathways including their temporal aspects. For instance, a human colon cancer cell network was built using cell signaling pathways to model and predict the phenotypic fate of the cell, and could predict changes to cell proliferation rates by inhibiting certain genes or proteins [109].

Data integration can be performed by intersecting different gene lists or by visualizing them together on a genome plot like Circos (Figure 9.4). Interrogating the affected genes for enrichment of Gene Ontology (GO) terms [110] or particular gene sets [111] along with pathway analysis using repositories and tools such as KEGG (the Kyoto Encyclopedia of Genes and Genomes) [112], DAVID (the Database for Annotation, Visualization and Integrated Discovery) [113] and IPA (Ingenuity® Systems Pathway Analysis, www.ingenuity.com) can provide evidence for the altered pathways that are driving the disease. Tools such as PARADIGM allow the combination of all types of high-throughput genomic and functional genomic datasets to infer a patient-specific mutated pathway [114]. A limiting factor in relying on existing databases for pathway analysis is the restriction of results to already known and annotated genes and networks.

When examining a cohort of samples, data integration and identifying commonly altered pathway(s) often becomes more significant than pinpointing recurrently mutated gene(s). Different samples could have mutations in various genes, all contributing to one biological process which, when disturbed, leads to tumorigenesis. TCGA pilot project, for instance, uncovered core mutated pathways in glioblastomas (GBM) by integrating sequence data, gene expression, copy number variation as well as epigenetic assessments [115]. Data integration in these cohorts can also identify genes which may be frequently altered through multiple mechanisms (point mutations, structural disruption, loss of copy, or hypermethylation of the promoter) but would not otherwise be identified through separate analysis of each data type.

By sequencing a patient's tumor and normal genomes, bioinformatic methods enable not only the discovery of aberrations and the potential oncogenic mechanism but also the determination of actionable targets and potential therapies [116,117]. Integration of mutation calls with drug databases such as DrugBank [118] or Therapeutic Target Database (TTD) [119], which curate drug−target interactions from the literature, will identify those mutations that might be inhibited or activated by the already approved drugs. The drug response can also be predicted *in silico* using resources with known pharmacogenetic information such as the Pharmacogenomics Knowledgebase (PharmGKB) [120]. However, the effectiveness of drug prediction methods depends on the completeness and accuracy of the existing knowledge resources. Some databases are populated with semi-manually curated data while others contain fully automated literature-mined interactions. In the latter cases, data integrity must be carefully monitored as auto-curation can lead to erroneous entries. The context in which the data in the database were generated should also be considered during the analysis. DrugBank, for example, stores all proteins known to bind to a specific drug, and thus contains results from biochemical binding assays that may not persist in more complex environments such as the human body; as such, protein−drug interactions identified in DrugBank may not always represent viable drug candidates for the treatment of cancer patients. In contrast, TTD focuses on collating target information for drugs currently in clinical trials but only contains primary targets annotated for most drugs. In short, the available bioinformatic databases and tools contain a wealth of information that can provide the foundation for data interpretation. However, the

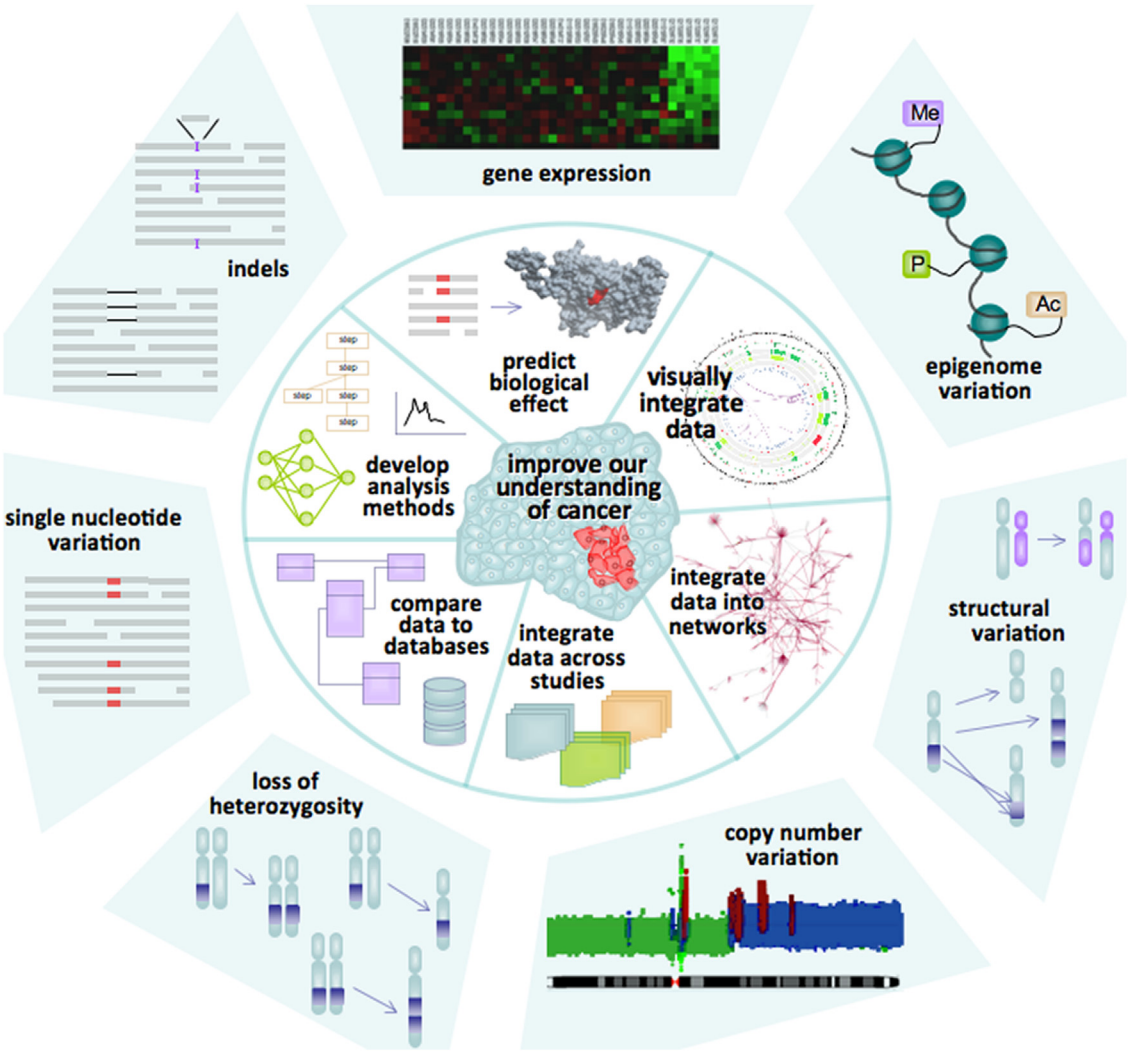

FIGURE 9.3 Identifying the perturbed pathways and networks driving the disease is integral to better understanding the drivers of cancer, exploring potential therapeutic targets and predicting drug response. Integration of data generated from both normal and cancer samples promises the unbiased examination of a cell's interconnected network of molecules.

accuracy and biases of these resources must be taken into consideration during analysis, especially when considering therapeutic intervention.

Data integration is not only applicable across all data types in one experiment but can also be applied to data across multiple platforms or analyses. Laboratories around the world are generating and analyzing data using various experimental designs, technologies and techniques. Bioinformatic methods have become key players for integration of these data. An example of such is the meta-analysis of microarray gene expression studies where the development of rigorous normalization methods can allow the vast amounts of gene expression data to be compared across platforms [121]. Data integration also requires the development of unified ontologies precisely to define and categorize biological concepts and data properties (http://www.geneontology.org/) [110].

Each data type independently can provide an overview of the alterations in a cancer genome. However, a more comprehensive view of the state of cell is obtained through their integration. Such a comprehensive analysis of diseased and normal cellular networks combined with computational modeling techniques would enable the identification of targets that can be inhibited in combination, or perhaps in a particular order, to maximize tumor cell death while sparing the healthy cells. For example, Iadevaia et al. modeled the IGF receptor signaling network and simulated the effects of inhibiting individual and combinations of targets. They predicted that optimal inhibition of MEK and mTOR would kill cancer cells but suboptimal inhibition would actually

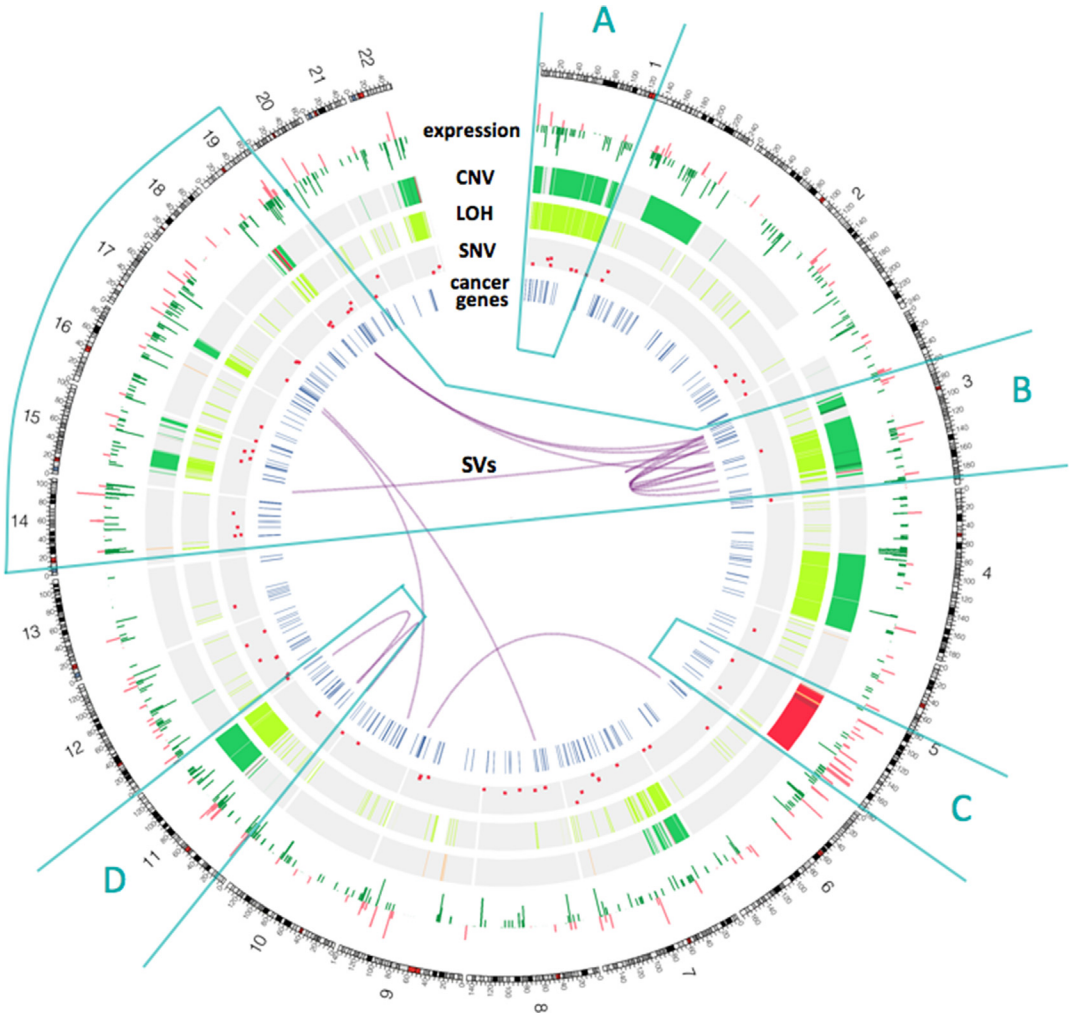

FIGURE 9.4 Data integration enables the interpretation of mutation calls and their biological significance. A hypothetical tumor-versus-normal dataset is plotted with Circos visualization software, showing examples of how integration can help interpret biological data. (A) An area with copy number loss (green CNV) and loss of heterozygosity (green LOH) agreeing with a one-copy deletion in the tumor genome. In this region, SNVs (red dots), which were heterozygous in the normal genome, may have become homozygous in the tumor. If only an inactivated allele of a tumor suppressor gene remains, for example, this SNV may contribute to oncogenesis. (B) A region showing varying copy number and a high density of structural variants may indicate extreme DNA rearrangement events such as chromothripsis. (C) Though copy number events do not necessarily correlate with expression results everywhere, regions where the two data types agree (red: CNV gain and high expression) may be of more importance and could provide a mechanistic basis for potential driver genes. (D) Integrating data can also help to verify if the data generated using different techniques agree with each other. In this region, for example, a deletion event (SV) correlates well with a loss of copy number (green CNV). In addition to the experimentally generated data, external datasets can also be included in such visualization plots. In this graph, one track for known cancer genes is added to the plot.

increase cancer cell viability. They subsequently validated this in cell proliferation assays [122].

CONCLUSION

Personalized cancer medicine, perhaps more appropriately termed "genomics medicine", aims to diagnose individuals based on their unique mutational profiles and provide therapies targeting those alterations while making predictions about the patient's response to treatment. Prior to the advent of NGS technologies, malignant tumors were diagnosed based on their anatomical site and histology characteristics; treatment options would then be based on patient's age, sex and/or family history. We now know that each cancer type is a collection of molecularly heterogeneous subtypes, and thus should not be treated identically. The power of NGS technologies in examining individual cancer genomes has given us the opportunity to identify different mutational profiles in patients with the same cancer type. For instance, *BRAF* activating mutations are found in 40–60% of melanoma patients while the rest carry wild-type copies of the gene [123]. Such knowledge allows for stratification of patients and targeted treatments; in this case, for instance, those

with the *BRAF* mutation benefit from the drug vemurafenib while the rest do not. This ensures administration of the right treatment to the right individuals and prevents unnecessary negative side effects in patients not likely to benefit from the drug. Genomic studies have also provided the opportunity to study rare cancers for which no treatment options exist. Tumors which are rare with respect to the tissue type and anatomical site they occur in could be quite common on the molecular level, harboring mutations in known cancer genes and pathways [116]. Such analyses allow for the delivery of already approved drugs to patients for whom no therapy options exist. With further advancements and the decreasing cost, high-throughput and comprehensive studies are starting to become tractable for clinical use [124]. Cancer genomic studies, especially large-scale initiatives such as TCGA, will provide us with a comprehensive overview of all somatic mutations in different cancer types and subtypes. These efforts will help in identifying and cataloging the key players in the disease. Maintaining databases for storing and linking genomic aberrations, clinicopathological characteristics, treatments and outcome of the disease for each patient will become inevitable in the near future. These databases need to be easily accessible by the scientific community and clinicians in order to advance research even further, thus one day enabling early diagnosis and targeted therapies in the clinic. Recently developed tools such as GeneInsight [125] provide the necessary communication gateways between research centers, medical laboratories and practicing clinicians, enabling the delivery of personalized cancer medicine.

Although bioinformatic tools that can more readily automate the data analysis and integration processes are currently in development, it will take much work in the future automatically to predict or annotate the somatic aberrations at a clinically usable level. To date, most studies have involved expert opinion where a panel of experts would discuss and choose the most actionable target(s) [124]. Exploiting the information from mutational analysis of cancers in developing therapeutic strategies also involves functional validation studies. Gene knockout methods, small interfering RNA (siRNA) experiments and selective overexpression of genes in model organisms, established cell lines and cell cultures can provide the opportunity for elucidating the functional role of mutations. However, to keep in pace with computational cancer genomics, these functional screening techniques need to be efficient and high-throughput.

GLOSSARY

Actionable target A mutated protein which is shown to be causally involved in disease initiation, progression and/or maintenance and which can be directly targeted by drugs.

Alignment Alignment is the process of mapping short sequence reads to the reference genome.

Assembly Assembly in genomics refers to the merging of short sequence reads into longer pieces of DNA contigs.

Bioinformatics The application of computer technology and programming to the field of biology.

Cancer genomics The field of study which examines cancers on the molecular level in the hope of identifying cancer-causing mutations and pathways.

Contig A contiguous sequence of nucleotides that is the result of assembling and joining overlapping reads; contigs provide the consensus sequence of the source DNA/RNA.

Copy number variation (CNV) Two copies of every gene are present in human cells, one inherited from each parent. Gain or loss of large regions of the genome leads to the gain or loss of gene copies. These are referred to as copy number variations.

Coverage Sequence coverage is the number of reads that span a specific position in the genome. The higher the coverage, the more confident the genotype calls.

Data integration Different data types (and mutation calls) from whole genome, exome, transcriptome or epigenome experiments are generated in cancer genomic studies. Combining the analysis results from all experiments rather than separate analysis of each can yield conclusive results on the state of the sample under study.

DNA sequencing Tools and techniques used for finding the order of nucleotide base pairs in a DNA molecule.

Driver mutations These are the somatic mutations in cancers which are responsible for providing the cell with growth advantage. As a result, they contribute to disease initiation, progression and maintenance.

Exome sequencing Sequencing only the protein-coding regions of the genome, the set of complete exons.

Expression analysis Refers to the analysis of the expression level of all or a subset of genes. Expression level can be quantified using the number of transcribed copies of each gene.

Functional validation Refers to proving the functional role of a driver mutation at the bench.

Genome The complete set of genetic material in a cell.

Germline mutations Mutations which are present in the gametes and are passed from one generation the next.

High-throughput sequencing Technologies which parallelize the sequencing of whole genomes by reading the sequence of millions of short DNA segments simultaneously.

Indels Indels refer to small insertions and deletions in the genome.

Next-generation sequencing technologies These are the novel technologies which have enabled high-throughput sequencing of complete genomes. Refer to high-throughput sequencing.

Paired-end sequencing The sequencing technique where the two ends of a DNA molecule of a known size are sequenced.

Paired-sample analysis Somatic mutation detection techniques which consider the tumor and matched normal samples simultaneously when identifying the most probable somatic aberrations.

Passenger mutations Any cancer cell could have hundreds of mutated genes, not all are contributing to the disease however. The majority are the result of the unstable cancer genome and do not provide the cell with any growth advantage. These are referred to as passenger mutations.

Pathway analysis Refers to identifying a pathway, a common cellular process governed by a complex of proteins and other molecules, which is affected by the accumulated mutations in the cell.

RNA-seq Also referred to as whole transcriptome shotgun sequencing (WTSS). It is the technique for finding the sequence of the transcribed regions of the genome. cDNA is made using the cell's RNA as template and is then sequenced.

Single nucleotide polymorphism (SNP) These are the variations that are present in the general healthy population. Any two individuals can have thousands, if not more, single nucleotides that vary in their genomes.

Single nucleotide variation (SNV) These are single nucleotide differences that are not present at high frequency in the population and might play a causal role in various diseases.

Somatic mutation Mutations which could be present in any cell type of the body except the germ cells (sperms and eggs). These mutations are not passed from one generation to the next.

Structural variation These refer to variations that encompass large areas of the genome and could include large insertions and deletions, inversions, duplications, translocations and gene fusions.

Transcriptome Complete set of transcribed molecules in a cell.

Verification Mutations identified through computational approaches need to be confirmed in the corresponding samples using techniques such as targeted capture and re-sequencing. This process is referred to as verification.

ABBREVIATIONS

AML Acute myeloid leukemia
BWA Burrows–Wheeler aligner
CCLE Cancer Cell Line Encyclopedia
CGCI Cancer Genome Characterization Initiative
CGH Comparative genomic hybridization
CGHub Cancer Genomic Hub
CGP Cancer Genome Project
CNV Copy number variation
COSMIC Catalogue Of Somatic Mutations In Cancer
DbSNP Database of Single Nucleotide Polymorphisms
ENCODE Encyclopedia of DNA Elements
FDA Food and Drug Administration
Gbp Giga base pair
GBM Glioblastomas
GO Gene ontology
GPU Graphics processing unit
HGP Human Genome Project
HPC High performance computing
ICGC International Cancer Genome Consortium
IGF Insulin-like growth factor
IGV Integrative Genomics Viewer
IPA Ingenuity Pathway Analysis
KEGG Kyoto Encyclopedia of Genes and Genomes
MAQ Mapping and assembly with quality
miRNA MicroRNA
NGS Next-generation sequencing
OMIM Online Mendelian Inheritance in Man
OregAnno Open regulatory annotation
PCGP Pediatric Cancer Genome Project
PCR Polymerase chain reaction
PDB Protein Data Bank
PharmGKB Pharmacogenomics Knowledgebase
PolyPhen Polymorphism phenotyping
SHRiMP Short read mapping package
SIFT Sorting Intolerant From Tolerant
siRNA Small interfering RNA
SNP Single nucleotide polymorphism
SNV Single nucleotide variation
SOAP Short Oligonucleotide Alignment Program
SV Structural variation
TARGET Therapeutically Applicable Research to Generate Effective Treatments
TCGA The Cancer Genome Atlas
TRANSFAC Transcription Factor Database
TTD Therapeutic Target Database
UCSC University of California Santa Cruz
WES Whole exome sequencing
WGS Whole genome sequencing
WGSS Whole genome shotgun sequencing
WTSS Whole transcriptome shotgun sequencing

REFERENCES

[1] A gene-centric human proteome project: HUPO – the Human Proteome organization. Mol Cell Proteomics 2010;9:427–9.
[2] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007;4:651–7.
[3] Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc 2006;1:729–48.
[4] Hirst M, Marra MA. Next generation sequencing based approaches to epigenomics. Brief Funct Genomics 2010;9:455–65.
[5] Wilson IM, Davies JJ, Weber M, Brown CJ, Alvarez CE, MacAulay C, et al. Epigenomics: mapping the methylome. Cell Cycle 2006;5:155–8.
[6] Jacinto FV, Ballestar E, Esteller M. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. BioTechniques 2008;44:35–43.
[7] Serre D, Lee BH, Ting AH. MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 2010;38:391–9.
[8] Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 2010;28:1097–105.
[9] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010;466(7303):253–7.
[10] Cokus SJ, Feng S, Zhang X, Chen Z, Merriman Z, Haudenschild CD, et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 2008;452:215–9.
[11] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 2008;133:523–36.

[12] Lister R, Pelizzola M, Dowen RH, Hawkins D, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315–22.
[13] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 2005;33:5868–77.
[14] Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 2009;27:353–60.
[15] Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 2009;27:361–8.
[16] Smith ZD, Gu H, Bock C, Gnirke A, Meissner A. High-throughput bisulfite sequencing in mammalian genomes. Methods 2009;48:226–32.
[17] International Cancer Genome Consortium. International network of cancer genome projects. Nature 2010;464:993–8.
[18] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
[19] International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.
[20] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73.
[21] Human Genome Structural Variation Working Group, Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al. Completing the map of human genetic variation. Nature 2007;447:161–5.
[22] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
[23] Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet 2008 [Chapter 10:Unit 10.11].
[24] Weisemann JM, Boguski MS, Ouellette BF. Sequence databases: integrated information retrieval and data submission. Curr Protoc Hum Genet 2001 [Chapter 6:Unit 6.7].
[25] Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, et al. BioMart Central Portal: an open database network for the biological community. Database (Oxford) 2011;2011:bar041.
[26] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010;11:647–57.
[27] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–7.
[28] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;18:1851–8.
[29] Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008;24:713–4.
[30] Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 2009;5:e1000386.
[31] Ferragina P, Manzini G. Opportunistic data structures with applications. Proceedings of the 41st Symposium on Foundation of Computer Science 2000;390–8.
[32] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
[33] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754–60.
[34] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;25:1966–7.
[35] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9.
[36] Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res 2008;18:324–30.
[37] Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 2008;18:810–20.
[38] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res 2009;19:1117–23.
[39] Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 2009;25:2283–5.
[40] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.
[41] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[42] Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 2010;26:730–6.
[43] Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res 2009;19:1124–32.
[44] Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 2012;28:907–13.
[45] Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012;28:311–7.
[46] Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics 2012;28:167–75.
[47] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865–71.
[48] Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS One 2009;4:e7767.
[49] Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res 2011;21:961–73.

[50] Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet 2010;19:R131–6.
[51] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
[52] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420–6.
[53] Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 2009;10:R23.
[54] Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 2009;19:1270–8.
[55] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009;6:677–81.
[56] McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 2009;19:1527–41.
[57] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453:56–64.
[58] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. *De novo* assembly and analysis of RNA-seq data. Nat Methods 2010;7:909–12.
[59] Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 2009;10:80.
[60] Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 2009;6:99–103.
[61] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9.
[62] Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 2009;19:1586–92.
[63] Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722–9.
[64] Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 2009;41:1061–7.
[65] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
[66] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 2009;37:D885–90.
[67] Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463:899–905.
[68] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[69] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
[70] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
[71] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Dai H, Hart AAM, Voskuil DW, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
[72] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.
[73] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907–12.
[74] Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 2003;13:1706–18.
[75] Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 2004;101:4164–9.
[76] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 2002;12:1611–8.
[77] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3.
[78] Holland RC, Down TA, Pocock M, Prlic A, Huen D, James K, et al. BioJava: an open-source framework for bioinformatics. Bioinformatics 2008;24:2096–7.
[79] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
[80] Stajich JE, Lapp H. Open source tools and toolkits for bioinformatics: significance, and where are we? Brief Bioinform 2006;7:287–96.
[81] Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, et al. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res 2012;40:D918–23.
[82] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res 2012;40:D84–90.
[83] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2012.
[84] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639–45.
[85] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.

[86] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–51.
[87] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
[88] Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007;7:233–45.
[89] Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics 2012;28:40–7.
[90] Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, et al. Integrative analysis of genome-wide loss of heterozygosity and mono-allelic expression at nucleotide resolution reveals disrupted pathways in triple negative breast cancer. Genome Res 2012;22:1995–2007.
[91] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153–8.
[92] Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet 2011;44:133–9.
[93] Downing JR, Wilson RK, Zhang J, Mardis ER, Pui CH, Ding L, et al. The Pediatric Cancer Genome Project. Nat Genet 2012;44:619–22.
[94] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[95] Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol 2007;8:R232.
[96] Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–74.
[97] Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics 2006;173:2187–98.
[98] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[99] Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 2009;37:D387–92.
[100] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–23.
[101] Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci USA 2007;104:20007–12.
[102] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306:636–40.
[103] Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform 2008;9:326–32.
[104] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 2010;38:D105–10.
[105] Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 2008;36:D107–13.
[106] Buratti E, Chivers M, Hwang G, Vorechovsky I. DBASS3 and DBASS5: databases of aberrant 3′- and 5′-splice sites. Nucleic Acids Res 2011;39:D86–91.
[107] Wang J, Zhang J, Li K, Zhao W, Cui Q. SpliceDisease database: linking RNA splicing and disease. Nucleic Acids Res 2012;40:D1055–9.
[108] Wang E. A roadmap of cancer systems biology. In: Wang E, editor. Cancer systems biology. Boca Raton, FL: CRC Press; 2010.
[109] Christopher R, Dhiman A, Fox J, Gendelman R, Haberitcher T, Kagle D, et al. Data-driven computer simulation of human cancer cell. Ann NY Acad Sci 2004;1020:132–53.
[110] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.
[111] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[112] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34:D354–7.
[113] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
[114] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 2010;26:i237–45.
[115] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.
[116] Jones SJ, Laskin J, Li YY, Griffith OL, An J, Bilenky M, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biol 2010;11:R82.
[117] Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov 2012;2:82–93.
[118] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. Nucleic Acids Res 2006;34:D668–72.
[119] Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. Nucleic Acids Res 2002;30:412–5.
[120] Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res 2002;30:163–5.
[121] Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med 2008;5:e184.

[122] Iadevaia S, Lu Y, Morales FC, Mills GB, Ram PT. Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. Cancer Res 2010;70:6704–14.

[123] Flaherty KT, Puzanov I, Kim KB, Ribas A, McArthur GA, Sosman JA, et al. Inhibition of mutated, activated BRAF in metastatic melanoma. N Engl J Med 2010;363:809–19.

[124] Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med 2011;3:111–21.

[125] Aronson SJ, Clark EH, Babb LJ, Baxter S, Farwell LM, Funke BH, et al. The GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. Hum Mutat 2011;32:532–6.