11

Analysis of Genomic Alterations in Cancer

Benjamin J. Raphael, Stas Volik, and Colin C. Collins

11.1 Introduction

Cancer is driven by a selection for mutations that include single nucleotide substitutions, short indels, and large-scale rearrangements of the genome [e.g., chromosomal inversions, translocations, segmental deletions, segmental duplications, and changes in chromosome copy number (aneuploidy and polyploidy)]. The frequency of these events varies greatly among tumors. For example, some tumors exhibit a large number of single nucleotide mutations but relatively normal chromosomal organization, while other tumors exhibit extensive chromosomal aberrations and rearrangement. In some types of cancer, these large-scale rearrangements produce changes in gene structure and regulation that are directly implicated in cancer progression and are targets for cancer therapeutics. A classic example is the Philadelphia chromosome [1], a 9;22 translocation observed in chronic myeloid leukemia. This translocation results in the ABL-BCR fusion protein [2] that is targeted by the drug Gleevec [3]. Another example is the chromosome in Burkitt's 8;14 translocation lymphoma. This translocation activates the *c-myc* gene by placing it under the control of a strong promoter of an immunoglobulin gene [4]. In contrast to these and other well-characterized translocations in leukemias and lymphomas, solid tumors frequently exhibit many chromosomal aberrations [5]. However, very few aberrations have been found to be recurrent across multiple patients, and thus it was believed that fusion genes like ABL-BCR were nonexistent in solid tumors.

However, more recent analysis has challenged this view [6] and two gene fusions were recently reported whose combined frequency exceeds 50% of tested prostate cancer patients [7]. These results suggest that additional fusion genes remain to be discovered or delinated.

During the past few decades, the vast majority of information about large-scale alterations in tumor genomes (e.g., as gain and loss of whole chromosomes, translocations, inversions, or large regions of duplication) has resulted from the application of cytogenetic and molecular cytogenetic techniques. These techniques are based on direct visualization of chromosomes and include chromosome banding, multiplex fluorescent in situ hybridization (mFISH) [8], and spectral karyotyping (SKY) [9]. Collectively these techniques have revealed numerous chromosomal aberrations, over 50,000 of which are recorded in the Mitelman database [10]. Despite this data, little is known about the detailed organization of tumor genomes or about the role of genome rearrangements in cancer progression. For example, the relative importance or prevalence of duplications in comparison to translocations and inversions in tumors is not known and the extent of variation in frequency of different events across different tumor types is unclear. The reason for this knowledge gap is that molecular cytogenetic techniques for identifying genome rearrangements have limited resolution (on the order of megabases) because they rely on isolation and analysis of metaphase chromosomes. This means that changes on a smaller scale will not be observed. Moreover, cytogenic techniques are relatively low-throughput, meaning that detailed studies of many samples are difficult. Also, most of these techniques require the isolation of metaphase chromosomes, which is challenging for some tissues.

In the era of genome sequencing, it is apparent that resequencing tumor genomes would provide the ultimate dataset for cancer mutation and rearrangement studies. However, it is presently unrealistic to sequence more than a few tumor genomes in view of the high cost of mammalian genome sequencing. Moreover, in contrast to the sequencing of the human genome, tumor genomes present unique computational and experimental challenges. First, assembly of a tumor genome by whole-genome shotgun sequencing is challenging because extensive segmental duplications present in many tumors presents a formidable fragment assembly problem. Second, solid tumors are a heterogeneous collection of cells with varying number and type of mutations. Thus, if one shotgun sequences DNA extracted from a tumor sample, one is not sequencing a single genome, but a population of different (albeit related) genomes. Despite these obstacles, the availability of a high-quality reference human genome sequence coupled with rapid advances in sequencing technology affords the opportunity for high-resolution sequence-based analysis of tumor genomes. One approach that mitigates the assembly problem is to restrict attention to protein coding regions and sequence only these regions to discover somatic point mutations

important in cancer. Notable efforts in this direction include [11–14] and the recently initiated Cancer Genome Atlas [15].

Here we focus on larger-scale rearrangements, duplications, and deletions. There are currently two sequence-based techniques used to examine these genomic alterations in tumors: comparative genomic hybridization to arrayed representations of the human genome and paired-end sequencing. CGH is restricted to the detection of changes in copy number in the tumor paired-end sequencing detects rearrangements while changes in copy number and point mutations. We note that both of these techniques have also been applied to assess inherited structural polymorphisms in the human genome [16, 17].

11.1.1 Measurement of Copy Number Changes by Array Hybridization

Array comparative genome hybridization (aCGH) [18] has become a dominant tool for the analysis of copy number changes in cancer. This technique involves the hybridization of differentially fluorescently labeled normal human and tumor DNA fragments to a set of genomic probes derived from normal human DNA. Measurements of tumor to normal fluorescence ratios at each probe identify locations in the human tumor genome (Figure 11.1) that are present in higher or lower copy than in the tumor genome. Comparative genomic hybridization was first developed as a cytogenetic technique for hybridizations to metaphase chromosomes [19], but the use of arrays has steadily improved the resolution of aCGH: earlier spotted clone arrays [20] have resolution at most 0.5–1 MB, but more recent arrays based on overlapping clones [21], PCR products [22], or oligonucleotides [23, 24] offer resolutions approaching 50 kb or less.

A major challenge in the interpretation of aCGH data is noise in the hybridization, and a variety of statistical techniques have been developed for this analysis. These techniques rely on the principle that if a duplicated or deleted region is large relative to the genomic spacing between probes on the array, then multiple adjacent probes will record the duplication or deletion. Thus measurements at probes from adjacent locations on the human genome typically will be correlated. Statistical methods exploit these correlations to transform the set of noisy probe measurements into contiguous segments of the genome that have normal or altered copy number in the tumor. Methods include change-point models [25], hidden Markov models [26], clustering [27], and a variety of other techniques, several of which are compared in [28].

Array-CGH has become a widespread tool in genomic analysis of cancer and has been used to: (1) identify candidates oncogenes and tumor-suppressor genes; (2) assay tumors for specific well-characterized aberrations such as amplification of *ERBB2*; and (3) correlate copy number profiles with prognosis, recurrence, or response to treatment. Pinkel and Albertson [29] review these applications of aCGH in cancer.

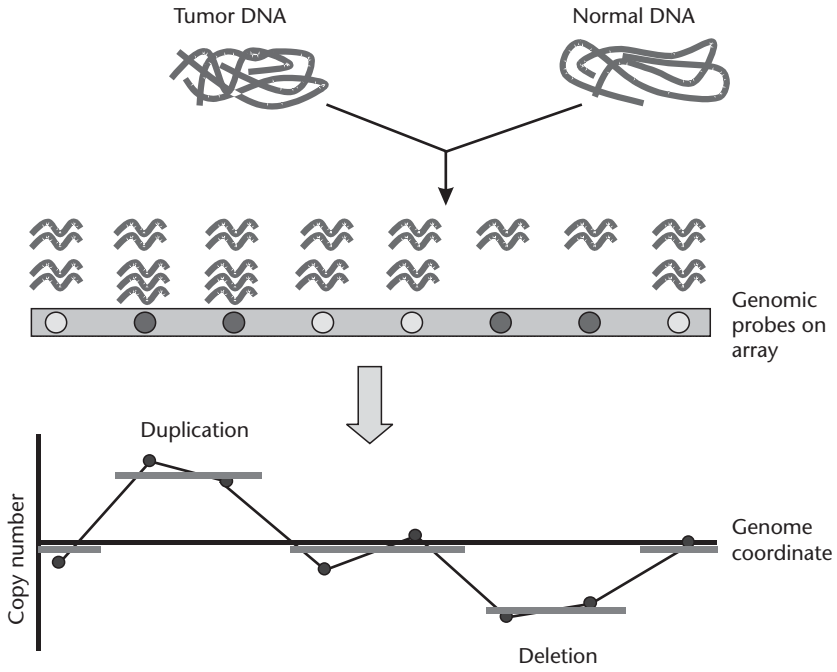

Figure 11.1 In array-CGH (aCGH), normal and tumor DNA (ideally from the same patient) are differentially labeled and hybridized to an array of genomic probes spaced across the reference human genome. The relative intensity measured at a probe indicates the copy number of the region in the tumor genome. Statistical methods are employed to filter the resulting copy number profile into duplicated segments, deleted segments, or segments with no change in copy number.

Array-CGH has some limitations as a tool for tumor genome analysis. First, aCGH does not detect rearrangements that have no effect on copy number including inversions and reciprocal (balanced) translocations. Second, although aCGH will reveal regions of the genome that are duplicated in a tumor, aCGH gives little information about the organization and locations of duplicated material within the tumor genome. For example, an aCGH experiment will not reveal whether two duplicated regions are located close together (or even on the same chromosome) in the tumor genome. In some tumor genomes, duplicated material from several disparate regions of the human genome is colocalized [30–32], and this knowledge is potentially useful for understanding altered regulation of genes in tumors. Finally, aCGH is impeded by genomic heterogeneity in the sample, arising either from contamination of tumor samples with normal (unmutated) cells or from heterogeneity in the alterations found within different cells of the tumor.

11.1.2 Measurement of Genome Rearrangements by End Sequence Profiling

Sequencing of tumor genomes overcomes some of the limitations of CGH, but as mentioned earlier, high-coverage shotgun sequencing of a large number of tumors is not yet practical. An approach called end-sequence profiling (ESP) [31] has proven to be effective for genome-wide analysis of rearrangements in tumor cells. ESP provides a balance between imprecise, but inexpensive cytogenetic technologies and very precise, but expensive, full-genome sequencing. ESP involves the sequencing of paired ends of tumor genome fragments and the mapping of these ends to the reference human genome sequence [Figure 11.2(a)]. ESP is able to reveal all types of rearrangements present in a tumor including inversions, translocations, transpositions, duplications, and deletions. ESP gives at least an order of magnitude more accurate representation of the tumor genome than cytogenetic techniques like SKY, and in addition yields detailed information about the organization of the tumor genome that is lacking in CGH. Moreover, ESP is less impeded by heterogeneity in the sample than CGH, since an end-sequenced fragment arises from a distinct piece of DNA from an individual tumor cell. Therefore, it is possible to overcome problems with heterogeneous samples or contamination by normal cell admixture by sequencing additional clones. Ultimately, the resolution of ESP is limited only by the number of clones sequenced and the size of each clone.

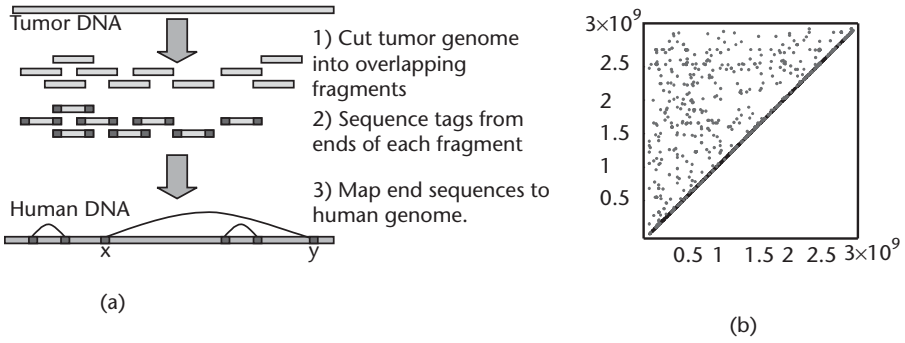

Figure 11.2 (a) In the end sequence profiling (ESP) technique, short tag sequences from the ends of fragments of the tumor genome are mapped to the human genome. Each mapped fragment is associated with a pair (x, y) of locations in the human genome. (b) The data from an ESP experiment consists of a set of ES pairs (x_1, y_1), ..., (x_n, y_n) represented as points in a two-dimensional plot. Typically, the distance between elements of an ES pair will approximately equal the length of a fragment (points near diagonal). However, since the tumor genome is a rearranged version of the human genome, there will also be a number of *invalid* ES pairs whose ends map far apart (points off diagonal). The goal is to reconstruct the organization of the tumor genome from the ES pairs and to find a plausible sequence of rearrangements that transform the human genome into the tumor genome.

ESP was first applied to a comprehensive study of the MCF7 breast cancer cell line [31, 32] and later to additional cell lines and primary turmors [33]. The following methodology was used. First, a bacterial artificial chromosome (BAC) library was constructed from the MCF7 cell line. That is, DNA from the MCF7 cell line was split into small fragments varying in size from 80–250 kb, and these pieces of DNA were cloned into BACs.[1] Second, the ends (≈500 bp) of each BAC were sequenced. Third, the resulting end sequences were mapped to the reference human genome. Only BACs with both end sequences mapping *uniquely* to the human genome were retained for further analysis. Each such BAC corresponds to a pair (x, y) of locations in the human genome where the end sequences map. In addition, since the end sequence may map to either DNA strand, each mapped end has a sign (+ or –) to indicate the strand. We call such a signed pair an *end sequence pair* (*ES pair*). Thus, the data from an ESP experiment consists of a set of ES pairs (x_1, y_1), ..., (x_N, y_N) [Figure 11.2(b)].

Typically, the distance between elements of a ES pair will equal the length L of a BAC clone (e.g., 80–250 kb), and the ends will have opposite, convergent orientations [i.e., an ES pair of the form $(+x, -(x+L))$]. We call such ES pairs *valid* pairs. However, since the tumor genome is a rearranged version of the human genome, there will also be a number of *invalid* pairs whose ends map far apart, or have the wrong orientation, or both. The valid and invalid pairs reveal information about the organization of the tumor genome. In particular, invalid pairs indicate distant regions of the human genome that are fused in the tumor, possibly revealing novel fusion genes [32]. However, in highly rearranged tumor genomes like MCF7, the complicated patterns of invalid pairs defy simple explanation and require the development of computational methods for analysis.

11.2 Analysis of ESP Data

The first step in the analysis of ESP data is to map the end sequences to the reference human genome sequence For end sequences of 500 bp, this step is easily accomplished using tools like MegaBlast [34] or BLAT [35]. The main challenge results from repeats and duplications in the human genome, which lead to nonunique mappings for end sequences. Clones with nonunique mappings can be removed from consideration, as the genomic region that they contain will likely be covered by other clones, assuming that the clone library is sufficiently large. Completion of end sequence mapping gives a set of ES pairs. The second

1. Note that the ESP methodology is flexible in terms of the cloning vector that is used. Fosmids, with insert size of ≈40 kb or plasmids with insert sizes of ≈2 kb can give increased resolution of the tumor genome, at the cost of a larger number of clones that is required with BACS.

step is to cluster the ES pairs to overcome experimental errors and identify clones spanning the same rearrangement. The primary source of experimental errors is chimeric BACs in the library. Chimeric BACs are produced by joining of two noncontiguous regions of DNA, and thus chimeric BACs will also correspond to invalid ES pairs. However, chimeric BACs are artifacts, rather than signs of real rearrangements. When an ESP project includes a sufficiently large number of BAC clones, chimeric BACs are easily distinguished from real rearrangements because the breakpoints of a rearrangement will likely be covered by two or more BACs. In contrast, because chimeric BACs typically combine two "random" segments of DNA, different chimeric BACs rarely will have ends mapped in close proximity. Thus, we define an *ES cluster* as a set of ES pairs whose entries are close enough that all ES pairs in a set could be explained by a single rearrangement event. That is, we say that ES pairs (x_1, y_1), ..., (x_n, y_n) form an ES cluster if there exist locations a and b such that:

$$l \le sign(x_i)(a - x_i) + sign(y_i)(b - y_i) \le L \text{ for } i = 1, \ldots, n$$

where l and L are the minimum and maximum clone sizes, respectively.

Then ES pairs arising from chimeric BACs are extremely unlikely to be members of ES clusters.

Having obtained a set of ES clusters, we can identify putative rearrangements in the tumor including inversions, translocations, and duplications (Figure 11.3). However, some ES clusters do not result from a single rearrangement of the human genome, but from multiple overlapping rearrangements [Figure 11.3(d)]. To analyze these overlapping rearrangements, we apply methods from comparative genomics to derive a putative tumor genome sequence and analyze genome rearrangements that transform the normal human genome into the tumor genome. In [36], in an early attempt to reconstruct tumor genomes from ESP data, we developed a computational approach with a few simplifying assumptions: (1) a sequence of inversions, translocations, chromosomal fissions, and chromosome fusions generates the tumor genome from the normal human genome; (2) no duplications occurred in the tumor genome; and (3) each BAC clone contains at most one rearrangement breakpoint. These assumptions allowed us to use a theoretical framework originally developed to study genome rearrangements that occur during species evolution. In this framework, both the human and tumor genomes are represented by integer permutations and the problem is to find a minimal sequence of rearrangement operations that transform one permutation into another. Under the assumption that only inversions, translocations, fission, and fusions occur, such a minimal sequence can be computed efficiently, specifically in time that is polynomial in the length of the permutation [37]. We applied this computational approach to ESP data from the

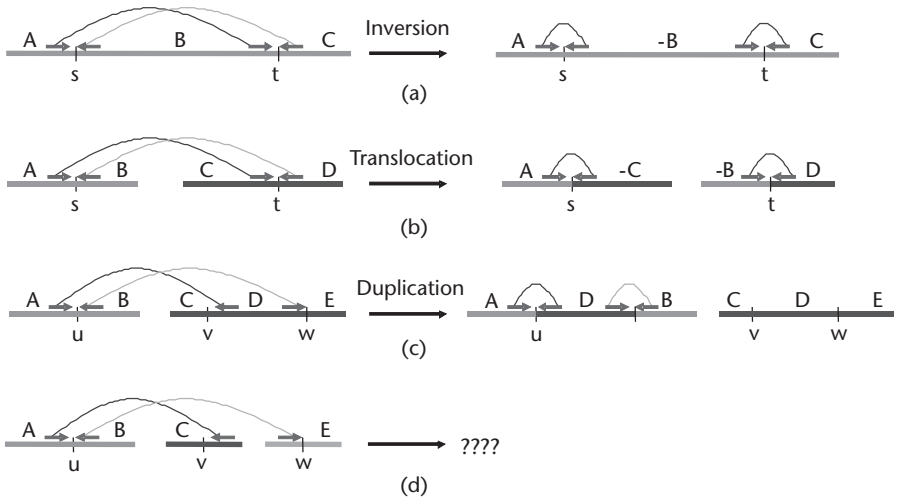

Figure 11.3 Locations and orientation of end sequence (ES) pairs suggest rearrangement events in the tumor genome including: (a) inversions on a single chromosome, (b) translocations between two chromosomes, (c) duplications or transpositions, or (d) compound events suggesting multiple rearrangements. Here, arrows indicate the locations and orientation of mapped end sequences, and arcs join end sequences (ES) that form an ES pair. Each of these events transforms the indicated invalid ES pair by rearranging the labeled segments of the genome.

MCF7 breast cancer cell line, and derived a putative reconstruction of the MCF7 genome (Figure 6 in [36]). This study produced the first high-resolution reconstruction of a tumor genome and directed further BAC sequencing experiments.

Of course, duplication and deletions are quite common in tumor genomes as shown by numerous CGH studies. The availability of ESP data allows us to study the organization of duplicated regions in tumors. In the MCF7 ESP data, we observed a complex pattern of ES pairs that suggested a process of overlapping rearrangements and duplications (see Figure 3b in [38]). We developed a computational technique to analyze duplications in this data using a model based on the biological process of duplication by amplisome[2] [38]. Amplisomes are essentially minimal units of duplication that replicate extrachomosomally and can reintegrate into chromosomes [39]. Our method reconstructs an amplisome sequence that explains the ESP data by finding the shortest path in a graph derived from the ES pairs. Using our method, we reconstructed a putative amplisome for the MCF7 genome (see Figure 5 in [38]).

2. Also referred to as episomes or double minutes, depending on their size.

While amplisomes have been observed in vitro and in vivo, they are only one mechanism by which tumor genomes evolve their new organization. A process called the breakage/fusion/bridge cycle also yields duplications at the ends of damaged chromosomes [40], and there is evidence that this process may active in human solid tumors [32, 33, 41]. The precise mechanisms that produce duplications in human tumors are not completely understood, and it is not known whether a tumor preferentially uses one mechanism of duplication or combines multiple mechanisms. ESP analysis of tumor genomes can help resolve these mysteries particularly in concert with the development of computational models of additional rearrangement and duplication mechanisms.

11.3 Combination of Techniques

It is likely that no single sequenced-based technique is optimal for analyzing all types of alterations in every tumor genome. However, different methods of analyzing tumor genomes should ideally produce concordant results. We recently compared ESP and array aCGH data for MCF7 and discovered that there was significant overlap between the genomic locations where aCGH identifies a change in copy number, and locations where ESP identifies rearrangements [33]. Of course, agreement between the two techniques is not expected to be perfect, both because of experimental noise and because aCGH cannot measure certain types of rearrangements. We note that even in the case of structural polymorphisms, there are discrepancies between different techniques [42]. Robust statistical methods to integrate measurements from different experimental techniques are needed.

11.4 Future Directions

An important consideration in the analysis of tumor genomes is the genomic heterogeneity found in a tumor. In principle, sequencing approaches like ESP have an advantage over aCGH in this regard. In the creation of a BAC (or other clone) library, DNA from multiple cells is pooled, but an individual clone does measure a rearrangement in a single cell, in contrast to aCGH, which averages over the cell population. ES pairs from rearrangement variants in different cells will be mixed together in the ESP data, but rearrangements that are common to all or most cells in a tumor population are more likely to be cloned and sequenced than sporadic rearrangements. Thus ESP is biased towards finding these common and potentially early rearrangement events in the development of the tumor. At the same time, deep sequencing via ESP can reveal rare events present in a subpopulation of tumor cells. A key question is to determine the

number of clones necessary to analyze populations of tumor cells with different amounts of heterogeneity; mathematical analyses, simulation studies, and pilot sequencing projects are needed to address this question. There is a parallel between sequencing a tumor and "community sequencing" or metagenomics approaches that simultaneously sequence an environmental sample containing a mixture of organisms [43, 44]. Community sequencing identifies rare organisms in a hetrogeneous mixture with deep sequencing just as ESP identified rare rearrangements in tumors with more sequenced clones. There is room for cross-fertilization of ideas between these two approches.

DNA sequencing technology continues to reduce in cost and improve in efficiency. Steady improvements in current technologies will make large-scale ESP studies more common. Moreover, other ESP-like strategies have been proposed including a paired-end sequencing technique that improve efficiency by concatenating multiple short paired-end tags into a single read in SAGE-like approach, yielding an order of magnitude more paired ends for the same number of sequenced reads [45]. Presently, the 18-bp end sequences produced by this technique are too short for tumor-genome rearrangement studies because too few of these short tags can be uniquely identified in the human genome. However, with slightly longer end sequences (e.g., 22–25 bp), enough will map uniquely to the human reference sequence to undertake effective ESP studies. The greatest promise in the near term lies in the application of the new generation of short-read sequencers that will be able to produce a significantly larger number of short-end sequence pairs in a cost-effective manner [46].

Such large-scale sequencing efforts and the future development of single-cell sequencing techniques will give an unprecedented catalog of tumor mutations, including both point mutations and large-scale alterations. Eventually, complete mutational analysis of tumors will become feasible. There will be a great demand for bioinformatic techniques to uncover sets of recurrent mutations and define similarity between highly mutated tumor samples. Similarity might mean more than possessing specific mutations or mutated genes in common; for example, different tumors might share similar mutated pathways. Ultimately, the knowledge gained from tumor genome studies will be used not only to discover gene targets for diagnostics and therapeutics, but also to better understand the temporal and population dynamics of the mutational process of tumor development.

References

[1] Rowley, J. D., "Letter: A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia Identified by Quinacrine Fluorescence and Giemsa Staining," *Nature,* Vol. 243, 1973, pp. 290–293.

[2] Heisterkamp, N., et al., "Localization of the c-ab1 Oncogene Adjacent to a Translocation Break Point in Chronic Myelocytic Leukaemia," *Nature*, Vol. 306, 1983, pp. 239–242.

[3] Druker, B. J., et al., "Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia," *N. Engl. J. Med.*, Vol. 344, 2001, pp. 1031–1037.

[4] Croce, C. M., et al., "Molecular Genetics of Human B- and T-Cell Neoplasia," *Cold Spring Harb. Symp. Quant. Biol.*, Vol. 51, Pt. 2, 1986, pp. 891–898.

[5] Albertson, D. G., et al., "Chromosome Aberrations in Solid Tumors," *Nat. Genet.*, Vol. 34, 2003, pp. 369–376.

[6] Mitelman, F., B. Johansson, and F. Mertens, "Fusion Genes and Rearranged Genes as a Linear Function of Chromosome Aberrations in Cancer," *Nat. Genet.*, Vol. 36, 2004, pp. 331–334.

[7] Tomlins, S. A., et al., "Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer," *Science,* Vol. 310, 2005, pp. 644–648.

[8] Speicher, M. R., S. Gwyn Ballard, and D. C. Ward, "Karyotyping Human Chromosomes by Combinatorial Multi-Fluor FISH," *Nat. Genet.*, Vol. 12, 1996, pp. 368–375.

[9] Schrock, E., et al., "Multicolor Spectral Karyotyping of Human Chromosomes," *Science,* Vol. 273, 1996, pp. 494–497.

[10] Mitelman, F., B. Johansson, and F. Mertens, *Mitelman Database of Chromosome Aberrations in Cancer*, 2006, http://cgap.nci.nih.gov/chromosomes/mitelman.

[11] Ley, T. J., et al., "A Pilot Study of High-Throughput, Sequence-Based Mutational Profiling of Primary Human Acute Myeloid Leukemia Cell Genomes," *Proc. Natl. Acad. Sci. USA,* Vol. 100, 2003, pp. 14275–14280.

[12] Stephens, P., et al., "A Screen of the Complete Protein Kinase Gene Family Identifies Diverse Patterns of Somatic Mutations in Human Breast Cancer," *Nat. Genet.,* Vol. 37, 2005, pp. 590–592.

[13] Sjoblom, T., et al., "The Consensus Coding Sequences of Human Breast and Colorectal Cancers," *Science*, Vol. 314, 2006, pp. 268–274.

[14] Greenman, C. P., "Patterns of Somatic Mutation in Human Cancer Genomes," *Nature,* Vol. 446, No. 7132, March 8, 2007, pp. 153–158.

[15] Kaiser, J., National Institutes of Health, "NCI Gears Up for Cancer Genome Project," *Science,* Vol. 307, 2005, pp. 11–82.

[16] Feuk, L., A. R. Carson, and S. W. Scherer, "Structural Variation in the Human Genome," *Nat. Rev. Genet.,* Vol. 7, 2006, pp. 85–97.

[17] Tuzun, E., et al., "Fine-Scale Structural Variation of the Human Genome," *Nat. Genet.,* Vol. 37, 2005, pp. 727–732.

[18] Pinkel, D., et al., "High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays," *Nat. Genet.,* Vol. 20, 1998, pp. 207–211.

[19] Kallioniemi, A., et al., "Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors," *Science,* Vol. 258, 1992, pp. 818–821.

[20] Pollack, J. R., et al., "Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays," *Nat. Genet.,* Vol. 23, 1999, pp. 41–46.

[21] Ishkonian, A. S., et al., "A Tiling Resolution DNA Microarray with Complete Coverage of the Human Genome," *Nat. Genet.,* Vol. 36, 2004, pp. 299–303.

[22] Dhami, P., et al., "Exon Array CGH: Detection of Copy-Number Changes at the Resolution of Individual Exons in the Human Genome," *Am. J. Hum. Genet.,* Vol. 76, 2005, pp. 750–762.

[23] Lucito, R., et al., "Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation," *Genome Res.,* Vol. 13, 2003, pp. 2291–2305.

[24] Barrett, M. T., et al., "Comparative Genomic Hybridization Using Oligonucleotide Microarrays and Total Genomic DNA," *Proc. Natl. Acad. Sci. USA,* Vol. 101, 2004, pp. 17765–17770.

[25] Olshen, A. B., et al., "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data," *Biostatistics,* Vol. 5, October 2004, pp. 557–572.

[26] Fridlyand, J., et al., "Hidden Markov Models Approach to the Analysis of Array CGH Data," *Journal of Multivariate Analysis,* Vol. 90, 2004, pp. 132–153.

[27] Wang, P., et al., "A Method for Calling Gains and Losses in Array CGH Data," *Biostatistics,* Vol. 6, 2005, pp. 45–58.

[28] Lai, W. R., et al., "Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data," *Bioinformatics,* Vol. 21, 2005, pp. 3763–3770.

[29] Pinkel, D., and D. G. Albertson, "Array Comparative Genomic Hybridization and Its Applications in Cancer," *Nat. Genet.,* Vol. 37, Suppl., 2005, pp. S11–S17.

[30] Guan, X. Y., et al., "Identification of Cryptic Sites of DNA Sequence Amplification in Human Breast Cancer by Chromosome Microdissection," *Nat. Genet.,* Vol. 8, 1994, pp. 155–161.

[31] Volik, S., et al., "End-Sequence Profiling: Sequence-Based Analysis of Aberrant Genomes," *Proc. Natl. Acad. Sci. USA,* Vol. 100, 2003, pp. 7696–7701.

[32] Volik, S., et al., "Decoding the Fine-Scale Structure of a Breast Cancer Genome and Transcriptome," *Genome Res.,* Vol. 16, 2006, pp. 394–404.

[33] Raphael, B. J., et al., " A Sequence Based Survey of the Complex Structural Organization of Tumor Genomes," unpublished document.

[34] Zhang, Z., et al., "A Greedy Algorithm for Aligning DNA Sequences," *J. Comput. Biol.,* Vol. 7, 2000, pp. 203–214.

[35] Kent, W. J., "BLAT—The BLAST-Like Alignment Tool," *Genome Res.,* Vol. 12, 2002, pp. 656–664.

[36] Raphael, B. J., et al., "Reconstructing Tumor Genome Architectures," *Bioinformatics,* Vol. 19, Suppl. 2, 2003, pp. II162–II171.

[37] Pevzner, P., *Computational Molecular Biology: An Algorithmic Approach*, Cambridge, MA: MIT Press, 2000.

[38] Raphael, B. J., and P. A. Pevzner, "Reconstructing Tumor Amplisomes," *Bioinformatics,* Vol. 20, Suppl. 1, 2004, pp. I265–I273.

[39] Windle, B. E., and G. M. Wahl, "Molecular Dissection of Mammalian Gene Amplification: New Mechanistic Insights Revealed by Analyses of Very Early Events," *Mutat. Res.,* Vol. 276, 1992, pp. 199–224.

[40] McClintock, B., "The Stability of Broken Ends of Chromosomes in *Zea Mays*," *Genetics,* Vol. 26, 1941, pp. 234–282.

[41] Chin, K., et al., "In Situ Analysis of Genome Instability in Breast Cancer," *Nat. Genet.,* Vol. 16, 2004, pp. 984–988.

[42] Eichler, E. E., "Widening the Spectrum of Human Genetic Variation," *Nat. Genet.*, Vol. 38, 2006, pp. 9–11.

[43] Venter, J. C., et al., "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science,* Vol. 304, 2004, pp. 66–74.

[44] Tyson, G. W., et al., "Community Structure and Metabolism Through Reconstruction of Microbial Genomes from the Environment," *Nature,* Vol. 428, 2004, pp. 37–43.

[45] Ng, P., et al., "Gene Identification Signature (GIS) Analysis for Transcriptome Characterization and Genome Annotation," *Nat. Methods,* Vol. 2, 2005, pp. 105–111.

[46] Bently, D. R., "Whole Genome Re-Sequencing," *Curr. Opin. Genet. Dev.,* Vol. 16, 2006, pp. 545–552.