Part 1.1

Analytical techniques: analysis of DNA

Chapter 1

Cancer genome sequencing

Abizar Lakdawalla, Jeffrey Fisher, Mostafa Ronaghi, and Jian-Bing Fan

Cancer is a disease of the genome. Each cancer is the result of a unique combination of germline and somatic mutations that could include nucleotide substitutions, insertions and deletions, chromosomal rearrangements and copy number changes in protein-coding or regulatory components of genes (1). In addition, cancer genomes may exhibit changes in DNA methylation and chromatin structure that can have a substantial impact on gene expression. A comprehensive catalog of all types of variants in a cancer opens new and unrivaled opportunities for understanding the mechanism of cancer onset and progression, predicting the response to therapeutics, and providing new biomarkers for diagnosis and prognosis.

The impact of next-generation DNA sequencing technology

The last 10 years have transformed our understanding of the cancer genome through a remarkable revolution in genome sequencing technology (2,3). In 2001, the Human Genome Project delivered the first draft of the human genome at a cost of $3B (4). By 2008, the cost had dropped to $2M (5). In 2012, a human genome can be sequenced for less than $5000. The dramatic increase in throughput and the drop in the cost of sequencing offers an unprecedented opportunity to comprehensively understand a cancer by determining all genomic variations, evaluating the impact of these DNA variations on the transcriptome (by RNA sequencing) and delineating the changes in the interactions of the genome with histones, enzymes, and transcription factors by sequencing DNA fragments associated with specific proteins (6).

With the increasing ease of sequencing a cancer genome, two important questions can now be posed: (i) what are the primary mutational events that trigger carcinogenesis?; and (ii) what are the cascades of genomic changes that lead to metastasis and to resistance to therapeutics? Studies based on second-generation sequencing technology have already revealed many unexpected etiologies for a number of cancers. For example, a significant proportion of glioblastoma multiforme seem to be triggered by mutations in isocitrate dehydrogenase 1 (7). Changes in the chromatin remodeling gene, ARID1A, have been associated with ovarian cancer (8); these genes could constitute potential therapeutic targets. Sequencing of DNA from a "basal-like" breast cancer sample, and the corresponding blood, brain metastasis, and xenograft samples all derived from the same patient indicated that secondary tumors may arise from a minority of cells within the primary tumor (9). A study based on sequencing from multiple tumors showed a remarkable phenomenon: a large number of genomic rearrangements occurring in a single cellular cataclysmic event with chromosomes crossing over at multiple points. The process, chromothripsis, was observed in ~25% of bone cancers and in ~2% of all cancers that were queried (10).

Second-generation DNA sequencing process

In 2005, Genbank contained about 50 gigabases (Gb) of data from all sequencing projects from all organisms. In 2012, the HiSeq 2500 (Illumina, Inc.), a second-generation sequencer, produces about 600 Gb of data in one run, with a daily throughput that exceeds the output of about 60 000 of the first-generation sequencer's output. This remarkable improvement in data throughput has been achieved by converting a labor-intensive serial process to a simple, highly parallel process.

In first-generation sequencing technologies, genomic DNA is fragmented and individual fragments are cloned into plasmids or phage to create a library with millions of individual clones. These plasmids are introduced into bacterial cells. Millions of isolated bacterial colonies are picked by robots, inoculated into media for growth, and processed for plasmid DNA isolations. Millions of individual sequencing reactions are performed on the isolated plasmid DNA, followed by capillary electrophoresis to generate sequence data for each plasmid. The process requires many technicians to work for several months in large laboratories to sequence a single human genome at a cost of approximately $20 million.

In second-generation sequencing, the serial process of growing and sequencing millions of individual clones is replaced by a highly parallel process where billions of DNA fragments are amplified and sequenced simultaneously, resulting in significant cost and time savings. The overall process consists of four steps:

Molecular Oncology, ed. Edward P. Gelmann, Charles L. Sawyers, and Frank J. Rauscher. Published by Cambridge University Press.

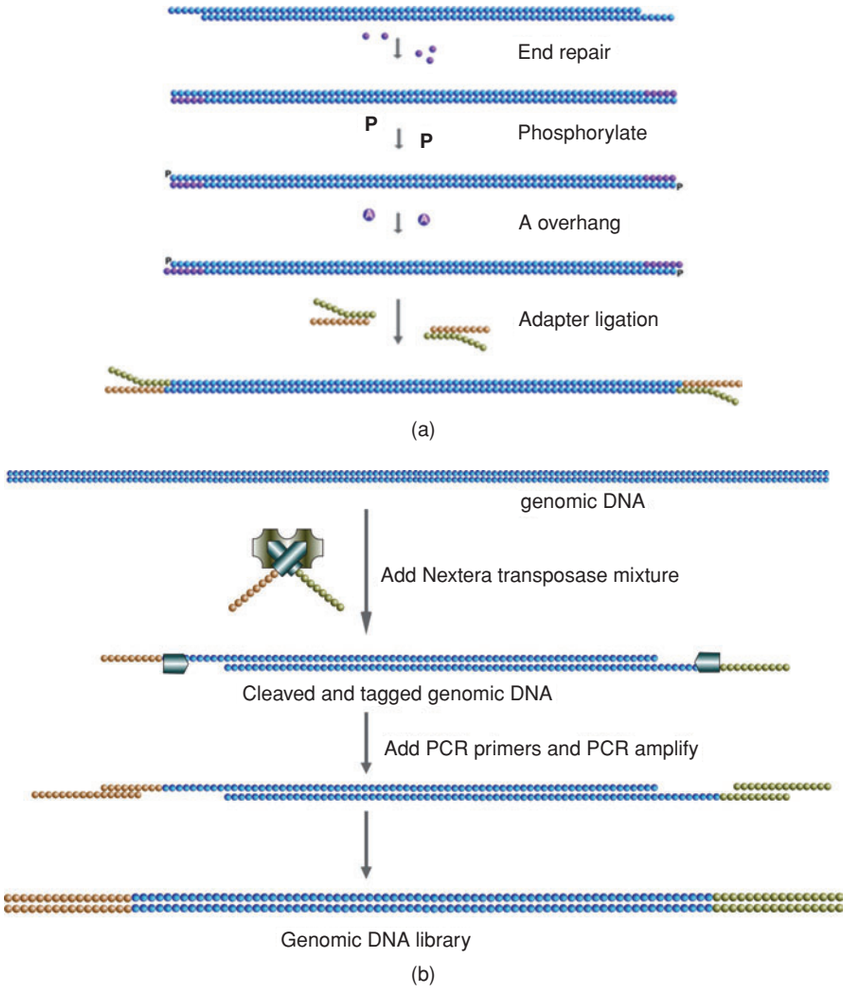

Figure 1.1 Preparation of a genomic DNA library (two methods). Second-generation sequencing libraries are made of short (500 bp) fragments with oligonucleotide adapters at both ends, using one of two methods: (a) gDNA is fragmented (mechanically or chemically), the ends are repaired and phosphorylated, and adapters ligated. (b) gDNA is incubated with the Nextera transposase mixture (containing enzyme, transposon end adapter oligonucleotide, buffers). Transposase simultaneously fragments the gDNA and inserts the adapter, covalently attaching it to the 5′ end of the DNA fragment. 3′ adapter sequences are added by limited-cycle PCR to produce a genomic library.

1. Creating a DNA library from genomic DNA;
2. Clonally amplifying the library in parallel in a single flow cell;
3. Sequencing all the clones simultaneously directly in the flow cell;
4. Analyzing the sequence.

These steps are described in more detail below.

1. Library preparation: isolated genomic DNA is fragmented into 500 bp segments by sonication. Blunt ends are created by polymerases and exonucleases, followed by phosphorylation and the addition of a single A base. A Y-shaped adapter is ligated to the ends of the DNA fragments to create the human genomic DNA library (Figure 1.1A). Alternatively, genomic DNA can be incubated with a transposase that carries DNA adapter sequences. The transposase simultaneously cleaves the DNA and ligates the adapter sequence to produce a fragmented library. A limited PCR reaction is used to synthesize the complementary adapter sequences at both ends of the DNA fragment, producing a library of genomic DNA fragments with adapter sequences at both ends (Figure 1.1B).
2. Clonal amplification: the genomic library with more than a billion individual fragments created in step 1 is clonally amplified in a hollow flow cell to produce DNA clones. The flow cell surfaces are pre-coated with oligonucleotides that are complementary to the adapter sequences added in step 1 (Figure 1.2). The denatured DNA library is introduced into the flow cell and individual fragments are allowed to hybridize to the flow cell oligonucleotides. The captured DNA fragments are copied by extension from the 3′ end of the flow cell oligonucleotide (Figure 1.3A), resulting in a copy of the original library fragment that is physically bound to the flow cell surface (Figure 1.3B). The bound DNA is allowed to loop over and hybridize to an adjacent oligonucleotide (Figure 1.3C). The bridged DNA is copied to produce a double-stranded DNA loop (Figure 1.3D). The dsDNA loop is denatured, producing two single-stranded DNA molecules (Figure 1.3E), that can

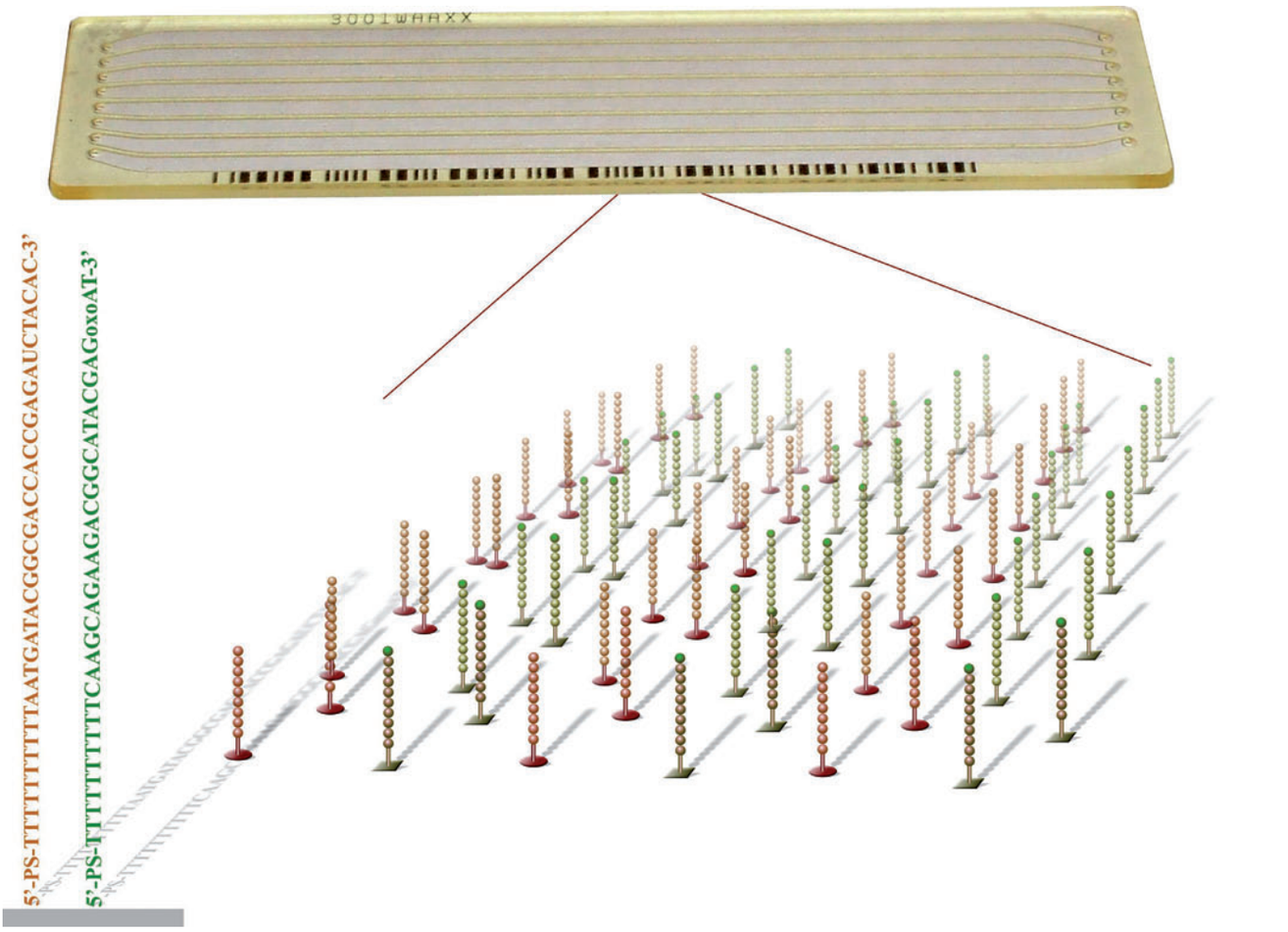

Figure 1.2 Sequencing flow cell. The surfaces of the eight-channel flow cell are coated with two oligonucleotides with each oligo containing a cleavable base. The oligo sequences are complementary to the adapter sequences, allowing the capture and amplification of a genomic library. The amplified clones or clusters will then be sequenced en masse.

loop over and hybridize to adjacent oligonucleotides to be copied again (Figure 1.3F). The bridge amplification process is repeated till a cluster is produced with about 2000 molecules (Figure 1.3G). The reverse strand is then cleaved off from each cluster (the flow cell oligonucleotides contain specific cleavable bases; Figure 1.3H), the free DNA ends are blocked to prevent unwanted extensions during the sequencing process and a sequencing primer is hybridized (Figure 1.3I). At the end of the cluster generation process, over one billion clusters are formed with each clone or cluster containing about a 1000 DNA molecules. The whole process is performed on a small automated cluster generation system, and takes about five hours.

3. Sequencing: the billions of clonal clusters in the flow cell are sequenced simultaneously and in parallel. An engineered DNA polymerase is utilized to add reversibly terminated, fluorescently labeled nucleotides to the sequencing primer (Figure 1.4A). The reversible terminator prevents the addition of more than one base at a time. The single added base is labeled with one of four colors; therefore each cluster will emit a fluorescent color depending on the added base. After dNTP incorporation, the clusters are imaged to record the color and location of each cluster (Figure 1.4B). The reversible terminator and the fluorescent dye are then removed, preparing the DNA molecules in the clusters for the next incorporation cycle (Figure 1.4A). The process is repeated about 150–250 times to build the sequence one base at a time. The DNA molecules are then flipped over by allowing the template strands to form loops by hybridizing to flow cell oligonucleotides. The reverse strand is re-synthesized by extension from the flow cell oligo. The forward strand is cleaved, leaving the reverse strand to be sequenced. This paired end sequencing strategy gives sequence information from both ends of each strand. Paired end sequencing produces twice the sequencing information from a library and facilitates an accurate alignment of sequenced

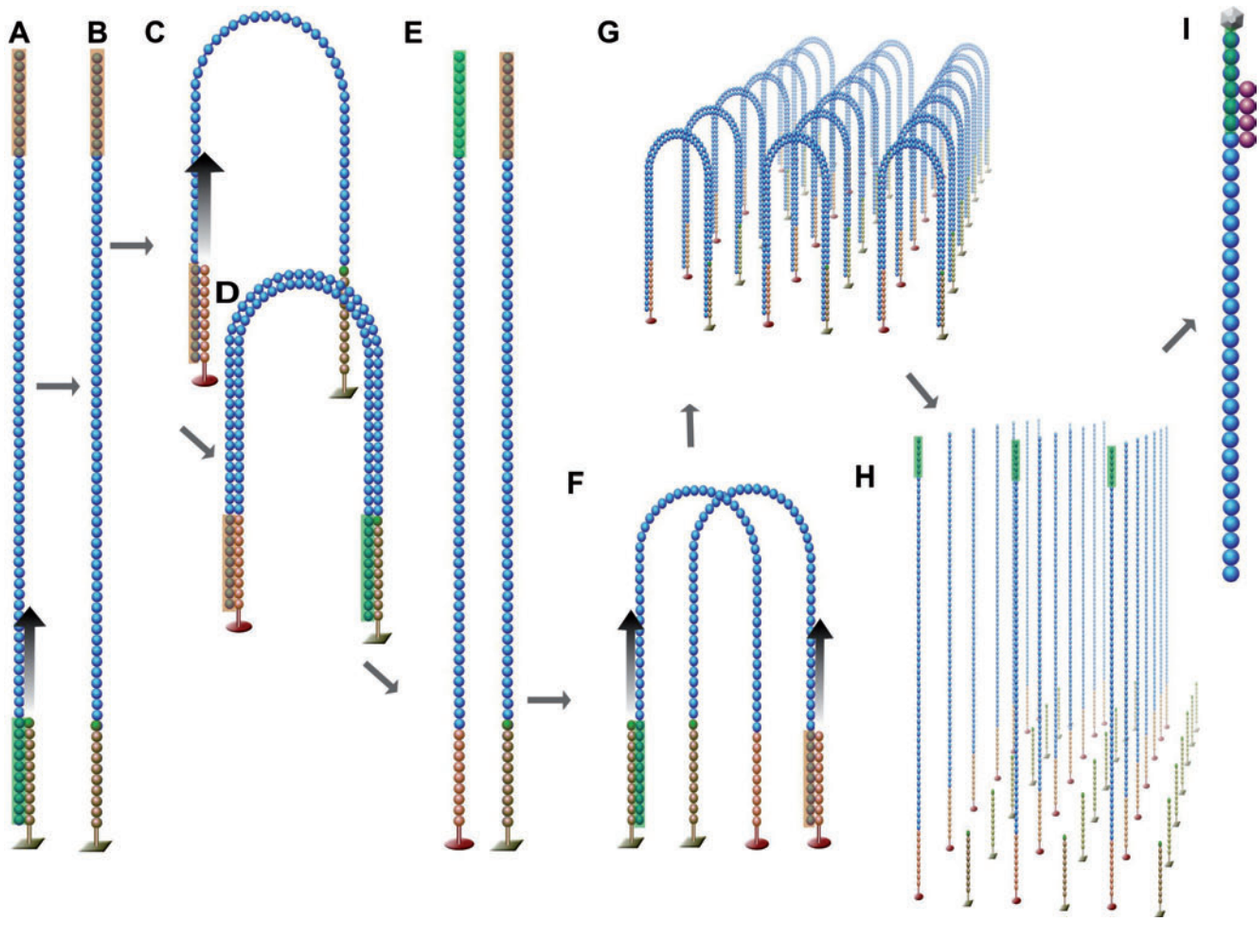

Figure 1.3 Generation of clonal clusters. Denatured DNA library fragments are captured onto the flow cell surface by hybridization to the oligonucleotides that coat the flow cell surface. The captured DNA fragment is copied by 3′extension (A) and the original fragment is denatured and discarded leaving a bound DNA fragment (B). Copied fragments are allowed to loop over and hybridize to an adjacent oligo (C). The single-stranded bridge is converted to a double-stranded molecule by polymerases (D). The dsDNA is denatured and the two strands (E) are allowed to hybridize to flow cell oligos and copied (F). The process is repeated till a tight cluster is formed (G). The cluster is denatured and the reverse strand is cleaved off (H). A sequencing primer is then added to the free ends of the DNA (I).

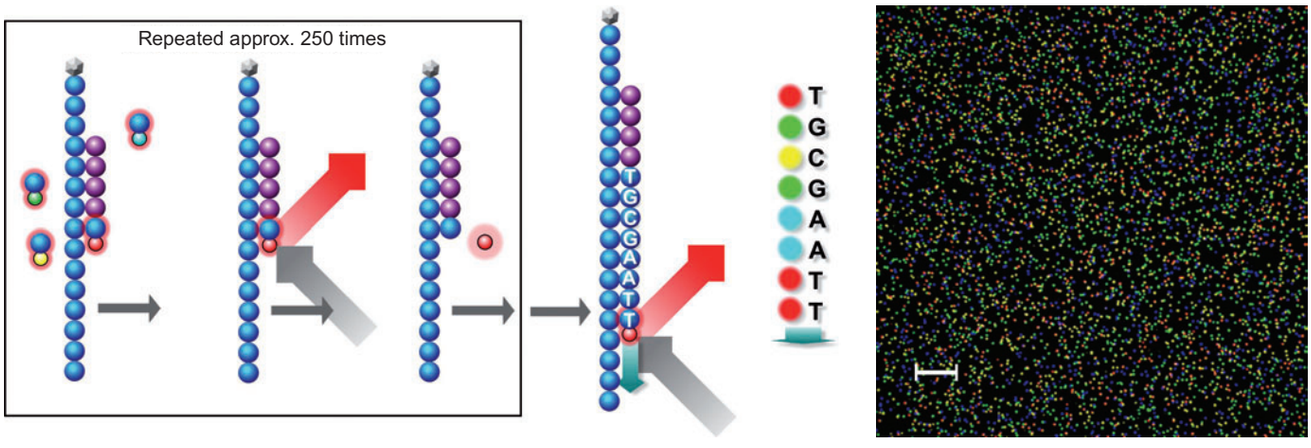

Figure 1.4 Sequencing of clusters. Clonal amplification produces more than a billion clusters attached to the flow cell surfaces. These clusters are sequenced in parallel by the addition of fluorescently labeled, reversibly terminated nucleotides in the presence of an engineered polymerase (A). After the addition of the single base, unincorporated nucleotides are washed off and the flow cell is imaged, producing a map (B) showing exactly which nucleotide was incorporated at each cluster (scale bar 10 um). After imaging, the terminator (and fluorescent dye) is removed, allowing the next base to be added. The process is repeated for about 250 cycles, building a 150 bp sequence one base at a time.

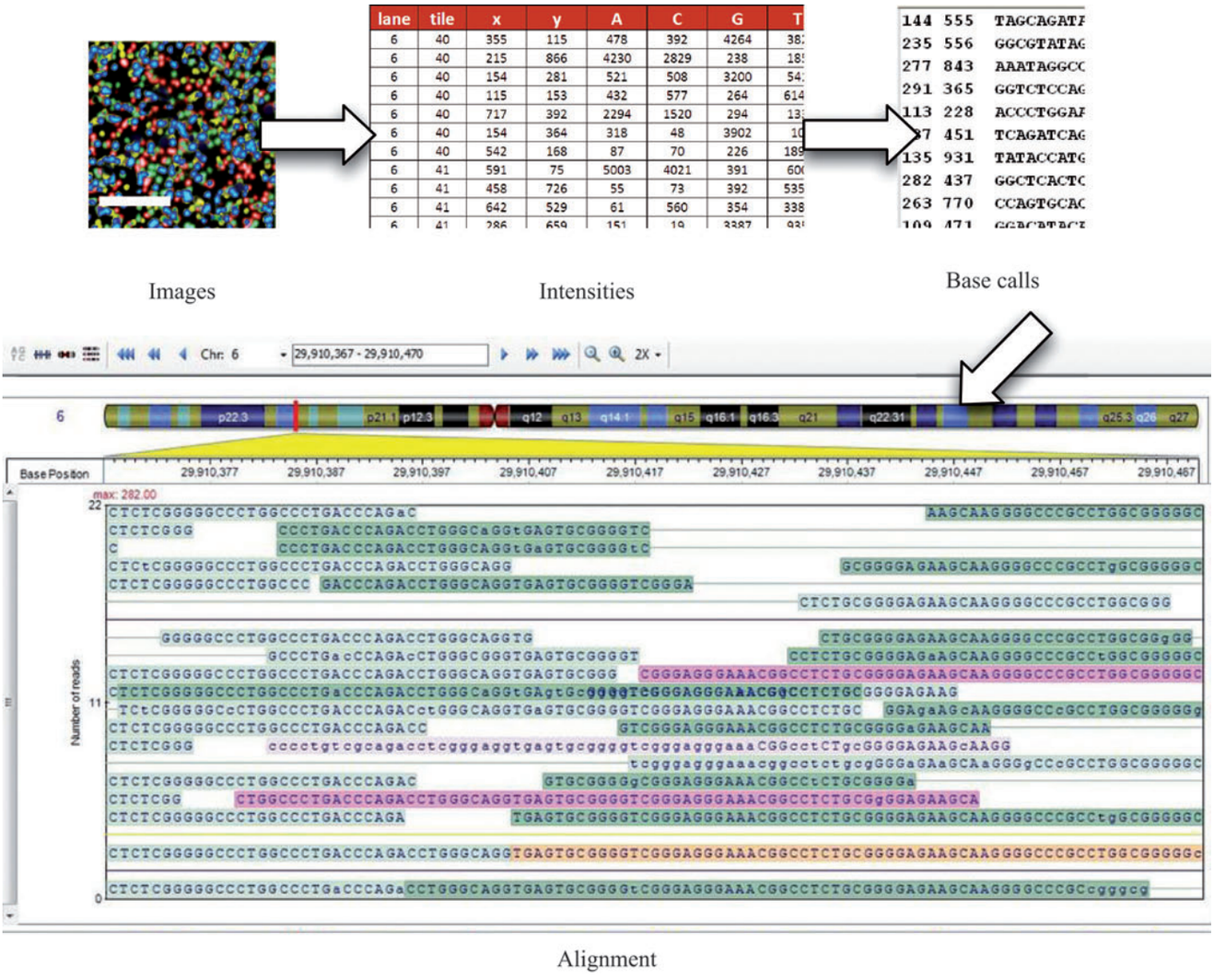

Figure 1.5 Data processing. The raw image data from the sequencer is converted to an intensity table that records the location of each cluster and the color intensity values (one color for each of the four bases). These numerical values are converted to a base call and used to assign a quality value for that base call. At the end of a sequencing run, the billions of clusters have each produced a 150 bp read, all of which are then aligned against a genome or against each other to produce overlapping contiguous fragments. The aligned reads are compared to the reference genome to derive a list of variants. The aligned reads are also converted to a format that can be imported into genome browsers for visualization.

fragments, since two reads separated by a known distance are now available.

4. Data analysis: the images captured during sequencing are converted to fluorescence intensity tables, which in turn are converted to base calls. The final reads (2 ×150–250 bp) are aligned to a reference genome and these alignments are converted to a format that can be visualized in genome browsers (Figure 1.5). The process of alignment involves a read being broken into smaller segments, called seeds. A seed is mapped to a reference genome. If a match is found, the location on the reference genome is recorded for that seed. If a match is not found, then another seed from the same read (32 bp downstream of the first seed) is now used to look for a match to the reference genome and so on. The process is repeated for the paired read from the same fragment to confirm the position of each DNA fragment on the reference genome. A genomic DNA library contains fragments from multiple copies of genomic DNA; depending on the sequencing library complexity, for every cluster that is formed in the flow cell there can be 10–10 000 other clusters whose sequences overlap significantly. Therefore each base will be read multiple times from independent clones, with the number of reads (read depth) being combined to make a high confidence base call (Figure 1.6). A change in the number of reads across the genome also provides a robust estimate of copy

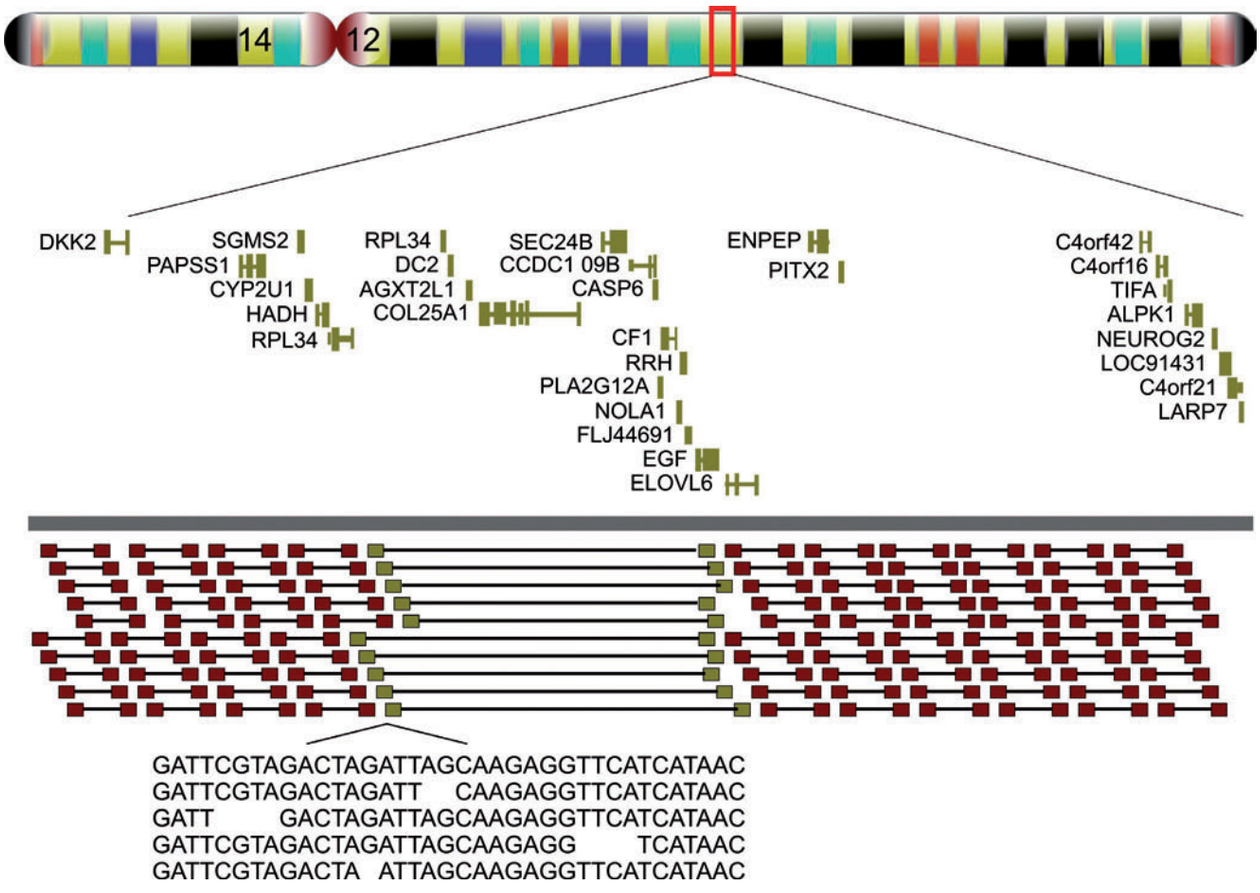

Figure 1.6 Detection of homozygous deletion. Simulated paired end sequencing data showing a genomic region where sets of the paired reads map at a greater distance than expected (green boxes) indicating a homozygous deletion. Alignment of the reads to a reference results in these reads seeming to be stretched apart, although in the sample these reads would be the correct distance because of the deletion event. Individual reads are shown below the paired ends showing the different fragments sequenced as individual clusters.

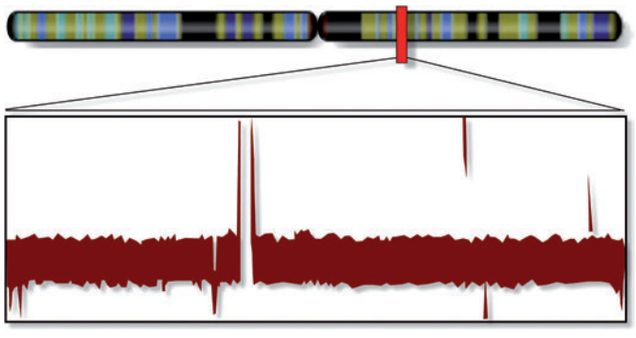

Figure 1.7 Copy number variation analysis. A dramatic change in the number of stacked reads in a region can indicate a copy number variation. Mapping the number of reads across the chromosomes therefore clearly shows all copy number variants at extremely high resolution. Zooming in to a region would also show base-level changes.

number variations (CNVs; Figure 1.7). Detecting CNVs by low depth sequencing, that is, each base is sampled at 1–5 times (1–5 × sequencing depth) provides an extremely effective and low-cost method to delineate all CNVs across the whole genome without any prior information on the location of a CNV.

Annotation and interpretation of genome sequencing data

In a typical cancer genome sequencing project, paired tumor and normal samples are sequenced to 40× average read depth, providing greater statistical accuracy for individual base calls. The sequence reads are assembled into a consensus sequence and compared against the reference to derive a list of variants. Anomalous regions are often resolved by local *de novo* assembly (that is, assembly without alignment to a reference). Single base substitutions are called from at least three independent high- quality sequence reads per allele. For small insertion and deletion (indel) detection, a local *de novo* assembly of all reads across the indicated region is carried out to form a consensus, and then followed with a gapped alignment of the consensus assembly. Larger structural variants are detected based on multiple independent anomalous paired end read alignments, i.e. pairs of reads where either the separation or orientation (or

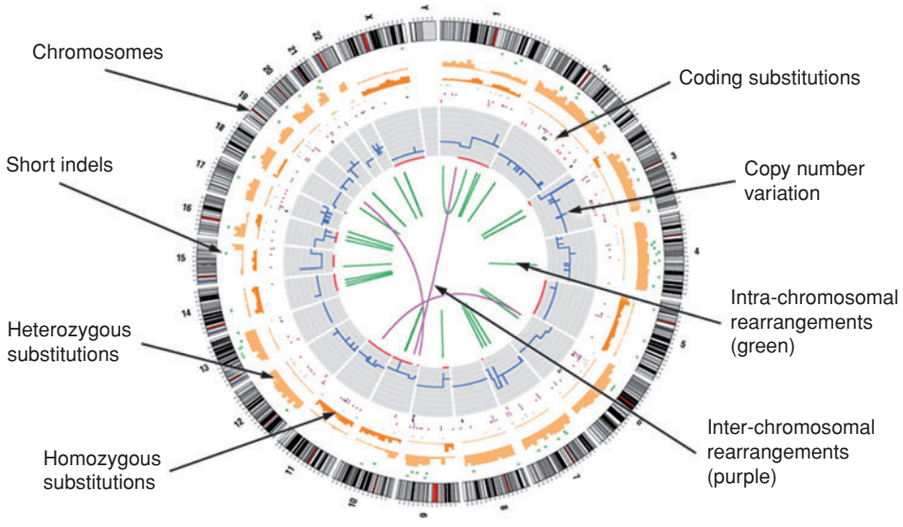

Figure 1.8 Catalog of COLO-829 somatic variants. The Circos plot displays all somatic variants in one compact form. The outermost ring represents the 23 human chromosomes arranged end-to-end in a circle. The inner notation shows the location of all the genetic changes which have arisen in a given type of cancer, in this case COLO-829, with each ring representing a different type of change: (from outside to inside) insertion–deletion events (indels), density of single-base substitutions (per 10 Mb), coding substitutions, copy number variants (CNV), and loss of heterozygosity (LOH). The lines criss-crossing the center of the plot show structural rearrangements, where whole sections have moved within a chromosome (green) or between chromosomes (purple).

both) deviate significantly from expectation, based on average DNA library fragment size. CNVs are identified by sequence read depth changes.

Germline variants are found in both the tumor and the normal sample, from the same individual. Somatic variants are those found only in the tumor. Somatic variants are then annotated with reference to gene, protein coding, and other information, and selected candidates (e.g. non-synonymous coding changes) are compared across samples to identify any common patterns that may reflect markers specific to the cancer. Visualization of the data is performed by importing the converted data files (BAM format) into genome browsers such as the UCSC and IGV genome browsers.

Somatic mutation profile of a cancer genome

In most cancer genomes, there could be thousands or tens of thousands of somatic variations, from single base mutations which might change the folding of a given protein to insertions or deletions that wipe out one or more entire genes. As one example, the complete catalog of somatic mutations in a melanoma cancer genome (COLO-829) is shown in a Circos plot (Figure 1.8). There were 33 345 single base substitutions, including 286 in the coding regions, and 1018 small indels, of which 14 were in the coding regions. There were also 37 structural rearrangements, of which 34 were intra-chromosomal (25 deletions, 6 insertions, 2 duplications, 1 complex change) and 3 were inter-chromosomal. In addition, 19 breakpoints in genes and 198 changes in DNA copy number were detected. Interestingly, the majority of somatic substitutions were C to T changes, a characteristic signature that differs from other studies. Furthermore, 92% of these changes were in a pyrmidine dimer (vs. 53% expected by chance). This pattern of C→T mutations correlates strongly with the known mutational mechanism associated with UV exposure that is the expected environmental mutagen in melanoma (11).

The significance of a genome variant is further evaluated based on the statistical concordance of the mutation in multiple samples with the same phenotype and the potential impact on biological function, for example, an introduction of a stop codon, a change in sequence that binds a transcription factor, changes in amino acids that alter protein function, production of novel proteins by gene fusion events, or changes in gene regulatory elements. These predictions help apportion the mutations into two broad categories. A small portion of the somatic mutations are termed "driver" mutations that directly impact oncogenic properties such as cell division, tissue invasion and metastasis, angiogenesis, and evasion of apoptosis (1). It is thought that only five or six driver mutations are required to convert normal cells into cancer cells (12). The identification of driver mutations will provide insights into cancer biology and highlight novel drug targets and diagnostic tests. The majority of somatic mutations in cancer genomes that do not contribute directly to cancer development are called "passengers" (6). They usually occur in a non-coding, non-regulatory region or coding region, but do not result in a change of amino acid in the protein (synonymous mutation). For example, in the case of COLO-829, of the 33K somatic substitutions identified, 23K were found between genes and another 9.5K were located in introns. Only about 1% of variants were found in coding regions and 1% in regulatory regions, with only five mutations located within candidate cancer genes (11).

An integrated view of cancer

Second-generation sequencing is being used extensively to provide a sensitive method to detect changes in the transcriptome. RNA sequencing of cancer transcriptomes provides a comprehensive view of the expression of coding and non-coding

Table 1.1 Cancer genome databases

Name	Link	Description	Maintained by
COSMIC	www.sanger.ac.uk/resources/databases/cosmic.html	Somatic mutations and CNV	Wellcome Trust Sanger Institute
ICGC	www.icgc.org/daco	Somatic mutations, expression, CNV, rearrangements	International Cancer Genome Consortium
TCGA	tcga-data.nci.nih.gov/tcga	Somatic mutations, expression, CNV, methylation	NIH
NCBI dbGAP	www.ncbi.nlm.nih.gov/gap	Genotype–phenotype analysis	NIH
Cancer Gene Census	www.sanger.ac.uk/genetics/cgp/census	Germline and somatic mutations implicated in cancers	Wellcome Trust Sanger Institute

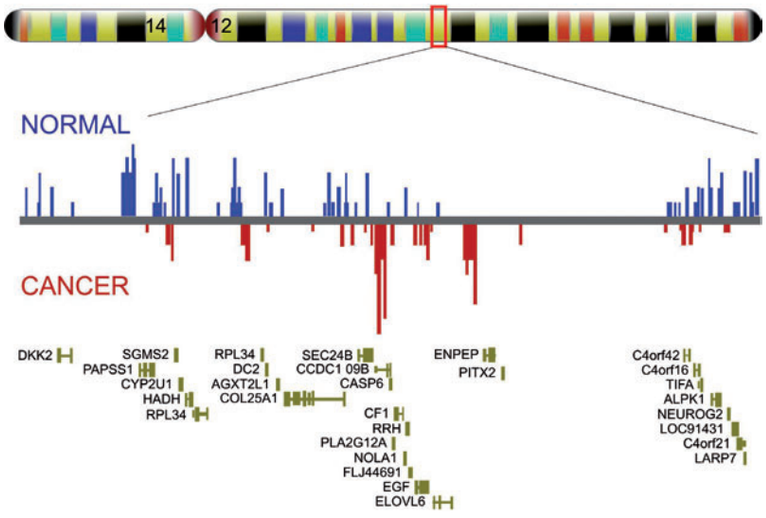

Figure 1.9 Analysis of gene expression by sequencing. RNA is converted to cDNA and sequenced as described. The sequence reads are then mapped to the genome. The number of reads mapping to a specific location provide an accurate estimate of the number of RNA molecules present in the original sample. As RNA sequencing is not based on a priori assumptions of gene models it can detect all transcripts, known and unknown. RNA sequencing data also clearly shows different splicing isoforms, gene fusion events, anomalous start and stop sites, and mutational events.

genes (Figure 1.9), perturbation in RNA splicing, and an accurate annotation of gene fusion events, RNA-editing, and coding cSNPs. Next-generation sequencing has also been applied very successfully to study changes in transcription factor binding profiles, modification of the chromatin structure, alteration of DNA methylation patterns, and the detection of cancer-associated small RNA in serum.

Currently, several systematic and large-scale cancer genome sequencing initiatives are supported by the USA NIH's Cancer Genome Atlas Pilot Project (TCGA; http://cancergenome.nih.gov; 13), the International Cancer Genome Consortium (ICGC; http://www.icgc.org; 14) and multiple other projects. Collectively, these efforts will collect more than 25 000 cancers from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe and analyze them at the genomic, epigenomic, and transcriptomic levels to reveal the repertoire of oncogenic mutations, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies (14). Several valuable public databases have been established to reflect progress in all these projects (Table 1.1).

Summary

The ability to sequence whole cancer genomes at low cost provides an unrivalled and previously inaccessible opportunity to comprehensively characterize individual cancers with high specificity and sensitivity. This has led to a new understanding of oncogenic mechanisms and identification of potent diagnostic and prognostic markers, which are used to classify a tumor and accurately predict the response to treatments (15). We predict an even broader application of whole-genome technologies to analyze the large libraries of archived tissue samples as well as samples that can be accessed non-invasively, such as circulating tumor cells. In the near future we anticipate a revolution in healthcare and personalized cancer treatment with accurate, high-throughput, low-cost sequencing at the heart of clinical practice (16).

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
2. Metzker ML. Sequencing technologies: the next generation. Nature Reviews Genetics 2010;11:31–46.
3. Mardis ER. A decade's perspective on DNA sequencing technology. Nature 2011;470:198–203.
4. Lander ES, Linton LM, Birren B, *et al.* Initial sequencing and analysis of the human genome. Nature 2001;409: 860–921.
5. Wheeler DA, Srinivasan M, Egholm M, *et al.* The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452: 872–6.
6. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458: 719–24.
7. Parsons DW, Jones S, Zhang X, *et al.* An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321:1807–12.
8. Wiegand KC, Shah SP, Al-Agha OM, *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. New England Journal of Medicine 2010;363:1532–43.
9. Ding L, Ellis MJ, Li S, *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 2010;464:999–1005.
10. Stephens PJ, Greenman CD, Fu B, *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011;144:27–40.
11. Pleasance ED, Cheetham RK, Stephens PJ, *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010;463:191–6.
12. Schinzel AC, Hahn WC. Oncogenic transformation and experimental models of human cancer. Frontiers in Bioscience 2008;13:71–84.
13. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.
14. Hudson TJ, Anderson W, Artez A, *et al.* International network of cancer genome projects. Nature 2010;464: 993–8.
15. Leary RJ, Kinde I, Diehl F, *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. Science Translational Medicine 2010;2:20ra14.
16. McDermott U, Downing JR, Stratton MR. Genomics and the continuum of cancer care. New England Journal of Medicine 2011;364:340–50.

Part 1.1 Analytical techniques: analysis of DNA

Chapter 2

Genome-wide association studies of cancer predisposition

Zsofia K. Stadler, Sohela Shah, and Kenneth Offit

Introduction

Until relatively recently the field of medical genetics has focused on the identification and treatment of rare, single-gene disorders that are usually associated with a high risk of a particular disease or trait. However, such high-penetrance susceptibility loci only account for a fraction of the familial aggregation of common diseases, and the genetic architecture of complex diseases, such as cancer, is probably better characterized by "polygenic" inheritance, wherein heritability is determined by the joint action of multiple genes. Since 2007, genome-wide association studies (GWAS) have resulted in a paradigm shift in the discovery of gene–disease associations and helped to identify multiple low-penetrance germline genetic variants for most common cancer types (1,2). Through a hypothesis-neutral genome-based approach, GWAS compare frequencies of common DNA variations, usually in the form of single-nucleotide polymorphisms (SNPs), in a large set of unrelated cases and controls to identify genetic variants associated with disease risk. GWAS have resulted in the mapping of susceptibility loci for many human diseases, including the identification of over 100 loci for cancer, most of which were not previously implicated in carcinogenesis.

This unprecedented rapid amassing of new genetic risk variants associated with cancer risk has generated hope that these germline markers may prove useful for cancer prevention, improve our understanding of cancer pathogenesis, and possibly lead to a new era of personalized genomics (3). However, as the majority of genetic variants confer only a modest risk of disease with effect size under 1.5 (Figure 2.1) and the biological mechanisms underpinning most associations are unknown, significant scientific barriers must be overcome before GWAS results can be meaningfully translated into patient care.

The genetics of cancer predisposition

In the 1980s and 1990s, genetic linkage studies identified a number of highly penetrant syndromes responsible for cancer predisposition within certain families. Molecular insight into breast–ovarian cancer (*BRCA1* and *BRCA2*), Lynch (mismatch repair genes), Li Fraumeni (*P53*) and Cowden (*PTEN*) syndromes, among others, has had a powerful impact on current clinical management of families affected by these syndromes. As such highly penetrant syndromes account for only a fraction of the familial risk of cancer, much of the heritability of cancer remains to be explained.

A strong hereditary component to cancer is supported by epidemiologic evidence from twin studies demonstrating concordance for most cancers among monozygotic compared with dizygotic twins and siblings (4). Population studies of familial clustering (5) suggest that genetic factors predispose to even environmentally induced diseases such as lung cancer. Using a candidate gene approach, pathways thought to be important in carcinogenesis have been investigated. In breast cancer, for example, a small number of candidate genes involved in response to DNA damage (*ATM* (6–8), *CHEK2* (9–11), *BRIP1* (12), and *PALB2* (13)) have been associated with a modest increase in breast cancer risk (approximately twofold). However, together with the known high-penetrance genes, all known breast cancer predisposition genes account for only about 25% of the familial aggregation of breast cancer in outbred populations (14). One explanation for this seemingly "missing" fraction of heritable disease is the common variant common disease (CVCD) hypothesis (15), which assumes that many different common genetic variants, each with a small effect size, collectively cause disease. Recent developments, including the completion of the Human Genome Project, the HapMap Project, and the emergence of high-throughput genotyping, have allowed further investigation of this hypothesis, through the GWAS approach.

Experimental design of genome-wide association studies

Genomic variation in GWAS reported to date is largely represented by single nucleotide polymorphisms (SNPs). GWAS rely on linkage disequilibrium (16) wherein SNPs are inherited in blocks, with many nearby SNPs being highly correlated. This enables the selection of one SNP (the tag SNP) that represents up to 50 000 surrounding base pairs. Through the HapMap

Molecular Oncology, ed. Edward P. Gelmann, Charles L. Sawyers, and Frank J. Rauscher. Published by Cambridge University Press.