Cancer genomics

Barbara L. Weber[1]

Abramson Family Cancer Research Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104
[1]Correspondence: weberb@mail.med.upenn.edu

The draft human genome sequence and the dissemination of high throughput technology provides opportunities for systematic analysis of cancer cells. Genome-wide mutation screens, high resolution analysis of chromosomal abberations and expression profiling all give comprehensive views of genetic alterations in cancer cells. From these analyses will come a complete list of the genetic changes that drive malignant transformation and of the therapeutic targets that may be exploited for clinical benefit.

Cancer is a disease of the genome, which is invariably altered at multiple sites in cancer cells. The goal of cancer research is to define these molecular defects and turn these discoveries into effective treatment and prevention regimens. Until recently, these efforts relied on laborious, and necessarily limited, one-gene-at-a-time approaches. However, technological advances, coupled with the draft human genome sequence, promise to both accelerate this research and fundamentally change how we think about cancer. Discrete molecular changes and the perturbation of associated pathways ultimately will be placed in context, enabling construction of a multidimensional map of the complex circuitry of cells and how it is altered in specific cancers. Subgroups of tumors will be defined by recurrent patterns of alterations. Tumor classification, now based on morphology, which is inexact, subject to marked interobserver variation, and not grounded in biologic relevance, will be replaced by a molecular classification scheme. Although to date, global expression profiling has been central to developing novel molecular classification schemes, a comprehensive mutation analysis, a whole genome catalog of submicroscopic chromosomal aberrations, or a global methylation scan will likely serve the same purpose. In fact, classification schemes are likely to have the most predictive value when data from multiple platforms are combined. Therapeutic choices will be specifically targeted at the altered pathways that define cancer subgroups, and will expand greatly as we develop a genome-wide catalog of cancer-related alterations. Thus, genomics is not just increasing the pace of cancer research, it is providing a completely new view of cancer cells, allowing us to see the whole cell at once, rather than bits of the cellular machinery in isolation.

Genomics will provide extremely detailed information and a global framework for understanding complex patterns in cells. Yet, as these complex datasets begin to emerge, so do the problems of interpreting them. What is functionally significant and what is noise? Which of the many differences between normal and tumor cells are the most promising molecular targets for development? What are the defining features of a predictive molecular classification scheme that can be reduced to clinical use? What are the limitations of and sources of error from currently available genomic technology, and how can the scientific community develop critical peer review of these data? And finally, how can these data be made publicly available, enhancing their usefulness by allowing data mining by individuals of diverse viewpoints and interests? Addressing these problems, and thereby taking full advantage of the power of genomics, will require the implementation of high throughput validation schemes, new paradigms for drug development, close collaboration between laboratory scientists and clinicians, and the development of standard nomenclature, data sharing formats, and quality-control measures—none of which exist at present.

The genomic tools discussed here may be considered in three broad categories: those that identify regions of the genome harboring genes altered in cancer, those that identify mutations in specific genes, and those that identify genes with an altered expression profile in cancer cells (Figure 1). Proteomic tools, those needed to catalog global protein expression profiles and posttranslational modifications, pose largely distinct and very complex challenges and are not considered here.

Which genes are altered in cancer?

Cancer arises from an accumulation of mutations in a series of genes over time. A panel of genes that are mutated in sequence in colon cancer has been proposed (Kinzler and Vogelstein, 1996), and several genes, such as p53, have been evaluated extensively in a wide range of tumor types. Yet to date, not one human cancer has been completely described with a clear picture of all the genes altered in that cancer. In addition, even a complete profile of the mutations in a given cancer represents only a single point in time. Cancers evolve, mutations accumulate, and cellular heterogeneity develops, complicating the analysis of tissue and accurate descriptions of individual clones. Cancers may also develop in the setting of field defects, where surrounding tissue, while appearing histologically normal, contains abnormalities that not only give rise to the tumor but may affect its behavior. Even completely normal stroma plays a role in tumor invasion and metastases, a effect that may vary as a function of molecular defects in a cancer cell, adding yet another layer of complexity to defining the alterations that result in cancer.

Thus, while a more complete list of genes mutated in cancers may be assembled by screening more cancers, and paralog searching, although limited (Futreal et al., 2001), may be of some use in the selection of candidate genes for further study, only whole genome approaches will provide a complete description, unbiased by what we think should or might be altered. An excellent example of a "post-genome" project designed with this in mind is the Cancer Genome Project (CGP), sited at The Sanger Center (Wellcome Trust Genome Campus, Hinxton, UK). The CGP employs a high throughput

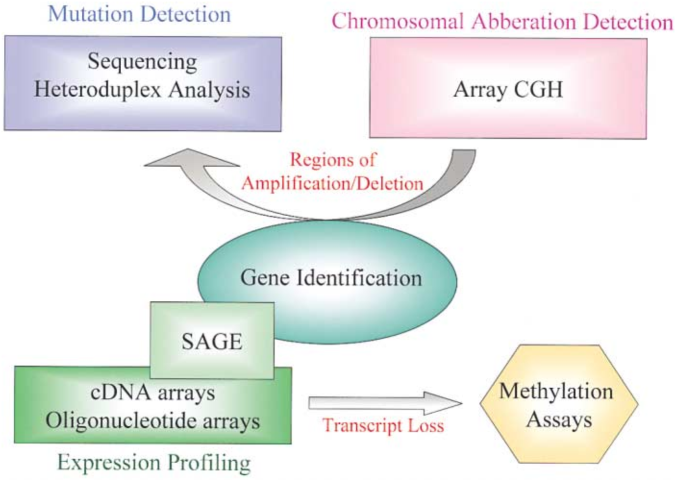

Figure 1. Genomic technologies for the analysis of cancer

Direct sequencing and heteroduplex detection provide the most accurate and flexible approaches to high throughput analysis of small intragenic mutations. Array comparative genomic hybridization (CGH) is the technology most amenable to identifying larger deletions and amplifications, but once identified, these regions must be analyzed to determine which genes have an effect on cancer development. Expression profiling will identify up- and downregulated transcripts that may occur in the absence of detectable DNA mutations. However, no technology has yet evolved as the clear solution to cataloging global methylation profiles. Downregulated transcripts detected using expression profiling may in turn be evaluated for promoter methylation, but this two-step process is cumbersome and will not identify effects of methylation other than gene silencing.

mutation detection scheme to evaluate all human genes for mutations in a large panel of tumors and cell lines with ultimate goal of answering the question, “which genes are mutated in human cancers?”

How many genes are altered in each cancer?

The exact number of alterations required for cancer to develop into clinical disease is not known. It is likely that the number varies by both type of alteration and tumor type, with estimates ranging from a single translocation in a hematologic malignancy, to three carefully selected mutations in an experimental model (Hahn et al., 1999), and to probably many more in adult solid tumors.

Providing a framework for considering which genes may be altered in cancer cells, Hanahan and Weinberg (2000) proposed six cellular capabilities they believe must be acquired by cancer cells. These include growth signal autonomy, evasion of apoptosis, insensitivity to antigrowth signals, sustained angiogenesis, limitless replicative potential, and the capacity to invade tissue and grow at metastatic sites. The means by which these capabilities may be acquired vary mechanistically and chronologically, and the number of mutations required to acquire a specific capability varies as well, but the end result is proposed as invariant. As such, this model allows for considerable flexibility in the absolute number of mutations that are required for cancer to develop. For example, a p53 mutation will both facilitate angiogenesis and evasion of apoptosis, a five-step model if all other capabilities are acquired individually, but in another cancer, several mutations may be required for the acquisition of each capability, such that in some cases, many mutations may be required before cancer fully develops.

The therapeutic implications of defining the number of disease-associated mutations in a cancer cell also remain unknown. One might expect that if a single mutation confers all six necessary capabilities of cancer development, then this tumor should be easy to treat if the appropriately targeted therapeutic agent can be delivered. But does this ever occur? And even if it does, is this reasoning correct? Can a single agent be expected to reverse all six properties? Do all six cancer-related capabilities need to be reversed, or is reversing one or two of them enough to eradicate the tumor? Again, attempts to answer these questions are unsupported by analysis of even a single human cancer in which one can confidently enumerate and define the critical cancer-causing mutations, and even a complete list of all mutations in a given cancer may be too simplistic. Considering mutations in the context of transcriptional and epigenetic changes and the surrounding tissue also may be required, and is the real promise of genomics—a comprehensive understanding of the complex cellular and host environment of cancer cells created by combining multiple platforms and analytic approaches.

How are genes altered in cancer?

Understanding mechanisms by which genes are altered is central to choosing analytic techniques and is probably necessary to develop diagnostic and therapeutic strategies. Base substitutions can activate proto-oncogenes (often by enabling ligand-independent proliferative signaling) and inactivate tumor suppressor genes, sometimes producing dominant negative effects. Large genomic deletions, small intragenic insertions and deletions, and promoter silencing by methylation or other epigenetic effects are primarily relevant to tumor suppressor genes. These changes may result in complete absence of protein, unstable or absent transcripts, a truncated but inactive product, or, as noted above, a truncated or mutated protein that may sequester wild-type protein or act as a competitive inhibitor of wild-type function.

In considering the significance of mutations in putative tumor suppressor genes, and thus their value as therapeutic targets, it is normally assumed that Knudsen’s two hit hypothesis applies (Knudsen, 1971). However, recent work suggests that in some cancer susceptibility syndromes, as well as in sporadic tumor formation, haploinsufficiency also may be important. Murine models of haploinsufficiency include monoallelic loss of PTEN in the promotion of prostate cancer progression (Kwabi-addo et al., 2001) and Smad4/DPC4 (+/–) mice with gastric polyps that progress to carcinoma in situ without loss of the remaining Smad4/DPC4 allele (Xu et al., 2000). Studies of P27KIP1 in acute lymphoblastic leukemia provide some evidence for haploinsufficiency in human cancers (Fero et al., 1998), but the identification of germline mutations CBFA2 in familial platelet disorder (FPD) provides perhaps the best evidence that gene dose is important for tumor suppression (Song et al., 1999). Individuals with FPD are heterozygous for germline mutations in CBFA2, with defects in both platelet number and function, despite the presence of one normal, expressed allele. Acute leukemias develop in 30%–50% of mutation carriers, without loss of wild-type CBFA2 expression. These effects should come as no surprise, given the numerous examples of gene dosage effects on the development and fitness of organisms, with trisomies and contiguous gene deletion syndromes being well-described examples. However, they do suggest that the common practice of dismissing candidate

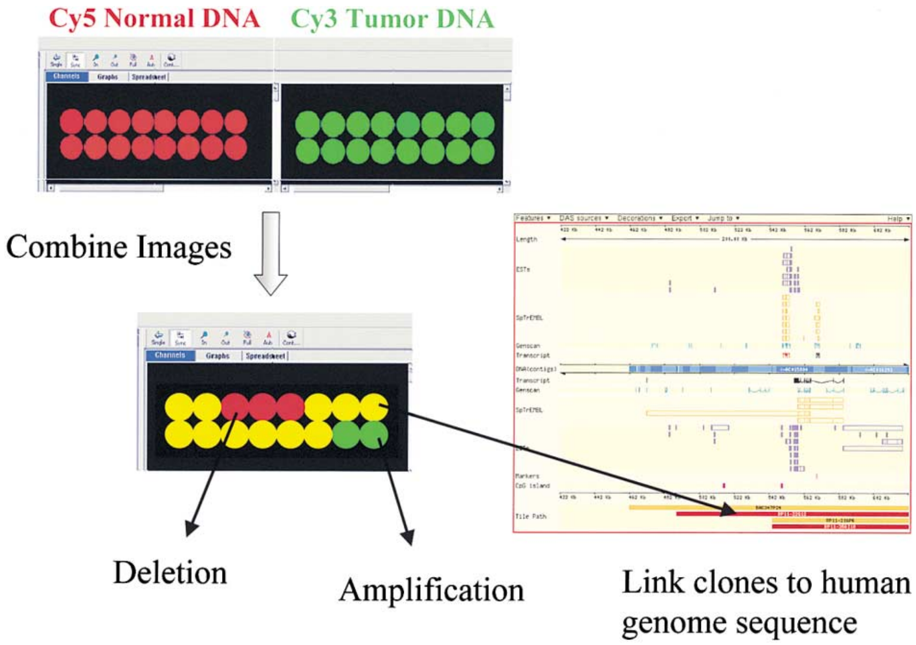

Figure 2. Array-based comparative genomic hybridization

First, tumor DNA is prepared and labeled with Cy3 (green) and cohybridized to an array made up of genomic clones with normal genomic DNA, labeled with Cy5 (red). A laser scanner reads spot intensities for each dye separately, and then the images are overlain to obtain the combined image. Spots appearing yellow when both tumor and normal DNA contain equal copy number DNA at the locus represented by a specific clone. Using this color scheme, amplified chromosomal regions appear green and deleted regions appear red. However, in reality, the relative intensities must be normalized to determine signal intensity representing diploid copy number, then the ratio of Cy3:Cy5 adjusted for unequal fluorescence, and finally ratios of normal to tumor DNA for each spot can be computed and plotted for visual analysis.

tumor suppressor genes when mutations are only identified in one allele may result in missing important molecular targets, and that we should pay close attention to transcripts with reduced but not absent expression levels.

Gene amplification (another example of gene dose effect) often affects growth factor receptor genes and gives tumor cells a selective advantage by enhancing response to levels of ligand that would not otherwise trigger proliferation. While amplification may generate many gene copies, data from studies of the MET oncogene in papillary renal carcinoma suggest that the combined effect of an activating point mutation coupled with only a small increase in copy number (3–4 copies per cell) may also play an key role in transformation (Fischer et al., 1998). Finally, oncogenes may be created by balanced chromosomal translocations, often juxtaposing the functional domain of a transcription factor with heterologous sequences that remove normal constraints on signaling, such as the bcr-abl oncogene created by the 9:22 translocation in chronic myelogenous leukemia.

While defining individual alterations is clearly important, combining data from multiple platforms is central to placing genetic alterations in context. As one example of tumor suppressor mutations that depend on cellular context, p53 and BRCA1 mutations are cooperative in tumorigenesis, and a p53 mutation, or one in a related checkpoint, may be required for the tumorigenicity of BRCA1 mutations (Bertwistle and Ashworth, 1998). This may be because loss of BRCA1 function with intact checkpoints leads to cell death, but in the absence of normal checkpoint function, cells survive and catastrophic genome instability develops. Yet even pairwise gene interactions provide a very simplistic example. Considering the thousands of genes that differ in expression between normal and tumor cells, it is likely that multiple interactions at each level of analysis are relevant to considering the biologic behavior of a cancer. One approach to looking at gene mutations in context is to evaluate the expression profile of a cancer from the standpoint of known mutations in that tumor, yet considering the number of genes that are mutated in various cancers and the enormous number of possible transcriptomes in human cancer, even this straightforward approach is a daunting task. Full characterization of tumor cells will require data sharing between investigators to generate sample sets with adequate statistical power and will likely produce models that need to be further refined with the addition of data from other genomic platforms, again increasing the complexity of the analysis.

Despite the complexities that develop when exponentially increasing the amount of data that describe a tumor, another advantage of integrating more than one platform is the ability to strengthen functional inferences. Using genome-wide expression profiling coupled with "interactome" data from *Saccharomyces cerevisiae* found in public databases, Vidal and colleagues have shown that genes with similar expression profiles are more likely to encode interacting proteins (Ge et al., 2001). While clustering transcripts with similar expression patterns suggests placement of gene products into identical or related pathways, the integration of genome-wide protein-protein interaction data with expression analysis enhances the inferences that can be made, as well as the resolution of those inferences. In this case, the combination of two datasets suggests where in a pathway proteins may act, and exact points at which cellular pathways intersect.

High resolution detection of chromosomal aberrations

Array comparative genomic hybridization

Array comparative genomic hybridization (CGH), while not useful for detection of small intragenic mutations, is extremely well suited to high-throughput, whole genome detection of chromosomal gains and losses at high resolution (Figure 2). Further streamlining this methodology, the availability of the draft human genome sequence means that chromosomal loci with copy number variation can be linked to the human genome sequence directly and immediately by the map location of clones on the array, and a list of candidate genes generated very quickly. The initial study used arrayed BAC and P1 DNA from chromosome 20 with a mean spacing of 3 Mb (Pinkel et al., 1998) and demonstrated the feasibility of detecting both gains and losses with single copy sensitivity using array CGH. Brown and Botstein subsequently demonstrated the utility of cDNA arrays for array CGH, which have the benefit of targeting analyses at expressed genes (Pollock et al., 1999). However, experience in several labs suggests that arrays constructed from large insert genomic clones result in more uniform hybridization kinetics than those made up of cDNAs, reducing the difficulty of normalizing signals. The current sensitivity of array CGH is limited largely by the spacing of genomic clones used to construct the arrays, but a BAC library with an average spacing of 1–2 Mb across the genome is now publicly available and the entire minimum tiling path of BAC clones is expected to become a public resource shortly (http://www.ncbi.nlm.nih.gov/genome/clone/ordering.html).

With adequate attention to normalization, array CGH is amenable to detection of both homozygous and hemizygous deletions that similarly may mark the location of tumor suppressor genes, defects previously detectable only through laborious analysis of individual candidate loci. Detection of large homozygous deletions is likely to be particularly useful in the identification of tumor suppressor loci—*p16*/*MTS1* (Kamb et al., 1994), *SMAD4* (Hahn et al., 1996), *PTEN* (Li et al., 1997), *hSNF5*/*INI1* (Versteege et al., 1998), and *RB* (Friend et al., 1986) all were mapped based on homozygous deletions. Metaphase FISH, or modifications such as SKY (spectral karyotyping), remain the best way to identify translocations; however, neither of these are amenable to high throughput approaches. Chromosomal regions with frequent loss of heterozygosity (LOH) are also thought to contain tumor suppressor genes. Thus, even in the absence of a homozygous deletion, array CGH is a powerful gene identification technique, as recently illustrated by Gray and colleagues (Hodgson et al., 2001) with an analysis of 84 murine islet cell tumors. The ratio of the CY3 and CY5 signal from genome regions with normal copy number (two for autosomes and female sex chromosomes, one for each sex chromosome in males) is used as a reference to define regions of the genome with amplifications (copy number >2) and deletions (<2). The analysis validated previous observations of recurrent LOH on mouse chromosomes 9 and 16, narrowed the candidate regions as defined by genotypic mapping, and identified previously unrecognized regions of recurrent allelic imbalance in regions syntenic to human chromosomes 12p11-13, 16q24.3, and 13q11-32 (losses) and 20q13.2 and 1p32-36 (gains).

While examples of novel tumor suppressor genes identified beginning solely with a region of LOH are hard to come by, there are many examples of known tumor suppressor genes where one allele is inactivated by a large deletion. These findings suggest that the technology needed to find cancer-related genes in large chromosomal deletions has been the limiting factor, not that regions of LOH do not harbor tumor suppressor genes. However, the genome instability and aneuploid nature of most cancer cells poses difficulties for array CGH—reduction in copy number from two to one is easily detectable with a robust normalization scheme, but not all LOH follows this simple model. Where there is mitotic recombination or chromosomal loss and reduplication, copy number remains unchanged, but only a single allele is present, and it may be a mutant allele. Other regions of LOH may be undetectable because loss of the wild-type alleles occurs in the presence of multiple chromosome copies; thus, the alleles may be present in ratios of 3:2 or 4:3, for example, possibly of functional significance, but below the level of detection (reductions in dose of at least 50%).

Array CGH has two features that make it particularly amenable to analysis of cancer cells. It is relatively insensitive to normal cell contamination, and high quality probes may be prepared from small amounts of archival material. Data from Gray and colleagues (Hodgson et al., 2001) show that it is possible to detect single copy changes in the presence of as much as 60% contamination with normal cells. This is particularly important in the analysis of tumors such as prostate and pancreatic cancer, where the tumor often is so interdigitated with normal stroma that even the use of laser capture microdissection may not be adequate to produce pure populations of tumor cells. Normal contamination also may occur from the lymphocytic infiltrates found in many adult solid tumors. But unlike expression profiling, contaminating normal cells should be invariant with respect to the measurement being taken (i.e., genome copy number); thus, if they do not compromise the ability to see copy number change, they will not complicate the analysis.

The other attractive feature of array CGH is the success of preparing high quality probes from small archival specimens. The need for unfixed tissue is a very limiting aspect of expression analysis of cancer cells. The ability to use paraffin-embedded tissue would greatly expand the number of specimens available for expression analysis, but high quality RNA generally is not extractable from archival material. However, array CGH, requiring DNA probes, does not suffer from this limitation. Pathology departments in most academic centers never discard archival material, and most community hospitals have at least 5 years of paraffin-embedded tumor blocks in storage. With appropriate approval, these blocks can be linked to medical records, facilitating studies that correlate clinical outcomes with patterns of chromosomal gains and losses. This may be particularly informative for the cancers that arise in the setting of a field defect, such as lung cancer due to carcinogen exposure, or breast cancer, possibly related to multiple cycles of cellular proliferation and regression. The resulting clonal expansion of morphologically normal adjacent breast tissue with unsuspected allelic imbalance may contain the earliest changes that initiate transformation. The ability to separate and analyze this tissue will be very informative in defining very early changes—an application where array CGH has particular advantage. Finally, many premalignant lesions can be fully analyzed by array CGH, again because of the ability to use small amounts of archival material. These lesions may be found alone or adjacent to invasive cancers, and provide the best reagents to define the accumulation of mutations during tumor development. They rarely have been available for analysis because of their small size and near impossibility of obtaining these lesions without fixative—the latter due to the need for clinical pathologists to examine an entire specimen rather than provide frozen tissue for research studies.

Array CGH thus has enormous potential to speed the identification new cancer genes, evaluate sequential genetic changes in tumor progression, and add to tumor classification schemes. Current limitations are the availability of this technology in only a few labs, the cost of commercial services (as much as $1000 a sample for high resolution arrays), and the relative lack of software to support the analyses. But the biggest unknown is the utility of the data that will be generated. How much will we miss buried in the heterogeneous and aneuploid collections of cells that make up cancers? Will cell lines, which have their own limitations, provide useful information, or will there be so much noise from secondary genetic changes that the false positive and negative rates will be prohibitive? Are there sufficient numbers of small homozygous deletions to war-

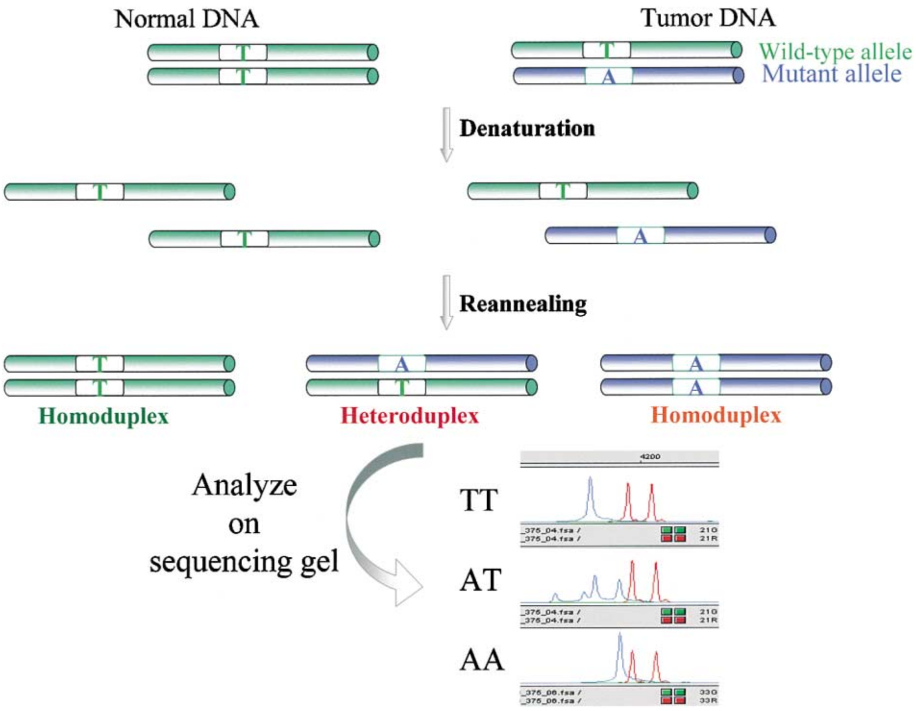

Figure 3. Heteroduplex-mediated mutation analysis

Heteroduplexes form following denaturation and reannealing of fragments from mixed normal and tumor DNA if the sequence varies between fragments. In this case, the normal sequence contains an adenine (A), which has been mutated in the tumor DNA to a thymidine (T). The necessity of mixing DNAs from normal and neoplastic tissue arises because allelic loss, frequently the means by which the remaining wild-type allele is lost, reduces a somatic mutation to homozygosity, precluding formation of heteroduplexes. Detection can be performed on standard sequencing gels, automated flat-gel sequencers, or capillary sequencers.

rant the analysis of extensive sample collections? And the biggest question: will the multiple regions of recurrent LOH described in the literature finally yield their long promised tumor suppressor genes with enhanced mapping resolution? These questions can only be answered as multiple investigators take up this new approach to analyzing cancer genomes, but the vast potential justifies a significant investment in this strategy.

High throughput mutation detection

Genomic sequencing

High throughput genomic sequencing is one way to define the genes that are mutated in cancer, but the technical hurdles are significant, and the problem of sorting disease-associated mutations from rare polymorphisms is enormous. Sequencing genomic DNA detects the most common alterations in tumor suppressor genes—base substitutions, small intragenic insertions, and deletions—with high sensitivity, and provides data in final form without an additional step. However, for most facilities, high throughput sequencing is a wholly impractical consideration for analyzing the genome of even a single cancer, due to the effort and cost required for such an undertaking. In addition, large deletions, rearrangements, and balanced translocations, while theoretically detectable with extensive sequencing of libraries prepared from individual tumors, will be missed, with rare exception. Finally, epigenetic changes such as promoter methylation are not detectable by sequencing without prior DNA modification, such as bisulfite treatment, that converts nonmethylated cytosine to uracil (Clark et al., 1994). Thus, despite the theoretical advantages of sequencing, it is not currently practical for large-scale mutation analyses, leading to the consideration of other strategies that are less labor intensive and less costly.

Heteroduplex detection

Heteroduplex detection circumvents many of the difficulties inherent in sequencing for mutation detection (Figure 3).

Several mutation detection techniques are suited to genome-scale application, including mismatch chemical cleavage (Cotton et al., 1988), single strand conformational analysis (Orita et al., 1989), and denaturation-based methods (reviewed in Fodde and Losekoot, 1994), but of these, conformation sensitive gel electrophoresis (CSGE) (Ganguly et al., 1993) is particularly well suited to high throughput analyses. Its simplicity makes it amenable to multiplexing and automated detection of variants, recent modifications have adapted it for use with capillary sequencers (Rozycka et al., 2000), and it is easily automated with standard robotics, further enhancing its applicability to genome-wide mutation scans. CSGE also is more sensitive to single base changes than techniques that rely on single strand conformational changes, and the speed of both the analysis and its interpretation provides significant advantages over sequencing.

There are two principal limitations to genome-wide CSGE analyses. First, genome-wide application of CSGE implies the availability of intron-based primers for every gene, a resource currently only available for a portion of the genome. In addition, the sequence polymorphisms that plague direct sequencing efforts also play havoc with CSGE if normal DNA from the same individual is not available for heteroduplex formation. Rapidly expanding polymorphism databases will aid in sorting disease-assorted mutations from rare polymorphisms, but this adds both uncertainty and an additional computational step. Nonetheless, for most investigators, few of whom will want to screen an entire genome for cancer-related mutations, CSGE is the easiest, fastest, most sensitive mutation detection technique available. Advances in mutation detection using microfabricated arrays, further improvements in sequencing, and possibly advances in mass spectrometry may eventually replace this approach, but are not widely available at present.

Expression profiling

While cumbersome for gene discovery, expression profiling is a powerful tool for describing cancer cells. It will identify therapeutic targets based on transcript level differences between normal and malignant tissue, and is very useful in developing molecular classification schemes that will ultimately drive therapeutic choices (reviewed in Lockhart and Winzeler, 2000). When used in experiments where cells are perturbed by a specific exposure, expression profiling will define cellular response, and in the setting of engineered mutations, it will

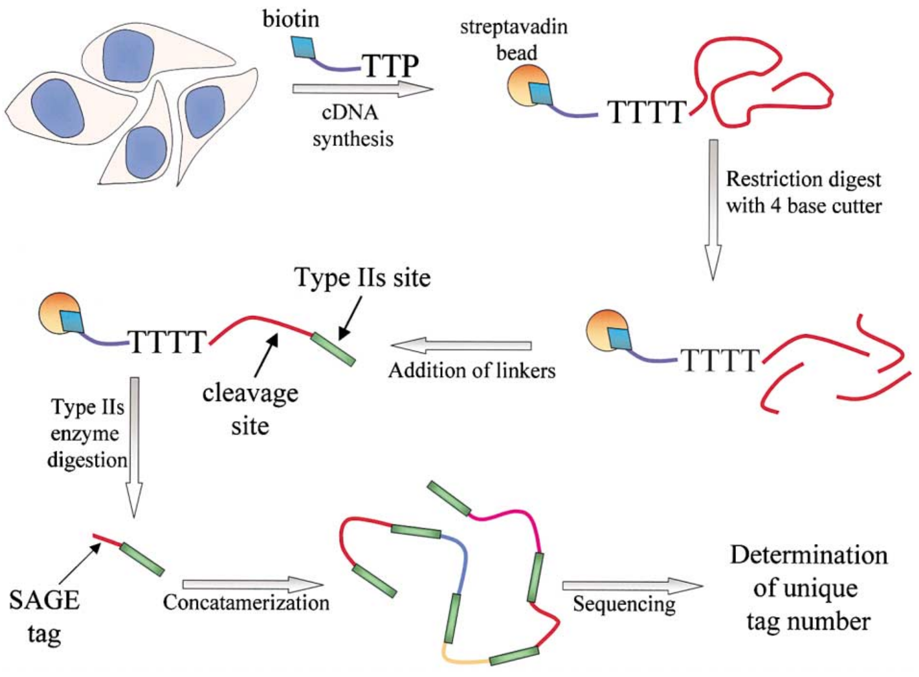

Figure 4. Preparation of a SAGE library

Data are generated from specific SAGE libraries as follows: (1) cDNA is prepared from tissue or cells using biotinylated dTTP; (2) cDNA is bound to streptavidin beads and cleaved into short fragments with a restriction enzyme recognizing a frequent four base pair sequence; (3) linkers with a type IIS restriction site are added, allowing cutting 15–20 bp from the IIS site and resulting in incorporation of 10–12 bp of 3′ cDNA into the linker fragment; (4) the fragments are concatemerized and cloned into a sequencing vector (allowing serial analysis of 30–35 transcripts per sequencing run); and (5) the SAGE library is sequenced, providing a comprehensive and quantitative picture of expressed genes without the need for any previous sequence data.

add to the development of contextual frameworks. Finally, differentially expressed transcripts identify genes silenced by methylation or other epigenetic mechanisms, as well as genes overexpressed in the absence of gene amplification. Both SAGE (*S*erial *A*nalysis of *G*ene *E*xpression) (Velculescu et al., 1995) and array-based expression profiling have been used to this end.

SAGE

SAGE (Figure 4) straddles the boundaries between gene discovery, mutation detection, and expression profiling (Velculescu et al., 1995), based on the principles that a 10 bp sequence tag will uniquely identify most human transcripts (assuming ~100,000 transcripts derived from ~35,000 unique genes) and that sequencing of SAGE libraries accurately reflects transcript number. Dismissed by some as prohibitively time-consuming, SAGE does have advantages. Library analysis can be automated and the analyses are amenable to high throughput approaches. In addition, SAGE libraries may be prepared from very small numbers of cells, facilitating application to premalignant tissue and small tumors. There is some indication that these “micro-SAGE” libraries are more representative than cRNA subjected to linear amplification for array analysis. (Ishii et al., 2000; Neilson et al., 2000) In a head to head comparison between SAGE and array-based expression profiling, there was excellent correlation in highly differentially expressed transcripts, and some evidence that SAGE is more sensitive in measuring low abundance transcripts and small differences in expression, both of which are likely be biologically significant. SAGE also has the advantage of being useful for gene discovery, a limitation of expression arrays that are comprised only of known genes or ESTs. This feature of SAGE is illustrated by a recent publication of the PRL-3 tyrosine phosphatase as a gene implicated in colon cancer metastasis (Saha et al., 2001). This observation was missed in other labs, not because of subtle expression differences but because PRL-3 is not represented on all commercially available expression arrays.

Once the genome is fully annotated, arrays are expanded to include all human genes, enhanced analytic techniques improve the ability of expression arrays to detect small differences, and databases to compare arrays between investigators are in place, it is unlikely that SAGE will be useful. Array-based transcriptional profiling is more amenable to large scale projects, considerably less time-consuming, and certainly more widely used. However, these advances will take several years, and for that time, SAGE remains a valuable and viable tool.

Expression arrays

Expression arrays are constructed of thousands of DNAs either spotted or synthesized onto glass slides (Lockhart et al., 1996; Schena et al., 1995). The DNAs may be collections of short oligos (e.g., 25mers), longer oligos (e.g., 60mers), or cDNAs of variable length. Spotted arrays employ fluorescent dye detection, commonly using a cohybridization strategy with Cy3-labeled cRNA from the test sample and a Cy5-labeled reference sample providing an intensity ratio, and thus a measure of relative expression. In contrast to array CGH, however, where the reference is genomic DNA, the optimal reference sample for expression profiling would contain detectable levels of every gene represented on the array within the linear range of the detection system, and thus does not exist for large arrays. In considering interpretation, relative expression differences of 2-fold or greater are generally considered significant, with lesser changes thought more likely to represent experiment variation.

However, while the use of reference samples and arbitrary significance thresholds for the interpretation of spotted arrays is a widely accepted experimental design, recent data suggest this design has multiple disadvantages and in fact may lead to erroneous conclusions (Jin et al., 2001). Perhaps most significantly, this design is inadequate for direct quantitative analysis of absolute transcript levels—essential if data from different experiments and labs are to be analyzed collectively. Jin et al. (2001) also provide convincing evidence that the common practice of setting arbitrary threshold ratios to assign significance has no basis in either biologic reality or in statistical theory. Without the use of a reference sample, which they point out provides no information of biologic interest, they evaluated expression differences that vary by age, sex, and strain in *D. melanogaster*. In these experiments, repetitive pairwise comparisons with age remaining the constant variable provided maximal power to detect age effects while still allowing for an analysis of the other two variables. Thus, two strains (Oregon [O] and Samarkand [S]), two ages (1 week and 6 weeks), and both sexes (M and F) were compared directly in pairs: OF1–OF6, OM1–OF6, SF1–SF6, and SM1–SM6. Each array was repeated six times, two of the six replicates with a dye swap. ANOVA, a variance analysis standard in statistical genetics, was employed for analysis. The findings are remarkable, and call into question much of the currently published array data. First, dye swapping had a marked effect, with multiple examples where opposite conclusions would be reached depending on the choice of dye (Cy3 or Cy5) for pair members. Second, the rigorous statistical analysis provided proof that arbitrary thresholds are not useful for assigning significance. After adjusting for both stochastic and biased variance, differences as low as 1.2 fold remained highly significant, and differences as great as 10-fold were dismissed as artifact. These data also highlight the fact that an arbitrary reference sample can bias a study toward unwarranted conclusions and limit power to detect changes in genes that are expressed at particularly high or low levels in the reference, as well as constrain data sharing between experiments that have not employed the same reference sample. These elegant analyses strongly suggest that we should rethink the widely used experimental design employing reference samples and arbitrary significance thresholds, replacing this design with one that allows definition of absolute units of expression, rather than ratios, and is most suited to the research question at hand. Thus, if looking for small effects, all pairs should be selected around the variable likely to have the least impact (age, in these experiments), as the power to see the effects of a specific variable is maximized if that variable remains constant in all pairwise comparisons. If the aim is to compare the effects of an exposure to a panel of different cells or tissues, then a looping or randomized pair design provides the best means of comparison.

Oligonucleotide arrays manufactured with twenty 25mers representing each gene on the array—10 perfect match sequences and 10 with a mismatch at the middle position—(Affymetrix, Inc) do have the advantage over spotted arrays of having been designed for use without a reference sample. The multiple oligomers provide some measure of replication, as target cRNA binds each independently, allowing an assessment of sensitivity and specificity; e.g., the composite signal generated from 9 perfect matches and 1 mismatch is weighted more heavily than a signal from 3 perfect matches and 6 mismatches. However, this design still does not correct for stochastic experimental variation nor bias in hybridization to specific probes, and thus does not address the problem of reliably detecting small differences between samples. The cost of these arrays constrains the application of multiple replicates, and leaves the validation of marked changes in individual genes to conventional approaches. In addition, genes that are not expressed in at least modest levels, and thus called “present,” cannot be evaluated at all in comparisons, and as with all arrays, incomplete genome annotation remains a significant limitation. Finally, the corporate policy of not providing investigators with the actual oligonucleotide sequences on the array completely eliminates the possibility of fully interpreting the primary data.

A potential technical improvement in the sensitivity and specificity of expression profiling (Hughes et al., 2001) may come from the use of single 60mers deposited using inkjet technology. This eliminates the need for masking in manufacture, reducing cost and greatly increasing flexibility. With careful selection of oligonucleotide sequence, this platform reportedly detects transcripts present at 0.1 copies per cell and discriminates any transcript in the genome that differs from another by five or more base pairs in the representative 60mer. The sensitivity, specificity, and flexibility of this system, particularly if used with the rigorous statistical analysis suggested by Jin el al. (2001), may make these arrays the best platform produced to date; however, they have yet to be made generally available so that they can be tested in academic labs.

Much has been made of the problems of data overload and the lack of hypothesis-driven research associated with expression profiling. However, neither of these criticisms accurately reflects either the power or the current limitations of this technology. Management of data derived from expression profiling is relatively straightforward—it is the lack of databases to aid in the generation of hypotheses and the extraction of biologically meaningful conclusions from the data that prove most problematic. Most investigators are experts in specific areas—cell cycle regulation, apoptosis, homeobox genes, and so on. How then can any one investigator wisely use data from experiments that provide information on virtually every cellular pathway? A badly needed resource is a database placing genes in the context of interconnected pathways, organized by relationships and providing links to relevant literature. Several databases for model organisms, such as Flybase, Wormbase, and Saccharaomyces Genome Database (SGD) incorporate sequence data, expression analyses, regulatory data, pathways, and phenotype. KEGG provides some mammalian pathway information, but no comprehensive database exists for either mice or humans. Links between these organism-specific databases, allowing comparative analysis, would add further to their value.

Another resource that is badly needed is a collection of expression profiles for public data mining. Steps toward implementing such a database have been taken by the Microarray Gene Expression Database group, who suggest minimal criteria for data in such a public repository (Brazma et al., 2001). However, we are far from the implementation of standardized terms format and content, means and agreements to deposit primary data, and even the ability to critically peer review global expression profiling data. It is incumbent on all investigators using this technique to contribute to these efforts to make their data available after publication (as is expected of sequence data) and in so doing, assist in creating a rigorous scientific standard for using and evaluating these data.

One example of the usefulness of large expression databases is illustrated in the use of expression profiling to assign function to uncharacterized open reading frames (ORFs) and to identify novel drug targets (Hughes et al., 2000). In this study, expression profiles from uncharacterized ORFs were compared to a (private) compendium of expression profiles from 300 known yeast mutants using a computational fingerprinting technique. Data from the mutant profiles first were clustered to define prominent expression patterns, identifying mutants with similar profiles and sets of coregulated genes. Comparing profiles from ORFs of unknown function to known mutants placed the uncharacterized genes in cellular pathways that directed subsequent biochemical experiments, none of which would have been assigned in a standard phenotype screen. These clusters were resilient to removing all transcripts encoding proteins with known function or their close paralogs, illustrating independence from knowledge of specific transcripts and a complete genome profile. In addition, none of the uncharacterized ORFs assigned into known pathways were up- or downregulated by more than 2-fold in any of the 300 compendium profiles, nor did profile assignments change when all transcripts with greater then 1.5-fold changes were masked, again highlighting the need for increased sensitivity from arrays.

Expression profiling is the most widely used genomic technology in current use. It has dramatically increased the amount of data recovered from single experiments and has initiated the process of molecular classification of tumors. Examples include breast cancers assigned to distinct categories based on presumed cell of origin within the breast (Perou et al., 2000), melanoma subtype suggested by expression of genes important in motility and invasion (Bittner et al., 2000), non-Hodgkin lymphoma clustered to more accurately reflect prognosis (Alizadeh et al., 2000), and the illustration that cancers that arise from a known event, such as a germline BRCA1 or BRCA2 mutation (Hedenfalk et al., 2001), can be separated from a more heterogeneous group of sporadic cancer-based expression profiles. However, all these clustering schemes employed relative small numbers of tumors, and at least in the case of lymphoma, marked expansion of the sample set yielded far too many tumor subsets to be clinically useful. The melanoma subtypes were derived from a set of 31 cancers with a uniformly bad prognosis, and the categories defined by the clusters did not correlate at all with extensively validated prognostic indicators such as depth. Are we correctly interpreting what these clusters are telling us? Can we exclude the possibility that these clusters are not a measure of factors such as normal cell admixture (clearly seen in the breast cancer profiling scheme) or sample handing differences? What are the sources of error and how might the clusters reflect this—particularly in light of the recent illustration of dye effect? How do we best eliminate predictors that solely reflect tissue of origin, and incorporate known clinical predictors to develop clustering schemes that are truly reflective of biological behavior and accurately reflect treatment response? While there is little doubt that molecular classification schemes that accurately predict tumor behavior will arise from expression profiling and other genomic platforms, it is clear that many more tumors will need to be analyzed, some aspects of the current clinical schemes will need to be incorporated, and modifications will continue as treatment response is added to the clustering as well.

Detection of epigenetic changes

Global methylation screens

Global methylation screens, still fraught with technical problems, are being developed to construct genome-wide profiles of CpG island methylation. Evidence that promoter methylation may be important in cancer comes from studies of methylation as a means of inactivating tumor suppressor genes, yet a tumor suppressor gene inactivated in cancer only by promoter silencing has yet to be confirmed, nor is there an example of a proto-oncogene activated only by promoter demethylation. The absence of such examples may be due to the limitations of a candidate gene approach, which even in the largest reported series—12 genes in more than 600 primary tumors (Esteller et al., 2001)—barely scratches the surface of what could be an average of 600 aberrantly methylated CpG islands in human cancer (Costello et al., 2000). It also could be that genes that cause cancer are rarely, if ever, altered only by methylation. If the latter is true, then we will not miss oncogenes and tumor suppressor genes, and thus important targets, in the absence of whole genome methylation data. However, if, as seems plausible, the means by which genes are altered can be exploited in developing therapeutic approaches, then these data may be extremely useful in developing drugs to reactivate silenced tumor suppressor genes.

Several global methylation assays are under development, including a high-density array of CpG islands (Yan et al., 2001). However, the approach requires a tedious, somewhat tricky sample preparation, and the current array of about 8,000 CpG islands is a significant underrepresentation of the estimated 45,000 that may be in the genome (admittedly not all of which are upstream regulatory elements of transcribed genes) (Costello et al., 2000). Restriction landmark genomic scanning (RLGS), a two-dimensional (2D) gel approach utilizing the methylation sensitivity of NotI sites for differential end-labeling of digested fragments prior to electrophoretic separation (Costello et al., 2000), shows promise as well. However, a limited number of CpG islands can be analyzed simultaneously with gel-based technology (1,184 in this study), and the image analysis of 2D gels with thousands of spots is time-consuming and subject to error. Possibly the most significant observation made to date using RLGS is that methylation profiles appear to be tissue type-specific, i.e., tumors can be clustered based on RLGS into groups that correlate with the tissue from which they arose. Whether it is a recapitulation of our current histological scheme or a more predictive analysis remains to be seen.

A number of other techniques, including methylation-sensitive arbitrarily primed PCR (Gonzalgo et al., 1997), methylated CpG island amplification (Toyota et al., 1999), and construction of a murine model with inducible DNA methyltransferase1 (DNMT1) (Jackson-Grusby et al., 2001) have been described to investigate global promoter methylation changes. The first two techniques detect promoter methylation directly, while the later is a clever example of the use of expression profile changes to detect promoters that have been silenced by methylation. In this model, profiles are generated before and after induction of DNMT1, with upregulated genes as candidates for having undergone demethylation, and thus tagged as methylated in the preinduction state. Breeding DNMT1 inducible mice to gene-specific tumor models and cataloging the genes methylated in those tumors using expression profiling after DNMT1 induction should yield interesting data on gene-specific methylation changes in cancer and simultaneously evaluate the potential for methylase inhibitors as cancer therapeutics. As no one technique yet has overcome the problems associated with determining global methylation profiles, this is a critical area for technical development and analysis.

Therapeutic successes with molecular targets

The use of cytotoxic agents has led to significant successes over the past fifty years in the treatment, and sometimes cure, of cancer. Childhood acute leukemias and nonseminomatous germ cell tumors were once uniformly fatal, but multidrug regimens result in cure rates that now approach 80% for some subtypes, including widely metastatic disease in the case of testes cancer. Some adult leukemias and lymphomas are curable, and combination chemotherapy for early breast cancer reduces recurrence rates, but we remain unable to uniformly prevent development of metastatic disease from most solid tumors in both children and adults. Even when curative, many treatments are associated with significant toxicity. In order to significantly improve on the current state of cancer treatment, it is essential that we develop therapies designed to specifically reverse the acquired capabilities that drive cancer cells. This approach, when sufficiently refined, not only should be effective in curing cancers, but should do so with limited toxicity. Several examples of cancer treatments based on molecular targets are already being used to treat patients, and they illustrate the principle that understanding the specific genetic defect that creates a molecular target is essential to the success of these agents.

The use of amplified proto-oncogenes as molecular targets is exemplified by herceptin, a blocking antibody directed at HER2/neu, a transmembrane tyrosine kinase growth factor receptor amplified in approximately 30% of breast cancers (reviewed in Harari and Yarden, 2000). While some toxicity occurs in normal cells with low levels of HER2/neu expression, the large differential in receptor number between normal and tumor cells gives a significant therapeutic index. Yet, why is herceptin only marginally effective in treating breast cancer? Presumably, high levels of HER2/neu expression are only part of the picture, and modulation of receptor expression further limits response. Both problems may be solved by applying genomic approaches to understand what other cellular pathways are disrupted in these cells, as well as elucidating the mechanisms that lead to herceptin resistance. The epidermal growth factor receptor (EGFR) is another molecular target commonly overexpressed in human tumors. ZD1839 (Iressa) is a quinazoline tyrosine kinase inhibitor selective for EGFR currently in clinical trials. Growth inhibition by ZD1839 occurs with dephosphorylation of EGFR, HER2, and HER3, dissociation of HER3 from PI(3)-kinase, and downregulation of Akt (Moasser et al., 2001).

Finally, the most recent example derived from molecular targeting is that of STI571, an Abl kinase inhibitor that targets the fusion protein formed by t(9; 22) (the Philadelphia chromosome) found in most cases of chronic myelogenous leukemia (CML) (Nowell and Hungerford, 1960). In Bcr-Abl, the coiled-coil domain of Bcr oligomerizes, leading to autophosphorylation and thus activation of the Abl kinase domain (McWhirter et al., 1993). In addition, the c-Abl DNA binding domain is replaced by Bcr sequences that mediate cytoplasmic sequestration, resulting in a form of Abl unable to induce apoptosis (Wang and Vigneri, 2001). Bcr-Abl thus has all the mitogenic activity of c-Abl, mediated through the Ras-Raf-ERK, JAK-STAT, and PI(3)kinase pathways, and none of the apoptotic activity of nuclear c-Abl, mediated through p73 (reviewed in Hunter and Blume-Jensen, 2001).

The success of treating CML with STI57 demonstrates the value of understanding the mechanism producing the cancer-causing mutation. STI 571 not only inhibits kinase-dependent growth signals but also induces apoptosis, because cytoplasmic retention of Bcr-Abl is partially dependent on the kinase activity of Abl (Druker et al., 1996; Horita et al., 2000). Thus, STI571 induces nuclear import of Bcr-Abl. This dual effect of STI571 raises the possibility of combining STI571 with a nuclear export inhibitor to enhance effectiveness and circumvent resistance (Wang and Vigneri, 2001). Also of significance in considering the future of targeting therapy is the dramatic efficacy of STI571 in treating tumors with related kinase mutations. Gastrointestinal stromal tumors (GIST), rare cancers with extremely rapid growth rates and a dismal prognosis, often have a mutation in c-kit, and respond dramatically to STI571 (Joensuu et al., 2001). These data are encouraging not only for patients with this rare disease, but also because they suggest that limited targets may need to be exploited, with activity of specifically targeted agents extending to other related molecules and thus active in multiple tumor types.

While drugs designed to specifically inhibit amplified, overexpressed, or activated proto-oncogenes are beginning to appear in the clinic, therapeutic agents that replace absent tumor suppressor gene activity are more troublesome, due to the complexities of functionally replacing an absent protein. Methylase inhibitors are being tested clinically, and show some promise as antitumor agents. Given that newly synthesized DNA must be actively methylated to maintain imprinting and tumor-specific methylation patterns, methylase inhibitors may be useful in reactivating tumor suppressor genes. However, methylation is thought to occur as a result of only a few enzymes, and gene-specificity of methylation in cancer is not well understood. Thus, these agents may not circumvent the problems of toxicity in the absence of a specific target and lack of tumor-specificity unless incorporated into a targeted drug delivery systems.

Conclusion

Innovations in microfabrication and capillary sequencing technology, coupled with a draft of the human genome sequence, have led to the development of several high throughput approaches to describe the genetic and epigenetic changes that contribute to cancer. While many cancers have yet to be even partially analyzed, and questions critical to designing successful therapeutics remain unanswered, it is now possible to envision a time when molecular phenotyping and targeted, individualized therapies that cure cancer will be a commonplace reality. Genomics is critical to these efforts, not only because it has exponentially increased data collection, but because it provides the opportunity not afforded by conventional approaches to see cancer from a global perspective, with specific alterations placed in context. Despite the current limitations, it is this feature of genomics that will enable a fundamental change in our understanding of cancer and thus our ability to cure patients.

References

Alizadeh, A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Bertwistle, D., and Ashworth, A. (1998). Functions of the BRCA1 and BRCA2 genes. Curr. Opin. Genet. Dev. *8*, 14–20.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406*, 536–540.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*, 365–371.

Clark, S.J., Harrison, J., Paul, C.L., and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. Nucleic Acids Res. *22*, 2990–2997.

Costello, J., Fruhwald, M.C., Smiraglia, D.J., Rush, L.J., Robertson, G.P., Gao, X., Wright, F.A., Feramisco, J.D., Peltomaki, P., Lang, J.C., et al. (2000). Aberrant CpG-island methylation has non-random and tumor-type-specific patterns. Nat. Genet. *24*, 132–138.

Cotton, R., Rodrigues, N.R., and Cambell, R.D. (1988). Reactivity of cytosine and thymine in single base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. Proc. Natl. Acad. Sci. USA *85*, 4397–4401.

Druker, B.J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G.M., Fanning, S., Zimmermann, J., and Lydon, N.B. (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl postive cells. Nat. Med. *2*, 561–566.

Esteller, M., Corn, P.G., Baylin, S., and Herman, J.G. (2001). A gene hypermethylation profile of human cancer. Cancer Res. *61*, 3225–3229.

Fero, M.L., Randel, E., Gurley, K.E., Roberts, J.M., and Kemp, C.J. (1998). The murine gene p27Kip1 is haplo-insufficient for tumour suppression. Nature *396*, 177–180.

Fischer, J., Palmedo, G., von Knobloch, R., Bugert, P., Prayer-Galetti, T., Pagano, F., and Kovacs, G. (1998). Duplication and overexpression of the mutant allele of the MET proto-oncogene in multiple hereditary papillary renal cell tumours. Oncogene *17*, 733–739.

Fodde, R., and Losekoot, M. (1994). Mutation detection by denaturing gradient gel electrophoresis (DGGE). Hum. Mutat. *3*, 83–94.

Friend, S., Bernards, R., Rogelj, S., Weinberg, R.A., Rapaport, J.M., Albert, D.M., and Dryja, T.P. (1986). A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature *323*, 643–646.

Futreal, A., Kasprzyk, A., Birney, E., Mullikin, J.C., Wooster, R., and Stratton, M.R. (2001). Cancer and Genomics. Nature *409*, 850–852.

Ganguly, A., Rock, M.J., and Prockop, D.J. (1993). Conformation-sensitive gel electrophoresis for rapid detection of single-base differences in double-stranded PCR products and DNA fragments: evidence for solvent-induced bends in DNA heteroduplexes. Proc. Natl. Acad. Sci. USA *90*, 10325–10329.

Ge, H., Liu, Z., Church, G.M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharyomyces cerevisiae.* Nat. Genet. *29*, 482–486.

Gonzalgo, M., Liang, G., Spruck, C.H., Zingg, J.M., Rideout, W.M., and Jones, P.A. (1997). Identification and characterization of differentially-methylated regions of genomic DNA by methylation-sensitive arbitarily-primed PCR. Cancer Res. *57*, 594–599.

Hahn, S., Schutte, M., Hoque, A.T., Moskaluk, C.A., da Costa, L.T., Rozenblum, E., Weinstein, C.L., Fischer, A., Yeo, C.J., Hruban, R.H., and Kern, S.E. (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. Science *271*, 350–353.

Hahn, W.C., Counter, C.M., Lundberg, A.S., Beijersbergen, R.L., Brooks, M.W., and Weinberg, R.A. (1999). Creation of human tumour cells with defined genetic elements. Nature *400*, 464–468.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Harari, D., and Yarden, Y. (2000). Molecular mechanisms underlying Erb/HER2 action in breast cancer. Oncogene *19*, 6102–6114.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., et al. (2001). Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. *344*, 539–548.

Hodgson, G., Hager, J.H., Volik, S., Hariono, S., Wernick, M., Moore, D., Albertson, D.G., Pinkel, D., Collins, C., Hanahan, D., and Gray, J.W. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinoms. Nat. Genet. *29*, 459–464.

Horita, M., Andreu, E.J., Benito, A., Arbona, C., Sanz, C., Benet, I., Prosper, F., and Fernandez-Luna, J.L. (2000). Blockade of the Bcr-Abl kinase activity induces apoptosis of chronic myelogenous leukemia cells by suppressing signal transducer and activator of transcription 5-dependent expression of Bcl-xl. J. Exp. Med. *191*, 977–984.

Hughes, T., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109–126.

Hughes, T., Mao, M., Jones, A., Burchard, J., Marton, M., Shannon, K., Lefkowitz, S., Ziman, M., Schelter, J., Meyer, M., et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. *19*, 342–347.

Hunter, T., and Blume-Jensen, P. (2001). Oncogenic kinase signaling. Nature *411*, 355–365.

Ishii, M., Hashimoto, S., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T., and Aburatani, H. (2000). Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profile analysis. Genomics *68*, 136–143.

Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., Dausman, J., Lee, P., Wilson, C., Lander, E., and Jaenisch, R. (2001). Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nat. Genet. *27*, 31–39.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.D., Passador-Gurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster.* Nat. Genet. *29*, 389–395.

Joensuu, H., Roberts, P.J., Sarlomo-Rikala, M., Andersson, L.C., Tervahartiala, P., Tuveson, D., Silberman, S., Capdeville, R., Dimitrijevic, S., Druker, B., and Demetri, G.D. (2001). Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. N. Engl. J. Med. *44*, 1052–1056.

Kamb, A., Gruis, N.A., Weaver-Feldhaus, J., Liu, Q., Harshman, K., Tavtigian, S.V., Stockert, E., Day, R.S., Johnson, B.E., and Skolnick, M.H. (1994). A cell cycle regulator potentially involved in genesis of many tumor types. Science *264*, 436–440.

Kinzler, K.W., and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. Cell *87*, 159–170.

Knudsen, A. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. USA *68*, 820–823.

Kwabi-addo, B., Giri, D., Schmidt, K., Podsypanina, K., Parsons, R., Greenberg, N., and Ittmann, M. (2001). Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. Proc. Natl. Acad. Sci. USA *98*, 11563–11568.

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Miliaresis, C., Rodgers, L., McCombie, R., et al. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science *275*, 1943–1947.

Lockhart, D., and Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. Nature *405*, 827–836.

Lockhart, D., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. *14*, 1675–1680.

McWhirter, J., Galasso, D.L., and Wang, J.Y. (1993). A coiled-coil oligomerization domain of bcr is essential for the transforming function of the Bcr-Abl oncoproteins. Mol. Cell. Biol. *13*, 7587–7595.

Moasser, M., Basso, A., Averbuch, S.D., and Rosen, N. (2001). The tyrosine kinase inhibitor ZD1839 (“Iressa”) inhibits HER2-driven signaling and suppresses the growth of HER2-overexpressing tumor cells. Cancer Res. *61*, 7184–7188.

Neilson, L., Andalibi, A., Kang, D., Coutifaris, C., Strauss, J.F., Stanton, J.A., and Green, D.P. (2000). Molecular phenotype of the human oocyte by PCR-SAGE. Genomics *63*, 13–24.

Nowell, P., and Hungerford, D.A. (1960). A minute chromosome in human granulocytic leukemia. Science *132*, 1497.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., and Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc. Natl. Acad. Sci. USA *86*, 2766–2770.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. *20*, 207–211.

Pollock, J., Westervelt, P., Kurichety, A.K., Pelicci, P.G., Grisolano, J.L., and Ley, T.J. (1999). A bcr-3 isoform of RAR-alpha-PML potentiates the development of PML-RAR-alpha driven acute promyelocytic leukemia protein (PML) in tumor suppression. J. Exp. Med. *193*, 521–529.

Rozycka, M., Collins, N., Stratton, M.R., and Wooster, R. (2000). Rapid detection of DNA sequence variants by conformation-sensitive capillary electrophoresis. Genomics *70*, 34–40.

Saha, S., Bardelli, A., Buckhaults, P., Velculescu, V.E., Rago, C., Croix, B.S., Romans, K.E., Choti, M.A., Lengauer, C., Kinzler, K.W., and Vogelstein, B. (2001). A phosphatase associated with metastasis of colorectal cancer. Science *294*, 1343–1346.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary cDNA array. Science *270*, 467–470.

Song, W., Sullivan, M.G., Legare, R.D., Hutchings, S., Tan, X., Kufrin, D., Ratajczak, J., Resende, I.C., Haworth, C., Hock, R., et al. (1999). Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. Nat. Genet. *23*, 166–175.

Toyota, M., Ho, C., Ahuja, N., Jair, K.W., Li, Q., Ohe-Toyota, M., Baylin, S.B., and Issa, J.P. (1999). Identification of differentially-methylated sequences in colorectal cancer by methylated CpG island amplification. Cancer Res. *59*, 2307–2312.

Velculescu, V., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484–487.

Versteege, I., Sevent, N., Lange, J., Rousseau-Merck, M.F., Ambros, P., Handgretinger, R., Aurias, A., and Delattre, O. (1998). Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. Nature *394*, 203–206.

Wang, J.Y., and Vigneri, P. (2001). Induction of apoptosis in chronic myelogenous leukemia cells through entrapment of Bcr-Abl tyrosine kinase. Nat. Med. *7*, 228–234.

Xu, X., Brodie, S.G., Yang, X., Im, Y.H., Parks, W.T., Chen, L., Zhou, Y.X., Weinstein, M., Kim, S.J., and Deng, C.X. (2000). Haploid loss of the tumor suppressor Smad4/Dpc4 initiates gastric polyposis and cancer in mice. Oncogene *19*, 1868–1874.

Yan, P., Chen, C.M., Shi, H., Rahmatpanah, F., Wei, S.H., Caldwell, C.W., and Huang, T.H. (2001). Dissecting complex epigenetic alterations in beast cancer using CpG island microarrays. Cancer Res. *61*, 8375–8380.