High-Content Analysis of Cancer Genome DNA Alterations

Yan Y. Degenhardt1, Richard Wooster1, Richard W. McCombie2, Robert Lucito2, and Scott Powers2,3

1 Translational Medicine, GlaxoSmithKline, King of Prussia, PA, 19406, USA
2 Cancer Genome Center, Cold Spring Harbor Laboratory, Woodbury, NY, 11797, USA

Summary

New technologies as well as concerted brute-force approaches have increased the content (number of genes) that can be characterized for genomic DNA alterations. Recent advances include the detection of activating point mutations in key kinase genes (BRAF, EGFR, PIK3CA) in multiple cancer types; preliminary insight into the entire repertoire of genes that can be mutated in cancer; the discovery of new oncogenes by high-resolution profiling of DNA copy number alterations; and the bioinformatic-driven discovery of oncogenic gene fusions. High-content promoter methylation detection systems have been used to discover additional methylated genes and have provided evidence for a stem cell origin for certain tumors. Some of these advances have had significant impact on the development and clinical testing of new therapeutics.

Introduction

There is considerable interest in being able to detect, catalog, and understand the entire set of DNA alterations in human cancers. The genetic model of cancer places primary importance on these changes, which result from recurring cycles of genomic alteration followed by clonal selection, and which predicts that the properties of a particular individual’s tumor are dictated by the specific DNA alterations within that tumor (the cancer genotype). Support for this model comes from experiments and clinical results that demonstrate that reversing the effect of even a single DNA alteration - such as inhibiting the aberrant tyrosine kinase activity of the BCR-ABL gene fusion protein by the small-molecule inhibitor imatinib - can have dramatic effects on the ability of the tumor cell to proliferate or even survive [1]. However the cancer genotype is not the only determinant, with the intrinsic properties of the tumor-initiating cell and the tumor microenvironment also playing key roles in cancer progression [2,3].

The cancer genotype includes both hereditary predisposition genes and the genes that are somatically altered by a variety of mechanisms including mutation, amplification, deletion, and translocation. Epigenetic alterations caused by DNA methylation or modifications of histones can also comprise a substantial portion of the final set of cancer-relevant DNA alterations.

In the earlier part of this decade, the only genome-wide technology for profiling the cancer genome was comparative genomic hybridization (CGH) (Box 1). Now there are several means...
to study DNA alterations on a genome-wide scale, including array CGH with very high genome resolution [4], new methylation detection systems [5], and last but not least, single-molecule sequencing technologies [6] (Box 1). In this review we focus on the recent biological discoveries that have been made with these new tools along with the large-scale efforts using older tools.

Mutational Profiling (Resequecing)

Large-scale Sanger dideoxy sequencing of targeted regions by the Sanger Institute and the Johns Hopkins group has led to two basic discoveries that both have direct translational relevance. The first was the discovery that \textit{BRAF} is frequently (40%) activated by point mutations in melanomas (Boxes 1,2) [7]. This is the second cancer characterized by frequent genomic activation of a kinase, with chronic myeloid leukemia (CML) and the \textit{BCR-ABL} translocation being the first. The success of kinase inhibitor therapy in CML has created intense interest in testing Braf pathway inhibitors in melanoma patients. The second discovery was that the \textit{PIK3CA} gene is frequently mutated in colorectal and other cancers (Box 2) [8]. If one includes mutations in all the pathway genes (\textit{PTEN}, \textit{PDK1}, \textit{AKT2} and \textit{PAK4}), the phosphatidylinositol–3-OH kinase pathway is mutationally activated in 40% of colorectal cancers [9]. These results have stimulated development and clinical testing of inhibitors of this pathway.

Independently of these large-scale efforts, several laboratories discovered that mutations in \textit{EGFR} occur frequently in lung cancer and predict a positive clinical response to EGFR inhibitors (Box 2) [10–12]. The frequency ranges from 10% in Caucasians to 40% in East Asian lung cancer patients [13]. Other \textit{EGFR} pathway genes are mutated in lung cancer including \textit{KRAS} (20%), and \textit{ERBB2}, \textit{BRAF}, and \textit{PIK3CA} (each 2–4%) [14]. Pao and co-workers discovered that lung cancers that harbor mutations in \textit{KRAS}, which encodes a downstream signaling component in the EGFR pathway, were resistant to treatment with EGFR inhibitors [15]. Another clear example of a genomic predictor of a negative response to a targeted therapeutic has recently been reported: oncogenic \textit{PIK3CA} mutations or low \textit{PTEN} expression in breast tumors are negative predictors for response to trastuzumab, the anti-Her2 monoclonal antibody used to treat breast cancer patients who harbor amplification of the \textit{HER2/ERBB2} gene [16].

The first genome-wide resequencing effort, which initially involved the coding regions of 14,000 genes [17] and was later expanded to include all of the reference sequence genes [18], has triggered controversy. The major issue is the claim that the Johns Hopkins group has discovered approximately 200 new cancer genes. Other groups claim that their statistical analysis was flawed and that in particular the small sample size (~20 tumors) is too small to distinguish driver mutations from passenger mutations [19].

However, the Sanger Institute independently came to a similar conclusion -that the repertoire of cancer genes that are altered by mutations is substantially larger than previously thought (Box 2) - after sequencing a much larger set of tumors (~ 200) for 518 protein kinase genes [20]. The authors addressed biological significance in their study by employing statistical analyses that are used in genome evolution studies, the premise of which is that the higher the ratio of non-synonymous to synonymous mutations, the more likely there has been biological selection. The number of kinase genes that can be mutated into cancer driver genes was estimated to be within 52 to 149, which is two to five times larger than the set of known cancer-causing kinase genes (~ 25) [21]. Many of these mutations occurred in the same functionally important protein regions that are mutated in \textit{BRAF}, providing additional evidence that they contribute to the cancer phenotype [20].
It is noteworthy that none of the 20,000 genes examined by the Johns Hopkins group are mutated at frequencies greater than 10% in colorectal or breast cancer besides those previously identified (TP53, PIK3CA, APC, KRAS, SMAD4, and CDC4) [17,18]. Assuming for the moment that there really is significant diversity in the set of genes that can be mutated to become driver genes in cancer, how will we move forward to develop target therapeutics? One possibility is to use pathway analysis to look for commonality. Along these lines, the set of genes rarely mutated in breast cancer are significantly enriched for components of the NF-kB pathway [18]. Thus NF-kB pathway inhibitors might be useful to treat a subclass of breast cancer patients identified as having mutations in NF-kB pathway genes, or, as we will see in the section that follows, amplification of the IKBKE gene.

Copy Number Alteration Profiling

Array-based comparative genomic hybridization (aCGH) studies now number over 500 but for the most part have not addressed functional relevance of the altered genes. Two approaches have emerged as being particularly relevant to cancer biology. One has been to focus on DNA copy number alterations that are tumor-type specific. MITF was identified and subsequently validated as a melanoma-specific amplified oncogene based on its selective amplification in melanoma [22]. Similarly, three genes (TTF1, NKX2-8, and PAX9) that are within a 150 kb region on 14q13 were found to be determinants of a lung-cancer specific amplicon [23]. All three of these genes act synergistically to promote and maintain tumorigenicity [23]. Occasionally, the TTF1 gene is amplified in the absence of NKX2-9 and PAX9 [24]. Both of these amplicons are relatively frequent oncogenic events: MITF is amplified in 20% of metastatic melanoma [22], and the TTF1/NKX2-8/PAX9 amplicon is the second most common amplicon in lung cancer (following MYC) and is amplified in up to 20% of lung cancers [23, 24] (Box 2).

Another productive approach that has emerged is to compare human DNA copy number alterations to those that occur during cancer progression in mouse models. The 11q22 amplicon, which is found in several tumor types including liver cancer, was found to be syntenic with an amplicon located at 9qA1 in murine model of liver cancer [25]. Two genes, cIAP1 and Yap, were overexpressed in amplified tumors from both species and acted cooperatively to accelerate and maintain tumorigenesis [25]. Similarly, the Nedd9 gene was found within a small amplicon in metastatic mouse melanomas and is amplified in human melanomas as part of a large gain of the p arm of chromosome 6 [26]. Nedd9 encodes a signaling protein that enhances invasion and metastasis when overexpressed in melanocytes [26]. In the future, introduction of genomic instability into mouse models will enable increased utility of this cross-species approach [27].

Recently, RAB25 on 1q22 and IKBKE on 1q32 were shown to be functional oncogenes in ovarian and breast cancers by both overexpression and RNAi studies [28,29]. Presumably, both RAB25 and IKBKE, the latter encoding an IkappaB kinase that activates the NF-kB pathway, are likely driver genes for the gain of the q arm of chromosome 1, which is one of the most frequent genetic events in human cancer – occurring in 50% of breast, liver, lung, ovarian and other tumor types. As of today, there is no systematic approach to identifying the driver genes for such large chromosomal gains or losses. This is a major black box in our understanding of the oncogenic events that underlie cancer progression.

In an innovative approach to finding the genetic underpinnings of a molecularly defined subclass of tumors, Chang and co-workers determined that the co-amplification and co-overexpression of MYC and a subunit of the COP9 signalosome (COP5) underlies the poor-prognosis wound signature seen in a subset of primary breast cancers [30]. Co-expression of
both MYC and COPS5 was necessary and sufficient to induce the wound repair signature in a normal mammary epithelial cell line.

Copy number profiling has also been used to look for molecular subclasses of common cancers. Profiling can distinguish three molecular subclasses of gliomas, one of which corresponds to a known clinical subtype but the other two are novel [31]. In breast cancer, where gene expression profiling has defined new molecular subclasses, copy number profiling has shown that within different gene-expression subclasses, tumors harboring amplicons show poorer prognosis than those that do not [32]. Multiple closely spaced amplicons on the same chromosome are highly correlated with aggressive disease and poor survival [33].

### Profiling Translocations and Gene Fusions

Gene fusions play key roles in the initial steps of hematological cancers and childhood sarcomas [34]. Over 300 genes are affected by gene fusion events in cancer and the majority of these were identified in hematological cancers. Although there are particular hematological cancers in which almost 100% of the tumors will have a specific gene fusion, e.g. BCR-ABL in CML or PML-RARA in acute promyelocytic leukemia, the frequency and patterns of gene fusions in hematological cancers is better typified by acute myeloid leukemia (AML). Only 20% of AML cases contain translocations, and these translocations are very diverse and constitute 109 different gene fusions [34]. Thus it is important keep these extremes in mind when considering the potential of discovering gene fusions in more common epithelial tumors.

In a truly amazing line of investigation, Chinnaiyan and colleagues analyzed the expression patterns of gene fusions in hematological cancers and determined an algorithm for finding candidate gene fusions using RNA expression data [35]. Among the top 10 candidate prostate cancer fusion genes were ERG and ETV1 —transcription factor genes known to be involved in fusions in other tumor types. They found that these genes were fused to 5′ promoter of the highly androgen-responsive gene TMPRSS2. Up to 80% of primary prostate cancers harbor TMPRSS2 fused to ERG, ETV1, or ETV4 (Box 2). These fusions are absent in normal prostate, proliferative inflammatory prostate, and benign prostate hyperplasia; but are detected in 20% of early neoplastic prostate lesions [36].

In addition, there was a recent serendipitous discovery of a gene fusion involving a tyrosine kinase gene (ALK) affecting 7% of non-small cell lung cancers (7% in an Asian population – this frequency is likely to be different in other races). This fusion, EML4-ALK, was detected in a cDNA library from a lung cancer following a transformation screen in NIH-3T3 cells [37]. This and the prostate cancer results raise the question as to why these gene fusions have not been detected beforehand? The answer appears to be the much greater technical and analytic problems associated with cytogenetic analysis of epithelial tumors compared to that of hematological cancers [34]. Thus there may be many more epithelial cancer gene fusions that remain to be discovered, either by informatic approaches or by experimental approaches such as high-throughput sequencing of both 5′ and 3′ ends of transcripts.

### Epigenetic Profiling

The importance of epigenetic alterations in cancer progression was shown years ago when methylation of the 5′ CpG-island of the p16/CDK2NA gene was proven to be responsible for its transcriptional silencing in 20%–40% of most common cancers [38,39]. Several cancer susceptibility genes, including BRCA1 and VHL, which cause familial forms of breast and kidney cancer, respectively, are silenced by methylation in a significant percentage of sporadic forms of the respective tumor types. 15% of sporadic breast cancers harbor methylated BRCA1 genes and their gene-expression profiles are identical to those of tumors from inherited
families where BRCA1 is mutated; both are completely distinct from those of other breast-cancer types [40].

Several groups have been developing array-based methods for genome-wide detection of methylation or other epigenetic alterations such as histone modifications [5,41,42]. These emerging methods have not yet yielded the depth of biological insight that methods that are lower in gene content, but technologically more robust, have yielded. Measurement of ~200 CpG islands in a panel of ~300 colorectal cancers with a PCR-based technology has provided convincing evidence that a subset of colorectal cancers – approximately 15%–have significantly greater CpG island methylation than other colon cancers, and that this subset is characterized by a methylated mismatch repair gene MLH1 and the sporadic form of microsatellite instability. Thus it appears that a methylator type of genomic instability proceeds the development of microsatellite instability (Box 2) [43]. This same platform was used to determine that stem cell Polycomb group targets are greater than 10-fold more likely to have cancer-specific promoter DNA hypermethylation than non-targets, supporting a stem cell origin of cancer in which reversible gene repression is replaced by permanent silencing [44].

**Conclusion**

Within the next five years it is likely that we will be able to detect and catalog the entire set of DNA alterations in a given cancer, which will be a significant milestone in the history of cancer research. New sequencing technologies will make it possible to sequence hundreds of cancer genomes, and will also enable ultimate resolution and genome coverage for detecting RNA expression, copy number and epigenetic alterations, as well as gene fusions. What will clearly lag behind is an understanding of the functional importance of all the DNA alterations, which will take several additional decades. In the meantime, cancer genome profiling will have important applications in pinpointing new targets, discovering resistance mutations to existing therapies, and discovering both positive and negative genomic predictors of response to specific therapies.

**Box 1. Milestones in Cancer Genome DNA Profiling**

<table>
<thead>
<tr>
<th>Year</th>
<th>Milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998 – 1999</td>
<td>Development of 1st generation arrays for profiling DNA copy number alterations</td>
</tr>
<tr>
<td>2001</td>
<td>Development of 1st generation arrays for profiling epigenetic alterations</td>
</tr>
<tr>
<td>2002</td>
<td>First major finding reported from large-scale resequencing projects</td>
</tr>
<tr>
<td>2004</td>
<td>Development of 1st generation single-molecule sequencers</td>
</tr>
<tr>
<td>2005</td>
<td>Bioinformatic approach developed that enables discovery of gene fusions</td>
</tr>
</tbody>
</table>

**Box 2. Major Discoveries Resulting from Cancer Genome DNA Profiling**

- Frequent mutational activation of EGFR, PIK3CA, and BRAF [7,8,11]
- Frequent translocations and interstitial deletions causing gene fusions in prostate cancer [35]
- Strong evidence for a methylator phenotype in colorectal cancer [43]
- Discovery of frequent tumor-type specific amplicons in melanoma and lung [22, 23,24]
Substantial diversity in the set of genes affected by putative driver mutations [17,20]

References


The results of the first cancer genome resequencing project that surveyed (most) all known genes. Although controversial, the finding that there are infrequent cancer-relevant mutations in a much larger set of genes than previously imagined is consistent with another report highlighted below.


The results of a major cancer genome resequencing project that surveyed all kinase genes in a relatively large sample set. Viewed together with the report highlighted above, this argues that there is a larger set of genes than previously imagined that have oncogenic potential in human cancer.


Perhaps the most dramatic demonstration of cooperating oncogenes within a single amplicon (the 14q13 amplicon in lung cancer), this paper highlights the increasing importance of performing functional analysis of cancer-relevant genes in parallel rather than individually.


The most comprehensive and extensive copy number alteration survey of human cancers to-date, this paper also describes the 14q13 amplicon present in a large percentage of lung cancers.


44•. Widschwendter M, Fieg F, Egle D, Mueller-Holzner E, Spizzo G, Markl C, Weisenberger DJ, Campan M, Young J, Jacobs I, et al. Epigenetic stem cell signature in cancer. Nat Genet 2007;9:157–158. [PubMed: 17200673] This study shows that Genes that are reversibly repressed in embryonic stem cells by Polycomb proteins are over-represented among genes that are permanently silenced in colon cancer; this link lends support to the theory that the tumor initiating cells of colon cancer are stem cells