The Genome Doctor, a Whole-Genome Sequencing Approach to Cancer

The enthusiasm of the clinical cancer research community has been renewed by clinically relevant demonstrations of molecular profiling arising from “whole genome sequencing” (WGS), leading to novel approaches of “personalised” cancer treatments. However, critics question the merits of such large-scale, expensive, time-consuming, data-harvesting, and descriptive approaches, citing the paucity of improvement in outcomes. We address here the challenges and potentials in the clinical application of whole-genome sequencing for cancer.

Why a whole genome sequencing approach to cancer?

Cancers ultimately result from the pathological impact of aberrations in DNA sequence. Human anomalous DNA sequences derive from exposure to mutagens, virus, inflammation, and even spontaneous mutation, and perpetuates through replication, resulting in the accumulation of tumour promoting alterations in the DNA sequence. An unbiased genome-wide mutation detection approach to cancer can identify somatically acquired sequence variants and functional mutations and thereby identify critical genes uniquely relevant to the development of a specific patient’s cancer. This strategy will ultimately provide the paradigm for the detection of germline mutations in non-neoplastic human genetic diseases, through genome-wide mutation detection approaches, which may fundamentally revolutionise cancer prevention and treatment through personal medicine.

Technology advancements make whole genome sequencing affordable

The Human Genome Project (HGP) applied a highly labour intensive process called Sanger sequencing to identify the original DNA sequences. The HGP launched in 1990 and formally completed the first draft sequence of an entire human genome in 2003[1,3]. The cost of HGP was about 5 billion dollars [4,5]. Since 2005, however, a new technology, called ‘next-generation sequencing’ or ‘massively parallel sequencing,’ has emerged and replaced Sanger sequencing [6]. This cut down not only the cost, but the time to complete the genome sequencing. For example, Complete Genomics Incorporation has developed a genome sequencing platform that achieves high accuracy and significantly reduced cost (currently > $10,000) with the ability to detect rare variants within 10 days [7]. Competition has motivated several companies to predict that within the very near future, delivery of a complete WGS project can be completed within 24 h with a cost of $1,000! [8,9]. These developments have prompted scientists to conceive the concept of using whole-genome sequencing in everyday clinical decision-making for personalized medicine [10].

In the USA, NIH has funded three institutions to apply WGS to studying human diseases – Baylor College of Medicine, the Broad Institute of Harvard and MIT, and the Washington University. In the United Kingdom, the Sanger Centre has decided to sequence the whole genome to study certain highly important diseases [11-15]. In Germany, Korbel and colleagues have published paediatric medulloblastoma genome data [16]. All of these sequencing centres have included a focus on cancer, applying the whole genome approach. We will discuss several recent publications to address the challenges and potentials for WGS for cancer.

The Washington University approach to acute myeloid leukaemia

Washington University researchers have published a series using the WGS approach to describe the acute myeloid leukaemia (AML) [17-22]. They reported a new mutation in a gene called DNM1 that may help identify patients with high risk of recurrent AML. They successfully used WGS to tailor a personalised treatment plan for a woman with a rare subtype of AML that responded well to a specific targeted therapy, sparing her from more aggressive stem-cell transplantation [23]. Her disease subtype could only be definitively identified through whole-genome sequencing.

The Johns Hopkins University approach to breast and colorectal cancers

At the Johns Hopkins University, Vogelstein and colleagues have published the consensus coding sequences of human breast and colorectal cancers from 11 breast and 11 colorectal cancer patients [24]. This group expanded the study by including analysis of the sequences of 20,857 transcripts from 18,191 human genes, including the great majority of encoding proteins. Any gene in the tumour that was mutated, but not in normal tissue from the same patient, was analysed in 24 other tumours. Selected genes were further analysed in 96 colorectal cancers to improve the definition of their mutation frequency and aid subsequent bioinformatic analyses [25]. They also found a new pathway for pancreatic neoplasia by identifying recurrent mutations at codon 201 of GNAS [26]. The data was derived from analyses of 113 patients with intraductal papillary mucinous neoplasm. KRAS and GNAS mutations can be used in the diagnosis and prognosis of patients with cystic pancreatic lesions.

Clinical Potential of WGS in Cancer Therapy

Vogelstein and colleagues have defined the first genomic landscape of two of the most common human cancers, breast and colon [24]. The majority of their identified 189 cancer-associated genes were not known to be mutated in tumours, prior to these results it was thought that the average tumour harboured 90 mutated genes. Unexpectedly, no gene was consistently mutated in either breast or colorectal tumours. Thus the number of mutations occurring during the evolution of human tumours is greater than previously estimated. The level of complexity suggests a two-step approach to cancer genome: first, a discovery screen, and second, a validation screen.
Discovered in the past, most cancer-causing mutations have been discovered by direct analysis, often providing evidence that specific genes control cell division. Current research indicates that these early efforts barely glimpsed the big picture. The whole genome approach has identified not only known colorectal tumour genes (TP53, APC, KRAS, SMAD4, FBXW7, EPHA3, SMAD2, and TGFBRII), but concomitantly uncovered their pathogenesis of other forms of cancer (GNAS, NF1, and RET) [25]. The approach also indicated other inherited cancer syndromes (familial adenomatous polyposis, neurofibromatosis, Li-Fraumeni syndrome, juvenile polyposis and multiple endocrine neoplasia).

**Validation Screen**

Vogelstein and his colleagues analyzed 13,023 genes to define the best studied genes in the human cancer genome. His 29-member team resequenced the protein-coding regions of the genes in 11 previously well-studied genes (EGFR, KRAS, TP53, SMAD4, and FBXW7 in colorectal cancer). The approach also implicated other inherited cancer genes (CAN genes; driver mutation genes) versus passenger mutation genes [34]. They argued that driver mutation genes were present at higher rates than the other known colorectal tumour genes (TP53 in breast cancer, APC, KRAS, TP53, SMAD4, and FBXW7 in colorectal cancer). Thus, they concluded that this genome approach did not implicate any additional cancer genes beyond the handful known from previous studies.

**Clinical relevance of the genomic approach**

Verhaak and colleagues [30] have published a framework for integrated genomic analysis that identifies clinically relevant subtypes of glioblastoma characterize by abnormalities in PDGFRα, IDH1, EGFR, and NF1. They catalogued recurrent genomic abnormalities in glioblastoma multiforme (GBM). Their gene expression-based molecular classification of GBM led to the construction of four subtypes – proneural, neural, classical and Mesenchymal, based on integrated multidimensional genomic data (patterns of somatic mutations and DNA copy number). Aberrations and gene expression of biomarkers (EGFR, NF1, and PDGFRα/IDH1) define the classical, mesenchymal, and proneural subtypes, respectively. Gene signatures of normal brain cell types show a strong relationship between subtypes and different neural lineages. Additionally, the response to conventional aggressive clinical therapy differs by subtype, with the greatest benefit in the Classical subtype and no benefit in the proneural subtype.

**Challenges**

Currently 90% of drugs fail in cancer patients. The relatively small number of new genes common to tumours reinforces concerns about the cancer genome approach. In the Johns Hopkins study, despite previously unknown mutated genes being discovered, the functional consequences of most of these and their actual role in tumourigenesis are unknown; even with that knowledge, we remain a long way from identifying new therapeutic targets.

**Large-scale efforts**

Vogelstein and his colleagues analyzed 13,023 genes to define the best studied genes in the human cancer genome. His 29-member team resequenced the protein-coding regions of the genes in 11 breast cancer samples and 11 colon cancer samples and found about 800,000 possible mutations. The team then redefined functional mutations in 189 cancer-associated genes by removing errors, normal variants, and changes that did not alter a protein.

**Heterogeneity**

All these studies show that cancer mutations exhibit daunting complexity and heterogeneity. A major issue in looking at the whole genome is that so many of the mutations known to date are not common to patients with a single type of cancer. Furthermore, many lie in areas of the genome where their unknown function and/or significance is unknown. Vogelstein’s team found that the average breast or colon tumour has 93 mutated genes, with at least 11 thought to be cancer-promoting. Their work yielded a total of 189 candidate cancer genes. The cancer genes differ between colon and breast cancers, suggesting that more steps or pathways leading to tumourigenesis of cancer than previously conjectured.

**Complex statistical analysis**

While Sjöblom et al. [31,32] reported 189 genes with an apparent statistically significant excess of mutations in breast and colorectal tumours, Rubin and Green [32] used a different set of criteria for statistical analysis for the same cancer genomic data and identified only a handful of previously known cancer genes (TP53 in breast cancer, APC, KRAS, TP53, SMAD4, and FBXW7 in colorectal cancer). Thus, they concluded that this genome approach did not implicate any additional cancer genes beyond the handful known from previous studies.

Eric Lander’s group at MIT/Harvard, after correcting the statistical analysis and using a background mutation rate that better fit their data, reported that these 189 novel cancer genes did not reach a 90% probability of being relevant to colon or breast cancer [33]. Forest and Cavet [34,35] argued against point probabilities algorithm that was used in the original paper [24] and they suggested using a value-based algorithm to reanalyse the same set of data. Their analysis led to the definition of only six (instead of 122) candidate cancer genes (CAN genes) in breast cancer and 28 (instead of 69) CAN genes in colorectal cancer.

Addressing the above criticism, the original study group introduced two experimentally derived concepts – candidate cancer genes (CAN genes; driver mutation genes) versus passenger mutation genes [34]. They argued that driver mutation genes identified by Sjöblom et al. [35] mutated at higher rates than the experimentally determined passenger mutation rate. Thus, they advocated that the focused functional studies are essential for determining cancer treatment strategies as guided by WGS approach, which can help identify those mutated genes that likely are subject to study in model organisms.

**Cost**

The US Cancer Genome Atlas is an ambitious $1.5 billion federal project to search systematically for genes mutated in dozens of cancer types [36]. The Hopkins study alone costs about $3 million per whole-genome sequence, which is rapidly dropping over the past three years. At Washington University, the 2008 AML study cost just over $1.5 million; over a third of it was devoted to developing the bioinformatics required to compare the tumour and normal genomes, whereas today, the cost is only $10,000.
Washington University researchers have sequenced the complete tumour and normal genomes of 150 patients. Washington University estimated a complete panel of current genome tests can cost up to $10,000 per patient, and that cost will rise as new prognostic genes are discovered and added to the panel of molecular profiling.

Future Directions

Current consensus argues for a two-stage experimental approach to cancer genome research, a discovery sequencing screen followed by a validation functional screen (Figure 1). Sequencing data can, in general, only point to candidate genes worthy of further functional study. Sequence data can only identify genes that are mutated at unusually high rates [25]. In general, such data cannot determine whether the higher rate is the result of higher intrinsic mutability or positive selection during tumourgenesis. At most, sequencing data should be used to prioritise candidate cancer genes on the basis of their mutation characteristics and frequency.

A whole genome approach to cancer identifies the genetic alterations in cancers on a genome-wide scale. The resulting compendium of genetic changes in individual tumours provides new opportunities for diagnosis and treatment of cancer in each patient. However, this approach raises significant new challenges in understanding the roles that these mutations actually play in cancer. Such an understanding is essential for patients with the mutations to have dramatic benefits.