REVIEWS

The cancer genome

Michael R. Stratton[1,2], Peter J. Campbell[1,3] & P. Andrew Futreal[1]

All cancers arise as a result of changes that have occurred in the DNA sequence of the genomes of cancer cells. Over the past quarter of a century much has been learnt about these mutations and the abnormal genes that operate in human cancers. We are now, however, moving into an era in which it will be possible to obtain the complete DNA sequence of large numbers of cancer genomes. These studies will provide us with a detailed and comprehensive perspective on how individual cancers have developed.

Cancer is responsible for one in eight deaths worldwide[1]. It encompasses more than 100 distinct diseases with diverse risk factors and epidemiology which originate from most of the cell types and organs of the human body and which are characterized by relatively unrestrained proliferation of cells that can invade beyond normal tissue boundaries and metastasize to distant organs.

Early insights into the central role of the genome in cancer development emerged in the late nineteenth and early twentieth centuries from studies by David von Hansemann[2] and Theodor Boveri[3]. Examining dividing cancer cells under the microscope, they observed the presence of bizarre chromosomal aberrations. This led to the proposal that cancers are abnormal clones of cells characterized by and caused by abnormalities of hereditary material. Following the discovery of DNA as the molecular substrate of inheritance[4] and determination of its structure[5], this speculation was supported by the demonstration that agents that damage DNA and generate mutations also cause cancer[6]. Subsequently, increasingly refined analyses of cancer cell chromosomes showed that specific and recurrent genomic abnormalities, such as the translocation between chromosomes 9 and 22 in chronic myeloid leukaemia (known as the 'Philadelphia' translocation[7,8]), are associated with particular cancer types. Finally, it was demonstrated that introduction of total genomic DNA from human cancers into phenotypically normal NIH3T3 cells could convert them into cancer cells[9,10]. Isolation of the specific DNA segment responsible for this transforming activity led to the identification of the first naturally occurring, human cancer-causing sequence change—the single base G > T substitution that causes a glycine to valine substitution in codon 12 of the *HRAS* gene[11,12]. This seminal discovery in 1982 inaugurated an era of vigorous searching for the abnormal genes underlying the development of human cancer that continues today.

Here we review the principles of our current understanding of cancer genomes. We look forward to the explosion of information about cancer genomes that is imminent and the insights into the process of oncogenesis that this promises to generate.

Cancer is an evolutionary process

All cancers are thought to share a common pathogenesis. Each is the outcome of a process of Darwinian evolution occurring among cell populations within the microenvironments provided by the tissues of a multicellular organism. Analogous to Darwinian evolution occurring in the origins of species, cancer development is based on two constituent processes, the continuous acquisition of heritable genetic variation in individual cells by more-or-less random mutation and natural selection acting on the resultant phenotypic diversity. The selection may weed out cells that have acquired deleterious mutations or it may foster cells carrying alterations that confer the capability to proliferate and survive more effectively than their neighbours. Within an adult human there are probably thousands of minor winners of this ongoing competition, most of which have limited abnormal growth potential and are invisible or manifest as common benign growths such as skin moles. Occasionally, however, a single cell acquires a set of sufficiently advantageous mutations that allows it to proliferate autonomously, invade tissues and metastasize.

The catalogue of somatic mutations in a cancer genome

Like all the cells that constitute the human body, a cancer cell is a direct descendant, through a lineage of mitotic cell divisions, of the fertilized egg from which the cancer patient developed and therefore carries a copy of its diploid genome (Fig. 1). However, the DNA sequence of a cancer cell genome, and indeed of most normal cell genomes, has acquired a set of differences from its progenitor fertilized egg. These are collectively termed somatic mutations to distinguish them from germline mutations that are inherited from parents and transmitted to offspring.

The somatic mutations in a cancer cell genome may encompass several distinct classes of DNA sequence change. These include substitutions of one base by another; insertions or deletions of small or large segments of DNA; rearrangements, in which DNA has been broken and then rejoined to a DNA segment from elsewhere in the genome; copy number increases from the two copies present in the normal diploid genome, sometimes to several hundred copies (known as gene amplification); and copy number reductions that may result in complete absence of a DNA sequence from the cancer genome (Fig. 2).

In addition, the cancer cell may have acquired, from exogenous sources, completely new DNA sequences, notably those of viruses such as human papilloma virus, Epstein Barr virus, hepatitis B virus, human T lymphotropic virus 1 and human herpes virus 8, each of which is known to contribute to the genesis of one or more type of cancer[13].

Compared to the fertilized egg, the cancer genome will also have acquired epigenetic changes which alter chromatin structure and gene expression, and which manifest at DNA sequence level by changes in the methylation status of some cytosine residues. Epigenetic changes can be subject to the same Darwinian natural selection as genetic events, provided that there is epigenetic variation in the population of competing cells, that the epigenetic changes are stably heritable from the mother to the daughter cell and that they generate phenotypic effects for selection to act on.

Finally, it should not be forgotten that another genome is harboured within the cancer cell. The thousands of mitochondria present each carry a circular genome of approximately 17 kilobases. Somatic mutations in

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2]Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. [3]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK.

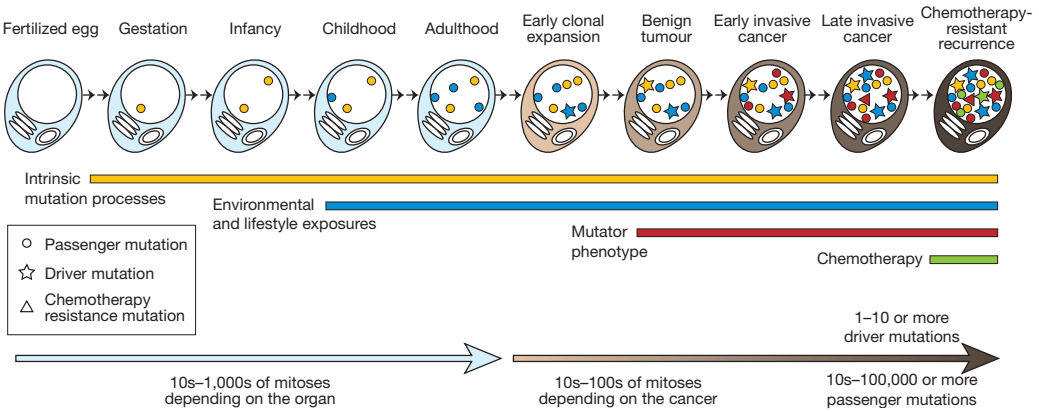

Figure 1 | The lineage of mitotic cell divisions from the fertilized egg to a single cell within a cancer showing the timing of the somatic mutations acquired by the cancer cell and the processes that contribute to them. Mutations may be acquired while the cell lineage is phenotypically normal, reflecting both the intrinsic mutations acquired during normal cell division and the effects of exogenous mutagens. During the development of the cancer other processes, for example DNA repair defects, may contribute to the mutational burden. Passenger mutations do not have any effect on the cancer cell, but driver mutations will cause a clonal expansion. Relapse after chemotherapy can be associated with resistance mutations that often predate the initiation of treatment.

mitochondrial genomes have been reported in many human cancers, although their role in the development of the disease is not clear[14].

Acquisition of somatic mutations in cancer genomes

The mutations found in a cancer cell genome have accumulated over the lifetime of the cancer patient. Some were acquired when ancestors of the cancer cell were biologically normal, showing no phenotypic characteristics of a cancer cell (Fig. 1). DNA in normal cells is continuously damaged by mutagens of both internal and external origins. Most of this damage is repaired. However, a small fraction may be converted into fixed mutations and DNA replication itself has a low intrinsic error rate. Our understanding of somatic mutation rates in normal human cells is still relatively rudimentary. However, it is likely that the mutation rates of each of the various structural classes of somatic mutation differ and that there are differences among cell types too. Mutation rates increase in the presence of substantial exogenous mutagenic exposures, for example tobacco smoke carcinogens, naturally occurring chemicals such as aflatoxins, which are produced by fungi, or various forms of radiation including ultraviolet light. These exposures are associated with increased rates of lung, liver and skin cancer, respectively, and somatic mutations within such cancers often exhibit the distinctive mutational signatures known to be associated with the mutagen[15]. The rates of the different classes of somatic mutation are also increased in several rare inherited diseases, for example Fanconi anaemia, ataxia telangiectasia, mosaic variegated aneuploidy and xeroderma pigmentosum, each of which is also associated with increased risks of cancer[16,17].

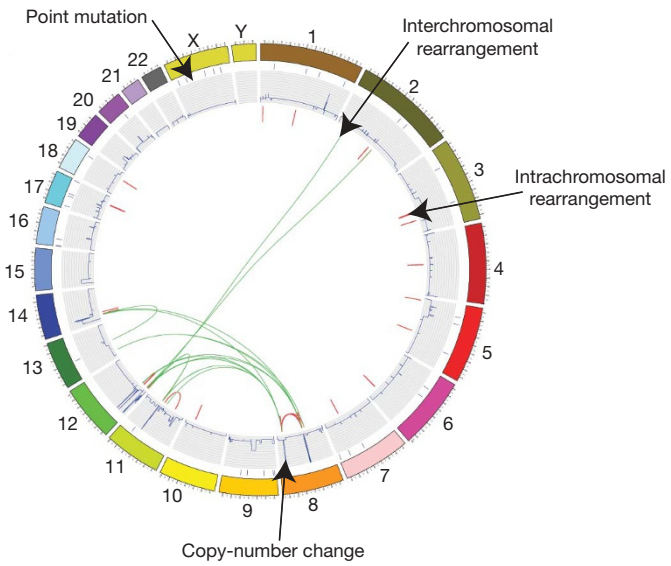

Figure 2 | Figurative depiction of the landscape of somatic mutations present in a single cancer genome. Part of catalogue of somatic mutations in the small-cell lung cancer cell line NCI-H2171. Individual chromosomes are depicted on the outer circle followed by concentric tracks for point mutation, copy number and rearrangement data relative to mapping position in the genome. Arrows indicate examples of the various types of somatic mutation present in this cancer genome.

The rest of the somatic mutations in a cancer cell genome have been acquired during the segment of the cell lineage in which predecessors of the cancer cell already show phenotypic evidence of neoplastic change (Fig. 1). Whether the somatic mutation rate is always higher during this part of the lineage is controversial[18,19]. For some cancers this is clearly the case. For example, colorectal and endometrial cancers with defective DNA mismatch repair due to abnormalities in genes such as *MLH1* and *MSH2*, exhibit increased rates of acquisition of single nucleotide changes and small insertions/deletions at polynucleotide tracts[20]. Other classes of such 'mutator phenotypes' may exist, for example leading to abnormalities in chromosome number or increased rates of genomic rearrangement, although these are generally less well characterized[20]. The merit of an increased somatic mutation rate with respect to the development of cancer is that it increases the DNA sequence diversity on which selection can act. However, it has been suggested that the mutation rates of normal cells may be sufficient to account for the development of some cancers, without the requirement for a mutator phenotype[18,19].

The course of mutation acquisition need not be smooth and predecessors of the cancer cell may suddenly acquire a large number of mutations. This is sometimes termed 'crisis'[21], and can occur after attrition of the telomeres that normally cap the ends of chromosomes, with the cell having to substantially reorganize its genome to survive.

Although complex and potentially cryptic to decipher, the catalogue of somatic mutations present in a cancer cell therefore represents a cumulative archaeological record of all the mutational processes the cancer cell has experienced throughout the lifetime of the patient. It provides a rich, and predominantly unmined, source of information for cancer epidemiologists and biologists with which to interrogate the development of individual tumours.

Driver and passenger mutations

Each somatic mutation in a cancer cell genome, whatever its structural nature, may be classified according to its consequences for cancer development. 'Driver' mutations confer growth advantage on the cells

carrying them and have been positively selected during the evolution of the cancer. They reside, by definition, in the subset of genes known as 'cancer genes'. The remainder of mutations are 'passengers' that do not confer growth advantage, but happened to be present in an ancestor of the cancer cell when it acquired one of its drivers (see Box 1).

The number of driver mutations, and hence the number of abnormal cancer genes, in an individual cancer is a central conceptual parameter of cancer development, but is not well established. It is highly likely that most cancers carry more than one driver and that the number varies between cancer types. On the basis of age–incidence statistics it has been suggested that common adult epithelial cancers such as breast, colorectal and prostate require 5–7 rate-limiting events, possibly equating to drivers, whereas cancers of the haematological system may require fewer[22]. These estimates are supported by experimental studies which show that engineering changes in the functions of at least five or six genes in normal primary human cells is necessary to convert them into cancer cells[23]. However, recent analyses of somatic mutation data from cancers indicate that the number of drivers might be much higher[24]. Ultimately, direct estimates of the number of drivers in individual cancers will be provided by identifying all the cancer genes and systematically measuring the prevalence of mutations in them.

One important subclass of driver is a mutation that confers resistance to cancer therapy (Fig. 1). These are typically found in recurrences of cancers that have initially responded to treatment but that are now resistant. Resistance mutations often confer limited growth advantage on the cancer cell in the absence of therapy. Some seem to predate initiation of treatment, existing as passengers in minor subclones of the cancer cell population until the selective environment is changed by the initiation of therapy[25,26]. The passenger is then converted into a driver and the resistant subclone preferentially expands, manifesting as the recurrence.

Box 1 | Driver and passenger mutations

All cancers arise as a result of somatically acquired changes in the DNA of cancer cells. That does not mean, however, that all the somatic abnormalities present in a cancer genome have been involved in development of the cancer. Indeed, it is likely that some have made no contribution at all. To embody this concept, the terms 'driver' and 'passenger' mutation have been coined.

A driver mutation is causally implicated in oncogenesis. It has conferred growth advantage on the cancer cell and has been positively selected in the microenvironment of the tissue in which the cancer arises. A driver mutation need not be required for maintenance of the final cancer (although it often is) but it must have been selected at some point along the lineage of cancer development shown in Fig. 1.

A passenger mutation has not been selected, has not conferred clonal growth advantage and has therefore not contributed to cancer development. Passenger mutations are found within cancer genomes because somatic mutations without functional consequences often occur during cell division. Thus, a cell that acquires a driver mutation will already have biologically inert somatic mutations within its genome. These will be carried along in the clonal expansion that follows and therefore will be present in all cells of the final cancer.

Some somatic mutations may actually impair cell survival. These will usually be subject to negative selection and hence be absent from the cancer genome. The traces of negative selection in cancer genomes are currently limited but it would be surprising if it was not operative.

A central goal of cancer genome analysis is the identification of cancer genes that, by definition, carry driver mutations. A key challenge will therefore be to distinguish driver from passenger mutations. The main strategy generally used exploits a number of structural signatures associated with mutations that are under positive selection. For example, driver mutations cluster in the subset of genes that are cancer genes whereas passenger mutations are more or less randomly distributed. This has been the approach adopted fruitfully in the past to identify most somatically mutated cancer genes in studies targeted at small regions of the genome.

Whole-genome sequencing, however, incorporating analysis of more than 20,000 protein-coding genes and unknown numbers of functional elements in intronic and intergenic DNA, presents a greater challenge, one rendered more daunting by the likelihood that passenger mutations in most cancer genomes substantially outnumber drivers. Because many cancer genes seem to contribute to cancer development in only a small fraction of tumours, large sample sets will have to be analysed to distinguish infrequently mutated cancer genes from genes with random clusters of passenger mutations. Furthermore, it is conceivable that some mutational processes are directed at specific genomic regions and thus generate clusters of passenger mutations that may be mistaken for drivers.

Therefore, all such signatures of positive selection need to be interpreted with caution. In practice, however, used in an informed and critical manner they will remain effective and reliable guides to the identification of cancer genes. Investigation of the biological consequences of putative driver mutations will often consolidate the evidence implicating them in oncogenesis and will provide insight into the subverted biological processes by which they contribute to cancer development.

The repertoire of somatically mutated cancer genes

The identification of driver mutations and the cancer genes that they alter has been a central aim of cancer research for more than a quarter of a century. It has been a remarkably successful endeavour, with at least 350 (1.6%) of the ~22,000 protein-coding genes in the human genome reported to show recurrent somatic mutations in cancer with strong evidence that these contribute to cancer development[27] (http://www.sanger.ac.uk/genetics/CGP/Census/). Most were identified by first establishing their physical location in the genome through low-resolution genome-wide screens, in particular cytogenetics for chromosomal translocations in leukaemias and lymphomas. A few were discovered using biological assays for transforming activity of whole cancer cell DNA and others through targeted mutational screens guided by biologically well-informed guesswork. Mutations in ~10% of these genes are also found in the germ line, where they confer an increased risk of developing cancer, and these were often initially identified by genetic linkage analysis of affected families. The size of the full repertoire of human cancer genes is a matter of speculation. However, studies in mice have suggested that more than 2,000 genes, when appropriately altered, may have the potential to contribute to cancer development[28].

The known cancer genes run the gamut of tissue specificities and mutation prevalences. Some, for example *TP53* and *KRAS*, are frequently mutated in diverse types of cancer whereas others are rare and/or restricted to one cancer type (http://www.sanger.ac.uk/genetics/CGP/cosmic/). In some cancer types, for example colorectal and pancreatic cancer, abnormalities in several known cancer genes are common. In contrast, in gastric cancer, relatively few mutations in known cancer genes have been reported.

Approximately 90% of the known somatically mutated cancer genes are dominantly acting, that is, mutation of just one allele is sufficient to contribute to cancer development. The mutation in such cases usually results in activation of the encoded protein. Ten per cent act in a recessive manner, requiring mutation of both alleles, and the mutations usually result in abrogation of protein function (these are sometimes known as tumour suppressor genes).

Patterns of mutation differ between dominant and recessive cancer genes. Recessive cancer genes are characterized by diverse mutation types, ranging from single base substitutions to whole gene deletions, which have the common outcome of abolishing the function of the encoded protein. In each dominantly acting cancer gene, however, the repertoire of cancer-causing somatic mutations is usually more constrained, both with respect to the type of mutation and its location in the gene. Missense amino acid changes (often restricted to certain key amino acids), in-frame insertions and deletions, and gene amplification are all common mutational mechanisms for activating dominantly acting cancer genes. Most, however, are activated through genomic rearrangement. This may join the sequences of two different genes to create a fusion gene or it may position the cancer gene adjacent to regulatory elements from elsewhere in the genome, resulting in abnormal expression patterns. Most of the known rearranged cancer

genes are operative in the relatively rare subset of cancers constituted by leukaemias, lymphomas and sarcomas. Recently, however, rearranged cancer fusion genes were discovered in more than half of prostate cancer cases[29] and in lung adenocarcinomas[30]. Their late discovery probably reflects the difficulty of identifying them amidst the jumble of passenger rearrangements present in many cancer genomes and hints that there are many more rearranged cancer genes to be found in common cancers.

Much of what we know about the biological pathways and processes that are subverted in cancer has originated from experiments exploring the functions of cancer genes. Certain gene families, notably the protein kinases, feature particularly prominently among cancer genes. Furthermore, cancer genes cluster on certain signalling pathways. For example, in the classical MAPK/ERK pathway[31] upstream mutations are found in cell-membrane-bound receptor tyrosine kinases such as *EGFR*, *ERBB2*, *FGFR1*, *FGFR2*, *FGFR3*, *PDGFRA* and *PDGFRB* and also in the downstream cytoplasmic components *NF1*, *PTPN11*, *HRAS*, *KRAS*, *NRAS* and *BRAF*. Recent exhaustive mutational analyses in gliomas have indicated that almost all cases have a mutation at one of the genes on these critical signalling pathways[32].

For some cancers, classification and treatment protocols are now defined by the presence of abnormal cancer genes. Acute myeloid leukaemia, for example, is subclassified on the basis of the presence of abnormalities involving specific cancer genes[33]. Each subtype has a characteristic gene expression profile, cellular morphology, clinical syndrome, prognosis and opportunity for targeted therapy. Moreover, because cancer cells are dependent on the abnormal proteins encoded by mutated cancer genes, they have become targets for the development of new cancer therapeutics. Flagships for this new generation of treatments include imatinib, an inhibitor of the proteins encoded by the *ABL* and *KIT* genes, which are mutated and activated, respectively, in chronic myeloid leukaemia[34] and gastrointestinal stromal tumours[35], and trastuzumab, an antibody directed against the protein encoded by *ERBB2* (also known as *HER2*), which is commonly amplified and overexpressed in breast cancer[36].

Early systematic sequencing of cancer genomes

Provision of the reference human genome sequence at the turn of the millennium offered new strategies and opportunities for surveying cancer genomes. Rather than depending on low-resolution maps, the highest possible resolution map, the DNA sequence itself, became available and has empowered investigation of cancer genomes in several ways. For example, much higher-resolution arrays have been developed, allowing finer mapping of copy number changes in cancer genomes leading to the identification of several new amplified cancer genes.

The availability of the human genome sequence has also raised the possibility that DNA sequencing itself could become the primary tool for exploration of cancer genomes. This has prompted several pilot experiments. So far, most have sequenced large numbers of PCR products to detect the base substitutions and small insertions and deletions (collectively termed 'point' mutations) present in the coding exons of protein-coding genes[32,37–44]. Typically, such studies have covered several hundred megabases of cancer genome with designs ranging from hundreds of genes analysed in a few hundred cancers to most of the ~22,000 protein-coding genes in 10–20 examples of a particular cancer class.

Several insights have been provided by these screens. They have brought success in the identification of point-mutated cancer genes including *BRAF*[45], *PIK3CA*[46], *EGFR*[47], *HER2* (ref. 48), *JAK2* (ref. 49), *UTX* (ref. 50) and *IDH1* (ref. 41). Some of these were unique discoveries, whereas others were simultaneously discovered in targeted mutational screens. Some were previously known cancer genes, but the discovery of point mutations highlighted new mechanisms and cancer types in which they are operative. Some were surprising and highlight the virtue of systematic and comprehensive screens, for example the discovery of the enzyme isocitrate dehydrogenase (*IDH1*), which constitutes part of the Krebs cycle of oxidative phosphorylation, as a cancer gene mutated in glioma[41]. Because many are kinases that are activated by the mutations found in cancer, they have prompted a wave of drug discovery to find inhibitors that may serve as anticancer therapeutics[51], some of which are already in clinical trials.

Exposing the landscape of the cancer genome

Important insights into the general parameters and patterns of somatic mutation in cancer have also emerged from these early studies. It appears that most somatic point mutations in cancer genomes are passengers[39]. Although this might have been predicted for mutations in intergenic and intronic DNA, it applies even in protein-coding exons. There is, however, statistical evidence in favour of many more driver mutations than can be accounted for by known cancer genes. These drivers appear to be distributed across a large number of genes, each of which is mutated infrequently, suggesting that the repertoire of somatically mutated human cancer genes is much larger than the ~350 currently catalogued[39,44]. Conceivably, these infrequently mutated cancer genes confer less selective growth advantage on a clone of cancer cells than more commonly mutated cancer genes, but other explanations can also be invoked. Some analyses also indicate that there may be as many as 20 driver mutations in individual cancers, considerably more than the 5–7 previously predicted[24].

Understanding of the prevalence and types of somatic mutation in cancer genomes has been greatly fostered by these studies. Some cancer genomes carry >100,000 point mutations whereas others have fewer than 1,000. Some of this variation can be accounted for by previous heavy mutagenic exposures or the existence of known DNA repair defects. However, in a subset of breast cancers there are large numbers of C-to-G base substitutions, almost always occurring at cytosines that follow a thymine, for which there is no obvious explanation and for which unknown exposures and/or mutator phenotypes are presumably responsible[42,43].

The effects of chemotherapy on the cancer genome have also been revealed by systematic sequencing experiments. For example, gliomas that recur after treatment with the DNA alkylating agent temozolomide have been shown to carry huge numbers of mutations with a signature typical of such agents[32,52,53]. The fact that the mutations could be detected at all indicates that these recurrences are clonal. Thus, these studies indicate that, although temozolomide only confers a short increased lifespan for the patient, almost all cells in a glioma respond and a single cell that is resistant to the chemotherapy proliferates to form the recurrence. Additional studies guided by these observations led to the identification of the underlying mutated resistance gene[52,53].

Beyond point mutations, some investigations have begun to explore the features of genomic rearrangements in common cancers, about which remarkably little is known. Early studies using conventional Sanger sequencing indicated that there is substantial complexity of rearrangement in these genomes[54,55]. The recent advent of massively parallel, second-generation sequencing technologies has enabled more comprehensive genome-wide screens revealing that some cancer genomes carry hundreds of somatically acquired rearrangements, whereas others carry very few. Moreover, the distinctive patterns of rearrangement found indicate that currently uncharacterized mutational processes may be at work[56].

Sequencing of cancer genomes in the future

The large-scale, systematic sequencing studies conducted so far have been constrained by the relatively low throughput and high cost of sequencing. They have therefore generally been restricted to components of the cancer genome (for example, coding exons), to small numbers of cancer samples or to a subset of the mutational classes present. In principle, however, all the structural classes of somatic mutation can be detected genome-wide by randomly fragmenting the cancer genome and sequencing large numbers of fragments such that each base in the reference human genome is covered several times by a sequence generated from the cancer. With a high enough level of coverage, essentially a full catalogue of somatic mutations from an individual

cancer genome can be obtained, including all point mutations, rearrangements and copy number changes. Mutations in the accompanying mitochondrial genomes of the cancer will also be collected. With further adaptation this could be extended to include epigenetic alterations and could be applied to the transcriptomes of cancers to investigate the first phenotypic effects of all these changes. This catalogue will include all the driver mutations and hence all the cancer genes operating in that cancer, whether they are protein-coding genes, non-coding RNA genes or more cryptic functional elements of the genome. Indeed, if known or unknown DNA viruses have contributed to oncogenesis these will also be discovered. The catalogue will also include all the passenger mutations that incorporate the signatures of previous exposures, DNA repair defects and other mutational processes the cancer has experienced over the decades during which it was evolving.

Until recently, this was an unattainable fantasy. However, the arrival of second-generation sequencing technologies promises a new era for cancer genomics. These platforms currently generate billions of bases of DNA sequence per week, yields that are predicted to increase rapidly over the next couple of years (Fig. 3). Several proof-of-principle studies have recently been published applying these technologies to cancer samples. These have demonstrated that the current generation of massively parallel sequencing platforms can identify the full range of somatically acquired genetic alteration in cancer, including point mutations on a genome-wide basis[57], insertions and deletions[57], copy number changes[56] and genomic rearrangements[56], as well as characterizing the cancer cell transcriptome[40,41]. Furthermore, these approaches have the potential to identify subclonal genetic diversity within the population of cancer cells[58], with particular relevance to the detection of subclones carrying drug-resistance mutations[59]. Indeed, one high-coverage cancer genome sequence has recently been reported[57] and several others will emerge during the course of 2009.

Even with the remarkable technological advances in sequencing, however, the parameters of experiments to catalogue all somatically acquired variants in a cancer genome are sobering. To obtain a complete catalogue of somatic mutations from an individual human cancer may require 20-fold sequence coverage of the cancer genome, and possibly more. Somatic mutations then have to be distinguished from inherited DNA variants. Although most inherited variants that are common in human populations (>5% allele frequency) have been discovered and are registered in databases, there are myriad rare inherited single nucleotide polymorphisms and structural variants that are not. In most cancer genomes these rare germline variants far outnumber the somatic mutations present. Therefore, for the foreseeable future at least, a high-coverage sequence of the normal genome from the same individual as the cancer will be an inescapable extra burden to allow identification of the somatic changes. Thus, more than 100,000,000,000 base pairs of DNA sequence will probably be required to identify the catalogue of somatic mutations in a single cancer genome.

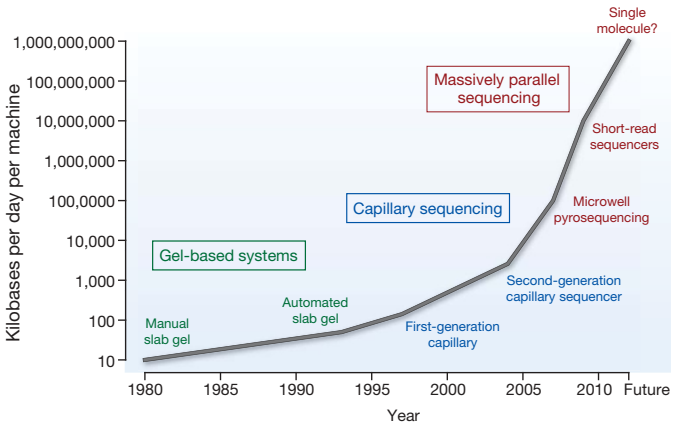

Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

Subsequently, it will be necessary to distinguish driver mutations from passengers (see Box 1). The power to distinguish clusters of driver mutations in cancer genes from chance clusters of randomly distributed passenger mutations will depend on how frequently a cancer gene is mutated and the prevalence of passenger mutations. To be confident of identifying a cancer gene that is mutated in ~5% of a particular type of cancer will require hundreds of cases to be sequenced. Each of the >100 cancer types will probably require similar sample sizes.

Coordinating the sequencing of cancer genomes

There is, therefore, much work to be done over the next few years. Ideally, it should be organized to maximize use of resources and harmonize the product. This is the mission of the International Cancer Genome Consortium (ICGC, see http://www.icgc.org/home). Building on the success of previous multinational, collaborative initiatives such as the Human Genome Project and the HapMap consortium, the aim of ICGC is to comprehensively characterize somatically acquired genetic events in at least fifty classes of cancer, including those with the highest global incidence and mortality, requiring high-coverage sequencing of 20,000 cancer genomes or more. The full catalogues of somatic mutation from each of these cancers will be integrated with expression and epigenetic profiles of the same cases and correlated with clinical features.

Projects under the ICGC imprimatur will adhere to predetermined standards and procedures for ethical approval, data release, intellectual property, sample quality, clinical annotation, data quality, data storage and sequencing completion. Most importantly, given the demanding nature of the task, the ICGC will coordinate studies to minimize duplication of effort and enable the most parsimonious deployment of resources.

The proposal to sequence large numbers of cancer genomes has generated controversy reminiscent of the debate before sequencing of the reference human genome almost 20 years ago. The experiments will be expensive and, to some extent, we cannot predict what will be found. However, the human genome is finite. Therefore, with further technological advances in DNA sequencing that are already in sight, this is a deliverable project that will comprehensively elucidate central questions relating to the nature of human cancer. The clinical and translational implications of such a body of work are profound. Beyond the identification of further potentially druggable cancer genes, a comprehensive catalogue of somatic mutations in carefully characterized clinical samples will generate new insights into the genetic patterns that underpin disease phenotype, prognosis, drug response and chemotherapy resistance. As the costs of sequencing whole cancer genomes drop towards US$1,000, routine sequencing in a clinical, diagnostic setting will become feasible. Such data may drive individualized therapeutic decision-making through the ability to predict prognosis, to choose therapeutic regimens known to have efficacy for the particular genetic subtype of cancer, to sensitively monitor response to therapy and to identify rare subclones harbouring drug-resistance mutations before therapy is even initiated. Individualized therapeutics will require individualized diagnostics.

The discussion is therefore not about whether to do the experiment, but when and how. In a manner similar to the Human Genome Project we have to coordinate the work internationally to maximize use of resources and minimize duplication of effort to generate a resource of high quality so that we only have to do it once, empowering cancer research with a lasting legacy for the future.

Forward look

Approximately 100,000 somatic mutations from cancer genomes have been reported in the quarter of a century since the first somatic

mutation was found in *HRAS*. Over the next few years several hundred million more will be revealed by large-scale, complete sequencing of cancer genomes. These data will provide us with a fine-grained picture of the evolutionary processes that underlie our commonest genetic disease, providing new insights into the origins and new directions for the treatment of cancer.

1. Garcia, M. *et al. Global Cancer Facts and Figures 2007* (ACS, 2007).
2. von Hansemann, D. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchows Arch. Path. Anat.* **119,** 299 (1890).
3. Boveri, T. *Zur Frage der Entstehung Maligner Tumoren*. 1–64 (Gustav Fischer, 1914).
4. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *J. Exp. Med.* **79,** 137–158 (1944).
5. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
6. Loeb, L. A. & Harris, C. C. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res.* **68,** 6863–6872 (2008).
7. Nowell, P. & Hungerford, D. A minute chromosome in human granulocytic leukemia. *Science* **132,** 1497 (1960).
8. Rowley, J. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–293 (1973).
   This is a pivotal paper describing the identification of the recurrent Philadelphia chromosome in chronic myeloid leukaemia.
9. Krontiris, T. G. & Cooper, G. M. Transforming activity of human tumor DNAs. *Proc. Natl Acad. Sci. USA* **78,** 1181–1184 (1981).
10. Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290,** 261–264 (1981).
11. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–152 (1982).
    This paper and the next are seminal papers identifying the first somatic point mutation in a human cancer. These findings launched a new era of molecular cancer genetics research.
12. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–149 (1982).
13. Talbot, S. J. & Crawford, D. H. Viruses and tumours — an update. *Eur. J. Cancer* **40,** 1998–2005 (2004).
14. Chatterjee, A., Mambo, E. & Sidransky, D. Mitochondrial DNA mutations in human cancer. *Oncogene* **25,** 4663–4674 (2006).
15. Olivier, M., Hussain, S. P., Caron de Fromentel, C., Hainaut, P. & Harris, C. C. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. *IARC Sci. Publ.* 247–270 (2004).
16. Kennedy, R. D. & D'Andrea, A. D. DNA repair pathways in clinical practice: lessons from pediatric cancer susceptibility syndromes. *J. Clin. Oncol.* **24,** 3799–3808 (2006).
17. Hanks, S. & Rahman, N. Aneuploidy-cancer predisposition syndromes: a new link between the mitotic spindle checkpoint and cancer. *Cell Cycle* **4,** 225–227 (2005).
18. Bodmer, W. & Loeb, L. A. Genetic instability is not a requirement for tumor development. *Cancer Res.* **68,** 3558–3561 (2008).
19. Loeb, L. A., Bielas, J. H., Beckman, R. A. & Bodmer, I. W. Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res.* **68,** 3551–3557 (2008).
20. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396,** 643–649 (1998).
21. Stewart, S. A. & Weinberg, R. A. Telomeres: cancer to human aging. *Annu. Rev. Cell Dev. Biol.* **22,** 531–557 (2006).
22. Miller, D. G. On the nature of susceptibility to cancer. The presidential address. *Cancer* **46,** 1307–1318 (1980).
23. Schinzel, A. C. & Hahn, W. C. Oncogenic transformation and experimental models of human cancer. *Front. Biosci.* **13,** 71–84 (2008).
24. Beerenwinkel, N. *et al.* Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.* **3,** e225 (2007).
25. Roche-Lestienne, C. *et al.* Several types of mutations of the *Abl* gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood* **100,** 1014–1018 (2002).
26. Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322,** 1377–1380 (2008).
27. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
28. Touw, I. P. & Erkeland, S. J. Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia. *Mol. Ther.* **15,** 13–19 (2007).
29. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).
    This is an important paper that, in addition to identifying a major somatic genetic contributor to prostate cancer, proved that highly prevalent gene fusions can drive common adult solid tumours.
30. Soda, M. *et al.* Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–566 (2007).
31. Johnson, G. L. & Lapadat, R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* **298,** 1911–1912 (2002).
32. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
    This is an important paper in a series of ever-larger studies designed to comprehensively screen cancer genomes for somatically mutated cancer genes.
33. Haferlach, T. Molecular genetic pathways as therapeutic targets in acute myeloid leukemia. *Hematology Am. Soc. Hematol. Educ. Prog.* **2008,** 400–411 (2008).
34. Druker, B. J. Translation of the Philadelphia chromosome into therapy for CML. *Blood* **112,** 4808–4817 (2008).
35. Sleijfer, S., Wiemer, E. & Verweij, J. Drug insight: gastrointestinal stromal tumors (GIST)—the solid tumor model for cancer-specific treatment. *Nature Clin. Prac. Oncol.* **5,** 102–111 (2008).
36. Mariani, G., Fasolo, A., De Benedictis, E. & Gianni, L. Trastuzumab as adjuvant systemic therapy for HER2-positive breast cancer. *Nature Clin. Prac. Oncol.* **6,** 93–104 (2009).
37. Bardelli, A. *et al.* Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300,** 949 (2003).
38. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
39. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
    This paper describes the heterogeneity of somatic mutation type and prevalence in human cancers, provides evidence for the concepts of passenger and driver mutations, and highlights the existence of many infrequently mutated cancer genes.
40. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321,** 1801–1806 (2008).
41. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).
42. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006).
    This is an important paper in a series of systematic large-scale screens of coding exons for point mutations in human cancer.
43. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature Genet.* **37,** 590–592 (2005).
44. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
45. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).
46. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304,** 554 (2004).
47. Paez, J. G. *et al. EGFR* mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304,** 1497–1500 (2004).
48. Stephens, P. *et al.* Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431,** 525–526 (2004).
49. Levine, R. L. *et al.* Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **7,** 387–397 (2005).
50. van Haften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nature Genet.* (in the press).
51. Sawyers, C. Targeted cancer therapy. *Nature* **432,** 294–297 (2004).
52. Hunter, C. *et al.* A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66,** 3987–3991 (2006).
53. Cahill, D. P. *et al.* Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.* **13,** 2038–2045 (2007).
54. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17,** 1296–1303 (2007).
55. Volik, S. *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA* **100,** 7696–7701 (2003).
56. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40,** 722–729 (2008).
57. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
58. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA* **105,** 13081–13086 (2008).
59. Thomas, R. K. *et al.* Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Med.* **12,** 852–855 (2006).

Acknowledgements We would like to thank N. Rahman for comments on the manuscript, G. Tang and B. Barrell for contributions to the figures and the Kay Kendall Leukaemia Fund and the Wellcome Trust for support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.R.S. (mrs@sanger.ac.uk).