Applications of next-generation sequencing in cancer

Pawan Upadhyay, Renu Dwivedi and Amit Dutt*

Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Navi Mumbai 410 210, India

The completion of the Human Genome Project established a baseline for human genome reference sequence allowing characterization of various alterations underlying several human diseases, including cancer and has brought the field of genomics to this unprecedented moment of a great scientific ferment. It has also sparked a concomitant revolution in sequencing technologies that have become a fundamental tool for genome analysis with potential to transform medical practice. As an emerging field, the next-generation sequencing (NGS) technology has stimulated rapid cataloguing of all alterations in cancer genomes and has enabled researchers to look at large-scale genome events such as chromosomal lesions and copy-number variations as well as small-scale aberrations represented by point mutations, small insertions and deletions. Several advancements with smaller and faster versions of available technologies have recently been introduced enabling more democratic usage of the technology. Here, we review the application of NGS technology in understanding the underlying goal to catalogue human cancer-causing somatic mutations.

Keywords: Cancer, genome analysis, next-generation sequencing, somatic mutations.

Introduction

CANCER is characterized by unrestricted proliferation of cells that are genetically modified to have acquired the ability to metastasize to distant organ sites[1]. In essence, cancer is a genetic disease arising from a stepwise accumulation of genetic and epigenetic alterations that deregulate multiple complex regulatory pathways of genes, proteins and biochemical components affecting cellular growth, division, migration and survival[1]. These alterations arise at somatic level in sporadic cancers or may be inherited through the germline in familial cancers, either activating proto-oncogenes or inactivating tumour suppressor genes.

While the proto-oncogene can be activated by a single mutation event, the inactivation of recessive tumour suppressor genes requires at least two hits, usually through a large deletion and a smaller mutational event such as a point mutation, small deletion or change in promoter methylation[2]. The characterization of the function of these tumour suppressors and oncogenes has led to not only the discovery of the biochemical pathways underlying the process of carcinogenesis, but also to an understanding of the normal homeostatic roles such pathways play in normal cells and tissues. For example, the insights to the mammalian cell-cycle process, apoptotic pathways, and growth factor signalling pathways were significantly driven with the discovery of cancer-associated genes such as the *RB1* tumour suppressor, the *Bcl2* oncogene, and oncogenic receptor tyrosine kinase receptors such as *PDGFR* and *HER2* (refs 3–5). These activated oncogenes conferring oncogenic addiction to cancer cells for maintenance of their malignant phenotype, are the Achilles' heel for the cells[6]. The cause of addiction could thus be an effective therapeutic target to kill the addicted cancer cells, a prototypical example being the use of imatinib (Gleevec), a small-molecule inhibitor of the BCR-ABL fusion protein in chronic myelogenous leukaemia. Imatinib has additional activity against the PDGFRA, PDGFRB and KIT receptor tyrosine kinase gene products[7–10].

Beyond imatinib, effective development of targeted therapeutics that can interfere with the function of oncogenic molecular targets remains sparse due to lack of our understanding of the biology of these molecular alterations that drive the tumourigenesis in these cancers. This trend is set to change. Beginning with the success of finding inhibitor of the BCL-ABL fusion protein in treating chronic myeloid leukaemia patients, more recently, we have witnessed the gradual realization of several genetically targeted therapeutics[11]. A growing number of such 'targeted therapeutics' are now in routine clinical use for the treatment of human malignancies, including imatinib (Gleevec), against BCR-ABL positive chronic myelogenous leukaemia (CML) and c-kit mutant gastrointestinal stromal cancers (GIST)[12,13]; trastuzumab (Herceptin), a monoclonal antibody targeting the extracellular domain of the HER2/neu oncogene that is effective against a subset of patients with HER2-amplified breast cancer[14]; and erlotinib (Tarceva), an inhibitor of the EGF receptor that is effective in a subset of non-small cell lung cancers harbouring activating mutations in the EGFR tyrosine kinase domain[15–17]. More recently, translocations of the *ALK* tyrosine kinase gene were found in lung adenocarcinoma patients[18]. These findings were rapidly translated to

*For correspondence. (e-mail: adutt@actrec.gov.in)

clinical practice, with a successful clinical trial of the ALK inhibitor crizotinib in patients with ALK-translocated lung adenocarcinoma[19]. An even greater number of targeted agents are currently being examined in phase II and phase III clinical trials, and it is expected that the coming years will see a tremendous increase in the number of such agents available for the treatment of human cancer. The ongoing discovery of the genetic basis of cancer can be translated into therapeutic advances, including the search for novel biomarkers that allow early diagnosis and improved monitoring of the disease to offer wider therapeutic choices and personalized medication for individual cancer patients.

Challenges in studying human cancer with conventional technologies

The human tumours are enormously genetically heterogeneous and complex that imposes huge technical restrictions to discern between cancer-causing genomic aberrations ('drivers') and innocent bystander mutations ('passengers') that have no oncogenic potential in the cells[20–22]. Several factors lead to the biological complexity of primary tumours: the inter-tumour heterogeneity arises due to different subtypes with distinct morphological phenotype, expression profiles, mutation and copy-number variation patterns – for example, 73 different combination possibilities of mutated cancer genes were recently found among 100 breast cancers[23–26]; intra-tumour sub-clonal genetic heterogeneity is driven by multiple distinct driver genetic events[27]; in addition to the tumour heterogeneity in solid cancers, normal DNA contamination further confounds the determination of allele fraction of individual driver genetic events. To overcome these biological complexities of tumour samples, the analysis for causal genetic event needs to be adjusted for the underlying ploidy, purity and copy-number alterations at chromosomal regions harbouring these events. For example, if 40% of the clinical specimen DNA is derived from cancer cells and if a mutation is present on 1 of 5 copies of a chromosome, the frequency of heterozygous mutation will be 8% in the sample. Detection of such low allele fraction in any tumour sample falls below the resolution of the classical Sanger sequencing technology that has limited ability to handle and analyse low allele frequencies not represented at 1 : 1 ratios.

A comprehensive understanding of the genetic events during tumourigenesis can, however, be gained only by integrating the mutational analyses at nucleotide level with analyses of copy-number alterations, methylation status and translocations. For instance, a tumour suppressor gene could be inactivated by point mutation or deleted and haploinsufficient or subject to promoter methylation – it might be deleted in 10% of patients, mutated in another 3%, promoter-hypermethylated in another 12% of patients, and out-of-frame fused with some other chromosomal region in 2% of patients. Combining this information would reveal that the gene is altered in 27% of patients, elevating its relevance threshold. Unfortunately, it is not technically feasible to interrogate the complete set of these genomic alterations in a tumour in a systematic and comprehensive manner using the classical Sanger sequencing or microarray-based approaches due to their limiting resolution. However, recent possibility of using a disruptive next-generation DNA sequencing technology allows reading billions of nucleotides in a single run enabling complete genome characterization of cancer. A typical NGS run can detect point mutations, copy-number alterations, LOH, infectious agents, epigenetic modifications, translocations and complex rearrangements in an unbiased manner.

Next-generation sequencing technology in cancer research

Recent advances in sequencing technologies and comprehensive methods to map cancer-associated copy number and structural aberrations now make it possible to consider enumerating all of the genetic alterations (mutations, copy-number changes, translocations, epigenetic modification and integration of infectious agents) harboured by a particular tumour. Until half a decade ago, almost all sequencing studies, including the Human Genome Project (HGP), relied heavily and almost exclusively on Sanger-based sequencing techniques[28]. However, since HGP, which cost an estimated US$ 300 million[29], considerable effort has been invested in creating technologies capable of sequencing an entire human genome in a timely and cost-efficient manner at resolutions to detect low-frequency variants, not possible before. Since then, the cost of a DNA sequencing reaction has decreased twice as fast as Moore's law[30], largely due to the development of high-throughput sequencing techniques with single-base resolution (Figure 1)[31,32]. This technological innovation is referred to as 'next-generation sequencing', or massively parallel sequencing[33–35]. The most commonly available platforms currently include Illumina's GAIIx and HiSeq machines (www.illumina.com), Roche's 454 sequencer (www.454.com), Applied Biosystem's Ion Torrent or Ion Proton machines (www.appliedbiosystems.com), and SMRT sequencing system introduced by Pacific Biosciences. As of 2013, a human genome with 30X average coverage costs US$ 5–10 thousand[36], that is expected to drop further to allow its routine application in clinical settings[29].

NGS technology has allowed elucidation of reference genome for various human populations (e.g. the HapMap[37] and 1000 Genomes[38] consortiums) and for cancer cells (e.g. The Cancer Genome Atlas (http://cancergenome.nih.gov/) and The International Cancer Genome Consortium[39]).

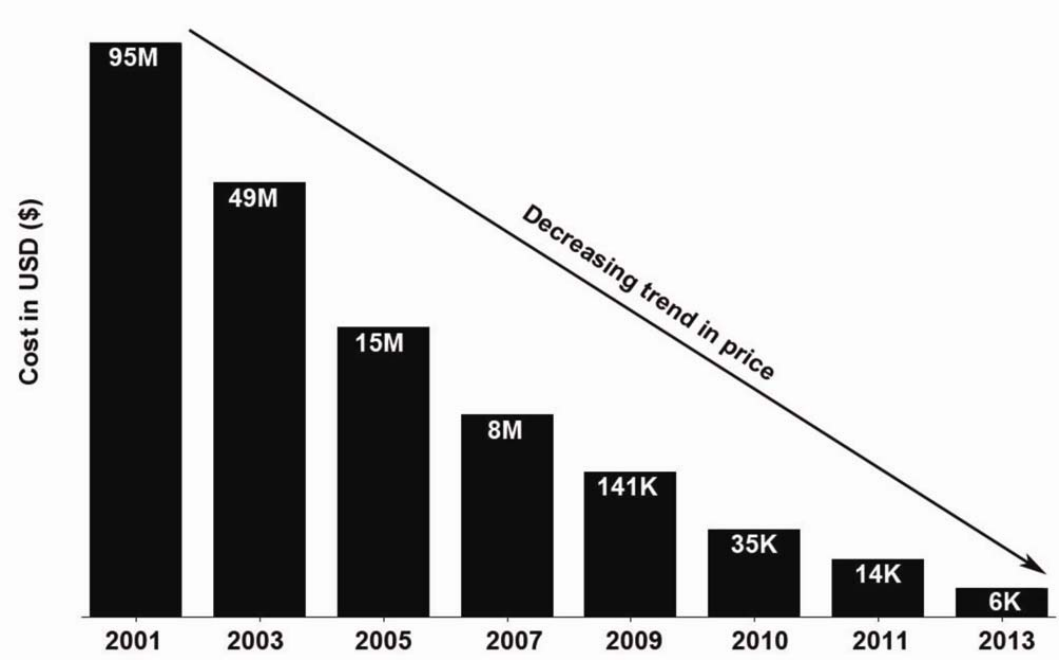

Figure 1. Descending trend of sequencing cost per human genome. (Bar graphs are not according to scale.)

These studies allow identification of variations underlying the cancer genomes, providing a comprehensive and high-quality set of common and rare polymorphisms and mutations. In fact, NGS has become a widespread tool for uses beyond sequencing of genomes[28,34] with specific applications such as whole exome sequencing (WES), which refers to sequencing of the DNA from all coding regions and other non-coding RNA sites such as miRNAs and untranslated regions (UTRs)[40,41], whole transcriptome sequencing (WTS) or RNA-Seq is used to measure gene expression, alternative splicing, allelic expression and to detect variants in expressed genes[42]. NGS has also improved existing technologies such as chromatin immune precipitation (ChIP) assays; where the bound DNA was previously hybridised to microarrays (ChIP-chip), these fragments can now be sequenced to determine the exact genomic sequence of the captured DNA and more sensitive expression measurements[28,34].

With this technology advancement, the field of cancer genomics can benefit not only from the complete human genome reference, but also from more specific individual cancer genomes. Moving ahead, third-generation sequencing, or single-molecule sequencing, currently emerging technologies will continue the drive to progress in this field (e.g. Pacific BioSciences SMRT sequencer[43], Ion Proton platform[44], GridION system from Oxford Nanopore[45]). These sequencing technologies with different features as compiled in Figure 2, enable a more precise definition of the genetic events occurring in a specific cell type since the source of DNA is from a single molecule sans PCR amplification to eliminate PCR incorporated artifacts. Currently, these technologies are not capable of sequencing an entire human genome using only one molecule, but considering the rate of development of sequencing technologies, these initiatives however look promising.

Applying next-generation sequencing to interrogate the cancer genome

Whole-genome sequencing (WGS) and WES technologies have been extremely successful in cataloguing all the different kinds of alterations of the cancer genome. The whole transcriptome approach (RNA-Seq) allows quantifying gene expression profiles and detecting the variant forms of alternative splicing, RNA editing and detection of novel fusion transcripts. Additionally, epigenetic modifications of the cancer genome, viz. DNA changes, histone methylation patterns can be determined by using Bisulfite-Seq and ChIP-seq applications in a massively parallel sequencing manner. As described above, an integrated analysis provides a high-resolution and a global view of the alterations underlying the cancer genome.

In recent years, many NGS-based efforts have underlined significant understanding of breast cancer[46–53], ovarian cancer[54], colorectal cancer[55,56], lung cancer[47], liver cancer[57], kidney cancer[27], head and neck cancer[58], melanoma[59], acute myeloid leukaemia (AML)[60,61], etc. (Table 1). As an interesting example, six studies by different groups reported their findings on a large breast cancer dataset: TCGA reported sequencing on 510 samples from 507

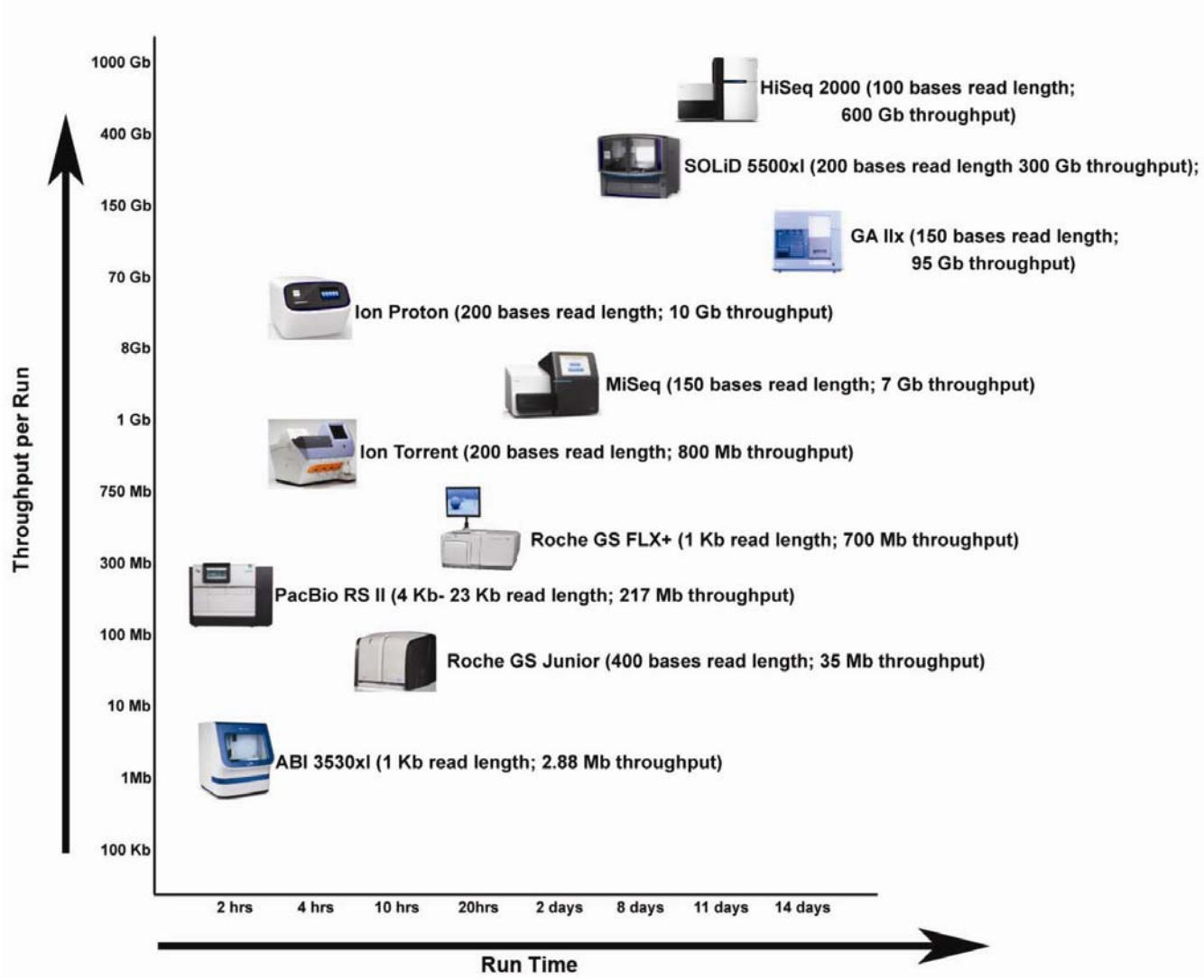

Figure 2. Advances in next-generation DNA sequencing technologies. (Axis is not according to scale.)

patients[47], As a compilation of these works, Banerji *et al.*[46] carried out exome sequencing on a set of 103 samples and WGS on 17 samples, Ellis *et al.*[48] did exome sequencing on 31 samples and WGS on 46 samples, Stephens *et al.*[53] sequenced exome of 100 samples; Shah *et al.*[51] performed WGS/WES and RNA sequencing on 65 and 80 samples of triple-negative breast cancers, and Nik-Zainal *et al.*[49] performed WGS on 21 tumour/normal pairs. Besides confirming recurrent somatic hallmark mutations in *TP53*, *GATA3* and *PIK3CA*, taken together these studies identified several other novel cancer-related mutations – mainly mutations of specific genes enriched in subtypes of breast cancers. Of note, mutations of *MAP3K1* were found to be frequently occurring inluminal A subtype[47,48].

NGS-based deep sequencing has also shed insights in distinguishing tumourigenic 'driver' mutations from their neutral 'passenger' counterparts, which occur as a result of decreased genomic stability but are not pathogenic. Several methods have already been applied to predict which missense mutations might be drivers, including CHASM[62], CanPredict[63], MutPred[64], KinaseSVM[65], SIFT[66], PolyPhen[67], MutationTaster[68] and MutationAssessor[69]. A recent study detailing exome sequencing of 72 colon tumour–normal pairs identified a magnanimous 36,303 missense somatic mutations. However, statistical analysis for significantly mutated genes using tools mentioned above led to only 23 candidates that included expected cancer genes such as *KRAS*, *TP53* and *PIK3CA* and novel genes such as *ATM*, which regulate the cell cycle checkpoint[56]. In another exome-based study, 224 lung tumour and normal pairs were sequenced. The study led to the identification of 15 highly mutated genes in the hypermutated cancers and 17 in the non-hypermutated cancers. Among the non-hypermutated cancers, novel frequent mutations in *SOX9*, *ARID1A*, *ATM* and *FAM123B* were detected besides the known *APC*, *TP53* and *KRAS* hallmark mutations. The downstream analysis of the mutations and functional roles of *SOX9*, *ARID1A*,

Table 1. Compilation of varying number of cancer genomes sequenced

Type of cancer	Exome	Transcriptome	Genome	Reference
Acute myeloid leukaemia			8	60
Acute myeloid leukaemia			24	61
Breast cancer	65	80	65	51
Breast cancer	510			47
Breast cancer	31		46	48
Breast cancer	103		17	46
Breast cancer	100			53
Breast cancer			21	60
Colon and rectal	224		97	47
Colon cancer	72	68	2	56
Head and neck cancer	32			58
Head and neck cancer	74			77
Hepatocellular	1		1	57
Melanoma			25	59
Ovarian carcinoma	316			54
Renal carcinoma	30			27
Squamous cell lung cancer	178	178	19	55

ATM and *FAM123B* suggested they are highly potential colorectal cancer-related genes. Non-hypermutated colon and rectum cancers were found to have similar patterns in genomic alternation. Further, whole genome sequencing of 97 tumours with matched normal samples identified the recurrent *NAV2-TCF7L1* fusion[47].

The future of next-generation sequencing in cancer research: technological limitation, biological noise and clinical utility of the data

Broadly, there are two major limitations to the application of NGS technology in cancer research: First, as the cost of sequencing drops, and with increasing yield output from diverse sequencing platforms, hundreds of millions or billions of reads generated are in order of hundreds of gigabytes in size, posing a huge challenge to store the high-throughput data. This problem gets further aggravated when dealing with tumour genome, where twice the amount of data need to be generated to sequence paired normal for each tumour sample to provide a baseline for subtraction to establish the somatic nature of underlying alterations in the tumour sample. Secondly, the reads from next-generation sequencers are much shorter than the ones from Sanger methods in length with relatively higher sequencing errors in the read, that vary depending on the choice of the NGS platform as shown in Figure 2. These short reads obtained from NGS platforms pose a major computational challenge for reference-independent *de novo* assembly for individual samples, that is desirable to understand the underlying genetic alterations in absolute rather than in comparison to healthy human reference genome. Powerful bioinformatics tools can partially resolve the constraints posed due to the computational infrastructure within an individual laboratory setting to perform such *de novo* assembly; a compromised routine analysis of the tumour genome with respect to healthy human reference genome is here to stay till an another leap in technological advancement with higher density of longer sequence reads is attained.

Besides the above stated technological and analytical limitations, mere identifying and enumerating the frequency of particular gene mutations while providing an essential list of genes implicated in cancer, all the same, does not provide an insight into the function of these genes. As such efforts proceed, the next key challenge is to determine which of the myriad genes implicated in such discovery efforts truly contribute to cancer initiation, progression, tumour maintenance and/or metastasis. Unfortunately, determining the functional role of a candidate oncogene or tumour suppressor gene has traditionally required labour-intensive, gene-specific approaches. As the number of genes that are found to be mutated and altered in a significant number of particular cancers increases, effective translation of cancer genome data, such as those derived from TCGA, into tangible clinical endpoints in a timely fashion demands that the cancer research community prioritize and focus its resources on the candidates with the highest potential for clinical relevance. Coupling these discoveries in cancer genome sequencing with systemic functional genomics screening would allow validation of novel therapeutic vulnerabilities in a high-throughput manner[70,71]. Recent dramatic technological advancements have enhanced our ability to characterize human cancer genome sequence and structure; this combined with the availability of a completed human reference genome and a rapidly expanding database of normal human genetic variation, have resulted in tremendous acceleration in the rate of discovery of new cancer gene targets[72,73], also transforming cancer diagno-

sis, clinical trial designing and treatment[74–76]. However, optimal exploitation of the clinical information contained a cancer genome will require comprehensive integration of these genomics alterations, clinical information of patients and precise interpretation by clinicians in the context of the disease. Indeed, the goal of compiling a complete catalogue of cancer-causing somatic mutations now appears feasible.

Box 1. First study describing landscape of somatic mutations in head and neck cancer, from India.

Recently, a comprehensive genomic characterization effort led by researchers from the National Institute of Biomedical Genomics and Advanced Centre for Training Research and Education in Cancer, Tata Memorial Centre, was carried out to study the Gingivo-buccal oral squamous cell carcinoma (OSCC-GB), as part of a larger international initiative – the International Cancer Genomics Consortium (ICGC). This study defines the first mutational landscape of Indian oral cancer patients using massively parallel whole-exome sequencing using NGS and copy number analysis using SNP array approach from DNA of blood and tumour samples across 110 patients[78]. Along with identification of hallmarks gene known to be frequently mutated in HNSCC such as: *TP53*, *FAT1*, *CASP8*, *HRAS* and *NOTCH1*, they have also discovered some significantly and frequently altered gene (*USP9X*, *MLL4*, *ARID2*, *UNC13C* and *TRPM3*) specific to OSCC-GB in 10–20% of the patient samples of Indian ethnicity.

1. Hanahan, D. and Weinberg, R. A., The hallmarks of cancer. *Cell*, 2000, **100**(1), 57–70.
2. Knudson Jr, A. G., Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA*, 1971, **68**(4), 820–823.
3. Friend, S. H. *et al.*, A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 1986, **323**(6089), 643–646.
4. Golub, T. R. *et al*., Fusion of PDGF receptor beta to a novel ets-like gene, *tel*, in chronic myelomonocytic leukemia with t(5;12) chromosomal translocation. *Cell*, 1994, **77**(2), 307–316.
5. Nunez, G. *et al*., Growth- and tumor-promoting effects of deregulated BCL2 in human B-lymphoblastoid cells. *Proc. Natl. Acad. Sci. USA*, 1989, **86**(12), 4589–4593.
6. Weinstein, I. B., Cancer. Addiction to oncogenes – the Achilles heal of cancer. *Science*, 2002, **297**(5578), 63–64.
7. Apperley, J. F. *et al*., Response to imatinib mesylate in patients with chronic myeloproliferative diseases with rearrangements of the platelet-derived growth factor receptor beta. *N. Engl. J. Med.*, 2002, **347**(7), 481–487.
8. Cools, J. *et al*., A tyrosine kinase created by fusion of the *PDGFRA* and *FIP1L1* genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. *N. Engl. J. Med.*, 2003, **348**(13), 1201–1214.
9. Heinrich, M. C. *et al*., Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J. Clin. Oncol.*, 2003, **21**(23), 4342–4349.
10. Rubin, B. P. *et al.*, Molecular targeting of platelet-derived growth factor B by imatinib mesylate in a patient with metastatic dermatofibrosarcoma protuberans. *J. Clin. Oncol.*, 2002, **20**(17), 3586–3591.
11. Druker, B. J. *et al.*, Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.*, 2001, **344**(14), 1031–1037.
12. Demetri, G. D., Targeting c-kit mutations in solid tumors: scientific rationale and novel therapeutic options. *Semin. Oncol.* (*Suppl. 17*), 2001, **28**(5), 19–26.
13. Tuveson, D. A. *et al.*, STI571 inactivation of the gastrointestinal stromal tumor c-KIT oncoprotein: biological and clinical implications. *Oncogene*, 2001, **20**(36), 5054–5058.
14. Pegram, M. D. *et al*., Phase II study of receptor-enhanced chemosensitivity using recombinant humanized anti-p185HER2/neu monoclonal antibody plus cisplatin in patients with HER2/neu-overexpressing metastatic breast cancer refractory to chemotherapy treatment. *J. Clin. Oncol.*, 1998, **16**(8), 2659–2671.
15. Lynch, T. J. *et al*., Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.*, 2004, **350**(21), 2129–2139.
16. Paez, J. G. *et al.*, EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 2004, **304**(5676), 1497–1500.
17. Pao, W. *et al*., EGF receptor gene mutations are common in lung cancers from 'never smokers' and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl. Acad. Sci. USA*, 2004, **101**(36), 13306–13311.
18. Soda, M. *et al*., Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, 2007, **448**(7153), 561–566.
19. Kwak, E. L. *et al*., Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med*., 2010, **363**(18), 1693–1703.
20. Marusyk, A. and Polyak, K., Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta*, 2010, **1805**(1), 105–117.
21. Merlo, L. M. *et al*., Cancer as an evolutionary and ecological process. *Nature Rev. Cancer*, 2006, **6**(12), 924–935.
22. Liu, E. T., Functional genomics of cancer. *Curr. Opin. Genet. Dev.*, 2008, **18**(3), 251–256.
23. Almendro, V. and Fuster, G., Heterogeneity of breast cancer: etiology and clinical relevance. *Clin. Transl. Oncol.*, 2011, **13**(11), 767–773.
24. Curtis, C. *et al.*, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 2012, **486**(7403), 346–352.
25. Russnes, H. G. *et al*., Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest.*, 2011, **121**(10), 3810–3818.
26. Yancovitz, M. *et al.*, Intra- and inter-tumor heterogeneity of BRAF (V600E))mutations in primary and metastatic melanoma. *PLoS ONE*, 2012, **7**(1), e29336.

27. Gerlinger, M. *et al*., Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med*., 2012, **366**(10), 883–892.
28. Shendure, J. and Ji, H., Next-generation DNA sequencing. *Nature Biotechnol.*, 2008, **26**(10), 1135–1145.
29. Metzker, M. L., Sequencing technologies – the next generation. *Nature Rev. Genet.*, 2010, **11**(1), 31–46.
30. Moore, G. E., Cramming more components onto integrated circuits. *Proc. IEEE*, 1998, **86**(2), 82–85.
31. Lunshof, J. E. *et al.*, Personal genomes in progress: from the human genome project to the personal genome project. *Dialogues Clin. Neurosci*., 2010, **12**(1), 47–60.
32. Meyerson, M., Gabriel, S. and Getz, G., Advances in understanding cancer genomes through second-generation sequencing. *Nature Rev. Genet.*, 2010, **11**(10), 685–696.
33. Hutchison 3rd, C. A., DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res*., 2007, **35**(18), 6227–6237.
34. Mardis, E. R., The impact of next-generation sequencing technology on genetics. *Trends Genet*., 2008, **24**(3), 133–141.
35. Reis-Filho, J. S., Next-generation sequencing. *Breast Cancer Res*. (*Suppl. 3*), 2009, **11**, S12.
36. Mamanova, L. *et al.*, Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 2010, **7**(2), 111–118.
37. International HapMap, C., The International HapMap Project. *Nature*, 2003, **426**(6968), 789–796.
38. Genomes Project, A map of human genome variation from population-scale sequencing. *Nature*, 2010, **467**(7319), 1061–1073.
39. International Cancer Genome, International network of cancer genome projects. *Nature*, 2010, **464**(7291), 993–998.
40. Coffey, A. J. *et al.*, The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.*, 2011, **19**(7), 827–831.
41. Ng, S. B. *et al*., Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 2009, **461**(7261), 272–276.
42. Wang, Z., Gerstein, M. and Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.*, 2009, **10**(1), 57–63.
43. Eid, J. *et al*., Real-time DNA sequencing from single polymerase molecules. *Science*, 2009, **323**(5910), 133–138.
44. Boland, J. F. *et al*., The new sequencer on the block: comparison of Life Technology's proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum. Genet*, 2013, **132**(10), 1153–1163.
45. Eisenstein, M., Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnol.*, 2012, **30**(4), 295–296.
46. Banerji, S. *et al.*, Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 2012, **486**(7403), 405–409.
47. Cancer Genome Atlas Research, Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 2012, **489**(7417), 519–525.
48. Ellis, M. J. *et al.*, Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 2012, **486**(7403), 353–360.
49. Nik-Zainal, S. *et al*., Mutational processes molding the genomes of 21 breast cancers. *Cell*, 2012, **149**(5), 979–993.
50. Nik-Zainal, S. *et al*., The life history of 21 breast cancers. *Cell*, 2012, **149**(5), 994–1007.
51. Shah, S. P. *et al*., The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 2012, **486**(7403), 395–399.
52. Stephens, P. J. *et al.*, Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 2009, **462**(7276), 1005–1010.
53. Stephens, P. J. *et al.*, The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 2012, **486**(7403), 400–404.
54. Cancer Genome Atlas Research, Integrated genomic analyses of ovarian carcinoma. *Nature*, 2011, **474**(7353), 609–615.
55. Cancer Genome Atlas, Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 2012, **487**(7407), 330–337.
56. Seshagiri, S. *et al*., Recurrent R-spondin fusions in colon cancer. *Nature*, 2012, **488**(7413), 660–664.
57. Totoki, Y. *et al*., High-resolution characterization of a hepatocellular carcinoma genome. *Nature Genet*., 2011, **43**(5), 464–469.
58. Agrawal, N. *et al*., Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science*, 2011, **333**(6046), 1154–1157.
59. Berger, M. F. *et al*., Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature*, 2012, **485**(7399), 502–506.
60. Ding, L. *et al.*, Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 2012, **481**(7382), 506–510.
61. Welch, J. S. *et al*., The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 2012, **150**(2), 264–278.
62. Wong, W. C. *et al*., CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, 2011, **27**(15), 2147–2148.
63. Kaminker, J. S. *et al*., CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res*., 2007, **35**(Web Server issue), W595–W598.
64. Li, B. *et al*., Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 2009, **25**(21), 2744–2750.
65. Torkamani, A. and Schork, N. J., Prediction of cancer driver mutations in protein kinases. *Cancer Res*., 2008, **68**(6), 1675–1682.
66. Ng, P. C. and Henikoff, S., SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*., 2003, **31**(13), 3812–3814.
67. Adzhubei, I. A. *et al.*, A method and server for predicting damaging missense mutations. *Nature Methods*, 2010, **7**(4), 248–249.
68. Schwarz, J. M. *et al.*, MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, 2010, **7**(8), 575–576.
69. Reva, B., Antipin, Y. and Sander, C., Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol*., 2007, **8**(11), R232.
70. Boehm, J. S. and Hahn, W. C., Towards systematic functional characterization of cancer genomes. *Nature Rev. Genet.*, 2011, **12**(7), 487–498.
71. Cheung, H. W. *et al.*, Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. USA*, 2011, **108**(30), 12372–12377.
72. Kallioniemi, A., CGH microarrays and cancer. *Curr. Opin. Biotechnol.*, 2008, **19**(1), 36–40.
73. Stratton, M. R., Campbell, P. J. and Futreal, P. A., The cancer genome. *Nature*, 2009, **458**(7239), 719–724.
74. Macconaill, L. E. and Garraway, L. A., Clinical implications of the cancer genome. *J. Clin. Oncol.*, 2010, **28**(35), 5219–5228.
75. Mestan, K. K. *et al*., Genomic sequencing in clinical trials. *J. Transl. Med*., 2011, **9**, 222.
76. McDermott, U., Downing, J. R. and Stratton, M. R., Genomics and the continuum of cancer care. *N. Engl. J. Med.*, 2011, **364**(4), 340–350.

77. Stransky, N. *et al.*, The mutational landscape of head and neck squamous cell carcinoma. *Science*, 2011, **333**(6046), 1157–1160.

78. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. *Nature Commun.*, 2013, **4**, 2873.

ACKNOWLEDGEMENTS. A.D. is supported by an Intermediate fellowship from the Wellcome Trust/DBT India Alliance (IA/I/11/2500278), by a grant from DBT (BT/PR2372/AGR/36/696/2011), and intramural grants (Seed-In-Air 2897, TMH Plan Project 2712 and IRB 92). P.U. is supported by an fellowship from CSIR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.