APPLICATIONS OF NEXT-GENERATION SEQUENCING

Advances in understanding cancer genomes through second-generation sequencing

Matthew Meyerson, Stacey Gabriel and Gad Getz

Abstract | Cancers are caused by the accumulation of genomic alterations. Therefore, analyses of cancer genome sequences and structures provide insights for understanding cancer biology, diagnosis and therapy. The application of second-generation DNA sequencing technologies (also known as next-generation sequencing) — through whole-genome, whole-exome and whole-transcriptome approaches — is allowing substantial advances in cancer genomics. These methods are facilitating an increase in the efficiency and resolution of detection of each of the principal types of somatic cancer genome alterations, including nucleotide substitutions, small insertions and deletions, copy number alterations, chromosomal rearrangements and microbial infections. This Review focuses on the methodological considerations for characterizing somatic genome alterations in cancer and the future prospects for these approaches.

Second-generation sequencing
Used in this Review to refer to sequencing methods that have emerged since 2005 that parallelize the sequencing process and produce millions of typically short sequence reads (50–400 bases) from amplified DNA clones. It is also often known as next-generation sequencing.

Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. Correspondence to M.M. email: matthew_meyerson@dfci.harvard.edu
doi:10.1038/nrg2841

A major near-term medical impact of the genome technology revolution will be the elucidation of mechanisms of cancer pathogenesis, leading to improvements in the diagnosis of cancer and the selection of cancer treatment. Thanks to second-generation sequencing technologies[1–5], recently it has become feasible to sequence the expressed genes ('transcriptomes')[6,7], known exons ('exomes')[8,9], and complete genomes[10–15] of cancer samples.

These technological advances are important for advancing our understanding of malignant neoplasms because cancer is fundamentally a disease of the genome. A wide range of genomic alterations — including point mutations, copy number changes and rearrangements — can lead to the development of cancer. Most of these alterations are somatic, that is, they are present in cancer cells but not in a patient's germ line[16].

An impetus for studies of somatic genome alterations, which are the focus of this Review, is the potential for therapies targeted against the products of these alterations. For example, treatment with the inhibitors of the epidermal growth factor receptor kinase (EGFR), gefitinib and erlotinib, leads to a significant survival benefit in patients with lung cancer whose tumours carry *EGFR* mutations, but no benefit in patients whose tumours carry wild-type *EGFR*[17–19]. Therefore, comprehensive genome-based diagnosis of cancer is becoming increasingly crucial for therapeutic decisions.

During the past decades, there have been major advances in experimental and informatic methods for genome characterization based on DNA and RNA microarrays and on capillary-based DNA sequencing ('first-generation sequencing', also known as Sanger sequencing). These technologies provided the ability to analyse exonic mutations and copy number alterations and have led to the discovery of many important alterations in the cancer genome[20].

However, there are particular challenges for the detection and diagnosis of cancer genome alterations. For example, some genomic alterations in cancer are prevalent at a low frequency in clinical samples, often owing to substantial admixture with non-malignant cells. Second-generation sequencing can solve such problems[21]. Furthermore, these new sequencing methods make it feasible to discover novel chromosomal rearrangements[22] and microbial infections[23–25] and to resolve copy number alterations at very high resolution[22,26].

At the same time, the avalanche of data from second-generation sequencing provides a statistical and computational challenge: how to separate the 'wheat' of causative alterations from the 'chaff' of noise caused by alterations in the unstable and evolving cancer genome.

This challenge is likely to be solved in part by systematic analyses of large cancer genome data sets that will provide sufficient statistical power to overcome experimental and biological noise[27,28].

In this Review, we discuss the key challenges in cancer genome sequencing, the methods that are currently available and their relative values for detecting different types of genomic alteration. We then summarize the main points to consider in the computational analysis of cancer genome sequencing data and comment on the future potential for using genomics in cancer diagnosis. Cancer genome sequencing is a rapidly moving field, so in this Review we aim to set out the principles and important methodological considerations, with a brief summary of important findings to date.

Cancer-specific considerations

Cancer samples and cancer genomes have general characteristics that are distinct from other tissue samples and from genomic sequences that are inherited through the germ line. These require particular consideration in second-generation sequencing analyses.

Characteristics of cancer samples for genomic analysis. Cancer samples differ in their quantity, quality and purity from the peripheral blood samples that are used for germline genome analysis. Surgical resection specimens tend to be large and have been the mainstay of cancer genome analysis. However, diagnostic biopsies from patients with disseminated disease tend to contain few cells — as surgical cure is not possible in these cases, minimizing biopsy size is a safety consideration. Therefore, the quantity of nucleic acids available may be limiting; obtaining sequence information from such biopsies will require decreasing the minimum inputs for second-generation sequencing. An alternative approach to sequencing from small samples is whole-genome amplification, but this method does not preserve genome structure and can give rise to artefactual nucleotide sequence alterations[29].

Nucleic acids from cancer are also often of lower quality than those purified from peripheral blood. One reason for this is technical: most cancer biopsy and resection specimens are formalin-fixed and paraffin-embedded (FFPE) to optimize the resolution of microscopic histology. Nucleic acids from FFPE specimens are likely to have undergone crosslinking and also may be degraded[30]. Second-generation sequence analysis of FFPE-derived nucleic acids can require special experimental [31] and computational methods to handle an increased background mutation rate[32,33]. A second reason for this difference in nucleic acid quality is biological: cancer specimens often include substantial fractions of necrotic or apoptotic cells that reduce the average nucleic acid quality, therefore, experimental methods should also be adapted to account for this. The many-fold coverage made possible by second-generation sequencing, however, can allow high-quality data to be produced from lower quality samples[21].

Finally, cancer nucleic acid specimens are less pure than specimens used to analyse the inherited genome, especially in terms of genomic DNA purity. The samples generally used for germline genome analysis — peripheral blood mononuclear cells — are known to be heterogeneous only at the rearranged immunoglobulin and T cell receptor loci in a subset of cells. By contrast, a cancer specimen contains a mixture of malignant and non-malignant cells and, therefore, a mixture of cancer and normal genomes (and transcriptomes). Furthermore, the cancers themselves may be highly heterogeneous and composed of different clones that have different genomes[34]. Cancer genome analytical models must take these two types of heterogeneity (cancer versus normal heterogeneity and within-cancer heterogeneity) into account in their prediction of genome alterations.

Structural variability of cancer genomes. Cancer genomes are enormously diverse and complex. They vary substantially in their sequence and structure compared to normal genomes and among themselves. To paraphrase Leo Tolstoy's famous first line from *Anna Karenina*: normal human genomes are all alike, but every cancer genome is abnormal in its own way.

Specifically, cancer genomes vary considerably in their mutation frequency (degree of variation compared to the reference sequence), in global copy number or ploidy, and in genome structure. These variations have several implications for cancer genome analysis: the presence of a somatic mutation is not enough to establish statistical significance as it must be evaluated in terms of the sample-specific background mutation rate, which can vary at different types of nucleotides (discussed further below). The analysis of mutations must also be adjusted for the ploidy and the purity of each sample and the copy number at each region. For example, if 50% of the tumour DNA is derived from cancer cells and a mutation is present on 1 of 4 copies of chromosome 11, the frequency of that mutation will be 12.5% in the sample. Similar considerations apply to the detection of somatic rearrangements.

To identify somatic alterations in cancer, comparison with matched normal DNA from the same individual is essential. This is largely owing to our incomplete knowledge of the variations in the normal human genome; to date, each 'matched normal' cancer genome sequence has identified large numbers of mutations and rearrangements in the germ line that had not been previously described[11–15,35]. In the future, the complete characterization of many thousands of normal human genomes may obviate this need for a matched normal sample.

Experimental approaches

Second-generation sequencing technologies are based on the simultaneous detection of nucleotides in arrayed amplified DNA products originating from single DNA molecules[36]. Specific methods include picotitre-plate pyrosequencing[3,5], single-nucleotide fluorescent base extension with reversible terminators[1] and ligation-based sequencing[2,4]. Thanks to advances in sequencing approaches that include these technologies, the number of bases that can be sequenced for a given cost has increased one millionfold since 1990, more than doubling every year, which is twice as fast as Moore's law for semiconductors[37].

First-generation sequencing
(also known as Sanger sequencing or capillary sequencing). The standard sequencing methodology used to sequence the reference human (and other model organism) genomes. It uses radioactively or fluorescently labelled dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. Various detection methods allow read-out of sequence according to the incorporation of each specific terminator (ddATP, ddCTP, ddGTP or ddTTP).

Whole-genome amplification
Various molecular techniques (including multiple displacement amplification, rolling circle amplification or degenerate oligonucleotide primed PCR) in which very small amounts (nanograms) of a genomic DNA sample can be multiplied in a largely unbiased fashion to produce suitable quantities for genomic analysis (micrograms).

Moore's law
The observation made in 1965 by Gordon Moore that the number of transistors per square inch on integrated circuits had doubled every other year since the integrated circuit was invented.

Table 1 | **Whole-genome sequencing studies of cancer**

Study	Method	Cancer type	Number of samples sequenced	Aberration type	Refs
Ley *et al.*, 2008	Deep single-end whole-genome sequencing	AML	1	Point mutations, insertions, deletions	10
Campbell *et al.*, 2008	Shallow paired-end whole-genome sequencing	Lung	2	Deletions, amplifications, tandem duplications, interchromosomal rearrangements	22
Stephens *et al.*, 2009	Shallow paired-end whole-genome sequencing	Breast	24	Deletions, amplifications, tandem duplications, interchromosomal rearrangements, inversions	39
Pleasance *et al.*, 2010	Deep paired-end whole-genome sequencing	Melanoma	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements	12
Pleasance *et al.*, 2010	Deep paired-end whole-genome sequencing	Small-cell lung	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements	13
Mardis *et al.*, 2009	Deep paired-end whole-genome sequencing	AML	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements	11
Shah *et al.*, 2009	Deep paired-end whole-genome sequencing	Breast	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements	15
Ding *et al.*, 2010	Deep paired-end whole-genome sequencing	Breast	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements, inversions	35
Lee *et al.*, 2010	Deep paired-end whole-genome sequencing	Lung	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements, inversions	14

AML, acute myelogenous leukaemia.

The application of second-generation sequencing has allowed cancer genomics to move from focused approaches — such as single-gene sequencing and array analysis — to comprehensive genome-wide approaches. Second-generation sequencing can be applied to cancer samples in various ways. These vary by the type of input material (for example, DNA, RNA or chromatin), the proportion of the genome targeted (the whole genome, transcriptome or a subset of genes) and the type of variation studied (structural change, point mutation, gene expression or chromosomal conformation). In this section, we briefly introduce the main approaches to second-generation sequencing of cancer and their associated experimental considerations. Chromatin immunoprecipitation followed by sequencing (ChIP–seq) is an important complement to cancer genomics but is not discussed as it has been reviewed elsewhere[38]. Key whole-genome sequencing studies to date are summarized in TABLE 1.

Compared with previous sequencing methods, which are analogue, second-generation sequencing is digital: it is possible to count alleles at any nucleotide or reads at any alignable position in the genome. Its digital nature gives rise to one of the key features of second-generation sequencing, the ability to over-sample the genome or other nucleic acid compartment that is targeted[10]. Over-sampling provides highly accurate sequence information by providing enough signal to overcome experimental noise and also allows detection of mutations and other genome alterations in heterogeneous samples such as cancer tissues.

Whole-genome sequencing. The first whole cancer genome sequence was reported in 2008, a description of the nucleotide sequence of DNA from an acute myeloid leukaemia compared with DNA from normal skin from the same patient[10]. Since then, six more complete sequences of cancer genomes together with matched normal genomes have been reported[11–15,35], and this number will grow rapidly.

Complete sequencing of the genome of cancer tissue to high redundancy, using germline DNA sequence from the same patient as a comparison, has the power to discover the full range of genomic alterations — including nucleotide substitutions, structural rearrangements, and copy number alterations — using a single approach[10–15,35]. Therefore, whole-genome sequencing provides the most comprehensive characterization of the cancer genome but, as it requires the greatest amount of sequencing, it is the most costly. Alternative, lower-cost approaches include shotgun sequencing with incomplete coverage (for example, less than 30-fold coverage; see below) — which is sufficient to identify somatic rearrangements in the genome[22,39] and copy number alterations[22,26] — and exome and transcriptome sequencing, which are described below.

The major potential of whole-genome sequencing for cancer is the discovery of chromosomal rearrangements. Previously, there were no systematic approaches to study solid tumours that have complex karyotypes. Therefore, until recently it was thought that chromosomal translocations were rare in epithelial tumours and found only in haematological malignancies in which they could be observed with cytogenetic methods[40,41]. However, the discoveries of the transmembrane protease serine 2 (*TMPRSS2*)–*ERG* translocations in prostate carcinoma[42] and the echinoderm microtubule-associated protein like 4 (*EML4*)–anaplastic lymphoma receptor tyrosine kinase (*ALK*) translocations in non-small cell lung carcinoma[43] have changed that view.

Chromatin immunoprecipitation
A technique used to identify the location of DNA-binding proteins and epigenetic marks in the genome. Genomic sequences containing the protein of interest are enriched by binding soluble DNA chromatin extracts (complexes of DNA and protein) to an antibody that recognizes the protein or modification.

Over-sampling
Reading the same stretch of DNA sequence many times to gain a confident sequence read-out.

Shotgun sequencing
Sequencing randomly derived fragments of the whole genome. The order and orientation of the sequences are determined by mapping individual reads back to a reference or through assembly of overlapping sequences into larger contigs of sequence.

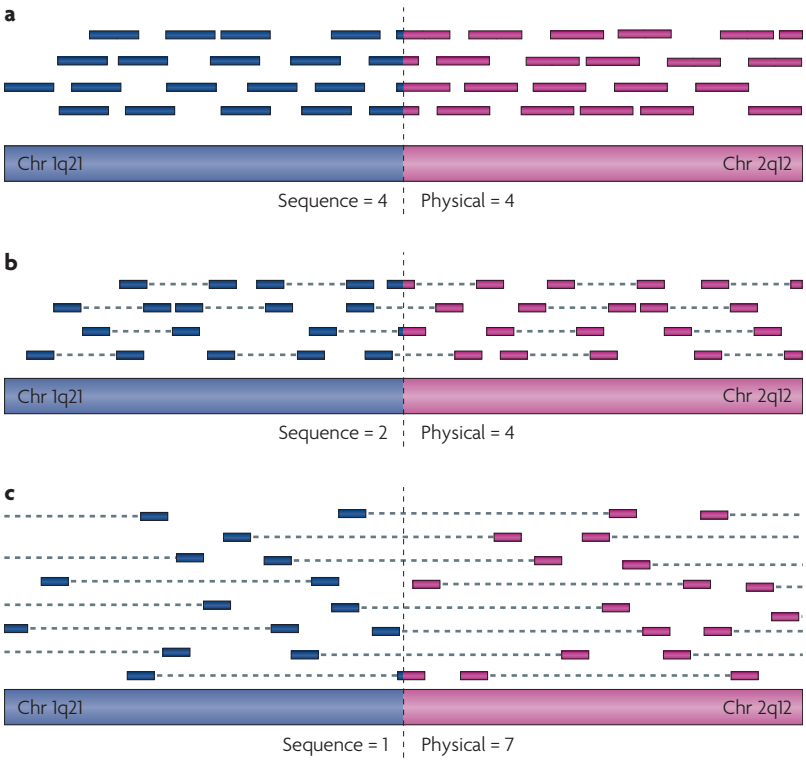

Figure 1 | **Depth of coverage and physical coverage.** To illustrate considerations regarding depth of coverage and physical coverage, a rearrangement between human chromosome 1q21 and chromosome 2q12 is shown. Sequenced DNA fragments are represented by coloured bars: single-end sequencing is shown in **a**; paired-end sequencing is shown in **b** and **c**, in which the bars and the dashed lines indicate the sequenced ends and unsequenced part, respectively. Blue bars map to chromosome 1 and purple bars to chromosome 2. Three different scenarios (**a**–**c**) are depicted that vary in the length of the DNA fragments that are sequenced. In each scenario, the sequence and physical coverage at the rearrangement site is shown below. Sequence coverage represents the number of sequenced reads that cover the site; this affects the ability to detect point mutations. Physical coverage measures the number of fragments that span the site; this affects the ability to detect the rearrangement, based on paired reads that map to different chromosomes. In cases in which the entire fragment is sequenced, as in **a**, the sequence and physical coverage are the same.

In addition to rearrangements between unique, alignable sequences, whole-genome sequencing may be able to detect other types of genomic alterations that have not been observable using previous methods. Among the most important of such events are somatic mutations of non-coding regions, including promoters, enhancers, introns and non-coding RNAs (including microRNAs), as well as unannotated regions. Other novel types of alterations in cancer may include rearrangements of repetitive elements, and recent studies have suggested that active retrotransposons in the human genome might contribute to cancer, so whole-genome sequencing would be informative in this regard[44,45].

Two important issues to consider when planning whole-genome sequencing experiments are depth of coverage and physical coverage. Sequence depth is measured by the amount of over-sampling: typically, to detect nucleotide alterations with high sensitivity, the 3 billion bases of the human genome are covered at least 30-fold on average, requiring the generation of 90 billion bases of sequence data per sample[10–15,35]. For cancer samples, this number needs to be increased to account for the decreased purity and often increased ploidy of each sample.

Physical coverage is important for detecting rearrangements and this detection is aided by analysis of 'paired reads'. In standard shotgun library methods, the fragments of DNA are typically 200–400 bases long, and second-generation sequencing technologies currently yield 50–100 base reads from each end of a fragment (known as paired reads). The expected distance between the paired reads is used to uniquely place the reads on the reference genome and unexpected read pairing can be used to detect structural anomalies.

The distance between the paired reads can be increased to thousands of bases by the creation of jumping libraries, which can be constructed by generating large circular fragments of DNA[4,13]. This leads to higher physical coverage of the genome with less sequence coverage and, consequently, lower cost. For example, with 3 kb spacing between pairs, the physical coverage of the genome is 10 times higher than with 300 bp inserts, so equivalent physical coverage can be obtained with 10 times less sequence coverage (FIG. 1). Although powerful for the detection of structural rearrangements, the jumping library approach has two main limitations. First, with less total sequence, the coverage at any given position is lower, therefore the sensitivity to observe base changes such as point mutations is correspondingly lower. Second, the jumping library approach requires large quantities of high-quality input DNA, which may not be possible with all clinical cancer samples, especially those derived from FFPE specimens.

Exome sequencing. Targeted sequencing approaches have the general advantage of increased sequence coverage of regions of interest — such as coding exons of genes — at lower cost and higher throughput compared with random shotgun sequencing, Most large-scale methods for targeted sequencing use a variation of a hybrid selection approach (FIG. 2): nucleic acid 'baits' are used to 'fish' for regions of interest in the total pool of nucleic acids, which can be DNA[46–49] or RNA[50]. Any subset of the genome can be targeted, including exons, non-coding RNAs, highly conserved regions of the genome or other regions of interest.

Analysis of selected sets of exons using capillary-based sequencing has been a powerful and effective approach to focus DNA sequencing efforts on the coding genes of greatest interest. For example, capillary sequencing of exons from specific gene families has led to the discovery of activating somatic mutations in various cancers, such as the *BRAF* serine–threonine kinase[51], the *EGFR*, *ERBB2*, fibroblast growth factor receptor 2 (*FGFR2*), *JAK2,* and *ALK* receptor tyrosine kinases[52–66], and the *PIK3CA* and *PIK3R1* lipid kinase subunits[28,67]. Whole-exome sequencing with capillary sequencing allowed the analysis of all known coding genes in colorectal, breast and pancreatic carcinomas and glioblastoma[68–71]. These studies have led to the discovery of somatic mutations in isocitrate dehydrogenase 1 (*IDH1*) in glioblastoma[69] and of germline mutations in the gene encoding partner and

Jumping library
A method of library construction in which the genome is divided into large fragments using a rare cutter enzyme. Fragments are circularized and DNA sequences are read from the ends of the fragment, without reading the intervening sequence.

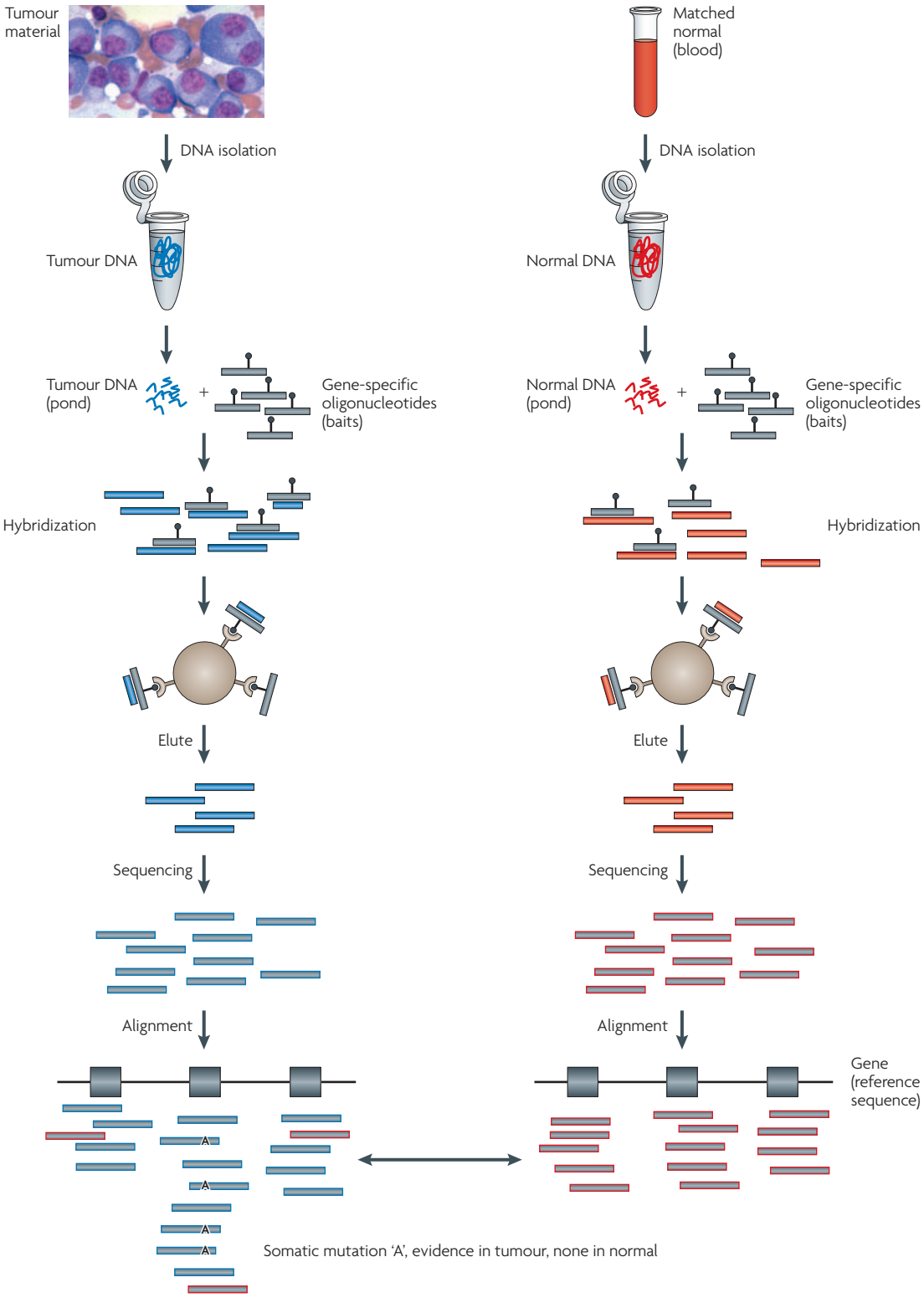

Figure 2 | **Sequence capture for cancer genomics.** A schematic diagram of hybrid selection to capture specific regions of the genome from tumour DNA (left panel, blue) and normal DNA (right panel, red). DNA from the starting material (the 'pond') is sheared and hybridized to oligonucleotides that are specific for the regions of interest (for example, exons in genes from a particular pathway or the whole exome; the 'baits'). The baits have a tag that allows them to be isolated (for example, by immobilization on beads). The captured DNA is eluted, prepared into sequencing libraries, sequenced and aligned to the bait sequences. Because this technique allows greater depth of coverage for the regions of interest, somatic mutations in the tumour DNA can be detected from admixed populations containing tumour and normal DNA-derived reads.

localizer of BRCA2 (*PALB2*) in patients with pancreatic carcinoma[72], among other important findings.

However, second-generation sequencing is a more efficient and comprehensive technology for whole-exome sequence analysis than capillary-based sequencing and is becoming increasingly routine[8,9]. Because the exome represents only approximately 1% of the genome, or about 30 Mb, vastly higher sequence coverage can be readily achieved using second-generation sequencing platforms with considerably less raw sequence and cost than whole-genome sequencing. For example, whereas 90 Gb of sequence is required to obtain 30-fold average coverage of the genome, 75-fold average coverage is achieved for the exome with only 3 Gb of sequence using the current state-of-the-art platforms for targeting[73]. However, there are inefficiencies in the targeting process. For example, uneven capture efficiency across exons can mean that not all exons are sequenced and some off-target hybridization can occur. These inefficiencies are likely to be ameliorated as sequencing and capture technology continue to improve.

The higher coverage of the exome that can be affordably achieved for a large number of samples makes exome sequencing highly suitable for mutation discovery in cancer samples of mixed purity. In addition, the hybrid selection approach will be particularly powerful for diagnostic analysis of the cancer genome; for diagnosis, there may be interest in sequencing specific oncogenes[74] and/or tumour suppressor genes at very high coverage in samples with a low percentage of tumour cells[21].

Transcriptome sequencing. Second-generation sequencing of the transcriptome (RNA-seq) — as cDNA derived from mRNA, total RNA or other RNAs such as microRNAs — is a powerful approach for understanding cancer. Transcriptome sequencing is a sensitive and efficient approach to detect intragenic fusions, including in-frame fusion events that lead to oncogene activation[6,7,75,76]. Transcriptome sequencing can also be used to detect somatic mutations but finding a matched normal sample for comparison is a challenge, as normal tissue is unlikely to express exactly the same genes as the tumour sample. Furthermore, mutation detection in genes expressed at low levels is hampered owing to lack of statistical power. Also, the possibilities of reverse transcriptase errors and RNA editing[15] need to be considered. Nevertheless, important somatic nucleotide substitution mutations have been discovered by transcriptome sequencing, most notably recurrent mutations in the forkhead box L2 gene (*FOXL2*) in ovarian granulosa cell tumours[77].

RNA-seq also allows analysis of gene expression profiles and is particularly powerful for identifying transcripts with low-level expression, which means that these transcripts can be included in tumour classification metrics[78]. RNA-seq may soon be competitive with oligonucleotide microarray technologies in terms of the cost and efficiency of gene expression analysis. Furthermore, transcriptome sequencing provides the advantage of not being limited to known genes but can also include the detection of novel transcripts, alternative splice forms and non-human transcripts.

Detecting classes of genome alterations

In contrast to previously available genome technologies, such as first-generation sequencing and array-based methods, second-generation sequencing methods can provide a comprehensive picture of the cancer genome by detecting each of the major alterations in the cancer genome (FIG. 3). Here we describe the analysis of each type of alteration briefly.

Somatic nucleotide substitutions and small insertion and deletion mutations. Nucleotide substitution mutations are the most common known somatic genomic alteration in cancer, occurring typically at the rate of about one somatic nucleotide substitution per million nucleotides[12,13,15,28,79]; insertion and deletion mutations are approximately tenfold less common in most cancer specimens. However, the rate of mutations varies greatly between cancer specimens. For example, ultraviolet radiation-induced melanomas have on the order of ten mutations per million bases[12] and hypermutated tumours with defects in DNA repair genes can reach rates of tens of mutations per million bases[28,79]. By contrast, haematopoietic malignancies can have less than one mutation per million bases[10,11]. Therefore, statistical analyses to assess mutation significance must take these sample-to-sample variations into account.

Various computational methods have been developed to determine the presence of somatic mutations using second-generation sequence data[80]. The detection of somatic mutations in cancer requires mutation calling in both the tumour DNA and the matched normal DNA, coupled with comparison to a reference genome and an assessment of the statistical significance of the number of counts of the mutation in the cancer sequence and its absence in the matched normal sequence. False-positive genome alteration calls are of two types: inaccurate detection of an event in the tumour, when the tumour and normal are both wild-type; and detection of a germline event in the tumour but failure to detect it in the normal. Different sources of noise contribute to the two types of false positives. The first type of error can be due to machine-sequencing errors, incorrect local alignment of individual reads and discordant alignment of pairs. Stochastic errors such as machine errors can be eliminated by high-level over-sampling of tumour and normal DNA sequence with sufficiently stringent statistical thresholds for mutation calling. The second type of false-positive mutation calls are caused by failures to detect the germline alleles that differ from the reference sequence in the normal sample, mostly owing to insufficient coverage.

In general, the most common cause of false-negative mutation calls is insufficient coverage of the cancer DNA. As discussed above, increased over-sampling may be required to overcome sample admixture, tumour heterogeneity and variations in ploidy (genome-wide and local).

The identification of candidate mutations associated with cancer then leads to two questions: is the specific mutation or the set of alterations in a particular gene statistically significant across all samples, and is the

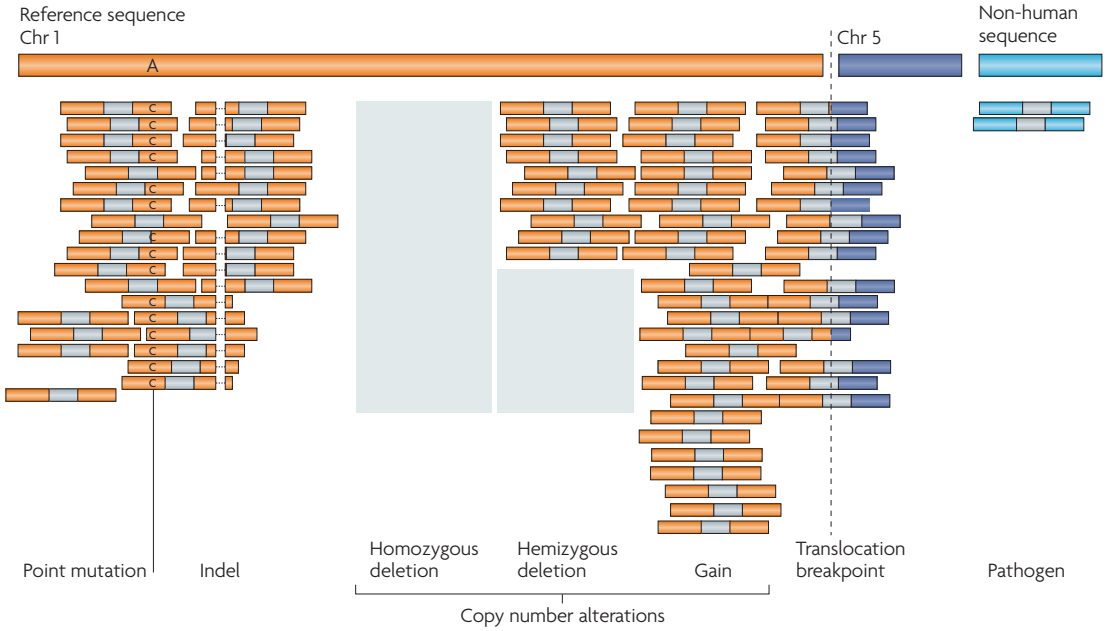

Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced fragments are depicted as bars with coloured tips representing the sequenced ends and the unsequenced portion of the fragment in grey. Reads are aligned to the reference genome (for example, mostly chromosome 1 in this example). The colours of the sequenced ends show where they align to. Different types of genomic alterations can be detected, from left to right: point mutations (in this example A to C) and small insertions and deletions (indels) (in this example a deletion shown by a dashed line) are detected by identifying multiple reads that show non-reference sequence; changes in sequencing depth (relative to a normal control) are used to identify copy number changes (shaded boxes represent absent or decreased reads in the tumour sample); paired-ends that map to different genomic loci (in this case, chromosome 5) are evidence of rearrangements; and sequences that map to non-human sequences are evidence for the potential presence of genomic material from pathogens.

alteration functionally significant? Across a set of samples, statistical significance of a gene (or an alteration) can be assessed by comparison to the sample-specific background mutation rates in the specific nucleotide context (for example, background rate of C-to-T transitions in CG dinucleotides often differs from the rate of mutations in A nucleotides[27,28]) and correcting for multiple hypotheses testing (that is, the higher chance of observing unlikely events when looking across many genes). Various computational tools have been developed to attempt to assess functional significance of a mutation. These tools predict the effect of an amino acid change on the protein structure and function, and some tools (for example, SIFT, CanPredict, PolyPhen and CHASM[81–85]) aim to distinguish 'driver' from 'passenger' alterations. In general, experimental validation of the function of mutations by approaches such as transformation assays[86] is the most powerful method; however, functional validation is limited because there is no suite of functional assays that are suitable for assessing all the types of pathways that can be altered in cancer.

Copy number. Array-based measurements have proven to be a powerful approach to determine the pattern of copy number alterations in cancer, from the gain or loss of chromosome arms to focal amplifications and deletions that might range from tens of kilobases to tens of megabases in size[87–94]. Sequence-based approaches to copy number were applied even before the development of second-generation sequencing technologies, using digital karyotyping approaches[95,96], which are based on sequencing large numbers of short sequence tags[97].

Second-generation sequencing methods offer substantial benefits for copy number analysis, including higher resolution (up to the level of the single-base insertion or deletion) and precise delineation of the breakpoints of copy number changes[22,26,98]. The digital nature of second-generation sequencing allows us to estimate the tumour-to-normal copy number ratio at a genomic locus by counting the number of reads in both tumour and normal samples at this locus. Unlike array-based measurements, counting sequences does not suffer from saturation and therefore allows accurate estimation of high copy number levels. It is, however, affected by sequencing biases caused by sequence context, such as GC content.

Genome-wide sequence-based methods are particularly valuable for copy number changes of between approximately 100 and 1000 bases — reflecting the maximum size that can be easily detected by PCR-based locus-specific sequencing and the minimum current resolution limit of array technologies, respectively. Copy number measurements by sequencing also allow definition of the sequence on the other side of the breakpoint.

Transformation assay
The measurement of cell phenotypes to assess oncogenic changes.

Digital karyotyping
A method to quantify DNA copy number. Short sequence-derived tags that cover the genome are used to read-out relative copy number.

High-coverage or low-coverage whole-genome sequencing is feasible for analysis of copy number alterations in cancer[22,26].

Chromosomal rearrangements. Before second-generation sequencing technology, there was no systematic method to identify chromosomal rearrangements in cancer genomes, except in cases of relatively simple genomes such as those of leukaemias, lymphomas and sarcomas, in which recurrent rearrangements could be detected by cytogenetics. The complex genomes of epithelial cancers proved refractory to such approaches. The first two major recurrent translocations to be identified in epithelial cancers, the *TMPRSS2–ERG* translocation in prostate cancer and the *EML4–ALK* translocation in non-small cell lung cancer, were discovered by informatic and functional approaches[42,43]. Second-generation sequencing analysis of genomes[13,22,39,99] and transcriptomes[6,7,75] has now been shown to allow systematic description of the rearrangements in a given cancer sample. Extension of these approaches to large numbers of samples should lead to the discovery of the major recurrent translocations in cancer. Indeed, a recent study has identified recurrent translocations of the *BRAF* and *CRAF* genes in prostate carcinomas[76].

Among the rearrangements that can be detected by second-generation sequencing[98] are: intrachromosomal rearrangements, including inversions, tandem duplications and deletions; insertions of non-endogenous sequences, including viral sequences; reciprocal and non-reciprocal interchromosomal rearrangements; and complex rearrangements, including combinations of these various events[39]. Using current technologies, rearrangements within long highly repetitive sequences such as Alu and LINE elements or centromeres present a major challenge and often cannot be detected.

Of the second-generation sequencing strategies, whole-genome sequencing is the most comprehensive but most costly approach, albeit made less sequence-intensive by the use of jumping libraries that provide high physical coverage. Transcriptome sequencing is highly cost efficient but is limited to the detection of coding fusion transcripts and would fail to detect, for example, the immunoglobulin–*MYC* rearrangements of Burkitt's lymphoma[100]. Exome sequencing has limited use for chromosomal rearrangement discovery, as it will only find those rearrangements that are within or near exons.

Microbe discovery methods. In addition to somatic alterations that are modifications of the normal human genome, many cancers are caused by microbial infections. A classic example is human papillomavirus, which can cause cervical cancer[101]. However, neither array methods nor directed sequencing approaches can identify new examples of microbial genomes that have inserted into the human genome in cancer samples.

Computational subtraction of the sequence from a sample from the human reference genome[25,95,102,103] can detect non-human sequences and thereby identify novel microbial infections associated with human disease. This method was theoretically demonstrated with first-generation sequencing methods but its practical application requires the higher sequence depth of second-generation sequencing. Computational subtraction of second-generation sequencing of the transcriptome has now been used successfully to discover the Merkel cell polyomavirus in a rare human skin cancer[23].

Some of the challenges in discovering novel microbial infections include low concentration of the microbial agent, possible 'hit-and-run' mechanisms of disease causation (in which the microbe is necessary for disease initiation but is no longer present in the cancer), quality issues with second-generation sequencing that cause artefacts that do not match the human genome and incompleteness of human genome reference sequences. These limitations, with the exception of the hit-and-run mechanism, are likely to be addressed by improvements in sequencing technology that decrease error rates and increase coverage.

Computational issues

Tools for computational analysis need to develop rapidly to keep up with the huge quantity of experimental data produced by second-generation sequencing. Many of the computational tools for second-generation sequencing were developed for analysing non-cancer samples for human population genetic studies[104–107]. Although computational analysis of cancer genomes can use many of these tools, there are additional challenges that are unique to cancer. The three main issues are: the need to simultaneously analyse data from the tumour and patient-matched normal tissue to identify rare somatic events (for example, somatic single nucleotide variations are ~1,000 times less frequent than germline variants); the ability to analyse very different and highly rearranged genomes; and the ability to handle samples with unknown levels of non-tumour contamination and heterogeneity within the tumour. TABLE 2 provides a summary of software packages.

Alignment and assembly. Before the data can be analysed, reads must be aligned to the specific chromosome, position and DNA strand from which they are most likely to have originated. At present, these alignments are performed against reference human genomes using methods developed for normal samples, such as MAQ[108], BWA[109], SSAHA2 (REF. 110), Bowtie[111], SOAP2 (REF. 112), SHRiMP[113], BFAST[114] and others. Methods differ in terms of their ability to accurately map noisy reads, paired-end reads, long versus short reads and in their computational efficiency. The choice of method depends on the sequencing platform, data quality and computational resources.

The uniqueness of every cancer genome and the difficulty of correctly assigning rearranged sequences from homologous regions mean that *de novo* assembly of cancer genomes, although computationally complex, is likely to become the most powerful approach.

Directed sequencing
Sequencing only subsets of the genome, for example, particular genes or regions of interest.

Table 2 | **Computational tools for cancer genomics**

Category	Method	URL	Comments	Refs
Alignment	MAQ	http://maq.sourceforge.net	Used by most cancer genome papers so far	108
	BWA	http://bio-bwa.sourceforge.net	Replacing MAQ. Considerably faster	109
	ELAND	http://www.illumina.com		117
	SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2	Used to validate location of reads	39,110
	Bowtie	http://bowtie-bio.sourceforge.net/index.shtml		111
	SOAP2	http://soap.genomics.org.cn		112
	SHRiMP	http://compbio.cs.toronto.edu/shrimp		113
	Corona Lite	http://solidsoftwaretools.com/gf/project/corona	Used for SOLiD	
	BFAST	http://bfast.sourceforge.net	Mainly used for SOLiD	114
Mutation calling	SNVMix	http://www.bcgsc.ca/platform/bioinfo/software/SNVMix		80
	CASAVA	http://www.illumina.com/software/genome_analyzer_software.ilmn		117
	Samtools	http://samtools.sourceforge.net		104
	Unified genotyper	http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper		107
	VarScan	http://varscan.sourceforge.net		105
Indel calling	Pindel	http://www.ebi.ac.uk/~kye/pindel		106
Copy number analysis	CBS	http://www.bioconductor.org https://r-forge.r-project.org/R/?group_id=702	CBS used on tumour/normal ratios calculated in fixed windows	118
	SegSeq	http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=182		26
Prediction of mutation functional effect	SIFT	http://blocks.fhcrc.org/sift/SIFT.html http://sift.jcvi.org		81
	Polyphen-2	http://genetics.bwh.harvard.edu/pph2		83
	XVAR	http://xvar.org		119
	CHASM			85
Visualization	CIRCOS	http://mkweb.bcgsc.ca/circos	Essentially all papers use CIRCOS to display genomic events	120
	IGV	http://www.broadinstitute.org/igv	IGV is used to display genomic events and for manual review	

A list of additional alignment methods with a brief description of each is constantly updated at http://en.wikipedia.org/wiki/List_of_sequence_alignment_software.

Mutation detection. As somatic genome alterations are rare, any method that detects mutations in cancer must do so with low false-positive rates. For example, as noted above, somatic single nucleotide variations occur at rates on the order of one somatic mutation per million bases, therefore the false-positive rate for any mutation detection method should ideally be considerably lower than 10^{-6} errors per base.

Recently, the first report of a method specific for somatic mutation calling, SNVMix, has been published[80], and additional tools are in development. Systematic analysis of false-positive and false-negative rates of the different methods based on real cancer data is yet to be performed. Alternatively, a naive somatic mutation caller can be built by independently applying a germline single-sample mutation caller to the tumour and normal data sets; somatic events are those detected in the tumour and not detected in the normal data. Germline analyses are more mature and there are a host of mutation-calling tools available, such as Samtools[104], UnifiedGenotyper[107], VarScan[105] and others.

Somatic mutation calling is more complex than germline mutation calling because cancer samples vary in their purity and ploidy. A key parameter defined for each mutation is its allelic fraction — the expected fraction of reads in the tumour that harbour the mutation among all reads that map to the same genomic location. The allelic fraction captures the local complexity of the tumour genome, the non-tumour contamination levels and any mutation-dependent experimental or alignment bias, and is also affected by the ploidy of the tumour and the copy number of the region. In germline analysis, most mutations have an allelic fraction of either 1/2 for heterozygous events or 1 for homozygous events. By contrast, somatic mutations may have any value above 0 up to 1. Clearly, the false-positive and false-negative rates not only depend on the coverage in the tumour

Free serum DNA
DNA that is cell-free and is circulating in the bloodstream. It typically refers to tumour DNA that can be isolated in the blood.

and normal tissue of the mutated site, but also strongly depend on the allelic fraction. Methods to determine the allele-specific copy number have been developed for SNP arrays[115] and are currently under development for second-generation sequencing data; these methods will allow determination of allelic fractions of mutations.

Validation of mutation and rearrangement calls. Accurate estimation of false-positive and false-negative rates is a challenge in itself. False-positive rates can be estimated by validation of the event using an orthogonal technology. For example, a common method for validation of single nucleotide variations and insertions or deletions is to use a genotyping assay such as mass spectrometric analysis[74]. However, this technology was designed for germline analysis and is not sufficiently sensitive to validate mutations with low allelic fractions. Therefore, current efforts are focused on applying deep targeted second-generation sequencing to validate the events. False-negative rates are even more complicated to estimate as one needs a set of known true-positive mutations for comparison, which are not readily available.

For validating rearrangements, the current methods require PCR amplification of the region surrounding the rearrangement followed by sequencing of this region. Therefore, they are not high-throughput. A developing concept is to capture the rearranged sites using a similar protocol to the exon capture approach and apply deep sequencing. Sequential application of validation methods to computational analysis of the genome will lead to iterative improvements in methods for the initial calling of genome alterations in cancer.

The future of cancer genomics

It is likely that second-generation sequencing methods will continue to transform cancer genomics, leading to the comprehensive discovery of all the major alterations in the cancer genome, followed by the application of comprehensive sequencing approaches to cancer diagnostics. Computational analysis will become a central part of these discovery and diagnostic efforts.

The major challenge will be to make biological sense of the mountains of genomic data. This will require computational, biological and clinical analyses of the genome data. The computational analyses will assess reproducibility and statistical significance; the biological analyses will assess links to pathways and the functional relevance of mutated genes to cancer; and the clinical analyses will assess relationships of genome alterations with cancer epidemiology, histology, prognosis and response to therapy.

Perhaps the biggest impact of second-generation sequencing of cancer genomes will be in cancer diagnostics. Comprehensive characterization of genomic abnormalities has not previously been feasible because of sample processing and purity requirements. Specifically, digital counting of mutant alleles helps to overcome the challenges of low tumour quantity, normal cell admixture in the tumour, heterogeneity of tumour genomes and variable ploidy. In the long run, second-generation sequencing is likely to allow diagnosis from ever smaller samples, eventually including circulating tumour cells[116] and free serum DNA[99]. Information databases that connect genomic findings with clinical parameters to ascertain the relevance of the genome alterations are also required to realize the potential of cancer genomics. These developments are likely to potentiate accurate genome-based diagnosis for an ever wider set of patients with cancer.

1. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
2. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
3. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
4. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
5. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
6. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009). **This paper demonstrates the power of second-generation transcriptome sequencing to identify rearrangements in coding genes.**
7. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **106**, 12353–12358 (2009).
8. Ng, S. B. *et al.* Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
9. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
10. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008). **This is the first publication describing whole-genome sequencing of a human cancer.**
11. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
12. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
13. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
14. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
15. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
16. Weir, B., Zhao, X. & Meyerson, M. Somatic alterations in the human cancer genome. *Cancer Cell* **6**, 433–438 (2004).
17. Mitsudomi, T. *et al.* Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised Phase 3 trial. *Lancet Oncol.* **11**, 121–128 (2009).
18. Mok, T. S. *et al.* Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**, 947–957 (2009).
19. Rosell, R. *et al.* Screening for epidermal growth factor receptor mutations in lung cancer. *N. Engl. J. Med.* **361**, 958–967 (2009).
20. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
21. Thomas, R. K. *et al.* Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Med.* **12**, 852–855 (2006).
22. 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).
23. Feng, H., Shuda, M., Chang, Y. & Moore, P. S. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* **319**, 1096–1100 (2008).
24. MacConaill, L. & Meyerson, M. Adding pathogens by genomic subtraction. *Nature Genet.* **40**, 380–382 (2008).
25. Weber, G., Shendure, J., Tanenbaum, D. M., Church, G. M. & Meyerson, M. Identification of foreign gene sequences by transcript filtering against the human genome. *Nature Genet.* **30**, 141–142 (2002).
26. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6**, 99–103 (2009).
27. Getz, G. *et al.* Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500 (2007).
28. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008). **This is the first paper from The Cancer Genome Atlas, which demonstrates the power of integrative analysis of multiple platforms for genomic analysis on a large series of cancer samples.**
29. Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216 (2006).
30. Gilbert, M. T. *et al.* The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? *PLoS ONE* **2**, e537 (2007).
31. Wood, H. M. *et al.* Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res.* **38**, e151 (2010).

32. Gallegos Ruiz, M. I. *et al.* EGFR and K-ras mutation analysis in non-small cell lung cancer: comparison of paraffin embedded versus frozen specimens. *Cell Oncol.* **29**, 257–264 (2007).
33. Marchetti, A., Felicioni, L. & Buttitta, F. Assessing *EGFR* mutations. *N. Engl. J. Med.* **354**, 526–528 (2006).
34. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20**, 68–80 (2010).
35. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
The first publication of the comprehensive sequencing of primary and metastatic tumour material from an individual.
36. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotech.* **26**, 1135–1145 (2008).
37. Pettersson, E., Lundeberg, J. & Ahmadian, A. Generations of sequencing technologies. *Genomics* **93**, 105–111 (2009).
38. Hoffman, B. G. & Jones, S. J. Genome-wide identification of DNA–protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. *J. Endocrinol.* **201**, 1–13 (2009).
39. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
This is the largest collection of samples for a single cancer type to be subject to whole-genome rearrangement analysis and documents the large sample-to-sample variability in frequency of events.
40. Rowley, J. D. Chromosome translocations: dangerous liaisons revisited. *Nature Rev. Cancer* **1**, 245–250 (2001).
41. Meyerson, M. Cancer: broken genes in solid tumours. *Nature* **448**, 545–546 (2007).
42. Tomlins, S. A. *et al.* Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
43. Soda, M. *et al.* Identification of the transforming *EML4–ALK* fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).
44. Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).
45. Huang, C. R. *et al.* Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**, 1171–1182 (2010).
46. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
47. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotech.* **27**, 182–189 (2009).
48. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
49. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
50. Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* **10**, R115 (2009).
51. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
52. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
53. 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.* **350**, 2129–2139 (2004).
54. 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* **101**, 13306–13311 (2004).
References 52–54 were the first publications to link therapeutic outcome in lung cancer to specific somatically acquired point mutations, and they suggest the value of systematic sequencing of kinase gene families.
55. Stephens, P. *et al.* Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431**, 525–526 (2004).
56. Baxter, E. J. *et al.* Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet* **365**, 1054–1061 (2005).
57. James, C. *et al.* A unique clonal *JAK2* mutation leading to constitutive signalling causes polycythaemia vera. *Nature* **434**, 1144–1148 (2005).
58. Kralovics, R. *et al.* A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N. Engl. J. Med.* **352**, 1779–1790 (2005).
59. 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).
60. Zhao, R. *et al.* Identification of an acquired JAK2 mutation in polycythemia vera. *J. Biol. Chem.* **280**, 22788–22792 (2005).
61. Dutt, A. *et al.* Drug-sensitive *FGFR2* mutations in endometrial carcinoma. *Proc. Natl Acad. Sci. USA* **105**, 8713–8717 (2008).
62. Pollock, P. M. *et al.* Frequent activating FGFR2 mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene* **26**, 7158–7162 (2007).
63. Chen, Y. *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **455**, 971–974 (2008).
64. George, R. E. *et al.* Activating mutations in ALK provide a therapeutic target in neuroblastoma. *Nature* **455**, 975–978 (2008).
65. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455**, 967–970 (2008).
66. Mosse, Y. P. *et al.* Identification of *ALK* as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930–935 (2008).
67. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304**, 554 (2004).
68. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
69. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
70. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
This paper described the first example of whole-exome sequencing of human cancers.
71. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
72. Jones, S. *et al.* Exomic sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. *Science* **324**, 217 (2009).
73. Bainbridge, M. N. *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol.* **11**, R62 (2010).
74. Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nature Genet.* **39**, 347–351 (2007).
75. Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427 (2010).
76. Palanisamy, N. *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nature Med.* **16**, 793–798 (2010).
77. Shah, S. P. *et al.* Mutation of *FOXL2* in granulosa-cell tumors of the ovary. *N. Engl. J. Med.* **360**, 2719–2729 (2009).
78. Morrissy, A. S. *et al.* Next-generation tag sequencing for cancer gene expression profiling. *Genome Res.* **19**, 1825–1835 (2009).
79. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).
80. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**, 730–736 (2010).
81. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
82. Kaminker, J. S., Zhang, Y., Watanabe, C. & Zhang, Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.* **35**, W595–W598 (2007).
83. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
84. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
85. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).
86. Hahn, W. C. & Weinberg, R. A. Rules for making human tumor cells. *N. Engl. J. Med.* **347**, 1593–1603 (2002).
87. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104**, 20007–20012 (2007).
88. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
This paper is an analysis of somatic copy number changes across 26 different human cancer types and points to regions commonly altered at significant levels across cancer types.
89. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
90. Bignell, G. R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14**, 287–295 (2004).
91. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
92. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007).
93. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**, 3060–3071 (2004).
94. Zhao, X. *et al.* Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res.* **65**, 5561–5570 (2005).
95. Tengs, T. *et al.* Genomic representations using concatenates of type IIB restriction endonuclease digestion fragments. *Nucleic Acids Res.* **32**, e121 (2004).
96. Wang, T. L. *et al.* Digital karyotyping. *Proc. Natl Acad. Sci. USA* **99**, 16156–16161 (2002).
97. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
98. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009).
99. Leary, R. J. *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
100. Dalla-Favera, R. *et al.* Human *c-myc onc* gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proc. Natl Acad. Sci. USA* **79**, 7824–7827 (1982).
101. Durst, M., Gissmann, L., Ikenberg, H. & zur Hausen, H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proc. Natl Acad. Sci. USA* **80**, 3812–3815 (1983).
102. Feng, H. *et al.* Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *J. Virol.* **81**, 11332–11340 (2007).
103. Xu, Y. *et al.* Pathogen discovery from human tissue by sequence-based computational subtraction. *Genomics* **81**, 329–335 (2003).
104. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
105. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
106. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
107. McKenna, A. H. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 19 Jul 2010 (doi:10.1101/gr.107524.110).
108. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
109. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
110. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).

111. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
112. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
113. Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**, e1000386 (2009).
114. Homer, N., Merriman, B. & Nelson, S. F. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* **4**, e7767 (2009).
115. LaFramboise, T. *et al.* Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput. Biol.* **1**, e65 (2005).
116. Maheswaran, S. *et al.* Detection of mutations in *EGFR* in circulating lung-cancer cells. *N. Engl. J. Med.* **359**, 366–377 (2008).
117. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
118. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
119. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **8**, R232 (2007).
120. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645 (2009).

Acknowledgements
We thank M. Lawrence and G. Saksena for careful review of the manuscript. We acknowledge support from The Cancer Genome Atlas programme of the National Cancer Institute, U24CA143867 and U24CA143845, and from the National Human Genome Research Institute, U54HG003067.

Competing interests statement
The authors declare competing financial interests: see web version for details.

FURTHER INFORMATION

Matthew Meyerson's homepage: http://research4.dfci.harvard.edu/meyersonlab
Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
BFAST: http://bfast.sourceforge.net
BWA: http://bio-bwa.sourceforge.net
CASAVA: http://www.illumina.com/software/genome_analyzer_software.ilmn
CBS: http://www.bioconductor.org; https://r-forge.r-project.org/R/?group_id=702
CIRCOS: http://mkweb.bcgsc.ca/circos
Corona Lite: http://solidsoftwaretools.com/gf/project/corona
ELAND: http://www.illumina.com
IGV: http://www.broadinstitute.org/igv
MAQ: http://maq.sourceforge.net
Pindel: http://www.ebi.ac.uk/~kye/pindel
Polyphen-2: http://genetics.bwh.harvard.edu/pph2
Samtools: http://samtools.sourceforge.net
SegSeq: http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=182
SHRiMP: http://compbio.cs.toronto.edu/shrimp
SIFT: http://blocks.fhcrc.org/sift/SIFT.html; http://sift.jcvi.org
SNVMix: http://www.bcgsc.ca/platform/bioinfo/software/SNVMix
SSAHA2: http://www.sanger.ac.uk/resources/software/ssaha2
SOAP2: http://soap.genomics.org.cn
Unified genotyper: http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper
VarScan: http://varscan.sourceforge.net
XVAR: http://xvar.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF