BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 5. 524–534
Advance Access published on 28 April 2010
doi:10.1093/bib/bbq013

Application of second-generation sequencing to cancer genomics

Keith Robison

Submitted: 1st March 2010; Received (in revised form): 26th March 2010

Abstract

New generations of DNA sequencing technologies are enabling the systematic study of genetic derangement in cancers. Sequencing of cancer exomes or transcriptomes or even entire cancer genomes are now possible, though technical and economic challenges remain. Cancer samples are inherently heterogeneous and are often contaminated with normal DNA, placing additional demands on informatics tools for detecting genetic variation. However, even low coverage sequencing data can provide valuable information on genetic rearrangements, amplifications and losses in tumor genomes. Novel recurrent oncogenic mutations and fusion transcripts have been discovered with these technologies. In some sequenced cancer genomes, tens of thousands of genetic alterations have been discovered. While this enables the detailed dissection of mutation classes, it also presents a formidable informatics problem of sorting active 'driver' mutations from inactive 'passenger' mutations in order to prioritize these for further experimental characterization.

Keywords: *cancer; sequencing-by-synthesis; genomics; mutations; copy number variation; genomic rearrangements*

INTRODUCTION

Most cancers possess extensive genome alteration ranging from a small number of point mutations to widespread aneuploidy [1]. Ongoing genomic instability results in chronic accumulation of mutations in tumors, an effect which can be heightened by DNA-damaging therapies. The study of such changes through DNA sequencing has a long history but has been limited by the high cost and limited throughput of DNA sequencing technologies. Second-generation sequencing instruments capable of generating vast quantities of sequencing data at modest cost have enabled a new round of studies scanning for somatic mutations across the exome, transcriptome or even the entire genome.

Second-generation sequencing technologies have been reviewed in detail elsewhere [2]. All of these technologies work via extension of a defined DNA primer, much like conventional Sanger sequencing. A key difference is that these technologies do not size separate fragments by electrophoretic mobility, but instead sequentially image the stepwise addition of nucleotides (or nucleotide blocks). This allows the packing of tens or hundreds of millions of sequencing targets onto an area the size of a microscope slide. Read lengths from many of these technologies are substantially shorter than from conventional Sanger sequencing, with the exception of the Roche 454 platform which can attain reads well over 700 nucleotides [2]. These short (35–100 nucleotides) reads present specific informatics challenges which will be touched on in this review.

Here I review the identification of somatic mutations and rearrangements in cancer genomes using second-generation sequencing technologies. This scope results in the omission of several other interesting topics at the intersection of sequencing and cancer, such as genome-wide analysis of the somatic epigenomics of cancer and studies to better understand inherited predispositions to cancer. Replacement of microarray technology for measuring mRNA and microRNA levels by sequencing is also gaining popularity [3], but will only be discussed

Corresponding author. Keith Robison, Infinity Pharmaceuticals Inc., 780 Memorial Drive, Cambridge MA 02139, USA.
Tel: 617-453-1339; Fax: 617-453-1001; E-mail: keith.robison@infi.com
Keith Robison is Informatics Lead Senior Scientist at Infinity Pharmaceuticals, a developer of novel cancer therapeutics. He earned his PhD from Harvard University and his BS from the University of Delaware.

when this information is obtained as a by-product of hunting for mutations.

Any review of second-generation sequencing technology is doomed to omit key new work which appears during the processing of the manuscript, as the field moves rapidly due to technological evolution, introduction of entirely new technology platforms and the application of these approaches to cancer. This review covers publications prior to March 2010.

SPECIAL ISSUES POSED BY CANCER SAMPLES

The acquisition of high-quality, appropriately consented samples is a significant challenge to cancer genomics [4]. Many tumors are not easily accessible for biopsy, particularly recurrent tumors. Histopathological slides and blocks represent a huge resource of cancer [4]. These room temperature stable archives of material have generally been preserved via formalin fixation and embedding in paraffin. Such 'FFPE' samples are potentially a rich source of information on cancer [5]. However, the process of fixing and embedding the samples subjects the DNA to many insults which may generate artifactual mutations in DNA sequencing data [6]. Furthermore, these samples are often small and extraction of DNA is inefficient. Whole genome amplification (WGA) may enable working with small amounts of input DNA, though WGA can introduce biases [7, 8], point mutations [9] and spurious rearrangements [8]. Advanced sample preparation methods can reduce the required amount of input DNA from many micrograms to nanograms [10, 11].

While cell lines yield pure samples, clinical cancer samples are generally mixed with normal cells. These may include surrounding tissue, fibroblasts (tumor stroma) and infiltrating lymphocytes. Hence, the DNA preparations are likely to contain a mixture of normal and tumor DNA, diluting the signal from the tumor [12]. Microdissection or laser capture approaches can improve the purity of tumor DNA, but often generate extremely tiny samples. The problem of normal DNA admixture is particularly acute for Sanger sequencing from uncloned DNA (e.g. PCR products), as the peaks in the electropherogram present the averages across the population of molecules [13]. Second-generation systems work on either individual DNA molecules or homogenous clusters of DNA amplified from a single DNA molecule [2]. Hence, they are less sensitive to this effect. However, admixture of normal DNA will increase the coverage requirements to detect somatic mutations [14].

Tumors themselves are generally not homogeneous. As noted before, a hallmark of cancers is the reduction in their ability to faithfully replicate their genome. In some tumors, such as those arising from hereditary defects in DNA replication, very high mutation rates are observed [15]. Hence, each tumor cell mitosis has a possibility of generating progeny genetically distinct from their parent. Furthermore, environmental insults (such as continued smoking in lung cancer [16] or sun exposure in non-metastatic melanoma [17]) have the opportunity to generate additional mutations. In addition, mutations can occur within a region of focal amplification, which will also result in dilution of the mutation within a large background of amplified wild-type DNA.

WHOLE GENOME SEQUENCING

Most second-generation sequencing systems generate very short reads in contrast to the near kilobase reads obtainable in very good Sanger sequencing data. Published data using the Illumina system have ranged from 31 [18] to 100 nucleotides [19]; using the SOLiD sequencing-by-ligation approach from 25 to 50 nucleotides [20], and the Helicos system from 20–50 nucleotides [21]. The 454 system is notable for substantially longer read lengths, with 200–500 nucleotide reads common. A variety of issues ultimately cause degradation of the sequencing signal and limit the amount of sequence which can be read from a single primer [2].

Read pair approaches increase the information obtainable from a single DNA molecule by obtaining sequence from both ends of the original fragment [20]. Paired end sequencing on the Illumina platform acquires sequence from both ends of a single DNA molecule. Upon completion of sequence acquisition from one primer, a series of steps strips the old primer from one end and replaces it with a primer on the other end from which a new read is acquired. In mate pair strategies, enzymatic or mechanical means are employed to replace the majority of the central DNA by a second universal priming site. Mate pair approaches enable sequencing both ends of initial fragments which are not practical to amplify within a sequencer. Large fragments are more

sensitive for detecting rearrangements, although they are also less precise at locating the breakpoints [22]. However, mate pair strategies can generate chimeric molecules [23], which can cause false positive rearrangement predictions.

While reads from the SOLiD, Illumina and Helicos system are limited in length, hundreds of millions of such reads can be obtained from a single sample in a single sequencer run. This immediately presents a bioinformatic challenge: how to efficiently map millions to billions of very short reads. Several dozen programs have been developed to address this problem, a topic covered elsewhere in this volume [24]. Read pair strategies introduce an additional element of information, as the two reads should (in the absence of genomic rearrangement) be situated from each other with distance and orientation appropriate to the library preparation. Read pairs also enable the unambiguous alignment of a read to a repetitive region, if its partner maps uniquely and this mapping is consistent only with a single repeat location. However, such alignments must be interpreted with caution, particularly if an inferred genomic feature is supported by only a small number of read pairs in the dataset. This feature is likely to be very valuable for sequencing genes with close paralogs or retrotransposed pseudogene copies. However, even with read pair strategies many reads cannot be mapped uniquely due to both reads being derived from repetitive sequences.

Even when aligning normal DNA to a human reference, natural genomic structural variation can prevent correct reads from aligning [25, 26]. Both germ-line SNPs [27] and somatic point mutants may reduce alignment sensitivity. Small indels are troublesome to detect, with adjacent indels potentially reducing the ability to detect nearby substitutions [28]. In a melanoma genome a known oncogenic 2-bp deletion was not detected automatically but could be found in the original data when specifically sought out [17]. However, a more recent study of a glioblastoma genome using longer reads on the same platform (SOLiD) had a high frequency of indel detection [29], suggesting that this problem can be addressed by a combination of improved sequencing protocols and new bioinformatics tools [30]. The sequenced melanoma genome also showed a high frequency of doublet mutations, in which two adjacent bases are mutated [17]. The possibility that doublet mutations may have led to the failure to map reads was not explored in this article. While most non-aligning DNA fails because of low quality or a highly repetitive nature, careful analysis of such reads may reveal infectious agents [31], structural variants [26, 32], highly mutated regions or regions otherwise well represented in the reference genome sequence.

An alternate approach is *de novo* genome assembly for each human DNA sample, but this approach requires enormous compute resources [33] and is unlikely to become routine. Localized reassembly of reads suspected to contain novel variants is a promising strategy, balancing variant detection sensitivity with compute expense [32–34].

A plethora of tools have emerged for converting aligned or assembled reads into variant calls, which has been reviewed elsewhere [35]. An assumption of diploidy will frequently be invalid in cancer genomes due to deletions or amplifications. The interaction of mutation and amplification processes as well as admixture can yield a wide range of allele frequencies [36]. Hence, tools developed for variant detection in normal tissue may not work well in clinical cancer samples. It is worth noting that three of the published whole cancer genome sequencing papers at this time have used pure material from tissue culture [16, 17, 29] and the other three used clinical samples with >80% tumor [36–38]. For truly broad application to cancer, validation of these approaches in lower tumor purity samples will be required. Variant detection tools developed for pooled samples may prove useful [14, 39, 40] and estimates of sample purity can be used to inform the identification of high-confidence mutations [37]. Variants identified in the tumor sample must be partitioned into germ-line variants and somatic mutations, generally by sequencing normal DNA from the same patient in parallel [16, 17, 37, 38]. Normal DNA is most typically obtained from a blood sample, skin biopsy or cheek (buccal) swab, which provide samples less likely to be contaminated with tumor DNA than adjacent tissue. However, in a leukemia study tumor-specific mutations were detected in a skin biopsy due to contamination with leukemic blast cells, underscoring the challenge of obtaining even normal DNA from some cancer patients and the value of computational approaches which can utilize sample purity information [37].

Cancer genomes often contain multiple rearrangements, including insertions, deletions, local segmental duplications and translocations. Identifying these changes and particularly their boundaries (breakpoints) is of great interest [22, 41], though

few of these rearrangements are recurrent [42, 43]. Breakpoints generated by rearrangements can be targeted by PCR assays, enabling patient specific blood assays potentially useful for monitoring treatment [43]. Two different strategies have generally been employed. First, with sufficient data it is possible to identify reads which contain the actual breakpoint. Bioinformatics tools such Pindel can reconstruct the sequence of a breakpoint by local reassembly of paired reads [34]. Read pair data may also be used to indirectly detect breakpoints. The most straightforward approach is to identify read pairs which map to different chromosomes or map in incorrect orientation on the same chromosome. An additional category is read pairs in the correct relative orientation but mapping in the wrong order due to a segmental duplication. Intrachromosomal alterations can be detected by identifying read pairs in the correct orientation but separated by unusually short or long distances. Because of the deep sampling nature of second-generation sequencing, tools such as MoDIL can identify short indels by the shift in sequence pair separation distance distribution which they induce [44]. However, the sensitivity and precision of such methods is limited by the tightness of the insert size distribution in a paired end library preparation. For example, long insert (3 kb or greater) mate paired libraries may falsely imply indels due to the challenge of precisely controlling the insert lengths during library preparation and can lead to false inference of rearrangements due to chimaeras [23]. Tools such as BreakDancer [45], PEMer [46] and VariationHunter [47] identify multiple types of rearrangements in a single run. However, rearrangements and indels identified by such programs should be weighed carefully in light of the number of reads supporting each inference, given the possibility of false positives arising from mismapped reads, particularly read pairs in which one read maps to repetitive sequence.

Paired read strategies can also enable the determination of local linkage of alleles (haplotypes) by chained inference. If the paired reads each contain a heterozygous variant, then those two variants are linked. Further pairs may enable a chain of linked variants to be inferred [48]. While to date this has been applied only to normal genomes [20], application to cancer data promises to yield significant insight. Linkage of multiple mutations may suggest cooperative effects, such as the evolution of chemotherapeutic resistance [49]. Linkage of tumor mutations to germline polymorphisms may suggest mechanisms of oncogenesis [50]. Somatic mutations or germline variants may also be used to detect unbalanced amplification or transcription of one allele [51].

Deviations from normal copy number in cancer are common. In addition to indicating the deletion or amplification of one or more genes, unbalanced amplification within genes can suggest unbalanced translocations. Second-generation sequencing data has demonstrated two key advantages over microarrays for copy number analysis. First, breakpoints can be located with a precision determined by sampling depth; higher resolution can simply be achieved with additional sequence data. Second, paired read strategies can identify balanced translocations in addition to copy number changes. Copy number estimation requires far less data than full genome scanning. High resolution (<1 kb) copy number maps have been achieved with 0.3× sequence coverage [52], whereas 30–40× coverage is required for whole-genome heterozygous variant discovery [20, 21, 26, 32, 33]. Even a few million mapped reads are sufficient to provide 15 kb precision, 3-fold better resolution than a commonly used 244 K microarray [5]. However, the detection sensitivity for breakpoints is dependent on sequence coverage; one study estimated only 50% of rearrangements were detected [53]. Use of longer insert mate pair libraries may enhance this sensitivity. Cross-referencing copy number information with rearrangement predictions from paired read strategies can be used to filter artifactual rearrangements [43].

Finally, from a bioinformatic perspective it is critical to consider the specific platform and sample preparation methods used to generate a cancer genome dataset. Each second-generation sequencing platform has its own characteristic error spectrum. For example, much study has gone into the problem of homopolymer tracts with the Roche 454 system [54]. Neighboring bases have been observed to influence sequencing errors in the Illumina system, notably a higher proportion of miscalls after G residues [55], and quality values generated by base callers may misestimate actual error rates [56]. The two-base encoding used by the SOLiD system should lead to very sensitive and reliable single nucleotide variant calling [29], though one recent study found that only 25% of small indels detected by SOLiD sequencing could be validated with Sanger sequencing [16]. Improved library methods have increased the uniformity of genome sampling with

regard to GC content [20, 57]. While no cancer genome datasets are available from the Helicos platform, a normal human sample sequenced with this system showed a high rate of indel errors owing to unobserved nucleotide incorporation events ('dark bases') [21]. The specific nature of these platform-specific biases will probably periodically shift as improved library preparation techniques and base calling software are developed in response to identified issues.

Even with the issues noted above, in the most recent studies 85% or more of single nucleotide variants identified by second-generation sequencing can be validated by Sanger sequencing [16, 17, 29]. A valuable benchmark for estimating false positive and false negative rates from second-generation sequencing is SNP microarray profiling of the same sample; concordances in excess of 95% have been observed in such analyses [37, 38].

TARGETED SEQUENCING

Whole genome sequencing (WGS) is becoming the gold standard for genome analysis. WGS has the advantage of detecting effectively all mutations, rearrangements and copy number changes. However, although costs continue to drop rapidly, WGS remains prohibitively expensive to apply on a grand scale, currently in excess of $35 000 per human sample in reagent costs [21], which does not include labor, equipment depreciation, bioinformatics and other real expenses. It is worth noting that as reagent costs per genome drop, informatics may become the dominant expense in human genome sequencing projects. Particularly with the need to assay many samples to detect recurrent mutations and co-occurring mutations, strategies which reduce cost and enable analyzing more samples are valuable.

Targeted sequencing refers to strategies for enriching the input to the sequencer for DNA regions of strong interest. In addition to reducing cost per sample, these approaches offer the possibility of much higher coverage of areas of interest, which may overcome issues of sample purity or quality. However, targeted methods may require substantially more input DNA than WGS [38]. These approaches generally either use hybridization of target DNA to designed oligonucleotides (or PCR products) or the specific amplification of targeted regions by PCR [58]. Each of these approaches has important considerations for cancer genomics and for the bioinformatic analysis of data derived by such approaches.

PCR methods are very well understood and can give very even coverage. Microfluidic setup enables small amounts of input material to be analyzed against many primer sets [59]. For cancer genomes, one liability of most PCR-based approaches is an inability to detect large rearrangements, insertions and deletions. Novel translocations are effectively invisible to PCR due to the need to define both primers in the reaction. Similarly, deletions will not be captured unless both primers in a reaction flank the deletion. Large insertions may not be amplified well if the insertion greatly enlarges the size of the PCR amplicon or exceeds the size of the input DNA fragments. PCR methods also can suffer from two specific informational problems when the PCR fragments are converted to sequencer-ready libraries by mechanical fragmentation. First, the ends of amplicons can be greatly overrepresented. Second, sequences derived from PCR primers will be enriched for errors due to oligonucleotide synthesis errors. If overlapping PCR amplicons are used, special care must be used when calling mutations in sequences which could have derived from primer regions. Finally, the possibility of allelic bias or dropout must be considered; variants affecting a primer binding site may suppress priming efficiency. Similarly, insertions within an amplicon may reduce or eliminate amplification and deletions internal to an amplicon may enhance its amplification. Clearly, deletions which destroy a primer binding site will eliminate amplification. Overlapping amplicons are one approach to minimizing these ascertainment biases [54].

Hybridization methods use oligonucleotides or DNA fragments either in solution or on solid supports to bind to targets of interest [58]. These methods can detect the full range of mutations and rearrangements and have also been scaled to the entire set of coding exons (exome) [60]. However, approaches to date have shown highly variable coverage of targeted regions, with a minority of regions receiving insufficient coverage for reliable variant detection. Hybrid selection can preserve copy number information, though multiple rounds of hybridization (which improve specificity) may degrade the copy number signal [60]. From a bioinformatic perspective, it is critical to align reads from targeted sequencing experiments to the entire genome and not just the targeted region. Not only are 15% or

more of the sequences from untargeted regions due to carry-over [60], but also it is critical to ensure that other regions are included to detect rearrangements and to correctly assign reads to genes rather than their paralogs or pseudogenes. Hybrid selection strategies, as with PCR, may exhibit bias for reference alleles when the variant position is contained within the targeting oligo [29].

Targeted sequencing can also be a valuable means to confirm mutations discovered by second-generation technologies as well as to follow novel mutations longitudinally. In a breast cancer study, the initial whole-genome sequencing was based on DNA from a metastasis. Non-synonymous coding variants identified from this sample were PCR amplified from the patient's normal DNA to distinguish somatic mutations from germline variants. The same PCR strategy was then used to determine which mutations had been present in the primary tumor [36].

Each second-generation sequencing read results from a single molecule or clone of identical molecules, enabling rare allele identification. This has been exploited to identify oncogenic [12] or chemotherapy resistance [61] mutations, with sensitivities of <0.1% have been reported [61]. In these cases, targeted sequencing is essential to assure a very high sampling depth for the region of interest. Ultimately, the inherent noise of the system will limit sensitivity [2, 14].

The clonal nature of second-generation sequencing can reveal the population structure of a tumor. Many B-cell malignanies arise from a B-cell which has already undergone a productive heavy-chain VDJ rearrangement. However, somatic hypermutation after the rearrangement will cause sequence divergence between cells. Targeted deep sequencing of the heavy chain locus and analysis by standard phylogenetic techniques demonstrated both a diversity of genotypes and a highly dominant clone. This suggests that the dominant clone became so through the acquisition of additional driver mutations after the initial oncogenic event [62]. Sequencing of rearranged immunoglobulin loci can also identify myeloproliferative disorders arising from multiple clones as well as be used to monitor residual disease [63]. In an AML patient, a similar approach demonstrated heterogeneity for the presence of the oncogenic FLT3 internal tandem duplication allele in both the original tumor and a relapse sample [37].

RNA-Seq

Whole transcriptome shotgun sequencing, also known as RNA-Seq, offers the opportunity to collect a range of information from a cancer sample. Newer sample preparation methods enable RNA-Seq from a few hundred nanograms [64] or even a few hundred picograms of total RNA [65].

As a mechanism to detect oncogenic point mutations, there has already been one spectacular RNA-Seq success. Granulosa cell tumors (GCT) of the ovary are rare cancers which can present either early or late in life and had been suggested to have a low rate of tumor mutation and rearrangement. RNA-Seq of only 16 ovarian tumors, four of which were GCTs, identified a recurrent mutation unique to the adult-type GCT samples. A validation set of 27 additional adult-type GCTs showed universal presence of the mutation whereas only one of eight juvenile GCTs contained this. Only two out of 60 other tumors contained the distinctive mutation, and these were both in a related tumor (thecoma) which may not be truly distinct from GCTs. The prior identification of germline FOXL2 loss-of-function alleles in premature ovarian failure provided evidence for the importance of this gene in normal ovarian development [66].

Gene fusions drive a number of tumors, with the 5′ fusion partner often providing transcriptional activation whereas the 3′ partner possesses inherent oncogenic potential [67]. Most fusions result from chromosomal aberrations, though some arise via transcriptional read-through from the 5′ fusion partner to a neighboring gene, some of which appear to recur in a tumor type [19, 42]. One approach for gene fusion detection relies on the 454 platform's long reads to identify breakpoint-crossing reads [68–71]. Alternatively, numerous short reads from the Illumina platform may cross a breakpoint [18]. The odds of detecting fusions to known or suspected fusion partners can be increased by hybrid selection, which also captures the greatest depth and diversity of fusion transcript isoforms [72]. Combination of long (454) and short (Illumina) can enable confident identification of rare fusion transcripts with the long read supplying a template for alignment of multiple short reads [73]. Paired reads offer even greater sensitivity for fusion transcript identification than individual reads [19, 74, 75]. Deep sequencing with paired ends is sufficiently sensitive to detect highly expressed fusion transcripts in pooled RNA [19].

RNA-Seq can also reveal non-mutational events. RNA-Seq yields expression information with a superior dynamic range to microarrays and with details of the transcript isoform distribution [3]. RNA-Seq may also suggest cases of imbalance in the transcription of two alleles of a gene [27, 51], suggesting the partial or complete silencing of one [36]. Comparison of RNA-Seq data with whole genome data from a breast tumor suggested that many sites had undergone RNA editing [36], a dimension of genetic diversity which has received little attention in the context of cancer.

While RNA-Seq can be a valuable tool for mutation and fusion discovery, its sensitivity will be limited by the expression level of the altered gene. The identification of point mutants in oncogenic genes of low expression level will remain challenging even with very deep transcriptome sequencing.

MUTATION SPECTRA AND CHARACTERIZATION

Whole genome sequencing offers an opportunity for the complete census of mutations within a tumor. A number of these mutations either have suggestive or direct evidence for participation in oncogenesis (Table 1); bioinformatic approaches to the challenge of identifying these will be discussed in the next section. In addition to identifying specific causative mutations, deep sequencing has identified a wide range of mutation loads, ranging from 10 missense mutations each in two leukemia genomes [37, 38] to nine times as many in a breast cancer [36] and small cell lung cancer [16] and 17 times as many in melanoma [17].

Careful analysis of the specific classes of mutations found and their distribution between transcribed and non-transcribed regions has revealed several trends. In melanoma, C > T/G > A changes overwhelmingly predominated and CC > TT/GG > AA mutations accounted for more than half of all doublet mutations, in agreement with known patterns of UV-mutagenesis [17]. Many of the remaining mutations were suggestive of oxidative damage [17]. In contrast, the small cell lung cancer sample had three major classes of mutations in a pattern in agreement with prior data from p53 mutations in the same tumor type [16]. CpG dinucleotides were enriched both for G > A transitions and G > C transversions, but these classes differed in their prevalence within CpG islands [16]. Conversely, GpA dinucleotides are mutated less frequently than expected by chance but TpA more frequently [16]. Single base insertions were most commonly gains of A or T but single base deletions favored C or G [16]. Overall, WGS of small cell lung cancer paints a complex picture of the mutagenic effects of cigarette smoke [16].

Mutations were much less prevalent in transcribed regions, illustrating the effects of transcription-coupled repair [17]. Furthermore, mutations were more likely to be found near the 3′ end of transcribed regions than the 5′ end, suggesting that abortive transcription contributes to more effective transcription-coupled surveillance [17]. Similarly,

Table 1: Cancer driver mutations discovered by second-generation sequencing

Tumor type	Method	Gene	Mutation type	Effect	Validation	Ref.
Lung, small cell	WGS	PVT1-CHD7	Fusion	Activating	Elevated expression in multiple samples	[16]
AML	WGS	IDH1	Mutation	Activating	Mutations in 16/188 samples	[38]
AML	WGS	ND4	Mutation	Unknown	Mitochondrial; highly enriched in three tumor samples	[38]
Prostate	RNA-Seq	HERPUD1-ERG	Fusion	Activating	Androgen-dependent expression	[19]
Prostate	RNA-Seq	FLJ35294-ETV1	Fusion	Activating	Androgen-dependent expression	[19]
Prostate	RNA-Seq	SLC45A3-ELK4	Fusion (read through)	Activating	Androgen-dependent expression; recurrence	[73]
Ovarian, Granulosa Cell Tumor	RNA-Seq	FOXL2	Mutation	Activating	Recurrence	[66]
Lung, Adenocarcinoma	RNA-Seq	R3HDM2-NFE2	Fusion	Activating	RNAi; elevated expression of NFE2 in multiple samples	[74]
Breast	WGS	ARFGEF2-SULF2	Fusion	Activating	RNAi	[92]
Melanoma	RNA-Seq	RBI-ITM2B	Fusion	Inactivating?	Recurrence	[42]
B-cell lymphoma		EZH2	Mutation	Inactivating	Recurrence	[93]

more highly expressed genes were less likely to be mutated than genes with low expression [16]. In melanoma, C > T and CC > TT mutations were highly strand-biased in transcribed regions, congruent with the bias of transcription-coupled repair for the transcribed strand [17]. Different mutational classes in the lung cancer sample showed different degrees of bias for transcribed regions, suggesting varying effectiveness of repair [16]. However, nearly two-thirds of the mutational deficit in transcribed regions does not fit the patterns expected for known mechanisms of nucleotide excision repair, suggesting that one or more additional mechanisms remain to be discovered [16, 17].

In melanoma the mutation spectrum for regions of copy-neutral loss of heterozygosity differed between heterozygous and homozygous mutations, suggesting different mutational processes [17]. Homozygous mutations (which must have formed before the chromosome reduplication that led to copy-neutral LOH) showed a pattern consistent with sun exposure whereas the heterozygous mutations (formed after reduplication) did not, suggesting that these had formed after metastasis [17]. As additional cancer genomes are sequenced, careful mining the combined data may identify additional mutational patterns inexplicable by known mutational or repair processes or which illuminate the temporal mutational history of tumors.

Given the large number of mutations generated by these studies, methods to differentiate biologically significant changes from irrelevant ones will remain critical. The COSMIC database records published somatic mutations in cancer, enabling the identification of recurrently mutated genes or sites [76]. General-purpose tools such as SIFT [77], MutPred [78] and CanPredict [79] use information from multiple alignments to estimate the effect of mutations. CHASM is trained specifically to find driver mutations based on random forest classification of 49 protein structure, evolutionary history and genomic context features [80]. Known cancer driver mutations may have different properties than phenotype-altering coding SNPs, which may be exploited for further driver discovery [81]. Structure-based tools attempt to identify substitutions likely to disturb the packing of amino acids within the 3D structure [82, 83]. Special-purpose tools have also been developed which focus on particular gene families recurrently mutated in cancer, such as protein kinases [84–86]. Other tools have specialized on particular features, such as signal peptides [87]. Analysis of biological networks may suggest the mechanism by which gene mutations induce oncogenesis [88]. Pleasance *et al*. [16] also identified one mutation potentially altering a transcriptional regulatory site, though the functional significance of neither this site nor the change has been experimentally tested. The flood of data from such projects will increase the need for tools to prioritize potential gene regulatory mutations for functional testing [89].

As with point mutations, many fusions may simply be 'passenger' events due to general genomic instability or the selection for amplification of an adjacent gene. The databases COSMIC [76] and ChimerDB [90] record known fusions from the literature. Some genes may be activated by fusion in some cancers and by other mechanisms in other [74]. Recurrent rearrangements or high expression of rearranged transcripts are clues to biological relevance. Computational methods for prioritizing fusion proteins for functional characterization have emerged. One approach relies on the observation that known oncogenic fusion protein partners are enriched for certain functional classes (such as kinases or transcriptional activators), an enrichment which is most pronounced for the 3′ partner [74]. An alternative approach notes that oncogenic fusion breakpoints occur in disordered regions of proteins and not well folded domains [67].

Despite this complexity, second-generation sequencing has already identified a number of candidates for oncogenic driver mutations which have supporting experimental or expression evidence (Table 1). Evidence supporting the causative nature of these mutations can include recurrence, analogy to well characterized driver mutations and systems biology inference of likely impact on oncogenic pathways. Ultimately, experimental validation will be required to validate these mutations and measure their impact on cancer initiation and progression.

CONCLUSIONS AND LOOKING FORWARD

Second-generation sequencing has already demonstrated great utility for identifying mutations in cancer, despite the application to date in only a limited number of samples. Cancer genomes will continue to present bioinformatics challenges in terms of the large degree of genomic alteration and the prioritization of genes for functional analysis.

Refinement of existing technologies and the introduction of new technologies [32, 91] will continue to reduce the cost of this approach, enabling it to be applied across the diverse spectrum of cancer types. As the price of sequencing drops to a several thousand dollars or less per sample [2, 32, 60, 91], cancer genomics will become a standard part of the diagnostic arsenal, enabling rational selection of therapeutics [12] and the tracking of tumor burden through minimally invasive methods [43]. This will bring new bioinformatics challenges to the forefront, such as choosing therapies based on genomic information and storing genomic information within standard electronic medical records.

Key Points

- Second-generation sequencing methods, including targeted sequencing, transcriptome sequencing and whole genome sequencing, are emerging as a key approach to survey genomic changes in cancer.
- Cancer genomes present specific bioinformatics challenges due to the short read lengths of most second-generation technologies.
- Second-generation sequencing discovery of mutations is outpacing experimental and computational methods for determining which are drivers contributing to disease.

Acknowledgements

The author wishes to thank John Keilty, Guillermo Paez for critical review of the manuscript and the anonymous reviewers for many helpful comments.

FUNDING

Infinity Pharmaceuticals.

References

1. Luo J, Solimini NL, Elledge SJ. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* 2009;**136**: 823–37.
2. Fuller CW, Middendorf LR, Benner SA, *et al*. The challenges of sequencing by synthesis. *Nat Biotechnol* 2009;**27**: 1013–23.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
4. Bathe OF. McGuire AL, The ethical use of existing samples for genome research. *Genet Med* 2009;**11**:712–5.
5. Schweiger MR, Kerick M, Timmermann B, *et al*. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. *PLoS One* 2009;**4**: e5548.
6. Agell L, Hernandez S, de Muga S, *et al*. KLF6 and TP53 mutations are a rare event in prostate cancer: distinguishing between Taq polymerase artifacts and true mutations. *Mod Pathol* 2008;**21**:1470–8.
7. Pinard R, de Winter A, Sarkis GJ, *et al*. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 2006;**7**:216.
8. Rodrigue S, Malmstrom RR, Berlin AM, *et al*. Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* 2009;**4**:e6864.
9. Rubin AF, Green P. Mutation patterns in cancer genomes. *Proc Natl Acad Sci USA* 2009;**106**:21766–70.
10. White RA, 3rd, Blainey PC, Fan HC, *et al*. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* 2009;**10**:116.
11. Meyer M, Briggs AW, Maricic T, *et al*. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res* 2008;**36**:e5.
12. Thomas RK, Nickerson E, Simons JF, *et al*. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med* 2006;**12**:852–5.
13. Nickerson DA, Tobe VO, Taylor SL. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* 1997;**25**:2745–51.
14. Druley TE, Vallania FL, Wegner DJ, *et al*. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 2009;**6**:263–5.
15. Loeb LA, Bielas JH, Beckman RA. Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res* 2008;**68**: 3551–7; discussion 3557.
16. Pleasance ED, Stephens PJ, O'Meara S, *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010;**463**:184–90.
17. Pleasance ED, Cheetham RK, Stephens PJ, *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010;**463**:191–6.
18. Morin R, Bainbridge M, Fejes A, *et al*. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 2008;**45**: 81–94.
19. Maher CA, Palanisamy N, Brenner JC, *et al*. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci USA* 2009;**106**:12353–8.
20. McKernan KJ, Peckham HE, Costa GL, *et al*. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* 2009;**19**:1527–41.
21. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009;**27**:847–52.
22. Bashir A, Volik S, Collins C, *et al*. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* 2008;**4**: e1000051.
23. Li R, Fan W, Tian G, *et al*. The sequence and de novo assembly of the giant panda genome. *Nature* 2010;**263**: 311–7.

24. Li H, Homer N. A survey of sequence alignment for next-generation sequencing. *Briefings in Bioinformatics* 2010, in press.
25. Lam HY, Mu XJ, Stutz AM, *et al*. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 2010;**28**:47–55.
26. Li R, Li Y, Zheng H, *et al*. Building the sequence map of the human pan-genome. *Nat Biotechnol* 2010;**28**:57–63.
27. Degner JF, Marioni JC, Pai AA, *et al*. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 2009;**25**:3207–12.
28. Krawitz P, Rodelsperger C, Jager M, *et al*. Microindel detection in short-read sequence data. *Bioinformatics* 2010; **26**:722–9.
29. Clark MJ, Homer N, O'Connor BD, *et al*. U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet* 2010;**6**: e1000832.
30. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. *PLoS One* 2009;**4**: e7767.
31. Weber G, Shendure J, Tanenbaum DM, *et al*. Identification of foreign gene sequences by transcript filtering against the human genome. *Nat Genet* 2002;**30**:141–2.
32. Drmanac R, Sparks AB, Callow MJ, *et al*. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010;**327**:78–81.
33. Li R, Zhu H, Ruan J, *et al*. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 2010;**20**:265–72.
34. Ye K, Schulz MH, Long Q, *et al*. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**:2865–71.
35. Dalca AV, Brudno M. Genome variation discovery with high-throughput sequencing data. *Briefings in Bioinformatics* 2010;**11**:3–14.
36. Shah SP, Morin RD, Khattra J, *et al*. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009;**461**:809–13.
37. Ley TJ, Mardis ER, Ding L, *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008;**456**:66–72.
38. Mardis ER, Ding L, Dooling DJ, *et al*. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 2009;**361**:1058–66.
39. Goya R, Sun MG, Morin RD, *et al*. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* 2010;**26**:730–6.
40. Koboldt DC, Chen K, Wylie T, *et al*. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;**25**:2283–5.
41. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009;**6**:S13–20.
42. Berger MF, Levin JZ, Vijayendran K, *et al*. Integrative analysis of the melanoma transcriptome. *Genome Res* 2010; **20**:413–27.
43. Leary RJ, Kinde I, Diehl F, *et al*. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med* 2010;**2**:20ra14–20ra14.
44. Lee S, Hormozdiari F, Alkan C, *et al*. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* 2009;**6**:473–4.
45. Chen K, Wallis JW, McLellan MD, *et al*. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 2009;**6**:677–81.
46. Korbel JO, Abyzov A, Mu XJ, *et al*. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;**10**:R23.
47. Hormozdiari F, Alkan C, Eichler EE, *et al*. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 2009;**19**: 1270–8.
48. Long Q, MacArthur D, Ning Z, *et al*. HI: haplotype improver using paired-end short reads. *Bioinformatics* 2009;**25**: 2436–7.
49. Pao W, Miller VA, Politi KA, *et al*. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med* 2005;**2**:e73.
50. Campbell PJ. Somatic and germline genetics at the JAK2 locus. *Nat Genet* 2009;**41**:385–6.
51. Tuch BB, Laborde RR, Xu X, *et al*. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One* 2010; **5**:e9317.
52. Chiang DY, Getz G, Jaffe DB, *et al*. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 2009;**6**:99–103.
53. Stephens PJ, McBride DJ, Lin M-L, *et al*. Complex landscapes of somatic rearrangement in human breast cancer genomes 2009;**462**:1005–10.
54. Harismendy O, Ng PC, Strausberg RL, *et al*. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009; **10**:R32.
55. Rougemont J, Amzallag A, Iseli C, *et al*. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 2008; **9**:431.
56. Dohm JC, Lottaz C, Borodina T, *et al*. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008;**36**:e105.
57. Quail MA, Kozarewa I, Smith F, *et al*. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 2008;**5**:1005–10.
58. Turner EH, Ng SB, Nickerson DA, *et al*. Methods for genomic partitioning. *Annu Rev Genomics Hum Genet* 2009; **10**:263–84.
59. Tewhey R, Warner JB, Nakano M, *et al*. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* 2009;**27**:1025–31.
60. Lee H, O'Connor B, Merriman B, *et al*. Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. *BMC Genomics* 2009;**10**:646.
61. Yauch RL, Dijkgraaf GJ, Alicke B, *et al*. Smoothened mutation confers resistance to a Hedgehog pathway inhibitor in medulloblastoma. *Science* 2009;**326**:572–4.
62. Campbell PJ, Pleasance ED, Stephens PJ, *et al*. Subclonal phylogenetic structures in cancer revealed by

ultra-deep sequencing. *Proc Natl Acad Sci USA* 2008;**105**: 13081–6.

63. Boyd SD, Marshall EL, Merker JD, *et al*. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* 2009;**1**: 12ra23.
64. Mamanova L, Andrews RM, James KD, *et al*. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 2010;**7**:130–2.
65. Ozsolak F, Goren A, Gymrek MA, *et al*. Digital transcriptome profiling from attomole-level RNA samples. *Genome Res* 2010;**20**:519–23.
66. Shah SP, Kobel M, Senz J, *et al*. Mutation of FOXL2 in granulosa-cell tumors of the ovary. *N Engl J Med* 2009;**360**: 2719–29.
67. Hegyi H, Buday L, Tompa P. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput Biol* 2009;**5**:e1000552.
68. Zhao Q, Caballero OL, Levy S, *et al*. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci USA* 2009;**106**:1886–91.
69. Guffanti A, Iacono M, Pelucchi P, *et al*. A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics* 2009;**10**:163.
70. Bainbridge MN, Warren RL, Hirst M, *et al*. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 2006;**7**: 246.
71. Sugarbaker DJ, Richards WG, Gordon GJ, *et al*. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc Natl Acad Sci USA* 2008;**105**:3521–6.
72. Levin JZ, Berger MF, Adiconis X, *et al*. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* 2009;**10**:R115.
73. Maher CA, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458**:97–101.
74. Wang XS, Prensner JR, Chen G, *et al*. An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat Biotechnol* 2009;**27**: 1005–11.
75. Ruan Y, Ooi HS, Choo SW, *et al*. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* 2007;**17**:828–38.
76. Forbes SA, Tang G, Bindal N, *et al*. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res* 2010;**38**:D652–7.
77. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**:1073–81.
78. Li B, Krishnan VG, Mort ME, *et al*. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics,* 2009;**25**:2744–50.
79. Kaminker JS, Zhang Y, Watanabe C, *et al*. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res* 2007;**35**:W595–8.
80. Carter H, Chen S, Isik L, *et al*. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009;**69**: 6660–7.
81. Talavera D, Taylor MS, Thornton JM. The (non)malignancy of cancerous amino acidic substitutions. *Proteins* 2010;**78**:518–29.
82. Hurst JM, McMillan LE, Porter CT, *et al*. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Hum Mutat* 2009;**30**:616–24.
83. Izarzugaza JM, Baresic A, McMillan LE, *et al*. An integrated approach to the interpretation of single amino acid polymorphisms within the framework of CATH and Gene3D. *BMC Bioinformatics* 2009;**10**(Suppl 8):S5.
84. Dixit A, Yi L, Gowthaman R, *et al*. Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS One* 2009;**4**:e7485.
85. Torkamani A. Schork NJ, Prediction of cancer driver mutations in protein kinases. *Cancer Res* 2008;**68**: 1675–82.
86. Izarzugaza JM, Redfern OC, Orengo CA, *et al*. Cancer-associated mutations are preferentially distributed in protein kinase functional sites. *Proteins* 2009;**77**:892–903.
87. Hon LS, Zhang Y, Kaminker JS, *et al*. Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. *Hum Mutat* 2009;**30**:99–106.
88. Torkamani A, Schork NJ. Identification of rare cancer driver mutations by network reconstruction. *Genome Res* 2009;**19**:1570–8.
89. Torkamani A, Schork NJ. Predicting functional regulatory polymorphisms. *Bioinformatics* 2008;**24**:1787–92.
90. Kim P, Yoon S, Kim N, *et al*. ChimerDB 2.0–a knowledgebase for fusion genes updated. *Nucleic Acids Res* 2010;**38**: D81–5.
91. Eid J, Fehr A, Gray J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**:133–8.
92. Hampton OA, Den Hollander P, Miller CA, *et al*. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res* 2009;**19**: 167–77.
93. Morin RD, Johnson NA, Severson TM, *et al*. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* 2010;**42**:181–185.