Annals of Oncology 22 (Supplement 1): i37–i43, 2011
doi:10.1093/annonc/mdq664

ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE
European Society for Medical Oncology

symposium article

Innovative technology for cancer risk analysis

S. Tommas*, K. Danza, B. Pilato & S. De Summa

National Cancer Centre 'Giovanni Paolo II', Bari, Italy

After completion of the Human Genome Project, analysis of genetic and genomic variations in different pathological states became possible. The capillary system based on Sanger methods is still very expensive in terms of time, cost and professionalism required. For this reason, the National Human Genome Institute proposed an 'advanced sequencing technology development' project with the aim of sequencing a genome in 1 day for $1000. Three validated platforms are commercially available and single molecule sequencing methods have been recently introduced, which are not only competitive in time and costs, but display greater accuracy than 'past generation' sequencing. Next generation technology allows, in a single experiment, the identification of copy number variation and large rearrangements, or detection of fusion transcripts analysis thus permitting the evaluation of cancer risk at multiple levels (genomic, transcriptomic, proteomic, epigenetic).

introduction

In the last few years, how to implement sequencing approaches leading to shortened time and lower costs, has become an intriguing challenge for many research groups. New methods, named 'next generation' sequencing, are not restricted to merely sequencing but allow full analysis of transcriptome, proteome and epigenetic alterations that occur in pathological states.

Cancers encompass disease states that include complex genetic alterations and their relationship to environmental factors included under the term 'life style'. Mutations leading to an altered protein structure and large genomic structural changes, e.g. genomic rearrangements and copy number variation, are the subject of current study. Traditional sequencing methods and array approaches have too many limitations to allow research work to respond in a timely manner to the questions arising from human genome sequencing.

The Human Genome Project (HGP) constituted a milestone that initiated a new era in the field of molecular biology. From the beginning of this fascinating challenge until the present, improvements in sequencing methods have added to its fascination. In the HGP, BAC-based sequencing was used. More than 20 000 BACs have been created, containing 100-kb fragments of human genome. BACs were amplified in bacterial culture, isolated and sheared to obtain 2- to 3-kb fragments, which were sequenced and reassembled by computer into other BACs. After completion of the HGP [1, 2], the BAC-based approach was replaced by faster assembly techniques. In whole genome sequencing (WGS), genomic fragments are directly sheared into different size classes and subcloned in plasmid and fosmid vectors (e.g. shotgun sequencing) [3]. New algorithms were created to assemble fragments in a more rapid way but highly polymorphic and repetitive regions cannot be easily assembled. At present, sequencing not only a genome but even a single gene is expensive, time consuming, requires high professionalism, special equipment and quality control.

The 'advanced sequencing technology development project', was initiated by National Human Genome Research Institute (NHGRI) in 2004 with the aim of sequencing a genome in 1 day for $1000. Many laboratories are actually committed to the creation of such 'next generation' sequencing methods.

At present, three validated platforms are available: Roche/454 FLX, Applied Biosystems Solid™ and Illumina/Solexa Genome Analyzer. Recently, two new sequencing systems have been announced: Helicos Heliscope ™ and Pacific Biosciences SMRT. From the HGP, which was estimated to have a cost of $3 billion [3], rates have become considerably lower. Resequencing with the Sanger method has a cost of about $10 million [4, 5]; with Roche/454 platform a 10-fold reduction in cost and 20-fold reduction in time have become possible while the Illumina system enables a human genome to be sequenced for $100 000 (Illumina, Analyst Day, 15 September 2007, Mandarin Oriental, New York, NY, USA) (Figure 1).

These new technologies made possible the opening of new fields and new applications in molecular biology and medicine, such as the precise analysis of RNA transcripts for gene expression and identification of DNA regions that interact with regulatory proteins, profiling of mRNA, small RNAs, chromatin structure and DNA methylation patterns. In analysis for hereditary cancer risk, next generation sequencing not only can reduce time and costs but, replacing array technology, can provide a complete spectrum of single nucleotide polymorphisms (SNPs) and of mutations still unidentified.

*Correspondence to: S. Tommasi, Clinical Experimental Laboratory, National Cancer Centre 'Giovanni Paolo II', Via Hahnemann 10, 70126 Bari, Italy. Tel: +39-080-5555527; E-mail: s.tommasi@oncologico.bari.it

	ABI Sanger	Roche 454/FLX	Illumina GA	SOLID	HELICOS	SMRT
Base accuracy	98%	97.4%	100%	99.7%	98.5%	99.999%
Cost/Mb	$500	$60	$2	$2	$1	<$1

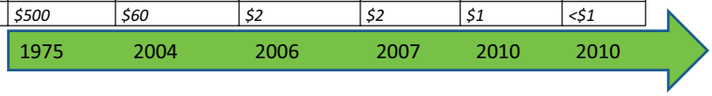

Figure 1. Base accuracy and cost per megabase using the next generation approaches with respect to the Sanger method.

the past and present generations of sequencing

In 1977, the Maxam and Gilbert method [6] and the Sanger technique [7] were introduced. The first method, based upon chemical degradation of DNA fragments, is no longer used because it uses toxic chemical compounds and because it is not suitable for automated sequencing. The Sanger enzymic dideoxy method needs a single-strand DNA fragment, a specific primer, a highly processive DNA polymerase and a mixture of four dNTPs and one ddNTP to block DNA synthesis. Initially, operators prepared four reactions, which were analyzed on gel; since the 1990s, Sanger DNA sequencing has become automated, with capillary-based systems [8, 9]. This system was used for the first sequencing of a genome locus, HPRT [10]. Overcoming the limitations of Sanger sequencing first used gels or polymers to separate labeled DNA fragments, as introduced by EMBL patent application (W. Ansorge, EMBL, Heidelberg, 1991) and then a process for sequencing nucleic acids without gel, sieving media on solid support and DNA chips (Verfahren zur Sequenzierung von Nukleinsauren ohne Gele, German Patent Application DE 4141178A1 and Corresponding Worldwide Patent Applications). This method was the first 'sequencing-by-synthesis' approach, consisting in detecting the next added labeled base (reversible terminator) by a CCD camera and dye removal for synthesis continuation. The EMBL application can be run on a great number of samples in parallel and on DNA chips, minimizing reaction volume.

A further refinement of this approach, the capillary system based on the Sanger method (e.g. ABI Sanger) became the 'official' technique of the HGP and it is the only method now used in laboratories. It has a cost of $0.50 per kilobase with a read length of ~1000 bp and a 99.9999% per-base accuracy.

To outperform this method, other attempts also based on chip approaches have been made. A fluorimetric SNP detection strategy based on the proportional hybridization of test and reference material with an oligonucleotide array platform has been designed to encompass the *BRCA1* entire coding region, SNPs and/or rearrangements (Monaco A, Tommasi S, Paradiso A US Patent Application # 61/068,182 filed 4 March 2008). This approach is able to discover known and unknown alterations including point mutations, deletions and insertions and large rearrangements using a single chip expressly designed using newly implemented software (Menolascina F, Tommasi S. Bevilacqua V, Paradiso A International patent pending # PCT/IB2007/054589 filed November 2007). This method has been validated [11] to be specific and accurate but is still expensive and time consuming to allow entire genome analysis.

the next generation of sequencing

Roche/454 FLX (http://www.454.com/)

This platform became commercially available in 2004 and it is based on a pyrosequencing method [12]. Library fragments are mixed with agarose beads carrying oligonucleotides complementary to a 454-specific adapter. These complexes are isolated into micelles containing polymerase chain reaction (PCR) reactants (oil–water micelles). In this way, after thermal cycling, ~1000 000 copies of each DNA fragment are present on the bead surface. Beads are arrayed into a picotiter plate, which contains a bead per well. After this step, beads containing ATP sulfurylase and luciferase are added to monitor sequencing reactions. Enzymes cause release of pyrophosphate, which labels dNTP and the light emitted is detected by a CCD camera. A limit is misinterpretation of homopolymer [13] but there are no substitution errors because incorporation is base specific at each step.

In a 7.5 h run ~100 million bases can be sequenced with a 99.5% raw base accuracy; this platform has a 200- to 250-bp read length [14]. A single run costs approximately $210 000.

Illumina Genome Analyzer (http://www.illumina.com/systems/genome_analyzer.ilmn)

Work by Turcatti and his group [15, 16] is the origin of this platform based on a 'sequencing-by-synthesis approach'. Amplification is realized with bridge PCR [15, 17]. Genomic DNA samples are prepared randomly and linked with an adapter at each end; forward and reverse primers are immobilized on a solid substrate by a flexible linker. All four nucleotides, fluorescently marked and 3'-OH blocked, are added. After this step, unused reactants are washed away and optics systems scan each lane. When an image is obtained, blocking groups are removed and the next cycle begins. The number of cycles is user defined and this allows a 25- to 35-bp read length with raw base-calling accuracy of ~98.5%. A common error of this system is base substitution. Cost of a run ($6300) is lower than one of the Roche/454 systems. In 2008, an upgrade (Genome Analyzer II) was introduced with a paired-end module and a new optics system that allows a reduction in run time for a 36-cycle run to 2 days for a single-read run and 4 days for a paired-end run [18].

Applied Biosystems Solid™ Sequencer (http://www.appliedbiosystems.com)

In 2007, the Solid system was introduced after work by McKernan [19] and Shendure [20]. The first step is the creation of an adapter-ligated fragment library. Emulsion PCR products with amplicons ligated on the surface of paramagnetic beads

[21] are immobilized on a solid substrate. This platform is based on 'sequencing-by-hybridation and ligation' with a DNA ligase approach [22, 23] (Macevicz, SC. DNA sequencing by parallel oligonucleotide extensions. US patent 5750341, 1998).

At each step, a population of octamer, fluorescently labeled at two central bases, is ligated to beads, optical systems detect labeled base and then fluorophore is removed. Two-base encoding allows errors to be avoided because each base is queried twice. This system has a 25- to 35-bp read length with a cost per run of about $7700.

Single-molecule sequencing: Helicos HeliScope™ Sequencer (www.helicosbio.com/) and SMRT Pacific Biosciences (http://www.pacificbiosciences.com)

The Helicos platform has two flow cells where billions of molecules of DNA are captured on a surface. Fluorescently labeled dNTP and polymerase are added to allow elongation of complementary strands. The next step is to wash away dNTP not incorporated, imaging, removal of fluorescent groups and another cycle begins. This system interrogates a single molecule without an amplification step, thus lowering costs. Current applications are: targeted resequencing, whole-genome resequencing, digital gene expression, DNA–protein interaction studies, and in the future, digital copy number variation and small RNA analysis could become possible. The error rate is reduced by two-pass sequencing, i.e. doubly read at the same position. Homopolymer read has high accuracy because analysis of every single molecule avoids the dephasing problem inherent in other methods due to asynchronous sequencing. In fact, in other sequencing approaches many molecules of dNTPs are incorporated in the same cycle and sequencing of templates with and without homopolymer is not synchronized.

In the Single-molecule Real Time Sequencing (SMRT) platform, sequencing is realized in a cell containing hundreds of zero-mode waveguides (ZMWs). A ZMW is a 'hole' of 10–50 nm in diameter that functions as a nanophotonic visualization chamber with a detection volume of 20 zeptoliters (10^{-21} l). A very high signal-to-noise ratio renders detection easy. This system allows resequencing and *de novo* sequencing. In both cases, templates are circular and a strand-displacing enzyme makes possible many reads of the same molecule, increasing the quality score.

Figure 2 summarizes the three methodological approaches.

applications of next generation sequencing

Next generation sequencing systems have been used in chromatin immunoprecipitation (ChIP) sequencing, a technique that allows identification of DNA regions that interact with protein by combining chromatin immunoprecipitation with next generation sequencing [24, 25], transcriptome studies [26, 27] and WGS [28–30]. Actually, there is a great interest in the sequencing of specific genes or of regions identified with SNP-association studies or of the exome [31, 32]. To better understand the power of next generation platforms, Harysmendy et al. [33] amplified six genomic regions coding for K^+/Na^+ voltage-gated channels from four individuals. Amplicons were mixed in equimolar quantities and

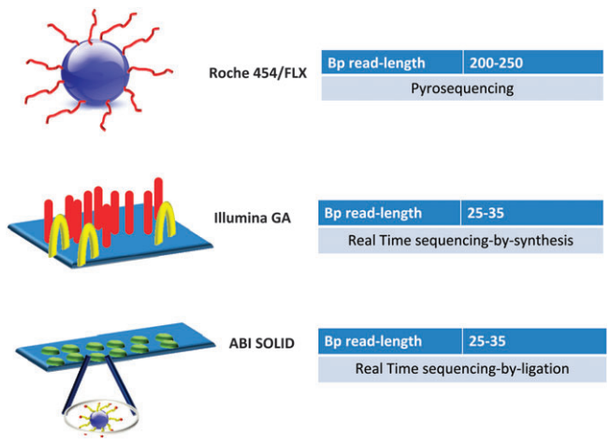

Figure 2. The three most validated next generation approaches. Base pair read length and technical models have been represented.

they were sequenced with three platforms and with ABI 3730XL or ABI Sanger. Accuracy was evaluated by comparing data of 80 SNPs assayed with Illumina Hap550 BeadChip. Genotype accuracy is 97.4% for Roche 454/FLX, 100% for Illumina, 97.7% for ABI Solid and 98% for ABI Sanger. However, comparison is not indicative because microarray does not identify novel variants. Comparison with ABI Sanger is much more important because this is the system actually used in almost every laboratory. Sequencing accuracy is 99.99% for three platforms; variant accuracy, i.e. the ability to identify a variant found with ABI Sanger, is 95% for Roche 454/FLX; 100% for Illumina; 96% for ABI Solid. The false-positive rate is 2.5% for Roche 454/FLX; 6.3% for Illumina; 7.8% for ABI Solid; the false-negative rate is 3.1% for Roche 454/FLX; 0% for Illumina; 0.9% for ABI Solid. Values of false-negative rates are comparable with those of ABI Sanger (i.e. 3%) [34]. Variant discrepancy—i.e. incorrect assignation of zygosity to a correctly identified variant—is 2% for Roche 454/FLX; 0% for Illumina; 3% for ABI Solid. Therefore, short-read Illumina and ABI Solid platforms are more sensitive than 454 but less specific [33]. Indel mutations are simply not detected with ABI Sanger but next generation platforms can make detection of these mutations easier by the independent sequencing of each allele.

cancer genome in the 'next generation' era

Genetic screening through targeted sequencing is a widespread diagnostic practice but it remains expensive and time-consuming. Pre-screening approaches (dHPLC and High Resolution Melting (HRM) techniques) allow a reduction in the number of sequencing reactions and they have been used in the study of acute lymphoblastic leukemia [35], lung cancer [36] and breast cancer [37]. The development in 1992 of comparative genomic hybridization marked the beginning of a new series of study in cancer genetics [38]. This approach is a molecular–cytogenetic method for the analysis of copy number changes (gains/losses) in the DNA content of enrolled subjects' DNA and often in tumor cells. The Progenetix

database collects karyotype information on many types of cancer [39]. For example, ovarian cancer genomic aberrations were annotated even if the identification of a target gene was difficult due to the complexity of such tumors. Mutations in PTEN and PIK3CA are more frequent in endometrioid subtype [40, 41], while high-grade serous tumors show loss of BRCA1 [42].

After completion of the HGP in 2003, analysis of whole-genome samples became possible for the identification of genetic variants that contribute to the onset of a disease. Genome-wide association studies (GWAS) generally requires two groups: cases, i.e. patients with disease, and controls, unaffected people. Frequent mutations are considered 'linked' to the specific disease. This technique allows the genetics of a disease to be investigated in a non-hypothesis-driven manner because the entire genome is screened but can lead to many false-positive results because of the large number of statistical tests [43].

SNPs detected by GWAS are listed in the HapMap Project with the exception of rare variants because such an approach identifies mutations with a frequency of >5%. In a three-stage GWAS, Thomas et al. [44] analyzed at stage 1, 528 173 SNPs in 1146 cases of invasive breast cancer and in 1142 controls present in HapMap Phase II [45]. At stage 2, 29 909 top SNPs were genotyped in 4547 cases and 4434 controls and at the final stage, 24 SNPs, 21 of which were chosen based on the results of preceding stages, were analyzed in 4078 cases and 5223 controls. In this way, they identified two new SNPs, 1p11.2 and 14p24.1. The first is located at the pericentromeric region containing a low-affinity Fc γ receptor family gene, *FCGR1B*, and downstream of the promoter of the *NOTCH2* gene. This study suggested an association between type 2 diabetes and postmenopausal breast cancer [46, 47]. The second SNP is within the *RAD51L1* gene, involved in break repair and the homologous recombination pathway. Ahmed et al. [48] identified two new breast cancer susceptibility loci, *SLC4A7* and *NEK10* on chromosome 3 and *COX11* on chromosome 17, in a four-stage WGA study. The first two steps, which identified seven breast cancer susceptibility loci, were conducted by Easton et al. [49] and the final step was designed to localize other loci involved in such diseases.

Next generation platforms allow genome characterization not only at a mutational level but also with the detection of large rearrangements and copy number variation in a single experiment [50, 51]. Resequencing is one of the applications possible with next generation approaches. The '1000 genomes project' (http://www.1000genomes.org), recently completed, and the Cancer Genome Atlas (http://cancergenome.nih.gov) provide a list of human variants, in particular focusing on rare ones. In cancer studies, the approach used is targeted sequencing (sequencing of target genes, including those present in regions that display copy number variations). However, a challenge is bridging the differences in approaches widely used between diagnostic and research laboratories.

Li et al. [52] compared error rates in SNP detection of the Illumina GA platform with Asian genome sequence with 36× high-quality data [53]. The Illumina platform has a phread-like quality score system to measure accuracy. Genotypes were inferred using a Bayesian statistical method, which is the same for Sanger sequencing [54] and introduced for next-generation sequencing (NGS) platform [55]. These investigators reported a reduced error rate of 0.5%–0.8% in reads of 35 bp and then higher accuracy than traditional Sanger sequencing. DNA copy number variations can be efficiently studied by the NGS platform as demonstrated by Castle et al. [56]. They developed a sequencing-based assay to study nuclear, mitochondrial and telomeric copy number using Illumina Genome Analyzer II. They prepared a genomic library from the UMC-11 cell line, a lung carcinoid-derived cell line, from blood of males and females not affected. Fragments, not selected to allow lab automation, were amplified using PCR and sequenced. They tested different read lengths and found that a 33-nucleotide length ensures lower costs and unambiguous read. Copy number variations, such as five times more mitochondria and four times less telomeric sequence were found with a lower error rate than array approaches.

The combination of next generation sequencing methods and GWAS, through linkage analysis of mutated loci in a particular population, can provide a new perspective in tumor biology and more predictive models.

SNP and mutation detection can be achieved through enrichment of genomic loci by microarray hybridization and next generation sequencing. By this approach, fragments of DNA capture probes printed on a microarray slide [57, 58]. Only probes that hybridized with fragments of interest are eluted and sequenced. Recently, enrichment of a DNA fragment library with short fragments (85–110 bp) was realized because many projects are focused on exons that have a maximum length of ~120 bp. In this reported study [59], the DNA sample used was first genotyped with Illumina HumanHap550+Genotyping BeadChip and 384 SNPs were present in the 1.69 Mb region analyzed; 96.4% of the region of interest was surveyed with a 4.2% false-negative rate and a positive rate of ~1 per 200 000 bp. Furthermore, 1197 novel variants were detected [59].

Development of bioinformatics tools to analyze output data from next generation platforms is very important. Recently, it was suggested that SNVMix models could correctly identify variants from next generation sequencing of a cancer genome [60]. In this study, two binomial models are proposed. In SNVMix 1, nucleotide base calls are assumed to be correct, and 'depth', i.e. the number of reads for each position, is the sum of a_j (number of reads that match the reference sequence) and b_j (number of reads that do not match). In SNVMix 2, the statistical significance of error in base calls and alignments is also considered. Maq [61] and SOAPSNP [62] approaches are also binomial models but they not consider that cancer genomes are in a pathological background (mixture of normal and tumor genotype). Goya et al. [60] tested models in a lobular breast cancer samples, comparing the accuracy of variant detection with the Maq model. They found that the accuracy of their models is higher and SNVMix 2 is more accurate than SNVMix 1 without achieving statistical significance.

beyond genome: transcriptome, proteome, methylome

Analysis of cancer risk is not limited to genome studies but it also includes changes at transcriptome, proteome and epigenetic levels. Next generation approach applications are summarized in Table 1.

Table 1. Next generation approaches can be used to investigate disease at different biological levels

Next generation applications		References
Genomic studies	SNP detection	[4, 29, 52, 55, 59, 60]
	Copy number variation	[50, 51, 56]
Transcriptome analysis	Splicing isoforms	[25, 26, 63, 64, 65, 66]
	Aberrant transcription	[67, 68]
Proteomic studies	Protein–DNA interactions	[23, 24, 69–72]
Epigenetic studies	Base methylation	[73, 74]

Transcriptome sequencing is one of the most interesting applications of next generation systems. Mane et al. [63] compared quantification of gene expression with results obtained with DNA microarray and quantitative RT-PCR using MAQC reference RNA samples [75, 76]. They used Roche/454 platforms, which provide data with high specificity and sensitivity and moreover, they discovered new splicing variants providing a more complex insight into the human transcriptome at lower cost. Microarrays do not allow identification of new variants; and therefore, sensitivity and specificity of next generation sequencing are important factors in cancer research to better understand tumorigenesis [77]. Morrissy et al. [64] developed a Tag-Seq method based on the Illumina Genome Analyzer, which is cheaper and more sensitive than LongSAGE [78] in cancer gene expression profiling. In both approaches, transcripts are cleaved by *Nia*III restriction endonuclease and then *Mme*I digestion generates a 21-bp tag. At this point in LongSAGE, tags are ligated to form ditags and sequenced, whereas in Tag-Seq, tags are directly sequenced. Difficulties in the SAGE method are due to the ligation step and cloning. Gene discovery of Tag-Seq is comparable to that of RNA-Seq [65, 66], also based upon Illumina Genome Analyzer, but Tag-Seq allows discrimination between sense and antisense transcripts. Known and novel sense–antisense gene pair expression was different in cancer state and controls and in particular they found antisense transcription at the BCL6 locus, which is involved in lymphomas. Libraries from a breast cancer patient and from grade II carcinoma epithelium samples show reduced sense-to-antisense ratio at this locus. Increase in antisense transcription can be due to hypomethylation because in epithelial carcinoma this locus is highly hypomethylated [79]. Transcriptome studies not only allow detection of expression changes but also identification of variants and rearrangements that have effects on gene functionality [67, 80]. Targeted sequencing through enrichment methods is the efficient solution also in transcriptome studies because it has been estimated that 40 million reads are necessary to obtain one fold coverage of an entire transcriptome [25]. Enrichment methods are molecular inversion probes (MIPs) [81, 82] and hybridization on a microarray surface [83, 84] or in solution [85]. Levin et al. [68], by solution hybrid selection and Illumina sequencing, studied 467 genes involved in tumorigenesis as reported in the Cancer Gene Census [86]. Before this study, it was reported that only 52 of all genes in analysis have splicing variants but Levin et al. [68] detected 177 genes with different splicing forms. Moreover, they identified two fusion transcripts, *BCR-ABL1* [87] and *NUP214-XKR3*.

Another important application is the study of interactions between proteins and DNA which are important in the regulation of gene expression. The ChIP approach was the first to be used [88] and recently a ChIP-chip method based on DNA microarray has been introduced [89]. The ChIP-chip approach is limited by its low signal-to-noise ratio and by the need for replicates to confirm discovery of a binding site. The ChIP-Seq technique replaces microarrays with next generation platforms [69]. Resolution is higher, <40 bp, and identification of new binding sites is possible, unlike the ChIP-chip approach.

Recently, the Wnt/β-catenin pathway, which is deregulated in colorectal cancers, was studied. The most common mutations are in the *APC* gene and such mutant cells contain a high level of β-catenin coactivator, which leads to deregulation of target genes. Bottomly et al. [70] created a Chip-Seq library to identify 2168 enriched β-catenin binding regions. They compared results with TCF4 binding regions detected using Chip-Seq [71, 72], finding that 786 regions are present in both libraries. Identification of the binding regions of the two transcription factors led to a better understanding of colorectal tumorigenesis at a higher resolution than other methods, e.g. SACO, which is a genome-wide approach that was used in similar work [90].

Not only targeted sequencing is predictive for cancer risk but also epigenetic alterations. Widschwendter et al. [73] showed that the DNA methylation pattern of peripheral blood cell DNA is predictive for breast cancer risk. Two alterations in methylation pattern were identified: it was observed that ER-α target genes are undermethylated when ER-α receptor is active, a factor that increases breast cancer risk [91, 92]; the second alteration is the hypermethylation of the promoter of polycomb group target genes [93–95]. Study of methylation pattern can be realized efficiently through SMRT sequencing in a more sensible and faster way than bisulfite conversion. It has been demonstrated that methylated bases alter polymerase kinetics, a factor that can be measured with primary sequence determination without additional steps [74].

conclusions

Analysis of cancer risk is an important need in research and diagnostic fields. Such studies require many resources in terms of costs, time and professionalism. At present, these limits can be overcome with next generation sequencing platforms. Combining these with GWAS permits the role of rare variants in cancer risk to be studied. Moreover, in a single experiment with next generation approaches, it is possible to identify copy number variations, large rearrangements and detection of fusion transcripts. Studies at genomic level are restrictive in the analysis of cancer because alterations at transcript, proteic and epigenetic levels have to be considered for an in-depth understanding of processes that lead to cancer development. However, there is an important aspect that needs to be underlined, which is the creation of bioinformatic models that allows large amounts of data arising from targeted sequencing with next generation platforms to be managed.

disclosures

The authors declare no conflict of interest.

references

1. Lander ES, Linton LM, Birren B et al. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860–921.
2. Venter JC, Adams MD, Myers EW et al. The sequence of the human genome. Science 2001; 291: 1304–1351.
3. Venter JC, Remington K, Heidelberg JF et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 2004; 304: 66–74.
4. Bentley DR. Whole-genome resequencing. Curr Opin Genet Dev 2006; 16: 545–552.
5. Levy S, Sutton G, Ng PC et al. The diploid genome sequence of an individual human. PLoS Biol 2007: 5 e254.
6. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA 1977; 74: 560–564.
7. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977; 74: 5463–5467.
8. Swerdlow H, Wu SL, Harke H et al. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. J Chromatogr 1990; 516: 61–67.
9. Hunkapiller T, Kaiser RJ, Koop BF et al. Large-scale and automated DNA sequence determination. Science 1991; 254: 59–67.
10. Edwards A, Voss H, Rice P et al. Automated DNA sequencing of the human HPRT locus. Genomics 1990; 6: 593–608.
11. Monaco A, Menolascina F, Zhao Y et al. 'Sequencing-grade' screening for BRCA1 variants by oligo-arrays. J Transl Med 2008; 6: 64.
12. Ronaghi M, Karamohamed S, Pettersson B et al. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 1996; 242: 84–89.
13. Harris TD, Buzby PR, Babcock H et al. Single-molecule DNA sequencing of a viral genome. Science 2008; 320: 106–109.
14. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008; 26: 1135–1145.
15. Fedurco M, Romieu A, Williams S et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 2006: 34 e22.
16. Turcatti G, Romieu A, Fedurco M et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res 2008: 36 e25.
17. Adessi C, Matton G, Ayala G et al. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 2000: 28 e87.
18. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008; 5: 16–18.
19. McKernan K, Blanchard A, Kotler L, Costa G. Reagents, methods, and libraries for bead-based sequencing. US patent application 20080003571 (2006).
20. Shendure J, Porreca GJ, Reppas NB et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005; 309: 1728–1732.
21. Dressman D, Yan H, Traverso G et al. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA 2003; 100: 8817–8822.
22. Brenner S, Johnson M, Bridgham J et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000; 18: 630–634.
23. Housby JN, Southern EM. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. Nucleic Acids Res 1998; 26: 4259–4266.
24. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007; 316: 1497–1502.
25. Bhinge AA, Kim J, Euskirchen GM et al. Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). Genome Res 2007; 17: 910–916.
26. Mortazavi A, Williams BA, McCue K et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008; 5: 621–628.
27. Wang ET, Sandberg R, Luo S et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008; 456: 470–476.
28. Wheeler DA, Srinivasan M, Egholm M et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008; 452: 872–876.
29. Bentley DR, Balasubramanian S, Swerdlow HP et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456: 53–59.
30. Wang J, Wang W, Li R et al. The diploid genome sequence of an Asian individual. Nature 2008; 456: 60–65.
31. Hodges E, Xuan Z, Balija V et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39: 1522–1527.
32. Porreca GJ, Zhang K, Li JB et al. Multiplex amplification of large sets of human exons. Nat Methods 2007; 4: 931–936.
33. Harismendy O, Ng PC, Strausberg RL et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 2009; 10: R32.
34. Bhangale TR, Rieder MJ, Nickerson DA. Estimating coverage and power for genetic association studies using near-complete variation data. Nat Genet 2008; 40: 841–843.
35. zur Stadt U. Rischewski J, Schneppenheim R, Kabisch H. Denaturing HPLC for identification of clonal T-cell receptor gamma rearrangements in newly diagnosed acute lymphoblastic leukemia. Clin Chem 2001; 47: 2003–2011.
36. Zhu W, Zou H, Beck A et al. Loss of heterozygosity in primary lung cancer using laser capture microdissection and WAVE DNA fragment analysis techniques. Med Sci Monit 2002; 8: BR95–99.
37. Tommasi S, Pilato B, Pinto R et al. Molecular and in silico analysis of BRCA1 and BRCA2 variants. Mutat Res 2008; 644: 64–70.
38. Kallioniemi A, Kallioniemi OP, Sudar D et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992; 258: 818–821.
39. Baudis M, Cleary ML. Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics 2001; 17: 1228–1229.
40. Campbell IG, Russell SE, Choong DY et al. Mutation of the PIK3CA gene in ovarian and breast cancer. Cancer Res 2004; 64: 7678–7681.
41. Obata K, Morland SJ, Watson RH et al. Frequent PTEN/MMAC mutations in endometrioid but not serous or mucinous epithelial ovarian tumors. Cancer Res 1998; 58: 2095–2097.
42. Press JZ, De Luca A, Boyd N et al. Ovarian carcinomas with genetic and epigenetic BRCA1 loss have distinct molecular abnormalities. BMC Cancer 2008; 8: 17.
43. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008; 299: 1335–1344.
44. Thomas G, Jacobs KB, Kraft P et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 2009; 41: 579–584.
45. Hunter DJ, Kraft P, Jacobs KB et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007; 39: 870–874.
46. Staiger H, Machicao F, Kantartzis K et al. Novel meta-analysis-derived type 2 diabetes risk loci do not determine prediabetic phenotypes. PLoS One 2008: 3 e3019.
47. Xue F, Michels KB. Diabetes, metabolic syndrome, and breast cancer: a review of the current evidence. Am J Clin Nutr 2007; 86: s823–s835.
48. Ahmed S, Thomas G, Ghoussaini M et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 2009; 41: 585–590.
49. Easton DF, Pooley KA, Dunning AM et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007; 447: 1087–1093.
50. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458: 719–724.
51. Shah SP, Morin RD, Khattra J et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 2009; 461: 809–813.
52. Li R, Li Y, Fang X et al. SNP detection for massively parallel whole-genome resequencing. Genome Res 2009; 19: 1124–1132.

53. Wang J, Wang W, Li R et al. The diploid genome sequence of an Asian individual. Nature 2008; 456: 60–65.
54. Marth GT, Korf I, Yandell MD et al. A general approach to single-nucleotide polymorphism discovery. Nat Genet 1999; 23: 452–456.
55. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008; 18: 1851–1858.
56. Castle JC, Biery M, Bouzek H et al. DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics 2010; 11: 244.
57. Albert TJ, Molla MN, Muzny DM et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007; 4: 903–905.
58. Summerer D, Wu H, Haase B et al. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. Genome Res 2009; 19: 1616–1621.
59. Mokry M, Feitsma H, Nijman IJ et al. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res 2010: 38 e116.
60. Goya R, Sun MG, Morin RD et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 2010; 26: 730–736.
61. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008; 18: 1851–1858.
62. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008; 24: 713–714.
63. Mane SP, Evans C, Cooper KL et al. Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. BMC Genomics 2009; 10: 264.
64. Morrissy AS, Morin RD, Delaney A et al. Next-generation tag sequencing for cancer gene expression profiling. Genome Res 2009; 19: 1825–1835.
65. Marioni JC, Mason CE, Mane SM et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008; 18: 1509–1517.
66. Rosenkranz R, Borodina T, Lehrach H, Himmelbauer H. Characterizing the mouse ES cell transcriptome with Illumina sequencing. Genomics 2008; 92: 187–194.
67. Maher CA, Kumar-Sinha C, Cao X et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009; 458: 97–101.
68. 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.
69. Robertson G, Hirst M, Bainbridge M et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007; 4: 651–657.
70. Bottomly D, Kyler SL, McWeeney SK, Yochum GS. Identification of β-catenin binding regions in colon cancer cells using ChIP-Seq. Nucleic Acids Res 2010.
71. Blahnik KR, Dou L, O'Geen H et al. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic Acids Res 2010: 38 e13.
72. Tuupanen S, Turunen M, Lehtonen R et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 2009; 41: 885–890.
73. Widschwendter M, Apostolidou S, Raum E et al. Epigenotyping in peripheral blood cell DNA and breast cancer risk: a proof of principle study. PLoS One 2008: 3 e2656.
74. Flusberg BA, Webster DR, Lee JH et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 2010; 7: 461–465.
75. Shi L, Reid LH, Jones WD et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006; 24: 1151–1161.
76. Canales RD, Luo Y, Willey JC et al. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol 2006; 24: 1115–1122.
77. Sugarbaker DJ, Richards WG, Gordon GJ et al. Transcriptome sequencing of malignant pleural mesothelioma tumors. Proc Natl Acad Sci USA 2008; 105: 3521–3526.
78. Saha S, Sparks AB, Rago C et al. Using the transcriptome to annotate the genome. Nat Biotechnol 2002; 20: 508–512.
79. Jiang L, Gonda TA, Gamble MV et al. Global hypomethylation of genomic DNA in cancer-associated myofibroblasts. Cancer Res 2008; 68: 9900–9908.
80. 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: 1886–1891.
81. Porreca GJ, Zhang K, Li JB et al. Multiplex amplification of large sets of human exons. Nat Methods 2007; 4: 931–936.
82. Turner EH, Lee C, Ng SB et al. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 2009; 6: 315–316.
83. Okou DT, Steinberg KM, Middle C et al. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007; 4: 907–909.
84. Hodges E, Xuan Z, Balija V et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39: 1522–1527.
85. Gnirke A, Melnikov A, Maguire J et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009; 27: 182–189.
86. Futreal PA, Coin L, Marshall M et al. A census of human cancer genes. Nat Rev Cancer 2004; 4: 177–183.
87. Wetzler M, Talpaz M, Van Etten RA et al. Subcellular localization of Bcr, Abl, and Bcr-Abl proteins in normal and leukemic cells and correlation of expression with myeloid differentiation. J Clin Invest 1993; 92: 1925–1939.
88. Solomon MJ, Larsen PL, Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell 1988; 53: 937–947.
89. Ren B, Robert F, Wyrick JJ et al. Genome-wide location and function of DNA binding proteins. Science 2000; 290: 2306–2309.
90. Yochum GS, Cleland R, Goodman RH. A genome-wide screen for beta-catenin binding sites identifies a downstream enhancer element that controls c-Myc gene expression. Mol Cell Biol 2008; 28: 7368–7379.
91. Widschwendter M, Siegmund KD, Muller HM et al. Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. Cancer Res 2004; 64: 3807–3813.
92. Leu YW, Yan PS, Fan M et al. Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer. Cancer Res 2004; 64: 8184–8192.
93. Widschwendter M, Fiegl H, Egle D et al. Epigenetic stem cell signature in cancer. Nat Genet 2007; 39: 157–158.
94. Ohm JE, McGarvey KM, Yu X et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet 2007; 39: 237–242.
95. Schlesinger Y, Straussman R, Keshet I et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet 2007; 39: 232–236.