Published in final edited form as:

Gut. 2013 June ; 62(6): 920–932. doi:10.1136/gutjnl-2011-301935.

Next generation sequencing and a new era of medicine

Graham Casey¹, David Conti¹, Robert Haile¹, and David Duggan²

¹Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, 90033

²The Translational Genomics Research Institute (TGen), Phoenix, Arizona 85004

Keywords

next generation sequencing; colorectal cancer; GI malignancies; genetics; genomics; epigenetics; germ-line; somatic; DNA; RNA

INTRODUCTION

The 10th anniversary of the publication of the first draft sequence of the human genome was celebrated recently, following the launch of the Human Genome Project in 1990 [1]. Now, 10 years later, a human genome can be sequenced in less than 10 days and soon will be sequenced routinely within a day. This remarkable progress is due to significant advances in DNA sequencing technologies from Sanger sequencing that has been dominant for nearly 30 years to next generation sequencing (NGS, also termed massively parallel sequencing). The ability to rapidly sequence human genomes and to generate genetic, transcriptomic, epigenetic data and other genome-wide data for a relatively small cost opens up numerous opportunities for translation into the clinic over the next few years. Can the knowledge gained through NGS advances lead to personalized medicine for cancer patients? In this review we provide an overview of NGS technologies and the challenges associated with handling massive amounts of data, data integration, visualization and analysis, followed by a discussion of the biological information that obtained through NGS within the context of colorectal cancer. Finally, we will discuss the remaining challenges for NGS to change clinical management of disease.

NEXT GENERATION SEQUENCING TECHNOLOGIES

The first draft of the human genome was generated using Sanger sequencing (considered first generation sequencing technology), a very successful but labor-intensive approach that required gel electrophoresis to size-separate dideoxy-terminated, PCR amplification products (originally radio-labeled nucleotides were used, but the change to fluorescent dye-labeled nucleotides greatly accelerated the application of this technology). While other

Corresponding author: Graham Casey, Ph.D. Tel 323-442-7865, Fax: 323-442-7886, gcasey@usc.edu.

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive license (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd and its Licensees to permit this article (if accepted) to be published in *Gut* editions and any other BMJ PGL products to exploit all subsidiary rights, as set out in our license.

sequencing methods existed (such as Maxam-Gilbert sequencing), Sanger sequencing was the mainstay sequencing method used worldwide for nearly three decades and has yielded many promising findings in cancer that have already entered the clinic [2]. First generation sequencing approaches such as Sanger sequencing relied on sequencing PCR amplification products together rather than sequencing individual DNA templates, resulting in poor quantitation particularly of DNA templates found at low frequencies (for example in sub population of cells). These two major limitations – the labor-intensive gel-based nature and poor quantitative data – along with extremely high cost presented insurmountable obstacles to widespread adoption of first generation sequencing for individual human genome sequencing.

NGS technologies address both of these limitations and rely on a combination of novel DNA template preparation, massively parallel sequencing of hundreds of millions to billions of individual DNA templates and image capture in real time, followed by genome sequence alignment, sequence assembly and variant detection. Hundreds of millions to billions of individual sequencing reactions (35–400bp in length) can be performed at one time. As a result of these technological improvements and high throughput, prices have fallen precipitously and a whole genome sequence now costs <\$5,000 and even less for lower coverage (for example <\$1500 for 10× genome coverage). This reduction in cost can be attributed to improvements in template capacity per instrument, sequencing chemistry and image capture.

There are a number of companies that have developed NGS technologies including Illumina's GAII, HiSeq2000, HiSeq1000, HiScanSQ, and MiSeq; Roche's Genome Sequencer FLX and GS Junior; and Life Technologies SOLiD, 5500, and Personal Genome Machine (PGM), and others (Complete Genomics, Helicos, and Pacific Biosciences). A summary of currently available NGS technologies is provided in Table 1 and a work-flow for NGS is provided in Figure 1. While next generation sequencers are usually grouped together in discussion as if they use similar technologies, the methods they employ are quite different. While all NGS methods rely on the generation of representative unbiased sources of nucleic acid templates from the complex genomes being interrogated, they differ in their chemistries (sequencing by synthesis versus sequencing by hybridization, for example), raw data capture (fluorescence emission, light emission, or pH change), as well as data analyses. A common feature is that they allow for the sequencing of up to billions of individual DNA templates in a single reaction.

Given that most image capture systems are not sufficiently sensitive to capture single fluorescent events, DNA templates need to be PCR amplified (exceptions include Helicos and Pacific Biosciences). The majority of NGS methods start with a non-biased approach to randomly fragment genomic DNA into smaller lengths that are used in turn to generate (in majority of NGS technologies) adapter-modified DNA templates for sequencing.

Two common methods to amplify DNA templates are used, solid-phase amplification and emulsion PCR (PCR amplification conducted in the aqueous phase of a water-in-oil emulsion). In solid-phase amplification (for example, used by Illumina sequencers) adapters are first ligated to DNA fragments through a series of reactions that modify the end of the

fragments. Both forward and reverse primers complementary to these adaptors are covalently attached to a support at high density, the adapter modified templates are captured by hybridization to the immobilized primers, and the single molecules are amplified in situ following looping and annealing of the free end of the DNA fragment to a complementary primer on the surface (see Figure 2). Repeated denaturation and primer extension results in localized amplification of single molecules in clusters across the surface of the slide. As a result, millions of spatially separated template clusters are generated and the NGS reactions can be initiated by hybridization using universal sequencing primers.

In emulsion PCR (used by 454 and Life Technologies, for example), adapters containing universal sequencing priming sites are ligated to the DNA fragments. Prior to emulsion PCR, DNA fragments are separated into single strands and captured onto beads under conditions favoring the capture of single DNA molecules on a single bead. Following repeated PCR amplification, many thousands of copies of the same molecules are generated (Figure 3). The beads can then be immobilized on a range of different substrates. The ensuing NGS methods rely on the spatial separation of the adapter modified templates now physically bound to a solid support and the repeated cycling and reverse incorporation of nucleotides and image capture of hundreds of millions of templates in parallel. With both amplification methods, the sequencing output signal is an aggregation of the nucleotides added to identical templates and is dependent upon how efficient is that addition process [3, 4, 5, 6, 7, 8, 9]. Current read lengths vary from 35nts to >400nts with consensus accuracies of 99.999% having been achieved from the very first publication [10].

The choice of sequencing strategies including fragment, paired-end, and mate-pair analyses allow for different types of genetic variation to be identified. All are capable of identifying single nucleotide variants (SNVs) and small insertion-deletion variants (indels). Increasing read lengths allow for more accurate variant calling due to the increased probability of uniquely mapping the read to the human genome. The use of paired-end and mate-pair sequencing allows for the identification of larger indels as well as additional structural variants including translocations and inversions. Finally, no single strategy or NGS platform can identify all genetic variants as it was recently shown that while the vast majority of identified variants between two NGS platforms were detected by both platforms (Illumina and Complete Genomics) variants unique to each platform were also identified and validated [11]. This would suggest that there still remains room for improvement in NGS technologies and variant calling methods especially for calling indels.

While the NGS technologies are now in common use, third generation sequencing technologies based on true single DNA molecule sequencing such as those offered by Helicos BioSciences HeliScope, Pacific Biosciences, and Complete Genomics are rapidly being developed. These technologies rely on single molecule DNA sequencing facilitated by extremely sensitive detection technologies. These methods do not require PCR amplification and as a result circumvent some of the problems associated with current NGS methods such as nucleotide incorporation errors. Third generation methods also have the advantage of lower DNA input. However third generation sequencers are susceptible to their own type of errors. Current error rates for the HeliScope (Helicos Biosciences) and Pacific Biosciences instrument are considered to be too high and the sequencing costs too expensive for most

applications. Complete Genomics with its “nanoball” hybridization based, fluorescent detection method has begun to change the NGS platform on numerous levels including accuracy, throughput, and cost. More detailed descriptions of solid-phase amplification, emulsion PCR, NGS and third generation sequencing technologies and associated limitations and challenges can be found in several recent informative reviews [3, 4, 5, 6, 7, 8, 9].

In all cases, precise DNA sequence alignment of data output is essential for correct data interpretation. Enormous amounts of data are generated in a relatively short time frame through NGS - approximately 4–6 whole human genomes at 30× raw coverage per HiSeq 2000 run from 6 billion sequencing reads, for example - in about one-and-a-half weeks. This data output provides daunting challenges that includes more than data storage and manipulation of the reads themselves, but also alignment to the correct location in the human genome using reference sequence, and assembling them into longer contiguous DNA stretches [12]. The statistical and computational needs required to interpret such large datasets pose significant challenges. There are an ever-increasing number of open source or commercial alignment algorithm software packages being developed [9, 13], but such algorithms need to be trained and validated for optimal performance. It is clear that as NGS gains momentum additional improved tools will be required to efficiently analyze larger and larger datasets, but current and emerging NGS technologies will face analytic challenges for a considerable amount of time. Computational analysis tools are also being rapidly developed for data analysis [14, 15]. While a considerable challenge today, this challenge is likely to diminish over time as more data is generated and made publicly available [16]. More detailed descriptions of mapping, aligning, variant calling, and computational analysis sequencing algorithms can be found elsewhere [17].

NEXT GENERATION SEQUENCING APPLICATIONS

Genetic variation is acquired through inheritance from parents or through acquired somatic mutation. A wide range of data can be generated using NGS that goes far beyond the identification of structural variants and mutation detection, and includes extensive genome-wide data on gene expression (RNA-seq) [18], genome-wide profiling of epigenetic alterations and modification of histone marks and chromatin structure [19], in addition to many other applications [20]. We summarize some of these applications with respect to GI malignancies in Table 2 and below.

Genome-wide analysis of DNA variation

The most popular approach for identifying inherited genomic sequence variation in recent years has been through the use of cost-effective commercial SNP arrays during a period when the application of NGS on a large scale was unrealistic. Commercial SNP array content and genome-wide representation of those arrays has grown substantially over the years using data deposited in the International HapMap Project (<http://hapmap.ncbi.nlm.nih.gov/>), an international collaboration begun in 2002 to develop a haplotype map of the human genome that describe the common patterns of human DNA sequence variation) and more recently data generated through the 1000 Genomes Project

(<http://www.1000genomes.org>, an international effort to produce a public catalog of human genetic variation across multiple ethnic populations) as reference panels. Unlike the SNP arrays, whole genome NGS offers the opportunity for greater resolution of variant complexity through the identification of every genetic variant (single nucleotide polymorphism, SNP or copy number variation, CNV) in a genome, but extensive studies will be required to test their functionality before the maximum impact of such studies are realized.

NGS has the potential to provide comprehensive genome-wide genetic information on the vast majority of genetic variants including rare variants that could not be obtained in any other manner. As costs for NGS continue to fall genome-wide studies become feasible. In the meanwhile, more limited NGS approaches are being employed that focus only on protein coding exons on a genome-wide scale (exome), such as exome sequencing or exome SNP arrays. These approaches target the less than 1% of the genome that may be expected to contain a higher proportion of risk variants. Use of exome arrays is a cost effective way to identify variants in exons. However, it provides no information on variants mapping to non-coding regions, and genome wide association studies (GWAS) (described further below) across multiple diseases have shown that the majority of risk variants identified to date map outside of exons, and most likely are in linkage disequilibrium (LD) with functional SNPs that likely map to regulatory regions such as enhancers or silencers that affect gene expression of genes in cis or trans [21, 22]. The only effective way to capture most if not all inherited genetic variation in an individual and to start to address in a meaningful way the missing heritability or “dark matter” will be through the use of whole genome NGS. It should, however, be noted that while NGS will provide important insight into individual level risk, risk prediction will remain a considerable challenge due to genetic heterogeneity of disease, interaction between gene $\times$ environment as well as gene $\times$ gene interactions and the potential role of epigenetics in disease.

Genome wide analysis of somatic mutations

Acquired somatic mutation seen in tumors range from single base substitutions, insertion or deletion (indels) of a small number of nucleotides to large copy number changes including deletions or amplification, and DNA rearrangements (fusion proteins such as TMPRSS2-ERG in prostate cancer). As somatic alterations are rare, avoidance of false positives is paramount. Somatic mutation calling is more challenging than germline variant calling because cancer samples are often highly heterogeneous in nature. Accuracy of calling depends upon the depth of genome sequence coverage (the number of times a region is independently sequenced) but also depends on allelic fraction. With improved NGS technologies single cell sequencing is now possible, improving accuracy but also overall cost. Validation methods will be necessary to confirm all identified mutations.

A growing number of somatic mutations have been reported for a wide range of different cancers. There are coordinated international efforts through the Cancer Genome Atlas (TCGA, <http://cancergenome.nih.gov/>, an international collaborative effort to generate and compile genetic and epigenetic data on multiple tumor types through genome-wide and next generation sequencing methods) and the International Cancer Genome Consortium (ICGC,

<http://www.icgc.org/>, and international effort to generate genome wide genomic, transcriptomic and epigenomic data in 50 different tumor types and/or subtypes across the globe) [23] to generate genome wide genetic and epigenetic data on a range of different tumor types. For each tumor type, up to 500 tumors will be sequenced by NGS producing an unprecedented amount of high quality data. TCGA recently published their findings in glioblastomas and ovarian cancers [16, 24] and data on CRC tumors will likely be published within the year.

Funded initiatives exist for the curation of genomic data resulting for tumor genome sequencing. As part of the Cancer Genome Project (CGP) the Catalogue of Somatic Mutations in Cancer (COSMIC) (<http://www.sanger.ac.uk/genetics/CGP/cosmic/>) is a curation of known somatic cancer mutations from the literature, and the Cancer Gene Census (CGC) (<http://www.sanger.ac.uk/genetics/CGP/Census/>) is a catalog of gene mutations causally implicated in cancer [25]. As of October 2011 COSMIC contains information on 52,524 unique somatic mutations across 19,885 genes, and CGC identifies 464 genes that contain causal mutations.

Targeted sequencing

While the costs of NGS are falling rapidly, they still remain high for large-scale whole genome sequencing projects. As a result, cost-efficient targeted NGS studies have matured more quickly. These approaches can be used to interrogate specific genomic regions, for example targeting only protein coding regions through exome sequencing (thereby eliminating nearly 99% of the genome) [26, 27, 28], targeting specific gene families (for example all kinase genes), or targeting specific genomic regions (for example to provide more detailed SNP information in regions identified through genome-wide association studies), or known disease causing genes such as those found in hereditary colon cancer. There are a variety of commercial methods available for efficiently targeting specific genomic regions or genes (for example, Agilent Technologies, Illumina, Roche-Nimblegen, and RainDance Technologies), coupled with NGS.

Transcriptome analysis

Relative expression levels of genes in cells/tissues are currently analyzed either individually using a variety of techniques such as real time quantitative PCR, or genome-wide using commercial or custom microarray platforms. Genome-wide microarray platforms have been successfully employed for many years not only for RNA extracted from fresh or snap frozen tissues but also for RNA extracted from formalin fixed paraffin embedded tissue blocks stored in many pathology departments. However NGS methods (RNA-seq) have several advantages over conventional methods for the analysis of the transcriptome [18]. RNA-seq provides an unbiased and detailed assessment of genome-wide gene expression as it does not rely on specific predesigned probes. In addition RNA-seq provides information on novel transcripts that were not previously recognized, as well as revealing information on alternatively spliced isoforms [29] and fusion events in cancer (for example *TMPRSS2-ERG*) [30]. Importantly, degraded RNA from formalin fixed paraffin embedded tissue blocks appears suitable for NGS applications [31, 32, 33]. RNA-seq also can provide data on novel classes of transcripts including micro RNAs, short RNAs, and long non-coding RNAs

(lncRNA) [34]. Importantly, unlike for microarray data, data generated through RNA-seq can be compared with RNA-seq data across studies as it provides a more reliable estimate of absolute expression levels based on the number of mapped sequence reads. RNA-seq can also be used to interrogate allele-specific gene expression [35] due to, for example, imprinting [36, 37] or functional differences in cis-acting regulatory elements [38, 39]. While there is significant potential for RNA-seq to generate large quantities of novel data there remain many challenges that require improvements in computational methods for unique read mapping, expression quantification and transcriptome reconstruction. Comprehensive discussion of such computational challenges associated with RNA-seq data analysis can be found elsewhere [40, 41].

Methylome analysis

Epigenetic modification plays an important role in transcriptional regulation and can involve either DNA methylation of CpG sites or modification of histone proteins (see next section). While commercial DNA methylation arrays are now providing important genome-wide scale information on methylation at CpG sites in cancers such as colorectal cancer [42], it is likely that within the next few years these array-based approaches will be superseded by NGS coupled with bisulphate treatment of DNA due to their unbiased nature and greater depth of coverage [43].

Genome-wide chromatin analysis

Genome-wide histone modifications through chromatin analyses are providing invaluable information on interactions between proteins and DNA [44, 45, 46]. These studies initially involved chromatin immunoprecipitation (ChIP) followed by hybridization to DNA microarrays (ChIP-chip) but this has been replaced by ChIP-seq (ChIP followed by NGS). A sample can be enriched for chromatin bound to specific transcription factors through cross-linking proteins and DNA followed by precipitation using specific antibodies. The enriched chromatin-bound DNA can then be sequenced by NGS and visualized to identify the genomic positions to which the transcription factor was bound. The number of reads for any region roughly translates into the abundance of sequences at that location. Coupling high throughput sequencing with DNase I hypersensitivity (DNase-seq) or formaldehyde assisted isolation of regulatory elements (FAIRE) [47] combined with NGS (FAIRE-seq) that identify regions of open chromatin structure potentially indicating regulatory sequences [48] can provide an immense amount of data on chromatin regulation of gene expression. Much of these data are being generated and curated through the ENCyclopedia Of DNA Elements (ENCODE) Project (<http://www.genome.gov/10005107>, an international effort designed with the goal of identifying all functional elements in the human genome sequence) [49]. While not directly applicable to clinical management, ChIP-seq approaches are providing unprecedented information on regulatory regions that will provide invaluable information on global regulation of gene expression that will inform cancer biology and understanding of diseases such as CRC.

Microbiome analysis

A symbiotic relationship exists between the human intestine and a large diverse population of microorganisms that lives within it [50, 51]. The gut microbiome has been directly

implicated in the pathogenesis of several different diseases including inflammatory bowel disease [52], where patients with inflammatory bowel disease show reduced diversity of specific species of bacterium compared to controls [53]. A distinct microbial colonization pattern was also reported recently between the gut microbiome of CRC patients and patients with polypectomies compared to controls [54, 55] and between CRC tumor tissues and adjacent normal colon epithelial tissue [50]. While a causal relationship between specific microbiomes and CRC has not been confirmed, studies suggest that a transgenomic relationship between the microbiome and gut exists that could have implications for therapeutic response as probiotic therapies have been found effective in disorders such as IBD [56]. NGS will provide more extensive and more representative data on the gut microbiome allowing hypothesis testing that may lead to the identification of novel drug targets [56]. It should be noted, however, that sequence data alone may not provide a complete picture of the gut microbiome as transcriptomic and proteomic analyses among others may also be needed for a comprehensive functional understanding of different microbiomes.

TRANSLATION TO THE CLINIC

NGS approaches have the potential within the next 10 years or so to fundamentally change cancer management, whether by risk assessment, tumor classification, or in the identification of novel prognostic markers leading to the development of new therapeutic targets. In the next few sections we will discuss the potential for NGS to improve management of cancer focusing mainly on CRC and highlight some of the remaining challenges.

Inherited susceptibility to CRC

Highly penetrant predisposing genes have been identified for several inherited CRC syndromes through genetic linkage studies, including the adenomatous polyposis colon cancer gene *APC* in Familial Adenomatous Polyposis (FAP) colon cancer, the mismatch repair genes in Lynch syndrome, and *SMAD4* and the BMP receptor *BMPRIA/ALK3* in juvenile polyposis and hereditary mixed polyposis syndrome [57, 58, 59]. There is strong evidence that additional familial CRC syndromes exist such as Familial Colorectal Cancer type X (FCCTX) where the predisposing gene(s) has yet to be identified. More recently, *MUTYH*-associated polyposis with its pattern of recessive inheritance has also been described [60, 61, 62]. Taken together mutations in these genes account for only a relatively small proportion (perhaps less than 5%) of the overall genetic burden of CRC and only a small fraction of familial risk.

In recent years, considerable effort has been directed towards the identification of common, low penetrance mutations through GWAS in an effort to explain more of the overall genetic burden. SNP array-based GWAS have provided important insight into the genetic basis of disease and as of June 2011, the Catalogue of Published Genome-Wide Association Studies lists 1,449 GWAS published hits at $p < 5 \times 10^{-8}$ for 237 traits. For CRC, 15 significantly associated genomic loci have been identified in populations of European descent [63, 64, 65, 66, 67, 68, 69, 70], but these SNPs explain only a little additional genetic risk for CRC and ~7–8% of the excess familial risk of CRC [69, 70, 71]. These and other studies suggest that

only a fraction of the risk variants for CRC have been identified and that many common variants remain undiscovered. For example, one study suggests that an additional 67 common (>5% minor allele frequency, MAF) variants may exist with similar effect size to those published [72]. A concurrent study suggests that as many as 172 common SNPs may contribute to CRC risk [67]. Both studies estimate that these common variants would explain only 17% of the phenotypic variance in CRC, suggesting that there are other causes of heritability, presumably including SNPs that are less common (1–5% MAF) as well as rare (<1% MAF) variants.

While ongoing meta and pooled GWAS analyses are likely to identify some of these “missing” loci, such studies are limited by the depth of SNP coverage available on commercial SNP arrays. The majority of GWAS have used SNP arrays that capture 60% or less of the common variant information (>5%) when compared to the 1000 Genomes Project and less than 50% of variants with a frequency of >2.5%. While new generation SNP arrays (2.5M and 5M Illumina arrays based on early 1000 Genomes Project data) will improve overall genome coverage, it is unclear if the current sequence depth coverage provided by 1000 Genomes Project captures all common and less common (>1%) SNPs. Although the latest exome SNP arrays allow for the genotyping of variants with a frequency <1%, the vast majority of variants <1% remain to be assayed and can only be assayed via NGS at this time.

The current clinical use of data arising through GWAS is limited since the relative risk associated with the known variants is low. GWAS, however, have the potential to highlight genes and pathways that are biologically relevant to the initiation of that disease. The importance of the β -catenin/APC pathway in CRC, for example, has been recognized for some time, as has the mismatch repair pathway in a subset of CRCs. GWAS data is providing further insight into the potential biology of CRC and a role for TGF β signaling. In the case of CRC, 6 of the 15 SNPs identified through GWAS map within or near genes involved in TGF β signaling (Figure 4), including *SMAD7* [63, 65], *GREM1* [73], *RHPN2*, the bone morphogenetic protein genes *BMP2* and *BMP4* [65] and most recently *LAMA5* that is required for the production of noggin, a secreted BMP antagonist [69]. TGF β is a compelling candidate pathway involved in CRC. TGF β proteins play a critical role in proliferation, differentiation, cell migration, adhesion, extracellular matrix production [74], and also energy balance, and several lines of evidence implicate the TGF β pathway in CRC tumorigenesis. For example, in addition to GWAS data, rare, high penetrance variants in other TGF β -related genes (*SMAD4* and the BMP receptor *BMPRI1A/ALK3*) have been reported in juvenile polyposis [75, 76, 77] and hereditary mixed polyposis syndrome [78]. In addition, somatic mutations of *SMAD4* and the TGF β receptor *TGFBR2* have been identified in CRC tumors [79, 80]. The cancer initiation properties of TGF β seem to be distinct from those of progression, as tumor cells eventually develop resistance to TGF β signaling and eventually develop activated TGF β signaling, leading to enhanced tumor growth and increased metastatic potential [81]. The potential to better understand CRC tumorigenesis improves as additional susceptibility loci are identified through GWAS, particularly with the application of targeted NGS.

Genetic testing for Lynch Syndrome

There are obvious clinical benefits to identifying Lynch syndrome cases including the opportunity for increased surveillance [82]. At present a labor intensive approach is typically used to test for Lynch Syndrome, including MSI testing and immunohistochemical staining analysis of DNA mismatch repair (MMR) proteins in tumor tissues, followed by sequential Sanger DNA sequencing of one or more MMR genes. Studies have shown that not all mutations can be identified by these methods including those involving large genomic deletions [83] that are typically identified using a different method (for example Southern blotting or Multiplex Ligation-dependent Probe Amplification, MLPA) [84]. NGS has the potential to eliminate both the sequential nature of sequencing and the need for discovery methods involving large genomic deletions as it can identify all types of genomic mutations from single base pair changes to copy number changes including large genomic deletions among others (e.g., structural variants). It is easy to envisage how application of an exome sequencing-based NGS strategy that includes all MMR genes (although PMS2 could pose a challenge due to the existence of multiple highly homologous genomic regions) could be applied that will reduce time, effort and costs associated with MMR gene testing. In the near future, we are likely to see the introduction of many more efficient and cost-effective genetic screening procedures.

One of the challenging features of genetic testing of cancer susceptibility genes such as the MMR genes is the classification of variants of unknown clinical significance (VUSs), and this challenge is likely to become even greater once NGS is adopted more widely. VUSs are genomic changes for which the effect on gene function is unclear. VUSs present a considerable clinical challenge, as a definitive answer cannot be provided to the patient. A wider adoption of NGS approaches will undoubtedly lead to the identification of many more VUSs but may also help clarify function of some VUSs through increased curation of MMR gene variants that can be compared across subjects and families [85]. There is clearly some movement towards a consistent reporting not only of MMR mutations but also of other susceptibility genes [86, 87, 88]. For MMR mutations, the International Society of Gastrointestinal Hereditary Tumors (InSiGHT, www.insight-group.org) has established a committee partnering with the Human Variome Project (www.humanvariomeproject.org) with the goal of curation and establishment of classification criteria for MMR gene variants [89]. The Leiden Open Variation Database (chromium.liacs.nl/LOVD2/colon_cancer/home.php) currently curates variant data for all the MMR genes as well as *APC*, *MUTYH*, and *EPCAM*. This “experience based” approach [86, 87, 88] is needed to integrate all the diverse sources of information such as family segregation, likelihood ratios, in silico prediction as well as functional evaluation (www.mmrv.info/) [90]. In addition to family information, in silico information and the development of suitable algorithms, data on mRNA allelic imbalance [89] as well as *in vitro* functional measures of MMR activity [91] may also be incorporated into this algorithm.

Tumor Mutation Profiling by NGS

Cancer is a genetic disease, where normal regulation of expression of critical genes is altered leading to aberrant gene expression or altered gene function leading to neoplastic transformation. Until recently relatively few molecular markers had been identified in CRC

tumors including *APC*, *TP53*, *BRAF*, *KRAS* and *PIK3CA*, as well as the microsatellite instability-high (MSI-H) and CpG Island Methylator Phenotype (CIMP) sub classes [2, 42, 58, 59, 92]. The identification of novel tumor markers through NGS should radically transform our understanding of CRC tumorigenesis.

The first complete cancer genome sequenced using NGS was published only 3 years ago and was from a patient with acute myeloid leukemia [93, 94], followed soon afterwards by studies of small numbers of breast [95] and lung [96] tumors and multiple cancer cell lines [97, 98]. It is only very recently that whole genome NGS data of large tumor datasets of glioblastoma and ovarian cancers have been published through the efforts of TCGA [16, 24].

The first comprehensive genome-wide analysis of gene mutations in CRC tumors using NGS was published in 2006 (with a follow up study from the same team in 2007) where an exome sequencing approach was used to analyze 11 CRC as well as 11 breast cancers [57, 58]. In this study, transcripts from 18,191 genes were sequenced using targeted NGS (described above). Acquired somatic mutations were seen in 848 genes and on average 80 mutations were found in any single CRC but as with other cancers, inter-tumor heterogeneity was considerable [16, 24, 57, 58].

The protein-altering mutation rate within this study was 3.1 mutations per Mb in the discovery set of 11 CRC tumors and 5.8 mutations per Mb in a validation set of 24 CRC tumors which is consistent with a recent sequencing study of 1,507 genes from 441 tumors comprising breast, lung, ovarian and prostate cancers using a high throughput tiling array approach [99]. There appears to be a wide range of protein-altering mutation rates between different cancers, with lung adenocarcinomas and squamous carcinomas having mutation rates of 3.5 and 3.9 per Mb, respectively and prostate cancer one of the lowest mutation rates of 0.33 per Mb, despite the fact that 75% of tumors harbored the *TMPRSS2-ERG* gene fusion. The reason for the different rates of somatic mutations of some cancers may relate to carcinogen exposure, such as tobacco carcinogen exposure and lung cancer and UV light exposure and melanomas, but for most cancers the reason for the relative mutation rates remain unclear but probably reflect biology of the cells of origin. It is expected that as more tumors are sequenced more accurate mutation rates will be developed for different tumor types as well as sub types of cancers.

It is clear that not all of the somatic mutations identified in these studies play a role in tumorigenesis and the terms “drivers” and “passengers” have been coined to indicate their respective roles [100, 101]. Drivers are those mutations that impact the growth and/or survival of a cell and have a positive selective advantage in a tumor cell population, whereas passengers are those mutations that confer little or no selective advantage to the cell and are linked to a cell or population of cells that expanded through clonal selection [102].

It was estimated that approximately 15 of the 80 or so somatic mutations identified in CRC tumors were drivers based on their occurrence in more than one tumor and when corrected for the size of the gene and average number of mutations expected by chance [58, 103]. It is becoming clear, however, that passenger mutations are not always randomly distributed and that their clustering can mimic driver mutations [104]. As a result, it is likely that many

mutations have been misclassified and a combination of functional and genetic approaches will be required to identify true drivers from the many passengers. Classification will also be improved as more validated somatic mutation data are entered into public databases such as CENSUS. It will be essential to distinguish drivers from passengers before these data can be used in the clinical arena [14].

High levels of mutation of some of the known genes involved in CRC were confirmed in the study of CRC tumors, including mutations in *APC* (>90%), *KRAS* (45%), and *TP53* (>50%), but several new mutated genes were identified, including the receptor tyrosine kinases *EPHA3* and *EPHB6*, but were mutated in only a little over 10% of CRC tumors [57, 58, 59]. Most other mutations occurred in <5–10% of tumors [57, 58]. This is consistent with other cancers when comparing across and within different cancers types for which NGS data are available. Only rarely are genes mutated at a high frequency in a tumor type such as *APC*, *KRAS* and *TP53* in CRC, and *BRAF* in melanomas. Few genes appear to be commonly mutated and most specific gene mutations occur infrequently [105]. The type of mutations seen in different tumor types and even within tumor types also varies considerably. If, as is likely, TCGA and other studies confirm this low frequency of mutations in any single gene this will pose a challenge to the optimistic view that we will identify novel common therapeutic targets for CRC. Individual genes, however, rarely drive tumorigenesis and it will be important to determine if there are any specific downstream biological pathways common to targets of mutation. For example in glioblastomas a large number of somatic mutations occurred in only a relatively small number of pathways [16].

At this time no genome-wide NGS CRC tumor data have been published, however, several NGS studies of CRC tumors are in press and it is expected that TCGA will complete a genome-wide analysis of somatic mutations in several hundred CRC tumors using NGS before the end of 2012. Only when those data are released will a clearer picture emerge of the range of acquired somatic mutations occurring in CRC tumors. These results should lead to additional studies to improve classification of CRC tumors and should identify potential diagnostic and prognostic markers, and therapeutic targets.

CRC tumor classification

The goal of tumor classification is to identify tumor sub types that can be used to help stratify patients beyond current pathological staging. One of the most successful applications of gene expression profiling currently in the clinic is in breast cancer where a small number of genes has been shown to be useful in the prediction of recurrence and to guide therapy [106, 107, 108, 109]. Similar approaches have been applied to CRC [110, 111, 112, 113, 114, 115, 116, 117] but with disappointing results [2]. In part this is likely the result of heterogeneity in CRCs but also the relatively small sample sizes of discovery sets in most of the studies. The success of such studies is likely to improve with the application of RNA-seq that offers the opportunity to interrogate a wide range of different RNA species as well as allowing a direct comparison of results across studies [29].

CRC can be classified using a number of different molecular features including epigenetic status [2, 118]. CRC tumors that exhibit CpG Island Methylator Phenotype high (CIMP-H, a high frequency of hypermethylation at cancer-specific CpG islands) define a subset of CRC

tumors that are strongly correlated with *BRAF* mutation status [92]. There is also some evidence that *KRAS* mutations strongly correlate with hypermethylation at a reduced number of CIMP loci [119, 120, 121]. In the most comprehensive study to date, Hinoue et al [42] screened 125 CRC tumors using the array based Illumina Infinium HM27 DNA methylation platform that interrogates promoter methylation at over 27,000 CpG sites of more than 14,000 genes. Four DNA methylation sub groups were identified with characteristic molecular and clinical features, including CIMP-H that correlated with *MLH1* hypermethylation and *BRAF* mutation, CIMP low tumors that were enriched for *KRAS* mutations, a CIMP negative sub group that shows a significantly higher frequency of *TP53* mutations and occurrence in the distal colon and a second CIMP negative sub group characterized by a low frequency of gene mutations and significantly enriched for rectal tumors [42]. It will be interesting to determine whether CRC subclassification based on methylation and molecular status correlates with clinical phenotype. TCGA will also soon publish genome wide studies of a large set of CRC tumors that will include methylation data using the Illumina Infinium HumanMethylation450 BeadChip (interrogates greater than 485,000 CpG islands) in addition to mutation status. In addition to methylome-defined classification of CRC tumors these data may also be useful in defining biomarkers for use in early screening of CRC development as some methylated markers are showing promise [122].

For each of these genome-wide approaches described (genomic, transcriptomic or methylomic), it is likely that a common set of molecular changes (for example, containing 50 or 100 targets or less) will be identified allowing the use of a smaller scale technology for routine clinical use. This would reduce time and the cost not only of running the assay but also in analysis. Some of these markers have the potential to be used as non-invasive molecular detection of disease. For example, there have been promising studies using cancer-associated aberrant methylation markers in the context of fecal DNA screening of patients. Such assays currently have a sensitivity of 46–77% for detecting early stage CRC and are superior in sensitivity of detection to fecal occult blood testing [59]. It remains to be determined whether sensitivity of these assays will be further improved with increased knowledge gained through NGS studies.

Predictive markers of response

We continue to under- or over-treat cancer mostly due to our poor overall understanding of tumor biology and limited knowledge of the normal processes of the cell. Appropriate application of NGS methods has the potential to greatly improve our knowledge based not only on tumor characteristics but also within the context of an individual's inherited genetic architecture. It will be a major challenge to translate the considerable amount of molecular data that will arise through NGS studies into clinically useful predictive or prognostic tests as many tumors including CRCs are heterogeneous and current classifications are not well refined even when histologic subtypes are defined. Published data are often conflicting as treatment recommendations are based on averages across poorly classified heterogeneous groups.

For CRC there have been some successes. MSI-H is a hallmark for mismatch repair defects that are found in tumors arising in Lynch syndrome patients as well as in patients with tumors that exhibit inactivation of the mismatch repair gene *MLH1* through promoter hypermethylation found in approximately 15% of sporadic forms of CRC. MSI-H status in CRC has been shown to be associated with better prognosis for CRC in a number of studies (reviewed in [2, 123]) and there is some suggestion that MSI-H status may predict improved adjuvant 5-fluorouracil based chemotherapy response, although some controversy still exists [2, 124]. Given the considerable but not exclusive overlap between MSI-H status and improved classification of CRC tumors based on CIMP status, attention should be given to future improvements in classifiers of CRC from NGS studies.

KRAS mutations are found in up to 50% of CRC tumors. *KRAS* is involved in EGFR signaling and mutations in *KRAS* lead to constitutive activation of MAPK signaling. Recent studies have shown that *KRAS* mutations in metastatic CRCs are predictive of resistance to anti-EGFR therapy with cetuximab or panitumumab [125, 126, 127, 128, 129, 130] recently reviewed in [124]. *BRAF* is a downstream effector of *KRAS*, and *BRAF* mutations also lead to constitutive activation of MAPK signaling. Data suggest that *KRAS* and *BRAF* mutations are mutually exclusive with *BRAF* mutations occurring in a smaller subset of CRC tumors (5–10%). *BRAF* mutation status does not appear, however, to be predictive of resistance to anti-EGFR therapy, and patients whose tumors harbor *BRAF* mutations seem to respond to anti-EGFR therapies (reviewed in [124]). As discussed above, *BRAF* mutations are found exclusively in tumors classified as CIMP-H [42, 92], and future studies will need to evaluate anti-EGFR therapies in the context of more robust classifiers identified through NGS studies.

While tumor characteristics are likely to be important predictors of response, inherited germline variants are also known to influence treatment response in CRC including genes such as thymidylate synthase involved in 5-fluorouracil response [2, 124]. There is also some evidence that inherited genetic variants in genes that encode genes involved in detoxifying enzymes may play critical roles in response to agents such as oxaliplatin and irinotecan [131, 132] and GWAS and particularly NGS genome-wide analyses of inherited variants have the potential to identify novel variants that affect drug metabolism and efficacy and pharmacogenetic response to therapy.

Novel drug targets

Unfortunately many drugs at best delay progression rather than eradicate cancer most likely due to the plasticity of tumors and the redundancy of some signaling pathways. For example, phase I trials of *BRAF* inhibitors were initially very encouraging in melanoma where *BRAF* (V600E) is mutated in over 50% of melanomas [133]. Many tumors subsequently recurred as they also expressed different *BRAF* isoforms that were not responsive to the targeted agent limiting the effectiveness of these promising *BRAF* inhibitors [134, 135, 136]. NGS offers the ability to dissect the heterogeneity of tumors and has the capacity to identify sub clones within a tumor that could be used to guide overall therapeutic approaches. For more information on drug targeting approaches see several recent reviews [137, 138, 139].

While ongoing studies are encouraging, for the majority of patients key targets have not been identified. Although NGS is expected to identify many potential targets both for treatment and possibly prevention, translating these genetic discoveries into highly effective drugs will be challenging. Extensive preclinical and careful clinical studies will be required to fully understand the mechanisms underlying the biology of response. Patient selection based on tumor mutation profiles for clinical drug testing will be critical for success [99]. Most cancer genes are mutated in less than 10% of tumors within a single type [57, 58]. As a result, a large repertoire of drugs combined with diagnostic tests will be required to optimally treat most patients with any type of cancer [140]. Use of NGS information in therapeutic design has been described in detail elsewhere [138, 139].

PERSONALIZED MEDICINE: REMAINING CHALLENGES

Whole genome sequencing of both inherited and acquired somatic mutations has already begun to change the way we diagnose disease. NGS offers the opportunity for unbiased analyses of genomes whether tumor or germ-line and a resultant improvement in our understanding of tumor biology. A growing number of Institutions and companies already offer clinical testing of inherited CRC syndromes and this is expected to grow substantially as NGS costs continue to come down.

Clearly there is considerable work ahead but NGS offers the potential for hypothesis driven research that can be conducted on valid molecular markers as a more rational approach to diagnosis and treatment. NGS is likely to reduce lengthy and expensive “diagnostic odysseys” owing to the nature of the technology and the cost-effective aspects of sequencing all known disease-causing genes for a particular disease at once rather than one by one. Sequencing of tumor DNA should lead to the application of more effective targeted therapies. These technologies will undoubtedly affect the way in which we manage disease.

As promising as the future looks, such benefits are currently only being realized in a few dozen organizations at most, although the numbers are growing. As NGS technologies are adopted or considered for broader adoption in the clinic, we need to be very conscious of the ethical, legal, and social implications of such studies and of the diverse disciplines necessary to conduct such studies. Genomics data can now be generated at a quality that is suitable for most scientific analyses but no standards exist for their use in clinical practice. Genomic research in humans raises important and challenging questions about the management and release of results, as increasing amounts of genomic data are uniquely identifiable and potentially could be linked to sensitive data about an individual’s health risks. This clearly raises ethical and legal questions about privacy and confidentiality and the potential risks and benefits of disclosing individual genomics data, and highlights the need for informed consent. Indeed, the decreasing costs of NGS have led to an increasing number of laboratories that can offer genetic testing and several companies now exist that offer direct-to-consumer genetic tests. Such companies can provide increasing amounts of personal genetic information but we do not have a very good understanding of the clinical implications for many of these genetic results, particularly germ-line data. Potentially important health data should not be provided without the full involvement of one’s physician or outside the context of existing clinical testing guidelines and genetic counseling,

particularly as there is a high potential for inaccurate risk estimates being provided by commercial entities.

The introduction of NGS into clinical care is likely to prove transformative with time. As we look to the future, it is in the context of the entire patient record where NGS is likely to have its greatest impact but significant challenges lie ahead before we will see full integration of NGS into routine clinical practice [141, 142].

Abbreviations

CGC	Cancer Gene Census
CGP	Cancer Genome Project
ChIP-seq	Chromatin Immunoprecipitation and sequencing: genome-wide profiling of histone marks
CIMP-H	CpG Island Methylator Phenotype, high
CNV	copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer
CRC	colorectal cancer
DNase 1-seq	DNase 1 hypersensitivity and genome-wide sequencing
ENCODE	ENCyclopedia Of DNA Elements Project
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements and genome-wide sequencing
FAP	Familial adenomatous polyposis cancer
FCCTX	Familial Colorectal Cancer type X
GWAS	genome-wide association study
ICGC	International Cancer Genome Consortium
Indels	insertions or deletions
InSiGHT	International Society of Gastrointestinal Hereditary Tumors
LD	linkage disequilibrium
lncRNA	long non-coding RNA
MAF	minor allele frequency
Methyl-seq	genome-wide profiling of epigenetic alterations
MLPA	Multiplex Ligation-dependent Probe Amplification,
MMR	DNA mismatch repair

MSI-H	microsatellite instability, high
NGS	next generation sequencing, also termed massively parallel sequencing
PGM	Personal Genome Machine
RNA-seq	RNA sequencing
SNP	single nucleotide polymorphism
TCGA	the Cancer Genome Atlas
VUSs	variants of unknown clinical significance

CITATIONS

1. Consortium IHG. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431:931–45. [PubMed: 15496913]
2. Walther A, Johnstone E, Swanton C, et al. Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer*. 2009; 9:489–99. [PubMed: 19536109]
3. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010; 11:31–46. [PubMed: 19997069]
4. Wong KM, Hudson TJ, McPherson JD. Unraveling the genetics of cancer: genome sequencing and beyond. *Annu Rev Genomics Hum Genet*. 2011; 12:407–30. [PubMed: 21639794]
5. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol*. 2009; 25:195–203. [PubMed: 19429539]
6. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008; 9:387–402. [PubMed: 18576944]
7. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008; 26:1135–45. [PubMed: 18846087]
8. Ross JS, Cronin M. Whole cancer genome sequencing by next-generation methods. *Am J Clin Pathol*. 2011; 136:527–39. [PubMed: 21917674]
9. Turner DJ, Keane TM, Sudbery I, et al. Next-generation sequencing of vertebrate experimental organisms. *Mamm Genome*. 2009; 20:327–38. [PubMed: 19452216]
10. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456:53–9. [PubMed: 18987734]
11. Lam HY, Clark MJ, Chen R, et al. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol*. 2011; 30:78–82. [PubMed: 22178993]
12. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2010; 8:61–5. [PubMed: 21102452]
13. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*. 2009; 6:S13–20. [PubMed: 19844226]
14. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*. 2010; 11:685–96. [PubMed: 20847746]
15. Mak HC. Trends in computational biology-2010. *Nat Biotechnol*. 2011; 29:45–9. [PubMed: 21221103]
16. TCGA TCGARN. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008; 455:1061–8. [PubMed: 18772890]
17. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20:1297–303. [PubMed: 20644199]

18. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009; 10:57–63. [PubMed: 19015660]
19. Wold B, Myers RM. Sequence census methods for functional genomics. *Nat Methods.* 2008; 5:19–21. [PubMed: 18165803]
20. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet.* 2010; 11:476–86. [PubMed: 20531367]
21. Pittman AM, Naranjo S, Jalava SE, et al. Allelic variation at the 8q23.3 colorectal cancer risk locus functions as a cis-acting regulator of EIF3H. *PLoS Genet.* 2010;6.
22. Pittman AM, Naranjo S, Webb E, et al. The colorectal cancer risk at 18q21 is caused by a novel variant altering SMAD7 expression. *Genome Res.* 2009; 19:987–93. [PubMed: 19395656]
23. Hudson TJ, Anderson W, Artez A, et al. International network of cancer genome projects. *Nature.* 2010; 464:993–8. [PubMed: 20393554]
24. TCGA TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature.* 2011; 474:609–15. [PubMed: 21720365]
25. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer.* 2004; 4:177–83. [PubMed: 14993899]
26. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet.* 2010; 19:R145–51. [PubMed: 20705737]
27. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet.* 2010; 42:30–5. [PubMed: 19915526]
28. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009; 461:272–6. [PubMed: 19684571]
29. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009; 25:1105–11. [PubMed: 19289445]
30. Maher CA, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature.* 2009; 458:97–101. [PubMed: 19136943]
31. 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. [PubMed: 19440246]
32. Wood HM, Belvedere O, Conway C, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res.* 2010; 38:e151. [PubMed: 20525786]
33. Kerick M, Isau M, Timmermann B, et al. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med Genomics.* 2011; 4:68. [PubMed: 21958464]
34. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature.* 2009; 458:223–7. [PubMed: 19182780]
35. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet.* 2010; 11:533–8. [PubMed: 20567245]
36. Babak T, Deveale B, Armour C, et al. Global survey of genomic imprinting by transcriptome sequencing. *Curr Biol.* 2008; 18:1735–41. [PubMed: 19026546]
37. Wang X, Sun Q, McGrath SD, et al. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One.* 2008; 3:e3839. [PubMed: 19052635]
38. Serre D, Gurd S, Ge B, et al. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.* 2008; 4:e1000006. [PubMed: 18454203]
39. Wittkopp PJ, Haerum BK, Clark AG. Independent effects of cis- and trans-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics.* 2008; 178:1831–5. [PubMed: 18245838]
40. Garber M, Grabherr MG, Guttman M, et al. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods.* 2011; 8:469–77. [PubMed: 21623353]
41. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet.* 2011; 12:671–82. [PubMed: 21897427]

42. Hinoue T, Weisenberger DJ, Lange CP, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* 2011
43. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 2009; 462:315–22. [PubMed: 19829295]
44. Jones PA, Baylin SB. The epigenomics of cancer. *Cell.* 2007; 128:683–92. [PubMed: 17320506]
45. Kouzarides T. Chromatin modifications and their function. *Cell.* 2007; 128:693–705. [PubMed: 17320507]
46. Li E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet.* 2002; 3:662–73. [PubMed: 12209141]
47. Giresi PG, Kim J, McDaniel RM, et al. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* 2007; 17:877–85. [PubMed: 17179217]
48. Boyle AP, Davis S, Shulha HP, et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell.* 2008; 132:311–22. [PubMed: 18243105]
49. ENCODE. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science.* 2004; 306:636–40. [PubMed: 15499007]
50. Marchesi JR, Dutilh BE, Hall N, et al. Towards the human colorectal cancer microbiome. *PLoS One.* 2011; 6:e20447. [PubMed: 21647227]
51. Holmes E, Li JV, Athanasiou T, et al. Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol.* 2011; 19:349–59. [PubMed: 21684749]
52. Marchesi JR, Holmes E, Khan F, et al. Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. *J Proteome Res.* 2007; 6:546–51. [PubMed: 17269711]
53. Frank DN, St Amand AL, Feldman RA, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A.* 2007; 104:13780–5. [PubMed: 17699621]
54. Scanlan PD, Shanahan F, Clune Y, et al. Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. *Environ Microbiol.* 2008; 10:789–98. [PubMed: 18237311]
55. Rehman A, Lepage P, Nolte A, et al. Transcriptional activity of the dominant gut mucosal microbiota in chronic inflammatory bowel disease patients. *J Med Microbiol.* 2010; 59:1114–22. [PubMed: 20522625]
56. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. *Genome Med.* 2011; 3:14. [PubMed: 21392406]
57. Sjoblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* 2006; 314:268–74. [PubMed: 16959974]
58. Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* 2007; 318:1108–13. [PubMed: 17932254]
59. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med.* 2009; 361:2449–60. [PubMed: 20018966]
60. Sieber OM, Lipton L, Crabtree M, et al. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med.* 2003; 348:791–9. [PubMed: 12606733]
61. Al-Tassan N, Chmiel NH, Maynard J, et al. Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet.* 2002; 30:227–32. [PubMed: 11818965]
62. Croitoru ME, Cleary SP, Di Nicola N, et al. Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. *J Natl Cancer Inst.* 2004; 96:1631–4. [PubMed: 15523092]
63. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet.* 2007; 39:1315–7. [PubMed: 17934461]
64. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24. 21. *Nat Genet.* 2007; 39:984–8. [PubMed: 17618284]

65. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet.* 2008; 40:1426–35. [PubMed: 19011631]
66. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23. 3. *Nat Genet.* 2008; 40:623–30. [PubMed: 18372905]
67. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet.* 2008; 40:631–7. [PubMed: 18372901]
68. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet.* 2007; 39:989–94. [PubMed: 17618283]
69. Houlston RS, Cheadle J, Dobbins SE, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13. 33. *Nat Genet.* 2010; 42:973–7. [PubMed: 20972440]
70. Tomlinson IP, Dunlop M, Campbell H, et al. COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer. *Br J Cancer.* 2010; 102:447–54. [PubMed: 19920828]
71. Hindorf LA, Gillanders EM, Manolio TA. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis.* 2011; 32:945–54. [PubMed: 21459759]
72. Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet.* 2010; 42:570–5. [PubMed: 20562874]
73. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13. 3 influence colorectal cancer risk. *Nat Genet.* 2008; 40:26–8. [PubMed: 18084292]
74. Massague J. TGFbeta in Cancer. *Cell.* 2008; 134:215–30. [PubMed: 18662538]
75. Howe JR, Bair JL, Sayed MG, et al. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet.* 2001; 28:184–7. [PubMed: 11381269]
76. Zhou XP, Woodford-Richens K, Lehtonen R, et al. Germline mutations in BMPR1A/ALK3 cause a subset of cases of juvenile polyposis syndrome and of Cowden and Bannayan-Riley-Ruvalcaba syndromes. *Am J Hum Genet.* 2001; 69:704–11. [PubMed: 11536076]
77. Friedl W, Uhlhaas S, Schulmann K, et al. Juvenile polyposis: massive gastric polyposis is more common in MADH4 mutation carriers than in BMPR1A mutation carriers. *Hum Genet.* 2002; 111:108–11. [PubMed: 12136244]
78. Cheah PY, Wong YH, Chau YP, et al. Germline bone morphogenesis protein receptor 1A mutation causes colorectal tumorigenesis in hereditary mixed polyposis syndrome. *Am J Gastroenterol.* 2009; 104:3027–33. [PubMed: 19773747]
79. Koyama M, Ito M, Nagai H, et al. Inactivation of both alleles of the DPC4/SMAD4 gene in advanced colorectal cancers: identification of seven novel somatic mutations in tumors from Japanese patients. *Mutat Res.* 1999; 406:71–7. [PubMed: 10479724]
80. Grady WM, Markowitz SD. Genetic and epigenetic alterations in colon cancer. *Annu Rev Genomics Hum Genet.* 2002; 3:101–28. [PubMed: 12142355]
81. Globe GC, Schieman WP, Lodish HF. Role of transforming growth factor beta in human disease. *N Engl J Med.* 2000; 342:1350–8. [PubMed: 10793168]
82. Ladabaum U, Wang G, Terdiman J, et al. Strategies to identify the Lynch syndrome among patients with colorectal cancer: a cost-effectiveness analysis. *Ann Intern Med.* 2011; 155:69–79. [PubMed: 21768580]
83. Casey G, Lindor NM, Papadopoulos N, et al. Conversion analysis for mutation detection in MLH1 and MSH2 in patients with colorectal cancer. *JAMA.* 2005; 293:799–809. [PubMed: 15713769]
84. Baudhuin LM, Mai M, French AJ, et al. Analysis of hMLH1 and hMSH2 gene dosage alterations in hereditary nonpolyposis colorectal cancer patients by novel methods. *J Mol Diagn.* 2005; 7:226–35. [PubMed: 15858146]

85. Woods MO, Williams P, Careen A, et al. A new variant database for mismatch repair genes associated with Lynch syndrome. *Hum Mutat.* 2007; 28:669–73. [PubMed: 17347989]
86. Spurdle AB. Clinical relevance of rare germline sequence variants in cancer genes: evolution and application of classification models. *Curr Opin Genet Dev.* 2010; 20:315–23. [PubMed: 20456937]
87. Plon SE, Eccles DM, Easton D, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat.* 2008; 29:1282–91. [PubMed: 18951446]
88. Arnold S, Buchanan DD, Barker M, et al. Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. *Hum Mutat.* 2009; 30:757–70. [PubMed: 19267393]
89. Kohonen-Corish MR, Macrae F, Genuardi M, et al. Deciphering the colon cancer genes--report of the InSiGHT-Human Variome Project Workshop, UNESCO, Paris 2010. *Hum Mutat.* 2011; 32:491–4. [PubMed: 21387463]
90. Ou J, Niessen RC, Vonk J, et al. A database to support the interpretation of human mismatch repair gene variants. *Hum Mutat.* 2008; 29:1337–41. [PubMed: 18951442]
91. Drost M, Zonneveld JB, van Dijk L, et al. A cell-free assay for the functional analysis of variants of the mismatch repair protein MLH1. *Hum Mutat.* 2010; 31:247–53. [PubMed: 20020535]
92. Weisenberger DJ, Siegmund KD, Campan M, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet.* 2006; 38:787–93. [PubMed: 16804544]
93. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature.* 2008; 456:66–72. [PubMed: 18987736]
94. 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. [PubMed: 19657110]
95. 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. [PubMed: 19812674]
96. Lee W, Jiang Z, Liu J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature.* 2010; 465:473–7. [PubMed: 20505728]
97. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature.* 2010; 463:191–6. [PubMed: 20016485]
98. 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. [PubMed: 20016488]
99. Kan Z, Jaiswal BS, Stinson J, et al. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature.* 2010; 466:869–73. [PubMed: 20668451]
100. Haber DA, Settleman J. Cancer: drivers and passengers. *Nature.* 2007; 446:145–6. [PubMed: 17344839]
101. Ashworth A, Lord CJ, Reis-Filho JS. Genetic interactions in cancer progression and treatment. *Cell.* 2011; 145:30–8. [PubMed: 21458666]
102. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature.* 2009; 458:719–24. [PubMed: 19360079]
103. Parmigiani G, Boca S, Lin J, et al. Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics.* 2009; 93:17–21. [PubMed: 18692126]
104. Bignell GR, Greenman CD, Davies H, et al. Signatures of mutation and selection in the cancer genome. *Nature.* 2010; 463:893–8. [PubMed: 20164919]
105. Beroukhi R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *Nature.* 2010; 463:899–905. [PubMed: 20164920]
106. Bertucci F, Cervera N, Birnbaum D. A gene signature in breast cancer. *N Engl J Med.* 2007; 356:1887–8. author reply -8.
107. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002; 347:1999–2009. [PubMed: 12490681]
108. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med.* 2004; 351:2817–26. [PubMed: 15591335]

109. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005; 365:671–9. [PubMed: 15721472]
110. Jiang Y, Casey G, Lavery IC, et al. Development of a clinically feasible molecular assay to predict recurrence of stage II colon cancer. *J Mol Diagn*. 2008; 10:346–54. [PubMed: 18556775]
111. Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J Clin Oncol*. 2004; 22:1564–71. [PubMed: 15051756]
112. Barrier A, Boelle PY, Roser F, et al. Stage II colon cancer prognosis prediction by tumor gene expression profiling. *J Clin Oncol*. 2006; 24:4685–91. [PubMed: 16966692]
113. Del Rio M, Molina F, Bascoul-Mollevis C, et al. Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. *J Clin Oncol*. 2007; 25:773–80. [PubMed: 17327601]
114. Lin YH, Friederichs J, Black MA, et al. Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. *Clin Cancer Res*. 2007; 13:498–507. [PubMed: 17255271]
115. Barrier A, Roser F, Boelle PY, et al. Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling. *Oncogene*. 2007; 26:2642–8. [PubMed: 17043639]
116. Arango D, Laiho P, Kokko A, et al. Gene-expression profiling predicts recurrence in Dukes' C colorectal cancer. *Gastroenterology*. 2005; 129:874–84. [PubMed: 16143127]
117. Grade M, Hormann P, Becker S, et al. Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. *Cancer Res*. 2007; 67:41–56. [PubMed: 17210682]
118. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*. 2007; 50:113–30. [PubMed: 17204026]
119. Ogino S, Kawasaki T, Kirkner GJ, et al. CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations. *J Mol Diagn*. 2006; 8:582–8. [PubMed: 17065427]
120. Shen L, Toyota M, Kondo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci U S A*. 2007; 104:18654–9. [PubMed: 18003927]
121. Yagi K, Akagi K, Hayashi H, et al. Three DNA methylation epigenotypes in human colorectal cancer. *Clin Cancer Res*. 2010; 16:21–33. [PubMed: 20028768]
122. Li M, Chen WD, Papadopoulos N, et al. Sensitive digital quantification of DNA methylation in clinical samples. *Nat Biotechnol*. 2009; 27:858–63. [PubMed: 19684580]
123. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. 2005; 23:609–18. [PubMed: 15659508]
124. Bohanes P, LaBonte MJ, Winder T, et al. Predictive molecular classifiers in colorectal cancer. *Semin Oncol*. 2011; 38:576–87. [PubMed: 21810517]
125. Jimeno A, Messersmith WA, Hirsch FR, et al. KRAS mutations and sensitivity to epidermal growth factor receptor inhibitors in colorectal cancer: practical application of patient selection. *J Clin Oncol*. 2009; 27:1130–6. [PubMed: 19124802]
126. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol*. 2008; 26:1626–34. [PubMed: 18316791]
127. Lievre A, Bachet JB, Boige V, et al. KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *J Clin Oncol*. 2008; 26:374–9. [PubMed: 18202412]
128. Benvenuti S, Sartore-Bianchi A, Di Nicolantonio F, et al. Oncogenic activation of the RAS/RAF signaling pathway impairs the response of metastatic colorectal cancers to anti-epidermal growth factor receptor antibody therapies. *Cancer Res*. 2007; 67:2643–8. [PubMed: 17363584]
129. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med*. 2008; 359:1757–65. [PubMed: 18946061]
130. Van Cutsem E, Kohne CH, Hitre E, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N Engl J Med*. 2009; 360:1408–17. [PubMed: 19339720]

131. Hoskins JM, Goldberg RM, Qu P, et al. UGT1A1*28 genotype and irinotecan-induced neutropenia: dose matters. *J Natl Cancer Inst.* 2007; 99:1290–5. [PubMed: 17728214]
132. Mishima M, Samimi G, Kondo A, et al. The cellular pharmacology of oxaliplatin resistance. *Eur J Cancer.* 2002; 38:1405–12. [PubMed: 12091073]
133. Flaherty KT, Puzanov I, Kim KB, et al. Inhibition of mutated, activated BRAF in metastatic melanoma. *N Engl J Med.* 2010; 363:809–19. [PubMed: 20818844]
134. Hatzivassiliou G, Song K, Yen I, et al. RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth. *Nature.* 2010; 464:431–5. [PubMed: 20130576]
135. Heidorn SJ, Milagre C, Whittaker S, et al. Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF. *Cell.* 2010; 140:209–21. [PubMed: 20141835]
136. Poulikakos PI, Zhang C, Bollag G, et al. RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. *Nature.* 2010; 464:427–30. [PubMed: 20179705]
137. Haber DA, Gray NS, Baselga J. The evolving war on cancer. *Cell.* 2011; 145:19–24. [PubMed: 21458664]
138. Stuart D, Sellers WR. Linking somatic genetic alterations in cancer to therapeutics. *Curr Opin Cell Biol.* 2009; 21:304–10. [PubMed: 19328671]
139. Sellers WR. A blueprint for advancing genetics-based cancer therapy. *Cell.* 2011; 147:26–31. [PubMed: 21962504]
140. McDermott U, Downing JR, Stratton MR. Genomics and the continuum of cancer care. *N Engl J Med.* 2011; 364:340–50. [PubMed: 21268726]
141. Freedman ML, Monteiro AN, Gayther SA, et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet.* 2011; 43:513–8. [PubMed: 21614091]
142. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet.* 2009; 10:353–8. [PubMed: 19434079]

Figure 1. Overview of the next generation DNA sequencing workflow

Nucleic acid sources from blood, tissue, and cell culture are all compatible with next generation sequencing (NGS). Nucleic acids including DNA and RNA from germ-line as well as somatic sources can be used in the NGS assays. Depending on the choice of experimental design, samples can be enriched for specific targets (e.g., the complete coding sequence of the human genome or exome, individual gene(s), or other custom needs); RNA specimens in most cases are converted to ds-cDNA at this time (exception being direct sequencing of RNA molecules on the Heliscope). Following fragmentation, if needed, nucleic acid samples are end repaired, A-tailed, linker-modified, and, in most cases, amplified prior to sequencing. Linker-modified and amplified nucleic acids can be sequenced using any number of next generation sequencing approaches depending on the needs of the study design. Raw sequence data is processed via a stabilizing but still dynamic pipeline; the highlights of the procedure are identified here. Numerous public and commercially available programs are used for variant calling, data visualization, and annotation. Finally, annotation of the identified variants along with other analyses can be used to improve the understanding of the genetic basis of disease. Of course, additional studies will likely be required for a complete understanding of the functional consequences of the identified variant(s). Specific details on all of the processes depicted here can be found in the references referred to in the text as well as references contained within.

*Depending on the RNA isolation protocol(s), total RNA, mRNA, miRNA, lncRNAs can be studied.

Figure 2. Solid phase amplification used by Illumina sequencers

In solid phase amplification forward and reverse primers are covalently attached to a support at high density. Adaptors complementary to these primers are ligated to DNA fragments, the adapter modified DNA fragment is then hybridized to the immobilized primers, and amplified in situ following looping and annealing of the free end of the DNA fragment to a complementary primer on the surface. Repeated denaturation and primer extension results in localized amplification of single molecules in clusters across the surface of the slide.

Figure 3. Emulsion PCR used by 454 and Life Technologies, for example

In emulsion PCR, adapters containing universal sequencing primers are ligated to the DNA fragments allowing a complex genome to be PCR amplified using a common set of PCR primers. Prior to the emulsion PCR, fragments are separated into single strands and captured onto beads under conditions favoring capture of single DNA molecules on a single bead, and following PCR amplification, several thousands of copies of the same fragment are generated. The spatially separated adapter modified templates are then physically bound to a solid support for NGS where the repeated cycling and reverse incorporation of nucleotides and image capture of hundreds of millions of templates is accomplished in parallel.

Figure 4. Overview of the core TGFβ pathway. TGFβ gene regions identified through GWAS include, *BMP2*, *BMP4*, *SMAD7*, *GREM1*, *LAMA5* and *RHPN2*

Table 1Next generation (2nd) and 3rd generation sequencing technologies.

Company	URL	Instrumentation *
2nd Generation:		
Complete Genomics	www.completegenomics.com	n/a [#]
Helicos Biosciences Corp.	helicosbio.com	HeliScope
Illumina, Inc.	www.illumina.com	Genome Analyzer IIx HiSeq 2000 HiSeq 1000 MiSeq HiScanSQ HiSeq2500
Life Technologies	www.lifetechnologies.com	SOLiD System 5500 Genetic Analyzer Ion Personal Genome Machine (PGM) Ion Proton
Raoche/454	www.454.com	GS FLX System GS Junior System
3rd Generation:		
Pacific Biosciences	www.pacificbiosciences.com	PACBIO RS

* see relevant web sites for the most recent information on instrumentation

[#] n/a not available

Table 2

Next generation sequencing: application to GI malignancies *

NGS approach	Application
<i>Genome-wide analysis of germ line DNA variation</i>	Identification of genetic factors for risk assessment. Such markers may also be useful as classifiers of GI malignancies and/or indicators of therapeutic response.
<i>Genome-wide analysis of somatic mutations in tumors</i>	Identification of somatic changes including gene mutations, indels, and chromosomal rearrangements such as translocations, that could be used to classify cancers and identify novel targets leading to individualized therapy.
<i>Targeted genomic sequencing</i>	Allows detailed and comprehensive interrogation of specific genomic regions or genes (for example mismatch repair genes) in a cost efficient manner.
<i>Transcriptome analysis</i>	RNA-seq provides an unbiased genome-wide evaluation of gene expression providing information on novel transcripts as well as alternatively spliced isoforms or fusion events in CRC and related GI malignancies.
<i>Methylome analysis</i>	Genome wide epigenetic analysis will help clarify the role of epigenetics in transcriptional regulation of key cancer genes. Such an approach could be used to identify markers of early detection and/or progression as well as monitoring of treatment response.
<i>Genome-wide chromatin analysis</i>	Provides data on genome-wide histone modifications through chromatin analyses yielding invaluable information on regulatory regions and global regulation of gene expression that will better inform biologic understanding of GI malignancies.
<i>Microbiome analysis</i>	Provides extensive and representative data on the gut microbiome allowing hypothesis testing that may lead to the identification of novel drug targets.

* further details and related references can be found in the text