A Genomic Analysis Workflow for Colorectal Cancer Precision Oncology

A genomic analysis workflow for colorectal cancer precision oncology

Giorgio Corti[1] *, Alice Bartolini[1] *, Giovanni Crisafulli[1,2], Luca Novara[1], Giuseppe Rospo[1], Monica Montone[1], Carola Negrino[1], Benedetta Mussolin[1], Michela Buscarino[1], Claudio Isella[1,2], Ludovic Barault[1,2], Giulia Siravegna[1,2], Salvatore Siena[3,4], Silvia Marsoni[4,5], Federica Di Nicolantonio[1,2, ◦,#], Enzo Medico[1,2, ◦,#] & Alberto Bardelli[1,2, ◦,#]

[1]Candiolo Cancer Institute, FPO-IRCCS, Candiolo (TO), Italy; [2]University of Turin, Department of Oncology, Candiolo (TO), Italy; [3]Niguarda Cancer Center, ASST Grande Ospedale Metropolitano Niguarda, Milano, Italy; [4]Department of Oncology and Haematology–Oncology, University of Milano, Milano, Italy; [5] FIRC Institute of Molecular Oncology (IFOM), Milan, Italy.

*** These authors contributed equally**

[◦,#] **Co-senior, co-corresponding**

Address for correspondence:
Federica Di Nicolantonio, Enzo Medico, Alberto Bardelli, Candiolo Cancer Institute, SP 142 km 3.95, 10060 Candiolo (TO), Italy
Federica Di Nicolantonio (federica.dinicolantonio@unito.it)
Enzo Medico (enzo.medico@unito.it)
Alberto Bardelli (alberto.bardelli@unito.it)

Declarations of interest: A Bardelli attended Guardant Scientific Advisory Boards

MicroAbstract

Accurate diagnosis and precision medicine of colorectal cancer (CRC) rely on patient-specific genomic maps. We present IDEA, an integrated DNA Next Generation Sequencing (NGS) and bioinformatic approach to determine the molecular landscape of CRC. First, genomic targets are pre-defined to obtain optimal sensitivity for tissue or blood samples. IDEA then pinpoints genetic variations with predictive and prognostic value, defines actionable targets and unveils drug resistance mechanisms in metastatic CRC patients. Results are presented in a final report, which includes clinically relevant information.

Abstract

Background: The diagnosis of colorectal cancer (CRC) is routinely accomplished through histopathological examination. Prognostic information and treatment decision are mainly determined by TNM classification, first defined in 1968. In the last decade, patient-specific CRC genomic landscapes were shown to provide important prognostic and predictive information. Therefore, there is a need for developing NGS and bioinformatic workflows that can be routinely used for the assessment of prognostic and predictive biomarkers. **Methods:** To foster the application of genomics in the clinical management of CRCs, IDEA workflow has been built to easily adapt to the availability of patient specimens and the clinical question that is being asked. Initially, IDEA deploys *ad-hoc* NGS assays to interrogate predefined genomic target sequences (from 600kb to 30Mb) with optimal detection sensitivity. Next, sequencing data are processed through an integrated bioinformatic pipeline to assess single nucleotide variants, insertions and deletions, gene copy-number alterations and chromosomal rearrangements. Finally, the overall results are gathered into a user-friendly report. **Results:** We provide evidence that IDEA is capable of identifying clinically relevant molecular alterations. For example, IDEA can detect primary and secondary resistance mechanisms to *ERBB2* blockade including sub-clonal *RAS* and *BRAF* mutations. When optimized to analyze circulating tumor DNA (ctDNA), IDEA can be used to monitor response and relapse in the blood of metastatic CRC patients receiving targeted agents. **Conclusions:** The IDEA workflow provides a flexible platform to integrate NGS and bioinformatic tools for refined diagnosis and management of advanced CRC patients.

Keywords: colorectal cancer, next generation sequencing, bioinformatics, genetic alterations, IDEA

Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed and the second most common cause of cancer death worldwide.[1] It arises from the sequential transformation of the normal intestinal epithelium into benign adenoma and then into an invasive adenocarcinoma. The gradual morphological transformation parallels with stepwise accumulation of genetic and epigenetic alterations.[2] Neutral evolution and short periods of genomic instability may lead to the concomitant occurrence of several molecular alterations and contribute to CRC polyclonal landscapes.[3-5]

The molecular landscape of CRC has prognostic relevance and affects the choice of therapeutic strategies, directing the rational deployment of targeted drugs directed against deregulated cellular processes to which CRC cells are dependent for their survival and proliferation.

A key factor in determing CRC cells continuos proliferation is aberrant activation of the Epidermal Growth Factor Receptor (EGFR) signaling pathway.[6, 7] Activation of EGFR, elicited by Epidermal Growth Factor (EGF), leads to sequential activation of intracellular signaling proteins, such as the Kirsten Rat Sarcoma viral oncogene homolog (KRAS), the B-Raf serine/threonine-protein kinase (BRAF) and the Extracellular signal-regulated kinase 1 (ERK1), conveying proliferative signal through regulation of gene expression.[8] Standard treatment of metastatic CRC (mCRC) patients is mainly based on cytotoxic chemotherapy with *ad hoc* addition of molecular-targeted regimens.[9] For example, anti-EGFR monoclonal antibodies, such as cetuximab and panitumumab, are administered as first- or second-line therapy in combination with chemotherapy. Even with the combination of these drugs, the median overall survival of mCRC patients does not go beyond 30 months.[10, 11] Furthermore, EGFR targeted inhibition is effective only in a molecularly-defined subgroup of patients. CRC tumors carrying activating mutations in *KRAS* or *BRAF* genes are usually refractory to EGFR blockade[12, 13] and anti-EGFR monoclonal antibodies are approved only for treatment of *RAS*/*BRAF* wild type tumors.[14-16] Notably, a sizable fraction of wild-type cases are intrinsically resistant to anti-EGFR treatments. Resistance is ofeten associated with alterations in genes (such as KRAS, *NRAS*, *ERBB2*, *EGFR*, *FGFR1*, *PDGFRA*, *MAP2K1* or *MET*) that lead to downstream or parallel signaling activation.[17-19] Unfortunately, even in responding patients, acquired resistance eventually emerges within 3 to 12 months of initiating therapies.[20-23] From a molecular standpoint, the unsuccessful outcome of anti-EGFR therapy is mainly related to the emergence of mutations in the *EGFR-RAS* pathway.[20, 24]

In addition to EGFR blockade, we previously demonstrated in preclinical models that amplification of *ERBB2* (encoding the HER2 tyrosine kinase receptor) is an effective therapeutic target in cetuximab-resistant tumors.[25] Based on these observations, HERACLES, a phase II trial aimed at

testing trastuzumab and lapatinib in patients with *ERBB2* amplified CRC, was performed with very encouraging results.[26]

In addition, molecular alterations leading to the constitutive activation of other receptor tyrosine kinases also play a pivotal role in colorectal tumorigenesis and drive primary resistance to EGFR targeted monoclonal antibodies. For example, fusions involving *ALK*, *RET*, *ROS1* and *NTRK* family genes occur in 0.2%-1% of the cases[27] and represent valuable therapeutic targets for highly selected mCRC patients.[28, 29]

As previously mentioned, CRC mutational landscape has also prognostic value. Mismatch repair proficient (MMRp) CRCs comprise 85% of the total cases and often arise in the left colon (add ref). Mismatch repair deficient (MMRd) cancers that carry defects in the DNA repair machinery, and preferentially arise in the right colon, account for the remaining 15% of cases. MMR deficiency causes insertions and deletions in regions of repetitive DNA sequences called microsatellites. For this reason, MMRd often leads to the onset of a phenotype called "microsatellite instability",[30, 31] which, importantly, is associated with favorable prognosis.[32, 33] Accordingly, the fraction of MMRd cases decreases to 5-7% in the metastatic setting.

In addition to molecular analyses performed on tissue samples, profiling circulating tumor DNA (ctDNA) offers unprecedented opportunities for genotyping, tracking minimal residual disease and monitoring the emergence of drug resistance in CRC and other tumor types.[34] In light of its predictive role in response to EGFR blockade, the analysis of *RAS* mutational status in ctDNA from mCRC patients has been approved by the European Society of Digestive Oncology (ESDO) and the European Society for Medical Oncology (ESMO) when tumor tissue is not readily available.[35] Liquid biopsies also allow tracking of clonal evolution during treatment, for example by detecting acquired alterations before disease progression is clinically manifest.[23] As compared to analyses performed on tissue, detection of mutant tumor DNA in blood is more challenging and requires dedicated methodologies.[34]

For the reasons discussed above, defining the complex genomic landscape of CRCs and identification of genetically distinct CRC subtypes involve deep molecular characterization of individual patients. This is necessary for diagnostic purposes and to properly tailor treatments. In this work, we describe multiple DNA Next Generation Sequencing (NGS) approaches that, coupled to computational and bioinformatic algorithms, allow determination of clinically relevant parameters in this clinical setting.

Materials and Methods

Patients samples

Metastatic CRC patients were treated with a dual HER2 blockade by trastuzumab and lapatinib within the HERACLES multicentre, open-label, phase 2 trial performed at four academic cancer centers in Italy, as previously described.[26, 36] The study was conducted according to the provisions of the Declaration of Helsinki and the International Conference on Harmonization and Good Clinical Practice guidelines. All patients provided written informed consent for participation to the study and associated procedures, including the molecular analyses described in this work.

Next generation sequencing: target enrichment and custom panel design

We carried out and optimized specific workflows for both DNA extraction and initial steps of library preparation (see Supplementary Material section) based on DNA specific features (Figure 1A-B). Independently from these initial steps, all libraries proceeded toward the enrichment of target regions. For Whole Exome Sequencing (WES, 45Mb) we used Nextera Rapid Capture Exome kit, or the latest equivalent TruSeq Rapid Exome Enrichment kit (Illumina Inc.). For the IRCC-TARGET and FUSION custom panels, we designed capture probes exploiting the DesignStudio tool available online (https://designstudio.illumina.com). In particular, for the IRCC-TARGET panel we identified a target covering all coding regions of 224 genes known to be involved in CRC tumorigenesis, progression, oncogenic signaling and sensitivity or resistance to targeted therapy, for a total of 603kb (Supplementary Table 1). Instead, the FUSION panel has been designed selecting the most frequent oncogenic kinases involved in fusions and the most frequently rearranged partners, identified on the basis of the available literature and TCGA database. Custom probes were designed to capture exons and introns of upstream (5') and downstream (3') partners. The panel also allowed enriching hot-spot mutations previously associated to EGFR blockade resistance in CRC, the entire promoter of EGFR ligands and, finally, all coding exons of genes known to be involved in CRC tumorigenesis (*PTEN, TP53, APC, CTNNB1*). The entire selected target regions encompassed 918kb (Supplementary Table 2). Upon quality assessment final libraries were then sequenced using Illumina MiSeq or NextSeq500 sequencers (Illumina Inc.).

Bioinformatic analysis of next generation sequencing data

All bioinformatic tools were run with default parameters unless otherwise specified. In the Quality Control (QC) module and "Mapping" phase, raw reads generated by the sequencer were aligned to the reference genome by *bwa-mem*[39] algorithm (version 0.7.13-r1126); PCR duplicates were marked using MarkDuplicates in the Picard tools suite[40] (v. 2.0.1). SAMtools[41] (v. 1.3.1) was used for reading, writing or viewing files in the SAM/BAM/CRAM format. The circular binary segmentation (CBS) algorithm, as implemented in the DNAcopy R module,[42] was used to cluster all gene copy-number alterations (CNA) in the dedicated module. Pindel[43] (v. 0.2.5b6) was used for local read realignment in the insertion/deletion (INDEL) module. Blat[44] (v. 35) was used for fine remapping of the reads in the FUSION module, with tileSize=11 and stepSize=5. We set that each reported fusion breakpoint must be supported by at least 10 reads and each fusion partner must have at least 15 mapped bases on the respective end of the read.

To carry out analyses for multiple patients at the same time, the bioinformatic workflow leverages a High Performance Computing (HPC) cluster composed of five nodes (1 master and 4 workers) running the SLURM workload manager. The use of an HPC cluster allows spreading jobs across nodes to significantly speed-up analysis as well as storing in a central location sequencing data, genome references, aligner indexes, annotations, genomic databases and analysis tools to ensure reproducibility. All custom scripts are available at https://bitbucket.org/ircc/idea.

Results

Next generation sequencing of colorectal cancer samples

The sample types to be analyzed thorugh NGS is variable and depend on logistic and clinical reasons. For instance, most often FFPE (formalin-fixed paraffin-embedded) -derived DNA samples are available for retrospective studies, while tumor heterogeneity, or longitudinal monitoring and detection of minimal residual disease, is often assessed using plasma circulating tumor DNA (ctDNA).[23, 45] Sometimes, fresh tissues (such as biopsies and preclinical models) can also be available (Figure 1A). It is therefore of outmost relevance to be able to process a variety of different samples. To this aim, below we outline specific guidelines for the initial steps of sample preparation.

First of all, we identified nucleic acid isolation procedure as a crucial step that requires tailored protocols and specific kits for each sample type (Figure 1C) to generate DNA of suitable quality and quantity for further analyses (Figure 1B).

After DNA extraction, samples authentication with Short Tandem Repeat (STR) analysis is perfromed to avoid misidentification and to verify the correct correspondence between samples belonging to the same patient/preclinical model.

Since DNA displays distinct characteristics depending on the starting material (Figure 1B), we adapted and optimized sample processing and sequencing protocols according to samples type.

In our experience, FFPE-derived DNA typically shows poor quality likely associated with processing steps for specimen preparation. In addition, the presence of DNA fragments of variable length makes this type of sample the most challenging to process with the NGS workflow. We obtained a remarkable improvements in the quality of final sequencing results by mechanical shearing DNA, thus rendering fragment length homogeneous and tailored for Illumina sequencers. The next steps involve End-repair and A-tailing of fragmented DNA molecules, both essential for subsequent ligation of adapter sequences (Figure 1D).

Unlike FFPE-derived DNA, ctDNA displays good quality due to the absence of chemical contaminants but is highly fragmented. In light of this, in our protocol, ctDNA is directly subjected to End-repair and A-tailing steps prior to adapter ligation (Figure 1D).

Finally, intact high-quality DNA is usually isolated from fresh or frozen tissue. We enzymatically fragment ctDNA by means of a transposon that cuts and simultaneously inserts sequencing adapters (Figure 1D).

In all cases, index sequences specific for each sample are inserted by means of a short PCR amplification, thus allowing to pool several samples in the same library.

The approaches outlined above allow isolation of DNA fragments of suitable length with ligated adapters. At this point, all samples could be directly subjected to whole genome sequencing

(WGS). However, as compared to WGS the analysis of specific regions of interest, such as whole-exome sequencing (WES) or custom panels, provides several advantages, as discussed later. We found that the capture-based approach is the preferred choice for enrichment of target regions, since it introduces less intrinsic biases compared to amplicon-based strategies.[46] In capture-based approaches, specifically designed biotinylated probes that hybridize to the corresponding target sequences are then captured by streptavidin magnetic beads (Figure 1E). The enriched libraries are then subjected to a final short PCR amplification and, afterwards, loaded on the sequencer.

Genotyping colorectal cancers in blood

We and others have shown that liquid biopsies can complement and, in some instances, provide more information then standard tissue biopsies.[34] Analyses of plasma samples offer the possibility to obtain a broad range of information on tumor heterogeneity and clonal molecular dynamics from a blood withdrawal. Notably, the workload required for plasma processing takes significantly less time than preparation of FFPE samples (Figure 2A). Importantly, to preserve ctDNA in plasma, blood samples must be processed within 2-4 hours from collection. After blood centrifugation and plasma isolation, ctDNA can be extracted within a few hours. Overall, ctDNA can be available for downstream analyses within 24-36 hours from sample collection. This aspect is important as in some instances, the timeline required to generate molecular maps staring from ctDNA or genomic DNA extracted from tissue has clinical relevance.

Target choice for optimal sensitivity

While the added value of NGS-based analysis in precision medicine is undisputed, optimal implementation of NGS strategy for diagnostic purposes is still being evaluated. The more commonly used NGS approaches are whole-genome sequencing (WGS), whole-exome sequencing (WES) and custom gene panels. Indeed, clinically actionable targets are typically localized in a small subset of genes,[6] which renders sequencing of defined genomic regions a valuable and cheaper alternative to WES or WGS. However, some features, such as determination of the microsatellite status (MSS/MSI) and the Tumor Mutational Burden (TMB), require sequencing of a sizeable fraction of the genome for reliable results, which may not always be achieved using targeted panels.[47]

The most important technical difference between a custom gene panel and either WGS or WES is related to the minimum frequency of the mutant allele that can be reliably discriminated, thus defining the limit of detection (LOD). The smaller the target (Figure 2B), the greater the sequencing depth you get, thus allowing to better recognize low frequency variations from the noise, that is due to errors introduced by sample preparation and base calling.[48] This is particularly relevant for liquid biopsy samples analysis: in plasma, tumor-derived DNA shed in the bloodstream is diluted by DNA

released by normal tissue;[49] accordingly, the optimal LOD for ctDNA should be below 1%. Other advantages of custom panels are related to the smaller number of sequenced bases, which allows for faster, cheaper and less demanding analyses compared to WES or WGS.

Over the past five years, we developed two target panels called IRCC-TARGET and FUSION panels. The IRCC-TARGET panel was designed to identify alterations in genes that are frequently mutated in CRC, while the FUSION panel is focused on the identification of translocations in CRC samples (for full description of custom panel targets, see Materials and Methods section). The FUSION panel is larger, owing to the need of sequencing introns-exons junctions to precisely identify the genomic breakpoint where the translocation occurs.[37, 50, 51] Once a translocation is identified, it can be exploited to design patient-specific PCR probes and track the rearrangement in ctDNA in longitudinal plasma samples.[37]

IDEA: bioinformatic workflow

The genomic landscape of CRCs encompasses several types of molecular alterations. We designed a comprehensive pipeline which is organized in specific modules (see Supplementary Material section), allowing for the identification of genomic alterations that are commonly associated with tumor onset, progression and drug resistance such as: i) single nucleotide variants (SNV), ii) insertions or deletions (INDEL), iii) gene copy-number alterations (CNA) and iv) fusions (FUSION). We assembled all the corresponding pipelines under a unique package, in which every module can be run independently, based on scientific or clinical requests (Figure 3 and Materials and Methods section). Furthermore, every section of the pipeline can be separately deployed, modified or upgraded.

The first step is the quality control (QC) module, which is mandatory for any type of analysis, as it evaluates that the quality of the sequencing data is appropriate to proceed with the subsequent bioinformatic analyses. Depending on the type of starting material, the NGS assay used and clinical needs, the above QC parameters are defined case-by-case. The next step (which is optional) is the data pre-processing module, which is aimed at removing or trimming unrelated. The mapping of raw reads to the reference genome and removal of PCR duplicates are then performed, representing a two-step process common to all analyses.

After this, IDEA proceeds to the identification of tumor specific molecular alterations. The overall aim is to detect somatic variations, i.e. variations that are related to the disease and not to germline. For this reason, all these analyses are typically performed following a comparison strategy, in which variations found in the germline DNA (healthy tissue or PBMC from the same patient) are subtracted from those present in the matched tumor sample. The only exception to this strategy relates to the FUSION module, in which we assume that germline DNA does not carry rearrangements and a single sample analysis is performed. Importantly, in those cases in which the normal sample is not available, we routinely assess Single Nucleotide Polymorphisms (SNP)

database (dbSNP)[52] in order to filter out known germline alterations. Since they can have clinical relevance, germline variants of known pathogenic impact (such as *BRCA1*, *BRCA2*, *MLH1*) are also monitored and can be reported upon request.

The SNV module includes custom scripts to identify and annotate each mutation, also providing genomic information for the final report. Specifically, amino acid and nucleotide changes, along with isoform accession number are reported. We also include the number of occurrences of the specific variant in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database.[53]

In the INDEL module, after comparing germline and tumor samples, the data are annotated and INDELs present only in the tumor sample and whose allelic frequencies are predefined with a customable threshold are listed. The COSMIC occurrence of individual INDELs is also reported.

The CNA module is designed to detect amplification or deletion of genomic regions in the tumor sample with respect to the matched germline. A custom algorithm clusters all the CNAs allowing definition of contiguous regions with similar increase or loss of copy number.

The FUSION module has been devised to pinpoint chromosomal rearrangements and accurate genomic DNA breakpoint; furthermore, a specific algorithm selects only fusions that could be correctly translated.

The last step of IDEA generates the FINAL REPORT (Figure 4) which has been designed to provide the most clinically relevant information at a glance and is tailored for a clinical readership. Additional details are made available for in-depth review of complex cases. In the FINAL REPORT, clinical information are listed, as well as sample characteristics and the technical specifications (i.e. NGS assay) indicated for the analysis (Figure 4A). The second part of the report summarizes overall genetic features of the analyzed sample, such as tumor mutational burden, microsatellite status and ploidy (Figure 4B). Results are combined in tables based on the type of molecular alterations (Figure 4C). For each variant, we first indicate the gene (symbol and full name) carrying the alteration, following the HGNC-approved gene nomenclature. This, together with the specific isoform accession number, avoids possible misunderstandings. Variants are sorted by their occurrence in the COSMIC database highlighting previously identified and potentially relevant alterations. Specific features for each DNA variation are listed as well: for instance, in the mutational analysis we report the SNV genomic position, the amino acid and nucleotide change, the substitution effect (synonymous, non-synonymous or stop gain/loss) and the allele frequency. Finally, the read depth (supporting reads) is indicated to evaluate the accuracy of variation calls. In the INDEL module the entire region affected by the variation is listed, together with its length, the effect, allele frequency and read depth. The FUSION module reports the partner genes involved in the translocation event and the read support. Furthermore, the FUSION module pinpoints the exact genomic breakpoint. Finally, for the CNA module, numerical gene copy-number for both normal and tumor samples is indicated and its fold-change in the tumor is also reported. A graphical

representation of the whole genome CN segmentation has been implemented, to better visualize amplifications and losses at the gene and/or chromosome levels (Figure 4).

Deploying IDEA to genotype colorectal tumors in tissue and blood

We previously reported the HERACLES clinical trial,[26] which was aimed at targeting HER2 with trastuzumab and lapatinib in *ERBB2*-amplified metastatic colorectal cancer (mCRC). Although 30% of the patients initially responded,[26] acquired resistance occurred in most of the cases. In collaboration with Guardant Health, we previously analyzed the blood of patients recruited in the HERACLES trial to identify putative mechanisms of primary and secondary resistance to trastuzumab and lapatinib.[36] To test the capabilities of the IDEA pipeline in a clinical setting, we performed NGS analyses on a subset of surgical tissue specimens and plasma samples collected during HERACLES treatment. In particular, we profiled using the IRCC-TARGET panel, FFPE samples and ctDNA from 10 patients belonging to HERACLES cohort. The bioinformatic pipeline was deployed to infer gene copy-number status of CRC patient tissues. Targeted capture sequencing allowed to reach high levels of both read depth and covered target in each sample, providing high quality and reliable results during the analysis. CN analysis revealed high-level of *ERBB2* amplification in 8/10 patients (CNA greater or equal to 3) (Figure 5) and a more modest increase in the other 2 patients. The analysis did not reveal recurrent copy number alterations in other genes (Figure 5).

Next, we studied plasma samples, to unveil possible variations associated with drug resistance. In patients who first achieved clinical benefit, but then relapsed, ctDNA profiles at progression and baseline were compared, while for progressive disease cases only the baseline ctDNA samples were analyzed (Figure 6). Using the IRCC-TARGET panel, we were able to identify genetic alterations in the ctDNA of all patients, including trunk variations in *TP53* and *APC*. We also detected drug-resistance related mutations in *KRAS*, *BRAF*, *PIK3CA*, *ERBB2* in 5/10 patient samples (Figure 6). Alterations in *RAS*/*RAF* highlighted the importance of the MAPK pathway as key mediator of resistance to anti-HER2 therapies. Among patients who experienced clinical benefit, emerging *KRAS* mutant clones and *BRAF* amplification were identified at progression. Other alterations detected at progression involved *ERBB2*, *EGFR*, *PIK3CA* and *PTEN* (Figure 6), suggesting an involvement of the PI3K-AKT pathway in the acquisition of resistance to dual HER2 blockade.

Discussion

It is now widely accepted that alterations in the DNA sequence underlie the development of neoplasms. The identification of mutated genes that are causally implicated in oncogenesis ('cancer genes') has been a major goal in medical sciences for the last three decades. The availability of the human genome sequence, coupled with the introduction of high throughput sequencing technologies, has created an unprecedented opportunity in the field of oncology. It is now possible to generate accurate genomic profiles from DNA isolated from cancer tissues and from blood. NGS technologies are already available in cancer centers and academic institutions. Considering that sequencing costs have dropped significantly in the past 5 years and are projected to further reduce, NGS-based diagnostic assays are becoming widely applicable and have been entering into clinical practice within the last few years. However, translating NGS data (raw sequencing 'reads') into a format that can be readily interpreted by pathologists and medical oncologists remains challenging and has not yet been standardized. To address this need, and using CRC as a test bed, we developed IDEA, a comprehensive analytical and computational pipeline. IDEA was conceived to identify somatic variants through the comparison of tumor and germline DNA samples. In this work, we have detailed each of the steps that, starting from a tissue fragment or a blood draw, are required to generate, process and analyze NGS data.

We show that IDEA can comprehensively identify several types of genetic alterations with clinical relevance including: single nucleotide variants, insertions and deletions, gene copy number alterations and rearrangements. To test IDEA in the clinical setting, we exploited an annotated set of mCRC patient samples collected during the clinical trial HERACLES. In this phase II experimentation, patients positive for HER2 overexpression showed a response rate of 30% to trastuzumab and lapatinib,[26] a paradigmatic example of precision oncology. When IDEA workflow was applied to DNA extracted from tissue and blood samples collected within HERACLES trial it could readily pinpoint *APC* and *TP53* mutations as well as *ERBB2* copy number alterations. When patients progressed, IDEA could detect the emergence of *KRAS* and *BRAF* mutant clones, as likely mechanisms of acquired therapy resistance. In summary, we developed and clinically validated a start-to-finish analytical and computational pipeline to analyze tissue and blood samples from CRC patients.

Conclusion

We have developed and integrated experimental and bioinformatic pipeline (IDEA) to support the diagnosis and clinical management of colorectal cancer patients. Using DNA Next Generation Sequencing and bioinformatic approaches, IDEA determines Single Nucleotide Variants (SNV), insertion/deletions (INDEL), Copy Number Alterations (CNA) and gene fusions, and ultimately provides a user-friendly report of clinical utility.

Clinical Practice Points

- Precision oncology is based on the concept that tumor-specific genomic landscapes provide important prognostic and predictive information. Therefore, there is an urgent need for developing NGS and bioinformatic workflows that can be routinely used in clinical settings. Furthermore, profiling circulating tumor DNA (ctDNA) offers unprecedented opportunities for early detection, tumor genotyping, tracking minimal residual disease and cancer evolution in colorectal and other tumor types.

- We developed IDEA, which integrates DNA Next Generation Sequencing (NGS) approaches and computational/bioinformatic algorithms, to precisely identify clinically relevant information in tissue and liquid biopsy of CRC patients. These include, single nucleotide variants, insertions or deletions, gene copy-number alterations and fusions. To test IDEA in the clinical setting, we exploited an annotated set of mCRC patient samples collected during the clinical trial HERACLES, unveiling primary and secondary mechanisms of resistance.

- IDEA is a start-to-finish analytical and computational pipeline to analyze tissue and blood samples from CRC patients. IDEA generates a user-friendly report with clinical utility.

Acknowledgments

The research leading to these results has received funding from: European Community's Seventh Framework Programme under grant agreement no. 602901 MErCuRIC (A.Bardelli); H2020 grant agreement no. 635342-2 MoTriColor (A.Bardelli); IMI contract n. 115749 CANCER-ID (A.Bardelli); AIRC 2010 Special Program Molecular Clinical Oncology 5 per mille, Project n. 9970 Extension program (A.Bardelli, E.M.); AIRC under IG 2018 – ID. 21407 project (F.D.N.); AIRC IG 2018 - ID. 21923 project (A.Bardelli); AIRC IG n. 17707 (F.D.N.); AIRC IG n. 16819 (E.M.); AIRC Special Program 5 per mille Metastases Project n 21091 (A.Bardelli,E.M, F.D.N.); Progetto NET-2011-02352137 (A.Bardelli,E.M, F.D.N.). Fondo per la Ricerca Locale (ex 60%), Università di Torino, 2017 (FDN); grant STRATEGY by Fondazione Piemontese per la Ricerca sul Cancro –ONLUS 5 per mille 2015 Ministero della Salute (FDN); Fondazione Piemontese per la Ricerca sul Cancro-ONLUS 5 per mille 2011Ministero della Salute (A.Bardelli, E.M., F.D.N.); Fondazione Piemontese per la Ricerca sul Cancro-ONLUS 5 per mille 2014 e 2015 Ministero della Salute (A.Bardelli); RC 2017 Ministero della Salute (FDN and LB); Roche per la Ricerca grant 2017 (G.S.). Ludovic Barault was the recipient of a MIUR-cofunded postdoctoral 'Assegno di Ricerca' from the University of Torino in 2018. Giulia Siravegna was supported by a 3-year FIRC-AIRC fellowship.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin.* 2015;65:87-108.

2. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell.* 1990;61:759-767.

3. Sottoriva A, Kang H, Ma Z, et al. A Big Bang model of human colorectal tumor growth. *Nat Genet.* 2015;47:209-216.

4. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat Genet.* 2016;48:238-244.

5. Cross WCh, Graham TA, Wright NA. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J Pathol.* 2016;240:126-136.

6. Wu X, Fan Z, Masui H, Rosen N, Mendelsohn J. Apoptosis induced by an anti-epidermal growth factor receptor monoclonal antibody in a human colorectal carcinoma cell line and its delay by insulin. *J Clin Invest.* 1995;95:1897-1905.

7. Van Emburgh BO, Sartore-Bianchi A, Di Nicolantonio F, Siena S, Bardelli A. Acquired resistance to EGFR-targeted therapies in colorectal cancer. *Mol Oncol.* 2014;8:1084-1094.

8. Ciardiello F, Tortora G. EGFR antagonists in cancer treatment. *N Engl J Med.* 2008;358:1160-1174.

9. Van Cutsem E, Cervantes A, Adam R, et al. ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Ann Oncol.* 2016;27:1386-1422.

10. Heinemann V, von Weikersthal LF, Decker T, et al. FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (FIRE-3): a randomised, open-label, phase 3 trial. *Lancet Oncol.* 2014;15:1065-1075.

11. Venook AP, Niedzwiecki D, Lenz HJ, et al. Effect of First-Line Chemotherapy Combined With Cetuximab or Bevacizumab on Overall Survival in Patients With KRAS Wild-Type Advanced or Metastatic Colorectal Cancer: A Randomized Clinical Trial. *JAMA.* 2017;317:2392-2401.

12. 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;26:1626-1634.

13. Di Nicolantonio F, Martini M, Molinari F, et al. Wild-type BRAF is required for response to panitumumab or cetuximab in metastatic colorectal cancer. *J Clin Oncol.* 2008;26:5705-5712.

14. Bardelli A, Siena S. Molecular mechanisms of resistance to cetuximab and panitumumab in colorectal cancer. *J Clin Oncol.* 2010;28:1254-1261.

15. Van Cutsem E, Lenz HJ, Köhne CH, et al. Fluorouracil, leucovorin, and irinotecan plus cetuximab treatment and RAS mutations in colorectal cancer. *J Clin Oncol.* 2015;33:692-700.

16. Peeters M, Oliner KS, Price TJ, et al. Analysis of KRAS/NRAS Mutations in a Phase III Study of Panitumumab with FOLFIRI Compared with FOLFIRI Alone as Second-line Treatment for Metastatic Colorectal Cancer. *Clin Cancer Res.* 2015;21:5469-5479.

17. Bertotti A, Papp E, Jones S, et al. The genomic landscape of response to EGFR blockade in colorectal cancer. *Nature.* 2015;526:263-267.

18. Bardelli A, Corso S, Bertotti A, et al. Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer. *Cancer Discov.* 2013;3:658-673.

19. Douillard JY, Oliner KS, Siena S, et al. Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. *N Engl J Med.* 2013;369:1023-1034.

20. Misale S, Yaeger R, Hobor S, et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature.* 2012;486:532-536.

21. Arena S, Siravegna G, Mussolin B, et al. MM-151 overcomes acquired resistance to cetuximab and panitumumab in colorectal cancers harboring EGFR extracellular domain mutations. *Sci Transl Med.* 2016;8:324ra314.

22. Russo M, Siravegna G, Blaszkowsky LS, et al. Tumor Heterogeneity and Lesion-Specific Response to Targeted Therapy in Colorectal Cancer. *Cancer Discov.* 2016;6:147-153.

23. Siravegna G, Mussolin B, Buscarino M, et al. Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nat Med.* 2015;21:795-801.

24. Diaz LA, Williams RT, Wu J, et al. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature.* 2012;486:537-540.

25. Bertotti A, Migliardi G, Galimi F, et al. A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discov.* 2011;1:508-523.

26. Sartore-Bianchi A, Trusolino L, Martino C, et al. Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial. *Lancet Oncol.* 2016;17:738-746.

27. Pietrantonio F, Di Nicolantonio F, Schrock AB, et al. ALK, ROS1, and NTRK Rearrangements in Metastatic Colorectal Cancer. *J Natl Cancer Inst.* 2017;109.

28. Medico E, Russo M, Picco G, et al. The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat Commun.* 2015;6:7002.

29. Pietrantonio F, Di Nicolantonio F, Schrock AB, et al. RET fusions in a small subset of advanced colorectal cancers at risk of being neglected. *Ann Oncol.* 2018;29:1394-1401.

30. Rodriguez-Bigas MA, Boland CR, Hamilton SR, et al. A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. *J Natl Cancer Inst.* 1997;89:1758-1762.

31. Germano G, Amirouchene-Angelozzi N, Rospo G, Bardelli A. The Clinical Impact of the Genomic Landscape of Mismatch Repair-Deficient Cancers. *Cancer Discov.* 2018.

32. Pritchard CC, Grady WM. Colorectal cancer molecular biology moves into clinical practice. *Gut.* 2011;60:116-129.

33. Papadopoulos N, Nicolaides NC, Wei YF, et al. Mutation of a mutL homolog in hereditary colon cancer. *Science.* 1994;263:1625-1629.

34. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. *Nat Rev Clin Oncol.* 2017;14:531-548.

35. Baraniskin A, Van Laethem JL, Wyrwicz L, et al. Clinical relevance of molecular diagnostics in gastrointestinal (GI) cancer: European Society of Digestive Oncology (ESDO) expert discussion and recommendations from the 17th European Society for Medical Oncology (ESMO)/World Congress on Gastrointestinal Cancer, Barcelona. *Eur J Cancer.* 2017;86:305-317.

36. Siravegna G, Lazzari L, Crisafulli G, et al. Radiologic and Genomic Evolution of Individual Metastases during HER2 Blockade in Colorectal Cancer. *Cancer Cell.* 2018;34:148-162.e147.

37. Siravegna G, Sartore-Bianchi A, Mussolin B, et al. Tracking a CAD-ALK gene rearrangement in urine and blood of a colorectal cancer patient treated with an ALK inhibitor. *Ann Oncol.* 2017;28:1302-1308.

38. Conway T, Wazny J, Bromage A, et al. Xenome--a tool for classifying reads from xenograft samples. *Bioinformatics.* 2012;28:i172-178.

39. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*2013.

40. Institute B. Picard Tools.

41. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078-2079.

42. Seshan VE OA. *DNAcopy: DNA copy number data analysis. R package*. 1.54.0 ed2018.

43. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.* 2009;25:2865-2871.

44. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res.* 2002;12:656-664.

45. Siravegna G, Geuna E, Mussolin B, et al. Genotyping tumour DNA in cerebrospinal fluid and plasma of a HER2-positive breast cancer patient with brain metastases. *ESMO Open.* 2017;2:e000253.

46. Samorodnitsky E, Jewell BM, Hagopian R, et al. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum Mutat.* 2015;36:903-914.
47. Cabel L, Proudhon C, Romano E, et al. Clinical potential of circulating tumour DNA in patients receiving anticancer immunotherapy. *Nat Rev Clin Oncol.* 2018;15:639-650.
48. Pfeiffer F, Gröber C, Blank M, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep.* 2018;8:10950.
49. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. *Clin Chem.* 2015;61:112-123.
50. Leary RJ, Sausen M, Kinde I, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med.* 2012;4:162ra154.
51. Russo M, Misale S, Wei G, et al. Acquired Resistance to the TRK Inhibitor Entrectinib in Colorectal Cancer. *Cancer Discov.* 2016;6:36-44.
52. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29:308-311.
53. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 2017;45:D777-D783.