Università degli Studi di Ferrara

DOTTORATO DI REICERCA IN
FARMACOLOGIA E ONCOLOGIA MOLECULARE

CICLO XXVI

COORDINATORE Prof. CUNEO, ANTONIO

USE OF NEXT-GENERATION SEQUENCING TO STUDY CODING AND NON-CODING RNA IN COLORECTAL CANCER

Settore Scientifico Disciplinare BIO/17

Dottorando
Dott. PALATINI, Jeffrey

Tutore
Prof. VOLINIA, Stefano

Anni 2011/2013

CONTENTS

6) FUTURE DIRECTION
7) LEGENDS
    a) Figures
    b) Tables
8) FIGURES/TABLES
9) REFERENCES
10) BIBLIOGRAPHY
11) ACKNOWLEDGEMENTS

ABSTRACT

The identification of novel mRNA and small RNA signatures of prognostic and diagnostic value in colorectal cancer (CRC) is primary focus of the thesis. The overall aim of the body this work is a deeper understanding of the molecular causes in the pathology of CRC and the identification of biomarkers, specifically mRNAs and other small non-coding RNAs with prognostic values in the clinical setting. These findings would in turn lead to an optimization of the therapeutic targets and ultimately to better clinical management of patients diagnosed with CRC.

Next-generation sequencing (NGS) is based on deep sequencing, which produces billions of short sequences at a time. NGS benefits biomedical research in several ways by interrogating whole or partially targeted genomes, transcriptomes and epigenomes, including non-coding RNAs (ncRNAs) and microRNAs (miRs). NGS is able to rapidly generate large amounts of sequence data at substantially lower cost and time respect Sanger Sequencing. I have been involved in the development and application of various novel techniques for the construction of sample libraries for NGS analysis. I have also worked with various methods of analysis of next-generation sequencing data of cancer samples.

In addition to NGS, I have also worked with numerous genomics technologies including, microarrays (both commercial and custom), NanoString, Real-Time PCR, protein arrays, and other genomics technologies to investigate not only colorectal cancer, but several other types of cancer including, but not limited to leukemia/lymphoma, breast cancer, head/neck cancer, osteosarcoma, and lung cancer.

MicroRNAs are non-coding RNA regulators of protein output by way of coding RNA disruption. MicroRNAs have been shown to be differentially expressed in many solid cancers, and they can be considered biomarkers for predictive signatures in cancer. The effects of microRNAs are exerted on cell pathology and physiology controlling translation of tens or even hundreds of different coding messengers and a unique messenger can be controlled by more than one

microRNA. In turn, one, or more, microRNAs, can disrupt entire physiological pathways.

Predictive markers are important in oncology as tumors of the same tissue of origin vary widely in their response to most available systemic therapies. Of all human cancers, colorectal cancer (CRC) is the third most commonly diagnosed cancer in the world at more than 500,000 new cases diagnosed per year.

Currently, the Tumor-Node-Metastasis (TNM) is currently the most effective and reliable predictor of CRC outcomes. However, recently new genetic alterations have been uncovered which could potentially be used to estimate prognosis in CRC, with several of them potentially representing predictive markers towards appropriate treatment regimens. Unfortunately, most of these biomarkers have failed validation in the clinical setting, with some notable exceptions being loss-of-function mutations in *KRAS, BRAF, SMAD4* and *TP53*. In addition, there are genetic alterations such as chromosomal instability (CIN), loss of heterozygosity (LOH), micro-satellite instability (MSI), that affect mismatch repair (MMR) genes, including *hMLH1, hMSH2, hMSH6*, and *PMS2*.
The overall predictive values of CIN and MSI remain controversial and the role of influence from mutations in other key genes involved in carcinogenesis still largely unclear.

Short RNAs were sequenced from paired colon adenocarcinoma and normal samples. The RNA sequences were aligned on the human genome by using multiple independent algorithms. All short RNA sequences were *de novo* merged into more than 250,000 distinct RNA contigs covering the human genome. These *de novo* short RNA contigs, or shortigs, were then matched to human genome annotation. Using this unbiased genome wide approach, all short RNAs were profiled in colon adenocarcinoma. Alongside known miRNAs[62], snoRNAs[63], and piRNAs[64], there were over 60 RNAs were differentially expressed from non-annotated shortigs, and represented candidates for novel cancer non-coding genes. RNA expression plots were obtained for each shortig, revealing RNA processing of precursor miRNAs or even of entire miRNA clusters. A number of discrepancies with miRBase annotations were detected. The dynamic range and

specificity of next generation sequencing allowed an unprecedented insight into miRNA and other non-coding RNA expression in colorectal cancer.

INTRODUCTION

Advances in Cancer

Cancer is a multi-systemic disease with complex and varyingly mechanisms underlying the propagation of uncontrolled cellular growth across many tissue types. The onset and progression of all cancer are directly related to changes in the genome, which deregulate the normal control and oversight of DNA replication and cellular growth. These regulatory changes can manifest themselves as mutations and structural changes in the DNA or at the level of RNA expression and even epigenetic modulation, thus making cancer a disease entirely of genetic origin.

In the past twenty years, progress in cancer research has had a significant impact on the diagnosis and prognosis of virtually all cancer of types. On the macro scale, methods for the detection, imaging and pathological screening of cancer have greatly improved and now routine examinations often lead to the early detection and in some cases the prevention of malignancy by removal of pre-cancerous masses. On a micro scale, many technological breakthroughs have had an explosive impact on our ability to study the genome (genomics) over the past three decades. Chiefly, the ability to better classify various cancers on a molecular basis has had the largest impact on cancer treatment and outcomes today. Advances in nucleic acid sequence and quantification, as well as the detection of epigenetic modulations throughout the genome have allowed micro classifications at the molecular level for various cancers. Additionally, on a broader scale, genomics research has led to better understanding of the fundamental mechanisms governing DNA replication and gene expression, particularly with the seminal revelation of the role of non-coding RNAs (ncRNAs) play in cancer and other disease progression by aberrant gene expression regulation.

Collectively, these novel advances have allowed for the personalization of therapeutic intervention across cancers and many other diseases on the basis of these sub-classifications and re-classifications. The emergence of molecular stratifications combined with better clinical data have contributed to more complex profile analyses and the construction of large databases of molecular profiles sub-

stratified on the basis of various genetic features (both somatic and hereditary), clinical outcomes, and treatment options. The expansions of these databases with continually increasing cohorts have ultimately proved invaluable in connecting molecular variation with disease manifestation and proliferation.

However, the expectation that increasing the size of the data sets will increase the power to detect true cancer-related genetic driving events from the background of a multitude of seemingly random mutations has been frustratingly elusive. In fact, recent results seem to suggest the opposite phenomenon. The larger sample sizes produce huge indexes of apparently significant cancer-associated genes implausibly[1]. Therefore, our ability to fully utilize these molecular indexes and leverage them toward better treatment and prognosis of cancer hinges on our increasing ability to analyze and interrogate them. This is currently the frontier in genomics-related research in cancer as well as other types of diseases.

Advances in Genomics Technologies

Sequencing

The field of genomics has been advancing at a rapid pace for the past two decades. There has been direct influence from the silicon chemistry industry and nano-fabrication processes on nucleic acid quantification along with novel advances in sequencing techniques built upon chain termination or chemical fragmentation, coupled with gel electrophoresis-based size separation methods originally developed by Nobel laureates Sanger and Gilbert[2-4]. The main difference between the Sanger and Maxam–Gilbert methods, was that the Sanger method employed the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators, instead of chemically fragmenting the DNA before separation as the Maxam–Gilbert method calls for. Ultimately, the Sanger Method proved to be a more efficient and safer method because it required less toxic chemicals and reduced amounts of radioactivity compared with the chemical fragmentation method. The original Sanger reaction required a labeled-DNA primer (radioactively or fluorescently), a single-stranded template, DNA polymerase enzyme, as well as deoxy- and di-deoxy-nucleotides. The template was divided into four aliquots, having an equimolar mix of the polymerase enzyme and each of the four

deoxynucleotides (dATP, dGTP, dCTP and dTTP). One of the four chain-terminating nucleotides in the di-deoxy form was added to each reaction to terminate DNA strand synthesis during the chain elongation step, which resulted in DNA fragments of various lengths. These fragments were pooled by reaction in separate wells (A, T, G, C) and subsequently separated on the basis of their size by denaturing them and migrating them through a poly-acrylamide gel slab by electrophoresis. After separation, the DNA bands were visualized either by either Ultra-violet light or autoradiography depending on the label used, fluorescence or radioactivity, respectively[2-4]. This method is also commonly referred to as First Generation Sequencing (FGS).

From here, Sanger-based DNA sequencing has been marked by several key advances, including read automation; capillaries and multi-capillaries; replaceable polymer gel matrices. The first of these advances in sequencing came as Hood et al[5] introduced primer-based sequencing with labeled di-deoxy-nucleotides (dye-terminator) in a single reaction, which allowed for the fragments to be read optically. This would serve as the technological basis for which the Human Genome was initially sequenced. This technique was further improved with modified polymerases and better fluorescent with energy-transfer dyes (e.g., ABI Prism)[6].

The next significant advance in Sanger-based sequencing was the introduction of the capillary electrophoresis (CE), which proved to be a significant alternative to the more cumbersome and slower alternative of large slab gel electrophoresis. The first commercially available instrument was developed by Brownlee et al[7,8]. This first instrument was capable of detection in the UV/VIS spectrums with automated sample injection and delivery coupled with an on-board computer capable of a simple, but automated high-resolution analysis of the separated and differentially labeled DNA fragments[9,10]. At this point, several instrument manufacturers begin to produce instruments based off this initial design. However, before these instruments could be used for true high-throughput analysis, some initial technical challenges involving thermal stability, the formation of bubbles in the gel matrices at the onset of electrophoresis, and most notably that the gel matrices were cross-linked needed to be addressed. Eventually, the formation of bubbles was addressed by a thermal adjustment and the problem of relatively

unstable cross-linked polymer gels was solved with the introduction of in-run replaceable gel chemistry. With these problems addressed, the result ultimately led an increase read accuracy and a marked improvement (several-fold decrease) in run times, which in turn, led to widespread use and distribution of these instruments.

It became clear soon after that if the three billion bases in the human genome sequence were to be decoded, a larger format instrument with greater throughput capability was necessary. The solution was developed by Mathies et al[11], which produced the first multi-array capillary cartridge breakthrough capable of ninety-six simultaneous sequencing reactions through ninety-six capillary threads. However, this new capillary system required a more robust and sensitive alternative to the previous detection capabilities. One that could handle the simultaneous detection of all the capillaries and the resulting increases in light-scatter. Dovichi et al developed laser-based detector, which used a flowing cuvette to sheath the severe light-scattering which occurred with simultaneous detection of the gel threads, while Yeung et al[12] first used an axial-beam excitation method which could focus the UV beam perpendicular across all the capillaries at once, which allowed for continuous CCD camera analysis across the entire array. This technology and these are methods are still primarily responsible for the majority of all Sanger-based sequencing today. The CE method remains the standard for sequence validation and CE instruments can be found in nearly institution, as well as private sequencing facilities, around the world.

Microarrays

High-throughput genomics first came in the form of high-density DNA microarrays. Pat Brown et al[13] first described the assembly of a custom fabricated microarray in the seminal publication in the journal *Science* in 1995 as a “high-capacity system to monitor the expression of many genes in parallel.” Prior to the development of this microarray technology, only a few gene transcripts could be PCR amplified and radio labeled to quantitate activity at a given time, in a method known as Northern Blot analysis. The basis of the microarray is that complimentary nucleic acid sequences will preferentially bind to each other (G>C and A>T) within a heterogeneous population of nucleic acids. The first microarray was produced by

using a robotic printer to 'blot' an array of thousands of cDNA molecules complimentary to RNA transcripts for thousands of genes, with a method similar to that of ink jet printing[13]. The array of small blots of double stranded cDNA were first immobilized on a specially coated glass microscope slide (microarray) and then denatured prior to hybridization with mRNA derived cDNA. Two different mRNA transcripts from two different cell populations were reverse-transcribed into cDNA and end-labeled with either a red or green fluorescent dye label. Equimolar ratios of each sample were denatured and were allowed to competitively hybridize to the microarray. After washing, the microarray is optically scanned at two wavelengths to excite both dyes and the resulting images are combined into a single image. Where one sample had strongly hybridized to the array, the ratio of signal color either red or green, was indicative of differential gene expression between the two samples. If both samples had hybridized similarly to the array an even color mixture of yellow indicated little transcript change between samples. This method was known as a two-channel competitive hybridization microarray.

Another important microarray technology was also being developed around the same time by Dr. Stephen Fodor and colleagues at Affymetrix, Inc., which implemented the use of light-directed chemolithography (similar to silicon wafer manufacturing) to place and secure millions of short oligo nucleotide sequences over a boron-slated glass surface to construct an ultra high-density microarray with significantly increased resolution compared to the aforementioned method developed by Brown. These types of microarray became known as GeneChips, with the first arrays being released to the public in 1996. GeneChips were sold as consumables, which required a microfluidics, station that carried out washing and staining steps and a special laser scanner originally developed in collaboration with Hewlett-Packard. In addition to the sizeable increase in resolution, this type of microarray technology relied on the complimentary binding of RNA to the oligonucleotides bound to glass surface of the array, instead of end-labeled cDNA employed by the Brown method. The GeneChip method initially relied on purified mRNA as the input source, which was then reversed-transcribed with a second-strand subsequently generated to produce cDNA. At this point, the resulting cDNA was invitro-transcribed back in cRNA using biotinylated ribonucleotides and RNA polymerase. The labeled targets were then stained with a fluorescent conjugate (streptavidin-phycoerythrin solution). This signified a major departure from the Brown method as targets were all non-differentially labeled which meant that every

target sample needed a independent measurement, as opposed to a competitive hybridization. This became known as single-channel hybridization, with a more stable method for labeling and fluorescence measurement. The Affymetrix GeneChips were also designed to serve as hybridization chambers and allowed the process of hybridization, washing/staining, and scanning procedures to be highly automated, reproducible, and in general, a faster method of microarray analysis compared to the Brown method[14].

Despite the many technological advantages that GeneChips have over custom manufactured microarrays, there are some advantages that custom arrays have over GeneChips, with the most obvious advantage being the ability to customize the content, which is critical for discovery work. It's also an inexpensive alternative to GeneChips for rapid validation of a relatively small amount of content over large cohorts. GeneChips, while being customizable, are essentially limited in the scope of discovery and the content design of the array is dependent on sequence information being available. However, it is generally cost-prohibitive for many facilities to employ customized GeneChips in the repertoire, for those projects, which require the use of microarrays.

Next-Generation Sequencing

The traditional Sanger-based sequencing and capillary electrophoresis have some significant disadvantages for ultra high-throughput applications. The efficacy of the separation is limited and begins to drop off after around a thousand bases due to sieving capabilities of the matrix, the reactions are limited in terms of cost efficiency for high-coverage discovery-based experiments where repetitive sequence coverage is paramount for the identification of rare genetic events, especially in the analysis of heterogeneous cell population. Therefore, Sanger-based CE sequencing has largely become import in more of a niche role, primarily for validation of clinical applications or small experiments. The challenge of handling the human genome needed a new approach. With the completion of the Human Genome Project in 2003, a project predominantly sequenced by the Sanger method at a cost of nearly three billion dollars over thirteen years, it became clear that faster and cheaper alternatives to FGS were necessary to take advantage of the newly cemented human template sequence. To interrogate sequence variation across scores of genes simultaneously in large cohorts

demanded a new way to sequence nucleic acid. Next-generation sequencing (NGS) is huge shift in paradigm to that of the Sanger chemistry. Instead of separating the DNA by electrophoresis of the chain-terminated products one by one, the hallmark of the current NGS technologies are that they take advantage of massively parallel sequencing of clonally amplified DNA molecules, which have been immobilized, in wells across a single or multiple flow cells[15]. The immobilized DNA fragments are sequenced either by polymerase-mediated oligonucleotide extension or by serial ligation of oligonucleotide complexes of billions of fragments in parallel. The other main characteristic of NGS is the need to construction libraries. Prior to clonal amplification the template DNA (or RNA) is carefully fragmented either at the whole-genome or partial genome (or transcriptome) level. From here, adaptor sequences are ligated to both end of the fragments (single or double-stranded) and pre-amplified to create libraries of cDNA fragments flanked by adaptor inserts. These adaptors are involved in fragment immobilization, clonal amplification, or the sequencing reaction itself. The process of library building remains the most important step in NGS technology and can be among the most difficult components of all the NGS platforms today.

The initial commercial platform for NGS on the market was launched as the GS-20 produced by 454 Life Sciences in 2005. The GS 20 was primarily based on the principles of pyro-sequencing and emulsification PCR chemistry (ePCR). Pyro-sequencing was first introduced by Nyren et al. in a 1993 landmark publication in which a method for sequencing by the detection of chemiluminescent pyrophosphates released during polymerase-directed deoxynucleoside triphosphate (dNTP) incorporation was described[16,17].
Subsequent refinement by Ronaghi et al.[18] a few years later would serve as principle technological basis for the first GS-20 NGS instrument from 454 LS.

In 2007, Roche Applied Science purchased 454 Life Sciences and launched the an updated version of the 454 instrument, now known as the GS FLX. The GS FLX continues to rely on the principle library preparation strategy involving the use of cPCR to massively clonally amplify DNA fragments. In addition, the GS FLX also uses the same pico-titer well plate system employed by the original instrument, in which a microplate has fiber-optic bundles etched into the plates surface which serve as the wells in which the sequencing reactions take place.

The post ePCR amplified library products are then deposited into the individual pico-titer wells with the sequencing chemistry necessary for the subsequent pyrophosphate sequencing reactions. Several iterative additions of free dNTPs are flowed in an orderly manner through the wells of the pico-titer plate and with the incorporation of every individual nucleotide a pyrophosphate is released. This release generates a localized, well-dependent, chemiluminescence that is captured by a charge-coupled CCD camera (Figure 1[15]). Images are gathered across the plate and software analyzed for their respective signal to-noise ratio and then the results are linearized into a common genetic sequence output[19]. The typical output from the GS FLX is approximately five-hundred million bases in total, with the average read length being >400 base-pairs. This represents the longest average read length among the three largest NGS platforms. The GS FLX has an enormous advantage in accurate sequence alignment (per read) over the other major platforms. It is for this reason, it remains the choice of de novo sequencing of small genomes, usually microbial for many projects[15,20]. Like all the NGS platforms, the sequence output is aligned to either a reference sequence and analyzed for differences, or a de novo assembly is made by stringing together overlapping sequences within the reads to produce reference scaffolds to be used as anchors when overlapping sequence is not available. There are some drawbacks to this platform. The cost is significantly higher per mb of output as compared with the other major platforms. Similarly, the relative output, or depth of coverage, per run makes this a difficult choice for human and other large genome sequencing projects, or studies with large cohorts. Although, the GS FLX with it's long sequence reads, is considered the most accurate in terms of sequence alignment, but because of the nature of the chemiluminescence from pyrophosphate sequencing, the GS FLX has trouble distinguishing long stretches of homopolymers. Theoretically, a stretch of 8 adenines (or any of the other three bases) would have twice the chemiluminescence as 4 adjacent adenines, however, in practice this is not always the case and this makes it difficult to call single-nucleotide polymorphisms (SNPs) in repeated sequence elements[19,21].

The next major NGS platform to market was developed in 2006, called the

Genome Analyzer, developed and manufactured by Solexa, a company founded by Shankar Balasubramanian and David Klenerman, in Great Britain and later acquired by Illumina (http://www.Illumina.com) later that same year. As the original goal of single molecule sequencing was not achieved the duo decided to capitalize on clonal sequencing of short DNA fragments immobilized onto microspheres. This became the basis for the "short read" platform in NGS[15]. As the term 'short read' suggests, short DNA fragments are clonally amplified and subsequently sequenced producing billions of short reads, which are then algorithmically assembled into contiguous linear sequences known as contigs or shortigs. These contigs can either be built using assembly scaffolds of overlapping contigs for de novo sequencing or by directly aligning to a reference sequence for re-sequencing and directed re-sequencing projects, such as RNA-seq or Methyl-seq. The library generation requires that template DNA be uniformly fragmented (either by sonication, chemical or enzymatic restriction) and size selected if genomic DNA is the template, or another selection protocol as necessary, if re-sequencing of RNA or the targeting of DNA is required. Like the GS FLX library preparation, the resulting fragments are end-repaired to generate 5'-phosphorylated blunt ends. The enzymatic Klenow fragment is used to add a single Adenine base to the 3'-end of the repaired DNA fragments. This allows the DNA fragments to be ligated to the oligonucleotide library adaptors with greater efficiency since these are manufactured with a single T base overhang at the 3'-end. The oligonucleotide library adaptors are complimentary to oligonucleotide anchors, which are immobilized on the surface of the glass sequencing slides within each of eight lanes or wells. These glass slides are known as ordered arrays or flow cells, and are optically transparent to allow for subsequent fluorescent detection (Figure 2[15]). A fundamental difference between the Illumina platform and the ePCR-based methodologies is that the clonal amplification of the DNA fragments takes place on instrument in the wells of the sequencing slides, as opposed to an emulsion of PCR reagents and template which occurs off-instrument. The Illumina method relies on templates hybridizing to the anchor oligos and 'arching' or 'bridging' over and hybridizing to adjacent anchors within each well as the PCR reagents are flowed through the cell, more arches or bridges are formed and the entire process forms clusters of clonally amplified product from each DNA template. This type of amplification is commonly referred to as bridge amplification, and is heavily dependant on proper dilution of the template and the number of amplification

cycles used to prevent crowding and allow each DNA fragment cluster to amplify and grow without contaminating surrounding templates. The resulting clusters usually generate close to a thousand clonally amplified molecules. Each well typically has enough space without significant steric strain to house approximately 50 x 10^6 individual clusters per flow cell (all lanes).

The sequencing reaction is based on the incorporation of four reversible fluorescent dye terminators in the presence of DNA polymerase. Each dye terminator is indicative of each of the DNA bases (A,T,G,C). This is called sequencing by synthesis and sequencing in this manner can be performed in either the forward or reverse direction, depending on the primer used, since each of the cluster's fragments have adaptors in both the forward and reverse directions. After the complimentary primers have been annealed, polymerase and an equimolar mixture of the labeled dye terminators are flowed the lanes of the flow cell with the addition of a labeled dNTP based on the complimentary sequence of each of the cluster's fragments. The resulting fluorescence is optically measured and analyzed in real-time, after each successive addition of a dye terminator, the label is chemically cleaved, washed and a subsequent addition continues where the previous reaction left off. The result is a 50-100 base-pair sequence beginning from either the 5'- or the 3'- end (or a 'paired-end' approach if both ends are sequenced), for each of the molecules in a cluster across the entire flow cell. The entire process takes approximately three to six days to complete, dependent on whether or not both ends of the template fragments are sequenced. A typical sequencing run of this type produces one to two billion bases per flow cell, per run[15,22]. Improvements to this system are continuing with great frequency and the Gb output is expected to grow significantly in the next couple of years while the run times are also continuing to decrease.

A major advantage of this type of short-read sequencing platform over the GS FLX is that is produces substantially more reads which in turn nets a deep depth of coverage (also known as deep sequencing). In addition, the overall output (Gb) is significantly greater than the long read platform with a similar run time, but less overall cost. For those experiments that require very deep sequencing, such as RNA-seq, small RNA-seq, methyl-seq and those where rare genomic events need to be examined, this short-read method has sizeable advantages over the long read GS FLX platform. There are also some disadvantages to this type of sequencing. For instance, it has been shown that the Illumina platform has greater

difficulty in accurately sequencing DNA with a higher G+C content that in turn creates a G+C bias across the read pool[23]. In addition, read accuracy begins to decrease as the size of the fragment increases, due to either incomplete blocking or incomplete cleavage of the labeled terminators resulting in the strand synthesis becoming 'out-of-phase' (also known as de-phased)[24,25]. This not only makes it difficult to correctly place a read into a reference source, but it also aberrantly creates false calls which are very difficult to measure from a bioinformatic point of view. This ultimately leads to a reduction in usable reads or to a high number of false negative or positive polymorphisms. However, better bioinformatic analysis techniques continue to improve the way these reads are analyzed, for example, it is worth noting that because these reads to progressively decrease in quality as the read progresses, it is possible to pinpoint the beginning pf the de-phasing by anchor the alignment of the reads into reduced read-length. This coupled with a 'paired-end' sequencing approach help facilitate better alignment into a reference source[26]. While this is currently an informatics technique developed to deal with better alignment of lower quality reads, it doesn't help to distinguish false calls within the read itself. Great strides in the field of bioinformatics are continuously being made and will ensure that the utility of the short-read platform will on increase as the technology moves forward into the future.

The last of the major NGS platforms, came to commercial market in 2007, was the SOLiD system, manufactured by Applied Biosystems (now Life Technologies). It is considered a short-read platform, but shares considerable technological aspects with both of the aforementioned platforms, both in terms of overall capability and methodology. The basis of this sequencing platform was derived from polony-sequencing, one of the first alternative sequencing strategies first described by the Church group (Harvard University) in their seminal publication in 2005[27]. This method is able to massively sequence millions of DNA template strands in parallel, which have been randomly fragmented into ~150-250-bp lengths. These fragments are clonally amplified by emulsification PCR. The fragments are then immobilized and sequenced by ligases and polymerases. Primers are anchored to the templates and discriminatory ligation of fluorescently labeled degenerative nonamers takes place based on sequence affinity to the template sequence. At each position of the template sequence a new set of labeled nonamers is

introduced and the excitation of particular fluorofor is indicative of the type of base present at that location in the template.

The library generation is very similar to the method used for the preparation of the GS FLX libraries in that adaptor sequences are ligated to both ends of the template sequence, which have been blunted with a single T overhang and phosphorylated prior to ligation. The 5' adaptor sequences are complimentary to attachment sites on the micro reactor beads used in both ePCR as well as attachment to the surface of the flow cell for immobilization. The 3' sequence serves as a complimentary primer site for the amplification steps; the ligation site for sequencing and also this is where the oligo-barcode is inserted if the use of multiplexing is desired. The template with the adaptors ligated is then mixed with PCR reagents in an aqueous solution that is suspended and stabilized within an oil emulsification. This entire mixture is then massively and clonally amplified in a large volume thermocycler. After amplification, the amplified template still attached to the micro reactor bead is then deposited onto a flow cell for subsequent on-instrument sequencing. One obvious advantage of this type of off-instrument amplification is that it is possible to evaluate and enumerate the general efficacy of the amplification step and possibly discarding or halting insufficient amplifications, prior to loading the sample on instrument for a lengthy and expensive sequencing run.

The chemical reactions employed by the SOLiD sequencing protocol are very different than the other major NGS platforms. First, the amplified libraries are deposited onto the flow cell, much like a gel, they are injected into well (or lanes) within the flow cell. Once loaded, a labeled primer in first annealed to the complimentary adaptor sequence at the 3'-end of the immobilized library DNA. Before any sequencing reactions take place, the fluorescence is measured and imaged producing a reference map for optical imaging and quality assessment of the following sequencing protocol. After the image index has be made, the fluorescent primers are stripped and a new set of non-labeled primers are annealed to the template in reverse orientation which presents a phosphate group at the 3'-end to which dye-coupled octamer probes are then ligated to, in contrast to the polymerase-mediated extension favored by both of the other major NGS platforms. The octamer probes that are used consist of a two base specific sequence followed by six degenerative nonamers (N) with one of four fluorescent labels affixed. There are sixteen possible combinations of the two-base specific

probes (four bases x four dyes), which, in the presence of ligase, are allowed to compete with each other to anneal at the phosphate group of primer based on complimentary sequence. The optical signal of the two-base specific probes are then imaged and then the dye-labeled nonamers portion of the probe is cleaved and a phosphate group is regenerated at the 5' end of the newly extended primer sequence. This is called a cycle and it is repeated ten times before complete cleavage of the initial primer sequence is cleaved off and washed. A new round of sequencing commences when a new primer, off-set by one base, is again annealed and the whole process is repeated for ten more cycles (Figure 3[15]). This continues through five sequencing primers (which takes about six days), generating billions of fifty-base pair reads. Because the primer is extended each cycle by the ligation of a two-base probe, and offset by one base through each successive round of sequencing, each base is independently interrogated and effectively sequenced twice. The result is a company reported 99.94% base call accuracy when sequence data are correctly de-convoluted in color space, which represents the highest call accuracy of the major NGS platforms. Subsequent studies seem to concur and indicate that the SOLiD platform does not appear to show the same G+C bias as the Illumina system in heavily GC-rich templates[28]. In addition, paired-end sequencing is possible on the SOLiD platform without the need to re-amplify the template, as in the case of the Illumina method. To sequence the 'other; end of the library fragment an additional set of primers are simply annealed to adaptor sequence at the opposite of the fragment and the entire sequencing process is then repeated. The SOLiD system has a similar overall output (Gb) per full instrument run to that of the Illumina, but generates about twice the amount of reads (~2.6 Billion) to achieve this. There are some obvious advantages of this short read system over the Illumina system. For example, because the systems routinely generates deeper sequence output it is ideally suited to interrogate small RNA libraries and whole transcriptome libraries where increased read length is less critical, while the amount of reads becomes more critical. The increase in accuracy due to the two-base encoding also becomes an important factor with the smaller read length in SOLiD. In addition, the SOLiD chemistry controls for out-of-phase reads by capping them to block extension. Another advantage SOLiD has over the Illumina platform is the run cost, especially for paired-end analysis. Because the libraries are generated off-instrument and only once, the amount of template to be sequenced can be tightly

controlled for both quality and amount, which also makes a multiplexing analysis much more efficient.

There are also some disadvantages of the SOLiD short read platform versus the Illumina system. The Illumina instrumentation provides much better 'walk-away' capability and faster run times. In addition, the library preparation for the Illumina system is significantly easier compared to the SOLiD or GS FLX platforms. The longer read lengths of the Illumina Genome Analyzer make it more ideally suited to whole large genome re-sequencing and large directed re-sequencing projects, while the SOLiD is better at aimed at projects like transcriptome analysis due to the greater number of sequence tags generate.

The next steps in NGS technology are poised to address two major areas of weakness for the three aforementioned platforms. The first area, Life Technologies and Illumina, both, have developed bench top sequencers focused on improvement of scalability, both in terms of speed and sample size –which also collaterally affects the price. These new instruments, The Personal Genome Machine (PGM) and the MiSeq, from Ion Torrent/Life Technologies and Illumina respectively, are making NGS accessible to most academic and private genomics facilities all over the world. With significant increases in speed and reduction of costs, these new instruments are bridging the gap between clinical and research applications for sequencing-based testing. Accuracy, however, remains a critical issue that needs further development before NGS replaces Sanger-based sequencing for most clinical sequencing facilities. The other future area of focus remains on the weakness of current NGS technology to interrogate very low input sources and for the need to build complex libraries, which are heavily reliant on several rounds of enzymatic reactions and harsh selection and clean-up methods[29]. The goal is to study cellular activity with native resolution at the molecular level. Pacific Biosciences has developed a platform that is capable of true single molecule sequencing, with no amplication of the input needed and can detect base modifications with read lengths averaging nearly one Kilobase. Reliability, high error rates, and cost are currently a monumental problem with the single-molecule real-time technology (SMRT) from Pacific Life Sciences for most facilities, but the ability to detect base modifications and the ability to sequence highly repetitive genomic regions with great alignment accuracy due to the long read lengths give this platform a trajectory for use in epigenetic studies where base modifications are routinely found in heavily repeated elements within many

complex genomes. Another technology on the horizon, which also addresses some of the same NGS constraints as the SMRT platform, is nanopore technology. A nanopore is, essentially, a nano-scale hole, which may consist of: 1) a biological molecule, such as a protein that forms a small pore in membrane lipid-bilayer 2) synthetic molecule such as graphene or a silicon derivative 3) a combination of both biological and synthetic. Essentially, a single strand of DNA, RNA or Protein can be passed through the nanopore with a current and sensors detect changes in the current's profile and can determine differences in bases, including single base modifications. Long reads are also feasible, but difficulty reading homopolymers may be an issue as with the PGM platform[30,31].

It is becoming increasing clear that NGS platforms currently available and on the horizon are destined to become highly specialized instruments filling niche roles in genomics that can take advantage of a platform's respective strengths, while essentially minimizing areas where they are less advantageous.

Colon and Rectal Cancers

The distinction between colon and rectal cancers is largely an anatomical distinction, at least at this point, and they are commonly referred to as a single disease in humans, colorectal carcinoma (CRC). The disease progresses from non-malignant polyps or lesions, which vary significantly in histomorphological characteristics. CRC has a large malignancy-related mortality rate in industrialized countries, annually killing more than half a million people worldwide, with a 5-year survival rate at approximately fifty percent. Metastasis to other major organs, such as the liver and lungs is often swift and is the primary cause of death, occurring in nearly twenty-five percent of patients at presentation[32,33]. The disease is a highly complex one, and yet, despite a large number of recent genome-wide sequencing studies that have revealed several genetic discoveries in CRC, to a large degree, the disease is not well characterized. A fully integrated view of the disease linking genetic alterations, epigenetic and transcriptional regulation at the coding and non-coding levels remains elusive for CRC.

Classifications of CRC

Because of the heterogenic nature of the disease, several histological and morphological CRC tumor variations are distinguishable and are indicative of a complex genotype-to-phenotype relationship in CRC tumorigenesis. However, the molecular causes and effects of such variability in these tumors need better understanding before effective therapeutic treatment options can be developed. Currently, the histological hierarchy, as defined by the World Health Organization (WHO) for CRC, stratifies tumors into adenocarcinoma or non-glandular variant classes, both driven by hereditary and somatic genetic factors, as well as clonal selection under lifestyle and environmental pressures[34]. The molecular classifications of tumorigenesis are either hereditary or non-hereditary, but much of the body of work of CRC studies focuses on hereditary tumorigenesis, due to the highly variable nature of CRC tumors, however, there are some pervasive molecular characteristics which span throughout sporadic CRCs and even some inherited cases. Most notable, is the prominent role of *APC* gene in sporadic adeno-carcinomas. Between 70%-80% of all sporadic CRCs exhibit an inactivating mutation in the *APC* gene, and nearly all mutations result in a truncated form of the APC-protein. Familial adenomatous polyposis and hereditary nonpolyposis colon cancers (HNPCC, also known as Lynch Syndrome) account for the vast majority of hereditary CRCs, which comprise nearly ten to fifteen percent of all CRC cases woldwide[35]. Current molecular classifications segregate both hereditary and sporadic tumors into three main categories, genomic instability (GIN), serrated neoplasias, and a newly anointed class for molecular characteristics that do not completely fit the two aforementioned classes.

Genomic instability is subdivided into three principle subclasses. The most common type of GIN, is chromosomal instability (CIN) which include chromosomal displacement or rearrangements, copy-number alterations, as well as mutations. For example, loss or partial loss of the 18q chromosomal region deleterious for genes such as *SMAD2*, *SMAD4,* or *DCC*, is found in up to 70% of primary CRCs, is a common molecular profile of CIN-related tumors[36]. Other characteristic molecular features of CIN-associated carcinomas are mutations in *APC* and *KRAS* genes. Although, CIN-related molecular lesions are known to be found in dysplastic foci, it has not been clearly demonstrated that CIN is responsible for malignancy or whether it simply a result. However, it believed that CIN acts as a molecular driver and promoter of neoplasia, but a single, putative driving CIN

event has yet to be identified[37,38]. Unfortunately, CIN-implicated carcinomas do not present an identifiable characteristic histomorphological profile, but they can be differentiated on the basis of tumor grade, necrosis, and the presence of extracellular mucin.

Another sub-class of GIN-related carcinomas is microsatellite instable (MSI). Microsatellites are genomic regions where short stretches of DNA sequence (or a single nucleotide) are repeated. There are hundreds of thousands of microsatellites scattered throughout the human genome. During DNA replication, mutations sometimes occur in some microsatellites causing misalignment of their repetitive subunits, which results in truncated or elongated strands, which are usually repaired by DNA mismatch-repair proteins. However, in tumors with a deficiency of these proteins, the repair mechanism often fails, or is incomplete. In CRC's with MSI, more than half of all microsatellites have mutations, consequently, making microsatellite instability an effective and straightforward marker of mismatch-repair deficiency[32,39]. It occurs in nearly all cases of HNPCC (or Lynch Syndrome) and is present in ~15% of sporadic cases of CRC[35,40]. MSI generally occurs when both alleles are knocked out by somatic inactivation or where there is an inherited germline mutation in one allele with an additional somatic inactivation of the other, but without any chromosomal abnormalities[41]. The normal mismatch repair function, which typically produces truncated alleles is either knocked out or is aberrant in MSI tumors. These types of tumors are not generally associated with *KRAS* or *TP53* gene mutations, but the *BRAF* status is considered a prognostic indicator with survival greatly improved in patients with MSI and *BRAF* intact[42]. From a pathology standpoint, MSI CRC specimens are often, but not always, heavily mucinous, littered with lymphocytes, and are inflamed at the tumor periphery which make them difficult to differentiate under a microscope[41,42].

The final subclass of GIN CRC's is the CpG island methylator phenotype (CIMP).

The CIMP class contains islands of CpG rich repeated elements, often found within or near promoter regions. In carcinogenesis, hyper-methylation of CpG islands is tantamount to transcriptional inactivation of genes with cell-cycle regulatory functions, such as tumor suppression, DNA mismatch repair, or apoptosis[43]. Typically, the genes that are most often associated with epigenetic modification in CIMP are *p16*, *MGMT*, and *hMLH1*. CIMP-classed tumors are further stratified based on their molecular profiles. For example, carcinomas with

frequent MSI and a dysfunctional *BRAF* gene are considered CIMP1, while those that are microsatellite stable, but exhibit frequent mutations in *KRAS* are considered CIMP2. In addition, microsatellite stable carcinomas where *TP53* is frequently mutated are generally CIMP(-)[44,32]. Proximal methylation of the *hMLH1* mismatch repair gene is a common characteristic of CIMP CRCs, with about half of all CIMP CRCs being microsatellite stable. In general CIMP CRCs are associated with mutations in *BRAF* and/or *BRAF* genes and a poor prognosis. Similar to the MSI class of tumors, CIMP carcinomas are difficult to differentiate on a histomorphological basis and despite methylation of the *hMLH1* promoter region, a characteristic phenotype is currently not well defined[45].

The serrated pathway is the next major class of CRCs with distinctive molecular and histomorphological characteristics. Tumors associated with the serrated neoplasia pathway are often characterized by an early promoting mutation in the *BRAF* gene. The subsequent increase of function of *BRAF* blocks or limits the activity in the apoptosis pathway in serrated polyps through an over-production of a serinethreonine kinase[46,37]. The histomorphological and molecular phenotype of serrated polyps or adenomas varies considerably, but are generally separated into two classes, sessile and traditional serrated adenomas/polyps. Sessile serrated adenomas/polyps (SSA/P) make up about 20% of all the serrated polyps, and have elongated L-shaped or anchor-shaped crypts and a large proliferative zone. SSA/Ps are typically found in the right hemicolon and are associated with progression to invasive adenocarcinomas[47,36].
Traditional serrated adenomas (TSA) generally have conventional adenoma characteristics with a serrated architecture, but they can also possess large column-like cell walls with a serrated architecture. They differ from SSA/Ps in that they exhibit left-sided localization, *KRAS* status (mutated in about 25%), and an increase in methylation frequency (notably, *MLH1* is not methylated).
Unfortunately a strong correlation of phenotype and genotype in not well defined in the serrated neoplasia pathway, but these polyps are highly malignant and are classed molecularly by the exhibition of *MSI*, *BRAF* or *KRAS* mutations and CIMP[47,48].

The CRCs which have newly characterized molecular mechanisms and pathways that can not be segregated into GIN or serrated neoplasias, have recently been

placed into a separate class of molecularly distinct carcinomas called additive molecular carcinomas[36,49]. These recent molecular findings include genetic and epigenetic alterations, but also non-coding RNA deregulation. These include, but are not limited to: histone modifications, loss of function for genes such as *TP53*, *TGF-beta* and *APC* with tumor suppressive roles; activation of the oncogenic RAS-RAFMAPK and P13K-Akt signaling pathways. Alterations in the non-coding transcripts of carcinogenesis are of particular interest, as CRCs have been described with deregulated microRNAs associated with both tumor suppressive and oncogenic activity. For example, down regulation of miR-143 and miR-145 and up-regulation miR-17-5, miR-31, and miR-183 have been identified in the carcinogenesis of colorectal lesions and polyps[37,49,50]. Molecular data characterizing the network of coding and non-coding RNAs and other genetic alterations are growing at a rapid pace with the technological developments in the capability to study genomics and these data should help to further stratify CRC classifications and to help clarify the current heterogeneity in histomorphological features.

Current Strategies in the Treatment of Colorectal Cancer

The most effective method for the diagnosis and predicting prognosis in the majority of patients diagnosed with colorectal adenocarcinomas continues to be the staging system from the American Joint Committee on Cancer (AJCC) or more commonly referred to as the TNM staging system[32]. The TMN is an acronym for the three criteria used in the system. The (T) refers to the growth extent of the primary *tumor*; (N) refers to the nearby spread to regional lymph *nodes*; (M) refers to the spread of tumors cells to distant organs and tissues, called *metastasis*. Based on the pathology of these criteria, tumors are grouped into four stages (i-iv), which in some cases can be further differentiated into sub-groups[34]. Stage I represents primary tumor growth with no spread to adjacent tissue or lymph nodes; stage II tumor growth has spread to the outer walls of the colon or rectum but has not penetrated them; stage III tumor growth has spread to nearby lymph nodes and/or adjacent fatty tissues, but not major organs; stage IV is the most serious stage as tumor growth has spread to distant organs[34,51]. Tumor biopsies are often graded on their histomorphological resemblance to normal colon or rectal cells/tissue as being “low grade” (similar to normal) or “high grade” (abnormal).

Low-grade tumors tend to progress slower than high-grade tumors, with better a prognosis. Grade is often considered with prescribed post-surgical adjuvant treatment with chemotherapeutics[34]. Patient response to chemotherapeutics and tumor classification (grade and stage), remain the prognostic gold standard directly correlated to patient outcomes. Most patients diagnosed with stage I-III, low-grade CRCs are usual treated with surgical options alone or sometimes in combinatorial treatment with chemotherapy and have been shown to have a five-year survival rate of 93.2% for stage I, 82.5% for stage II, and 59.5% for stage III patients. These survival rates are in stark contrast to those patients diagnosed with stage IV carcinomas, which typically have higher-grade tumors and a five-year survival of 8.1%[32-34]. For those patients who are at risk for developing metastasis or primary reoccurrence (stages II, III) or those that have been diagnosed with metastatic tumors (stage IV), adjuvant chemotherapy is typically used as the post surgical treatment strategy. However, despite the widespread use of chemotherapeutics to treat late stage carcinomas, the molecular mechanisms that determine clinical response in patients remains unclear. As a result, significant portions of those patients who are prescribed chemotherapy derive no tangible benefit from this treatment and are potentially at greater risk of toxic over treatment. In addition to problems associated with toxicity, there is a substantial financial burden on the health care system for continued ineffective treatment regiments[51-53]. It is critical to gain a better understanding of the underlying molecular mechanisms involved in both carcinogenesis and patient response. It is also essential to identify better prognostic markers so that we can better segregate patients into those who are most likely to benefit from current adjuvant therapy and to design more effective regiments in the future for those who do not benefit from current strategies. These are the primary goals for the continued and future clinical management of patients diagnosed with CRC.

Molecular Biomarkers in CRC

Although the TNM staging system is the most effective predictive tool in the clinical management of CRC, the use of the TNM system alone is not very effective at determining the efficacy of the adjuvant therapies. Recent developments in genomics and the ability to perform genome-wide association studies with next-generation sequencing have provided a critical boost to molecular data with

respect to clinical outcomes and carcinogenesis in CRC. The use of molecular markers is gaining popularity as an effective tool to predict clinical response in patients with various treatment regiments. These markers have also provided the possibility for future targets of therapeutic intervention. Several potential biomarkers have been described; yet only mutations in *KRAS* have been largely used as clinical predictors in the treatment of CRC (Table 1[59]). This is largely because most of the biomarkers studied thus far have failed to definitively produce molecular signatures that validate their respective clinical outcomes[54].

In other cancers, such as, certain leukemias and breast cancers molecular gene expression profiles have either been approved or are in the process of approval by the US Food and Drug Administration (FDA) as decision-making tools in support of particular cancer treatment options[54].

The use of the *APC* gene which, when deactivated, disrupts the *APC/WNT* pathway, looks promising as a potential early marker of carcinogenesis in adenomas, as inactivation usually occurs in the normal epithelium. Additionally, the status of *TP53* and *TGF*-β/*SMAD4* have shown promise as biomarkers, because the loss of either or both has been implicated in the enabling of clonal expansion of tumor cells in the invasive adenocarcinomas [55,56].

The members of the Ras family of genes are group of three proteins that operate downstream of several receptor tyrosine kinase (RTK) growth factors (e.g., epidermal growth factor receptors (EGFR); mitogen-activated protein kinase (MAPK) and PI3K pathways) and ultimately impose regulation of cellular growth in normal cells. They have been implicated in many cancers, including colorectal cancer where nearly 40% of adenocarcinomas have been shown to have somatic mutations in *KRAS*[54]. However, mutations in KRAS are also commonly found in polyps and adenomas that rarely progress to malignancy so KRAS is not a marker of requirement in the progression from colorectal adenoma to carcinoma, but it clearly helps drive the development of advanced CRCs and is associated with poor prognosis of the disease[57].

Mutations in the BRAF gene are closely associated with an altered form of the typical adenoma-carcinoma progression known as CIMP where the DNA is highly methylated. Therefore, it has significant potential to be used a diagnostic classification marker, however, it is also associated with poor outcomes and so it also has potential as an important prognostic indicator of survival[58].

To a lesser degree, mutations in TP53, CMYC, PTEN, AKT, PIK3CA, SOX9 and

SMAD2/4 genes are found in small sub-classes of CRC and are potentially prognostic, but further clinical analysis is necessary before their worth as prognostic biomarkers can be determined[60] (Table 1[59]).

To date the majority of gene expression sequencing studies on CRC have focused primarily on coding mRNA expression, and thus many non-coding regulatory elements, such as microRNAs that may have significant prognostic value as biomarkers may have been overlooked. However several microRNAs have been described seemingly with significant potential as biomarkers in various stages of carcinogenesis in colorectal cancer. For example, miR-17-92, miR-135, and miR-145 have been implicated in early progression from normal epithelial tissue to the formation of adenomas. In addition, let7, miR-18a, miR-21, miR-126, miR-143, miR-34a-c, and miR-483-3p[61] all seem to play an important role in the progression of adenoma to carcinoma, but their relative prognostic values need further clinical study[59] (Figure 4[59]).

As the study of potential CRC biomarkers and expression profiles expands and as significant improvements are made into the design and clinical patient data collection, molecular biomarkers will become a routine and an effective predictive method to support decisions on the future clinical management of CRC.

METHODOLOGIES

Small RNA library generation and sequencing.

Written informed consent was obtained for all patients, and the institutional review board (IRB) approved the study. Serial cryosections were obtained from all tumors. The first and last cryosections of each series were used to verify tumor cell content. Samples were only included in this study if the tumor cell content was >70 %. Cryosections not used for histological analysis were transferred to TRIzol, and total RNA was extracted using the miRNEASY kit (Qiagen) according to the manufacturer's recommendations. For all samples, 2 μg of extracted total cellular RNA was size selected by gel electrophoresis and excision to preserve to <40nt RNA faction for subsequent library preperation. Libraries for deep sequencing were prepared from the size-selected total RNA according to the manufacturer's protocol [SREK (small RNA expression Kit), Life Technologies, Foster City, CA.], with one notable exception: during the library amplification, only 12 rounds of PCR were used as opposed to the 15 that are called in the protocol. The reason for this was to reduce the amplification noise and adaptor amplification, as the majority of target lengths were atypically small ~17-25bp for PCR templates. Library integrity was monitored using a Bioanalyzer (Agilent). Template bead preparation, emulsion PCR and deposition steps were performed according to the standard protocol, and slides were analyzed on a SOLiD system Version 3.0 (Applied Biosystems).

Sequence processing and mapping

Mapping of SOLiD reads was performed using both the small RNA pipeline (Life Technologies) and PASS. The small RNA pipeline and PASS were used to extract counts and extensions of miRNA in small RNA reads, from 18 nucleotides in length. When matching to either miRNome (precursor sequences from miRBase) we recorded only perfect matches. When we aligned the short RNA reads to the whole genome, we recorded alignments with up to one mismatch. Only reads with at least 3 sequenced reads per sample were inputted in the SQL database. Raw digital expression values (read counts) were obtained by summing the number of

reads that mapped to one of the reference databases, human genome hg19, miRBase release 16.0, viral or bacterial genomic sequences from NCBI. The confidence in the correct assignment of short reads to miRNAs or other genomic locations was increased by discarding reads mapping to more than 4 loci (as almost all known human miRNAs are equally or less repetitive). For merging of short RNA reads into short RNA contigs (shortigs), we considered each mapped read with at least 3 counts in each sample, either normal or cancer, maintaining the strand. We merged into a single transcriptional unit all the reads within a distance of less than 100 nucleotides. We also assigned to each shortig a score which was the sum of the distinct reads (not the counts of reads) for each sample/patient (i.e. 1 read in 5 patients = 5, 3 distinct reads in 5 patients = 15. Thus this score does not take in account transcriptional activity (i.e. counts of reads per sequence per sample). We finally retained the 270,216 merged shortigs with consistent transcription by using a score threshold of 5.

To quantify the short RNA reads we used two modifications of RPKM scaling[85], based on the read count of each analyzed sample. For short RNA reads , the index consisted of the read count divided to the number of (millions of) mapped miRNA reads in the sample (RPMM). As the length of the short reads was almost constant, we did not use here the division by length in kilobase. Quantiles normalization was used after RPMM scaling. Thresholding was at equal to, or less than, 5 RPMM. Allowed percent absent values for each short sequence, were 85%. Datasets with less than 1 million matched reads were not analyzed further. For the quantification of short RNA contigs, the length in kilobases was used to standardize the reads per million matched miRNAs, thus defining RPKMM. RPKMM was used in place of RPKM, because we used size selected RNA as starting material, rather than un-fractionated RNA.

Additional miRNA detection techniques

Total RNA (20ng) was reverse transcribed using the RT stem–loop primer system (Applied Biosystems), enabling miRNA-specific cDNA synthesis. Subsequent RT-PCR with Exiqon LNA kit was also used.

Northern blot analysis was performed as previously described[66].

NanoString assays were performed as described by the manufacturer.

Statistical analysis

R (http://www.r-project.org) and BRB Array tools were used for statistical and clustering analysis. Filtering of expression tables was performed as follows. Reads or contigs were not analyzed further when less than 20% of the expression data had at least a 1.5-fold change in either direction from the median value and when percent of data missing or filtered out exceeds 50%. We decided to use quantiles normalization on the short RNA reads, after RPMM standardization, as all the evaluated parameters indicated a clear improvement. 2969 short RNA reads were differentially expressed in colon adenocarcinoma (p-value< 0.05), with a global p-value of: 0.007. The shortigs identified by at least 1 significant short read were further studied by summing the RPMMs of all the spanning short reads. For each shortig RPKMM were obtained (reads per kilobase per million miRNAs). Paired t-test was performed on the selected RNA shortigs (same filtering conditions as for short RNA reads). Classification prediction was performed using different models (diagonal linear discriminant, nearest neighbors, and nearest centroids) and incorporated non-coding RNA shortigs that were differentially expressed at the 0.05 significance level, as assessed by the random variance t-test. The misclassification error was estimated for each model by using leave-one-out cross-validation[86].

RESULTS AND DISCUSSION

Small RNAs are regulatory class of evolutionarily conserved, non-coding RNAs (ncRNAs) that are involved in the regulation of gene expression. Alterations in the expression these small non-coding elements have been shown over the past several years to contribute to the disruption of messenger RNA (mRNA) expression and ultimately to the pathogenesis of most, if not all, human malignancies. These alterations can be caused by various mechanisms, including deletions, amplifications or mutations involving miRNA loci, epigenetic silencing or the de-regulation of transcription factors that target specific miRNAs[65]. Some miRNAs have very strong association with cancer[66]. Among them, miR-21 is over-expressed in most tumor types[67]. Over-expression of miR-21 in mouse leads to a pre-B malignant lymphoid-like phenotype, demonstrating that mir-21 is a genuine oncogene[68]. When miR-21 is inactivated, the tumors regress completely in a few

days, partly as a result of apoptosis. These results demonstrate that tumors can become addicted to onco-miRs. On the other hand some miRNAs are strongly down-regulated in cancer, such for example miR-145[69,70] that can regulate the quiescent versus proliferative phenotype of smooth muscle cells[71]. Consequently, there are some important applications of miRNAs with clinical relevance. First, miRNAs have been proposed as biomarkers in early diagnosis of cancer by non-invasive techniques. Second, because malignant cells show dependence on miRNAs, which in turn control, or are controlled by, multiple protein-coding cancer genes, these small molecules provide opportunities for the development of RNA-based therapies. The advantage of a miRNA approach is based on its ability to concurrently target multiple effectors involved in cell differentiation, proliferation and survival[72].

Our goal was that of systematically sequence all short RNAs in colon adenocarcinoma, including miRNAs, to assess the absolute expression and diagnostic significance of the expanding classes of non-coding RNAs. First, to test our procedure, we used the short RNA reads to measure only miRNAs, as described in miRBase[73]. We used two different methods, the Small RNA Pipeline (SRP)[74] and PASS[75]. The algorithms were both implemented in a pipeline which funneled all the alignments from the patient cohorts into a SQL database. miRNAs were identified by next generation sequencing with perfect match to the miRBase precursor sequences. Table 2 shows the miRNAs which discriminate between colon adenocarcinoma and normal tissues (p-values <0.01). Fold change was the ratio of geometric means of RPMM (reads per million of matched miRNAs) in adenocarcinoma vs. normal paired samples from the same patient. Mature and isomiRNA forms (i.e. different mature reads for the same miRNA) were annotated according to miRBase 16. Only the isomiRNAs detected by both the SRP and the PASS algorithms are listed in Table 2. The miRNAs identified with the two methods were essentially overlapping, albeit PASS was more sensitive, with a gain of 53% in the number of significant iso-miRNAs (393 in PASS vs. 256 in SRP). The results for the individual pipelines are reported in Table 3. There was a very good correlation between the miRNA counts using the two algorithms (Figure 8, adjusted R square = 0.96). When the ends of iso-miRNAs used were plotted along the precursor sequences, it was apparent that the seed region in the isomiRNAs was either entirely or partially identical to that of the canonical mature form. Figure 9 and figure 10 both show the respective ends' usage graphs for miR-

21 and miR-145. These iso-miRNAs were therefore expected to share the same biological activity of the respective mature forms. We then plotted for each miRNA the cumulative RNA counts along the precursor sequence. The plots for 397 distinct miRNA are available online (http://aqua.unife.it/miRNAplots). Among them we detected discrepancies between 150 miRNAs and their miRBase definitions. The miRNAs for which discrepancies were present are listed in Table 4 and were particularly frequent among the most recently discovered miRNAs. In detail, 9 miRNAs were expressed mainly from the opposite strand of the precursor, 20 were identified by different mature form coordinates, and 19 putative miRNAs revealed an aberrant expression plot much unlike that of a classical miRNA. Further, we measured higher expression of the star over the mature form in 33 miRNAs.

A major aim in our work was to identify novel non-coding RNAs, beyond known miRNAs, having a diagnostic value in colorectal adenocarcinoma. We thus aligned all the sequenced short RNAs to the whole human genome, using the PASS pipeline, and without *a priori* distinguishing miRNAs from other genes. In order to accurately measure the expression levels and to correctly map short RNAs we used only perfect matches to the human genome. The trials we did allowing even only 1 mismatch in fact resulted in a large number of mapping ambiguities. Perfect matching would in principle be a problem for the determination of SNPs, but we disregarded this, because we were not concerned with presence of SNPs at this time, due to the very short size of the reads it is very difficult to very SNPs occurring at the level of RNA processing. When only perfect matches were considered in the alignments of the colon samples, a total of 477,595 distinct short reads were mapped. If strand was not considered, we identified 476,882 distinct loci; i.e. less than 1000 loci were transcribed on both strands. Perfect match yielded about 66% of mapped reads, while allowing for 1 mismatch the aligned reads were more than 99%. Again, we decided not to use the alignments with one mismatch because, even if mapping to just one locus, they vastly increased the number of loci and decreased the statistical significance of data, hinting at an increase in noise, rather than information. Among all the distinct RNA reads we needed to identify those produced by defined and consistent transcriptional units, and remove lone RNA molecules detected only in very few samples or at very low level. To attain this goal, we used all the mapped short RNA reads in our samples to define a genome wide map of short RNA loci with consistent transcriptional

activity. We did so by preserving RNA transcription strands. By merging the reads with close spatial contiguity, we identified 270,216 short RNA contigs (hereafter named shortigs) with consistent transcriptional activity in cancer and normal samples. All miRNAs expressed in colon were present among these 270K shortigs (a Genome Browser custom track with all the genomic coordinates for the 270K shortigs can be obtained at http://aqua.unife.it/ShortigShapes). About 42% of the shortigs were found to be overlapping to non annotated ESTs, 4% mapped to repeats and only 12% mapped to highly conserved sequences (phastConsElements46way UCSC table). Nevertheless when the these sequences were compared to ESTs from other organisms (XenoEST table), a large portion (91%) found an homologue, showing that most of the shortigs are transcribed in some organisms. The shortigs were annotated according to UCSC Genome Browser. Since it was possible that for short RNAs the traditional gene models were too conservative, we used the Gencode[76] gene model alongside the established UCSC annotation system. A graphical representation of a shortig is shown for the miR-17-92 cluster (Figure 5A).

We finally proceeded to identify the RNA shortigs with diagnostic value in colon carcinoma. Reads per kilobase per million miRNAs were used to measure expression of the shortigs. T-test and permutations were performed to assess p-values and false detection rates. Overall, 129 RNA shortigs were differentially expressed in colon carcinoma, with p-values <0.05 in paired t test (Table 3). This non-coding RNA signature was also efficient in predicting cancer and normal colon samples. The leave-one-out cross-validation method was used to compute the misclassification rate: > 90% with diagonal linear discriminate analysis, nearest neighbors or nearest centroids (Table 5). Fifty-six of the shortigs (43.4%) coincided with miRNA precursors or clusters (Figure 5B). Among the remaining RNAs, 52% were novel RNAs with unknown function (45 are intragenic and 22 are intergenic). 59 RNA contigs were up-regulated and 70 down-regulated in colon adenocarcinoma. The respective ratio in miRNAs was 19 vs. 37, with a slight excess of miRNA down-regulation in comparison to the non-miRNA counterpart. We performed detailed manual inspection for the genome location of all differentially expressed novel RNAs. Four additional annotated shortigs were thus identified: a couple of tRNA-like genes (chr11:65273440-65273625+ at 11q13.1 and chr10:69524258-69524366+ at 10q21.3), a novel snoRNA (SNORD19B, at 3p21.1) and a piRNA (piR-51810 at 8q13.2). Seven novel RNA genes were

located very close (<1 kb) to regions of chromatin modification as reported by the Encode[77] tracks. Among the shortigs located within introns of coding genes, there was an equal distribution between sense and antisense orientations with respect to the host gene. The expression range, subdivided by RNA classes, was highest for miRNAs, with a 2 logs difference in the maximal values (Figure 5C). The steepness of the curves in cancer and normal samples were similar within the different RNA classes, with the exception of the small snoRNAs/piRNA class, showing a noticeable bump in cancer. The tree representing the cluster analysis of the short RNA contigs is shown in Figure 6. For each of the 129 diagnostic shortigs we generated an expression plot with cumulative RPKMM along the contig. Plots for two miRNAs, a piRNA and a snoRNA are shown in Figure 7. It is interesting to note that the peak of the snoRNA within the GNL3 host gene was larger than the individual sequencing reads. All the shortig plots are available online at http://aqua.unife.it/ShortigShapes. Some miRNAs were sporadically, or not at all, reported in colon cancer, like up-regulated miR-135b, miR-503, miR-183[78], miR-182 and down-regulated miR-129, miR-137b, miR-9, miR-138, miR-218, were present in the short RNA signature. Expression of p53 responsive miR-215[79], implicated in cell-cycle arrest, was also decreased in colon adenocarcinoma. Many of these miRNAs had lower RPKMM in comparison to the most prominent miRNAs in the colon adenocarcinoma signature, like miR-21[80] and miR-145[81] (Table 2). It is possible that low abundance miRNAs might have often been discriminated against in the studies performed with less sensitive or robust detection techniques. The abundant miR-29a was over-expressed in CRC, while the related low abundance miR-29b was down-regulated (Table 6). When we validated the expression of these two miRNAs, with RT-PCR and Northern blot the signal of miR-29b (low RPKMM) was confounded with that of miR-29a (high RPKMM), and thus erroneously called as up-regulated. Only by using LNA-based PCR and Nanostring we could validate the down-regulation of miR-29b. The summary of the validation of a set of miRNAs by using different technical platforms pointed out to next-generation sequencing as the most robust detection method. In order to substantiate the relevance of these non-coding RNAs in cancer we studied their presence in chromosomal areas associated to copy number variations. We used the comparative genomic hybridization (CGH) Progenetix database[82] to identify regions of amplification or deletion in a large number of cancer samples. We called an area as amplified or deleted, if there was a

corresponding 3-fold change in the number of abnormalities detected over a cytoband. A number of amplifications were associated to over-expressed RNA shortigs (Table 7). In particular, miR-135b, miR-183, miR-182, miR-21, miR-29a and miR-25 were in frequently amplified regions (cytobands with number of amplified cases over number of deleted cases > 3). RNA shortigs which were also amplified, included SNORD12, SNORD54, SNORD78, SNORD123 and other novel short RNAs.

CONCLUSION

We used short RNA sequences in the range between 18 and 35 nucleotides to *de novo* assemble non-coding RNA genes and measure their activity in colon adenocarcinoma and normal cells. The technique we used covered contiguous stretch of RNAs, and allowed the identification of miRNA precursors, or even primary RNAs from miRNA cluster, like the miR-17-92 locus on chromosome 13[83,84]. The quantification of short RNAs was also used to derive a transcriptional profile along each RNA contig or shortig. This allowed us to map the mature and star forms along each human miRNA precursor. We detected a number of expressed isoforms for each miRNA, but they did not seem to change the targeting specificity determinant, the seed region. The slope on the 5' of mature miRNAs was usually steeper than that on the 3' end. Finally, some of the differentially regulated short RNAs are routinely used as normalizers in various molecular biology assays due to their seemingly stable and ubiquitous expression, which is alarming. For instance, U48, which is often employed as a normalizer in PCR and other assays, was noticeably varied in expression between groups. It is apparent that such usage might significantly affect the outcome of any genome wide assay. The key finding in this study, was that we demonstrated that non-coding RNA is differentially regulated in colorectal adenocarcinoma, but is not limited to mRNAs and miRNAs alone, as previously surmised, but also includes deregulation at the snoRNA and piRNA levels as well. These effectors can exert key control over vast number of cellular functions, such as alternative splicing and gene silencing[64].

FUTURE DIRECTION

We have obtained an additional 24 paired normal/adenocarcinoma genomic DNA and total RNA samples from the National Cancer Institute (NCI) to continue work on profiling coding and non-coding RNAs in CRC, as well as targeted DNA re-sequencing of a cancer-specific mutanome using SureSelect (Agilent Technologies) bait library that we have designed which includes genome loci for all microRNA primary sequences; exons from all microRNA processing genes (e.g, *DROSHA*, *DICER*, etc.) and the exons from approximately ~450 cancer-associated genes. We believe there significant mutations and/or SNPs present in these loci that could have both prognostic and diagnostic impact on colorectal cancer that either would not be present at the RNA level or would be very difficult to find and/or validate at the RNA level.

In addition, to these, we designed nearly 38,000 small and large non-coding target sequences throughout the genome with an emphasis on fragile sites, breakpoints and other genomic translocation hotspots. The intention is to use the RNA samples for both mRNA and small RNA transcriptome studies to further characterize the difference in expression profiles between normal colon and adenocarcinomas. The 48 additional samples contain 32 (16-paired) new and previously unused samples, while 16 (8-paired) of the 48 samples are the same sample we have used in these studies. We intend to both cross validate these findings as well increase our statistical significance on the latest miRbase build.

LEGENDS

Figures

Figure 1. Roche 454 GS FLX sequencing.

Template DNA is fragmented, end-repaired, ligated to adapters, and clonally amplified by emulsion PCR. After amplification, the beads are deposited into picotiter-plate wells with sequencing enzymes. The picotiter plate functions as a flow cell where iterative pyrosequencing is performed. A nucleotide-incorporation event results in pyrophosphate (PPi) release and well-localized luminescence. APS, adenosine 5_-phosphosulfate.

Figure 2. Illumina Genome Analyzer sequencing.

Adapter-modified, single-stranded DNA is added to the flow cell and immobilized by hybridization. Bridge amplification generates clonally amplified clusters. Clusters are denatured and cleaved; sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Post-incorporation fluorescence is recorded. The fluorofor and block are removed before the next synthesis cycle.

Figure 3. Applied Biosystems SOLiD sequencing by ligation.

Top: SOLiD color-space coding. Each interrogation probe is an octamer, which consists of (3_-to-5_ direction) 2 probe-specific bases followed by 6 degenerate bases (nnnzzz) with one of 4 fluorescent labels linked to the 5_ end. The 2 probe-specific bases consist of one of 16 possible 2-base combinations. Bottom: (A), The P1 adapter and template with annealed primer (n) is interrogated by probes representing the 16 possible 2-base combinations. In this example, the 2 specific bases complementary to the template are AT. (B), After annealing and ligation of the probe, fluorescence is recorded before cleavage of the last 3 degenerate probe bases. The 5_ end of the cleaved probe is phosphorylated (not shown) before the second sequencing step. (C), Annealing and ligation of the next probe. (D), Complete extension of primer (n) through the first round consisting of 7 cycles of ligation. (E), The product extended from primer (n) is denatured from the adapter/template, and the second round of sequencing is performed with primer (n _ 1). With the use of progressively offset primers, in this example (n _ 1), adapter bases are sequenced, and this known sequence is used in conjunction with the

color-space coding for determining the template sequence by de-convolution (see Fig. 1 in the online Data Supplement). In this technology, template bases are interrogated twice.

Figure 4. The role of microRNAs (miRNAs) in colorectal cancer (CRC) pathogenesis. Selected miRNAs that show altered expression in CRCs, along with their potential messenger RNA targets, are indicated. The scheme is based on the genetic model for colorectal cancer highlighted by Fearon & Vogelstein[25]. Abbreviations: APC, adenomatous polyposis coli; CASP3, caspase 3; CDK4,6, cyclindependent kinase 4,6; ECM, extracellular matrix; CTGF, connective tissue growth factor; DCC, deleted in colorectal carcinoma;
EGFR, epidermal growth factor receptor; EMT, epithelial-to-mesenchymal transition; ICAMs, intercellular adhesive molecules; MMPs, matrix metalloproteinases; mTOR, mammalian target of rapamycin; PDCD4, programmed cell death 4; PTEN, phosphatase and tensin homolog; RECK, reversion-inducing cysteine-rich protein with kazal motifs; SIRT1, sirtuin 1; TGFβRI/II, transforming growth factor βreceptor I/II; TIMP3, tissue inhibitor of metalloproteinase 3; TSP1, thrombospondin 1; uPAR, plasminogen activator,
urokinase receptor; ZEB1/2, zinc-finger E-box binding homeobox 1

Figure 5. A) The plot represents the short RNA contig for the miR-17-92 cluster on chromosome 13. This miRNA cluster is regulated by Myc[22, 23] and is activated in adenocarcinoma. Each line in the plot represents a different sample (cancer in black and normal colon in red). All 6 miRNAs in the locus were correctly identified, as indicated by the UCSC annotation in red. The EvoFold prediction shows that the lighter green areas separate the two peaks, corresponding to the mature and star forms). B) The expression range of the different short RNA classes in cancer and normal colon. The RPKMM indicates the RNA level for each differentially expressed shortig. C) The pie chart shows the distribution of the differentially expressed RNA contigs according to their annotation classes. miRNAs account for less than 50% of the non-coding short RNA

Figure 6. Cluster analysis of short RNA contigs differentially expressed in colon adenocarcinoma (t-test, p-value<0.01). Annotations are according to UCSC hg19 assembly. Antisense RNAs are indicated 'as' after the gene symbol. Non annotated loci are indicated just by the chromosomal location.

Figure 7. Cumulative RPKMM short RNA contig plots for two miRNAs, a snoRNA and a piRNA. All the plots for the 129 significant shortigs in colon adenocarcinoma are available online at http://aqua.unife.it/ShortigShapes. miR-21 and miR-135b are up-regulated microRNAs in cancer, while the piR-51810 piRNA is down-regulated. SNORD19B.2-201 is depicted in the bottom right plot and is over-expressed. Fold-changes indicated in the figure are for unpaired comparison.

Figure 8. Small RNA Pipeline and PASS scatter plot

The fold changes of differentially expressed miRNAs are very consistent across the two profile determined by the two different methods implemented for colon adenocarcinoma (adjusted R square = 0.96). The ABI Life Technologies small RNA pipeline method is indicated as SRP.

Figure 9. Ends usage in the distinct differentially regulated miR-21 isomiRNAs. The mature miR-21 form as reported in miRBase extends from chromosome 17:57918634 to 57918655 with a length of 22 nt. While the most commonly used end is at 57918655, most of the 5'ends are either on the mature form or 1-2 nucleotides 3'. Thus at the 5' end, the seed region is either identical to

the standard miR-21 or generally 1 to 2 nucleotides shorter. This isomiRNAs are therefore expected to have very similar biological activity to that of the mature form. Strikingly the most used 3' end is 1 nt 3' of the canonical site. The frequency indicated on the bar does not correlate to the RPMM of the relative reads, but to the usage in the distinct isomiRNAs.

Figure 10. Ends usage in the distinct differentially regulated miR-145 isomiRNAs. The mature miR-145 form as reported in miRBase extends from chromosome 5: 148810224 to 148810246 with a length of 23nt. The most commonly used end is at 148810224, as reported by miRBase, and most other 5'ends are 1 or 2 nucleotides 3'. Only 1 of the 15 recorded isomiRNAs extends 1 nucleotide longer than the canonical form. At the 5' end, the seed region is either identical to the standard miR-145 or generally 1 to 2 nucleotides shorter. This isomiRNAs are therefore expected to have very similar biological activity to that of the mature form. The frequency indicated on the bar does not correlate to the RPMM of the relative reads, but to the usage in the distinct isomiRNAs.

Figure 11. hg19 loci complexity, i.e. number of distinct reads loci per sample. Sample IDs are on the X axis. Shown data are for alignments with either perfect matching or only 1 mismatch (and 1 genomic hit).

Figure 12. hg19 mapped reads complexity per sample, i.e. number of total counts of mapped loci. Sample IDs are on the X axis. The three worst samples (123, 131, 132) were excluded from the analysis.

Figure 13. Measure correlations between different miRNA detection platforms A) SOLiD vs. Nanostring (9/10 concordant trends within CRC/Normal) R = 0.98, p-value =2.73E-12 B) SOLiD vs. stem loop RT-PCR (7/9 concordant trends within CRC/Normal) R = 0.51, p-value =0.06 C) Nanostring vs. RTPCR (7/9 concordant trends within CRC/Normal) R = 0.41, p-value =0.15

Tables

Table 1. Recurrent somatic mutations in oncogenes and tumor-suppressor genes in colorectal cancer (CIMP, CpG island hypermethylation phenotype; CRC, colorectal cancer; MSI-H, high-frequency microsatellite instability; MSS, microsatellite stability).

Table 2. Short non-coding RNA contigs discriminate colon adenocarcinoma from normal colon. Paired t-test was performed on 254 RNA shortigs identified by 2969 short RNA reads with p-value <0.01. Among them, 129 shortigs had diagnostic values in colon adenocarcinoma with a miclassification error of less than 0.01 and are listed in Table 1. RPKMM were used for quantification (reads per kb per million miRNAs). Although the alignment pipeline did exclude reads mapping to multiple loci, a filter for repetitive elements was performed by using RepeatMasker. As a quality control check, only samples with runs of more than 1 reads million matched to hg19 were used in the statistical analysis.

Table 3. Differentially expressed miRNAs in colon adenocarcinoma, as determined by the small RNA pipeline (SRP) and PASS. miRNAs were identified by next generation sequencing using the small RNA pipeline (Life Technologies, Foster City) with perfect match and miRBase precursor sequences (p-values <0.01). Fold change is the ratio of geometric means of reads RMPM (per million) after quantiles normalization. Mature and isomiRNA forms are annotated according to miRBase 16. miRNAs are sorted by fold changes between tumor and normal tissues. Only perfect matches were recorded (no mismatches allowed). Multivariate permutations test was computed based on 1000 random permutations and parametric p-values are reported, alongside false detection rates (FDR). The isomiRNAs corresponding to the mature form, as reported by miRBase, are shown. Additional isomiRNAs are listed when more expressed or had higher fold changes than the mature form. Only the isomiRNAs detected by both the small RNA (SRP) and the PASS pipelines are listed. In particular, the PASS pipeline produced an expression matrix with 3630 different isomiRNAs, after filtering (threshold of 5 RPMM and a minimum presence in 25% of the samples). Of these, 2289 were retained after filtering, when less than 20 % of expression data have at least a 1.5 -fold change in either direction from isomiRNA's median value. Only

very few genes had values, which average, was below threshold in one of the two classes. When quantiles normalization was performed on the top of RPMM counts, 2280 short RNAs were retained after filtering (9/2289 = 0.4% less than without quantiles normalization). Quantiles normalization yielded 393 significant miRNAs at $p<0.01$ and FDR <0.01, slightly more than without normalization. Overall we decided to use quantiles normalization after RPMMstandardization, as all the parameters indicated an improvement. When using the small RNA pipeline by Life Technologies, again with no mismatches, we identified 1538 isomiRNAs and 256 isomiRNAs with p-value < 0.01.

Table 4. miRNA annotation discrepancies with miRBase. The expression plots for along each precursor miRNA in the colon samples, subdivided in the discrepancy classes, can be downloaded at http://aqua.unife.it/ShortigShapes. In the cases of wrong strand of precursor, the count profiles were plotted for each of the two strands.

Table 5. The performance of colon adenocarcinoma classification using the short non-coding RNA signature. Leave-one-out cross-validation method was used to compute the misclassification rate.

Table 6. Colon adenocarcinoma and normal colon miRNA quantification using different detection techniques (average values). Values indicate averages in each patient cohort for each platform. SOLiD is cohort 1, Nanostring, RT-PCR, Northern blot are platform 2 and Microarrays are cohort 3. Cohort 3 has benign adenomas as controls, in place of normal colon samples. miRNAs with discordant trends are highlighted in yellow. The difference in the miR-145 levels between adenocarcinoma and benign adenoma might be related to early epigenetic differences between adenomas and normal colon tissues.

Table 7. Correspondence between over-expressed non-coding RNAs and amplification by CGH in cancer (Progenetix database[87]). The table lists only the loci for which at least 3 fold excess of amplification over deletion were reported in the same cytoband.

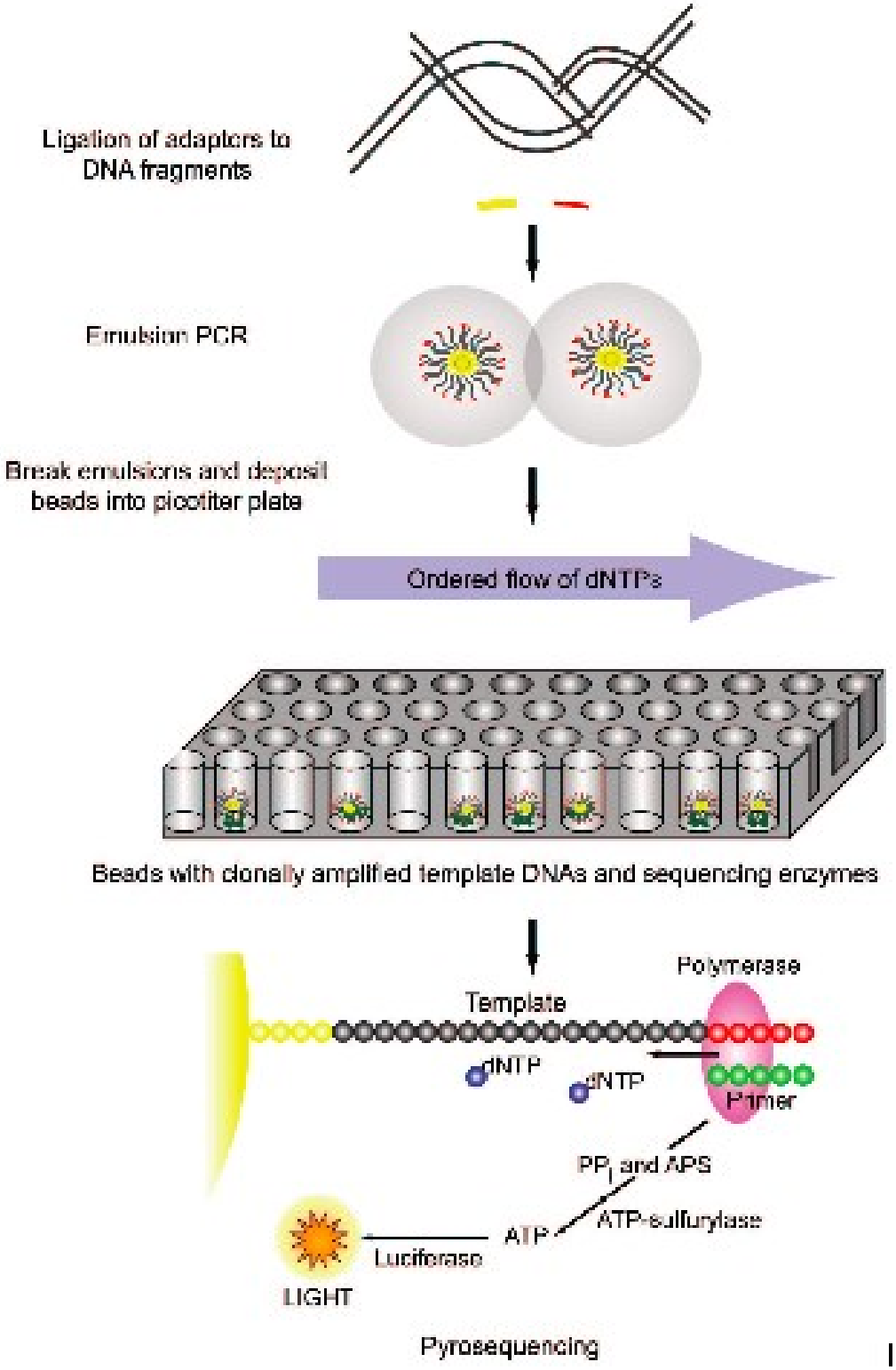

Figure 1. Roche 454 GS FLX sequencing.

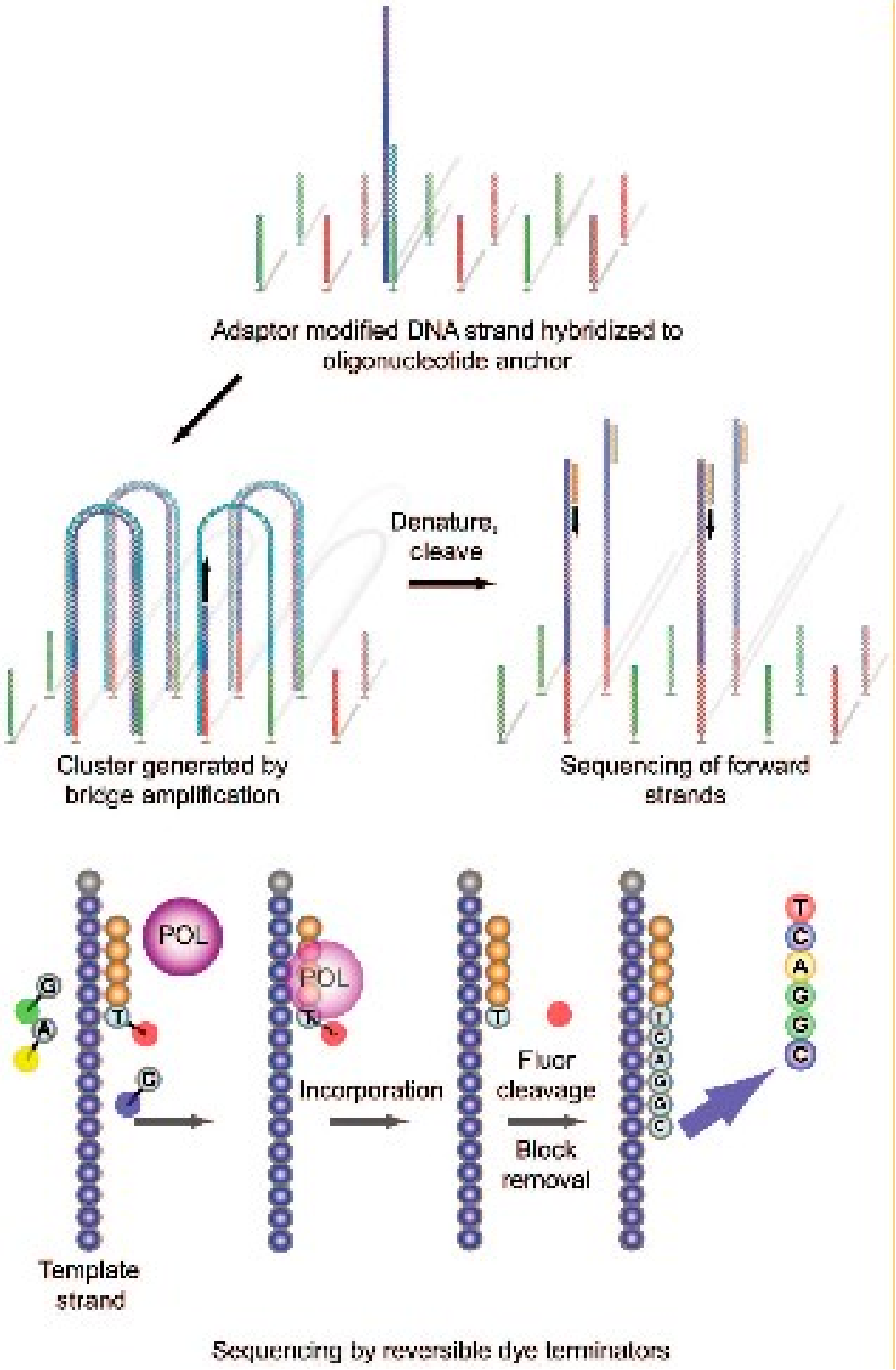

Figure 2. Illumina Genome Analyzer sequencing.

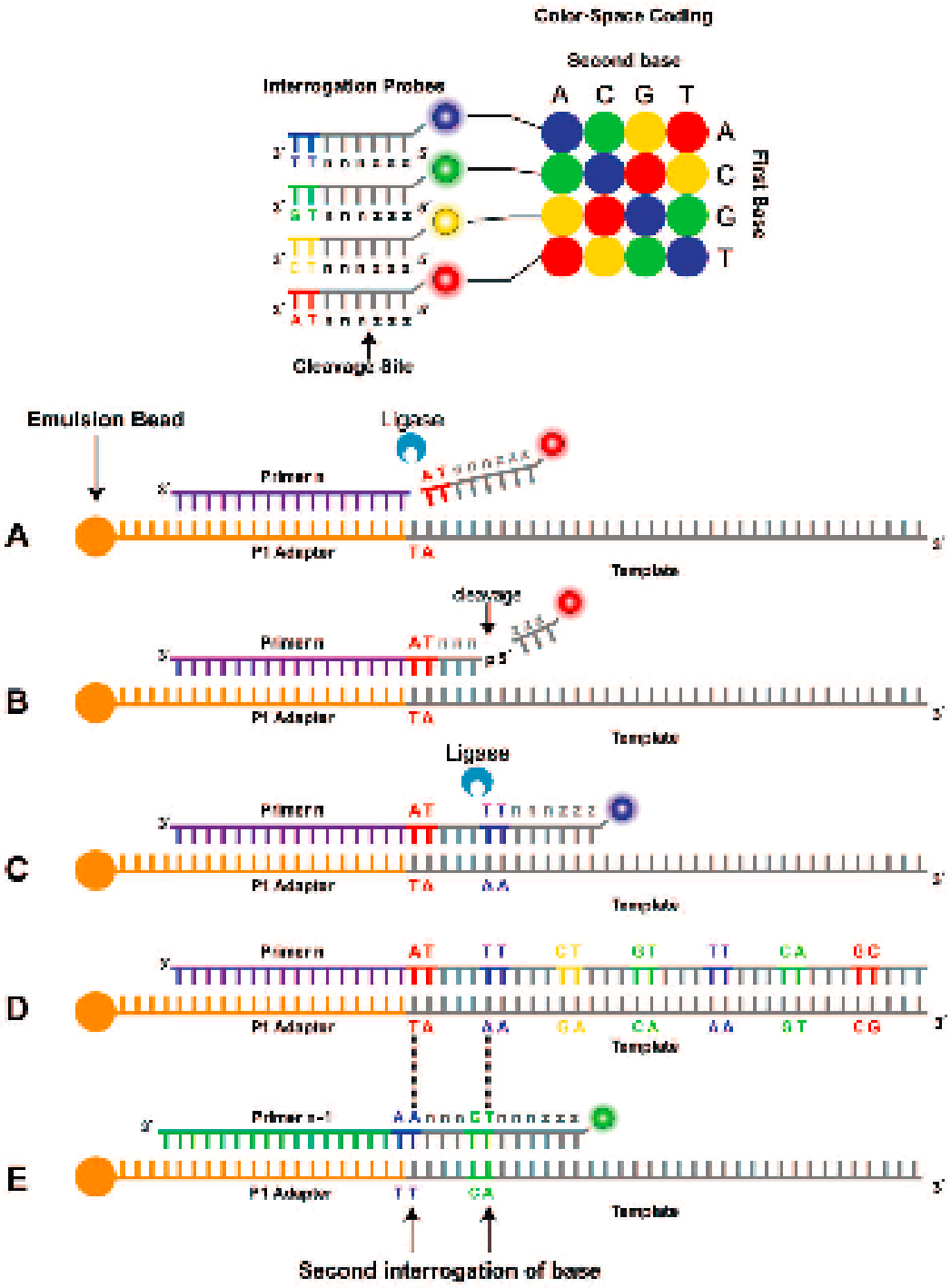

Figure 3. Applied Biosystems SOLiD sequencing by ligation.

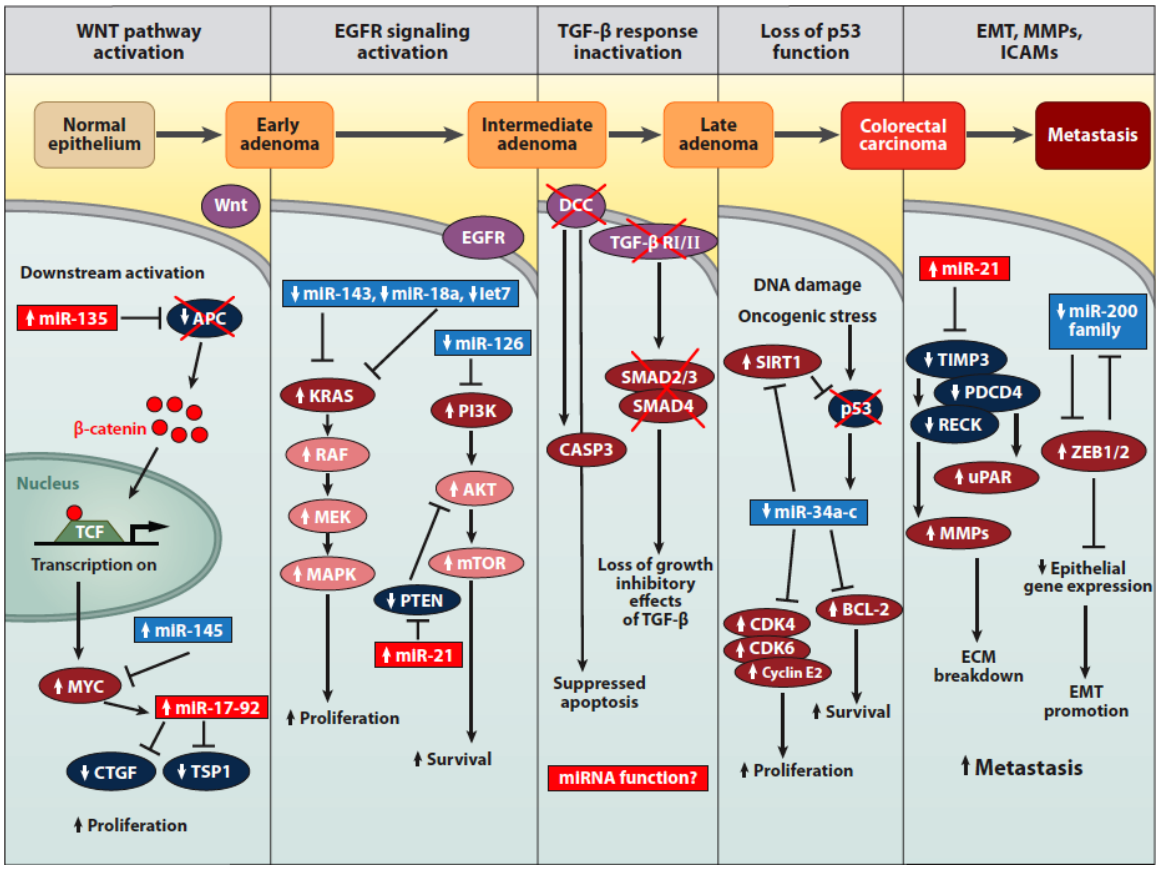

Figure 4. The role of microRNAs (miRNAs) in colorectal cancer (CRC) pathogenesis.

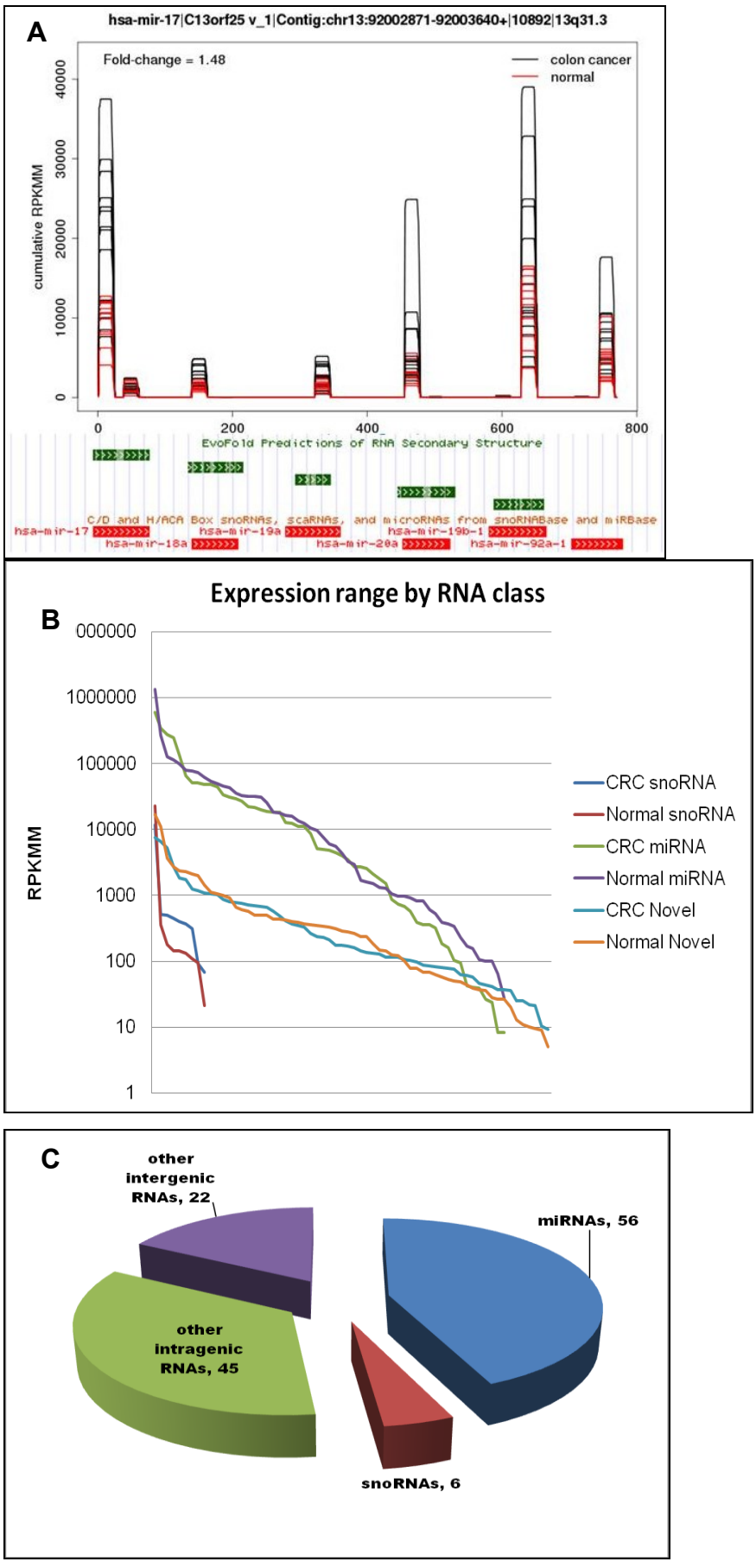

Figure 5. RNA Shortigs in colorectal adenocarcinomas

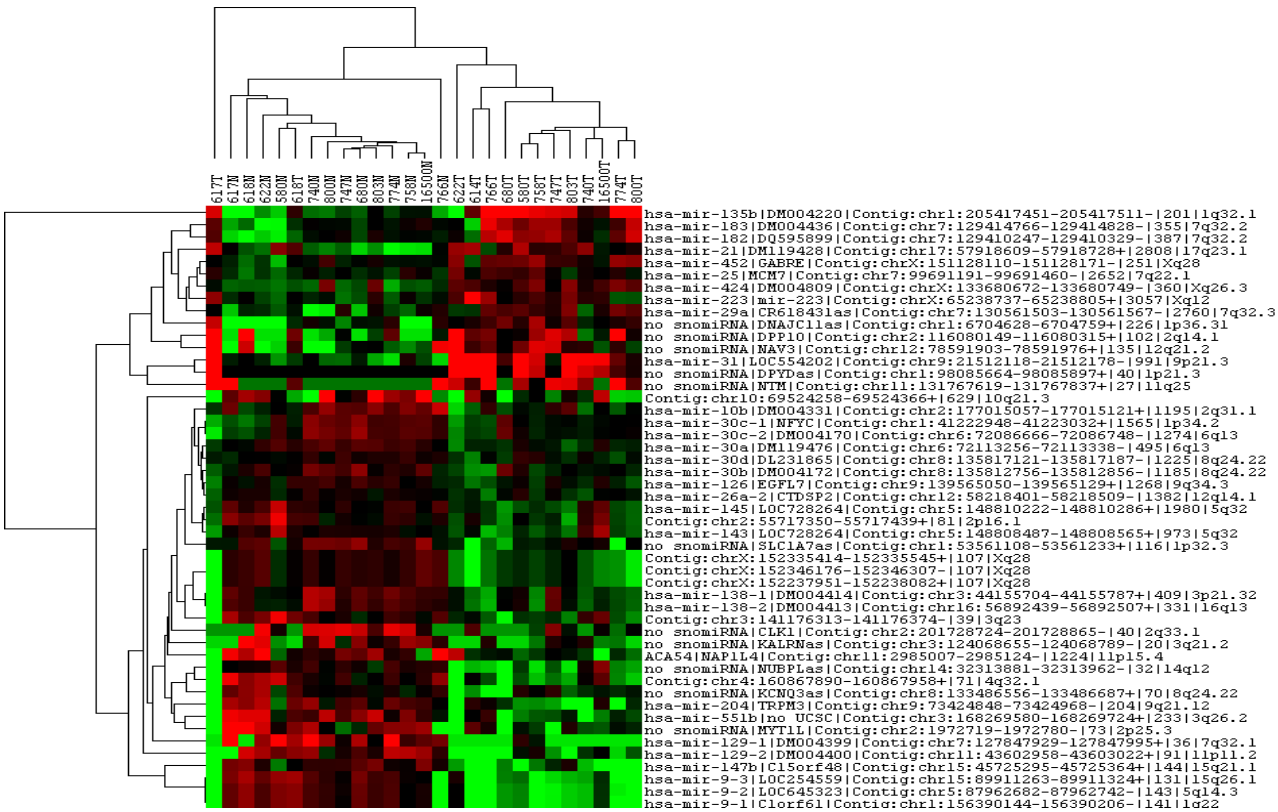

Figure 6. Cluster analysis of short RNA contigs differentially expressed in colon adenocarcinoma (t-test, p-value<0.01).

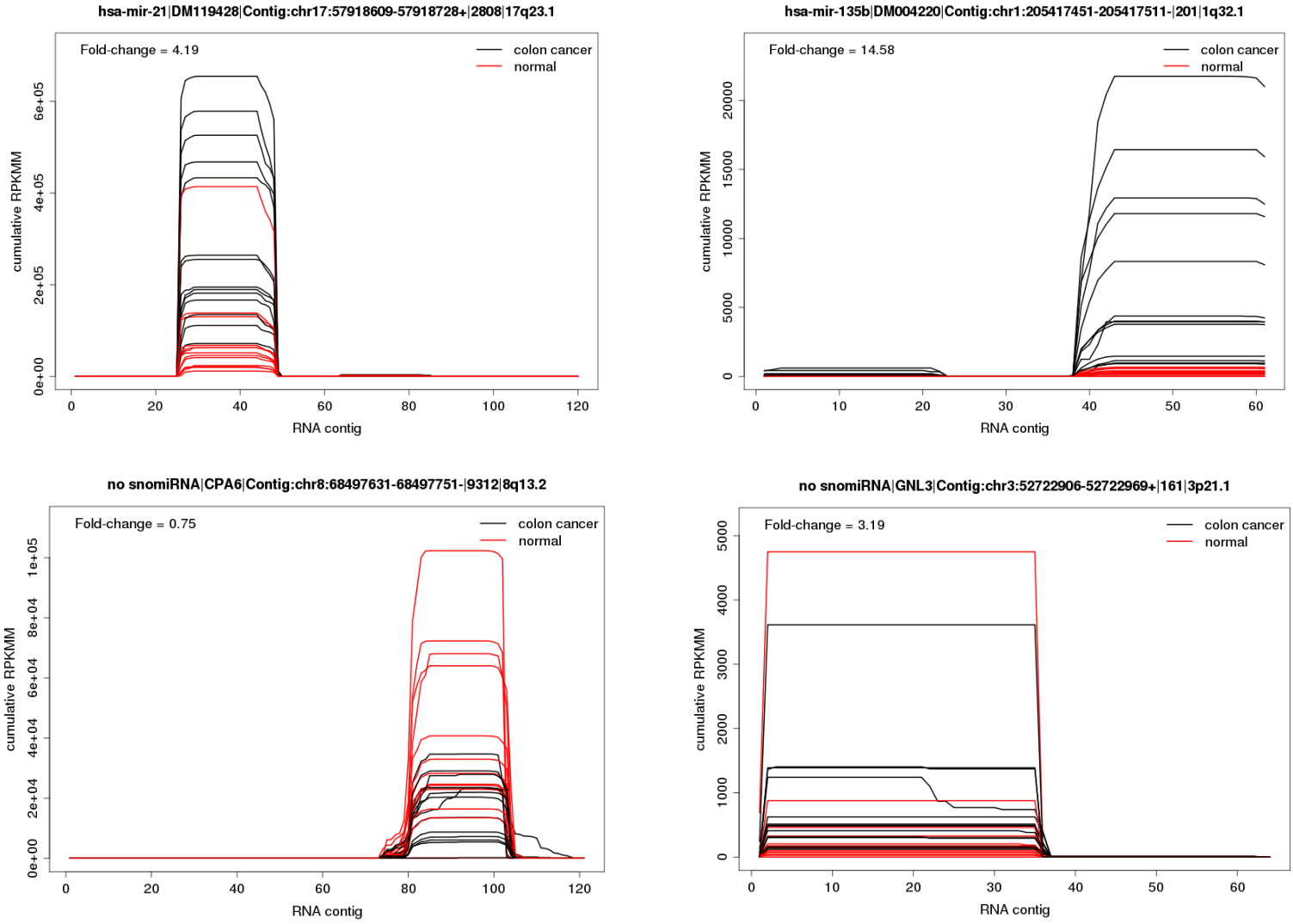

Figure 7. Cumulative RPKMM short RNA contig plots for two miRNAs, a snoRNA and a piRNA.

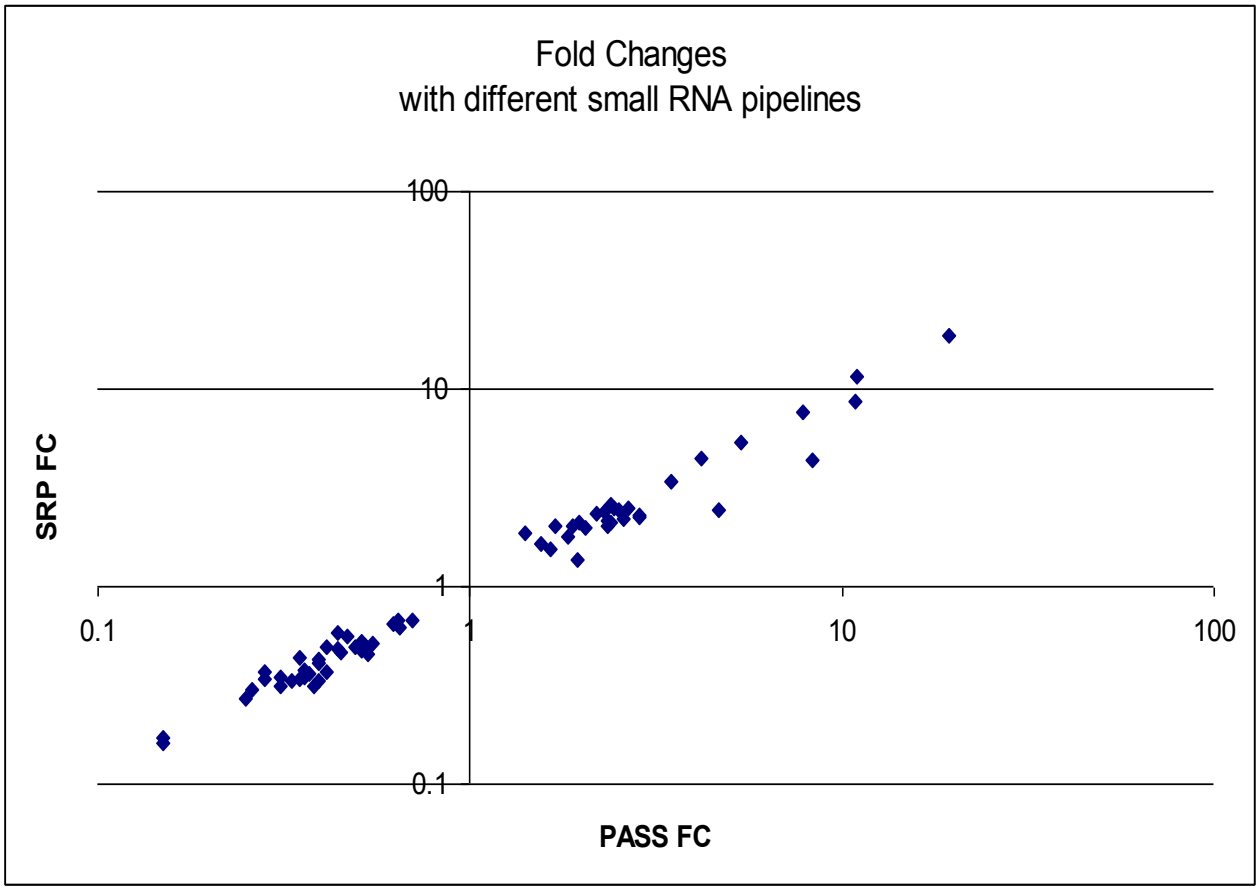

Figure 8. Small RNA Pipeline and PASS Scatter Plot

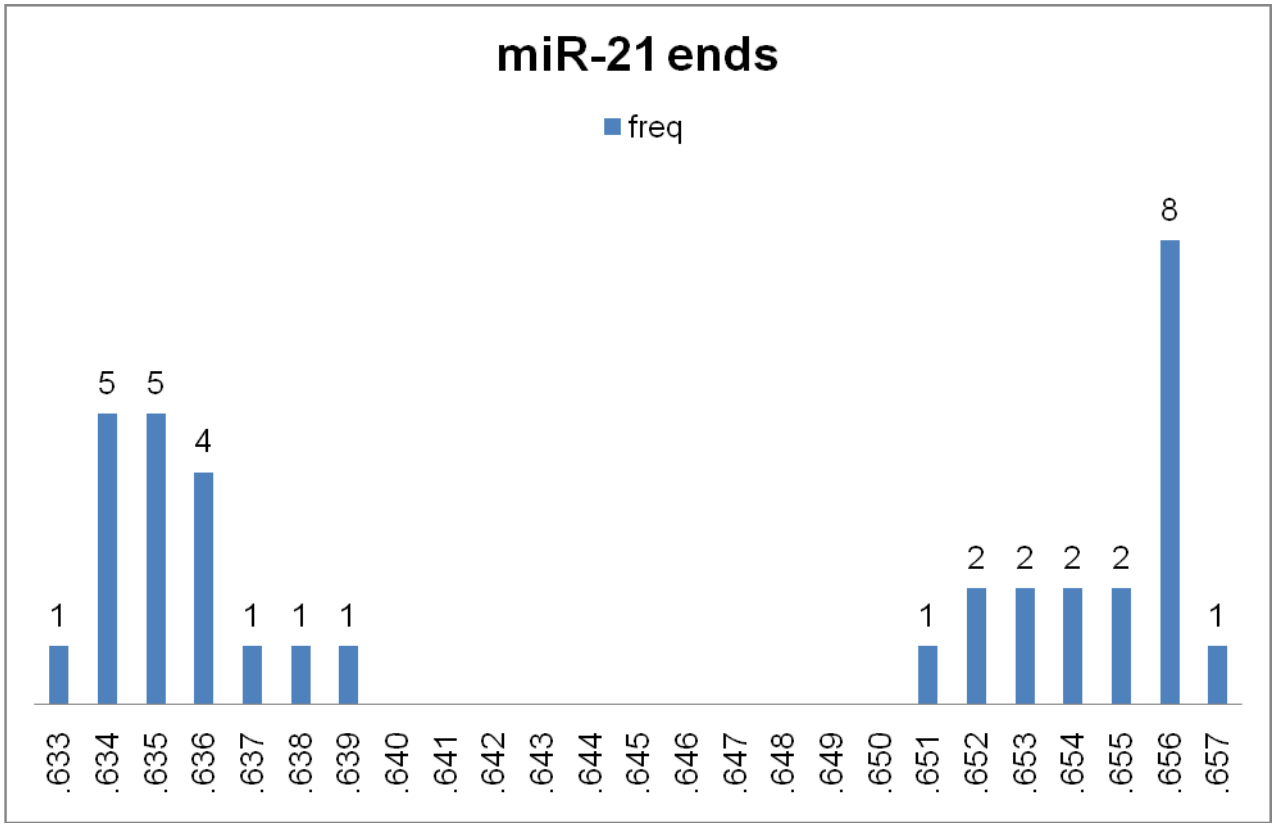

Figure 9. Ends usage in the distinct differentially regulated miR-21 isomiRNAs.

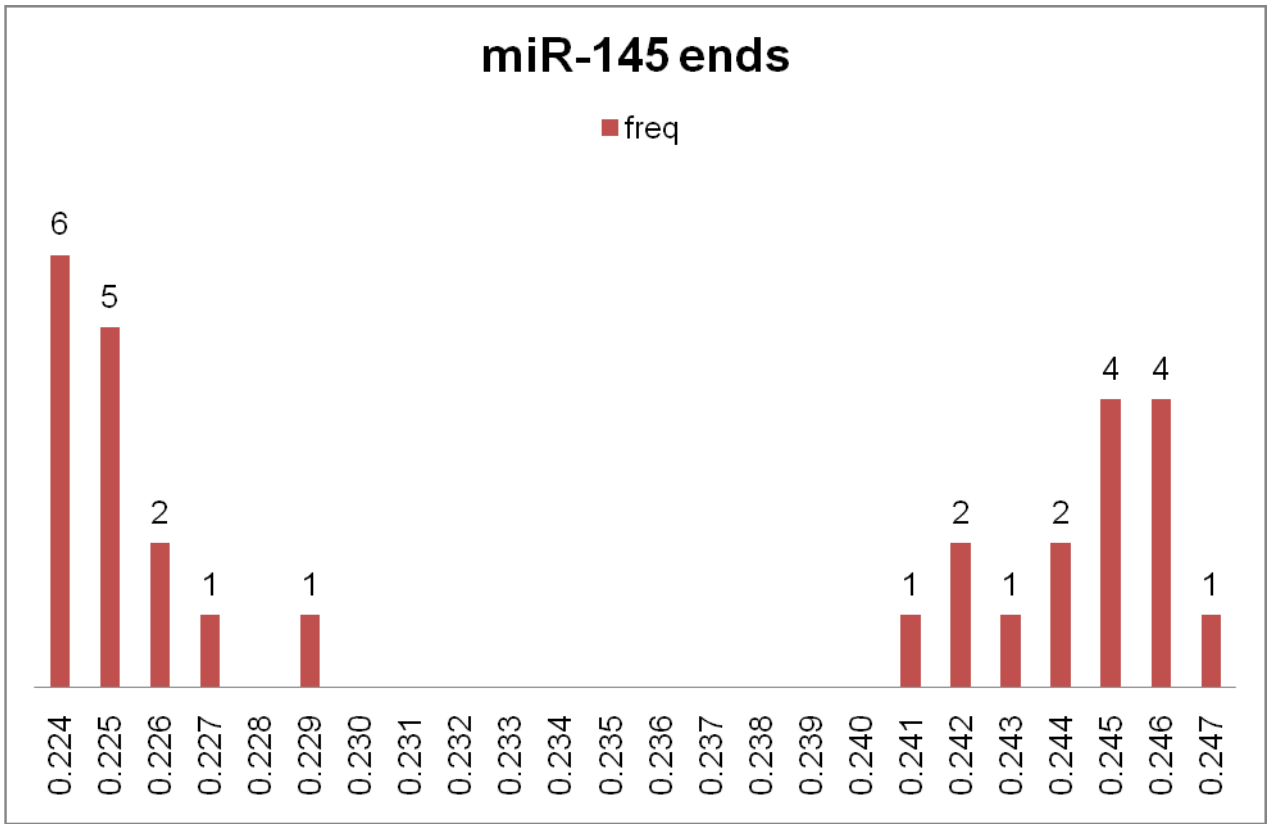

Figure 10. Ends usage in the distinct differentially regulated miR-145 isomiRNAs.

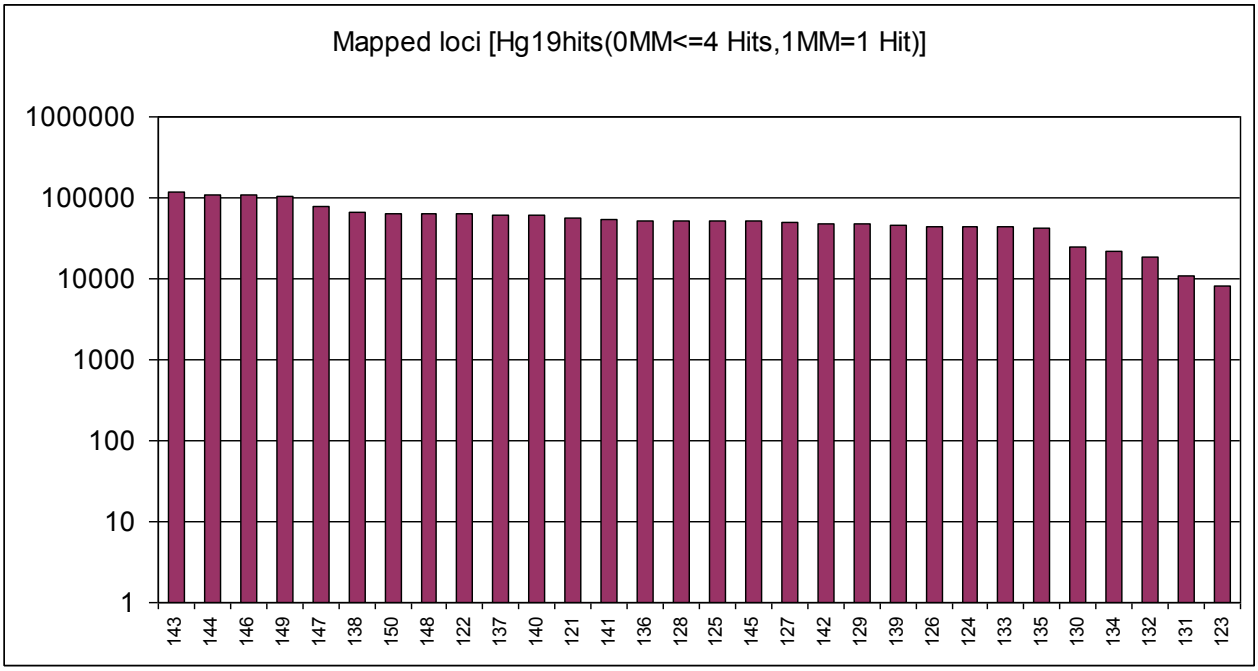

Figure 11. hg19 loci complexity

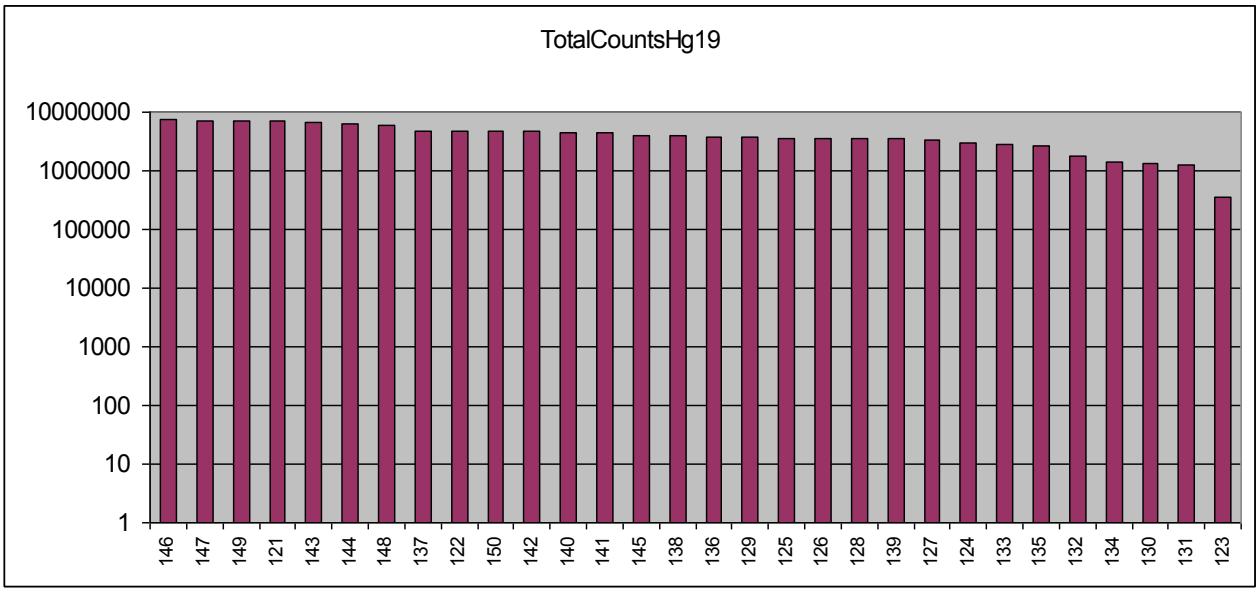

Figure 12. hg19 mapped reads complexity per sample

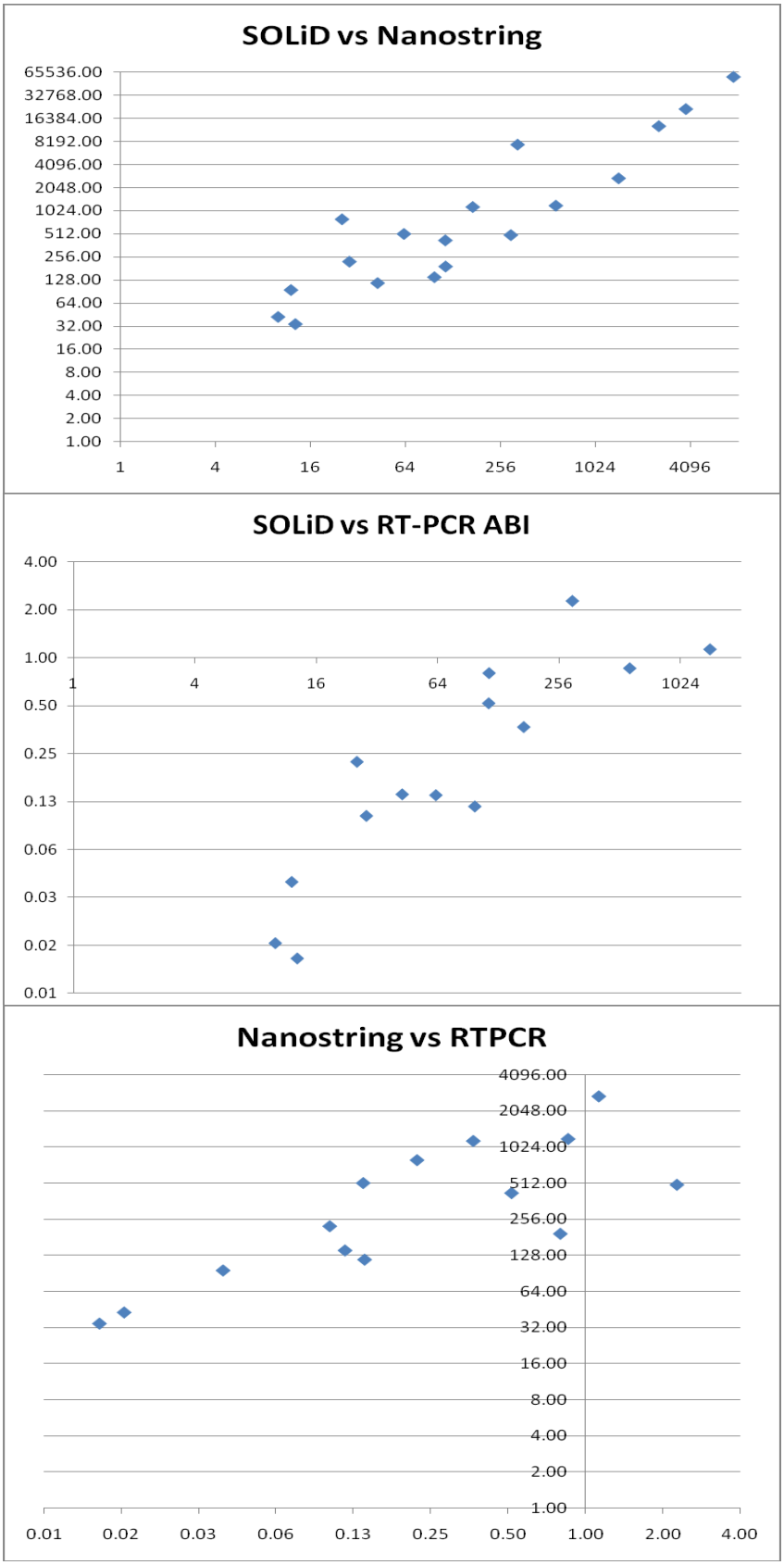

Figure 13. Measure correlations between different miRNA detection platforms

Table 1. Recurrent somatic mutations in oncogenes and tumor-suppressor genes in colorectal cancer

Gene	Type of mutation	Estimated frequency of alterations
Oncogenes		
KRAS	Point mutations (codons 12, 13, 61)	40% (>75% of mutations are at codon 12)
NRAS	Point mutations (codons 12, 13, 61)	<5%
PIK3CA	Point mutations activating kinase activity	15–25%
BRAF	Point mutations activating kinase activity (e.g., *V600E*)	5–10% (mutations linked to CIMP-positive CRCs)
EGFR	Gene amplification	5–15%
CDK8	Gene amplification	10–15%
CMYC	Gene amplification	5–10%
CCNE1	Gene amplification	5%
CTNNB1	Stabilizing point mutations and in-frame deletions near N terminus	<5%
NEU (HER2)	Gene amplification	<5%
MYB	Gene amplification	<5%
Tumor-suppressor genes		
p53	Point mutation, allele loss	60–70% (>95% of point mutations are missense)
APC	Frameshift, point mutation, deletion, allele loss	70–80% (nearly all mutations lead to truncated proteins)
FBXW7	Nonsense, missense, deletion	20%
PTEN	Nonsense, deletion	10%
SMAD4	Nonsense, missense, allele loss	10–15%
SMAD2	Nonsense, deletion, allele loss	5–10%
SMAD3	Nonsense, deletion	5%
TGFβIIR	Frameshift, nonsense	10–15% (>90% of MSI-H CRCs have mutations)
TCF7L2	Frameshift, nonsense	5% (mutations in both MSI-H and MSS CRCs)
ACVR2	Frameshift	10% (>80% of MSI-H CRCs have mutations)
BAX	Frameshift	5% (often one allele in ~50% of MSI-H CRCs)

Table 2. Short non-coding RNA contigs discriminate colon adenocarcinoma from normal colon.

Permutat. p-value	FDR	Geom mean of RPKMM in CRC	Geom mean of RPKMM in normal colon	Fold-change (in paired samples)	Unique id
0.0035	0.0141	2698.2	167.84	16.85	hsa-mir-135b|DM004220|Contig:chr1:205417451-205417511-|201|1q32.1
1e-04	0.00874	20109.25	1273.26	14.42	hsa-mir-31|LOC554202|Contig:chr9:21512118-21512178-|991|9p21.3
0.011	0.0476	265.21	39.27	6.95	NAV3|Contig:chr12:78591903-78591976+|135|12q21.2
0.0017	0.0165	35.96	5	6.66	DPYDas|Contig:chr1:98085664-98085897+|40|1p21.3
0.0012	0.0114	84.61	12.97	6.63	DPP10|Contig:chr2:116080149-116080315+|102|2q14.1
0.0012	0.0141	173.77	26.67	6.63	DNAJC11as|Contig:chr1:6704628-6704759+|226|1p36.31
0.0112	0.0396	161.37	27.08	5.66	hsa-mir-503|MGC16121|Contig:chrX:133680362-133680443-|83|Xq26.3
0.001	0.00874	46.14	9.01	5.44	NTM|Contig:chr11:131767619-131767837+|27|11q25
0.0069	0.0772	370.89	68.75	5.35	Contig:chr1:71163226-71163286+|37|1p31.1
0.0168	0.0427	60.89	10.09	5.22	Contig:chr8:91319432-91319523+|82|8q21.3
0.0122	0.0427	497.04	107.86	4.82	HBII-99|C20orf199|Contig:chr20:47897225-47897303+|740|20q13.13 (SNORD12)
6e-04	0.017	2086.07	528.36	4.71	hsa-mir-183|DM004436|Contig:chr7:129414766-129414828-|355|7q32.2
5e-04	0.00836	249530.61	53513.51	4.4	hsa-mir-21|DM119428|Contig:chr17:57918609-57918728+|2808|17q23.1
0.0036	0.0138	81.58	20.34	4.33	Contig:chr2:120046346-120046439+|45|2q14.2
6e-04	0.00874	4965.72	1321.4	4.23	hsa-mir-182|DQ595899|Contig:chr7:129410247-129410329-|387|7q32.2
0.0188	0.0481	107	26.85	4.05	Contig:chr4:145226185-145226296-|59|4q31.21
0.0033	0.0211	36.68	9.41	4.02	Contig:chr4:31160576-31160713+|49|4p15.1
0.0033	0.0332	41.86	11.06	4.01	Contig:chrX:125606345-125606856-|1238|Xq25
0.0226	0.055	1056.97	288.37	4	Contig:chr1:95751777-95751856+|154|1p21.3
0.0112	0.0476	509.75	144.92	3.87	GNL3|Contig:chr3:52722906-52722969+|161|3p21.1 SNORD19B.2-201
0.0328	0.0874	2712.94	821.82	3.78	hsa-mir-7-2|DL233857|Contig:chr15:89155087-

					89155148+|532|15q26.1
0.0021	0.0127	393.91	105.95	3.46	hsa-mir-542|no UCSC|Contig:chrX:133675393-133675453-|165|Xq26.3
0.0125	0.0358	757.3	238.96	3.38	Contig:chr4:149864853-149864937-|83|4q31.23
0.0505	0.0945	313.82	94.6	3.36	mgh28S-2411|TAF1D|Contig:chr11:93464671-93464739-|2013|11q21
0.0207	0.055	67.5	21.53	3.22	U48|C6orf48|Contig:chr6:31802953-31803094+|411|6p21.33 (SNORD48)
0.0034	0.0531	183.7	64.89	3.2	hsa-mir-301b|no UCSC|Contig:chr22:22007279-22007339+|125|22q11.21
0.0148	0.0439	135.98	43.19	3.09	Contig:chr12:67479721-67479859-|127|12q14.3
0.0102	0.0758	397.51	131.27	3.04	SNORD123|SNORD123|Contig:chr5:9548950-9549016+|157|5p15.31
0.044	0.0764	238.29	77.99	2.97	TNPO1as|Contig:chr5:72112291-72112420-|74|5q13.2
0.0419	0.074	80.41	27.99	2.9	AlphaTFEB|Contig:chr11:65273440-65273625+|153|11q13 .1 tRNA-like small RNA
0.007	0.0465	368.64	144.1	2.89	U78|GAS5|Contig:chr1:173834685-173834912-|1691|1q25.1 (SNORD 78)
0.0012	0.0114	341020.75	127487.84	2.82	hsa-mir-29a|CR618431as|Contig:chr7:130561503-130561567-|2760|7q32.3
0.0187	0.0681	775.13	264.34	2.8	RMRP|Contig:chr9:35657748-35658017-|1396|9p13.3
0.001	0.00874	4403.7	1668.82	2.6	hsa-mir-424|DM004809|Contig:chrX:133680672-133680749-|360|Xq26.3
0.023	0.0587	445.01	181.12	2.6	U54|RPS20|Contig:chr8:56986395-56986461-|1073|8q12.1 (*U54 small nucleolar RNA, C/D box 54)*
0.0047	0.0192	2722.5	1126.25	2.55	IGF2BP3as|Contig:chr7:23403166-23403283+|425|7p15.3
0.0015	0.0165	87.27	35.99	2.5	PKN2|Contig:chr1:89273813-89273899+|52|1p22.2
0.0366	0.0747	158.44	67.67	2.43	MBNL1|Contig:chr3:152171516-152171596+|53|3q25.2
0.0096	0.0465	351.36	145.23	2.4	EXOC4|Contig:chr7:133294036-133294151+|200|7q33
3e-04	0.00874	3486.17	1600.76	2.39	hsa-mir-452|GABRE|Contig:chrX:151128110-151128171-|251|Xq28
0.0535	0.0945	701.67	336.19	2.38	NEGR1|Contig:chr1:72028786-72028897-|238|1p31.1
0.0259	0.0626	102.45	49.6	2.3	SAMD13|Contig:chr1:84774303-

					84774485+|138|1p31.1
0.0462	0.0851	5293.22	2374.46	2.24	COL25A1as|Contig:chr4:109814955-109815022+|541|4q25
0.0304	0.0764	1813.87	909.69	2.21	RIMS1|Contig:chr6:73094787-73094904+|453|6q13
0.0379	0.0758	857.67	419.48	2.16	MAPK10as|Contig:chr4:87139084-87139183+|304|4q21.3
0.0073	0.0476	96.72	48.37	2.13	OC90|Contig:chr8:133070320-133070492-|189|8q24.22
0.0361	0.0764	1193.81	603.93	2.08	ZNF639|Contig:chr3:179041259-179041338+|201|3q26.33
0.0349	0.0711	670.69	352.46	2.05	C14orf106as|Contig:chr14:45707338-45707417+|125|14q21.2
0.0021	0.0116	274428.61	114525.36	2.01	hsa-mir-223|mir-223|Contig:chrX:65238737-65238805+|3057|Xq12
0.0429	0.0874	1086.09	569.25	1.99	C1GALT1|Contig:chr7:7222241-7222309+|338|7p22.1
0.0228	0.0531	567.14	307.86	1.92	MAML2as|Contig:chr11:95805368-95805452+|109|11q21
0.0215	0.0614	333.57	183.66	1.87	Contig:chr5:102929579-102929698+|198|5q21.2
0.0456	0.0874	18108.22	10626.54	1.77	hsa-mir-210|AK123483|Contig:chr11:568109-568191-|1025|11p15.5
0.0035	0.0193	26820.93	15978.52	1.72	hsa-mir-25|MCM7|Contig:chr7:99691191-99691460-|2652|7q22.1
0.0128	0.0476	1495.81	907.99	1.67	hsa-mir-18b|no UCSC|Contig:chrX:133304072-133304136-|207|Xq26.2
0.0074	0.0531	50602.77	32689.47	1.65	hsa-mir-17|C13orf25 v_1|Contig:chr13:92002871-92003640+|10892|13q31.3
0.0293	0.0632	11044.19	7602.76	1.49	hsa-mir-455|COL27A1|Contig:chr9:116971728-116971789+|680|9q32
0.0132	0.0374	43915.07	31734.6	1.48	hsa-mir-34a|EF609116|Contig:chr1:9211749-9211817-|1268|1p36.22
0.0347	0.076	210.81	146.56	1.46	CNTN5|Contig:chr11:99340489-99340585+|116|11q22.1
0.0206	0.0547	21903.25	31146.21	0.75	hsa-mir-30d|DL231865|Contig:chr8:135817121-135817187-|1225|8q24.22
0.0196	0.0531	33454.87	49179.49	0.71	hsa-let-7g|WDR82|Contig:chr3:52302294-52302375-|2126|3p21.1
0.0035	0.0531	66171.79	98154.65	0.7	hsa-mir-29c|EU154352|Contig:chr1:207975207-207975289-|2633|1q32.2
0.0236	0.0592	11203.91	16475.98	0.7	hsa-mir-27b|C9orf3|Contig:chr9:97847717-

					97847811+|865|9q22.32
0.0185	0.0513	31263.99	44951.36	0.69	hsa-mir-26a-2|CTDSP2|Contig:chr12:58218401-58218509-|1382|12q14.1
0.02	0.0547	12261.58	17633.12	0.69	hsa-mir-30e|NFYC|Contig:chr1:41220042-41220109+|1983|1p34.2
0.0288	0.0675	21684.21	31899.16	0.69	hsa-mir-101-1|DM004381|Contig:chr1:65524121-65524202-|1526|1p31.3
0.0313	0.0697	29642.04	43478.06	0.69	hsa-mir-101-2|RCL1|Contig:chr9:4850309-4850369+|1286|9p24.1
0.0043	0.00958	7429.98	11020.66	0.67	THRB|Contig:chr3:24318273-24318349-|482|3p24.2
0.0189	0.0546	48447.78	72786.57	0.67	hsa-mir-574|FAM114A1|Contig:chr4:38869677-38869739+|1397|4p14
0.0228	0.055	566.19	831.26	0.67	hsa-mir-766|SEPT6|Contig:chrX:118780723-118780803-|694|Xq24
0.0231	0.0589	48359.11	76764.62	0.66	hsa-mir-126|EGFL7|Contig:chr9:139565050-139565129+|1268|9q34.3
0.0029	0.0127	51540.85	78901.72	0.65	hsa-mir-140|WWP2|Contig:chr16:69967005-69967071+|3293|16q22.1
0.0053	0.0203	660.77	1048.59	0.64	Contig:chr11:128169158-128169268+|69|11q24.3
8e-04	0.00874	18155.8	35920.96	0.54	hsa-mir-30b|DM004172|Contig:chr8:135812756-135812856-|1185|8q24.22
0.0179	0.0531	173.02	329.15	0.54	CDKAL1|Contig:chr6:20962410-20962533+|339|6p22.3
0.0059	0.0192	11665.82	22794.29	0.53	CPA6|Contig:chr8:68497631-68497751-|9312|8q13.2 piRNA piR-51810
0.0012	0.00874	12852.72	25320.18	0.52	hsa-mir-30c-1|NFYC|Contig:chr1:41222948-41223032+|1565|1p34.2
5e-04	0.00874	4842.86	9787.8	0.5	hsa-mir-30a|DM119476|Contig:chr6:72113256-72113338-|495|6q13
0.0165	0.0531	1754.78	3675.42	0.5	Contig:chr1:233917132-233917242+|201|1q42.2
0.0147	0.0427	133673.79	264827.3	0.49	hsa-mir-143|LOC728264|Contig:chr5:148808487-148808565+|973|5q32
6e-04	0.00874	8590.34	18134.65	0.48	hsa-mir-30c-2|DM004170|Contig:chr6:72086666-72086748-|1274|6q13
0.0059	0.0203	1021.62	2117.51	0.48	Contig:chr2:101260946-101261043-|126|2q11.
4e-04	0.00874	484.84	1006.14	0.47	IGSF21|Contig:chr1:18627390-18627531+|188|1p36.13
0.0368	0.0745	128.36	280.09	0.46	Contig:chr3:117186707-117186828+|99|3q13.31
0.0026	0.0141	592371.61	1341828.71	0.43	hsa-mir-145|LOC728264|Contig:chr5:148810222-148810286+|1980|5q32

0.0063	0.0203	1235.79	2731.32	**0.43**	**Contig:chr2:55717350-55717439+|81|2p16.**
0.0245	0.0758	18971.83	62582.06	**0.41**	**hsa-mir-215|IARS2as|Contig:chr1:220291218-220291278-|543|1q41**
0.0011	0.0114	24.78	57.74	**0.4**	**LENG8|Contig:chr19:54969382-54969562+|69|19q13.42**
0.0017	0.0114	167.43	368.94	**0.4**	**NEBLas|Contig:chr10:21338351-21338478+|204|10p12.31**
7e-04	0.0192	723.04	2264.05	**0.4**	**SLC1A7as|Contig:chr1:53561108-53561233+|116|1p32.3**
0.0321	0.071	2567.62	6095.22	**0.4**	**hsa-mir-3065as|AATK|Contig:chr17:79099685-79099746-|513|17q25.3**
2e-04	0.00874	5176.84	13599.17	**0.39**	**hsa-mir-10b|DM004331|Contig:chr2:177015057-177015121+|1195|2q31.1**
0.0032	0.017	806.48	1977.92	**0.39**	**AK311257as|Contig:chr8:142405380-142405453-|108|8q24.3**
0.0026	0.0138	6542.9	16867.98	**0.36**	**PDGFC|Contig:chr4:157834741-157834874-|243|4q32.1**
3e-04	0.0476	362.9	968.71	**0.36**	**hsa-mir-511-1|MRC1|Contig:chr10:17887121-17887181+|154|10p12.33**
3e-04	0.0476	362.9	968.71	**0.36**	**hsa-mir-511-2|MRC1|Contig:chr10:18134050-18134110+|154|10p12.33**
0.01	0.0587	144.97	388.46	**0.34**	**CTNNA3|Contig:chr10:68504527-68504640-|152|10q21.3**
0.0066	0.0657	77.19	238.65	**0.34**	**AF086303|Contig:chr6:74832055-74832154+|39|6q13**
0.0012	0.0268	43.73	120.91	**0.33**	**GRSF1as|Contig:chr4:71702781-71702900+|87|4q13.3**
0.0507	0.0945	228.4	666.75	**0.33**	**CALD1|Contig:chr7:134630474-134630553+|104|7q33**
0.0373	0.0758	21.86	63.34	**0.32**	**C14orf25|Contig:chr14:38091955-38092100+|23|14q21.1**
0.0061	0.0975	3957.24	12273.91	**0.32**	**hsa-mir-195|DM004261|Contig:chr17:6920946-6921007-|611|17p13.1**
0.0028	0.095	1786.95	5501.52	**0.31**	**hsa-mir-218-1|SLIT2|Contig:chr4:20529922-20529983+|308|4p15.31**
0.0054	0.0681	320.95	1050.77	**0.3**	**hsa-mir-149|GPC1|Contig:chr2:241395374-241395475+|226|2q37.3**
0.0376	0.0711	132.83	429.73	**0.3**	**CCBL2as|Contig:chr1:89454772-89454870+|124|1p22.2**
0.0317	0.0711	36.64	124.12	**0.3**	**KCNQ3as|Contig:chr8:133486556-**

					133486687+|70|8q24.22
0.0051	0.0427	21.28	77.71	**0.28**	**AKAP6as|Contig:chr14:32953303-32954323-|3013|14q12**
0.0045	0.0587	410.29	1440.61	**0.27**	**Contig:chr3:141176313-141176374-|39|3q23**
0.0044	0.0614	863.68	3271.92	**0.27**	**hsa-mir-138-2|DM004413|Contig:chr16:56892439-56892507+|331|16q13**
3e-04	0.0114	116.08	502.99	**0.26**	**Contig:chrX:152237951-152238082+|107|Xq28**
3e-04	0.0114	116.08	502.99	**0.26**	**Contig:chrX:152335414-152335545+|107|Xq28**
3e-04	0.0114	116.08	502.99	**0.26**	**Contig:chrX:152346176-152346307-|107|Xq28**
3e-04	0.0192	93.43	359.79	**0.26**	**ACA54|NAP1L4|Contig:chr11:2985007-2985124-|1224|11p15.4 (H/ACA Box snoRNA)**
0.0028	0.0531	736.48	2911.79	**0.25**	**hsa-mir-138-1|DM004414|Contig:chr3:44155704-44155787+|409|3p21.32**
0.0012	0.0114	9.37	36.8	**0.23**	**CLK1|Contig:chr2:201728724-201728865-|40|2q33.1**
0.0119	0.0476	25.3	101.55	**0.22**	**MYT1L|Contig:chr2:1972719-1972780-|73|2p25.3**
0.0045	0.0233	10.3	53.22	**0.2**	**KALRNas|Contig:chr3:124068655-124068789-|20|3q21.2**
0.0012	0.0138	62.93	358.71	**0.18**	**NUBPLas|Contig:chr14:32313881-32313962-|32|14q12**
0.0103	0.033	76.62	439.66	**0.15**	**Contig:chr10:69524258-69524366+|629|10q21.3 Possible tRNA**
4e-04	0.00874	23.61	158.02	**0.14**	**hsa-mir-551b|no UCSC|Contig:chr3:168269580-168269724+|233|3q26.2**
4e-04	0.00874	38.77	334.94	**0.14**	**hsa-mir-9-3|LOC254559|Contig:chr15:89911263-89911324+|131|15q26.1**
4e-04	0.0203	94.79	610.21	**0.14**	**hsa-mir-204|TRPM3|Contig:chr9:73424848-73424968-|204|9q21.12**
0.0074	0.0306	670.58	4354.08	**0.14**	**hsa-mir-1-1|C20orf166|Contig:chr20:61151518-61151581+|178|20q13.33**
2e-04	0.00874	26.47	230.09	**0.13**	**hsa-mir-129-2|DM004400|Contig:chr11:43602958-43603022+|91|11p11.2**
4e-04	0.00874	40.32	370.3	**0.13**	**hsa-mir-9-1|C1orf61|Contig:chr1:156390144-156390206-|141|1q22**
4e-04	0.00874	42.59	384.46	**0.13**	**hsa-mir-9-2|LOC645323|Contig:chr5:87962682-87962742-|143|5q14.3**
6e-04	0.0124	58.2	399	**0.13**	**Contig:chr4:160867890-160867958+|71|4q32.1**
< 1e-07	0.00874	103.83	1502.1	**0.087**	**hsa-mir-147b|C15orf48|Contig:chr15:45725295-45725364+|144|15q21.1**

0.0025	0.0114	8.37	100.42	**0.087**	**hsa-mir-129-1|DM004399|Contig:chr7:127847929-127847995+|36|7q32.1**

Table 3. Differentially expressed miRNAs in colon adenocarcinoma, as determined by the small RNA pipeline (SRP) and PASS.

p-value (SRP)	FDR (SRP)	Geom mean of RPMM in colon adeno-carcinoma (SRP)	Geom mean of RPMM in normal colon (SRP)	Fold-change (SRP)	Fold Change (PASS)	Chromosomal coordinates	miRNA
8.7e-06	0.000892	362.31	18.7	**19.37**	18.75	hsa-mir-31@9:21512156-21512177 22(-)ncl:1	**hsa-mir-31**
6.53e-05	0.00279	866.85	79.13	**10.95**	11.45	hsa-mir-31@9:21512157-21512177 21(-)mature	**hsa-miR-31**
0.0081944	0.0531	1903.77	175.2	**10.87**	8.69	hsa-mir-21@17:57918634-57918656 23(+)ncl:1	**hsa-mir-21**
0.0028611	0.0287	23622.22	2839.43	**8.32**	4.39	hsa-mir-29a@7:130561507-130561528 22(-)mature	**hsa-miR-29a**
0.0002439	0.00708	140.08	17.82	**7.86**	7.67	hsa-mir-135b@1:205417489-205417511 23(-)mature	**hsa-miR-135b**
2.36e-05	0.00173	394.61	74.04	**5.33**	5.33	hsa-mir-224@X:151127102-151127123 22(-)ncl:1	**hsa-mir-224**
1.03e-05	0.000968	85.35	18.39	**4.64**	hsa-mir-222@X:45606443-45606462 20(-) ncl:-1 FC= 2.44	hsa-mir-222@X:45606440-45606462 23(-)ncl:2	**hsa-mir-222**
2e-06	0.000342	43.84	10.5	**4.17**	4.48	hsa-mir-224@X:151127103-151127123 21(-)mature	**hsa-miR-224**
0.0017424	0.0225	349.81	100.72	**3.47**	3.39	hsa-mir-182@7:129410287-129410310 24(-	**hsa-miR-182**

						)mature	
0.001873	0.0227	163.08	56.95	**2.86**	2.26	hsa-mir-199a-1@19:10928105-10928127 23(-)	**hsa-mir-199a-1**
0.002398	0.0269	55.07	19.41	**2.84**	2.31	hsa-mir-183@7:129414806-129414827 22(-)	**hsa-mir-183**
5.34e-05	0.00259	3394.71	1283	**2.65**	2.48	hsa-mir-21@17:57918635-57918656 22(+)	**hsa-mir-21**
7.34e-05	0.00289	751.43	290.27	**2.59**	2.2	hsa-mir-25@7:99691194-99691215 22(-) mature	**hsa-miR-25**
0.001496	0.021	13137.08	5244.31	**2.51**	2.45	hsa-mir-223@X:65238779-65238800 22(+)mature	**hsa-miR-223**
0.0009978	0.0163	8333.54	3411.05	**2.44**	2.48	hsa-mir-223@X:65238779-65238801 23(+)ncl:1	**hsa-mir-223**
0.0066569	0.047	46.16	19.36	**2.38**	2.6	hsa-mir-1247@14:102026699-102026720 22(-)mature	**hsa-miR-1247**
0.0017122	0.0223	3676.75	1542.77	**2.38**	2.11	hsa-mir-17@13:92002872-92002892 21(+)ncl:-2	**hsa-mir-17**
0.0002082	0.00658	1633.86	698.22	**2.34**	2.16	hsa-mir-20a@13:92003326-92003346 21(+)ncl:-2	**hsa-mir-20a**
0.0008038	0.0138	100.42	42.85	**2.34**	2.04	hsa-mir-301a@17:57228509-57228532 24(-)ncl:1	**hsa-mir-301a**
0.0030172	0.0288	236.34	102.7	**2.3**	2.38	hsa-mir-424@X:133680711-133680731 21(-)ncl:-1	**hsa-mir-424**
0.0003488	0.00838	1535.96	702.85	**2.19**	2.35	hsa-mir-18a@13:92003010-92003032 23(+)mature	**hsa-miR-18a**
0.003218	0.0302	13539.22	6614.35	**2.05**	2.0	hsa-mir-	**hsa-**

						17@13:92002872-92002894 23(+)mature	miR-17
0.0051651	0.042	36.02	18.41	1.96	hsa-mir-455@9:116971768-116971787 20(+)ncl:-2 FC 2.12	hsa-mir-455@9:116971766-116971787 22(+)	hsa-mir-455
0.0014047	0.0202	28.14	14.47	1.94	1.36	hsa-mir-487a@14:101518831-101518852 22(+)mature	hsa-miR-487a
2.33e-05	0.00173	5635.89	3003.41	1.88	2.01 (substantial difference in RPMM)	hsa-mir-23a@19:13947407-13947429 23(-)ncl:2	hsa-mir-23a
0.0096939	0.0587	184.73	100.88	1.83	1.79	hsa-mir-452@X:151128150-151128171 22(-)mature	hsa-miR-452
0.0065205	0.0464	16.96	10	1.7	hsa-mir-552@1:35135216-35135235 20(-)ncl:-1 FC 2.03	hsa-mir-552@1:35135215-35135236 22(-)ncl:1	hsa-mir-552
0.0042764	0.0361	62.5	37.85	1.65	hsa-mir-494@14:101496018-101496039 22(+)mature FC 1.54	hsa-mir-494@14:101496018-101496038 21(+)ncl:-1	hsa-mir-494
0.0096361	0.0587	236.89	152.99	1.55	hsa-mir-92a-1@13:92003615-92003636 22(+)mature FC 1.63	hsa-mir-92a-1@13:92003616-92003636 21(+)ncl:-1	hsa-mir-92a-1
0.003562	0.0326	14.01	10	1.4	1.88	hsa-mir-106b@7:99691628-99691646 19(-)ncl:-2	hsa-mir-106b
0.0090538	0.0571	235.74	336.12	0.7	0.68	hsa-mir-28@3:188406582-188406602 21(+)ncl:-1	hsa-mir-28
0.0035127	0.0324	398.7	609.36	0.65	0.62	hsa-mir-29c@1:207975210-207975230 21(-)ncl:-1	hsa-mir-29c
0.0051476	0.042	4113.68	6470.64	0.64	0.68	hsa-mir-23b@9:97847547-97847567 21(+) mature	hsa-miR-23b

0.00937	0.0582	268.1	432.4	**0.62**	0.65	hsa-mir-30d@8:135817163-135817183 21(-)ncl:-1	**hsa-mir-30d**
0.0042674	0.0361	293.22	533.37	**0.55**	0.52	hsa-mir-140@16:69967045-69967066 22(+)	**hsa-mir-140**
0.0026118	0.0276	12.54	23.52	**0.53**	hsa-mir-143@5:148808541-148808561 21(+) mature FC 0.45	hsa-mir-143@5:148808540-148808558 19(+)ncl:-2	**hsa-mir-143**
0.0007865	0.0138	1776.19	3447.62	**0.52**	0.50	hsa-mir-30b@8:135812813-135812834 22(-)mature	**hsa-miR-30b**
0.0007886	0.0138	14.56	28.55	**0.51**	0.53	hsa-mir-138-2@16:56892439-56892459 21(+)ncl:-2	**hsa-mir-138-2**
0.000338	0.00838	34866.92	68651.32	**0.51**	0.47	hsa-mir-145@5:148810224-148810246 23(+)mature	**hsa-miR-145**
0.006862	0.048	529.5	1077.2	**0.49**	0.49	hsa-mir-30c-1@1:41222972-41222994 23(+)mature	**hsa-miR-30c**
0.0026229	0.0276	21.39	45.69	**0.47**	0.56	hsa-mir-190@15:63116206-63116227 22(+)	**hsa-mir-190**
2.5e-06	0.000385	5073.82	11329.89	**0.45**	0.46	hsa-mir-125a@19:52196521-52196542 22(+)ncl:-2	**hsa-mir-125a**
0.0003146	0.00806	584.97	1336.92	**0.44**	0.58	hsa-mir-26a-1@3:38010904-38010922 19(+)ncl:-3	**hsa-mir-26a-1**
0.0006179	0.0121	66.22	151.6	**0.44**	0.48	hsa-mir-28@3:188406582-188406601 20(+)ncl:-2	**hsa-mir-28**
0.0016075	0.0217	18.8	45.8	**0.41**	0.37	hsa-mir-218-1@4:20529922-20529942 21(+)mature	**hsa-miR-218**
0.0018569	0.0227	18.17	43.82	**0.41**	0.49	hsa-mir-221@X:45605651-	**hsa-mir-**

						45605670 20(-)ncl:-3	**221**
0.0009366	0.0155	19.2	49.1	**0.39**	0.43	hsa-mir-149@2:241395432-241395454 23(+)mature	**hsa-miR-149**
0.0030133	0.0288	40.4	104.46	**0.39**	0.41	hsa-mir-29b-1@7:130562229-130562249 21(-)ncl:-2	**hsa-mir-29b-1**
0.0006389	0.0121	1060.9	2738.76	**0.39**	0.33	hsa-mir-378@5:149112430-149112450 21(+)mature	**hsa-miR-378**
5.39e-05	0.00259	174.5	463.98	**0.38**	0.31	hsa-mir-378c@10:132760901-132760921 21(-)ncl:-4	**hsa-mir-378c**
0.0002963	0.00786	262.92	702.03	**0.37**	0.36	hsa-mir-10b@2:177015057-177015078 22(+)ncl:-1	**hsa-mir-10b**
9.07e-05	0.00332	78.23	216.81	**0.36**	0.38	hsa-mir-30a@6:72113299-72113319 21(-)ncl:-1	**hsa-mir-30a**
0.0007891	0.0138	566.16	1575.27	**0.36**	0.35	hsa-mir-497@17:6921298-6921318 21(-)mature	**hsa-miR-497**
8.87e-05	0.00332	11417.82	32609.4	**0.35**	0.44	hsa-mir-145@5:148810224-148810245 22(+)ncl:-1	**hsa-mir-145**
0.0013086	0.0192	100.82	290.54	**0.35**	0.34	hsa-mir-192@11:64658632-64658654 23(-)ncl:2	**hsa-mir-192**
0.0023399	0.0266	437.33	1309.05	**0.33**	0.33	hsa-mir-150@19:50004089-50004110 22(-)mature	**hsa-miR-150**
4.5e-06	0.000628	11.49	36.96	**0.31**	0.35	hsa-mir-147b@15:45725296-45725318 23(+)ncl:1	**hsa-mir-147b**
1.3e-06	0.00025	2431.76	7891.04	**0.31**	0.31	hsa-mir-378@5:149112430-149112451 22(+)ncl:1	**hsa-mir-378**
4.13e-05	0.0023	22.8	80.87	**0.28**	0.34	hsa-mir-138-1@3:44155726-44155749 24(+)ncl:1	**hsa-mir-138-1**

0.0001377	0.00471	16.74	60.53	**0.28**	0.37	hsa-mir-204@9:73424947-73424968 22(-)mature	**hsa-miR-204**
0.0007098	0.0132	53.07	203.44	**0.26**	0.3	hsa-mir-338@17:79099687-79099708 22(-)	**hsa-mir-338**
6.47e-05	0.00279	58.2	231.64	**0.25**	0.27	hsa-mir-139@11:72326147-72326168 22(-)mature	**hsa-miR-139-5p**
0.0053629	0.0431	325.54	2100.76	**0.15**	0.16	hsa-mir-215@1:220291257-220291278 22(-)ncl:1	**hsa-mir-215**
0.0025041	0.0273	154.67	1047.92	**0.15**	0.17	hsa-mir-215@1:220291258-220291278 21(-)mature	**hsa-miR-215**

Table 4. miRNA annotation discrepancies with miRBase.

miRNA (- indicates antisense)	**Discrepancy**
hsa-let-7c	star form with wrong coordinates
hsa-mir-100	not detectable star form
hsa-mir-103-1	miRNA with un-annotated star form
hsa-mir-103-1-as-	miRNA with un-annotated star form
hsa-mir-103-2-as-	miRNA with un-annotated star form
hsa-mir-107	miRNA with un-annotated star form
hsa-mir-1-1	miRNA with un-annotated star form
hsa-mir-1185-1	wrong mature coordinates
hsa-mir-1201	major/star forms swap
hsa-mir-1234	wrong mature coordinates
hsa-mir-1248	aberrant precursor processing
hsa-mir-1259	aberrant precursor processing
hsa-mir-125a	miRNA with un-annotated star form
hsa-mir-127	miRNA with un-annotated star form
hsa-mir-1271	miRNA with un-annotated star form
hsa-mir-1273	aberrant precursor processing
hsa-mir-1273-	aberrant precursor processing
hsa-mir-1273d	aberrant precursor processing
hsa-mir-1273d-	aberrant precursor processing
hsa-mir-1285-1	aberrant precursor processing
hsa-mir-1285-1-	aberrant precursor processing
hsa-mir-129-1	major/star forms swap
hsa-mir-129-2	major/star forms swap
hsa-mir-1303	major/star forms swap
hsa-mir-1306	major/star forms swap
hsa-mir-1307	miRNA with un-annotated star form
hsa-mir-130a	star form with wrong coordinates
hsa-mir-133a-1	miRNA with un-annotated star form
hsa-mir-133a-2	miRNA with un-annotated star form
hsa-mir-134	miRNA with un-annotated star form
hsa-mir-138-2	not detectable star form
hsa-mir-140	major/star forms swap
hsa-mir-142	miRNA with un-annotated star form

hsa-mir-146b	miRNA with un-annotated star form
hsa-mir-147b	miRNA with un-annotated star form
hsa-mir-149	not detectable star form
hsa-mir-151	miRNA with un-annotated star form
hsa-mir-152	miRNA with un-annotated star form
hsa-mir-154	major/star forms swap
hsa-mir-181b-1	miRNA with un-annotated star form
hsa-mir-181b-2	miRNA with un-annotated star form
hsa-mir-1826	miRNA with un-annotated star form
hsa-mir-188	miRNA with un-annotated star form
hsa-mir-190	miRNA with un-annotated star form
hsa-mir-193a	major/star forms swap
hsa-mir-194-1	miRNA with un-annotated star form
hsa-mir-196a-1	miRNA with un-annotated star form
hsa-mir-196a-2	not detectable star form
hsa-mir-1972-1	wrong strand of precursor
hsa-mir-1972-1-	wrong strand of precursor
hsa-mir-1972-2	wrong strand of precursor
hsa-mir-1972-2-	wrong strand of precursor
hsa-mir-1975	aberrant precursor processing
hsa-mir-1979	aberrant precursor processing
hsa-mir-199a-1	miRNA with un-annotated star form
hsa-mir-199a-2	miRNA with un-annotated star form
hsa-mir-199b	miRNA with un-annotated star form
hsa-mir-203	miRNA with un-annotated star form
hsa-mir-210	miRNA with un-annotated star form
hsa-mir-2110	miRNA with un-annotated star form
hsa-mir-212	miRNA with un-annotated star form
hsa-mir-215	miRNA with un-annotated star form
hsa-mir-2355	miRNA with un-annotated star form
hsa-mir-28	miRNA with un-annotated star form
hsa-mir-296	miRNA with un-annotated star form
hsa-mir-299	miRNA with un-annotated star form
hsa-mir-301a	miRNA with un-annotated star form
hsa-mir-3065	wrong strand of precursor
hsa-mir-3065-	wrong strand of precursor
hsa-mir-3074	wrong strand of precursor

hsa-mir-3074-	wrong strand of precursor
hsa-mir-3120	wrong strand of precursor
hsa-mir-3120-	wrong strand of precursor
hsa-mir-3130-1	aberrant precursor processing
hsa-mir-3130-1-	aberrant precursor processing
hsa-mir-3130-2	aberrant precursor processing
hsa-mir-3130-2-	aberrant precursor processing
hsa-mir-3130-3	aberrant precursor processing
hsa-mir-3130-3-	aberrant precursor processing
hsa-mir-3145	wrong mature coordinates
hsa-mir-3159	aberrant precursor processing
hsa-mir-3159-	aberrant precursor processing
hsa-mir-3184-	wrong strand of precursor
hsa-mir-323	wrong mature coordinates
hsa-mir-323b	wrong mature coordinates
hsa-mir-324	major/star forms swap
hsa-mir-330	major/star forms swap
hsa-mir-331	major/star forms swap
hsa-mir-337	miRNA with un-annotated star form
hsa-mir-339	miRNA with un-annotated star form
hsa-mir-33a	major/star forms swap
hsa-mir-342	major/star forms swap
hsa-mir-34b	major/star forms swap
hsa-mir-361	miRNA with un-annotated star form
hsa-mir-362	miRNA with un-annotated star form
hsa-mir-365-2	not detectable star form
hsa-mir-369	miRNA with un-annotated star form
hsa-mir-375	miRNA with un-annotated star form
hsa-mir-376b	miRNA with un-annotated star form
hsa-mir-376c	miRNA with un-annotated star form
hsa-mir-381	miRNA with un-annotated star form
hsa-mir-382	miRNA with un-annotated star form
hsa-mir-409	major/star forms swap
hsa-mir-423	major/star forms swap
hsa-mir-4284	aberrant precursor processing
hsa-mir-4286	aberrant precursor processing
hsa-mir-4297-	wrong strand of precursor

hsa-mir-452	star form with wrong coordinates
hsa-mir-454	major/star forms swap
hsa-mir-455	major/star forms swap
hsa-mir-483	major/star forms swap
hsa-mir-485	miRNA with un-annotated star form
hsa-mir-486	aberrant precursor processing
hsa-mir-486-	aberrant precursor processing
hsa-mir-490	major/star forms swap
hsa-mir-500	major/star forms swap
hsa-mir-501	miRNA with un-annotated star form
hsa-mir-502	major/star forms swap
hsa-mir-511-1	major/star forms swap
hsa-mir-511-2	major/star forms swap
hsa-mir-532	miRNA with un-annotated star form
hsa-mir-542	major/star forms swap
hsa-mir-548d-1	wrong mature coordinates
hsa-mir-548d-1	wrong strand of precursor
hsa-mir-548d-1-	wrong strand of precursor
hsa-mir-548d-2	wrong mature coordinates
hsa-mir-548d-2-	wrong mature coordinates
hsa-mir-548h-2	wrong mature coordinates
hsa-mir-548q	wrong strand of precursor
hsa-mir-548q-	wrong mature coordinates
hsa-mir-548q-	wrong strand of precursor
hsa-mir-548t	wrong mature coordinates
hsa-mir-548t-	wrong mature coordinates
hsa-mir-548v	major/star forms swap
hsa-mir-550-1	major/star forms swap
hsa-mir-550-2	major/star forms swap
hsa-mir-552	miRNA with un-annotated star form
hsa-mir-558	wrong mature coordinates
hsa-mir-558-	wrong mature coordinates
hsa-mir-566	wrong mature coordinates
hsa-mir-566-	wrong mature coordinates
hsa-mir-574	major/star forms swap
hsa-mir-574-	wrong mature coordinates
hsa-mir-582	miRNA with un-annotated star form

hsa-mir-582-	wrong mature coordinates
hsa-mir-590	miRNA with un-annotated star form
hsa-mir-616	major/star forms swap
hsa-mir-619	aberrant precursor processing
hsa-mir-619-	aberrant precursor processing
hsa-mir-625	major/star forms swap
hsa-mir-625-	major/star forms swap
hsa-mir-642	miRNA with un-annotated star form
hsa-mir-652	miRNA with un-annotated star form
hsa-mir-654	major/star forms swap
hsa-mir-660	miRNA with un-annotated star form
hsa-mir-671	miRNA with un-annotated star form
hsa-mir-769	miRNA with un-annotated star form
hsa-mir-874	miRNA with un-annotated star form
hsa-mir-886	aberrant precursor processing
hsa-mir-92a-2	not detectable star form
hsa-mir-935	wrong mature coordinates
hsa-mir-941-1	aberrant precursor processing
hsa-mir-941-2	aberrant precursor processing
hsa-mir-941-3	aberrant precursor processing
hsa-mir-942	wrong mature coordinates
hsa-mir-98	miRNA with un-annotated star form

Table 5. The performance of colon adenocarcinoma classification using the short non-coding RNA signature

Pair ID	Mean Number of genes in classifier	Diagonal Linear Discriminant Analysis Correct?	1-Nearest Neighbor	3-Nearest Neighbors Correct?	Nearest Centroid Correct?
CRC16500	121	YES	YES	YES	YES
CRC580	112	YES	YES	YES	YES
CRC617	126	YES	YES	YES	YES
CRC618	144	NO	YES	YES	YES
CRC622	114	YES	YES	YES	YES
CRC680	115	YES	YES	YES	YES
CRC740	118	YES	YES	YES	YES
CRC747	113	YES	YES	YES	YES
CRC758	109	YES	YES	YES	YES
CRC766	124	YES	YES	YES	YES
CRC774	109	YES	YES	YES	YES
CRC800	116	YES	YES	YES	YES
CRC803	116	YES	YES	YES	YES
Mean percent of correct classification		**92**	**100**	**100**	**100**

Table 6. Colon adenocarcinoma and normal colon miRNA quantification using different detection techniques (average values)

miRNA	SOLiD	RT-PCR (Stem-loop)	Microarrays (Adenocarcinoma vs. Benign Adenoma)	RT-PCR (LNA)	Northern Blot	NanoString
miR-29a Control	328.17	2.46	147.04	6.16	4866397.79	7491.28
miR-29a Cancer	2574.11	2.01	526.25	13.45	9790703.92	13026.55
miR-31 Control	12.85	0.01	Ns			34.49
miR-31 Cancer	97.62	0.12	Ns			140.73
miR-135b Control	10.00	0.02	224.61			42.74
miR-135b Cancer	25.4	0.22	392.36			798.77
miR-223 Control	572.94	0.86	250.21			1200.20
miR-223 Cancer	1432.18	1.13	2137.13			2724.59
miR-224 Control	12.06	0.04	65.26			95.77
miR-224 Cancer	42.62	0.14	142.17			117.97
miR-497 Control	170.68	0.37				1154.83
miR-497 Cancer	62.61	0.14				514.62
miR-148a Control	102.88	0.29	Ns			1368.73
miR-148a Cancer	38.67	0.27	Ns			2843.71
miR-215 Control	114.44	0.52			851396.62	424.10
miR-215 Cancer	28.28	0.10			651444.00	224.35
miR-378 Control	297.47	2.27				497.43
miR-378 Cancer	114.89	0.80				193.85
miR-145 Control	7640.96		830.76		10165254.98	57050.29
miR-145 Cancer	3815.46		1765.8		3911435.14	21756.85

Table 7. Correspondence between over-expressed non-coding RNAs and amplification by CGH in cancer

Short RNA contig	Cases of Amplif.	Cases of Deletion	Fold A/D	Comments
mir-135b |Contig:chr1:205417451-205417511-|201|1q32.1	4407	551	8.0	miRNA
Contig:chr8:91319432-91319523+|82|8q21.3	4407	826	5.3	
HBII-99 |Contig:chr20:47897225-47897303+|740|20q13.13	3581	551	6.5	SNORD12
mir-183 | Contig:chr7:129414766-129414828-|355|7q32.2	3856	826	4.7	miRNA
mir-21 | Contig:chr17:57918609-57918728+|2808|17q23.1	3856	1102	3.5	miRNA
mir-182 | Contig:chr7:129410247-129410329-|387|7q32.2	3856	826	4.7	miRNA
Contig:chr12:67479721-67479859-|127|12q14.3	2204	551	4	
SNORD123|SNORD123|Contig:chr5:9548950-9549016+|157|5p15.31	2755	551	5	SNORD123
U78|GAS5|Contig:chr1:173834685-173834912-|1691|1q25.1	4407	275	16	SNORD78
hsa-mir-29a|Contig:chr7:130561503-130561567-|2760|7q32.3	3856	826	4.7	miRNA
U54|RPS20|Contig:chr8:56986395-56986461-|1073|8q12.1	3738	850	4.4	SNORD54
IGF2BP3as|Contig:chr7:23403166-23403283+|425|7p15.3	4097	623	6.6	overlap on Human EST AV7298
MBNL1|Contig:chr3:152171516-152171596+|53|3q25.2	2935	896	3.3	Overlaps H- Inv v7.0 gene predictions (HIT000005157)
EXOC4|Contig:chr7:133294036-133294151+|200|7q33	3900	863	4.5	
OC90|Contig:chr8:133070320-133070492-|189|8q24.22	5076	763	6.7	In a region with CTCF TFBS poi from ENCODE
ZNF639|Contig:chr3:179041259-179041338+|201|3q26.33	3224	934	3.5	Burge lab RNA seq colon

|C1GALT1|Contig:chr7:7222241-7222309+|338|7p22.1	3986	729	5.5	Overlaps ESTs, possibly an exon
mir-25 |Contig:chr7:99691191-99691460-|2652|7q22.1	4232	787	5.4	miRNA

REFERENCES

1. Khoury, M.J., Clauser, S.B, et al. Population Sciences, Translational Research and the Opportunities and Challenges for Genomics to Reduce the Burden of Cancer in the 21st Century. *Cancer Epidemiol Biomarkers Prev.* 2011 October; **20**(10): 2105–2114
2. Karger, B.L., Guttman, A. DNA Sequencing by Capillary Electrophoresis. *Electrophoresis*. 2009 June ; **30**(S1): S196–S202
3. Gilbert W, Maxam A. *Proc Natl Acad Sci U S A*. 1973; **70**:3581–3584
4. Sanger F, Nicklen S, Coulson AR. *Proc. Natl. Acad. Sci. USA*. 1997; **74**:5463–5467
5. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, et al. *Nature*.1986; **321**:674–679.
6. Glazer AN, Mathies RA. *Curr Opin Biotechnol*. 1997; **8**:94–102.
7. Jorgenson JW, Lukacs KD. *Science*. 1981; **53**:266–272.
8. Kasper TJ, Melera M, Gozel P, Brownlee RG. *J Chromatogr*. 1988; **458**:303–312.
9. Guttman, A.; Paulus, A.; Cohen, AS.; Karger, BL., et al. *Electrophoresis* 1988; ***VCH***:151-159
10. Guttman A, Cohen AS, Heiger DN, Karger BL. *Anal. Chem*. 1990; **62**:137–141.
11. Huang XC, Quesada MA, Mathies RA. *Anal. Chem*. 1992; **64**:2149–2154.
12. Taylor JA, Yeung ES. *Anal. Chem*. **1992**:1741–1749.
13. Schena, M., et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995)
14. Lipshutz,R., Fodor,S., Gingeras,T. and Lockha synthetic oligonucleotide arrays. *Nature Genet*
15. Voelkerding, K., Dames,S., Durtschi, J. Next-G Basic Research to Diagnostics. *Clinical Chemi*
16. Nyren P, Pettersson B, Uhlen M. Solid phase DNA mini-sequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem*. 1993;**208**:171–5.
17. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*. 1996;**242**:84–9.

Figure 12. Measure correlations between different miRNA detection platforms

18. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science*.1998;**281**:363–5.
19. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;**437**:376–80.
20. Pearson BM, Gaskin DJ, Segers RP, et al. The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828). *J Bacteriol*. 2007;**189**:8402–3.
21. Huse SM, Huber JA, Morrison HG, et al. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. 2007;**8**:R143.
22. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008;**456**:53–9.
23. Quail MA, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods.* 2008;**5**:1005–10.
24. Dohm JC, Lottaz C, Borodina T, et al. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 2008;**36**:e105.
25. Erlich Y, Mitra PP, delaBastide M, et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods*. 2008;**5**:679–82.
26. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*. 2008;**40**:722–9.
27. Shendure,J. Porreca,G. Reppas,N., Church, G. et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome *Science*. 2005; **309**;5741:1728-1732
28. Hawkins, RD., Hon, G., Ren, B., Next-Generation Genomics: an Integrative Approach. *Nat Rev Genet.* 2010; **11(7)**: 476–486.
29. Head, S., Komori, HK., LaMere, S., et al. Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques*. 2014; **56**(2):61-77.
30. Maitra, R., Kim, J., Dunbar, W. Recent advances in nanopore sequencing. *Electrophoresis*. 2012; **33**(23):3418-28.
31. Bragg, L., Stone, G., Tyson, G. Shining a light on dark sequencing:

characterising errors in Ion Torrent PGM data. *PLoS Comput Biol*. 2013; (**4**):e1003031.

32. Bianchi, P., Laghi, L., Delconte, G., Malesci, A. Prognostic Value of Colorectal Cancer Biomarkers. *Cancers.* 2011; **3**:2080-2105
33. Parkin, D., Bray, F., Ferlay, J., Pisani, P. Global cancer statistics, 2002. *CA Cancer J. Clin.* 2005; **55**:74-108.

34. American Cancer Society. *Cancer Facts & Figures 2014* Atlanta, Ga: American Cancer Society; **2014**.
35. Lynch, H., de la Chapelle, A. Hereditary colorectal cancer. *N Engl J Med*. 2003; **348**:919-32.

36. Kaemmerer, E., Klaus, C., Jeon, MK., Gassler, N. Molecular classification of colorectal carcinomas: The genotype-to-phenotype relation. *World J Gastroenterol.* 2013;**19**(45):8163-8167
37. Kanthan R, Senger JL, Kanthan SC. Molecular events in primary and metastatic colorectal carcinoma: a review. *Patholog Res Int* 2012; **2012**: 597497
38. Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. *Gastroenterology* 2010; **138**: 2059-2072
39. Peltomaki, P. Role of DNA mismatch repair defects in the pathogenesis of human cancer. *J. Clin. Oncol.* 2003; **21**:1174-1179.

40. Jass, J., Young, J., Leggett, B. Evolution of colorectal cancer: Change of pace and change of direction. *J. Gastroenterol. Hepatol.* 2002; **17:**17-26.

41. Moreira L**.**, Balaguer F., Lindor N, de la Chapelle A, Hampel H. et al. Identification of Lynch syndrome among patients with colorectal cancer. *JAMA.* 2012; **308**:1555-1565
42. French A., Sargent D., Burgart L. et al. Prognostic significance of defective mismatch repair and BRAF V600E in patients with colon cancer. *Clin Cancer Res.* 2008; **14**: 3408-3415
43. Goel, A., Nagasaka,T., Arnold C. et al. The CpG island methylator phenotype and chromosomal instability are inversely correlated in sporadic colorectal cancer. *Gastroenterology.* 2007; **132**: 127-138

44. Suehiro Y, Wong CW, Chirieac LR, Kondo Y. et al. Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. *Clin Cancer Res.* 2008; **14**: 2560-2569

45. Ogino, S.; Nosho, K.; Kirkner, G.; Kawasaki, T.; et al. CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut.* 2009; **58**: 90-96.

46. Leggett B, Whitehall V. Role of the serrated pathway in colorectal cancer pathogenesis. *Gastroenterology.* 2010; **138**:2088-2100

47. Torlakovic E., Gomez J., Driman D. et al. Sessile serrated adenoma (SSA) vs. traditional serrated adenoma (TSA). *Am J Surg Pathol.* 2008; **32**:21-29

48. East J., Saunders B., Jass J. Sporadic and syndromic hyperplastic polyps and serrated adenomas of the colon: classification, molecular genetics, natural history, and clinical management. *Gastroenterol Clin North Am.* 2008; **37**: 25-46

49. Goel A, Boland C. Recent insights into the pathogenesis of colorectal cancer. *Curr Opin Gastroenterol.* 2010; **26**: 47-52

50. Hur K, Cejas P, Feliu J, Moreno-Rubio J. et al. Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis
Gut. 2014; **63**:4 635-646

51. O'Connell, J.; Maggard, M.; Ko, C. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *J. Nat. Cancer Inst.* **2004**; *96*: 1420-1425.

52. Midgley, R.S.; Yanagisawa, Y.; Kerr, D.J. Evolution of nonsurgical therapy for colorectal cancer. *Nat. Clin. Pract. Gastroenterol. Hepatol.* 2009; **6**: 108-120.

53. Sobin, L.H.; Fleming, I.D. TNM classification of malignant tumors, fifth edition. *Cancer.* 1997; **80**: 1803-1804.

54. Locker, G.Y.; Hamilton, S.; Harris, J.; Jessup, J.M.; Kemeny, N.; Macdonald, J.S.; Somerfield, M.R.; Hayes, D.F.; Bast, R.C., Jr. ASCO 2006

update of recommendations for the use of tumor markers in gastrointestinal cancer. *J. Clin. Oncol.* 2006; **24**:5313-5327.

55. Wood LD, Parsons DW, Jones S, Lin J, Sj ¨oblom T, et al. The genomic landscapes of human breast and colorectal cancers. *Science. 2007;* **318**:1108–13
56. Leary RJ, Lin JC, Cummins J, Boca S, Wood LD, et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl. Acad. Sci. USA. 2008;* **105**:16224–29
57. Malumbres M, Barbacid M. RAS oncogenes: the first 30 years. *Nat. Rev. Cancer. 2003;* **3**:459–65

58. Rajagopalan H, Bardelli A, Lengauer C, Kinzler K, Vogelstein B, Velculescu V. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. *Nature.* 2002*;* **418**:934
59. Fearon, E. Molecular Genetics of Colorectal Cancer. *Annu. Rev. Pathol. Mech. Dis.* 2011; **6**:479–507
60. TCGA, Kucherlapati, R., et al. Comprehensive molecular characterization of human colon and rectal cancer*. Nature.* 2009; **487**:330–337

61. Veronese, A. et al. Oncogenic role of miR-483-3p at the IGF2/483 locus. *Cancer Res*. 2009; **70**:3140-3149

62. Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature reviews* **12**, 99-110.
63. Mendes Soares, L.M. & Valcarcel, J. The expanding transcriptome: the genome as the 'Book of Sand'. *The EMBO journal* **25**, 923-931 (2006).
64. Senti, K.A. & Brennecke, J. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet* **26**, 499-509.
65. Croce, C.M. Causes and consequences of microRNA dysregulation in cancer. *Nature reviews* **10**, 704-714 (2009).
66. Volinia, S. et al. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 2257-2261 (2006).

67. Volinia, S. et al. Reprogramming of miRNA networks in cancer and leukemia. *Genome research* **20**, 589-599 (2010).
68. Medina, P.P., Nolde, M. & Slack, F.J. OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. *Nature* **467**, 86-90.
69. Chivukula, R.R. & Mendell, J.T. Abate and switch: miR-145 in stem cell differentiation. *Cell* **137**, 606-608 (2009).
70. Sachdeva, M. et al. p53 represses c-Myc through induction of the tumor suppressor miR-145. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 3207-3212 (2009).
71. Cordes, K.R. et al. miR-145 and miR-143 regulate smooth muscle cell fate and plasticity. *Nature* **460**, 705-710 (2009).
72. Garzon, R., Marcucci, G. & Croce, C.M. Targeting microRNAs in cancer: rationale, strategies and challenges. *Nature reviews* **9**, 775-789.
73. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic acids research* **36**, D154-158 (2008).
74. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382 (2009).
75. Campagna, D. et al. PASS: a program to align short sequences. *Bioinformatics (Oxford, England)* **25**, 967-968 (2009).
76. Harrow, J. et al. GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7 Suppl 1**, S4 1-9 (2006).
77. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
78. Sarver, A.L., Li, L. & Subramanian, S. MicroRNA miR-183 functions as an oncogene by targeting the transcription factor EGR1 and promoting tumor cell migration. *Cancer research* **70**, 9570-9580.
79. Braun, C.J. et al. p53-Responsive micrornas 192 and 215 are capable of inducing cell cycle arrest. *Cancer research* **68**, 10094-10104 (2008).
80. Schetter, A.J. et al. MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. *Jama* **299**, 425-436 (2008).
81. Michael, M.Z., SM, O.C., van Holst Pellekaan, N.G., Young, G.P. & James, R.J. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol Cancer Res* **1**, 882-891 (2003).

82. Liu, J., Bandyopadhyay, N., Ranka, S., Baudis, M. & Kahveci, T. Inferring progression models for CGH data. *Bioinformatics (Oxford, England)* **25**, 2208-2215 (2009).
83. O'Donnell, K.A., Wentzel, E.A., Zeller, K.I., Dang, C.V. & Mendell, J.T. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839-843 (2005).
84. He, L. et al. A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828-833 (2005).
85. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).
86. Simon, R., Radmacher, M.D., Dobbin, K. & McShane, L.M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **95**, 14-18 (2003).
87. Liu, J., Bandyopadhyay, N., Ranka, S., Baudis, M. & Kahveci, T. Inferring progression models for CGH data. *Bioinformatics (Oxford, England)* **25**, 2208-2215 (2009).

BIBLIOGRAPHY

1. Nakanishi H, Taccioli C, **Palatini J**, Fernandez-Cymering C, Cui R, Kim T, Volinia S, Croce CM. Loss of miR-125b-1 contributes to head and neck cancer development by dysregulating TACSTD2 and MAPK pathway. Oncogene. 2014 Feb 6;33(6):702-12.

2. Previati M, Manfrini M, Galasso M, Zerbinati C, **Palatini J**, Gasparini P, Volinia S. Next generation analysis of breast cancer genomes for precision medicine. Cancer Lett. 2013 Oct 1;339(1):1-7.

3. Fassan M, Volinia S, **Palatini J**, Pizzi M, Fernandez-Cymering C, Balistreri M, Realdon S, Battaglia G, Souza R, Odze RD, Zaninotto G, Croce CM, Rugge Md Facg M. MicroRNA Expression Profiling in the Histological Subtypes of Barrett's Metaplasia. Clin Transl Gastroenterol. 2013 May 16;4:e34.

4. Meng W, McElroy JP, Volinia S, **Palatini J**, Warner S, Ayers LW, Palanichamy K, Chakravarti A, Lautenschlaeger T. Comparison of microRNA deep sequencing of matched formalin-fixed paraffin-embedded and fresh frozen cancer tissues. PLoS One. 2013 May 16;8(5):e64393.

5. Jones KB, Salah Z, Del Mare S, Galasso M, Gaudio E, Nuovo GJ, Lovat F, LeBlanc K, **Palatini J**, Randall RL, Volinia S, Stein GS, Croce CM, Lian JB, Aqeilan RI. miRNA signatures associate with pathogenesis and progression of osteosarcoma. Cancer Res. 2012 Apr 1;72(7):1865-77.

6. Volinia S, Galasso M, Sana ME, Wise TF, **Palatini J**, Huebner K, Croce CM. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci U S A. 2012 Feb 21;109(8):3024-9.

7. Lenze D, Leoncini L, Hummel M, Volinia S, Liu CG, Amato T, De Falco G, Githanga J, Horn H, Nyagol J, Ott G, **Palatini J**, Pfreundschuh M, Rogena E, Rosenwald A, Siebert R, Croce CM, Stein H. The different epidemiologic subtypes of Burkitt lymphoma share a homogenous micro RNA profile distinct from diffuse large B-cell lymphoma. Leukemia. 2011 Dec;25(12):1869-76.

8. Fassan M, Volinia S, **Palatini J**, Pizzi M, Baffa R, De Bernard M, Battaglia G, Parente P, Croce CM, Zaninotto G, Ancona E, Rugge M. MicroRNA expression profiling in human Barrett's carcinogenesis. Int J Cancer. 2011 Oct 1;129(7):1661-70.

9. Goparaju CM, Blasberg JD, Volinia S, **Palatini J**, Ivanov S, Donington JS, Croce C, Carbone M, Yang H, Pass HI. Onconase mediated NFKβ downregulation in malignant pleural mesothelioma. Oncogene. 2011 Jun 16;30(24):2767-77.

10. Kim T, Veronese A, Pichiorri F, Lee TJ, Jeon YJ, Volinia S, Pineau P, Marchio A, **Palatini J**, Suh SS, Alder H, Liu CG, Dejean A, Croce CM. p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. J Exp Med. 2011 May 9;208(5):875-83.

11. Sana ME, Iascone M, Marchetti D, **Palatini J**, Galasso M, Volinia S. GAMES identifies and annotates mutations in next-generation sequencing projects. Bioinformatics. 2011 Jan 1;27(1):9-13.

12. Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, Marchesini J, Mascellani N, Sana ME, Abu Jarour R, Desponts C, Teitell M, Baffa R, Aqeilan R, Iorio MV, Taccioli C, Garzon R, Di Leva G, Fabbri M, Catozzi M, Previati M, Ambs S, Palumbo T, Garofalo M, Veronese A, Bottoni A, Gasparini P, Harris CC, Visone R, Pekarsky Y, de la Chapelle A, Bloomston M, Dillhoff M, Rassenti LZ, Kipps TJ, Huebner K, Pichiorri F, Lenze D, Cairo S, Buendia MA, Pineau P, Dejean A, Zanesi N, Rossi S, Calin GA, Liu CG, **Palatini J**, Negrini M, Vecchione A, Rosenberg A, Croce CM. Reprogramming of miRNA networks in cancer and leukemia. Genome Res. 2010 May;20(5):589-99.

13. Godlewski J, Nowicki MO, Bronisz A, Nuovo G, **Palatini J**, De Lay M, Van Brocklyn J, Ostrowski MC, Chiocca EA, Lawler SE. MicroRNA-451 regulates LKB1/AMPK signaling and allows adaptation to metabolic stress in glioma cells. Mol Cell. 2010 Mar 12;37(5):620-32.

14. Visone R, Rassenti LZ, Veronese A, Taccioli C, Costinean S, Aguda BD, Volinia S, Ferracin M, **Palatini J**, Balatti V, Alder H, Negrini M, Kipps TJ,

Croce CM. Karyotype-specific microRNA signature in chronic lymphocytic leukemia. Blood. 2009 Oct 29;114(18):3872-9.

15. Volinia S, Mascellani N, Marchesini J, Veronese A, Ormondroyd E, Alder H, **Palatini J**, Negrini M, Croce CM. Genome wide identification of recessive cancer genes by combinatorial mutation analysis. PLoS One. 2008;3(10):e3380.

16. Morrison C, **Palatini J**, Riggenbach J, Radmacher M, Porcu P. Fine-needle aspiration biopsy of non-Hodgkin lymphoma for use in expression microarray analysis. Cancer. 2006 Oct 25;108(5):311-8.

17. Guzman J, Yu JG, Suntres Z, Bozarov A, Cooke H, Javed N, Auer H, **Palatini J**, Hassanain HH, Cardounel AJ, Javed A, Grants I, Wunderlich JE, Christofi FL. ADOA3R as a therapeutic target in experimental colitis: proof by validated high-density oligonucleotide microarray analysis. Inflamm Bowel Dis. 2006 Aug;12(8):766-89.

18. Lemon WJ, **Palatini JJ**, Krahe R, Wright FA. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. Bioinformatics. 2002 Nov;18(11):1470-6.

ACKNOWLEDMENTS

I would like to thank my family for their support. I would like to thank my Ferrara family, Drs. Stefano Volinia, Marco Galasso, and ME Sana . To Profs. Cuneo and Bernardi. I wish to thank my lab staff at The Genomics Shared Resource at The Comprehensive Cancer Center at The Ohio State University and my colleagues at OSU. I would also like to thank Drs. Ralph Krahe and Kimmo Virtaneva for planting the first seeds of science in my heart, without their early guidance this would not be possible. I wish to thank Dr. Curtis Harris at the NCI for providing us with the rare paired normal/adenocarcinomas that were used in these studies.

Most of all I would like to thank Prof. CM Croce and his lab and staff at OSU, for without his and their support none of these studies would have been possible.