Università del Piemonte Orientale

Dipartimento di Scienze della Salute

Dottorato di Ricerca in Scienze e Biotecnologie Mediche

XXXI Ciclo (A.A. 2015-2018)

Development of two new approaches for NGS data analysis of DNA and RNA molecules and their application in clinical and research fields

Francesco Favero

Supervised by Prof. Umberto Dianzani

Co-Supervised by Prof. Flavio Mignone

Index

1.0 Introduction

1.1 General Introduction

During the last years the sequencing of nucleic acids has been a key technology to understand the structure of genes, the mechanism of splicing, and even to discover new molecules like miRNAs and lncRNAs. The first sequencing technology has been Sanger Sequencing; this technique dominated for almost two decades and led to a number of monumental successes, including the complete sequencing of the human genome [1]. Even if Sanger sequencing is a powerful method to understand the genomic complexity of various organisms, it carries several limitations because, for example, it requires an high amount of time to sequence a relative low number of bases. Even if NGS (Next-Generation Sequencing) produces reads usually shorter than the canonical 1,000 bp of length of the typical Sanger sequences, the massive depth of coverage, i.e. multiple reads over the same template DNA region, compensates for the limitations of short reads. Indeed, the average higher coverage of NGS guarantees a more accurate result compared to Sanger sequencing. Also, NGS technology has got the ability to produce an enormous volume of data cheaply. Nowadays it is possible to sequence an entire genome in few days or even in few hours [2]. These NGS technologies have enabled the possibility to sequence many new genomes, to analyze genomic diversity [3], to discover pathogenic variants [4] and to perform extensive studies of transcription, gene regulation and epigenetics in many species [5-6]. However the main issue of this technology is the high-throughput of data itself, since it is not simple to extract a biological meaning from such a great amount of data, and several computing algorithms and approaches are needed for interpreting any single output of a sequencing run. The instrument returns in output very large files containing a list of the fragmented DNA sequenced called “reads” in which the DNA from input samples is coded and it may be not trivial at all to manage

this output and to perform all the analysis steps required to extract any useful information: this is the reason why bioinformatics analysis became an essential part of the process. Moreover, the field of application for NGS techniques is wide and there is no a “universal” way to analyze the data. Depending on the design of the project it would be needed to use different software for the analysis or to create a new custom one.

1.2 Chemistry of NGS

The preparation of the samples and the sequencing reactions differs in many ways based on the kind of study, the nature of the sample and the platform of sequencing used. For each platform, the input is a double-stranded DNA library consisting of short fragments flanked by adapters of known (and platform-specific) sequence. The most used platforms are produced by Illumina, TermoFisher and Applied Biosystems SOLiD. Here we present the details of Illumina platform, the platform that the data analyzed in this study. However, every currently available platform of NGS uses a step of amplification of DNA sample to output a great amount of reads in order to amplify the signal.

After the extraction of the DNA or the retrotrascription of RNA in cDNA, the sample must be subjected to some step before the sequencing reaction. The sample could be DNA, cDNA or the product of a previous PCR amplification (in the *Amplicon Based* sequencing). Generally the steps to prepare a sample and obtain a DNA library ready to be sequenced are the following:

1) Fragmentation or target amplification.
2) Adapter ligation.
3) Size selection.
4) Quality control and quantification.

1) Fragmentation or target amplification

NGS platform cannot receive in input a large molecule of DNA and directly sequence all of it. For this reason, in the library preparation step, long nucleic acids molecules must be fragmented in small molecules of variable dimension, depending on the platform used (usually into a length of 150–500 base pairs for Illumina platforms. This fragmentation is usually obtained by mechanical approach or enzymatic digestion. If a mechanical approach is chosen, that could be nebulization that consists in using a disposable device driven by pressurized air, or in using ultrasounds to fragment the DNA with Covaris instrument. Covaris fragments the DNA in pieces of different dimension varying the intensity and the duration of the acoustic waves (150 – 5,000 bp) [7]. An alternative method of library preparation is Illumina's Nextera tagmentation technology that incorporate in a single step fragmentation and adapter ligation step: in this method a transposase enzyme fragments and inserts adapter sequences into dsDNA in the same time [7]. Normally, if Nextera tagmentation technology is not used, nucleic acids after fragmentation must be attached to adapters in order to proceed in the library preparation.
The original DNA can also be previously targeted and amplified by PCR, instead of being fragmented, in order to obtain a sequencing of a specific DNA region. The product of the PCR is then normally processed and sequenced.

2) Adapters Ligation

DNA fragments must be blunted and 5' phosphorylated while 3' ends must be Adenosine-tailed in order to properly attach adapters at 5' and 3'. For this goal, the

fragments must be treated using different enzymes: First, a mixture of T4 polynucleotide kinase, T4 DNA polymerase, and Klenow Large Fragment is used for the blunt and for the 5' phosphorylation. Next, the 3′ ends are A-tailed using either Taq polymerase or Klenow Fragment without exonuclease activity (exo-) [7]. After this, a reaction of the enzyme T4 Ligase attaches the adapters to the 5' and 3' properly modified ends of DNA fragments. However, the adapters ligation step can vary a lot depending on the used protocol but it consists always in a PCR reaction. Adapters are molecules ligated with barcodes (also called indexes). Barcodes are oligonucleotides between 6 and 12 bp that are unique, specific for each sample; they are used to associate each read only with a specific sample and are ligated to each fragment of DNA in this experimental step.

3) Size Selection

A size selection is needed before the DNA can be sequenced. Usually the size selection step can be done through magnetic beads.

4) Quality control and quantification

After the size selection step the library is ready to be loaded on the flow cell and to be sequenced. However, a good practice is to perform a titration run in order to verify the integrity of the library and, more importantly, anyway, verify the DNA concentration of the different samples in order to load the right amount of DNA in the flow cell. The volumes and the concentration to load vary from one platform to another.

-Bridge PCR

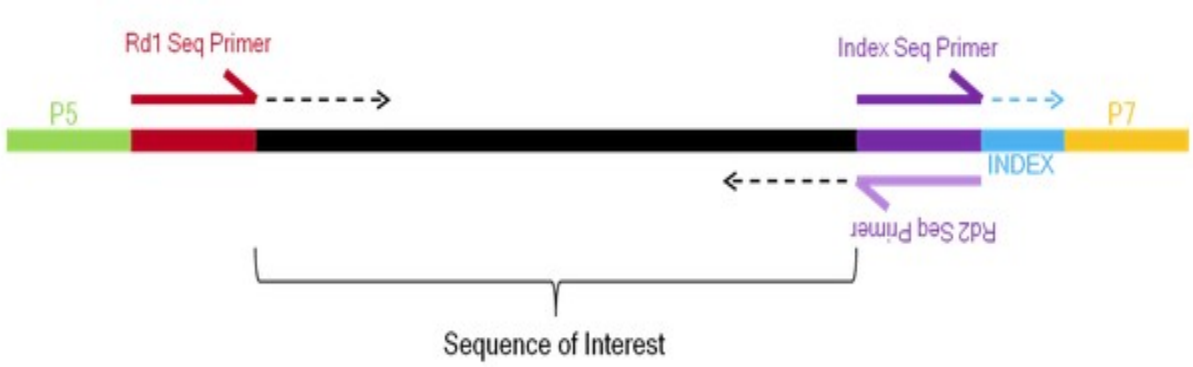
After the size selection and the quality control step, the template DNA, ligated with adapters, P5 and P7 fragments is loaded into the flow cell in which every single molecule of sample is amplified in the *Bridge PCR* step, also called *In Situ* PCR. First, the DNA template hybridizes into the surface of the flow cell thanks to the complementary sequence of P5 and P7 to the oligos present on the surface of the flow cell. After the hybridization many cycles of amplification are performed by the instrument in the *Bridge PCR* step. At the end of this step, each nanowell will contain many clones of the original DNA template that will be sequenced.

-Sequencing Step

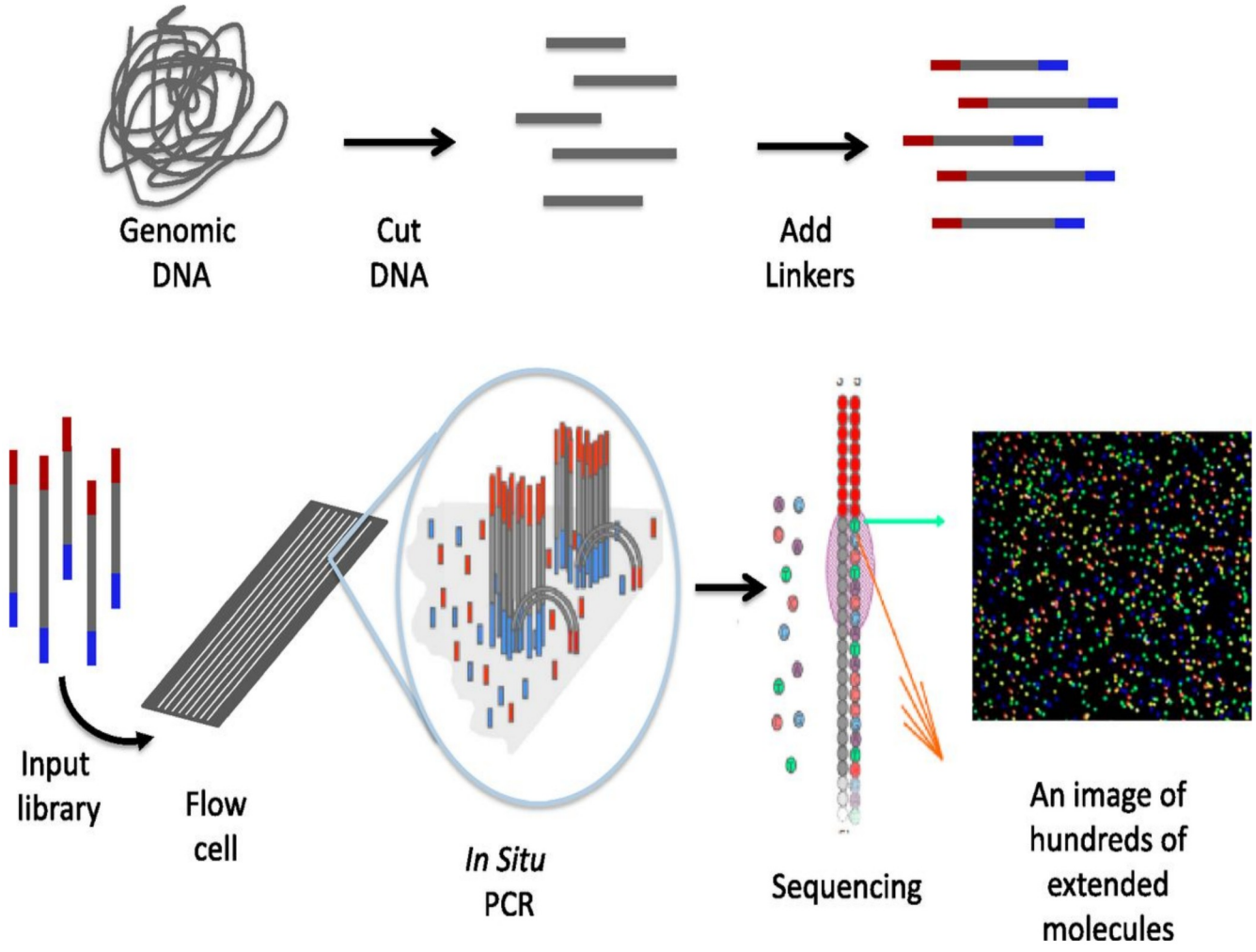
During the sequencing step the clusters are sequenced. After the end of *Bridge PCR* step sequencing primers are added into the flow cell and a sequencing PCR is performed. The sequencing primer anneal to its the complementary region and start to synthesize a new strand of each amplified molecule in the first round of sequencing step. In this PCR, however, called sequence-by-synthesis PCR, that is performed at the same time for each cluster, the Taq polymerase incorporates nucleotides that carry a fluorescent molecule that emits light at different wave lengths depending on the base carried (A, C, T or G). When a fluorescent molecule is added, the camera of the instrument reads and interprets the base incorporated. Each base is added one per round. After the reading of the incorporated base, the fluorescent molecule is washed away and another fluorescent base is added. The second step of sequencing starts at the point in which the barcode primer is added and the barcode is sequenced. After, the reverse primer is added and the read is sequenced in reverse with the same principle if you are using a paired-end approach. Note that for reads with double index, the total number of sequencing steps is four, because a new step with a new primer is needed to sequence the other index. However, every time a base

is included and sequenced, this occurs at the same time for all the cluster on the flow cell, in this way, millions of clustering are sequenced at the same time.

The single-end sequencing works exactly like the paired-end sequencing but it ends after the sequencing of the region of interest and the index (or both index 1 and index 2) without sequencing in reverse the read.



Scheme of the sequencing of each molecular of cluster for a read with a single index [8].



General scheme of Illumina Sequencing procedure [9]

1.3.1 Somatic Variant Analysis

NGS has been also widely used in oncology to discover the key mutations in cancer and has been applied in the field of Molecular Diagnostics

Molecular diagnostics is a procedure that consists in screening a patient at molecular level, particularly analyzing his DNA or RNA, and ascertain if these molecules carry a particular mutation that could have a possible effect on the structure or the

expression level of the proteins produced by the cell. An application of NGS in molecular diagnostics is, for example, the identification of clinically relevant variants present in a biopsy of cancer cells.

The study of the mutation of DNA is pivotal in oncology. The subclonal mutations in cancer it's very important not only to understand the biological evolution of the disease but also to predict the patients' response to treatment [10]. These subclonal mutations are generally present only in a small number of reads in each NGS run, and in low frequency rate. For this reason they could be generally difficult to identify with NGS technology and with currently available bioinformatics tools because they tend to have a frequency similar to the background noise, mostly generated by sequencing errors of each NGS run.

Sequencing errors are one of the most common issues in NGS sequencing and the most typical source of noise background. Although provided accuracy is usually very high (in the range of 98% to 99.9%) each platform has got its specific error model that must be considered when performing analysis. Therefore sequencing errors leads to the insertion of SNP and indels mutation in a random number of reads during the sequencing process. Above all, the background noise could be generated by errors in sequencing process, but also by contamination in the sample or by errors of the PCR sequencing reaction that is performed in targeting sequencing prior to the sequencing step. These facts could lead to call a variant that is not present in the original sequence, in this case a false positive result may be generated. In any case this problem can be compensated by sequencing depth, that represents how many times each nucleotide from input sample is sequenced: since it is very unlikely to get the same errors in all reads, sequencing more and more time the same region of the DNA aims to lower the impact of a single error. Adequate coverage, or sequencing depth, is critical for an accurate calling of true variants and today it is not a problem to achieve very high sequencing depths. The main method used to achieve a sufficient sequencing depth is called *Amplicon-based sequencing* in which the original DNA is

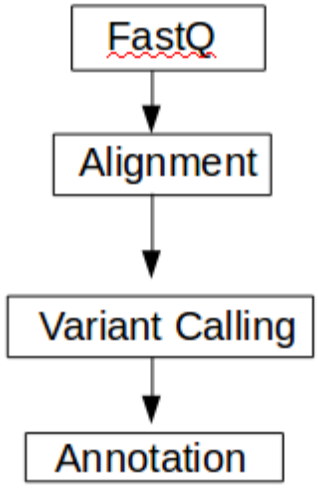
amplified by PCR prior to be sequenced, generating a list of PCR amplicons that later will be sequenced. This process is useful in order to obtain an higher final number of reads and therefore major sequencing depth for each single nucleotide in the interested target region. Sequencing a specific region, like a specific gene, instead of performing a whole exome or a whole genome sequencing is a preferable approach in this situation to obtain an higher sequencing depth in the region of interest.

Variant calling software can distinguish true subclonal variants after the alignment of the reads to a reference genome [11-19]. The alignment aims to associate each read with some coordinates of the reference and to probe if in that region there is a mutation in the sequenced sample. This can lead to the discovery of SNP and indels in the region of interest. But after the alignment step a huge list of variants with vary frequency is found; many of them consists in noise background. For this reason the subsequent analysis step is the variant calling step. Many of the current available tools for variant calling can't identify accurately somatic point mutations, and are generally applied in WES (Whole Exome Sequencing) or in WGS (Whole Genome Sequencing) and not in ultra-deep sequencing based on PCR amplicons [20]. WES and WGS are techniques of sequencing that allow to explore a large portions of a genome, or even sequencing it all, but with relative low degree of read depth, generally around 100-150x for WES and 30x for WGS [21]. Most part of the currently available variant calling tools, for example SomaticSniper, are calibrated to run at these read depths [22]. The variant calling tools are often used in the study of oncogenes and oncosuppressor genes such as KRAS, NRAS, BRAF and EGFR because they are very important for the diagnosis and prognosis, and for the treatment of many malignancies such as colorectal cancer [23] because these genes often contain hotspot missense mutations, which are the focus of variant calling tools [24-25]. However, the available variant callers, such as GATK [24] and VarScan2 [25] are designed to call indels and SNV but often they lack in sensibility or in precision.

For all these reason we used a custom pipeline of analysis to analyze our data described in chapter “Materials and Methods”.

1.3.2 NGS data analysis workflow: Detection of somatic DNA mutations

The analysis of somatic variants from NGS data is complex and characterized by multiple steps. The following scheme sums up every step.



Workflow of NGS analysis of somatic variants

FastQ: The workflow of each analysis starts from the FastQ file that contains the list of reads sequenced by the instrument.

Alignment: This step is computing challenging and consists in the comparison of each read to a reference sequence that usually for amplicon-based approach is a small sequence like a single gene.

Variant Calling: In this step of the analysis, true variants must be distinguished from false variant (noise background).

Annotation: In this step variants must be analyzed one by one, compared to the known variants databases such as dbSNP [26] and ClinVar [27] to obtain more information on them and ascertain, for example, if a specific variant is already known and associated with a certain pathology.

1.3.3 BRCA1 and BRCA2

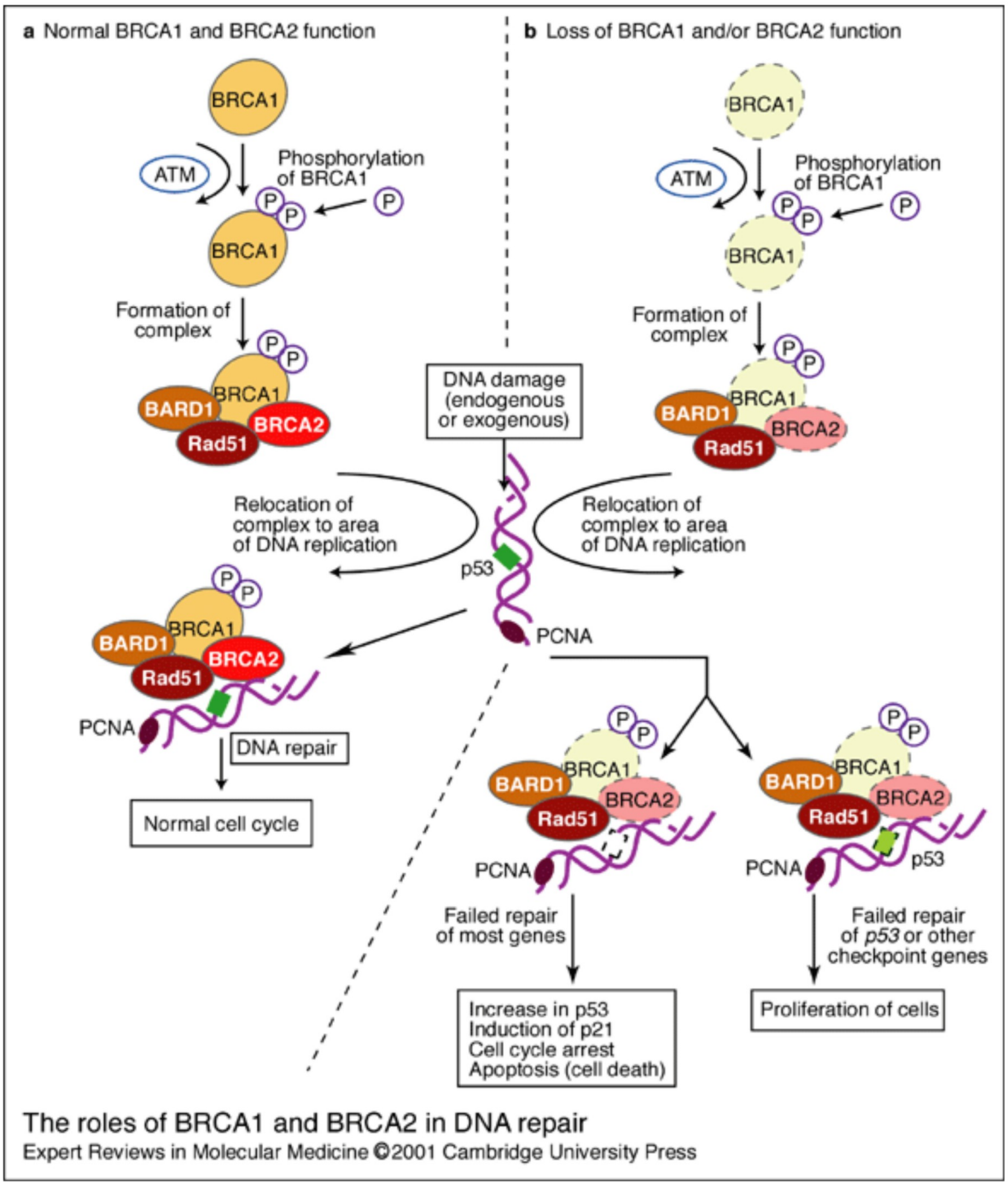
BRCA1 and BRCA2 are genes located respectively on chromosome 13 and 17. These genes codify for two tumor suppressor proteins. Unfortunately, it is common that a mutated version of these genes is present on patients in germline state, these mutations are responsible for the hereditary breast and ovarian cancer. It is sufficient a single loss of one of the two alleles of these two genes to produce a considerable higher risk to develop cancer. Patients that inherited a mutated version of BRCA1 or BRCA2 have got a risk of 50-80% of developing breast cancer and of 30–50% of developing ovarian cancer. However, mutated BRCA1 and BRCA2 are not only correlated with breast and ovarian cancer, but also prostate cancer and pancreatic cancer. It has been calculated, for example, that for the mutation of BRCA2, the relative risk of developing pancreatic and prostate cancer are up to 10-fold and 20-fold respectively [28].

BRCA1 and BRCA2 are involved in some mechanisms of DNA damage repair, in particular in the mechanism called Homologous Recombination (HR) and in the mechanism of Non-Homologous End Joining (NHEJ). Homologous Recombination uses a template strand, the other chromosome, as template to guide repair. Since a template is present, this process is almost error free; it is activated by BRCA1 and is performed by a crossing-over mechanism between the portion of damaged DNA and the correspondent sequence on the other chromosome. Non-Homologous

recombination instead is activated by both BRCA1 and BRCA2 and doesn't need a template DNA to be performed, but, since a benchmark sequence is not given, it is more likely that in this process of repair a mutation is inserted [29].

DNA damage can be generated by errors during the replication process, radiations or other genotoxic compounds, like oxygen free radicals [30], and can be very dangerous if it occurs during the replication of a cell because, if not repaired, the damage can be inherited and consolidated in the cell lineage. For this reason, during cell cycle there are several checkpoints, particularly in G1/S and G2/M phases, in which sensor proteins check the integrity of the DNA, to repair it if needed, before going on with the replication.

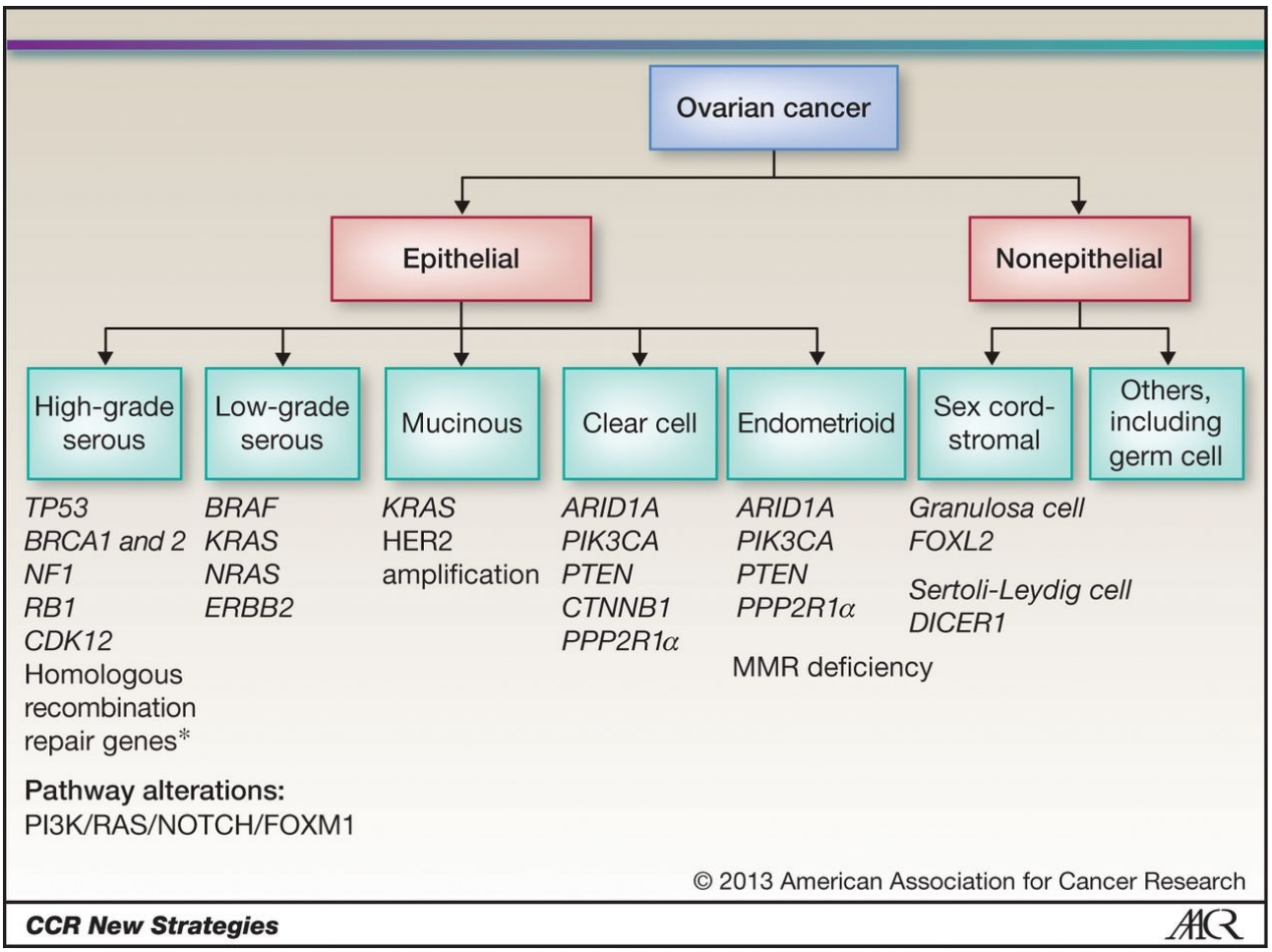
BRCA1 and BRCA2 can interact each other in a complex with RAD51 and other proteins for the activation of HR pathway [31]. BRCA1 and BRCA2 interact also with the protein ATM and with TP53 that has got a central key in the physiology of cell cycle [32]. If BRCA1 or BRCA2 are mutated, the pathway of TP53 is deregulated. With a deregulation of TP53, the normal cell cycle is disrupted and also the apoptosis is deregulated. If these pivotal functions are not correctly performed, the cell gains an advantage in survival and doesn't die even if it is present a damage in the DNA. With the accumulation of DNA damage and the downregulation of TP53 pathway, the cell could gain malignant features; if TP53 is downregulated, the checkpoint controls of the DNA damage are skipped and the cell continues to proceed in the cell cycle even when DNA damage is present.



The roles of BRCA1 and BRCA2 in DNA repair

Expert Reviews in Molecular Medicine ©2001 Cambridge University Press

1.3.4 Ovarian cancer and PARP inhibitors



Classification of ovarian cancer

Breast cancer is the most common cancer in woman worldwide and is often based on a mutation of BRCA1 or BRCA2 like ovarian cancer. Ovarian cancer causes more deaths in the United States than any other type of female reproductive tract cancer, with an estimated 22,430 new cases and 15,280 deaths in 2007 [33]. Ovarian cancer can be classified in various way depending on the cell of origin and the degree of differentiation. A scheme of the main classification of ovarian cancer is shown in the figure from the AARC of 2013 presented in this paragraph: the figure shows the main subtypes along with the main altered genes of ovarian cancer. The sex-cord stromal ovarian cancer derives from a component of the stromal cells of the ovary:

the granulosa cells, thecal cells and fibrocytes. Generally this kind of ovarian cancer derives from the cells that secrete hormones. Germ cells are the cells that will become gametes. Epithelial carcinoma derives from cells on the surface layer that covers the ovary or other tracts of the reproductive system like fallopian tubes. Serous carcinomas may have a complex admixture of cystic and solid areas with extensive papillations, or they may contain a predominantly solid mass with areas of necrosis and hemorrhage [34]. Mucinous ovarian cancers are cystadenomas that usually occur as a large, multiloculated cystic mass with mucus-containing fluid and have other peculiar features [35]. Clear Cell Carcinoma of the ovary (CCCO) is an epithelial ovarian carcinoma but shows peculiar features, mains are: CK7+ and CK20+, absence of mutations in BRCA1, BRCA2, p53, but presence ofmutations in ARID1A, PIK3CA and pathway of mTOR up-regulated [36].

In this research, we analyzed data only from patient diagnosed with high-grade serous ovarian carcinoma (HGSC) that is the only subtype that may carry a mutation of BRCA1 and/or BRCA2. We sequenced only patients with HGSC that developed drug resistance against traditional chemotherapy used in first line treatment; they have been screened for BRCA1 and BRCA2 mutations with our NGS analysis. If BRCA1 or BRCA2 were mutated, with a mutation that alters the function or the expression of the protein, the patient has been treated with PARP inhibitors: the mechanism of action of these drugs is highly related with the DNA mechanism of repair.

DNA mechanisms of repair are divided in two main groups:

-Single-Strand: this mechanism repairs the damage without using any template for repairing the damage. This class is divided in many submechanisms (base excision repair, nucleotide excision repair and mismatch repair).

-Double-Strand: this mechanism uses the other filament of DNA as template for replacing the bases damaged. This is divided in Nonhomologous end-joining (NHEJ) mechanism and Homologous Recombination (HR).

All mechanisms of DNA repair of the single strand break depend on PARP (Poly-ADP-Ribose Polymerase) proteins. PARP protein family is a superfamily of 18 proteins [37]. PARP inhibitors like Olaparib, already approved by FDA and EMA [38-39] activate they effect threw a mechanism called tumor-selective cytotoxicity. PARP inhbitors in particular target PARP1 and PARP2 that are two protein with a zinc-finger domain that binds the damaged DNA. The role of PARP to ADP-ribosylates itself and its ADP-ribosylation aims to recruit other proteins including PARP itself in the damage site of the DNA and promotes the repair of the damaged region [40]. PARP inhibitors block both PARP1 and PARP2. Since these proteins are fundamental to repair the damage of the single strand of the DNA, these inhibitors block this pathway. PARP inhibitors work for both hereditary and somatic forms of ovarian cancer. However, the drug doesn't have a systemically effect since in normal cells BRCA mutations are often in heterozygosis, and these mutations have no a phenotypic effect, and the cells, with a reduced amount of BRCA are still able to survive even when PARP1 or PARP2 are inhibited since SSB pathway is blocked, the replication fork is blocked for the great accumulation of SSB and the SSB becomes a DSB but the cell can still repair the DSB through HR. On the contrary, in cancer cells BRCA1 and BRCA2 are often mutated in homozygosis and HR is impossible to be activated. This lead to a significant synthetic lethality of PARP inhibitors in cells where BRCA1 and BRCA2 are mutated in homozygosis since they have no other way to repair the DSB caused by the accumulation of SSB [41-42]. As final effect, even if these drugs are usually administrated orally, their toxic effect is limited to cancer cells.

However, since PARP normally inhibits the pathway of the error-prone DSB repair of NHEJ mechanism, the inhibition of PARP leads to an aberrant activation of this mechanism that accumulates even more errors in DNA structure [43]. So, with PARP inhibitors not only the SSB cannot be repaired, and this brings to more DSB errors,

but when a DSB occurs cannot be repaired with HR mechanism, that is impaired, but is attempted to be corrected with NHEJ mechanism, that is prone to errors. Because all of these factors, it is very likely that a cancer cell would accumulate enough damage in DNA structure to die.

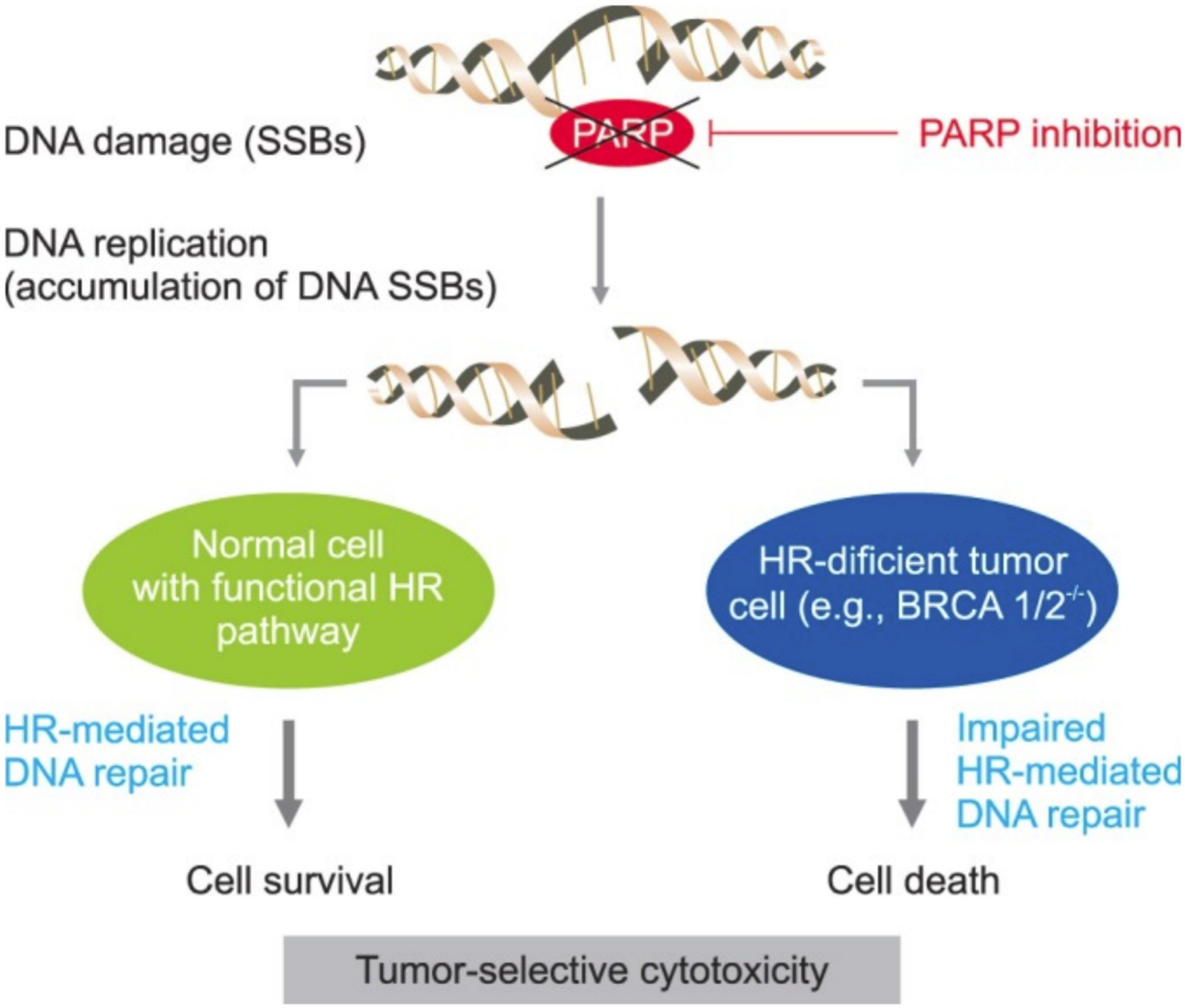


Illustration of Tumor-selective cytotoxicity of PARP inhibitors [44]

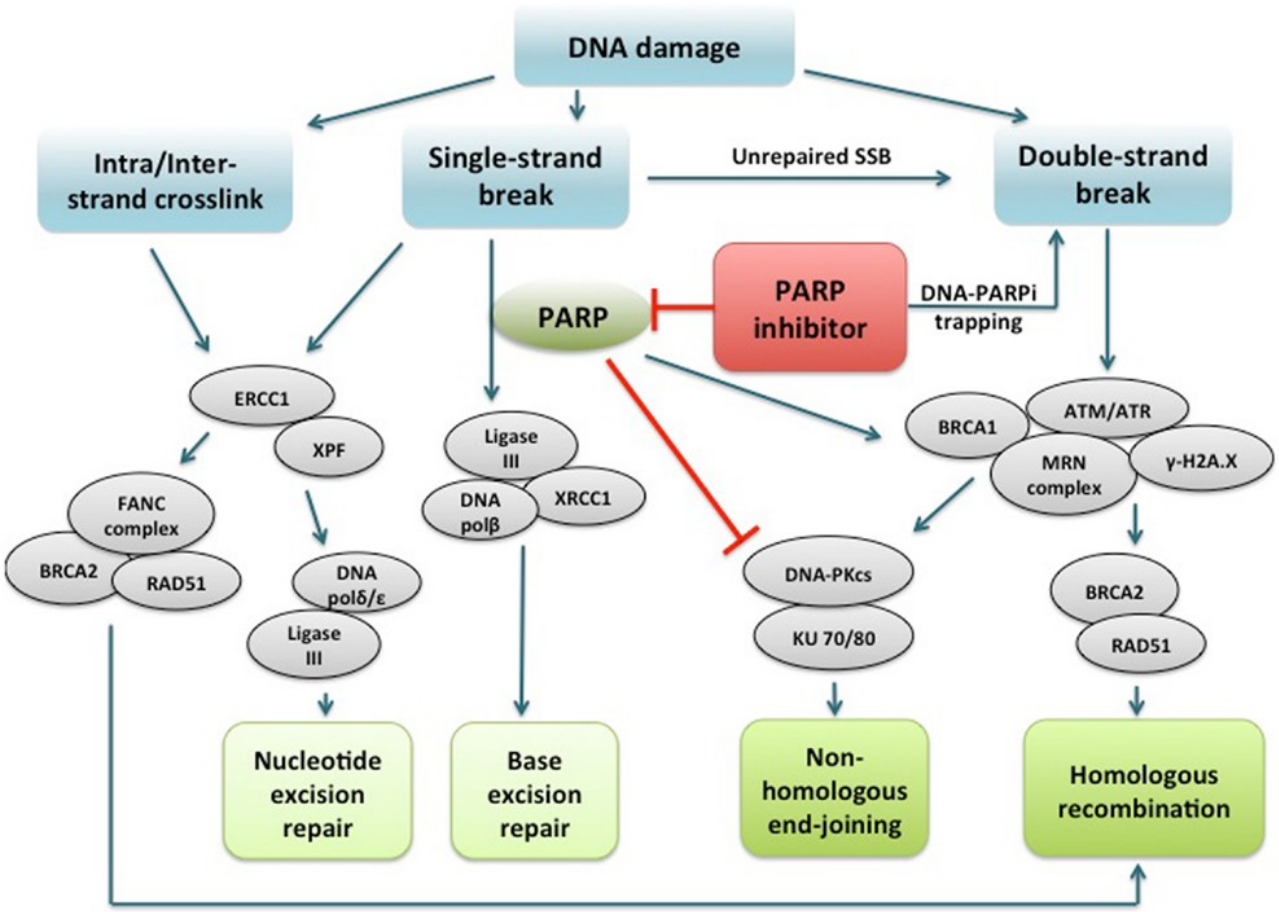


Illustration of DNA damage pathways and PARP inhibitors activity [45]

1.4.1 RNA-Seq

RNA-seq (also called Whole-Transcriptome shotgun sequencing (WTSS) [46]) consists in the sequencing of all the transcripts produced in a specific biological sample. It is a powerful technique that aims to identify and quantify all the transcripts present in a controlled experimental situation. With a RNA-seq analysis it is possible to obtain different pieces of information about, besides protein coding genes, miRNAs, lncRNAs and other non-coding RNAs. The data analysis approach of an RNA-seq depends on the availability of a reference genome of the organism or the organisms in exam. If a reference genome is available, it is possible to perform a *reference based* analysis, else it is necessary to perform a *De novo assembly.* RNA-seq can be applied on several kinds of study such as evaluation of nucleotide variations, evaluation of methylation patterns [47], micro-RNA analysis, and analysis of Differential Expressed Genes (DEG). In general in every study of DEG, after the extraction from the biological sample, the RNA is fragmented and retrotranscripted into cDNA and then sequenced on a NGS platform. After this step, the reads are mapped to a genome or transcriptome, if a *reference based* approach is used, otherwise in the *de novo* approach the reads are reassembled with specific software. Finally, the expression levels for each gene or protein isoform is estimated by statistical analysis and DEG can be obtained [48-49]. After this step of analysis, it is necessary to give a peculiar biological significance to the obtained DEG, starting from the original biological question [50]. With the increasing popularity of RNA-Seq technology, many software and pipelines have been developed to analyze these data but they are usually designed for the *Reference based* method. This approach is certainly the most efficient and accurate but it is applicable only if high quality annotated references are available; this could be true for human genome, but not for all species of mammals, plants, bacteria, fungi or other organisms. When this

condition is not fulfilled, the only way to extract some valid information from RNA-seq data is the *De novo assembly* approach. In this methodology, reads are not aligned to a reference genome but assembled one to another to recreate the original full length transcripts. There are some tools like Bridger [51], Oases [52] and Trinity [53] that can assembly *de novo* reads from RNA-seq, but although they work with a good level of sensibility, they produce an high number of false positive that result in the lack of specificity.
This new software is called STAble and it will be explained in details in “Materials and Methods” chapter; STAble is thought to be applied in Metatransciptomics studies.

1.4.2 The human microbiota and its impact on human health

Today NGS technology is one of the most used technique for the study of the microbiota [54]. This study is usually divided in two main fields: metagenomics and metatranscriptomics. Both tries to extract information directly from the sample of interest, without previously culturing the organisms. Metagenomics tries to identify which microorganisms are present usually sequencing the 16S rRNA [55], that is conserved between different species, and aligning reads to databases of annotated sequences, such RDP [56]; with this procedure in some, mostly rare, cases it is possible to perform classification even at strain level. To obtain a taxonomic classification, it is also possible to do a shotgun metagenomics sequencing and amplify all the DNA present in the sample (including fungi and some viruses) that aims to classify more species and go deeper in the analysis even at strain level, but with a greater effort in term of computational performance and sequencing cost since greater coverage is needed to perform this kind of analysis. Instead Metatranscriptomics aims to identify a set of transcripts expressed by the

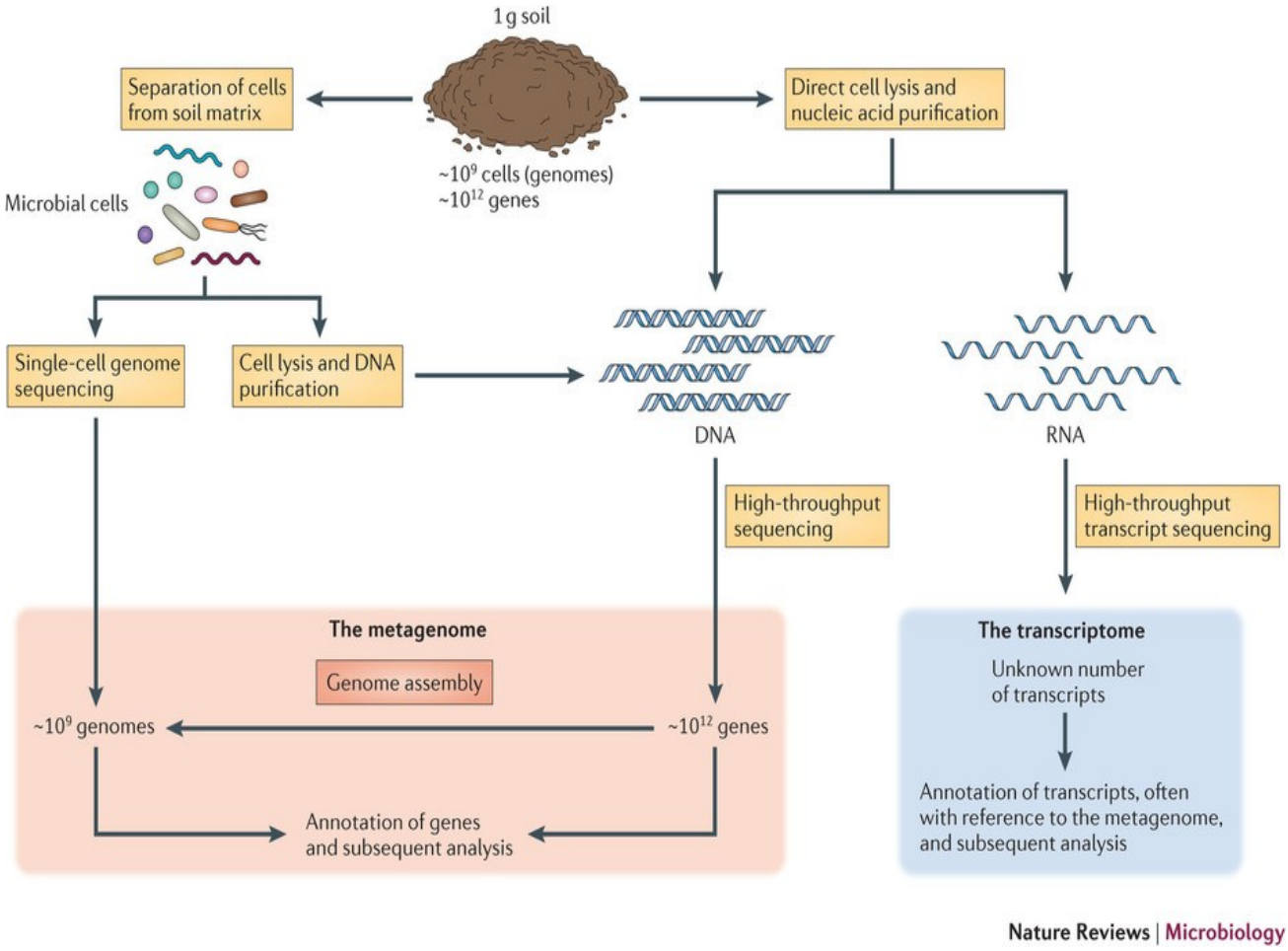
microorganism population sequencing the entire RNA retrotranscripted in cDNA [57].

The study of microbiota is very important for human health since more than 10^{14} microrganisms (bacteria, fungi, protozoa, viruses) are resident on the human body but the most well study group is the one of bacteria [58]. Even if different microrganisms are present in different district of the human body, the most study environment is the gut. The vast majority of the bacteria of the human gut microbiota are *Firmicutes* and *Bacteroidetes* [59-60] and they reside in the distal part of the gut [61]. The gut microbiota contributes to the host's homeostatis with the biosynthesis of essential amino acids and vitamins [62] while the host gives to the resident microorganisms nutrients and a controlled environment in which they can live, recreating a perfect symbiotic relationship. Plus, microbiota is very important to coordinate immune system, since studies conducted using germ-free mice suggest that the microbiota directly promote local intestinal immunity through its effects on toll-like receptor (TLR) expression, antigen presenting cells, differentiated T cells, lymphoid follicles and the promotion of systemic antibody expression [63-66]. Even if the above mentioned states are linked to physiological events, the influence of microbiota is pivotal even in pathological situation since a lot of disease have been linked to an altered gut microbiota. For example obesity is usually correlated by an altered intestinal Bacteroides:Firmicutes ratio [67]. Above all, if it is clear that a wealth state of gut microbiota is essential to maintain a good homeostasis of the digestive tract, the influence of gut microbiota is not only linked to the digestion and assimilation processes or to the immune system, since a recent study evidenced how the gut bacteria can produce significant amounts of amyloid proteins and lipopolysaccharides, which are key players in the pathogenesis of Alzheimer’s disease [68]. Also, it has been shown that individuals with metabolic disorders such as obesity and diabetes have much more probability to have intestinal dysbiosis compared to healthy individuals [69-70]. Since microbiota plays a critical role in all

these systems, both in health and in pathological condition, and since diet can rapidly influence the composition of microbiota population [71], altering the balance of power between species i.e. advantaging or disadvantaging pathogen species, it is evident that the interest in understanding the microbiota population is currently increasing. Since every single patient has got a different physiological gut microbiota, the potential of tailored medicine in this field would be enormous: the ultimate goal would be, given a pathological situation, administer the best treatment (drug or diet) to realign the patient's altered microbiota to a condition of physiological homeostasis. The impact of diet is very powerful on gut microbiota because, for example, intestinal small chain fatty acids have been shown to directly increase the abundance of T regulatory cells in the gut and to protect against allergic airway inflammation [72-75]; also they may inhibit the transcription factor NF-κB, leading to a decrease secretion of several pro-inflammatory cytokines [76]. The importance of an intervention on microbiota to threat metabolic disease, and potentially other kind of pathological states, it is proved by several situations. For example *Lactobacillus* seems to have an important role in treat obesity. Particularly *Lactobacillus* has been shown to alleviate obesity-associated metabolic complications interacting with *Bacteroidetes* and *Firmicutes* and also modulating host immunity [77-79]. Therefore acting on the gut microbiota can be extremely important to drive a patient towards an health situation.

Although during the last years more and more databases of human symbiotic microorganisms are being generated, like The Human Microbiome Project (HMP) [80] that keeps a curated collection of sequences of microorganisms associated with the human body, including eukaryota, bacteria, archaea and viruses, from both shotgun and 16S sequencing projects and other specialized databases have being produced comprising only members of the human intestinal microbiota [81-82], there is still not a database which contains full sequenced genomes of all the species that composes the human microbiota. For this reason STAble works with a *de novo*

assembly approach to be applied for the study of potentially every microorganism, even for those whose reference is not available.



Differential approaches for sequencing microbial populations with NGS technology

1.4.3 STAble core algorithm: Reconstruction of transcripts module

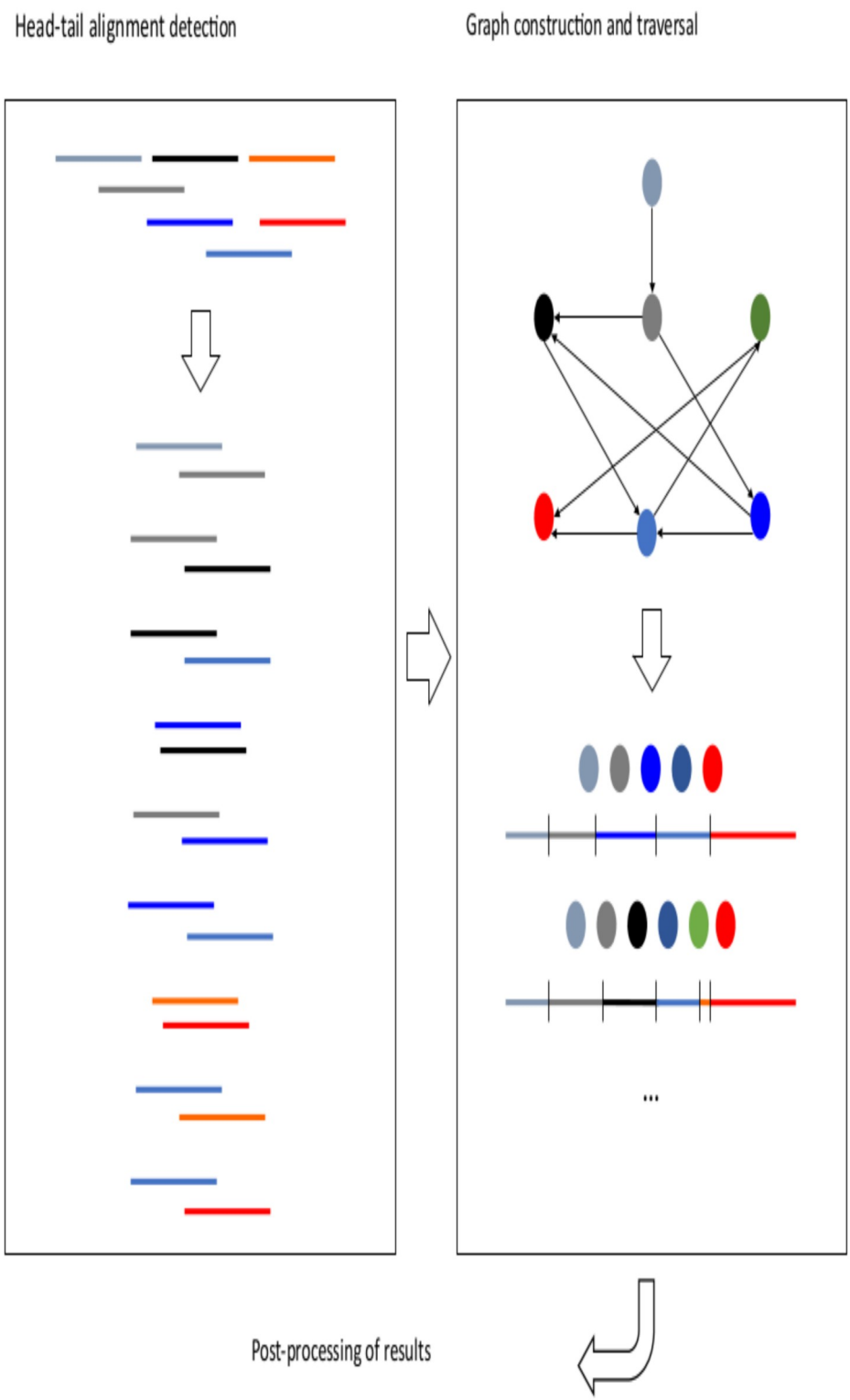
STAble algorithm has been originally developed in order to perform *de novo* reconstruction of transcripts on bacterial samples.

In the first module, STAble uses a new approach in order to reconstruct transcripts from raw reads. The central idea is that, instead of using small k-mers, STAble uses the entire length of the reads to reconstruct transcripts, performing a head-tail

alignment of the reads, in order to reduce the number of false positive results. STAble performs three main processes in order to reconstruct transcripts from reads:

1. Efficient detection of head-tail alignments.
2. Construction and traversal of an unweighted directed graph.
3. Post processing of results.

The head-tail alignment step between reads is a fundamental process because if two reads align each other in head-tail position with a good score, they can be assembled into a longer contig. The alignment starts from a k-mer called anchor, that is used to start the alignment. The default length of anchor is 11 and an anchor is considered valid only if consists in a sequence that contains all four nucleotides in order to filter low homopolymerich regions or other low complexity regions. Moreover, in this first alignment step a number of mismatches are allowed to be aligned in order to deal with sequencing errors (default value is 10% of overlap length). There are other important parameters that play an essential role in this step of analysis like. All of them, explained in details in [83], are important to avoid incorrect alignments that can lead to false positive results or for avoiding to obtain chimeric reconstructions and to reduce the complexity of the next steps of the analysis.
Then, all the alignments and all the reads are built up in a De-Brujin graph that must be correctly followed to obtain a list of correctly reconstructed transcripts. This second part is very complex from a computational point of view. Briefly, each node of this graph represents a read and each arc represents an head-tail alignment between two of them. Every time a node is traversed, the partial reconstructed transcript is elongated until no more head-tail alignments are available.
Benchmark data of this first module of STAble can be found in Appendix B.


Head-tail alignment detection
Graph construction and traversal
...
Post-processing of results

Stable first module workflow: The left panel shows the head-tail alignments between reads, the right panel shows the building and the crosswalk of the directed unweighted graph to reconstructtranscripts. After this step, transcripts are post-processed before returning the output.1.4.4 Post-Processing

Due to high redundancy of NGS data it is possible that different paths that are completely unrelated in terms of graph nodes represent the same biological sequence, so we decide to implement a clustering algorithm, *vsearch* in order to collapse redundant results that they have the same biological meaning [84].

2.0 Materials and Methods

2.1 Perl scripts

Since FASTQ files, the standard output format of NGS platforms, contain hundreds of thousands of reads, it is impossible to manually manage this kind of files. For this reason, a large part of this research work was based on the usage of a programming language studied for the manipulation of huge text files, “Perl” [85]. Perl is a modulated, interpreted programming language that integrates some functions of C, sed, awk, sh and however, even if it is mainly projected for scanning huge texts files and extracting pieces of information from them, it is also a good language for system management.

2.2.1 Variant Calling GUI – Amplicon Suite

The new variant calling algorithm presented in this work is integrated in a software for the analysis of SNP and indels called Amplicon Suite. The software is composed by an intuitive GUI (Graphical User Interface) that can be used to automatically process raw data.Amplicon Suite integrates every step of NGS sequencing data analysis, starting from alignment step to the annotation of the identified variants.

2.2.2 Variant Calling algorithm

Our variant calling algorithm is based on the detection of outliers in a distribution of values as described in [86]. Since the vast majority of NGS variants in an NGS run is part of the noise background, the software takes this fact in consideration and builds

up a suitable two-way contingency table for the detection of outliers. The input parameters that the algorithm takes in are the number reads that confirm each variant and their sequencing depth. A one-tailed binomial distribution is constructed by the algorithm to detect the outliers.

2.2.3 Simulated Datasets

The variant callers were tested on simulated dataset to carfully benchmark performances with a controlled environment.

Simulated datasets were generated selecting 96 target regions on BRCA1 and BRCA2 genes similar to the target regions amplified in real samples (see below). Reads were generated using ART [87] as Illumina 150 bp paired end with Miseq quality profile. Briefly, we generated a number of 100.000 normal and 100.000 mutated reads per amplicon. In mutated reads we inserted a total number of 800 SNPs and 500 indels mutations. For each amplicon we inserted a number of known mutations. . A custom perl script took normal and mutated reads and shuffle them in order to give a different number of reads for each amplicon, simulating various degrees of sequencing depth for BRCA1 and BRCA2 genes for each amplicon. We put mutations in each sample at a custom depth and frequency only in one amplicon at a time. This choice has been made to control the number of mutation for each sample, in order to recreate a number of mutations per sample that could be found in real datasets, and to test if our system was good to identify mutations occurring in every amplicons of BRCA1 or BRCA2 genes. This process has been made several times inserting mutations in each gene and in each amplicon, simulating a total number of about 1300 samples. Mutations were inserted in a depth varying from from 0 to 10.000, with controlled frequencies of 1%, 3% and 10%.

When we run our algorithm to analyze these simulated data and distinguish between true variants and noise background we divided the variants in two groups, separating

SNPs from indels, since it is known that in NGS runs they have a different noise background. We didn't perform this division when we run the other algorithms since they are supposed to work directly on the BAM file that contains all the aligned reads of the sample in exam.

2.2.4 Other Variant Calling software

We compared the performance of our algorithm to the best variant calling software according to the analysis of Xu et al. [20] and VarDict [88].
All of the variant calling algorithms have generally been run, if not otherwise specified with standard parameters suggested in documentation or with the commands suggested by Xu et al [20]. In particular, we calculated the sensibility and the precision of each tool, testing it on the simulated data that we generated with ART simulator. In particular we tested:
-VarDict (v1.07) [88]: VarDict is an ultra sensitive variant caller for both single and tumor-normal paired samples. VarDict implements several novel features such as amplicon bias aware variant calling from targeted sequencing experiments, rescue of long indels by realigning soft clipped reads and better scalability than many Java based variant callers.
We run VarDict with the "single sample analysis" option. we simulated our datasets in sam format and we converted them into a bam file using Samtools [89], after this conversion we sorted and indexed the BAM file using the same tool, then we analyzed the bam file using VarDict and we obtained the result in VCF format. The minimum variant frequency is a needed parameter; we run the software always with the minimum possible value for that parameter (1%). This software can find both SNP and indels mutations.

-SomaticSniper [22] version v1.0.2: SomaticSniper calculates the Bayesian posterior probability of each possible joint genotype across the normal and cancer samples. This software works only making a differential analysis between the “normal” sample and the “tumor” samples; the software compares these two samples and highlights the variants that are present in the tumor sample but not in the normal sample. We generated the input files into a SAM file and then we converted it into a BAM file using Samtools [89]. After this conversion we used the same tool to sort and index the BAM file, then we run SomaticSniper with standard parameters on the sorted and indexed BAM file and we obtained the result in VCF format. This software cannot call indel mutations.

-Mutect [90] version v1.1.4: MuTect is an algorithm that detects somatic point mutations in NGS data using a Bayesian classifier approach and can perform a single sample analysis. We run this algorithm with its default parameter settings using java 6.

We decided to not use the standard Reject filter in order to avoid a too restrictive variant calling analysis: This filter would remove false positives (FP) caused by nearby misaligned small indel events. This software cannot identify indel mutations.

-VarScan2 [25] version v2.3.6: VarScan2 needs two files for working: normal sample file and tumor sample file. This software automatically call variants on the base of coverage and quality of the variant itself separately in the normal file and in the tumor file. This tool is based on the Fisher’s exact test. We generated the input file using SAMtools mpileup command, starting from a standard indexed and sorted BAM file, and then we run the program with standard parameters but calibrating the software setting to the minimum frequency of the variant it had to find in order to increase the performance of the software knowing the frequency of the inserted mutations. VarScan2 can point out SNP and indels mutations. We run it with the specific option of “min-var-freq”, in order to improve the performance in precision.

-GATK [22] version GenomeAnalysisTK 2.3.9: This software can call variants separately from tumor sample and normal tumor and it doesn't need two distinct BAM files to work. We converted simulated data to a SAM file and obtained the BAM from it using Samtools [89]. With the same tool we sorted and indexed the BAM file, and after, we used PicardTools function AddOrReplaceReadGroups to generate the head of the file that GATK needs for working. Using Java 6, we launched the algorithm with standard parameters.
-Strelka [91] version v.1.0.7: Although Strelka is another software suggested for variant calling by Xu and colleagues [20], this algorithm is very complicated to use and requires a deep configuration before starting the analysis for any single sample. Because of this problem we decided to not compare our system to Strelka algorithm.

We also considered to compare the performance of our system of analysis against Strelka2 [92], Mutect2 (that is just an update of Mutect [90]), SNVsniffer [93] and JointSNVMix [94] as somatic variant calling softwares but they rely on paired tumor-normal samples for calling variants. For this reason we chose to not test them on our datasets.
All of the variant calling algorithms have generally been run, if not otherwise specified with standard parameters suggested in documentation or with the commands suggested by Xu and colleagues [20].

2.2.5 Statistical analysis of simulated datasets

We evaluated the performance of our algorithm and the performance of the software in terms of sensibility and precision.

Sensibility=TP/TP+FN (True Variant called/(True Variant called+True Variant not called)).

Precision= TP/TP+FP (True Variant called/(True Variant called+Error called)).

2.2.6 Real dataset

We analyzed a cohort of 474 patients from Policlinico Gemelli of Rome, all diagnosed with ovarian cancer classified as High-grade serous carcinoma (HGSC). Patients signed the consensus for the treatment of their data in research field. The library preparation and the sequencing reaction have been performed following the instructions of BRCA-Devyser Kit protocol for the sequencing of BRCA1 and BRCA2 genes. Several variants discovered by our algorithm have been later validated using Sanger Sequencing method.

More information about the kit can be available at [95].

2.3 STAble methods

We prepared several lines of code in order to add two new modules to the main core of STAble. These two modules consist in a minor module written in Perl that associate the reconstructed transcripts to a list of known mRNAs and count them estimating their abundancy, putting this information in a contingency table, and a more complex module that perform a flux-balance analysis starting from the informations contained in the contingency table. After the writing of the two new modules, we added them at the original STAble core algorithm. We then tested it on a real RNA-seq dataset as described in our work [83]. Briefly we downloaded reads of

some metatranscriptomic data extracted from sheep gut available in SRA from Kamke and colleagues work [96] and applied the first module of STAble on them in order to obtain a list of reconstructed mRNAs. We then downloaded bacterial FASTA sequences of orthologous genes of several pathways (glycolysis/gluconeogenesis, butanoate metabolism, methane metabolism, carbon fixation pathways, phosphotransferase system) from KEGG orthology database [97]. We chose these pathways because Kamke and colleagues already investigated these data and they highlighted them as the most important and expressed in their experiment.

Briefly, the reconstructed transcripts were aligned to bacterial genes using BLAST algorithm accepting matches with at least 92% of similarity and allowing up to 20 nucleotides of mismatches over flanking regions. After this step the raw reads were realigned against reconstructed transcripts using BLAST. Then, reads used to reconstruct each transcript are associated with orthologous genes and a contingency table has been obtained with some custom Perl scripts. The contingency tables with read count for each orthologous gene has been processed with metabolic models, with a process similar to flux balance analysis, to interpret gene expression: the process adopted is described in [98]. The metabolic model used for the flux balance analysis is a general metabolic model of *E. Coli*.

3.0 Aim of the study

The aim of this study is focused on two main areas of NGS analysis data: RNA-seq (with a specific interest in meta-transcriptomics) and DNA somatic mutations detection.

We developed a simple and efficient pipeline for the analysis of NGS data derived from gene panels to identify DNA somatic point mutations.

In particular we optimized a somatic variant calling procedure that was tested on simulated datasets and on real data. The performance of our system has been compared with currently available tools for variant calling reviewed in literature.

For RNA-seq analysis, in this work we tested and optimized STAble, an algorithm developed originally in our laboratory for the *de novo* reconstruction of transcripts from *non reference based* RNA-seq data. At the beginning of this study, the first module of STAble was already been written. The first module is the one which reconstructs a list of transcripts starting from RNA-seq data. The aim of this study, particularly, consisted in adding a new module to STAble, developed in collaboration with Cambridge University, based on the flux-balance analysis in order to link the metatranscriptomic analysis to a metabolic approach. This goal has been achieved in order to study the metabolic fluxes of microbiota starting from metatranscriptomic data.

4.0 Results

4.1.0 Variant Calling results overview

At the beginning of this study, we tested the existing variant calling software on the simulated data that we generated with ART, in order to investigate their performance in finding SNPs and indels and verify if they could be applied in the analysis of our data. The results of these tests can be found in Appendix A. Since we were not satisfied of these results, we proceeded with the development of our algorithm of variant calling and we proceed to test it against the other software on simulated data. ASince the performance of our algorithm on simulated data have been satisfying, later we proceeded to use it to analyze real data.

4.1.1 Variant Calling results on simulated datasets

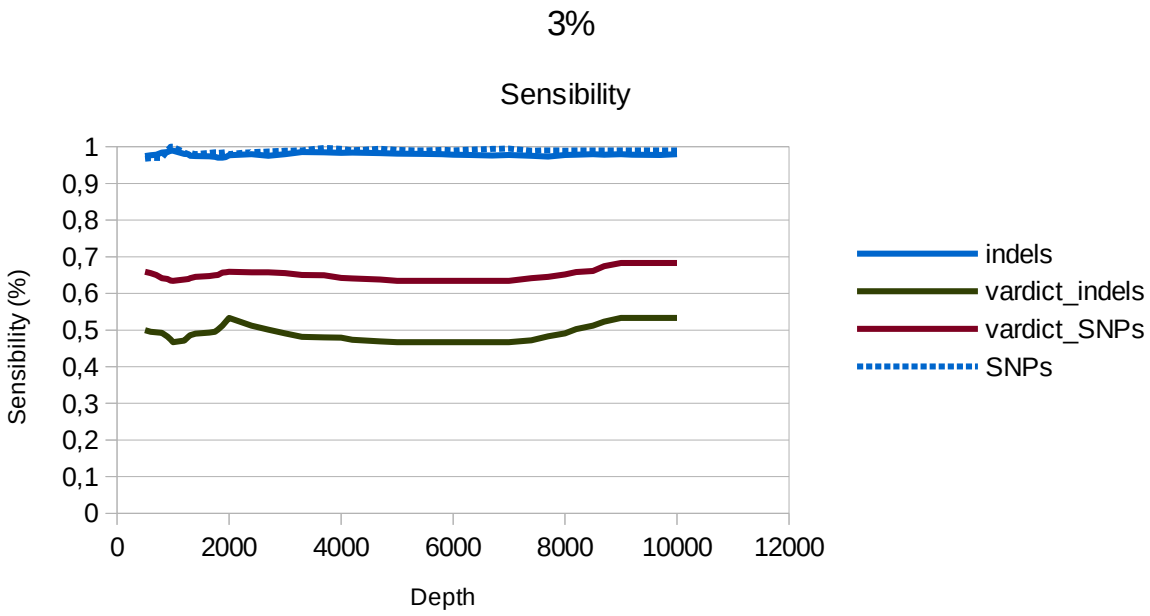
In the following graphs we show a deep analysis of the performances of our algorithm against the other variant calling software. **Figures 1-2** show the performance of our algorithm in finding mutation inserted at vary frequency rate against the best software tested at each condition. The data demonstrate that our algorithm has got better performance in terms of sensibility and its results are more stable and reliable since the sensibility increases directly proportionally with sequencing depth. The software have in some point better performance than our algorithm in terms of precision, but at the same time they show far worse results in terms of sensibility.

In **Figure 3** it is shown a further comparison of the performance of the software against our algorithm in finding somatic mutations at 1% of frequency in our simulated data at a depth of 2000. SomaticSniper and GATK are not shown in this

graph, because at 2000 of depth their sensibility and precision are equal to zero in finding variants at 1% of frequency.

In **Figure 4** it is shown a further comparison of the performance of the software against our algorithm in finding somatic mutations at 1% of frequency in our simulated data at a depth of 5000. VarDict loses performance at this read depth regarding indels mutation. SomaticSniper and GATK are not shown in this graph, because at 5000 of depth their sensibility and precision are equal to zero in finding variants at 1% of frequency.

In the end, our algorithm shows better performance, more stable and more reliable results increasing read depth of sequencing on simulated datasets. It has got a far better sensibility than other variant calling softwares at a cheap cost in precision.



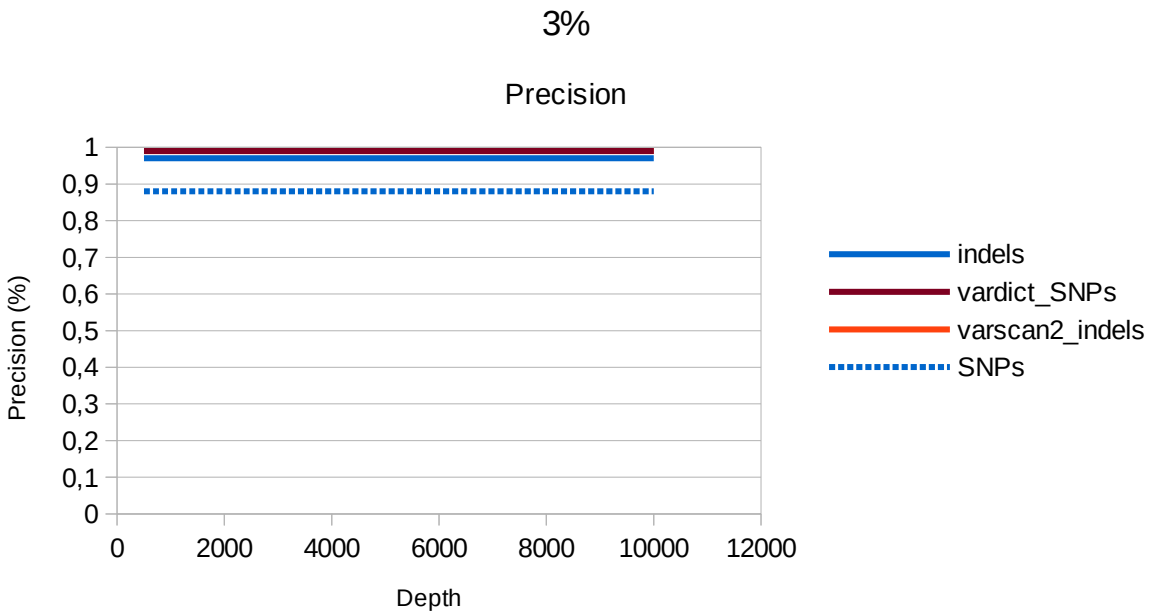


Figure 1: Performance of our algorithm on simulated datasets (mutations at 3% of frequency rate). Our algorithm is far superior than any other software in terms of sensibility (VarDict is the best among the other software at this condition). In terms of precision our performance is slightly worse than the other software (all around 99% of precision) except for transversions (our precision is around 88%), but this drop in precision is far highly compensated by the major sensibility. (VarDict is the best among the other software even if it lacks in precision in finding indels compared to VarScan2).

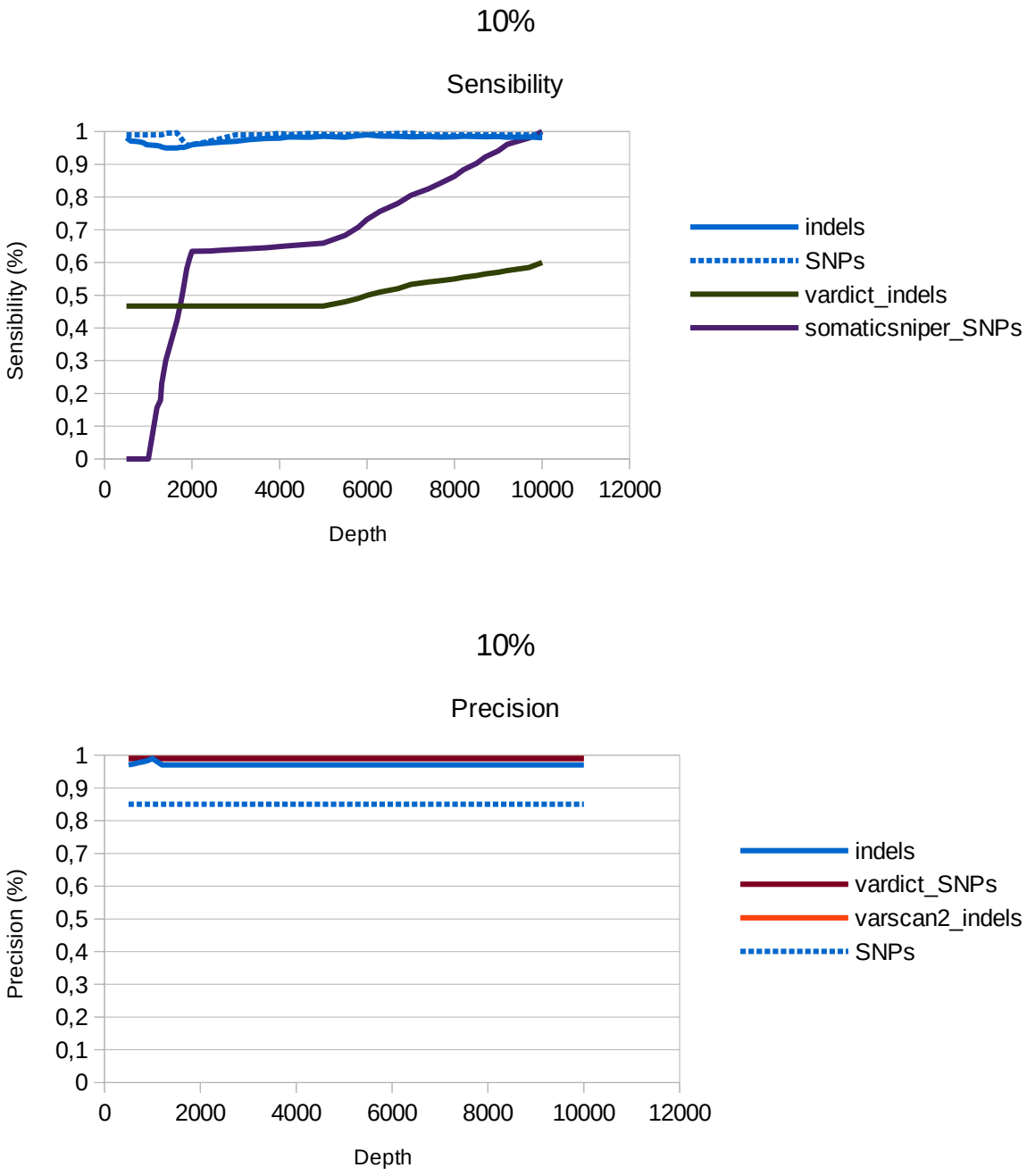


Figure 2: Performance of our algorithm on simulated datasets (mutations at 10% of frequency rate). Our algorithm is far superior than any other software in terms of sensibility. SomaticSniper is the best among the other software at this condition (sensibility at 10%) except for indels; SomaticSniper cannot point out indels mutation while VarDict can. In terms of precision our performance is slightly worse than the other software (all around 99% of precision, our precision is around 85%), but this drop in precision is far highly compensated by the major sensibility. VarDict is the best

among the other software (regarding the precision at 10%) even if it lacks in precision in finding indels compared to VarScan2 at this condition.

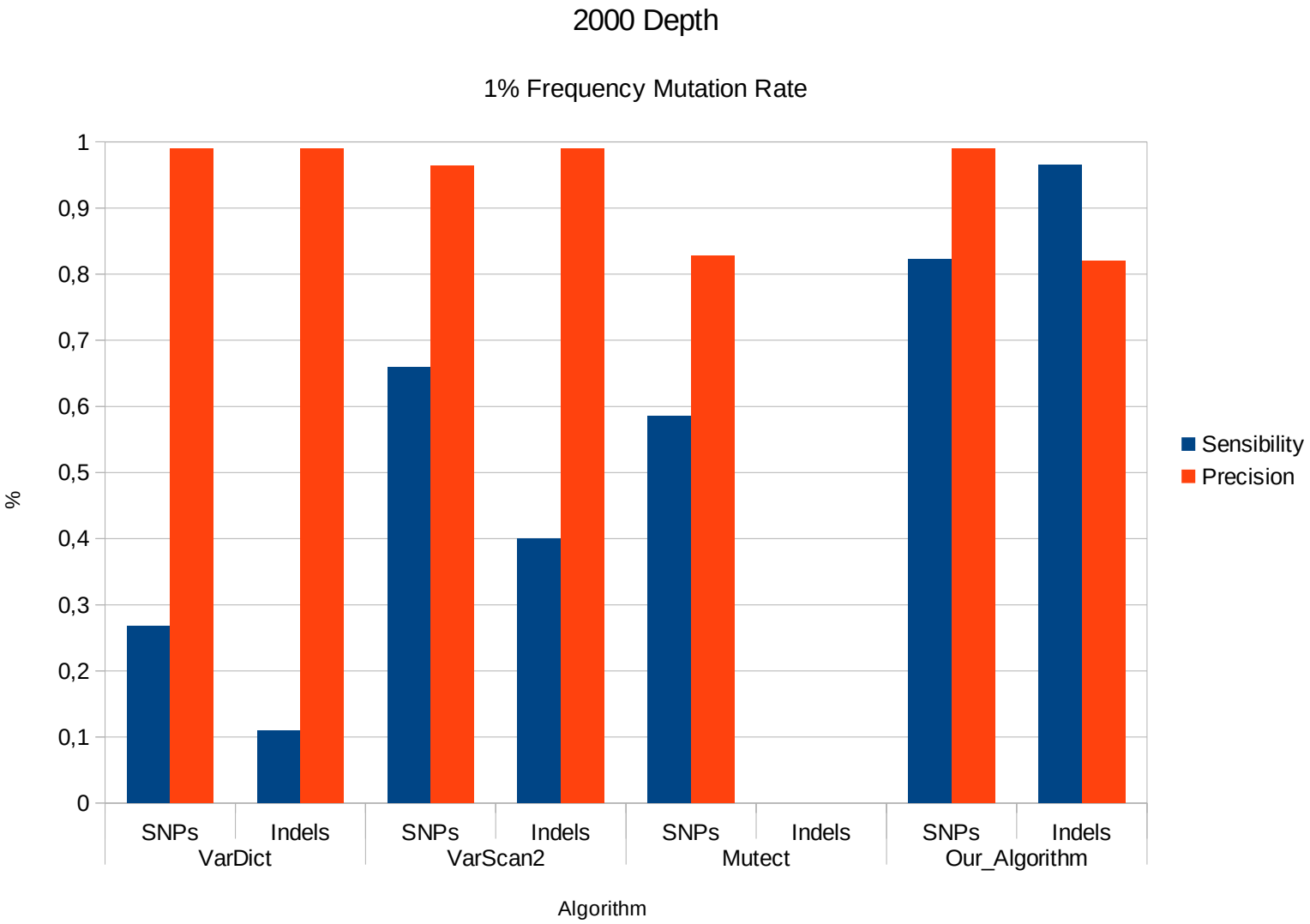


Figure 3: Comparison of the performance of the best software (Mutect, VarScan2 and VarDict) against our algorithm at 2000 depth in finding somatic mutations at 1%. Indels of Mutect is zero because this software cannot call indels.

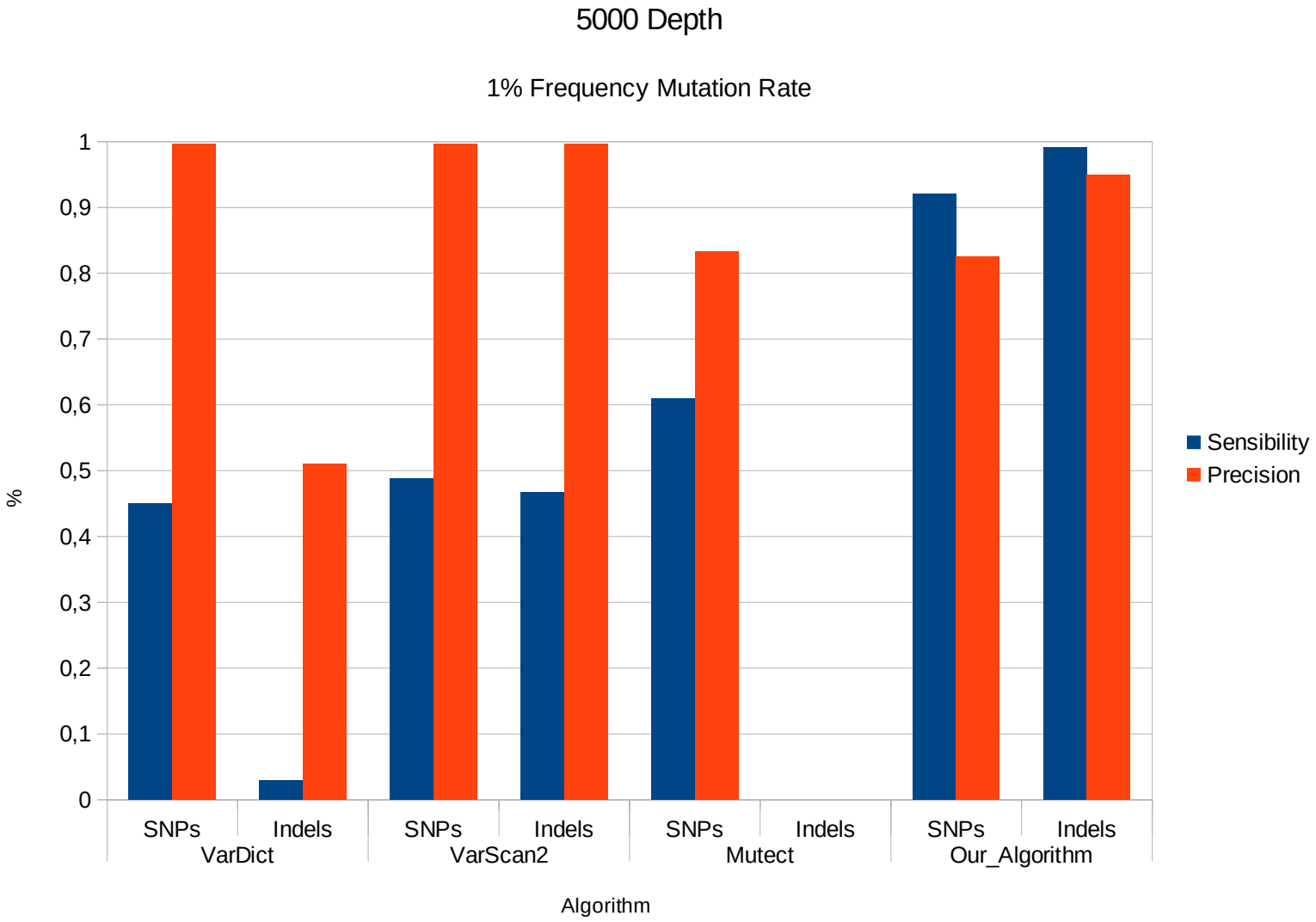


Figure 4: Comparison of the performance of the best software (Mutect, VarScan2 and VarDict) against our algorithm at 5000 depth in finding somatic mutations at 1%. Indels of Mutect is zero because this software cannot call indels.

4.1.2 Variant Calling tests on real dataset

Thanks to a collaboration with Gemelli hospital, we were able to analyze with our algorithm the data obtained from the sequencing of a total court of 474 patients affected by HGSC. When we tested our algorithm we found that a surprisingly high number of variants passed the filter and were reported as true. **Figure 5** shows the results obtained with a sample but results are very similar for all the analyzed samples. Red dots represent the variants that passed our variant caller.

Such an high number of true somatic variant in every sample is biologically very unlikely, moreover we noticed that many of the variants were recurring among samples. However this number of positive variant and their frequency is coherent with artifacts introduced duringthe preparation of the sample. Indeed as demonstrated in [21], during amplicons generation PCR may insert a number of errors. Error rate also depend on the specific base and it has ben demonstrated that it isdifferent between transitions and transversions.

For this reason when decided to apply two changes to our procedure: firstly we decide to apply our algorithm on transitions (G:A and T:C variants) and transversions (all other SNPs) separately since, as demonstrated in [21] they show a different level of noise background. Secondly we decided to introduce a second level of filtering to remove variants present in “many” samples at the “same” frequency. Indeed it is unlikely to find the same variant at very similar frequency in many samples. The second level works as follow: variants with a recurrence in less than of 5% of samples are considered true; more recurring variants are considered true only if their allele frequency is more than 2 standard deviation from the average VAF of the same variants in all the samples.

By applying this second filter only variants marked with an “X” in the figure were considered positive and have been validated with Sanger Sequencing.

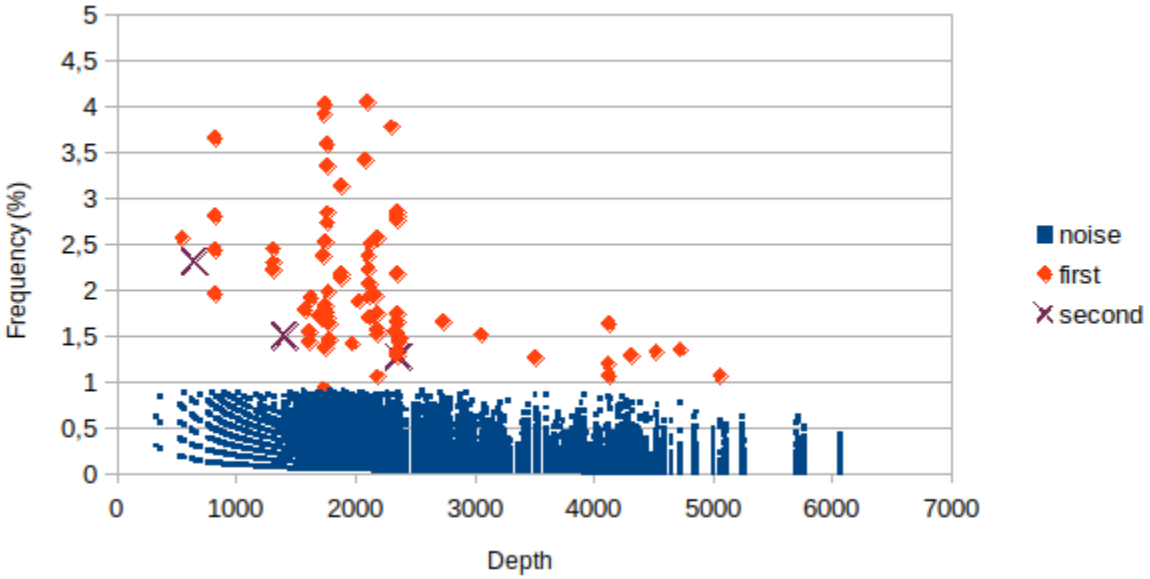


Figure 5: An example of variant distribution in one of the samples in exam. “x” axis represents read depth while “y” axis represents the frequency of variants in percentage. Blue variants are filtered out by the first filter. Red variants remain after the first filtering. Variants indicated with an “X” passed the two filters.

Table 1 shows the filtering steps power on a subset of all the samples that we analyzed. Here we show only the first 49 samples that we analyzed. At the begging, after the alignment step every sample carries a huge number of variants,shown in the “unfiltered” column; The vast majority of them comes from sequencing errors that are filtered out by the first filter. The second level of filtering eliminates the remaining falls positive calls. What remains after this filter is a number of somatic variants, one per sample in average, that is reasonable from a biological point of view.

Table 2 shows the list of variants found in our real dataset that have been validated with Sanger Sequencing

Sample	Unfiltered	First Filter	Second Filter
sample1	61602	197	1
sample2	67991	105	1
sample3	64207	137	1
sample4	65658	138	2
sample5	65673	118	1
sample6	65107	151	2
sample7	62811	138	0
sample8	60308	141	1
sample9	64066	137	0
sample10	66367	116	1
sample11	77551	137	0
sample12	83748	171	4
sample13	75931	138	2
sample14	73224	116	0
sample15	77617	136	2
sample16	59898	213	3
sample17	60523	161	0
sample18	75449	117	2
sample19	58931	159	0
sample20	59673	151	1
sample21	72060	142	1
sample22	80219	117	0
sample23	63829	126	1
sample24	62167	246	2
sample25	69860	161	0
sample26	76783	148	2
sample27	73932	155	1
sample28	82567	216	9
sample29	80834	113	3
sample30	77496	141	2
sample31	62135	133	0
sample32	66175	117	0
sample33	65311	127	0
sample34	71847	118	0
sample35	66752	168	1
sample36	65483	119	0
sample37	64406	123	0
sample38	66487	108	1
sample39	61731	117	0
sample40	64754	122	0
sample41	68625	114	1
sample42	62796	125	0
sample43	72496	103	3
sample44	70211	135	0
sample45	75539	125	0
sample46	63040	114	1
sample47	65527	120	3
sample48	74137	165	4
sample49	60427	139	2

Table 1: This table shows the subsequent levels of filtering the data in our pipeline, showing the number of variants that remain after the application of the filters of our pipeline.

Table 2

Sample	HGVS	Gene	Frequency (%)
27	c.8165C>T	BRCA2	5
41	c.3262A>G	BRCA1	12
38	c.1244A>G	BRCA2	25
21	c.4264_4273delGAGACTTCTG	BRCA2	8
35	c.5313delC	BRCA1	19
20	c.5158_5163delACCCAG	BRCA1	21

Figure 6 show all the validated variants with Sanger method found in our 49 samples. VAF=Variant Allele Frequency.

Panel A: c.8165C>T on gene BRCA2.

Panel B: c.3262A>G on gene BRCA1.

Panel C: c.1244A>G on gene BRCA2.

Panel D: c.4264_4273delGAGACTTCTG on gene BRCA2.

Panel E: c.5313delC on gene BRCA1.

Panel F: c.5158_5163delACCCAG on gene BRCA1.

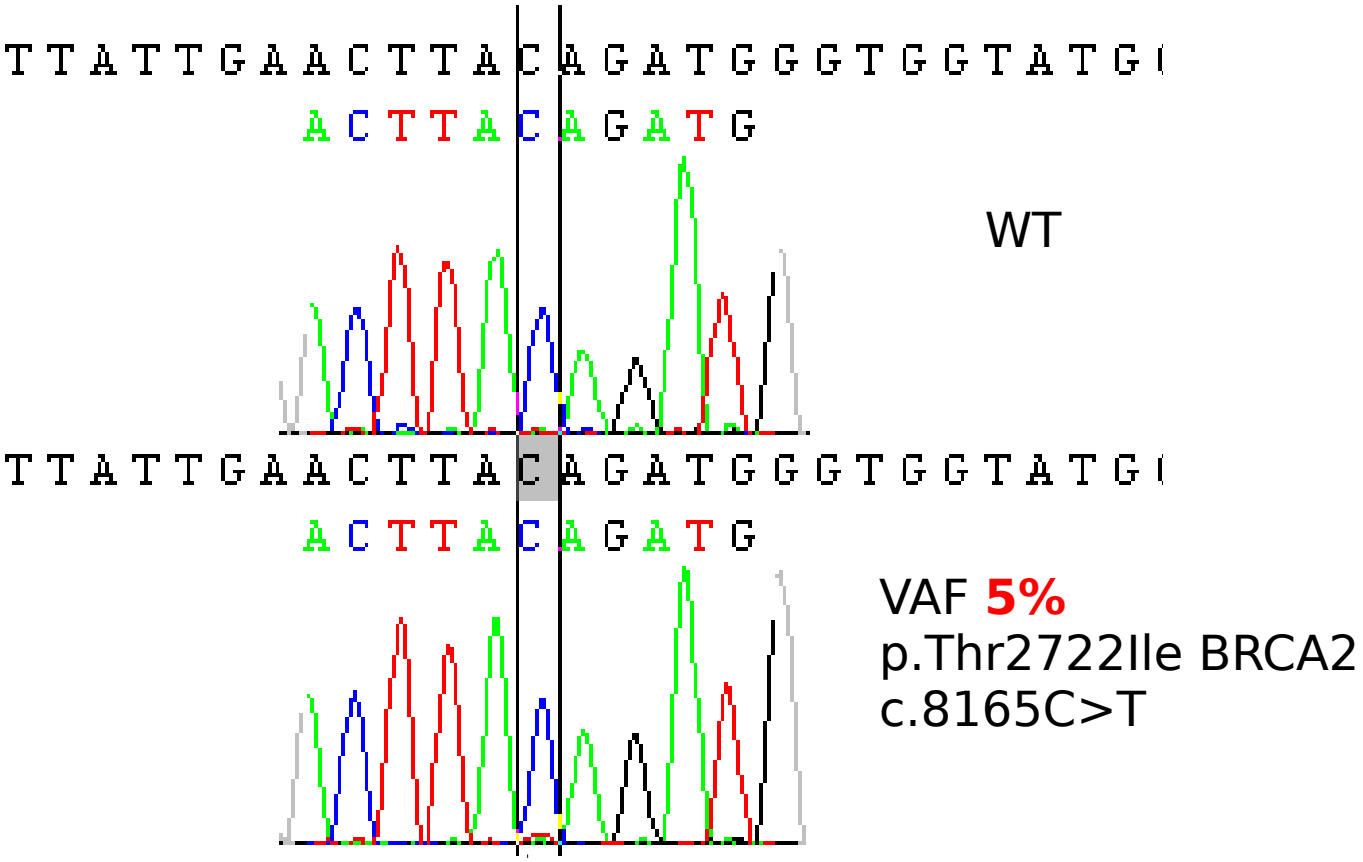


Figure 6 Panel A

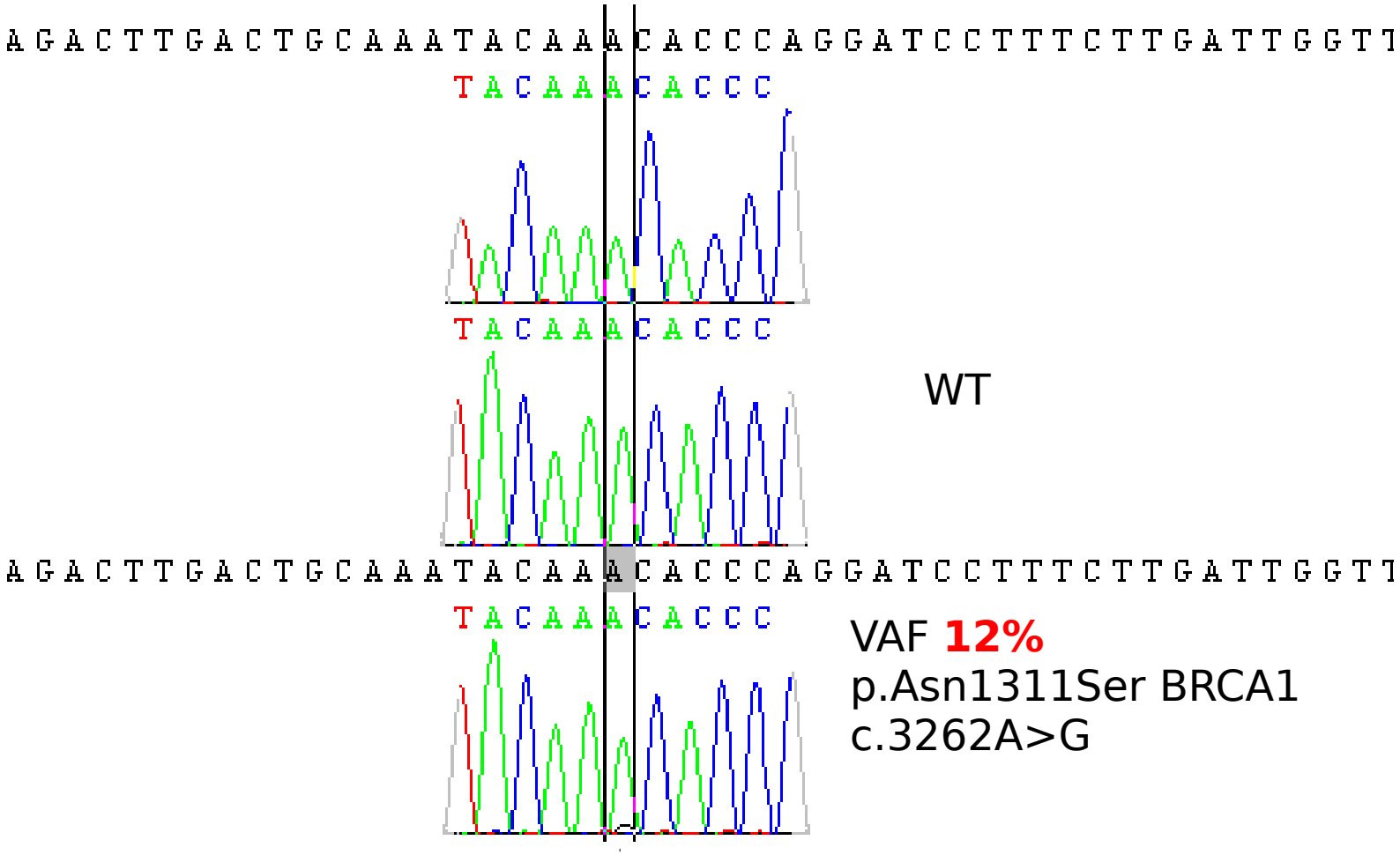


Figure 6 Panel B

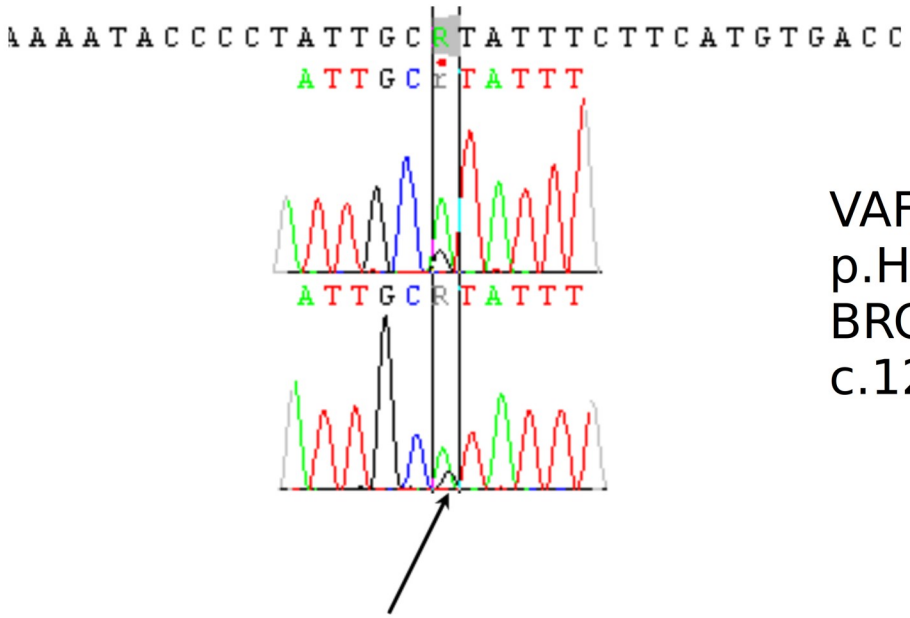


Figure 6 Panel C

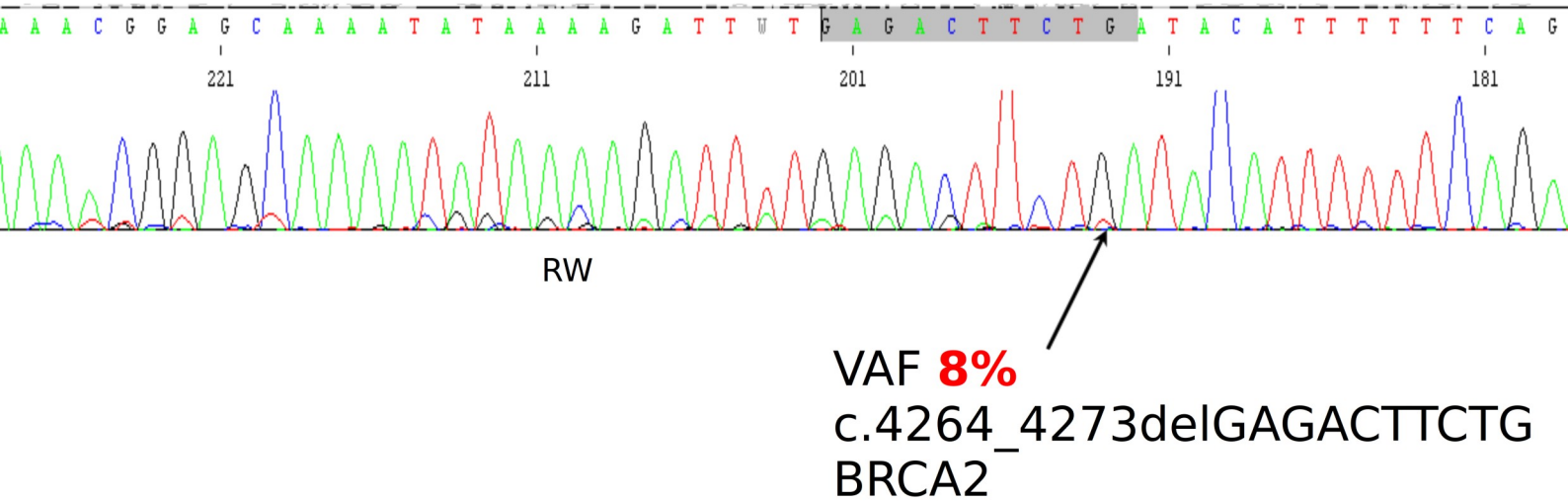
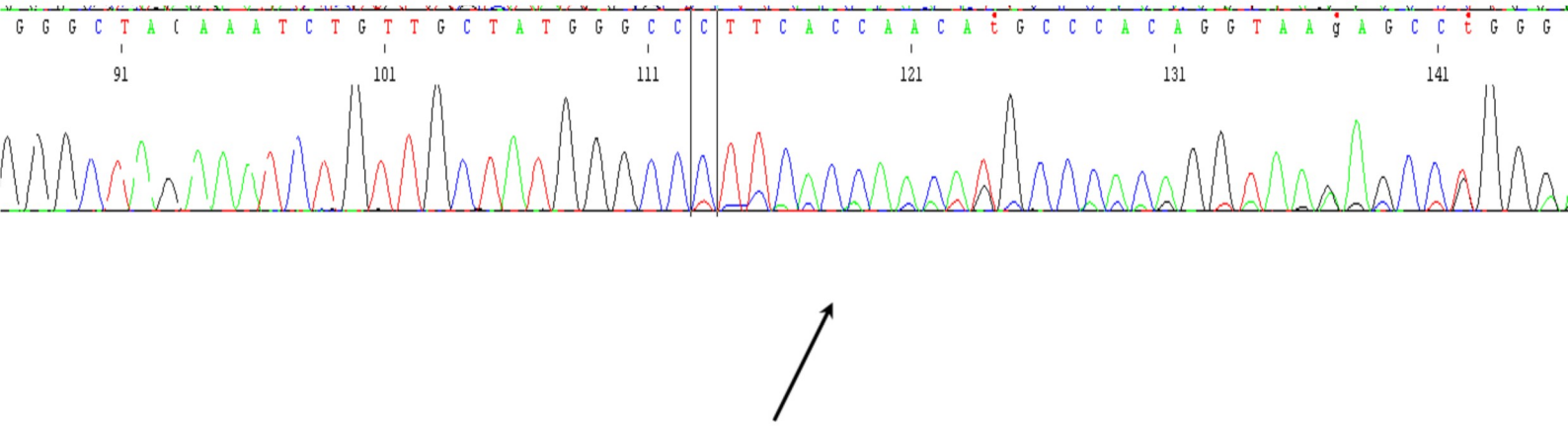
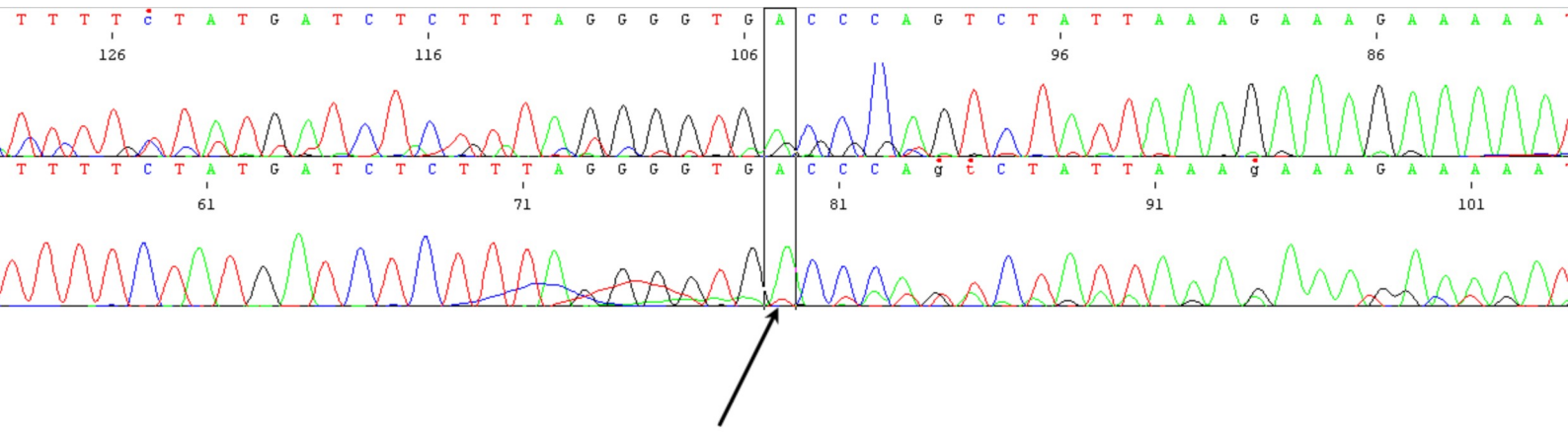


Figure 6 Panel D



VAF **19%**
BRCA1
c.5313delC

Figure 6 Panel E



VAF **21%**
BRCA1
c.5158_5163delACCCAG

Figure 6 Panel F

4.2 STAble results

We downloaded original RNA reads are from Kamke and colleagues of HMY and LMY bacteria [96]. LMY and HMY RNAs have been separately reconstructed with STAble. Reconstructed transcripts have been then aligned using BLAST algorithm against our custom reference genes as described in “Materials and Methods” section. The original reads of the samples have then been aligned to the reconstructed transcripts to estimate the abundance of each reconstructed transcript as explained in Materials and Methods section. The data obtained from this analysis flowed into two distinct contingency tables, one for HMY and one for LMY, that have been the input files for the last module of STAble: the flux-balance analysis.

Table 3 and **Table 4** show a sum up the features of the samples used in our study. The tables contain the number of the reads of each sample, the number of unused reads for each sample, i.e. the reads that have been not used for the reconstruction of any transcript, the number of the reconstructed transcripts for each sample and the number of the reads associated to the mRNAs. Unused reads have been aligned later against Blast NCBI database and we verified that they were bacterial rRNA sequence. That could be explained with consideration that samples had not been well cleaned from rRNA sequences.

Figure 7 shows a scheme of the modules whom STAble is composed. Module A has been written before the start of this study while modules B and C have been written as part of this research project. Module C has been written in collaboration with Cambridge University.

Figure 8 shows a graphic example of a pathway that we reconstructed in our study, particularly the pathway of Butanoate Metabolism. We mapped the raw reads count associated with each gene of the pathway, coloring in red the genes more expressed in LMY animals and in green the genes more expressed in HMY animals. These raw data seems to indicate the same results suggested by Kamke and colleagues: the

formation of butyrate from succinate is more abundant in LMY animals or negatively correlated to methane yield animals as suggested by Kamke and colleagues [96].

The raw reads count associated to each reconstructed transcripts, of each pathway, was the input of the module C, the flux-balance analysis module.

Table 5 and **Table 6** showed the results of the last module of STAble, the one which predict metabolic switches after the reconstruction of a dataset of transcripts. Original RNA reads are from Kamke and colleagues of HMY and LMY bacteria [96]. Results obtained from our metabolic network analysis are consistent with data about differences in usage of Glycolysis/Gluconeogenesis and Butanoate Biosynthesis pathways described in the paper (data not shown). Interestingly our analysis identified new pathways that are independent from the original set of transcripts used to feed the metabolic model network. Indeed, our metabolic network analysis identified that both in LMY and HMY bacteria, transport channels are highly expressed. Moreover, the performed analysis revealed carbohydrate metabolism as dominating followed by amino acid metabolism, results in agreement with those reported by Hinsu and colleagues that described functionally active bacteria and their biological processes in rumen of buffalo (*Bubalus bubalis*) adapted to different dietary treatments [99].

Table 3: Data of the original number of reads, the reads discarded by the assembler and the number of reconstructed transcripts for LMY animals.

Table 3

Sample	Total N° Reads	N°_Unused_Reads	N° Reconstructed mRNAs	N°_Reads_Associated_to_mRNAs
SRR873454_LOW	29084026	13669375	615298	10555304
SRR873451_LOW	33318400	13114978	640845	14678617
SRR873453_LOW	45075982	18756225	795115	18788929

Table 4: Data of the original number of reads, the reads discarded by the assembler and the number of reconstructed transcripts for HMY animals.

Table 4

Sample	Total N° Reads	N°_Unused_Reads	N° Reconstructed mRNAs	N°_Reads_Associated_to_mRNAs
SRR873461_HIGH	33752384	18069111	673936	11360308
SRR873463_HIGH	26822560	14526596	560288	8166309
SRR1206249_HIGH	31366260	15950930	650016	10646926

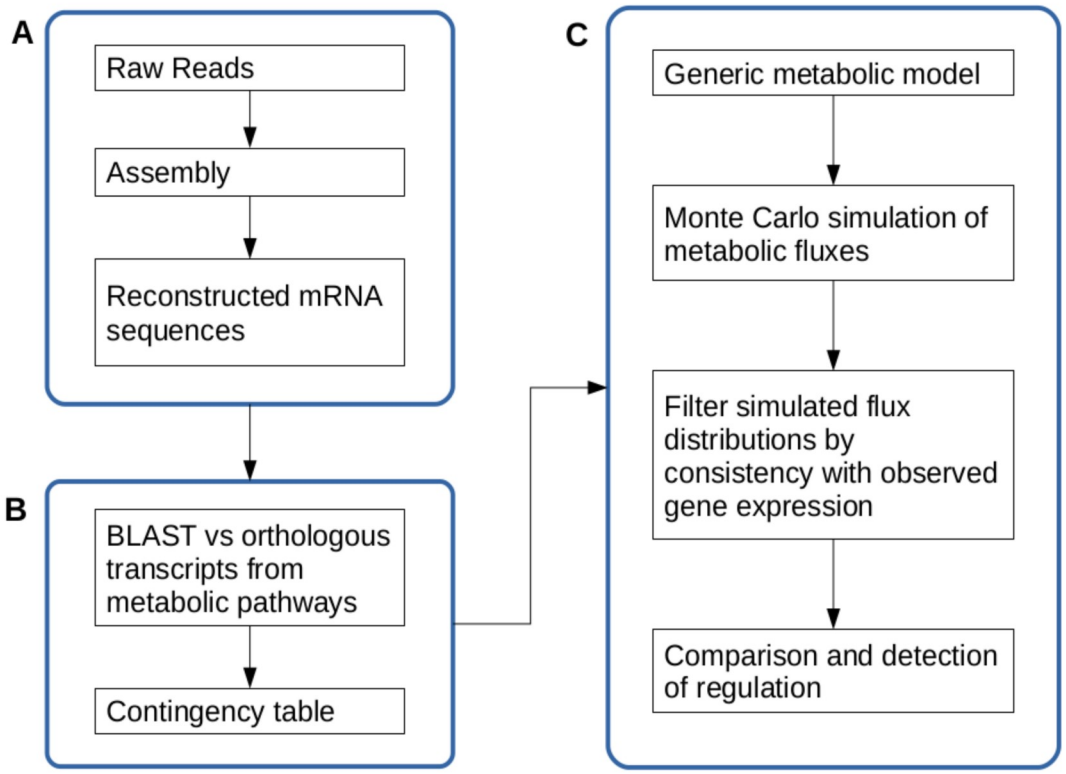


Figure 7: Scheme of STAble final workflow

A: First core STAble module of reconstruction of transcripts, mainly written in C. This module has already been developed in our lab before the beginning of this study.

B: Module written in Perl. In which reconstructed transcripts were assigned to orthologous transcripts included in several metabolic pathways.

C: Metabolic flux module. A flux balance analysis is performed starting from the contingency table that contains the reconstructed transcripts blasted against their orthologous genes. Output is a list of genes predicted to be up or down regulated.

A and B module have been developed by our laboratory, while C module has been developed mainly by Cambridge University.

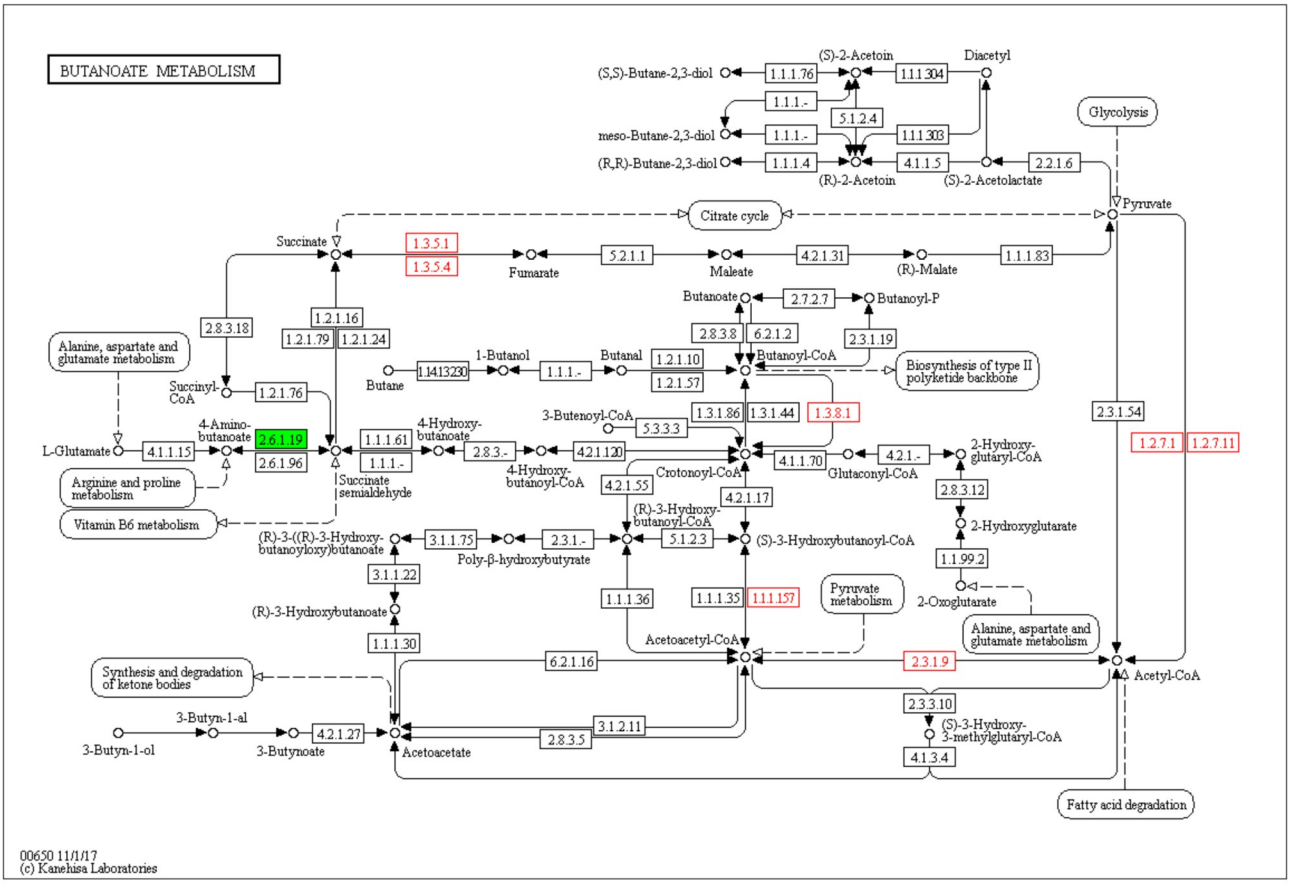


Figure 8: This figure shows the reconstructed transcripts in our LMY samples (red) and our HMY samples (green).

Table 5: List of all bacterial metabolic reactions identified in high methane yield animals.

Table 6: List of all bacterial metabolic reactions identified in low methane yield animals.

Table 5

abbreviation	Subsystem	officialName
NADH16pp	Oxidative Phosphorylation	NADH dehydrogenase (ubiquinone-8 & 3 protons) (periplasm)
PROt2rpp	Transport	L-proline reversible transport via proton symport (periplasm)
PROt4pp	Transport	Na+/Proline-L symporter (periplasm)
GLCP2	Glycolysis/Gluconeogenesis	glycogen phosphorylase
GLCS1	Glycolysis/Gluconeogenesis	glycogen synthase (ADPGlc)
GLGC	Glycolysis/Gluconeogenesis	glucose-1-phosphate adenylyltransferase
THRt2rpp	Transport	L-threonine reversible transport via proton symport (periplasm)
THRt4pp	Transport	L-threonine via sodium symport (periplasm)
INSt2pp	Transport	inosine transport in via proton symport (periplasm)
INSt2rpp	Transport	inosine transport in via proton symport reversible (periplasm)
PPCSCT	Alternate Carbon Metabolism	Propanoyl-CoA: succinate CoA-transferase
SUCOAS	Citric Acid Cycle	succinyl-CoA synthetase (ADP-forming)
TALA	Pentose Phosphate Pathway	transaldolase
ACCOAL	Alternate Carbon Metabolism	acetate-CoA ligase (ADP-forming)
GLUt4pp	Transport	Na+/glutamate symport (periplasm)
PPAKr	Alternate Carbon Metabolism	Propionate kinase
PTA2	Alternate Carbon Metabolism	Phosphate acetyltransferase
THFAT	Folate Metabolism	Tetrahydrofolate aminomethyltransferase
FOMETRi	Folate Metabolism	Aminomethyltransferase
ADK3	Nucleotide Salvage Pathway	adentylate kinase (GTP)
FBA3	Pentose Phosphate Pathway	7-bisphosphate D-glyceraldehyde-3-phosphate-lyase
PFK_3	Pentose Phosphate Pathway	phosphofructokinase (s7p)
URAt2pp	Transport	uracil transport in via proton symport (periplasm)
URAt2rpp	Transport	uracil transport in via proton symport reversible (periplasm)
GLYt2pp	Transport	glycine transport in via proton symport (periplasm)
GLCP	Glycolysis/Gluconeogenesis	glycogen phosphorylase
NDPK1	Nucleotide Salvage Pathway	nucleoside-diphosphate kinase (ATP:GDP)
CA2t3pp	Inorganic Ion Transport and Metabolism	calcium (Ca+2) transport out via proton antiport (periplasm)
CAt6pp	Inorganic Ion Transport and Metabolism	calcium / sodium antiporter (1:1)
PPKr	Oxidative Phosphorylation	polyphosphate kinase
URIt2pp	Transport	uridine transport in via proton symport (periplasm)
URIt2rpp	Transport	uridine transport in via proton symport reversible (periplasm)
NADH18pp	Oxidative Phosphorylation	NADH dehydrogenase (demethylmenaquinone-8 & 3 protons) (periplasm)
FRD3	Citric Acid Cycle	fumarate reductase
ALAt2pp	Transport	L-alanine transport in via proton symport (periplasm)
ALAt2rpp	Transport	L-alanine reversible transport via proton symport (periplasm)
GLYt2rpp	Transport	glycine reversible transport via proton symport (periplasm)

Table 6

abbreviation	Subsystem	officialName
ALATA_L	Alanine and Aspartate Metabolism	L-alanine transaminase
THMDt2pp	Transport	thymidine transport in via proton symport (periplasm)
THMDt2rpp	Transport	thymidine transport in via proton symport reversible (periplasm)
NAt3pp	Inorganic Ion Transport and Metabolism	sodium transport out via proton antiport (cytoplasm to periplasm)
VPAMTr	Valine, Leucine and Isoleucine Metabolism	Valine-pyruvate aminotransferase
VALTA	Valine, Leucine and Isoleucine Metabolism	valine transaminase
SUCDi	Oxidative Phosphorylation	succinate dehydrogenase (irreversible)
GLUABUTt7pp	Transport	4-aminobutyrate/glutamate antiport (periplasm)
ABUTt2pp	Transport	4-aminobutyrate transport in via proton symport (periplasm)
GLYt4pp	Transport	glycine transport in via sodium symport (periplasm)
GLUt2rpp	Transport	L-glutamate transport via proton symport reversible (periplasm)
GLDBRAN2	Glycolysis/Gluconeogenesis	glycogen debranching enzyme (bglycogen -> glycogen)
GLYCLTt2rpp	Transport	glycolate transport via proton symport
GLYCLTt4pp	Transport	glycolate transport via sodium symport (periplasm)
ACt2rpp	Transport	acetate reversible transport via proton symport (periplasm)
ACt4pp	Transport	Na+/Acetate symport (periplasm)
ADK1	Nucleotide Salvage Pathway	adenylate kinase
PTAr	Pyruvate Metabolism	phosphotransacetylase
ACKr	Pyruvate Metabolism	acetate kinase
ACS	Pyruvate Metabolism	acetyl-CoA synthetase
SERt2rpp	Transport	L-serine reversible transport via proton symport (periplasm)
SERt4pp	Transport	L-serine via sodium symport (periplasm)
GLCtex	Transport	glucose transport via diffusion (extracellular to periplasm)
PRPPS	Histidine Metabolism	phosphoribosylpyrophosphate synthetase
PPM	Alternate Carbon Metabolism	phosphopentomutase
R15BPK	Alternate Carbon Metabolism	Ribose-1,5 bisphosphokinase
R1PK	Alternate Carbon Metabolism	ribose 1-phosphokinase
GLCtexi	Transport	D-glucose transport via diffusion (extracellular to periplasm) irreversible
ADNt2pp	Transport	adenosine transport in via proton symport (periplasm)
ADNt2rpp	Transport	adenosine transport in via proton symport reversible (periplasm)
ASPt2pp	Transport	L-aspartate transport in via proton symport (periplasm)
ASPt2rpp	Transport	L-aspartate transport in via proton symport (periplasm) reversible
INDOLEt2pp	Transport	Indole transport via proton symport irreversible (periplasm)
INDOLEt2rpp	Transport	Indole transport via proton symport reversible (periplasm)
FBA	Glycolysis/Gluconeogenesis	fructose-bisphosphate aldolase
PFK	Glycolysis/Gluconeogenesis	phosphofructokinase
ICHORS	Cofactor and Prosthetic Group Biosynthesis	isochorismate synthase
ICHORSi	Cofactor and Prosthetic Group Biosynthesis	Isochorismate Synthase
HPYRI	Alternate Carbon Metabolism	hydroxypyruvate isomerase
HPYRRx	Alternate Carbon Metabolism	Hydroxypyruvate reductase (NADH)
TRSARr	Alternate Carbon Metabolism	tartronate semialdehyde reductase
CYTDt2pp	Transport	cytidine transport in via proton symport (periplasm)
CYTDt2rpp	Transport	cytidine transport in via proton symport reversible (periplasm)
FRD2	Citric Acid Cycle	fumarate reductase
NADH17pp	Oxidative Phosphorylation	NADH dehydrogenase (menaquinone-8 & 3 protons) (periplasm)
EX_h(e)	Exchange	H+ exchange
EX_fe3(e)	Exchange	Fe3+ exchange
EX_fe2(e)	Exchange	Fe2+ exchange
Htex	Transport	proton transport via diffusion (extracellular to periplasm)
FEROpp	Inorganic Ion Transport and Metabolism	ferroxidase
FE3tex	Transport	iron (III) transport via diffusion (extracellular to periplasm)
FE2tex	Transport	iron (II) transport via diffusion (extracellular to periplasm)
GLBRAN2	Glycolysis/Gluconeogenesis	4-alpha-glucan branching enzyme (glycogen -> bglycogen)
EX_o2(e)	Exchange	O2 exchange
EX_h2o(e)	Exchange	H2O exchange
O2tex	Transport	oxygen transport via diffusion (extracellular to periplasm)
H2Otex	Transport	H2O transport via diffusion (extracellular to periplasm)
CRNDt2rpp	Transport	D-carnitine outward transport (H+ antiport)
CRNt2rpp	Transport	L-carnitine outward transport (H+ antiport)
CRNt8pp	Transport	L-carnitine/D-carnitine antiporter (periplasm)
ALAt4pp	Transport	L-alanine transport in via sodium symport (periplasm)

5.0 Discussion and conclusion

During the PhD., the research activity has been divided in two main field of analysis of NGS data: the development of a new variant calling algorithm for the analysis of somatic mutations (SNP and indels) and the analysis of RNA-Seq data. This second part ended with our publication [83] and consisted in the update of STAble, a *de novo* assembly software.

As regard the first part of the research, the variant calling algorithm has been integrated in Amplicon Suite, a software that is currently used by clinicians for the analysis of NGS data. The algorithm can perform single sample analysis, because it can calibrate itself on the base of the noise background of the single sample without the need of particular settings before running the analysis or the need of a control sample. This means that could be used on different kind of data and on different experimental condition and can potentially work on different gene panels. Here we presented only data from a single gene panel, particularly data obtained with the use of the BRCA-Devyser Kit but our laboratory is already analyzing, for example, some data from CFTR gene using the same algorithm with good results. Other software can be good for germline analysis and in WES or WGS but it is very difficult to use them for a single sample somatic variant analysis since many of them like VarScan2 need a control sample to calibrate themselves before starting the analysis; often, a control sample cannot be easily obtained in clinical conditions, or could be too much expensive to obtain and analyze it in terms of time or cost. For these reasons we developed our algorithm. First, we developed an algorithm that could eliminate the first level of noise background, and we tested it on simulated datasets generated with ART simulator. To test our algorithm and compare its performance against the other variant calling software we considered to use available datasets like for examples

those provided by the DREAM group [100], but anyone of the datasets that we found on the available databases was conform to our need. We wanted something that could represent in a very similar manner the panel genes that we needed to analyze. Because of this unmet necessity and since we wanted to have a complete control on our data, in order to perform a precise test of the algorithms and the other software we chose to simulate the data and to test the performance of the variant calling software in silico. But when we tested the algorithm on real sample we realized that applying only a first level of filter was not sufficient to detect somatic true variants, since after the application of the first level filter we found still a lot of number of variant on the samples tested, even more than one hundred per sample, a number that is not plausible from a biological point of view. This can be explained because ART when produces simulated reads, it mainly introduces the errors of the sequencing platform but cannot simulate the errors induced by the PCR in the amplicon based sequencing [21]. The error introduced by the PCR divides the noise background of SNPs in two main groups that we call transitions (T:C and G:A mutations) and transversions (other transitions plus transversions). Since these two groups of mutations shows difference noise background due to the error introduced by the Taq polymerase during the amplification step, as described in [21], we decided to divide the input variants of our variant calling algorithm in three groups when we applied it on real datasets: transitions, transversions and indels (data not shown) to take in consideration even the error introduced by the Taq polymerase. Above this correction, to further reduce the lack in terms of precision of our simulated data, we decided to insert a new level of filter on our pipeline, a filter of second level: since variant calling software filter the variants on the base of the frequency it is impossible for them to filter variants with an high frequency that could be, however, the product of a contamination. Indeed, if a variant is present in an high number of samples always at the same frequency is likely an error, due to a contamination or a PCR error, or due to other technical problems occurring in the phase of preparation of the library, since it

is extremely unlikely that a somatic variant, that is basically a rare event, is present in many samples always at the same frequency. The criteria we used to filter variants in the second step of filtering are the followings: variants that are present in more than 5% of the samples always at the same frequency are filtered out because they are considered part of the noise background, while variants that are present in more than 5% of the samples that have got a frequency equal or higher to the average of frequency among the other samples plus 3 times the standard deviation are considered interesting and are not filtered out.
In the final report of our pipeline the clinician can examine each variant that pass the filters and can choose which variants are the most interesting from their clinical point of view. The clinical has got a full control of the variants discarded from the second filter, indeed the option of *Variant Distribution* permits to examine the presence of that variant among the samples in our database with a bar chart in which on the x axis there is the frequency in percentage, and on y axis there is the number of samples in which that variant is present. This information could admit to bypass the second level of noise background that other variant calling software cannot point out.

In the second part of this research work we focus on the analysis of RNA-Seq data. This second part ended with our publication [83] and consisted in the update of STAble, a *de novo* assembly software. STAble is born as a new standalone RNA-seq *de novo* assembler since our group decided, prior to the start of this research project, to develop a new tool for the analysis of RNA-seq data, developing the first module of STAble. This module is based on the principle of head-tail alignments between reads and the construction and the traversal of a graph in order to reconstruct full length transcripts, as described in [83]. But then we evaluated that the reconstruction of full length transcripts was not enough informative from a biological point of view. We decided to add a new part in our workflow of analysis that uses the information given by transcripts expressed by microbiota to build-up a metabolic network. This

last module is based on a flux-balance analysis approach and has been developed by Pietro Liò and his group in Cambridge University. More details about this module can be found in [98]. The aim of this flux-balance analysis is to identify switches in metabolic pathways of the microbial population in the sample in exam from the normal metabolic flux that is represented by a model organism. In our test we used the normal metabolic flux of *E. Coli* as model. The output of the analysis is a list of predicted altered, downregulated or upregulated genes, grouped in pathways. This approach admits to analyze deeply the flux of metabolism of an entire microbiota and opens the theoretical possibility to intervene not balancing the bacterial population with physiological species to restore the healthy condition, like happens in many therapeutic approach, but particularly in adopting a specific drug that target the altered gene or genes, or pathway that are upregulated or downregulated in that patient and in that specific moment. This idea, that in this moment is mainly experimental, could contribute to find new clinical approaches to treat dysbiosis. Although we tested STAble on different datasets of human and microbiota, for the moment, the complete pipeline of analysis that includes the flux-balance analysis module has been tested only on a single RNA-seq dataset, the one of Kamke and colleagues [96]. We wanted to test it on more datasets but even if many RNA-seq datasets from human microbiota can be available, we needed to find one dataset that were been extensively analyzed, even with transcript abundancy, to have a good quality benchmark dataset to test our second module to verify the performance of our system. Our specific work in this project, above testing again the first module and optimize some parameters, was to reconstruct the transcripts from RNA-seq data of Kamke and colleagues with the first module of STAble, align the reconstructed transcripts against a list of orthologous genes of several bacterial metabolic pathways downloaded from KEGG database and build-up a contingency table with the read counts associated with each ortologhous gene that is the input file for the flux-balance analysis module, as described in [83]. After the tests, we found that our results were

comparable with ones declared by Kamke and colleagues (data not shown), but our work predicted that in the samples of Kamke and colleagues were expressed new pathways and proteins like transport membrane proteins and membrane channels, which Kamke and colleagues did not find in their analysis. Since our system predict the same results that Kamke and colleagues validated in their experiments, but also predicted the alteration of the metabolic fluxes of different pathways and the alteration of expression of different proteins, these results, even if they are only preliminary, and would need to be experimentally validated, leaded us to the conclusion that STAble could be an optimal instrument to analyze switches in the metabolism of microbiota and an optimal instrument to predict the expression of new proteins and reactions not directly inferable by raw metatranscriptomic data.

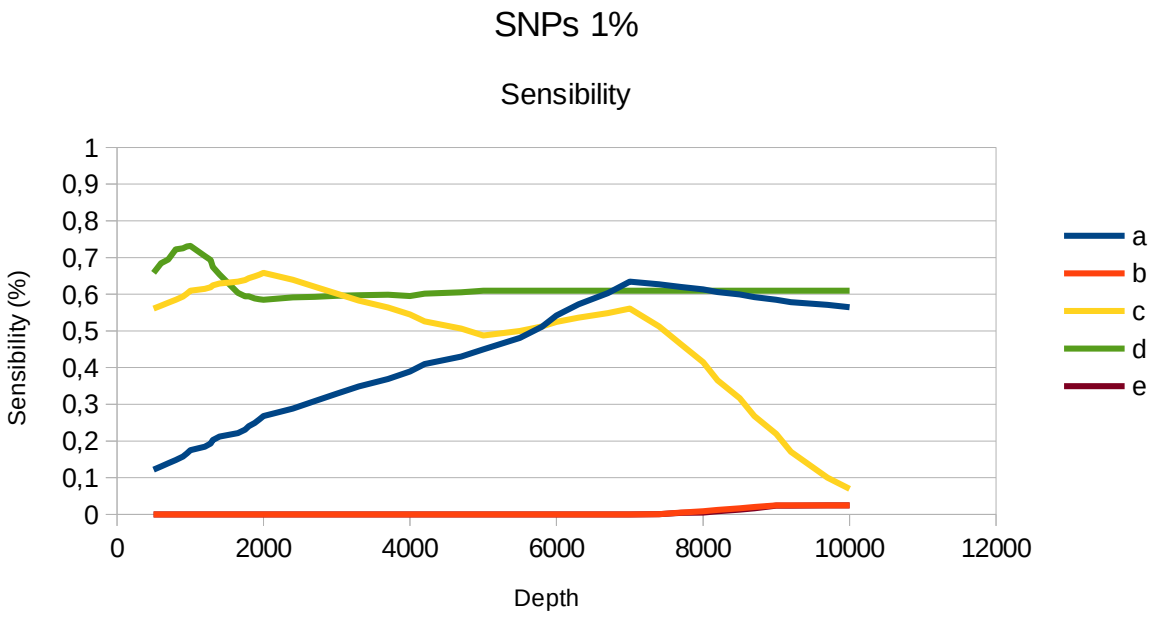
Above all, the main output data of this last module is not only a list of single genes up or down-regulated, but also a group of pathways that can be more easily correlated with a biological meaning than a list of isolated genes. This could help to build-up a different approach in the future, even in clinical applications, for the treatment of dysbiosis.

6.0 Appendix A

Prior to the development of our variant calling algorithm, we tested the other variant calling software on simulated data, and we evaluated if they were able to identify both SNPs and indels on our simulated datasets calculating sensibility and precision for each condition. Results for SNPs are shown in **Figures a-b-c,** results for indels are presented in **Figures d-e-f.** Software like SomaticSniper and Mutect are not able to identify indels. GATK either could not identify indels when we tested it on our datasets.

GATK obtained bad results at each level of depth or at each level of mutation frequency. SomaticSniper generally shows a low level of precision which increases

with raising of read depth. Mutect works well at medium-high level of depth, and it can identify low frequency somatic mutations in SNPs with 1% and 3% of frequency at different read depths. VarDict works more or less with the same performance of Mutect. Mutect and SomaticSniper cannot call indel mutations. Generally, these software show irregular and unstable results, particularly VarScan2 which sensibility generally drops at different read depths. Also, in several point conditions during our tests their precision and sensitivity is zero because they could not call any variant in that experimental condition.



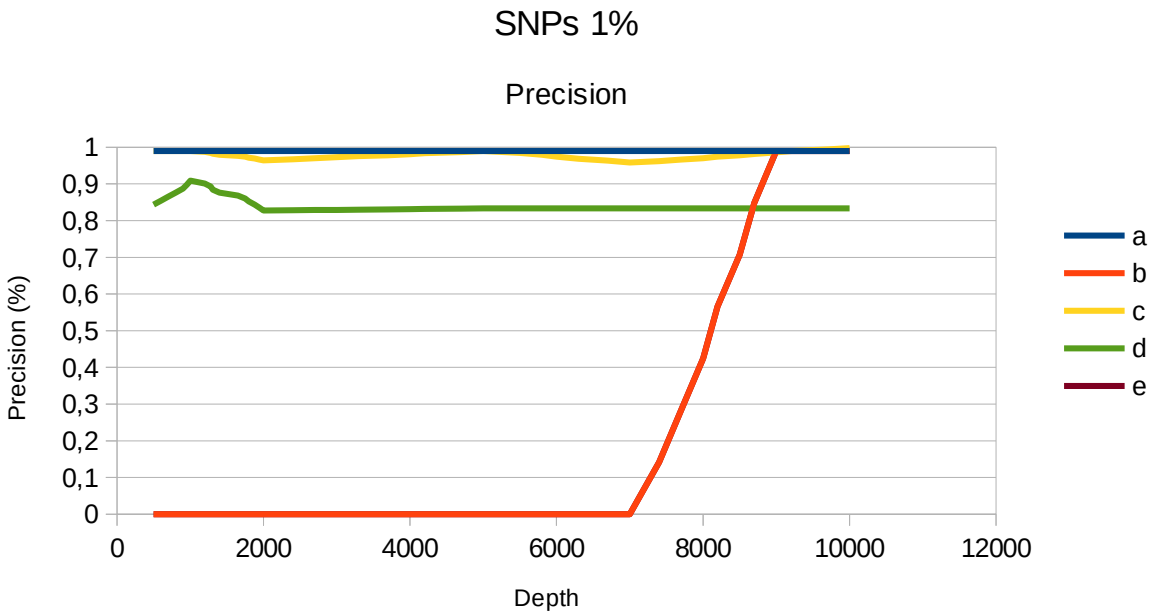
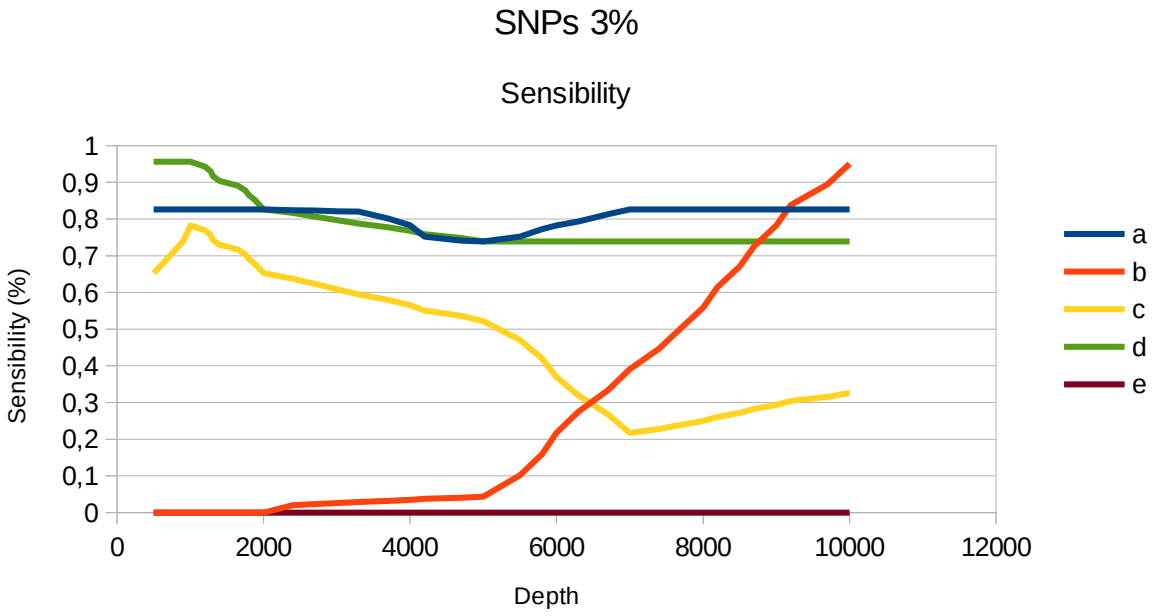


Figure a: Performance of the software in analyzing SNPs at 1%. a=VarDict; b=SomaticSniper; c=VarScan2; d=Mutect; e=GATK.
VarScan2 sensibility drops at high values of read depth.



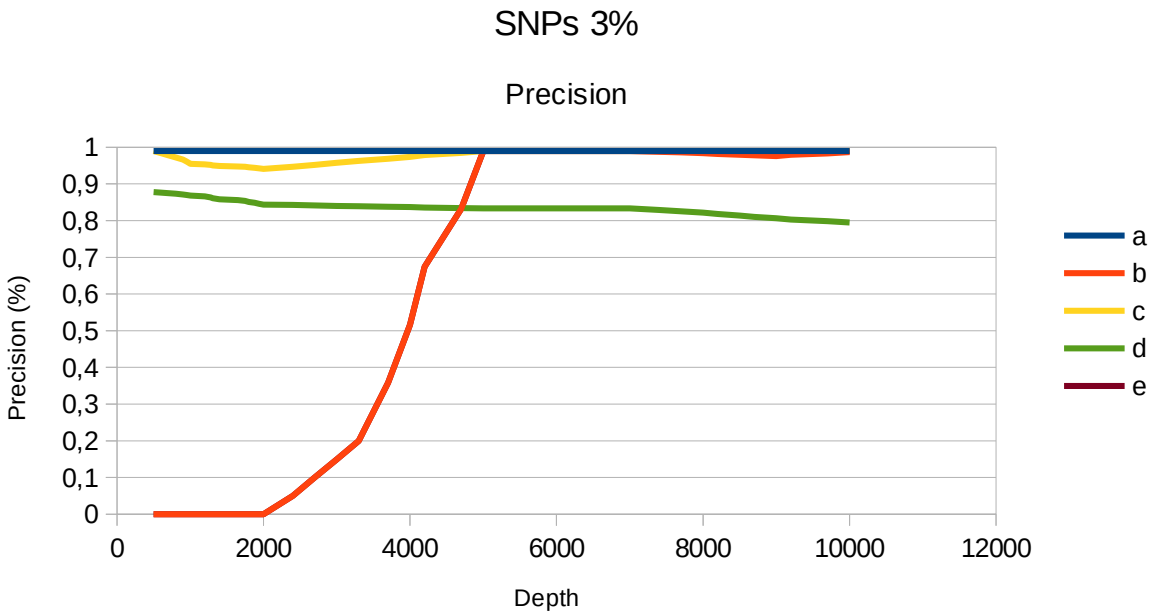
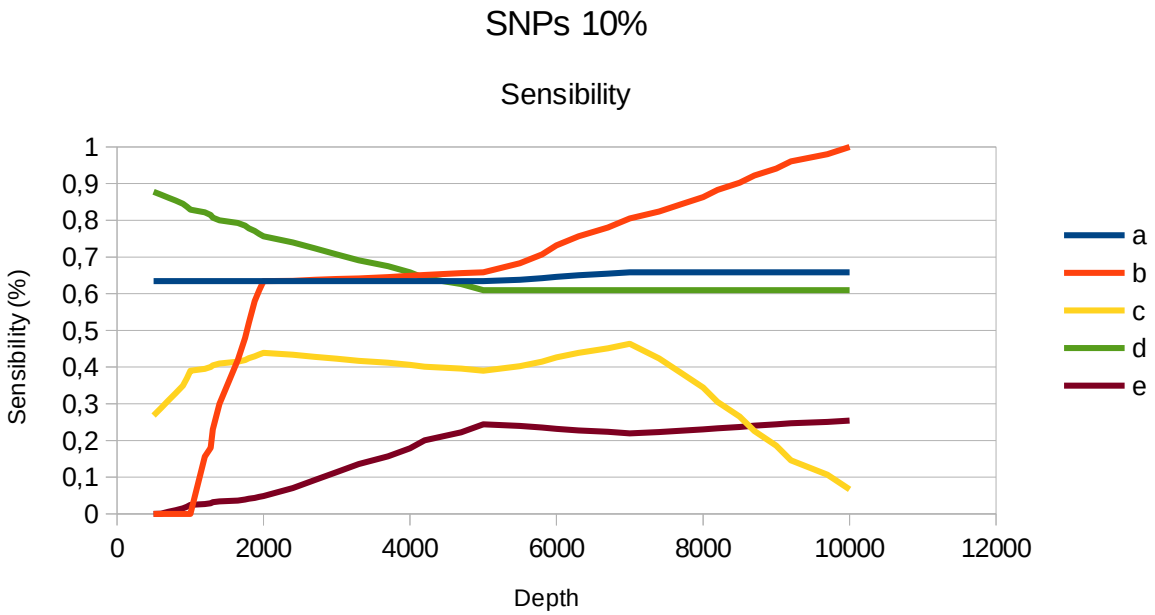


Figure b: Performance of the software in analyzing SNPs at 3%. a=VarDict; b=SomaticSniper; c=VarScan2; d=Mutect; e=GATK.
In this conditions VarScan2 sensibility shows the most unstable result dropping at 1000 of read depth but increasing at very high level of read depth.



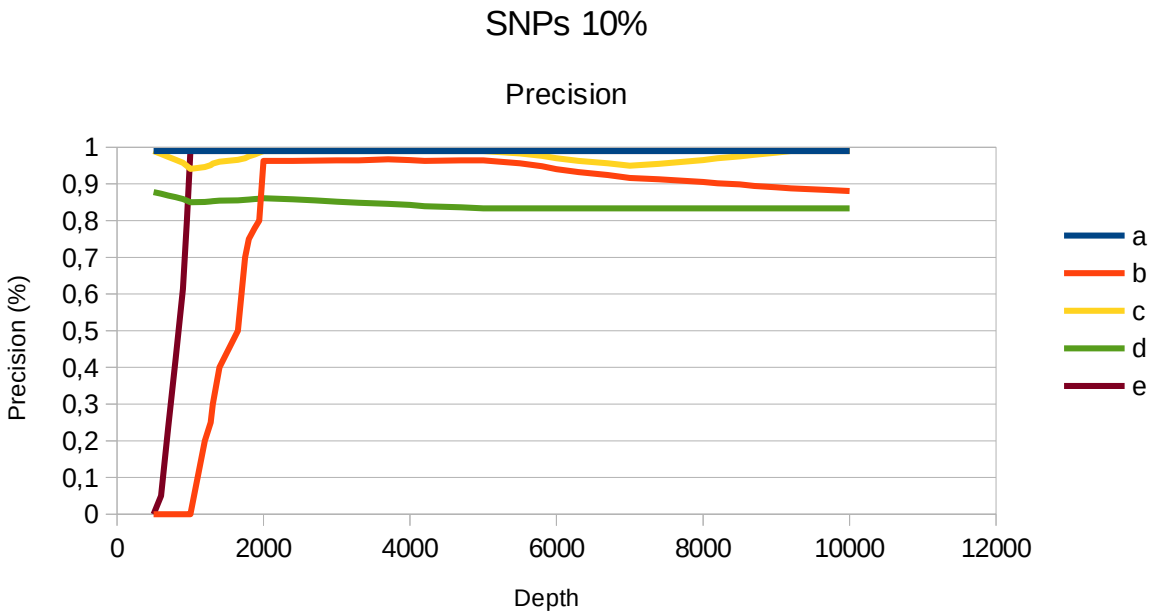
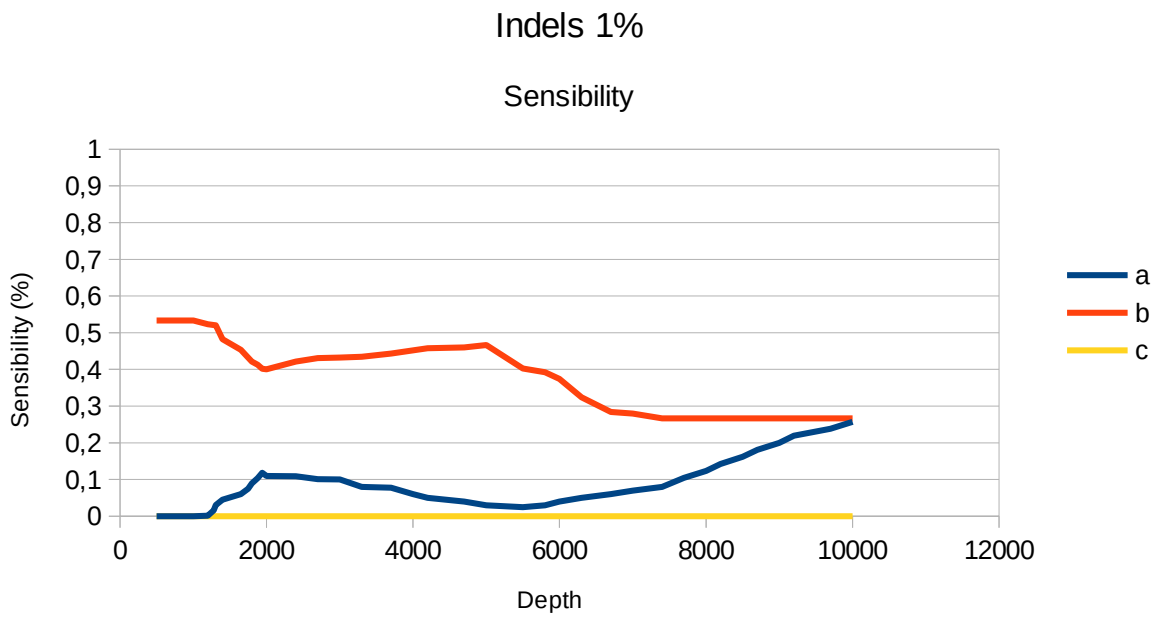


Figure c: Performance of the software in analyzing SNPs at 10%. a=VarDict; b=SomaticSniper; c=VarScan2; d=Mutect; e=GATK.
VarScan2 sensibility drops again at 7000 of read depth while other software reach a plateau or a more defined trend.



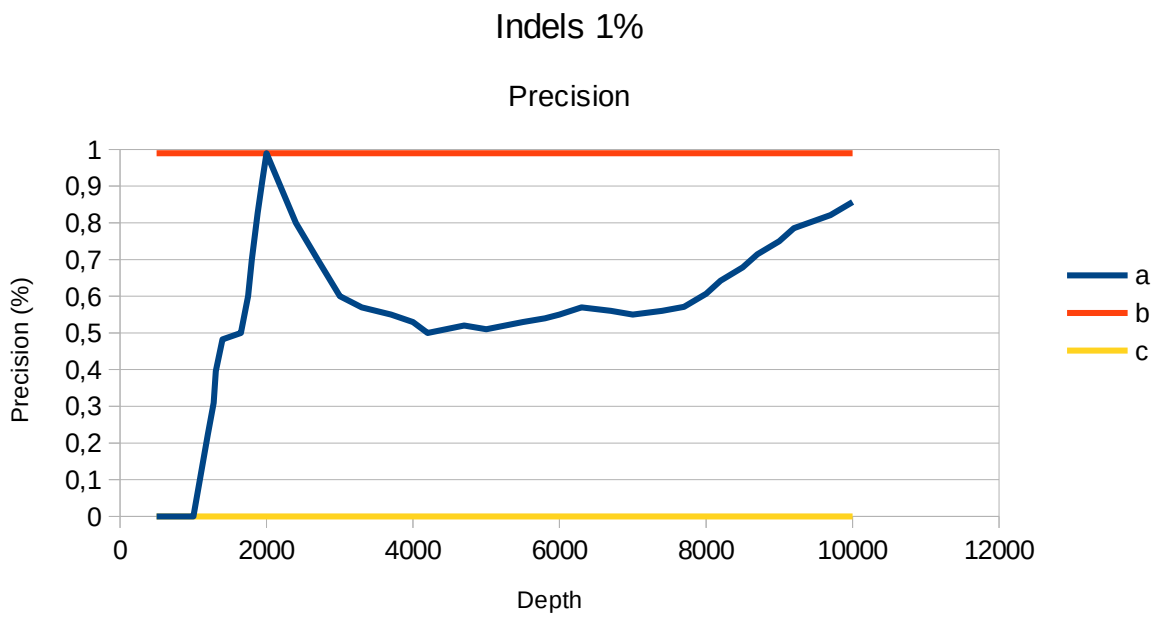
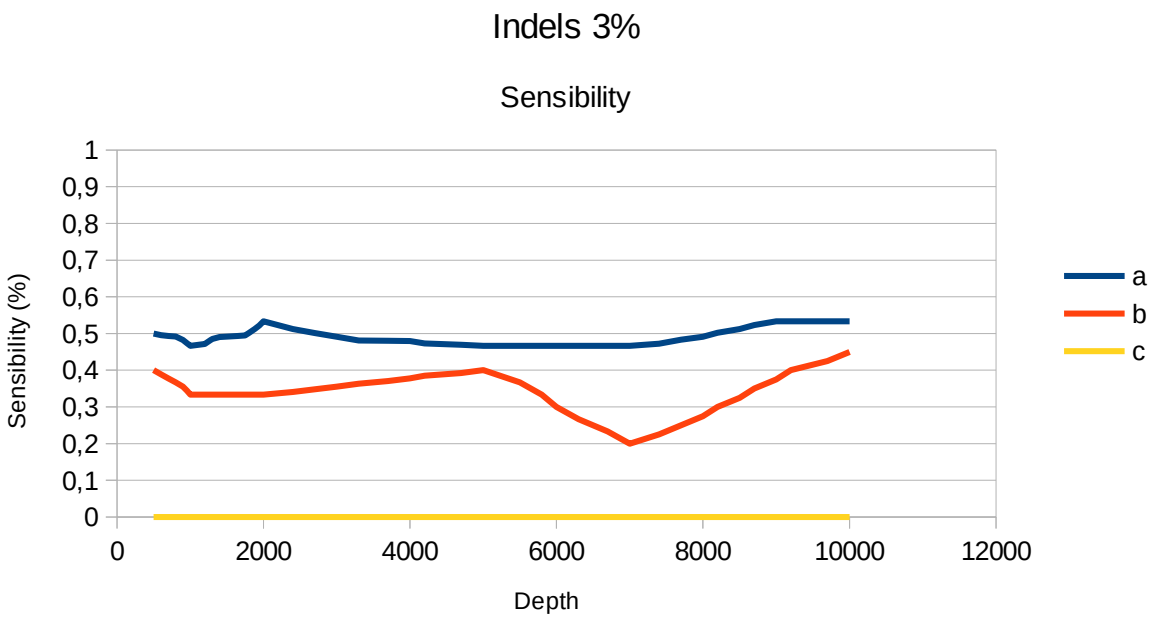


Figure d: Performance of the software in analyzing indels at 1%. a=VarDict; b=VarScan2; c=GATK.
VarDict shows very poor performance in terms of sensibility while its precision performance is not stationary.



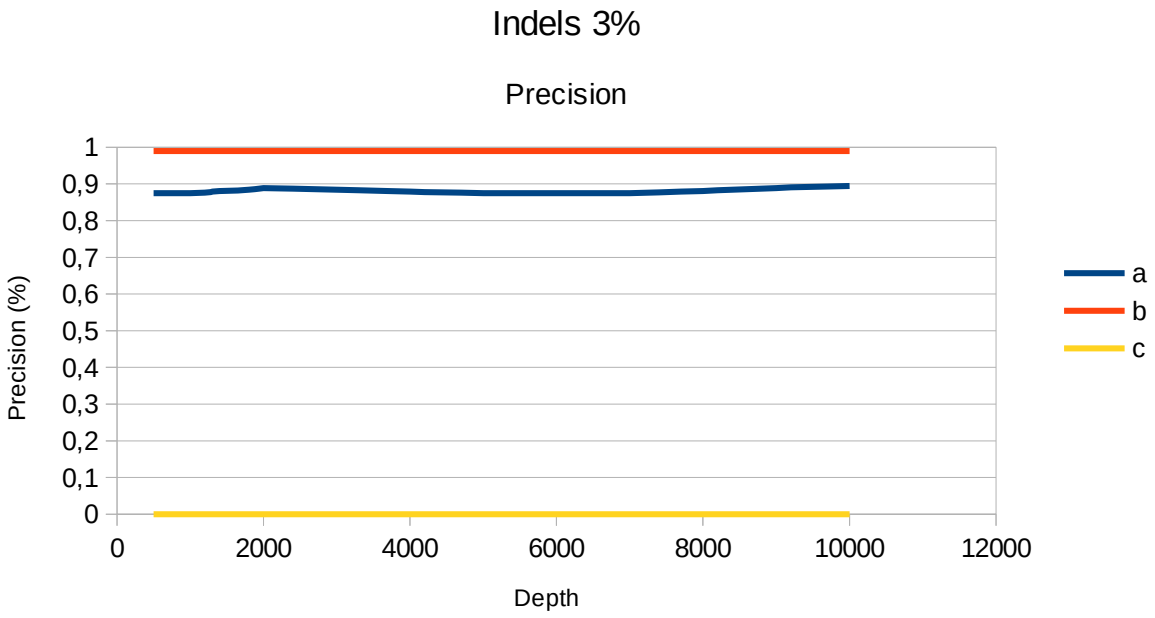
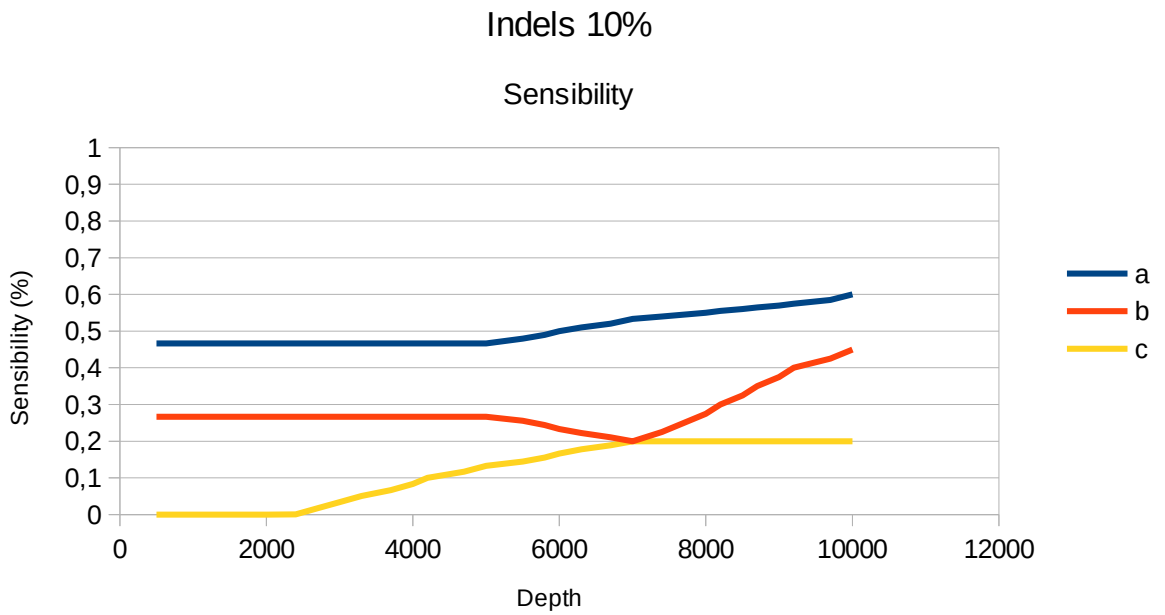


Figure e: Performance of the software in analyzing indels at 3%. a=VarDict; b=VarScan2; c=GATK.



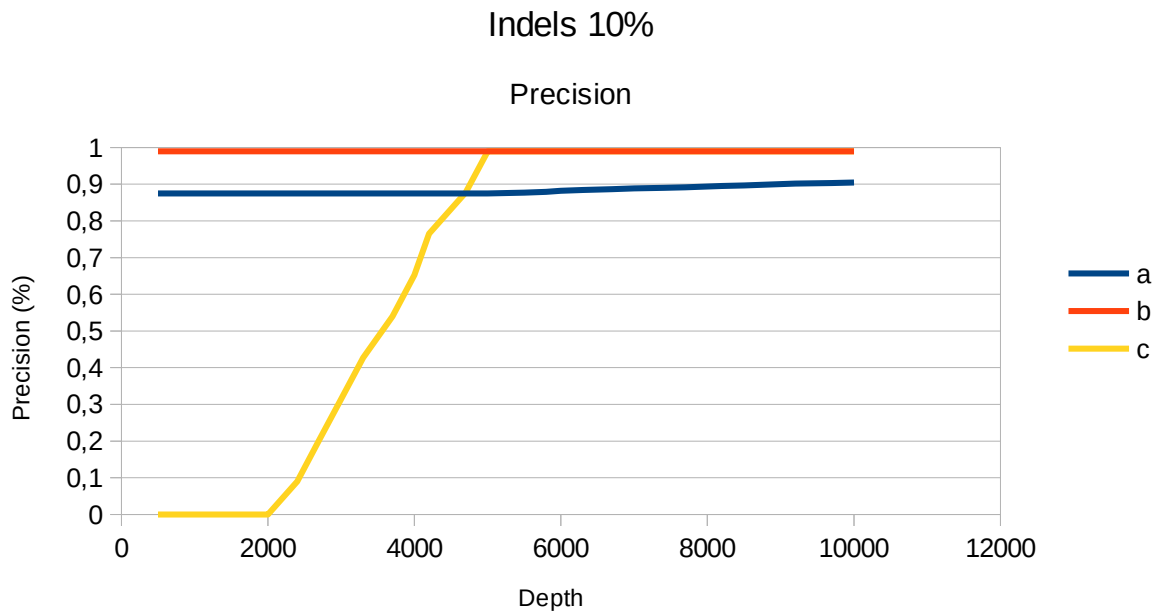


Figure f: Performance of the software in analyzing indels at 10%. a=VarDict; b=VarScan2; c=GATK.

7.0 Appendix B

STAble Benchmarks data

STAble has been tested on simulated datasets and has been optimized for working both on bacterial transciptome but also on human genomes and to reconstruct even transcripts with splicing variants.

The original work of STAble consisted in validate the performance of STAble and comparing it against the other available *de novo* assembly software on simulated data because it was necessary to know unambiguously what was true and what was false in benchmark tests. With real data the correctness of reconstructions has been usually assessed by the alignment of the reconstructed transcripts to a genome or to a database of known transcripts, so as long as reconstructions were compatible with the reference were considered as true. With simulated datasets instead, more hard filters have been applied by accepting as true only those transcripts that were effectively present in database used for simulation: this opportunity allowed to characterize a new type of false positive that has been called False positive of class A (FPA): FPA group contains the reconstructions that are compatible with genome but do not correspond to any of the transcripts used to perform sequencing simulation. Chimeric reconstructions that are not compatible with genome instead have been called False positive of class B (FPB).
A total of 4 datasets have been simulated to test STAble performance against Trinity, Oases and Bridger using ART and simulating Illumina reads of 150 bp with 20x of coverage, single end, platform HiSeq 2500. Reconstructed transcript have been aligned to reference of really expressed genes using BLASTN [101]. Only the transcripts that were fully included in the reference sequences have been accepted as true positive reconstruction. If a reference sequence was reconstructed of at least the

90% of its own length, it was labeled as full reconstructed. GMAP have also been used to distinguish between FPA and FPB: false positive reconstructions that were compatible with genome were labeled as FPA, else they were labeled as FPB.
The results of the comparison between STAble, Bridger and Oases in reconstructing the simulated datasets are shown in Table A and Table B.
Reconstructing dataset A Oases showed the highest sensibility but also the lowest precision since it showed the highest number of false positives. On the other hand STAble showed a good sensibility comparable to other software but producing the lowest number of false positives, only three transcripts. The same behaviour of STAble can be seen on dataset B. However, in this case Oases and Trinity showed a slightly higher number of transcripts reconstructed at 70% and 100% than STAble, but at the same time they also showed the highest rate of false positives.
Since datasets C and D were produced using bacterial transcripts we could not produce FPA results since splicing is not present in bacterial genomes. Anyway in dataset C STAble showed the highest sensibility while minimizing false positive ratio.
Results on dataset D underlines the notable technical features of STAble: it run with only 8 GB of RAM and completed the assembly work. While, analyzing the same dataset, even on a computer equipped with 48 GB of RAM other assemblers could not complete the assembly.

Table A: Results on 200 (Dataset A) and 6309 (Dataset B) random human transcripts. STAble returned the most reliable set of results showing a sensibility comparable to other assemblers while producing only 3 false positives.

Table B: 11815 (dataset C) and 43578 (dataset D) mixed bacterial transcripts. STAble shown the best sensibility while producing the lowest false positive ratio alongside with Trinity. Due to absence of alternative splicing in bacterial transcriptome it is not

possible to produce FPA class errors. With dataset D it has been not possible to compare results of STAble with existing assemblers as they terminated with an out of memory error.

Legend:
Assembler: Name of the assembler.
# of results: Total number of reconstructed transcripts.
# of FP: Number of False Positive results.
FPA: False Positive class A.
FPB: False Positive class B.
100%: Number of full reconstructed transcripts.
70%: Number of transcripts reconstructed at 70%.
S100: Percentage of full reconstructed transcripts.
S70: Percentage of transcripts reconstructed at 70%.
FPR: False Positive Ratio.

Table A

Dataset A

Assembler	# of results	# of FP	FPA	FPB	100%	70%	S100	S70	FPR
STAble	227	1	0	1	152	161	76%	81%	0.44%
Bridger	210	58	30	28	143	148	72%	74%	28%
Oases	321	106	89	17	159	165	80%	83%	33%
Trinity	258	56	48	8	157	167	79%	84%	22%
Dataset B									
Assembler	# of results	# of FP	FPA	FPB	100%	70%	S100	S70	FPR
STAble	8906	2285	1053	1232	3295	4179	52%	66%	26%
Bridger	5697	1820	945	875	2728	3315	43%	53%	32%
Oases	16895	5722	2835	2887	3550	4156	56%	66%	34%
Trinity	8300	2543	2223	320	3603	4315	57%	68%	31%

Table B

Dataset C							
Assembler	# of results	# of FP	100%	70%	S100	S70	FPR
STAble	13985	983	10007	10263	85%	87%	7%
Bridger	5873	253	8510	9075	72%	77%	4%
Oases	5579	268	6687	8603	57%	73%	5%
Trinity	7597	145	9136	9565	77%	81%	2%
Dataset D							
Assembler	# of results	# of FP	100%	70%	S100	S70	FPR
STAble	134110	1040	20800	35424	48%	81%	0.8%

8.0 References

[1] International Human Genome Consortium. *Finishing the euchromatic sequence of the human genome.* Nature. 2004; 431(7011):931-945.

[2] https://www.illumina.com/

[3] The 1000 Genomes Project Consortium. *A global reference for human genetic variation.* Nature. 2015; 526(7571):68-74.

[4] Nik-Zainal S, Davies H, Staaf J, et al. *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature. 2016; 534(7605):47-54.

[5] The Encode Project Consortium. *An integrated encyclopedia of DNA elements in the human genome.* Nature. 2012; 489(7414):57-74.

[6] Celniker SE, Dillon LA, Gerstein MB, et al. *Unlocking the secrets of the genome.* Nature. 2009; 459(7249):927-930.

[7] Head SR, Komori HK, LaMere SA, et al. *Library construction for next-generation sequencing: Overviews and challenges.* BioTechniques. 2014; 56(2):61-passim.

[8] http://nextgen.mgh.harvard.edu/CustomPrimer.html

[9] Johnsen JM, Nickerson DA, and Reiner AP. *Massively parallel sequencing: the new frontier of hematologic genomics*. Blood. 2013; 122(19):3268-3275.

[10] Ding L, Ley TJ, Larson DE, et al. *Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.* Nature. 2012; 481(7382):506-510.

[11] Cancer Genome Atlas Research Network. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways*. Nature. 2008; 455(7216):1061-1068.

[12] Cancer Genome Atlas Research Network. *Integrated genomic analyses of ovarian carcinoma.* Nature. 2011; 474(7353):609-615.

[13] Banerji S, Cibulskis K, Rangel-Escareno C, et al. *Sequence analysis of mutations and translocations across breast cancer subtypes.* Nature. 2012; 486(7403):405-409.

[14] Stransky N, Egloff AM, Tward AD, et al. *The mutational landscape of head and neck squamous cell carcinoma.* Science. 2011; 333(6046):1157-1160.

[15] Ding L, Getz G, Wheeler DA, et al. *Somatic mutations affect key pathways in lung adenocarcinoma*. Nature. 2008; 455(7216):1069-1075.

[16] Berger MF, Hodis E, Heffernan TP, et al. *Melanoma genome sequencing reveals frequent PREX2 mutations.* Nature. 2012; 485(7399):502-506.

[17] Cancer Genome Atlas Network. *Comprehensive molecular characterization of human colon and rectal cancer.* Nature. 2012; 487(7407):330-337.

[18] Walter MJ, Shen D, Shao J, et al. *Clonal architecture of secondary acute myeloid leukemia.* N Engl J Med. 2012; 366(12):1090–1098.

[19] Nik-Zainal S, Van Loo P, Wedge DC, et al. *The life history of 21 breast cancers.* Cell. 2012; 149(5):994-1007.

[20] Xu H, DiCarlo J, Satya RV, et al. *Comparison of somatic mutation calling methods in amplicon and whole exome sequence data.* BMC Genomics. 2014; 15:244.

[21] Brodin J, Mild M, Hedskog C, et al. *PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data.* PLoS One. 2013; 8(7):e70388.

[22] Larson DE, Abbott TE, Wilson RK. *Using SomaticSniper to Detect Somatic Single Nucleotide Variants.* Curr Protoc Bioinformatics. 2015; 15(155):15.5.1-15.5.8.

[23] Zhang J, Zheng J, Yang Y, et al. *Molecular spectrum of KRAS, NRAS, BRAF and PIK3CA mutations in Chinese colorectal cancer patients: analysis of 1,110 cases.* Sci Rep. 2015; 5:18678.

[24] McKenna A. Hanna M, Banks E, et al. *The Genome Analysis Toolkitart: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res. 2010; 20(9):1297-1303.

[25] Koboldt DC, Zhang Q, Larson DE, et al. *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome Res. 2012; 22(3):568-576.

[26] Sherry ST, Ward MH, Kholodov M, et al. *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res. 2001; 29(1):308-311.

[27] Landrum MJ, Lee JM, Riley GR, et al. *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Research. 2014; 42(Database issue):D980-D985.

[28] Foulkes WD. *Molecular origins of cancer: inherited susceptibility to common cancers.* N Engl J Med. 2008; 359(20):2143-2153.

[29] Mao Z, Bozzella M, Seluanov A, et al. *Comparison of nonhomologous end joining and homologous recombination in human cells.* DNA repair. 2008; 7(10):1765-1771.

[30] Hoeijmakers JH. *Genome maintenance mechanisms for preventing cancer.* Nature. 2001; 411(6835):366-374.

[31] Hussain S, Wilson JB, Medhurst AL, et al. *Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways.* Hum Mol Genet. 2004; 13(12):1241-1248.

[32] Rasti M and Azimi T. *TP53 Binding to BRCA1 and RAD51 in MCF7 and MDA-MB-468 Breast Cancer Cell Lines In vivo and In vitro*. Avicenna J Med Biotechnol. 2015; 7(2):76-79.

[33] Jemal A, Siegel R, Ward E, et al. *Cancer statistics.* CA Cancer J Clin. 2007; 57(1):43-66.

[34] Berek JS, and Bast RC Jr. *Epithelial Ovarian Cancer.* In: Kufe DW, Pollock RE, Weichselbaum RR, et al., editors. Holland-Frei Cancer Medicine. 6th edition. Hamilton (ON): BC Decker; 2003. Available from: https://www.ncbi.nlm.nih.gov/books/NBK12433/

[35] Brown J and Frumovitz M. *Mucinous tumors of the ovary: current thoughts on diagnosis and management.* Curr Oncol Rep. 2014; 16(6):389.

[36] Fujiwara K, Shintani D, Nishikawa T. *Clear-cell carcinoma of the ovary.* Ann Oncol. 2016; 27(Suppl 1):i50-i52.

[37] Amé JC, Spenlehauer C, de Murcia G. *The PARP superfamily.* Bioessays. 2004; 26(8):882-893.

[38] http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/003726/WC500180153.pdf

[39] https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/206162lbl.pdf

[40] Plummer R. *Poly(ADP-ribose) polymerase inhibition: a new direction for BRCA and triple-negative breast cancer?* Breast Cancer Res. 2011; 13(4):218.

[41] Tutt AN, Lord CJ, McCabe N, et al. *Exploiting the DNA repair defect in BRCA mutant cells in the design of new therapeutic strategies for cancer.* Cold Spring Harb Symp Quant Biol. 2005; 70:139-148.

[42] Helleday T. *The underlying mechanism for the PARP and BRCA synthetic lethality: clearing up the misunderstandings.* Mol Oncol. 2011; 5(4):387-393.

[43] Montoni A, Robu M, Pouliot É, et al. *Resistance to PARP-Inhibitors in Cancer Therapy.* Front Pharmacol. 2013; 4:18.

[44] Eskander RN and Tewari KS. *Beyond angiogenesis blockade: targeted therapy for advanced cervical cancer.* J Gynecol Oncol. 2014; 25(3):249-259.

[45] O'Sullivan CC, Moon DH, Kohn EC, et al. *Beyond Breast and Ovarian Cancers:PARP Inhibitors for BRCA Mutation-Associated and BRCA-Like Solid Tumors.* Front Oncol. 2014; 4:42.

[46] In Seok Y. and Sangwoo K. *Analysis of the Whole Transcriptome Sequencing Data: Workflow and Software.* Genomics Inform. 2015; 13(4):119-125.

[47] Cokus SJ, Feng S, Zhang X, et al. *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature. 2008; 452(7184):215–219.

[48] Oshlack A, Robinson MD, Young MD. *From RNA-seq reads to differential expression results.* Genome Biol. 2010; 11(12):220.

[49] Wang Z, Gerstein M, Snyder M. *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet. 2009; 10(1):57-63.

[50] Li P, Piao Y, Shon HS, et al. *Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data.* BMC bioinformatics. 2015; 16(1):347.

[51] Chang Z, Li G, Liu J, et al. *Bridger: a new framework for de novo transcriptome assembly using RNA-seq data.* Genome Biol. 2015; 16:30.

[52] Schulz MH, Zerbino DR, Vingron M, et al. *Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels.* Bioinformatics. 2012; 28(8): 1086-1092.

[53] Grabherr MG, Haas BJ, Yassour M, et al. *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol. 2011; 29(7):644-652.

[54] Rintala A, Pietilä S, Munukka E, et al. *Gut Microbiota Analysis Results Are Highly Dependent on the 16S rRNA Gene Target Region, Whereas the Impact of DNA Extraction Is Minor.* J Biomol Tech. 2017; 28(1):19-30.

[55] Shah N, Tang H, Doak TG, et al. *Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics.* Pac Symp Biocomput. 2011; 165-176.

[56] Cole JR, Wang Q, Fish JA, et al. *Ribosomal Database Project: data and tools for high throughput rRNA analysis.* Nucleic Acids Res. 2014; 42(Database issue):D633-D642.

[57] Gilbert JA, and Hughes M. *Gene expression profiling: metatranscriptomics.* Pac Symp Biocomput. 2011; 733:195-205.

[58] Gill SR, Pop M, DeBoy RT, et al. *Metagenomic Analysis of the human distal gut microbiome.* Science. 2006; 312(5778):1355-1359.

[59] Flint HJ, Duncan SH, Scott KP, et al. *Interactions and competition within the microbial community of the human colon: links between diet and health.* Environ Microbiol. 2007; 9(5):1101-1111.

[60] Walker AW, Ince J, Duncan SH, et al. *Dominant and diet-responsive groups of bacteria within the human colonic microbiota.* ISME J. 2011; 5(2):220-30.

[61] Walter J and Ley R. *The human gut microbiome: ecology and recent evolutionary changes.* Annu Rev Microbiol. 2011; 65:411-429.

[62] Bäckhed F, Ley RE, Sonnenburg JL, et al. *Host-bacterial mutualism in the human intestine.* Science. 2005; 307(5717):1915-1920.

[63] Lundin A, Bok CM, Aronsson L, et al. *Gut flora, toll-like receptors and nuclear receptors: a tripartite communication that tunes innate immunity in large intestine.* Cell Microbiol. 2008; 10(5):1093-1103.

[64] Lee YK and Mazmanian SK. *Has the microbiota played a critical role in the evolution of the adaptive immune system?* Science. 2010; 330(6012):1768-1773.

[65] Belkaid Y and Hand TW. *Role of the microbiota in immunity and inflammation.* Cell. 2014; 157(1):121-141.

[66] Noverr MC and Huffnagle GB. *Does the microbiota regulate immune responses outside the gut?* Trends Microbiol. 2004; 12(12):562-568.

[67] Turnbaugh PJ, Ley RE, Mahowald MA, et al. *An obesity-associated gut microbiome with increased capacity for energy harvest.* Nature. 2006; 444(7122):1027-1031.

[68] Pistollato F, Sumalla Cano S, Elio I, et al. *Role of gut microbiota and nutrients in amyloid formation and pathogenesis of Alzheimer disease.* Nutr Rev. 2016; 74(10):624-634.

[69] Ley RE, Turnbaugh PJ, Klein S, et al. *Microbial ecology: human gut microbes associated with obesity.* Nature. 2006; 444(7122):1022-1023.

[70] Karlsson F, Tremaroli V, Nielsen J, et al. *Assessing the human gut microbiota in metabolic diseases.* Diabetes. 2013; 62(10):3341-3349.

[71] David LA, Maurice CF, Carmody RN, et al. *Diet rapidly and reproducibly alters the human gut microbiome.* Nature. 2014; 505(7484):559-563.

[72] Schäffler H, Herlemann DP, Alberts C, et al. *Mucosa-attached bacterial community in Crohn's Disease coheres with the clinical disease activity index.* Environ Microbiol Rep. 2016; 8(5):614-621.

[73] Arpaia N, Campbell C, Fan X, et al. *Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation.* Nature. 2013; 504(7480):451-455.

[74] Furusawa Y, Obata Y, Fukuda S, et al. *Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells.* Nature. 2013; 504(7480):446-450.

[75] Trompette A, Gollwitzer ES, Yadava K, Bru C, et al. *Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis.* Nat Med. 2014; 20(2):159-166.

[76] Lecomte V, Kaakoush NO, Maloney CA, et al. *Changes in gut microbiota in rats fed a high fat diet correlate with obesity-associated metabolic parameters.* PLoS ONE. 2015; 10(5):e0126931.

[77] Aller R, De Luis DA, Izaola O, et al. *Effect of a probiotic on liver aminotransferases in nonalcoholic fatty liver disease patients: a double blind randomized clinical trial.* Eur Rev Med Pharmacol Sci. 2011; 15(9):1090-1095.

[78] Nardone G, Compare D, Liguori E, et al. *Protective effects of Lactobacillus paracasei F19 in a rat model of oxidative and metabolic hepatic injury.* Am J Physiol Gastrointest Liver Physiol. 2010; 299(3):G669-676.

[79] Compare D, Coccoli P, Rocco A, et al. *Gut–liver axis: The impact of gut microbiota on non alcoholic fatty liver disease.* Nutr Metab Cardiovasc Dis. 2012; 22(6):471-476.

[80] Turnbaugh PJ, Ley RE, Hamady M, et al. *The human microbiome project*. Nature. 2007; 449(7164):804-810.

[81] Ritari, J, Salojärvi, J, Lahti, L, et al. *Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database.* BMC Genomics. 2015; 16:1056.

[82] Forster SC, Browne HP, Kumar N, et al. *HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes.* Nucleic Acids Res. 2016; 44(Suppl D1):D604-D609.

[83] Saggese I, Bona E, Conway M, **Favero F,** et al. *STAble: a novel approach to de novo assembly of RNA-seq data and its application in a metabolic model network based metatranscriptomic workflow.* BMC Bioinformatics. 2018; 19(Suppl 7):184.

[84] Rognes T, Flouri T, Nichols B, et al. *VSEARCH: a versatile open source tool for metagenomics.* PeerJ. 2016; 4:e2584. ECollection 2016.

[85] https://www.perl.com/pub/1999/03/pm.html/

[86] Mignone F and Rapallo F. *Detection of outlying proportions.* J of Appl Stat. 2017; 45(8):1382-1395.

[87] Huang W, Li L, Myers JR, et al. *ART: a next-generation sequencing read simulator.* Bionformatics. 2012; 28(4):593-594.

[88] Lai Z, Markovets A, Ahdesmaki M, et al. *VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.* Nucleic Acids Res. 2016; 44(11):e108.

[89] Li H, Handsaker B, Wysoker A, et al. and 1000 Genome Project Data Processing Subgroup. *The Sequence alignment/map (SAM) format and SAMtools.* Bioinformatics. 2009; 25(16):2078-2079.

[90] Cibulskis K, Lawrence MS, Carter SL, et al. *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol. 2013; 31(3):213-219.

[91] Saunders CT, Wong WS, Swamy S, et al. *Strelka:accurate somatic small-variant calling from sequenced tumor-normal sample pairs.* Bioinformatics. 2012; 28(14):1811-1817.

[92] Kim S, Scheffler K, Halpern AL, et al. *Strelka2: Fast and accurate variant calling for clinical sequencing applications.* Nat Methods. 2018; 15(8):591-594.

[93] Liu Y, Loewer M, Aluru S, et al. *SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations.* BMC Syst Biol. 2016; 10(Suppl 2):47.

[94] Roth A, Ding J, Morin R, et al. *JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data.* Bioinformatics. 2012; 28(7):907-913.

[95] https://devyser.com/products/oncology/devyser-brca-ngs/

[96] Kamke J, Kittelmann S, Soni P, et al. *Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a Sharpea-enriched microbiome characterised bylactic acid formation and utilisation.* Microbiome. 2016; 4(Suppl 1):56.

[97] Kanehisa M, Furumichi M, Tanabe M, et al. *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res. 2017; 45(Suppl D1):D353-D361.

[98] Conway M, Angione C, Liò P. *Iterative Multi Level Calibration of Metabolic Networks.* Current Bioinformatics. 2016; 11(Suppl 1):93-105.

[99] Hinsu AT, Parmar NR, Nathani NM, et al. *Functional gene profiling through metaRNAseq approach reveals diet-dependent variation in rumen microbiota of buffalo (Bubalus bubalis).* Anaerobe. 2017; 44:106-116.

[100] http://dreamchallenges.org/project/icgc-tcga-dream-somatic-mutation-calling-challenge/

[101] Altschul SF, Gish W, Miller W, et al. *Basic local alignment search tool.* J Mol Biol. 1990; 215(3):403-410.