EUROPEAN SCHOOL OF MOLECULAR MEDICINE (SEMM)
UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

PhD in Molecula Medicine – XXVI cycle

Human Genetics

MOLECULAR ALTERATIONS IN HUMAN GENETIC DISEASES THROUGH NEXT GENERATION SEQUENCING TECHNOLOGIES

Dr. Valeria D'Argenio

Academic years: 2011-2014

EUROPEAN SCHOOL OF MOLECULAR MEDICINE (SEMM)
UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

PhD in Molecula Medicine – XXVI cycle

Human Genetics

MOLECULAR ALTERATIONS IN HUMAN GENETIC DISEASES THROUGH NEXT GENERATION SEQUENCING TECHNOLOGIES

Tutor

Prof. Francesco Salvatore

Internal Supervisor

Prof . Giuseppe Castaldo

Esternal Supervisor

Dr. Penelope Bonnen

PhD Student

Dr. Valeria D'Argenio

Academic years: 2011-2014

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

NGS	Next Generation Sequencing
sstDNA	single stranded DNA
emPCR	emulsion PCR
PPi	pyrophosphate
ATP	adenosine triphosphate
CCD	Charge-Coupled Device
PTP	picotiterplate
WES	Whole exome sequencing
HCM	hypertrophic cardiomyopathy
DCM	dilated cardiomyopathy
ARVC	arrhythmogenic right ventricular cardiomyopathy
LVNC	left ventricular noncompaction
RCM	restrictive cardiomyopathy
LQTS	long QT syndrome
SQTS	short QT syndrome
CPVT	catecholaminergic polymorphic ventricular tachycardia
SCD	sudden cardiac death
IBD	Inflammatory bowel disease
CD	Crohn disease
UC	ulcerative colitis
CD	Celiac disease
MWT	Maximal ventricular wall thickness
ECG	electrocardiographic
LVH	left ventricular hypertrophy
PCDAI	Pediatric Crohn's Disease Activity Index
BT	before therapy
AT	after therapy
GFD	gluten free diet
HCDiffs	high confidence nucleotide differences
AllDiffs	all nucleotide differences
P	probability
TP	true positive
TN	true negative
FP	false positive
FN	false negative
OTUs	operational taxonomic units

FIGURES INDEX

TABLES INDEX

ABSTRACT

Next Generation Sequencing (NGS) technologies have greatly impacted every field of molecular research, reducing costs and simultaneously increasing throughput of DNA sequencing. These features, together with technology's flexibility, have open the way to a variety of applications, especially for the study of the molecular basis of human diseases.

So far, several analytical approaches have been developed to selectively enrich regions of interest from the whole genome, both to identify germinal and/or somatic sequence variants. All of these have assessed their potential in research area and are now being improved also in routine molecular diagnostics. Thanks to the improvement due to NGS methods introduction, also the metagenomic field has achieved very exciting results, increasing our knowledge about the microbiome and its mutually beneficial relationships with the human host. If microbiome plays a role in the maintenance of a healthy status, it is conceivable to suppose that its quantitative and/or qualitative alterations could lead to pathological dysbiosis, as shown in an increasing number of intestinal and extra-intestinal diseases.

The aim of this project was to use NGS-based strategies to study the molecular basis of human diseases. In particular, two analytic approaches were used: DNA sequence capture and metagenomics.

A DNA sequence capture approach was used to analyze a large panel of genes possibly related to inherited cardiomyopathies. The obtained results indicate that this approach is useful to analyze, in a time and cost effective manner, heterogeneous diseases allowing the identification not only of the disease-causing mutation, but also of other variants involved in

disease-phenotypic expression. Finally, methods reliability was higher than traditional, currently used techniques. All the above data indicate that this validated NGS-based approach can be used to improve the molecular analysis of inherited cardiomyopathies, such as of other inherited diseases, also in routine diagnostic settings.

With regard to metagenomics, a 16S rRNA pyrotag analysis was carried out to deeply investigate the gut microbiome composition of Crohn and celiac diseases. Specific microbial signatures were identified in the patients. Moreover, the effects of Crohn nutritional therapy on gut microbial composition were also verified. These results suggest a role of gut microbiome in diseases pathogenesis and could, in turn, make possible to develop novel diagnostic, prognostic and, most important, therapeutic strategies.

Taken together, all the above results indicate that the used NGS-based procedures can be easily applied to increase our understanding of the molecular basis of human diseases and that they can be useful also for routine diagnostic purposes.

Keywords

Next generation sequencing, molecular diagnostics, metagenomics, inherited cardiomyopathies, Crohn disease, celiac disease.

I. INTRODUCTION

The process of determining the exact order of the nucleotides in a DNA molecule or in a genome is defined as "DNA sequencing". This simple definition is enough to understand how, techniques able to perform this operation have radically changed the course of molecular research in all its fields of application.

In the last 30 years, the so-called Sanger sequencing has been the most widely used sequencing technology worldwide [1]. This method, developed in the '70s by Frederick Sanger, thanks to continuous improvements in technology performances, reached its peak with the Human Genome Project (HGP), which, in 2001, elucidated the first entire human genome [2,3]. Now, the Sanger sequencing procedure is completely automated; however, it is a single amplicon method and, as such, it is expensive and time consuming. Starting from these points, there was the need to develop novel sequencing methods able to overcome Sanger limits and answer to the increasing requests of great amounts of high quality sequencing data, both in a faster and cheaper fashion.

To address the above mentioned issues, in the last ten years, novel technologies, called "next generation sequencing" (NGS), have become available. NGS methods have dramatically increased the throughput of DNA sequencing, simultaneously reducing its costs [4]. Just to give the idea of what this means, it took more than 10 years to elucidate the first human genome sequence and it cost 3 billion $. Using NGS instruments, the entire genome sequence of an individual has been elucidated in only 1 year and at a much lower cost [5]. It is expected that the sequencing of the entire

genome of an individual will cost about 1,000 $ in a near future (Figure 1) [6].

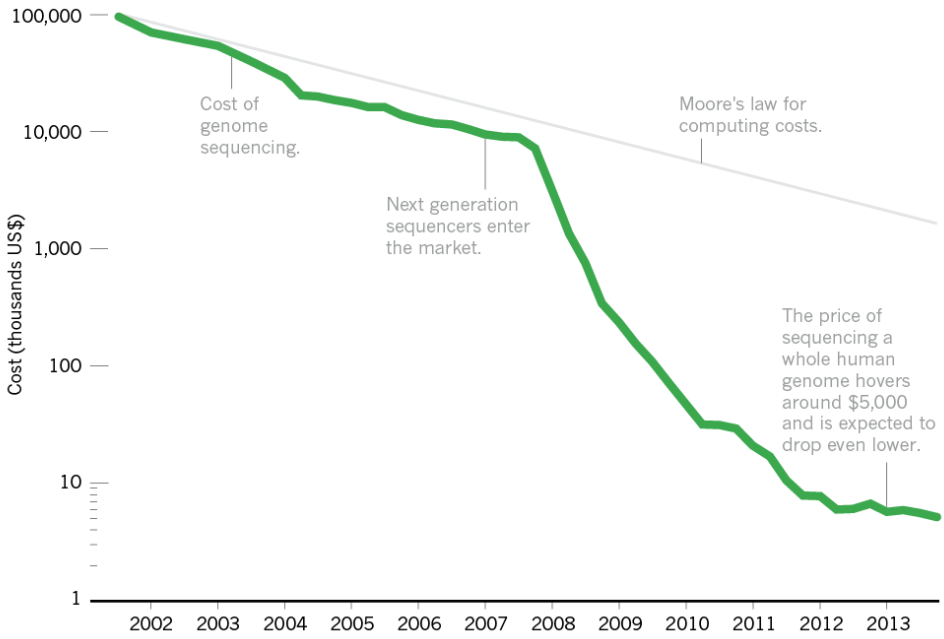


Figure 1. **Impressive reduction of DNA sequencing costs.** The graph shows that the introduction of NGS dramatically dropped sequencing costs respect to the hypothetical prevision of the Moore's Law, based on the prevision of exponential computing costs reduction [6].

These aspects, coupled to the technologies versatility, have open the way to the massive NGS diffusion in every field of molecular research and to the begin of the so called "-Omics" era.

1.1 Next Generation Sequencing Technologies

Next Generation Sequencing technologies (NGS) have been defined as the next phase of DNA sequencing evolution since they allow single laboratories to sequence entire genomes in a few time and at very competitive costs [7].

Different NGS platforms have been developed so far, each one showing specific advantages and limitations. However, until now three of them has had the largest diffusion: i) the Roche 454 Genome Sequencer FLX (http://www.my454.com/); ii) the Illumina HiSeq (http://www.illumina.com); and iii) the Life SOLiD (http://www.lifetechnologies.com/). In general, samples' preparation through these platforms is made up of three main analytic steps: i) the generation of a single stranded DNA (sstDNA) library; ii) the library amplification; and iii) the sequencing reactions. In addition, data analysis using specific bionformatic pipelines has to be carried out (Figure 2).

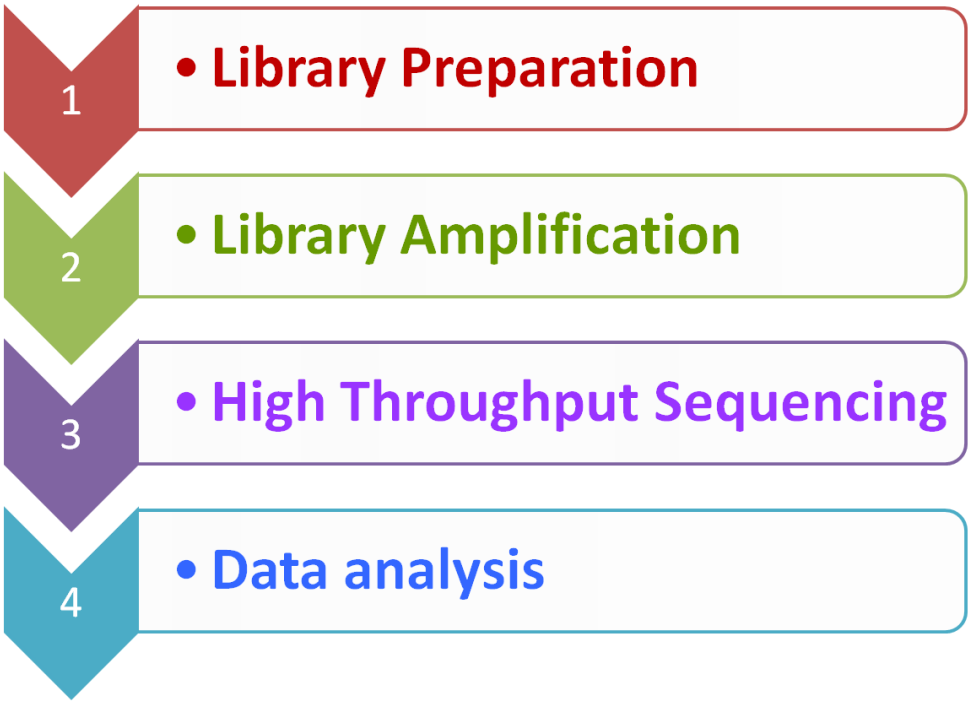


Figure 2. Overview of the sample preparation workflow through NGS platforms.

The library preparation is the most flexible step in the entire procedure. It depends on the kind of project, on the nature of the biological

samples used as source, and on the question/s you are looking to address. Several methods have been used for NGS library preparation. Traditionally, the analysis of an entire genome is based on the isolation of the DNA and its fragmentation using different methods. Otherwise, to analyze selected genomic regions, a number of enrichment procedure, both PCR-based and PCR-independent, are available [8]. Finally, it is also possible to analyze the whole RNA or RNA subpopulations [9]. At the end of the library preparation (independently from the used protocol) a population of sstDNA fragments, homogeneous in size and compatible with sequence reads length, is obtained. These fragments are modified by the ligation of specific adapters (different for each NGS platform), which are required as primers for the subsequent amplification and sequencing reactions. The chemistry used to carry out these two steps are different for each NGS platform. In particular, the library amplification can be obtained by emulsion PCR (emPCR) or by solid-phase amplification. In the emPCR, after adapters ligation, DNA fragments are captured on the surface of specific beads and amplified into the droplets of a water-in-oil emulsion. The ratio between library fragments and capture beads is carefully evaluated (titration assay) to allow the capture of one DNA molecule per bead. In this way, at the end of the amplification cycle, each bead will carry million of copy of the same DNA fragment clonally amplified. The emulsion increase the amplification specificity since each bead is amplified in its own droplet, reducing cross contamination risks. After the amplification, the emulsion is chemically broken to recover and enrich the DNA carrying beads (Figure 3, panel A) [10]. In the solid-phase amplification, the adapted DNA fragments are hybridized to a glass slide to obtain clonally amplified clusters. In this case, high density forward and reverse primers are immobilized on the slide and

the ratio between them and the library fragments is crucial for cluster density and, consequently, for sequencing throughput [8]. In brief, ssDNA fragments are hybridized to the glass slide surface, extended by polymerase and denatured to obtain ssDNA fragments covalently ligated to the slide surface. These obtained ssDNA fragments flip over to form a bridge by hybridizing to the primer on the slide and are used as template by the polymerase (bridge amplification) (Figure 3, panel B).

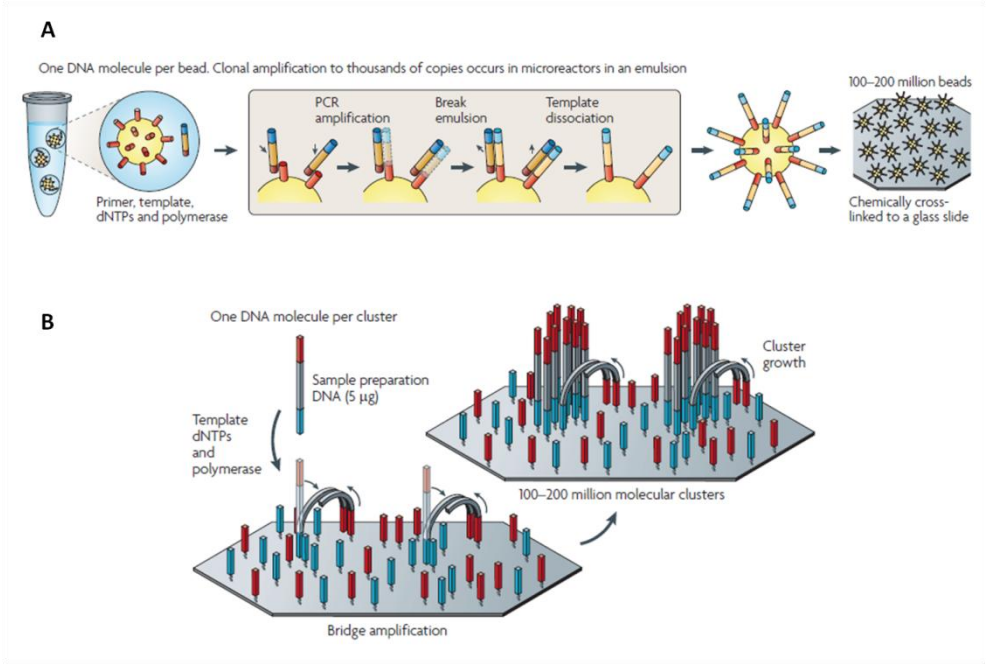


Figure 3. Library amplification's strategies. DNA library fragments are immobilized on the surface of DNA capture beads and amplified into an emulsion (A), or hybridized on the surface of a glass slide to do a bridge amplification (B) [8].

After the amplification, the libraries are ready to be sequenced. Different sequencing chemistries have been developed and optimized by each NGS platform, such as pyrosequencing, reversible terminator strategy and sequencing by ligation. In the pyrosequencing chemistry the four

nucleotides are eluted one at time in a fixed order. When a nucleotide complimentary to the template is added, it is incorporated by the polymerase in the elongation strand, releasing a molecule of PPi. The latter is used by the sulphurylase to convert the adenosine phosphosulfate in ATP which is used by the luciferase to oxidate luciferin. The emitted light is registered by a CCD camera and converted in sequence data (Figure 4, panel A) [10]. In the reversible terminator strategy, the four nucleotides are labeled each with a different fluorescent dye, so that they can be eluted all together during each sequencing cycle. The nucleotide complimentary to the template is incorporated and, after fluorescence signals registration, it is cleaved together with the terminating group becoming available for nucleotide adding in the next sequencing step (Figure 4, panel B) [11]. Finally, the sequencing by ligation use a mixture of fluorescently labeled octamer oligonucleotides that hybridize to the sequence adjacent to the primed template. These octamers are two-base-encoded probes in which the combination of the first two bases is assigned to a specific dye. So, after the hybridization and fluorescence registration, the last three bases of the octamer are cleaved to start again with the octamers elution. In this way, at the end of the sequencing cycle, two bases each five nucleotides have been read on the template strand. At this point, the primer is removed and a second ligation cycle is performed using a "n-1 primer". This primer resetting procedure is repeated more time to fill in the gaps in the template. This allows each base to be read twice, in two independent ligation reactions and using two different primers, ensuring a great accuracy in base calling (Figure 4, panel C) [12].

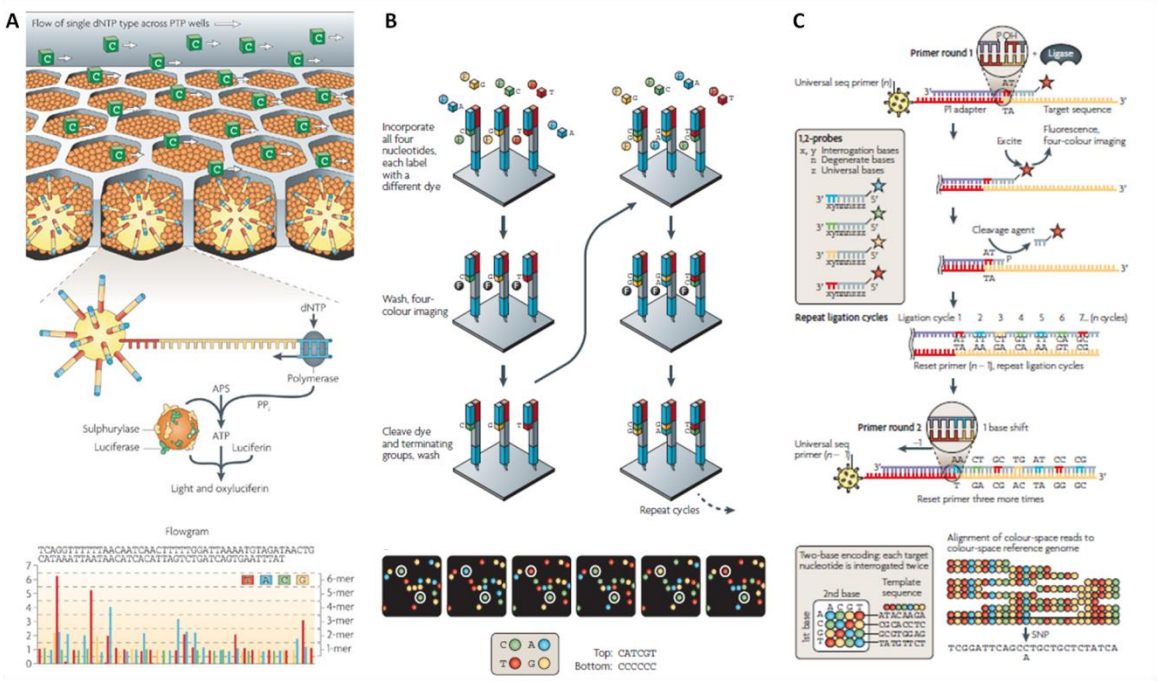


Figure 4. NGS sequencing chemistries. The pyrosequencing process is based on the production of light when a nucleotide complimentary to the template is added. Since the four nucleotides are not differentially labeled, they are eluted one at time in a fixed order to ensure correct base calling. In the 454 Roche procedure, the amplified library fragments are annealed on the surface of a DNA capture bead. These beads (each carrying million of copy of the same DNA library fragment clonally amplified) are deposited into the wells of a fiber optic slide (PTP) together with beads carrying the pyrosequencing enzymes. In this way, sequencing reactions occur simultaneously in all the PTP wells (A). The reversible terminator strategy uses four differentially labeled nucleotides eluted together in each sequencing cycle. In the Illumina system, the amplified library fragments are annealed on the surface of a glass slide (flow cell) to obtain clusters simultaneously sequenced (B). Finally, the sequencing by ligation procedure use a mixture of fluorescently labeled octamers coupled to a dual color code. In the SOLiD procedure the library amplification has been optimized from a bead to a slide system (C) [8, 10-12].

The different technical features of each NGS platform account for their own strengths and pitfalls, especially in terms of sequencing reads length, base calling accuracy and sequencing throughput (Table 1).

Table 1. Comparison of the currently available NGS platforms features.

NGS Platform	Library amplification	NGS chemistry	Read lenght (bp)*	Run throughput (Gb)*	Run time(Days)
Roche 454 GS FLX	emPCR	Pyrosequencing	700[†]	0.7	1
Illumina HiSeq	Solid-phase	Reversible terminator	2x125[§]	600	10
Life SOLiD	Solid-phase	Sequencing by ligation	2x50[§]	160	8
Roche 454 Junior	emPCR	Pyrosequencing	700[†]	0.07	0.5
Illumina MiSeq	Solid-phase	Reversible terminator	2x300[§]	15	2.2
Life Ion Torrent	emPCR	H^+ Ion semiconductor	400	2	0.34

NGS, Next generation Sequencing; bp, base pair; Gb, gigabase.
*Reported considering the highest performances actually available for each platform. [†]Average read's length (up to 1,000 bp). [§]Paired-end sequencing.

The progressive optimization and standardization of NGS procedure has lead to a continuous increase of sequencing productivity/run and a reduction of sequencing costs, as mentioned above [6]. This, in turn, has increased the request for routine NGS-based applications also in clinical settings for diagnostic purposes [7]. In this view, the so called NGS "bench-top" instruments, such as the Roche 454 Junior, the Illumina MiSeq and the Life IonTorrent, have been launched on the market promising the same sensitivity and accuracy of the biggest instruments, but with instruments costs and sequencing run time markedly reduced. From a technical point of view, while the Junior and the MiSeq systems use the same chemistries of the larger versions, the IonTorrent is completely different from its

corresponding larger platform. In fact, it couples the emPCR protocol for library amplification to a hydrogen ion semiconductor chip for sequencing reactions. In particular, the clonally amplified DNA beads are loaded into the wells of this semiconductor device, where the four unlabeled nucleotides are loaded. When there is an incorporation event, a hydrogen ion is released and the consequent voltage difference is recorded (Figure 5).

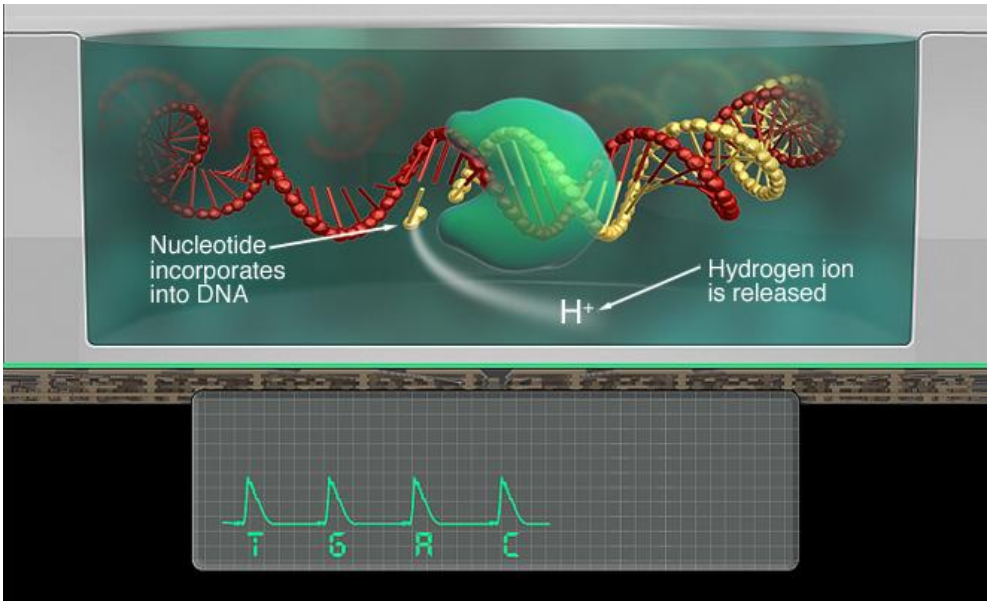


Figure 5. IonTorrent sequencing chemistry. A nucleotide incorporation into the strand of DNA in elongation produce the release of a hydrogen ion. The ion charge modify the pH, which is detected by ion sensor and converted in sequence data.

Taken together, these bench-top platforms are featured by a fast turnaround time and a great flexibility in sequencing throughput, representing a real chance to combine NGS advantages to a routine laboratory use. Their main features in terms of productivity are summarized in Table 1.

In addition to all the above mentioned sequencers, usually defined as "second generation", novel technologies also called "third generation"

sequencers are being developed. The main feature of these novel sequencing technologies is that they avoid the library amplification step. Therefore, library fragments are directly sequenced at a single molecule level with the great advantage to avoid PCR biases. Different third generation platforms are available, such as the SeqLL HeliScope (http://seqll.com/), the Pacific Biosciences PacBio (http://www.pacificbiosciences.com) and the Oxford Nanopore (http://www.nanoporetech.com) [13-16]. Also in this case, different chemistries have been developed and characterize each platform also in terms of capability (Table 2). Even if the potential of these novel technologies is really exciting, to date they are not yet currently and massively used. It is to be expected that, once the protocols will be standardized and the analytic performances improved, they will overwhelm the market.

Table 2. Comparison of the Third generation sequencers features.

Platform	Sequencing chemistry	Read lenght (bp)*	Run throughput (Gb)*	Run time(Days)
SeqLL HeliScope	Reversible terminator	32†	28	1.2
Pacific Bioscience PacBio	Real time	10,000†	0.5†	80.2
Oxford Nanopore	Electronic Sensing	NA	NA	NA

bp, base pair; Gb, gigabase; NA, not available.
*Reported considering the highest performances actually available for each platform. †Average value. §Paired-end sequencing.

1.2 Next Generation Sequencing Technologies applications

The great technological advances following the diffusion of NGS-based approaches gave new impetus to several research topics. It is

important to underline that the NGS platforms are featured by a large flexibility. Several biological samples can be used as source and different kind of projects can be realized in a way and in time not easily imaginable before NGS revolution. This novel awareness has involved also the biomedical research area with the aims to markedly accelerate the search for genetic causes of human diseases and to answer previously difficult-to-answer questions regarding disease pathogenesis. In recent years, several NGS-based approaches have been described and validated to improve the study of the molecular basis of human diseases with the aim to: (i) develop novel, sensitive, accurate, cost- and time-effective pipelines for molecular diagnostics; and (ii) highlight the mechanisms involved in diseases development to identify novel diagnostic, prognostic and therapeutic markers [7]. In fact, NGS technologies can be easily applied to the study of the human genome through a variety of approaches and until now, they have been successfully used to analyze target regions of the human genome, ranging in size from the entire exome to a restricted number of genes or a single amplicon [17-19]. In addition to nucleotide variants detection, NGS-based strategies are useful also to study the DNA methylation status, both at single gene or at genome-wide level [20]. Finally, metagenomics has been really improved by NGS advent [21].This phenomenon is so diffused that is conceivable to suppose that novel NGS-based strategies are still developing and that these technologies will became even more routinely, especially for diagnostic purposes, considering the progressive protocol simplification, the operator “hand on” work reduction, and the advantages of the “bench-top” NGS platforms. In addition, the integration of data obtained using several NGS-based strategies could represent an additional advantage to better understand the mechanisms involved in diseases development and, in turn to

identify actionable target for a better patients identification, stratification and treatment.

Actually, two kinds of NGS-based approaches has having a huge diffusion and showing their potentialities for the study of the molecular basis of human diseases, targeted DNA sequence capture and metagenomics, and will be discussed more in detail.

1.3 Targeted DNA sequence capture

Targeted DNA sequence capture is a NGS-based approach for the simultaneous analysis of genomic target regions selectively enriched by the whole DNA.

NGS techniques allow the study of entire genomes faster and cheaper than conventional Sanger sequencing [22-23]. However, the entire sequencing of a large number of samples is not yet feasible for routine use due to the cost, time and infrastructures required. Thus, different approaches to specifically enrich target genomic regions, simultaneously allowing samples barcoding for sequence multiplexing, have been developed and can be classified in PCR-based and PCR-alternative strategies. Both of these can be used for NGS libraries preparation and the choose of the most appropriate one depend on the size of the target regions, the number of samples to be analyzed, the costs and time required and on the biological questions to be addressed (Table 3).

Table 3. NGS-based enrichment strategies for DNA sequence variants identification.

Enrichment System	DNA input	Required Time*	Sensitivity	Specificity	Max target size
Long PCR	5ng/ amplicon	4-5h	High	High	Depend on amplicon length
Multiplex PCR	5ng/ multiplex	3-4h	High	High	Depend on amplicon length and multiplexing
Microdroplet PCR	1.5ug	48h[†]	High	High	Up to 20,000 genomic loci
WES	500ng-2ug	92h	High	>60%[§]	50-75Mb[§]
Targeted capture	500ng-2ug	92h	High	>60%[§]	Up to 50 Mb of custom regions

WES, whole exome sequencing; Mb, megabase.

PCR has been the most widely used pre-sequencing strategy to date, since it is perfectly compatible with Sanger sequencing and also with all the NGS instruments: at the end of the amplification, the resulting amplicons have NGS-platform-specific adapters ligated to their ends. This represents a library that is suitable for the downstream sequencing reactions [17]. Since barcode sequence-tags can be also added during this step, samples multiplexing is also allowed [24]. In general, PCR-based strategies, including also multiplex PCR and long range PCR, are useful to analyze one or a few genes [25-31]. Otherwise, since PCR amplification is too laborious for large scale NGS downstream applications, it risks to be a bottleneck in the sample preparation workflow.

The so called "DNA sequence capture" approach is a PCR alternative strategy able to overcome PCR limitations and is an excellent way to isolate large or highly dispersed regions from a pool of DNA molecules [32]. Sequence capture is essentially based on hybrid capture

reactions for the selective enrichment of targeted genomic regions. Specific capture probes can be synthesized to enrich the regions of interest from the whole genome, thus obtaining a captured, adapted and barcoded library for NGS applications [18,33,34]. More in details, DNA fragments hybridize to the capture probes synthesized on DNA microarray glass slides in array-based hybridization method [18], while, biotinylated DNA or RNA probes are used in liquid-phase hybridization. The non-targeted DNA fragments are washed away and the enriched DNA is recovered and used for high throughput sequencing (Figure 6).

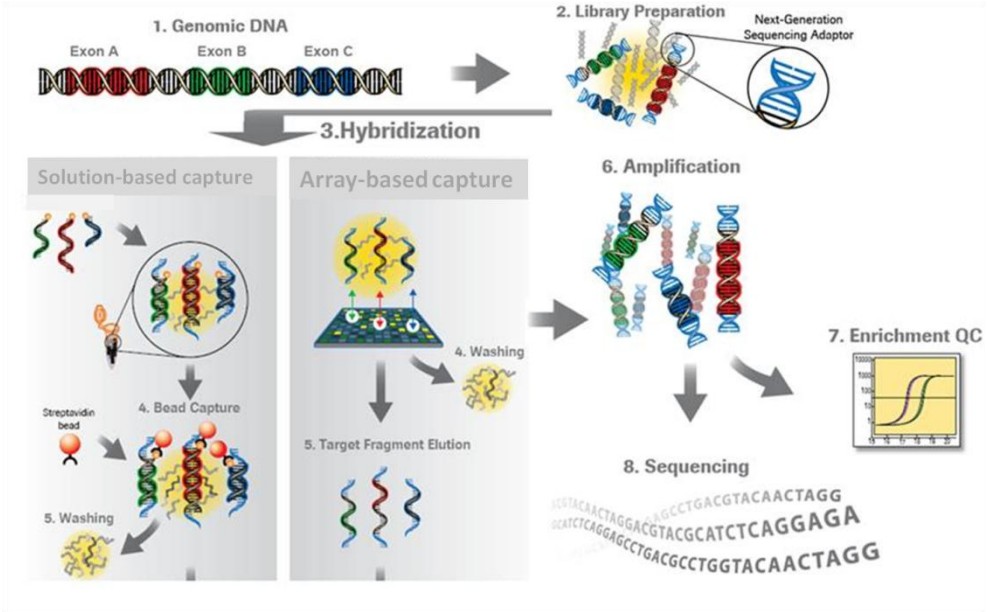


Figure 6. Schematic view of the DNA sequence capture procedure.

An alternative technology (solution-based) is a noted example of enrichment system featured by selective circularization-based method which is a further development of the principle of selector probes used in several diagnostic approaches [35,36]. In brief, genomic DNA is fragmented by restriction enzyme digestion and circularized by hybridization to probes

whose ends are complementary to two non-contiguous stretches of a target region (Figure 7).

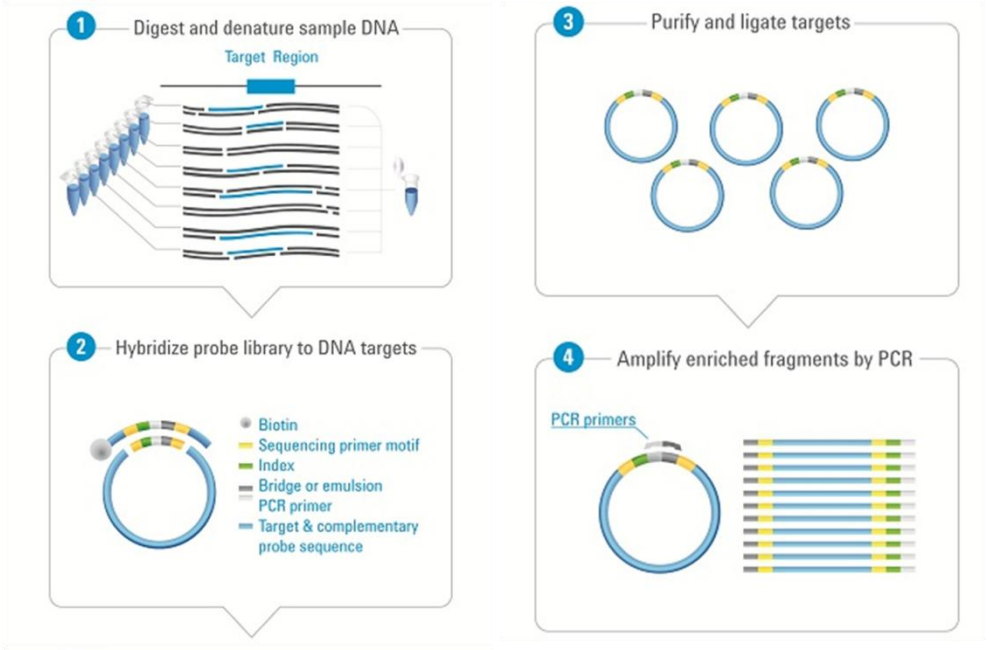


Figure 7. Circularization-based procedure for selective DNA enrichment.

Even if PCR-based enrichment methods have the benefit of even coverage and high specificity, DNA sequence capture has several advantages able to overcome it. Hybridization is less sensitive to contamination and also mismatches are less harmful. In addition, PCR specificity depends on reaction optimization and primer design: large rearrangements in genomes, for example, may be undetectable unless primer pairs in flanking regions. Unlike PCR, drawbacks of hybridization capture are the need of relatively large amount of high-quality DNA and the loss of target molecules during library preparation. Therefore, this approach is especially suitable for the study of large genomic regions, either contiguous or not contiguous, including the entire exome.

Whole exome sequencing (WES) is defined as the selective sequence the coding regions of a genome, to discover rare or common variants associated with a disorder or phenotype [37,38]. As a result, WES is an attractive and practical approach for the study of coding variants related to rare Mendelian disorders and of many disease-predisposing SNPs throughout the exome [39-42].

To reduce sequencing time and costs and avoid the drawbacks related to data analysis, the target enrichment of specific genomic regions of interest seems to be an attractive alternative. Until now, target enrichment-based strategies, have been used for sensitive nucleotide variants identification [43,44], validation of novel diagnostic tools [45-47] and drug resistance/sensitivity profiling [48-50]. This method is useful for the study of complex families with different genotype/phenotype correlation and to identify all the at risk subjects [51].

Considering all the above, and also that target enrichment technologies are easy to use, they appear really suitable for the study of the molecular basis of genetic diseases, both for research and diagnostic purposes.

1.3.1 Inherited cardiomiopathies

Inherited cardiomyopathies are a group of heterogeneous genetic diseases usually classified according to functional and morphology abnormalities of the cardiac muscle. It includes hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), arrhythmogenic right ventricular cardiomyopathy (ARVC), left ventricular noncompaction (LVNC), and restrictive cardiomyopathy (RCM). In addition, another group of inherited cardiomyopathies is featured by a primary involvement of

cardiac electric transmission (channelopathies) and includes long QT syndrome (LQTS), short QT syndrome (SQTS), Brugada syndrome, and catecholaminergic polymorphic ventricular tachycardia (CPVT). All these diseases are featured by a high clinical and genetic heterogeneity [52].

Clinical presentation ranges from asymptomatic to severe rapidly worsening forms and the same symptoms can be the expression of different diseases. The age of onset is also extremely variable: sometimes an inherited cardiac disease is diagnosed before birth, while other subjects can show clinical signs later in the adulthood [52]. Most relevant, all of these diseases can predispose to the development of malignant arrhythmias, heart failure and sudden cardiac death (SCD). It has been estimated that the majority of SCD in young individuals can be related to the presence of one of the above mentioned diseases [53]. Therefore, the correct identification of the molecular alterations responsible for inherited cardiomyopathies is crucial for the correct patient's management and for the identification of all the at risk subjects within the affected families.

The cardiomyopathies' clinical variability is reflected in the heterogeneity of their genetic basis. To date, tens of genes, showing different mechanisms of inheritance and incomplete penetrance, have been related to the onset of each inherited cardiomyopathy. However, they explain only a variable proportion of all cases suggesting the existence of other, still unknown disease-causative genes [52]. In addition, the highly variable phenotipic expression, also in the presence of the same causative mutation and in the same family, has strongly suggested a role for additional inherited variants, in the same gene or in independently inherited genes, able to act as phenotype-modifiers. Finally, molecular variants in the same gene have been related to different cardiomyopathies.

Taken together, all the above mentioned issues explain the difficulties in the correct molecular diagnosis of these overlapping diseases. Therefore, NGS-based approaches, enabling the simultaneous analysis of large number of genes, promise to overcome these limitations. Increasing evidences are assessing the potentialities of NGS-based approaches for the study of inherited cardiomyopathies [52]. van de Meerakker et al identified, in a DCM family, a novel mutation in the alpha-tropomyosin gene through the targeted enrichment of a panel of 23 candidate genes (array-based enrichment) [54]. In another study, the custom enrichment of 16 genes was able to identify a compound heterozygosis in an infant affected by LVNC, confirming the potential of NGS as fast diagnostic tool [55]. Similar approaches have been used in different patient settings showing their reliability in terms of accuracy and sensitivity also for diagnostic pulposes [56-62] Finally, post-mortem WES was able to identify the causative mutation in a young women death for SCD [63]. Taken all together, these findings assess the feasibility of NGS for the study of complex inherited diseases, such as inherited cardiomyopathies and suggest their use as routine diagnostic procedure in a molecular biology laboratory.

1.4 Metagenomics

Metagenomics is defined as the field of molecular research that studies the complexity of microbiomes, i.e. the entire collection of all the genomic elements of a specific community of microrganisms, including bacteria, archaea, viruses, and some unicellular eukaryotes, living in a specific environment (microbiota).

In the last few years, metagenomics literature has grown exponentially, essentially due to technological improvements related to the

introduction of NGS. This has, in turn, greatly improved our understanding about the role of human microbiota in the healthy status and its possible implications in diseases pathogenesis [64].

If we consider the human body as an environment, the human microbiota is the entire collection of microorganisms living on the surface and inside our body (Table 4) [65-68].

Table 4. Human microbiota composition across the five most extensively studied body sites. Interestingly, the oral and gut microbiota have the highest microbial diversity, while the urogenital tract has the smallest bacterial diversity [from 64].

	Human microbiota (10 times more microbial than human cells: 10^{14} vs 10^{13})	
Human Microbial Habitats	**Most represented Phyla and their relative abundance (%)**	**Number of species**
Oral cavity	*Firmicutes* (36.7), *Bacteroidetes* (17.3), *Proteobacteria* (17.1), *Actinobacteria* (11.9), *Fusobacteria* (5.2)	>500
Skin	*Actinobacteria* (52), *Firmicutes* (24.4), *Proteobacteria* (16.5), *Bacteroidetes* (6.3)	~300
Airways	*Actinobacteria* (55), *Firmicutes* (15), *Proteobacteria* (8), *Bacteroidetes* (3)	>500
Gut	*Firmicutes* (38.8), *Bacteroidetes* (27.8), *Actinobacteria* (8.2), *Proteobacteria* (2.1)	>1,000
Urogenital tract[a]	*Firmicutes* (83), *Bacteroidetes* (3), *Actinobacteria* (3)	~150

[a]mainly female.

These communities are required for human physiology, immune system development, digestion and detoxification reactions [60,70]. In this view, humans can be defined as "superorganisms" made of two genomes, one inherited from parents and the other acquired, i.e., the microbiome [71]. Differently from the inherited genome, which is almost stable during lifetime, the microbiome is extremely dynamic and influenced by age, diet, hormonal cycles, travel, therapies, and illness [72-77].

Most of the human adult microbiota lives in the gut. Only in the human colon microbial cell density exceed 10^{11} cells/g contents, being equivalent to 1-2 kg of body weight [78]. In addition, it has been estimated that the human gut microbiome accounts for more than 5 million different genes [79]. Even if over 1,000 different species colonize the human gut [21], they belong to a small number of phyla: *Firmicutes*, *Bacteroidetes* and *Actinobacteria*, followed by the less represented *Proteobacteria*, *Fusobacteria*, *Cyanobacteria* and *Verrucomicrobia* [70]. To date, a number of functions have been associated to the gut microbiome, including polysaccharide digestion, immune system development, defense against infections, synthesis of vitamins, fat storage, angiogenesis regulation, and behavior development [69,70,80,81]. Therefore, it is conceivable to suppose that alterations of the human gut microbiome can play a role in disease development and more our understanding on its role will grow up, more it will became possible to use the human microbiome for diagnostic purposes or as target for novel therapies.

All the above, explain the boom of metagenomics worlwide. NGS plays its role in this scenario, since it allows the qualitative and quantitative analysis of a specific microbiome without selection biases and constraints associated with cultivation methods. Different NGS-based strategies can be used for metagenomic purposes, as shown in Figure 8.

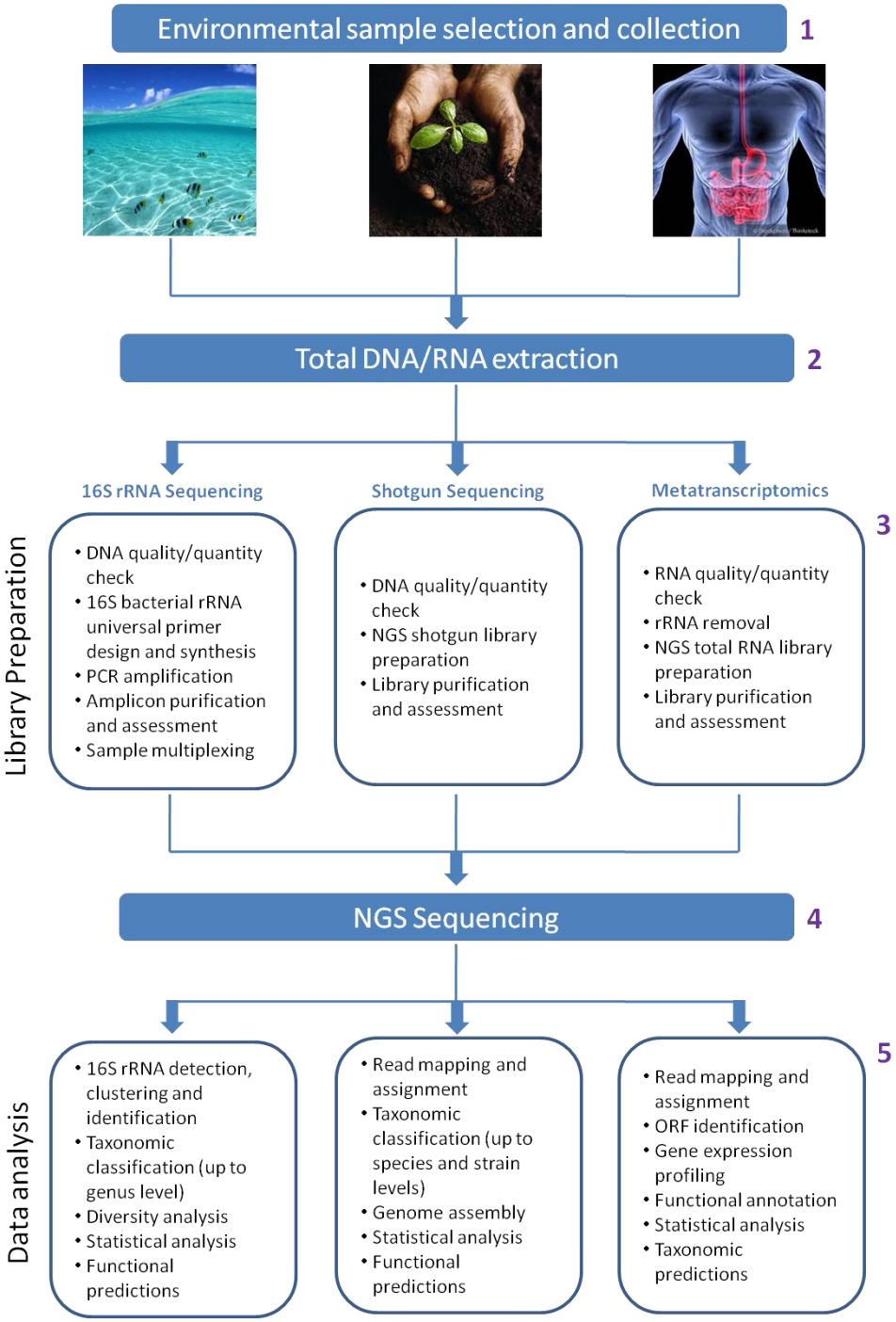


Figure 8. Next Generation sequencing-based approaches for metagenomics. Starting from an environmental sample of interest (1), total DNA and/or RNA are extracted (2). Three different sample preparation strategies can be used depending on the project aims: 16s rRNA Sequencing, Shotgun Sequencing, and Metatranscriptomics (3). Usually, the 16S rRNA procedure allows sample multiplexing while a higher coverage is required for the others. After sequencing (4), specific bioinformatic pipelines are used for data analysis (5) [from 64].

Shotgun approaches allow a comprehensive view of entire microbial communities both at DNA or RNA level, with taxonomic assignment up to the species. However, they still have technical limitations, most of which related to sequence data analysis. Therefore, the most used approach is the targeted resequencing of specific sequence tags useful for phylogenetic purposes, like the bacterial 16S rRNA gene. The latter gene has a peculiar structure characterized by hypervariable regions spaced by ultra-conserved regions [82]. Universal primers which anneal on the conserved regions can be used to amplify, in a single PCR reaction, virtually all the bacteria present in a target environment, and to unequivocally identify them at the end of sequencing [83].

This approach has been successfully used to analyze the microbial community richness of the human gut microbiome and its relationship with specific diseases, such as obesity and immune-related and inflammatory diseases [84-93]. It is now well known that the gut microbiome of obese subjects differ from that of not obese and that such differences could play a role in altering gut permeability and induce inflammatory reactions [91-93]. Similar mechanisms seems to be applicable to an increasing number of intestinal and extra-intestinal diseases, including inflammatory bowel diseases and celiac disease [84-87].

1.4.1 Gut microbiome and inflammatory bowel diseases

Inflammatory bowel disease (IBD) are chronic inflammatory disorders of the gastrointestinal tract with increasing worldwide incidence [94]. IBD clinical features are severe, including diarrhea, weight loss and debilitating abdominal pain, and result in substantial morbidity and impairment in quality life [95]. IBD also increase the risk for colon cancer

development [96]. The affected intestinal areas are usually characterized by transmural inflammation associated to lymphoid hyperplasia, submucosal edema, ulcerative lesions and fibrosis which can be assessed by colonoscopy [97]. The two main forms of these disorders are Crohn disease (CD) and ulcerative colitis (UC). IBD causes are still unclear. Even if host genetics play a key role, mostly in CD than in UC onset [98], environmental factors are also involved [99]. Numerous evidences emphasize the role of gut microbiome in triggering and perpetuating typical IBD chronic inflammation. The most widely currently accepted hypothesis on IBD pathogenesis is that, in genetically predisposed subject, an altered host immune response against luminal agents, such as the gut microbiota, could result in the IBD chronic inflammation [97,100].

It has been established that microbes rapidly colonizes the gut after birth [101] and this process is influenced by different factors, such as the heredity of the mother, the immediate living environment, the feeding practices, microbial infections, and the host's genetics [102]. As discussed above, this mutual relationship between the human host and its microbiota seems required for healthy status acquisition and maintenance.

Several studies have identified significant alterations in the gut microbiota composition of IBD patients with respect to healthy individuals [103,104]. Therefore, there is a growing interest in understanding whether interactions between intestinal microbes and innate immunity could influence IBD expression and act as triggers of the inflammation, since this could open the way to novel diagnostic and therapeutic opportunities. In addition, since nutritional therapy is effective in pediatric Crohn disease, it has been suggested that it may induce its beneficial effects by modifying the microbiome composition [105].

1.4.2 Gut microbiome and celiac disease

Celiac disease (CD) is a chronic, multi-factorial, inflammatory disorder of the small intestine that involves interactions between genetic and environmental factors [106]. In genetically susceptible individuals (HLA-DQ2/DQ8 carriers) ingestion of gluten leads to an abnormal intestinal immune response involving both adaptive and innate immunity, which is characterized by failure to establish and/or maintain tolerance to dietary peptides in wheat, barley, and rye (particularly to wheat gliadin) [107]. This abnormal immune activity damages the small intestine, which typically shows villous atrophy, crypt hyperplasia, and an increased number of lymphocytes within both the epithelium and the lamina propria [108].

However, more recent studies have challenged the gluten-hypersensitivity dogma and suggest that exposure to gluten may not be the only factor contributing to the onset of CD [109,110]. Indeed, traditional pathogenic mechanisms of gluten hypersensitivity do not explain why: i) the frequency of DQ2/DQ8 molecules in the general population is about 30%, but only 1to 3% of individuals develop CD [106]; ii) CD is increasingly diagnosed in adulthood, many years after introduction of gluten in the diet [111]; iii) the prevalence of CD prevalence is rapidly increasing in the western world (although consumption of gliadin-containing food has not increased) [112].

All these observations support the idea that other environmental factors, as the gut microbiome, could play a role in CD pathogenesis.

A number of studies have identified significant alterations in the composition of the gut microbiota in CD patients with respect to healthy individuals [113,114]. As for other diseases, also in this case the relationship between intestinal microbes and innate immunity through their

role in promoting inflammation and impairing mucosal barrier functions could clarify CD pathogenesis and give novel opportunities for patients care.

II. AIMS

The aim of this PhD project was to use next generation sequencing (NGS)-based strategies to study the molecular basis of human diseases with the specific aims to: i) increase diagnostic sensitivity, respect to commonly used techniques; ii) identify novel disease-causing genes; iii) identify phenotype modifier genes able to justify the heterogeneity of clinical signs; and iv) clarify pathogenetic mechanisms involved in disease development. To do these, we used two different NGS-based approaches: targeted DNA sequence capture and metagenomics.

Targeted DNA sequence capture was used to assess the feasibility of this approach for the study of a large panel of candidate target genes possibly related to complex (highly heterogeneous from both molecular and clinical points of view) inherited diseases. We studied, as model disease, the inherited cardiomyopathies. Particularly, we carefully evaluated results reliability to assess methods portability into routine diagnostic settings. In addition, we evaluated the feasibility of pooling together barcoded DNA samples from various patients in order to reduce the cost and time of the procedure.

Metagenomics, through the 16S bacterial rRNA analysis, was used to verify the presence of a specific dysbiosis into the composition of human gut microbiome in association to specific diseases. In this case, we used as model two inflammatory diseases: Crohn and celiac disease. In particular, in Crohn disease we aimed to assess the effects of nutritional therapy on the microbiome composition.

III. MATERIALS AND METHODS

3.1 Patients selection and biological samples collection

3.1.1 Targeted DNA Sequence Capture

Three unrelated individuals with a clinical diagnosis of hypertrophic cardiomyopathy (HCM) were selected for the present project with the aim to validate a targeted DNA sequence capture approaches for the study of inherited cardiomyopathies. HCM was defined as unexplained left or left and right ventricular hypertrophy in the absence of any potential cause of cardiac hypertrophy. It was diagnosed by echocardiographic evidence of increased wall thickness, two standard deviations or more above the upper reference limit of healthy individuals matched for age, sex, and body surface area. Maximal ventricular wall thickness (MWT) was defined as the greatest thickness in the various segments and measured as absolute value and as z-score for age and body surface area; z-score reference value < +2 [115]. Normal reference ranges of specific electrocardiographic (ECG) features were: PR: 0.12-0.20 sec; QRS: 0.06-0.10 sec; QTc: ≤ 440 msec in men, ≤ 460 msec in women.

Patient 1 was diagnosed with unexplained left ventricular hypertrophy (LVH) and cleft mitral valve during intrauterine life. HCM with asymmetrical LVH was confirmed at birth (MWT anterior interventricular [IV] septum was 8 mm at; z-score +7). This patient was followed-up at the Cardiomyopathy Clinic of the Monaldi Hospital (Naples, Italy) with regular ECG and echocardiographic evaluations. Verapamil therapy was started and the left ventricular wall thickness progressively normalized during childhood. No significant progression of LVH was seen

at follow-up. This is similar to a previous report [116]. At the last clinical evaluation, at 15 years of age, she showed the mitral cleft with concomitant mild-to-moderate regurgitation; her septal thickness was nearly normal (MWT at anterior IV septum was 12 mm; z-score +3.1). Patient 2 (now 8 years old) was diagnosed at birth with non-obstructive HCM and multiple septal ventricular defects (MWT at anterior IV septum at 4 months age was 8.3 mm; z-score +6.1). During follow-up, in the absence of cardiological therapy, LVH did not progress and all ventricular septal defects, except one, closed [116]. Her ECG showed signs of LVH and a long QTc (490 msec), which was considered "a secondary effect" of the cardiac hypertrophy. At last follow-up, MWT at the anterior IV septum was 8 mm (z-score +1.4). Patient 3 was diagnosed at birth with a valvular pulmonary stenosis, and underwent a surgical valvulotomy at 14 years of age. He was diagnosed with diabetes mellitus and polycythemia at 44 years of age. One year later, he had the first episode of atrial fibrillation/flutter. He also showed a first-degree atrioventricular block, an incomplete right bundle branch block (QRS 110 msec), and a long QTc (QT 495 msec). Echocardiography showed non obstructive HCM (MWT at anterior IV was 15 mm; z-score +4.4). The patient and his younger daughter also suffer from minor depressive disorders. Patient 3 underwent biochemical investigation and a muscle biopsy to exclude that LVH was of a secondary nature (biochemical evaluation: normal lactate/pyruvate ratio; normal CK, AST, ALT; muscle biopsy: negative COX and SDH fibers, normal mitochondria, no inclusion bodies). During follow-up, recurrent episodes of atrial fibrillation/flutter and a progressive end-stage evolution ("burn out") were observed; thus the patient underwent orthotropic heart transplant in 2011 at 51 years of age (MWT at anterior IV was 17 mm before heart transplantation; z-score +5.8).

Two of these patients (1 and 2) had previously been evaluated by using a DHPLC/Sanger test to search for causative mutations within the exons of 8 sarcomeric genes: *MYH7, MYBPBC3, TNNI3, TNNT2, TPM1, ACTC, MYL2* and *MYL3* [115]. The molecular analysis was extended also to the patient's relatives in order to study mutations segregation in each family. We obtained informed consent from each patient and family member, according to the procedures of the institutional review boards of the participating institutions.

Blood samples were obtained from each subject. Total DNA was isolated from peripheral blood using the Nucleon BACC3 Genomic DNA Extraction Kit (GE Healthcare, Life Sciences) according to the manufacturer's instructions. Then, samples quantification was done through the NanoDrop 2000c Spectrophotometer (Thermo Scientific).

3.1.2 Metagenomics

With regard the gut microbiome characterization of Crohn disease, two child where enrolled for the project. In particular, one child was affected by Crohn disease and underwent nutritional therapy, the other was a sex- and age-matched not affected subject. The Crohn disease patient was a 14-year-old boy diagnosed with active Crohn disease (Pediatric Crohn's Disease Activity Index, PCDAI=50). After colon endoscopy, he underwent nutritional therapy consisting of a daily powder constituted by proteins, antioxidants and anti-inflammatory fats (Alicalm formula, Nutricia Advanced Medical Nutrition) for 8 weeks. After this time, a clinical re-evaluation revealed disease remission (PCDAI= 0). Endoscopic ileum mucosal samples at diagnosis (BT-patient) and after therapy (AT-patient) were collected for DNA extraction. An ileum tissue sample was obtained

also from a 15-year-old boy affected by a gut polyp and without familiarity for Crohn disease.

To study the mucosal gut microbiome of celiac disease (CD), 15 active CD patients, 10 clinical controls and 6 at gluten free diet (GFD) CD patients were recruited among patients attending the Departments of Gastroenterology of the Universities of Salerno and of Roma-Tor Vergata, Italy. The exclusion criteria for enrolment were: any known food intolerance apart from gluten, IgA deficiency, treatment with antibiotics, proton pomp inhibitors, antiviral or corticosteroids, or assumption of probiotics in the 2 months before the sampling time. Duodenal biopsies from all the enrolled individuals were collected during diagnostic endoscopy procedures.

Total DNA was extracted from all collected biopsies (3 mg/sample, duodenum for CD and ileum for Crohn disease respectively) using the QIAamp DNA mini Kit (Qiagen,Venlo, Netherlands), following the manufacturer's instructions.

3.2 NGS Library Preparation

3.2.1 Targeted DNA Sequence Capture

Gene selection and microarray design. A list of target candidate genes was produced by selecting all the genes related to cardiomyopathy onset and the genes coding for ionic channels, membrane receptors, growth factors and inflammatory and transcriptional factors [117-122]. Thus, 202 genes of interest were identified (Appendix 1), and a list with their refseq IDs was generated. Starting from this list, 2,250 genomic targets were selected, each corresponding to one or more exons, plus a flanking region of 500 bp at the 5' and 3' ends of each exon. These genomic coordinates,

including chromosome start and stop positions, were used to define unique probes for the final set of 4,790 probed regions (one or more for each target) for a total of about $3.9x10^6$ bp. The microarray was provided by Roche Nimblegen Inc. (Madison, WI, 454 Optimized Sequence Capture 385K Array). Probe uniqueness was assessed by comparing the probe sequence to the genome [123].

Enrichment of target sequences. The selected regions were captured according to the manufacturers' protocols using the NimbleGen Hybridization System (Nimblegen and Roche, NimbleGen Arrays User's Guide: Sequence Capture Array Delivery v3.2) [18]. The procedures is based on four main steps: i) preparation of a library of adapted DNA fragments; ii) library hybridization on the custom capture array; iii) enriched library fragments recovery; and iv) enrichment assessment. In brief, 21 µg of genomic DNA from each sample were sheared into small fragments (range size: 300-500 bp) through nebulization. After fragments ends polishing and purification (AMPure Beads, Agencourt), specific adapters (Adaptor A and B) were ligated to each fragments. After quality (DNA chip1000, BioAnalyzer, Agilent) and quantity assessment (NanoDrop, Thermo Scientific), each library has been amplified and hybridazed on the custom array for 72h. After incubation, stringent washes were performed to remove the unbound fragments, while the enriched fragments were eluted and amplified. Sample enrichment was evaluated according to the manufacturer's instructions, i.e. by measuring, through quantitative PCR analysis, the relative fold-enrichment of 4 control loci present in each capture array. The analysis was carried out on a LightCycler 480 real time PCR system (Roche).

3.2.2 Metagenomics

An aliquot of the tissue DNA was used for PCR amplification and sequencing of bacterial 16S rRNA gene. To deeply investigate the bacterial composition of gut mucosal samples, a 548 bp amplicon, spanning from V4 to V6 variable regions of the 16S rRNA gene, was amplified using the 519F and the 1067R primers [124]. Both primers were modified to obtain fusion primers so that each one contained at the 5' end a universal 454 adaptor (adaptor A for forward primer and adaptor B for the reverse) and a specific 10 nucleotide tag/sample (Table 5).

Table 5. List of the primers' sequences used to amplify, by NGS methodology, the bacterial 16S V4-V6 region. For each couple the sequences are reported from 5' to 3', both on forward and reverse primers. Each primer is a fusion primer resulting from the following sequences (from left to right): i) the 454-Roche adaptor sequences (adaptors A and B) required for emulsion amplification and sequencing reactions (upper case characters); ii) the sample-specific 10 nucleotide tag sequences (MID 1-10) required to univocally tag each individual subject (underlined characters); iii) the primers' template-specific sequences (bold characters).

16S Primers for whole microbiota amplification	
F (5'-3')	**R (5'-3')**
Adaptor A MID (1-10) Template Specific Primer	**Adaptor B MID (1-10) Template Specific Primer**
CGTATCGCCTCCCTCGCGCCATCAG-ACGAGTGCGT-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-ACGAGTGCGT-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-ACGCTCGACA-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-ACGCTCGACA-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-AGACGCACTC-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-AGACGCACTC-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-AGCACTGTAG-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-AGCACTGTAG-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-ATCAGACACG-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-ATCAGACACG-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-ATATCGCGAG-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-ATATCGCGAG-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-CGTGTCTCTA-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-CGTGTCTCTA- **TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-CTCGCGTGTC-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-CTCGCGTGTC- **TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-TAGTATCAGC-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-TAGTATCAGC-**TGACGACAGCCATGC**
CGTATCGCCTCCCTCGCGCCATCAG-TCTCTATGCG-CAGCAGCCGCGGTAATAC	CTATGCGCCTTGCCAGCCCGCTCAG-TCTCTATGCG- **TGACGACAGCCATGC**

F: forward; R:reverse.

PCR reactions were carried out using 25 µl of H_2O, 20 µl of 2.5X HotMaster PCR mix (Eppendorf), 1.5 µl of each primer 10 µM and 60 ng of DNA. The amplifications were performed on a DNA ENGINE Chassis (Biorad) at the following conditions: 2 min at 94°C, 30 cycles of 94°C for 40 s, 50°C for 40 s, and 65°C for 40 s, and a final extension of 70°C for 7 min. After visualization by agarose gel electrophoresis, each PCR products was individually purified using magnetic purification beads (AMPure Beads, Agencourt), assessed for quality on a Bioanalyzer 2100 (DNA 1000 chip, Agilent) and quantified using the Quant-it PicoGreen dsDNA kit (Invitrogen). Equimolar amounts of each amplicon were pooled together to obtain multiple amplicon libraries, each containing a total of 5 mixed subjects.

3.3 NGS Library Amplification and sequencing

Both the NGS-based approaches used in the present project were carried out using the GS FLX System (454 Roche). Therefore, the obtained DNA libraries were amplified through emPCR and sequenced using the pyrosequencing chemistry, according to the manufacturer's specifications [10].

3.4 Bioinformatics

3.4.1 Targeted DNA Sequence Capture

Read mapping and variant detection. The obtained sequencing reads were mapped by using the GS Reference Mapper software package (454 Roche, version 2.6), with the default parameters. Variants were identified by comparing assembled vs reference genomic sequence (hg19, http://genome.ucsc.edu/cgi-bin/hgGateway). The procedure produces a list of high confidence nucleotide differences (HCDiffs), as well as a larger list of all nucleotide differences (AllDiffs), identified by a less stringent approach. The GS Reference Mapper uses a combination of flow signal and quality score information together with the type of nucleotide change to determine if a variant is to be treated as a high-confidence variant. The strategy used to select HCDiffs is based on the following criteria: 1) there must be at least 3 non-duplicate reads with the difference, with at least 5 bases on both sides of the difference, and a few other isolated sequence differences in the read; 2) there must be both forward and reverse reads showing the difference, unless there are at least 7 reads with quality scores over 20 (or 30 if the difference involves a homopolymer of 5 nt or more); 3) not all overcalls/undercalls are reported as insertions/deletions, but only those where the difference is the consensus of the sequenced reads. The AllDiff strategy is less stringent and requires at least 2 non-duplicate reads, without restrictions in terms of forward/reverse strands. To facilitate downstream analyses, the complete list of variants (AllDiffs/HCDiffs), together with relevant information concerning target coverage, were imported in a relational database based on PostgreSQL (http://www.postgresql.org).

Genotype calling and statistical analyses. For each candidate polymorphic site, the genotype was assigned according to a procedure for evaluation of statistical significance of deviations from the predicted distribution of reference-supporting/alternative-supporting read frequency in the case of alternative-homozygous (a/a), heterozygous (r/a) and reference-homozygous (r/r) subjects. The procedure was based on a binomial exact test where the probability (P) of producing the observed frequencies is calculated starting from three alternative hypotheses [125], corresponding to a ratio of 0.99, 0.5 and 0.01 for a/a, r/a and r/r respectively, as one would expect for a sequencing error rate of 0.01 [126]. The hypothesis with the highest P-value was taken as the result; a quality score was calculated as the logarithm with base 10 of the ratio between the highest and second best P-value. P-values were calculated by using the *binom.test* function from the R statistical software environment (http://www.r-project.org/).

Assessment of the NGS methodology. The sensitivity and specificity of the described NGS-based procedure was assessed by evaluating its ability to identify all base changes present in a subset of eight genes, previously associated with HCM[115]. This test was carried out on two patients in which the same eight genes had previously been analyzed[115], by using a combination of DHPLC and Sanger sequencing. For both procedures, the analysis was targeted to 128 exons extended by 60 flanking bases on each side, for a total of 34,067 bp (Table 6).

Table 6. Summary of the features of the genomic regions analyzed by both NGS and DHPLC/Sanger methods. These regions were used to evaluate NGS reliability with best estimate method.

	NGS	DHPLC/Sanger
	Target	
Target exons	128	128
Target bases	34,067	34,067
	Experimental Design	
Segments	2245	125
Bases (n)	3,897,552	38,013
Targets in design	128	128(54*)
Target bases in design	34,067	29,014

*number of targets partially covered.

We evaluated both the NGS-based and DHPLC/Sanger procedures by testing their ability to correctly detect all base changes present in the patients' sequence. No technique clearly better than the two procedures being tested is yet available to be used as a "gold standard". Therefore, for each patient a "best estimate" sequence was obtained by combining and reviewing for each position all the available experimental data. Specifically, when automated sequencing data univocally identified a nucleotide by all methods, it was accepted as such; in all other cases comparative analysis of the results was carried out by hand. When data from Sanger and NGS sequencing were both available, but divergent, the results could often be reconciled by visual inspection of electropherograms (Sanger) and multiple alignment of NGS reads, and taking into account the most common interpretation mistakes, namely small peaks, too few or conflicting reads and so on. In these cases the resulting nucleotide was accepted and included in the best estimate sequence; when base and genotype assignment could not be unambiguously obtained, the position was not accepted and therefore

removed from the "best estimate" sequence. This effort produced a sequence that best fits the deepest level of analysis performed, and that was used as a reference to assess the sensitivity and specificity of the NGS and the DHPLC/Sanger procedures, by counting as true positive (TP), true negative (TN), false positive (FP) or false negative (FN) the results available from the each of the two procedures, taken independently. Sensitivity and specificity were then calculated according to the conventional formulas: Sensitivity = TP/(TP+FN); Specificity = TN/(TN+FP). Confidence intervals for sensitivity and specificity were estimated by using the Pearson-Klopper method, and the R statistical software environment.

The aim of the DHPLC approach we used was to identify rare exon sequence changes; therefore some variants identified by NGS were difficult to detect or undetectable as they are located: a) close to the ends of the amplicons analyzed by DHPLC; b) homozygous; or c) within untranslated sequences. To better assess DHPLC sensitivity making the results of calculations comparable to those of other reports, these calculations were repeated by excluding these variants, as reported under Results.

Variant annotation. The potential consequence of nucleotide substitutions, insertions and deletions on transcripts and other functional elements was examined using the "Variant Effect Predictor" tool [127], version 2.5, and the release 67 (May 2012) of transcripts and regulatory regions annotated in Ensembl (http://www.ensembl.org). The same tool was used to map variant loci to previously described SNPs. The allele frequency in each of the 1,000 Genomes super-populations (EUR, which is the population closest to our three subjects, ASN, AFR and AMR) was

annotated, when available, by using a custom developed script based on Ensembl's perl API (http://www.ensembl.org) [128].

3.4.2 Metagenomics

16S rRNA barcoded amplicon sequences were analyzed using QIIME v. 1.8.0 [129]. Sequences were quality filtered and demultiplexed using default QIIME parameters. The filtered reads were assigned to operational taxonomic units (OTUs) using an open-reference OTU picking approach on the basis of sequence similarity using UCLUST [130] against the Greengenes database (v. 13_5) [131] at 97% identity. A representative set of sequences was taken for each OTU and taxonomic classification was performed with the Ribosomal Database Project (RDP) classifier 2.2 [132]. Next, the representative sequences were aligned using PyNAST [133] to the Greengenes Core reference alignment [134] and a phylogenetic tree was built using FastTree [135]. The phylogenetic tree was used for downstream phylogenetic community analyses. Community diversity analyses at a rarefaction depth of 854 sequences/sample were performed using QIIME related scripts. Principal coordinates analyses (PCoA) were generated from UniFrac weighted and unweighted distance matrices [136].

In order to assess whether specific taxa were significantly differentially abundant across study groups, we used the analysis of variance (ANOVA) test together with a Bonferroni correction using QIIME scripts. OTUs with p-value ≤ 0.05 after Bonferroni adjustment were considered statistically significant. To compare the alpha diversity across sample groups, a non-parametric two sample t-test was run using 999 Monte Carlo permutations to calculate the p-values. Analysis of Similarity Statistics (ANOSIM) [137] and non-parametric multivariate ANOVA (Adonis) [138]

using Unifrac distance matrices were used to test the significance of differences in the beta diversity.

IV. RESULTS

4.1 Targeted DNA Sequence Capture

NGS analysis. Each of the three selected patients and their genomic pool DNA sample were separately analyzed as described under Materials and Methods. Target DNA enrichment of the 202 selected genes was carried out using a array-based hybridization method. Capture efficiency in the three patients, as assessed by quantitative PCR analysis, typically ranges between 280 and 330 fold, well above the suggested 200 fold threshold indicative of an acceptable capture prior to sequencing. The pooled DNA sample yielded a 390-fold enrichment. The amount of captured DNA was measured by spectrophotometry, and the average yield was 10 μg per sample. Two sequencing runs were performed with each sample loaded in one large PicoTiterPlate region. This procedure yielded on average 203 Mb/sample, with a number of sequencing reads around 650,000/sample with a very similar read length of ~330 bp (Table 7). The same analytic strategy was used for the DNA pool. The pooled sample was sequenced in duplicate on a PicoTiterPlate in a single run. The procedure yielded more than 307 Mb, being equivalent to 1,067,389 sequencing reads with an average size of 298 bp (Table 7). For all samples, over 90% of the reads were unambiguously mapped on the human genome reference sequence, while a fraction ranging between 72.8% and 76.5% of all reads fell within the targeted regions. Overall, this corresponds to an average enrichment of the targeted DNA of 595-fold, calculated as the ratio between the fraction of reads mapped on the targets and the fraction of the human genome represented by them. With this approach, a variable length of DNA

sequence, adjacent to both ends of the targets, was also usually determined, in addition to targeted sequences. Average sequencing depth within the target area ranged from 32.8x to 36.1x for the three patients (Table 7), while the number of "covered" targets varies depending on the depth that is defined as "acceptable", and that is used as a threshold (Figure 9). If a minimum accepted depth of 10x is chosen, at least 91.8% of the targeted bases were covered for all patients (Figure 9A). These values are similar to, or even higher than, the coverage reported in other studies aimed at detecting sequence variations in inherited diseases [139-144]. Most low depth or uncovered bases were located within a small number of targets, ranging between 32 and 63 for the three patients. These targets were completely unsequenced, probably because of low efficiency capture in the enrichment procedure (Figure 9B). The average sequencing depth of the pooled sample was 45.4x, as expected from duplicate analysis (Table 7).

Table 7. Overview of the entire sequencing and annotation procedure. The table shows the results obtained for patients P1, P2, P3 and the pool containing DNA from the three patients plus a control. The target overall gene regions (4Mb) include the 202 genes selected in this study (see Materials and Methods and Appendix 1).

	P1	P2	P3	Pool
qPCR enrichment (fold)	319	333	288	390
Total Sequenced bases (bp)	197,796,849	215,720,477	197,375,419	307,280,122
Total Sequenced reads (n)	629,332	687,163	627,170	1,067,389
Average reads length (bp)	328.41	325.54	329.43	298.02
Reads mapped on genome (n/%)	584,128/92.8	646,491/94.1	582,455/92.9	1,002,244/93.9
Reads mapped on targets (n/%)	458,343/72.8	525,851/76.5	474,050/75.6	793,459/74.3
Target DNA enrichment (fold)	578	607	600	579
Average depth (fold ± s.d.)	32.8 ± 16.3	34.2 ± 17.3	36.1 ± 18.9	45.4± 21.4
Target coverage with depth >0 (%)	99	99	99	99
Target coverage with depth >8 (%)	93	94	93	96
Target coverage with depth >15 (%)	84	86	85	91
Total identified bp differences (AllDiffs)	20,500	19,654	20,904	30,580
High Confidence bp differences (HCDiffs)	6,894	6,157	6,371	9,115
Variants in final set (all/novel)	3,350/1,456	2,721/985	2,902/1,204	-

qPCR, quantitative PCR analysis; P1, P2, P3, patients; HC, high confidence; n, absolute number; s.d., standard deviation.

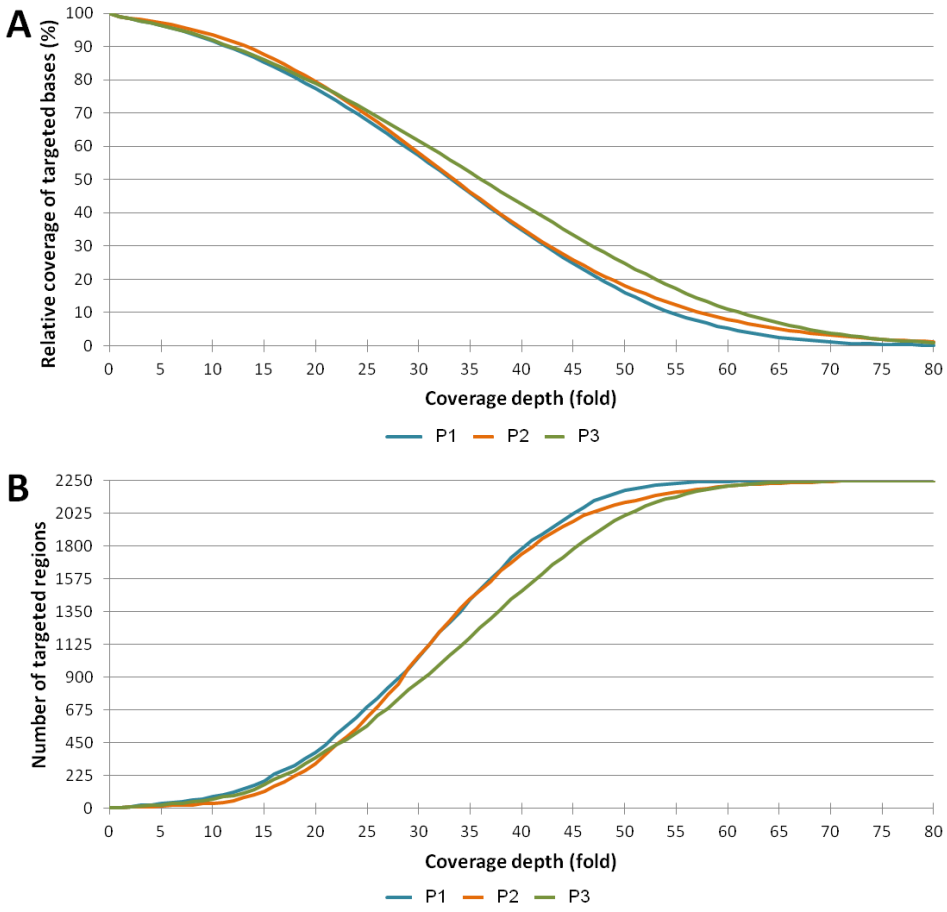


Figure 9. Coverage of targeted bases and regions. A) Relative coverage of bases contained within targeted regions is shown for each of the three individuals. B) Number of target regions vs average coverage depth, right cumulated. The color code is referred to each of the three patients that show a very similar behavior.

Variant detection, genotype calling and variant annotation: prioritization pipeline. Candidate variants were detected by using the GS Reference Mapper software. As reported in Table 7, of the many hypothetical variants reported in the AllDiff list, 6,894, 6,157 and 6,371 were selected by the HCDiff selection procedure in patients 1, 2 and 3 respectively. For each HCDiff variant, the genotype was defined according to a significance evaluation test (see Materials and Methods). This test was used to identify additional potential detection errors, i.e; r/r loci that

erroneously passed the HCDiff filter. In fact, a small number of variant loci (92, 69, 77 respectively) are better designated as reference-homozygous (r/r) (Table 8).

Table 8. Genotype assignment. Genotype was assigned for the various high confidence (HC) bp differences (Diffs), based on binomial exact test and reported separately for each patient.

Subject	HCDiffs	HCDiffs r/r*	HCDiffs r/a†	HCDiffs a/a‡
P1	6,894	92	3,714	3,088
P2	6,157	69	3,500	2,588
P3	6,371	77	3,626	2,668

*Homozygous for reference allele; †Heterozygous; ‡Homozygous for non-reference (alternative) allele.

In each individual, between 1,036 and 1,530 variants correspond to putatively novel alleles not reported in the Ensembl variation database (release #67). A frequency filter was used to exclude "variants" that are either the most frequent allele in populations close to the population of our three subjects (southern Italian), or present with a frequency too high to be consistent with pathogenicity within these or other populations. For each subject, the number of variants with a non-reference allele frequency greater than a given threshold value (0.5, 0.1, 0.05 and 0.01) is reported in Table 9.

Table 9. Evaluation of common allele frequency. High Confidence reference/alternative (r/a), alternative/alternative (a/a) is the set of HC variants refined by removing variants which did not pass the binomial exact test. EUR indicates the number of variants which could be discarded because the supposed alternative allele has frequency greater than the indicated threshold in the 1000 Genomes phase 1 EUR (Europeans) super population; similarly ASN !EUR, AFR !EUR, AMR !EUR and ALL !EUR are variants, where the supposed alternative allele frequency is greater than the indicated threshold, for ASN (East Asian), AFR (African), AMR (Ad Mixed American) and ALL (all together) super populations respectively. For the non-European populations only variants not already above threshold in the EUR super population have been counted.

Subject	HC r/a, a/a	Allele frequency	EUR	ASN !EUR	AFR !EUR	AMR !EUR	All !EUR
P1	6,802	>0.50	1,854	276	356	139	517
		>0.10	3,312	80	119	35	170
		>0.05	3,452	66	55	42	91
		>0.01	3,589	15	12	22	30
P2	6,088	>0.50	1,767	254	336	128	496
		>0.10	3,166	107	160	95	200
		>0.05	3,367	38	83	25	99
		>0.01	3,515	7	36	18	44
P3	6,294	>0.50	1,790	361	354	163	547
		>0.10	3,238	65	121	57	160
		>0.05	3,392	34	73	37	88
		>0.01	3,540	4	18	14	26

P, patient; HC, high confidence; r/a, reference/alternative; a/a, alternative/alternative.

Variants in which the alternative allele frequency is greater than 0.05 in the EUR super-population were filtered out to produce the “Final” set (3,350, 2,721 and 2,902 variants for patient 1, 2 and 3 respectively). This multistep selection/annotation procedure is reported in Figure 10.

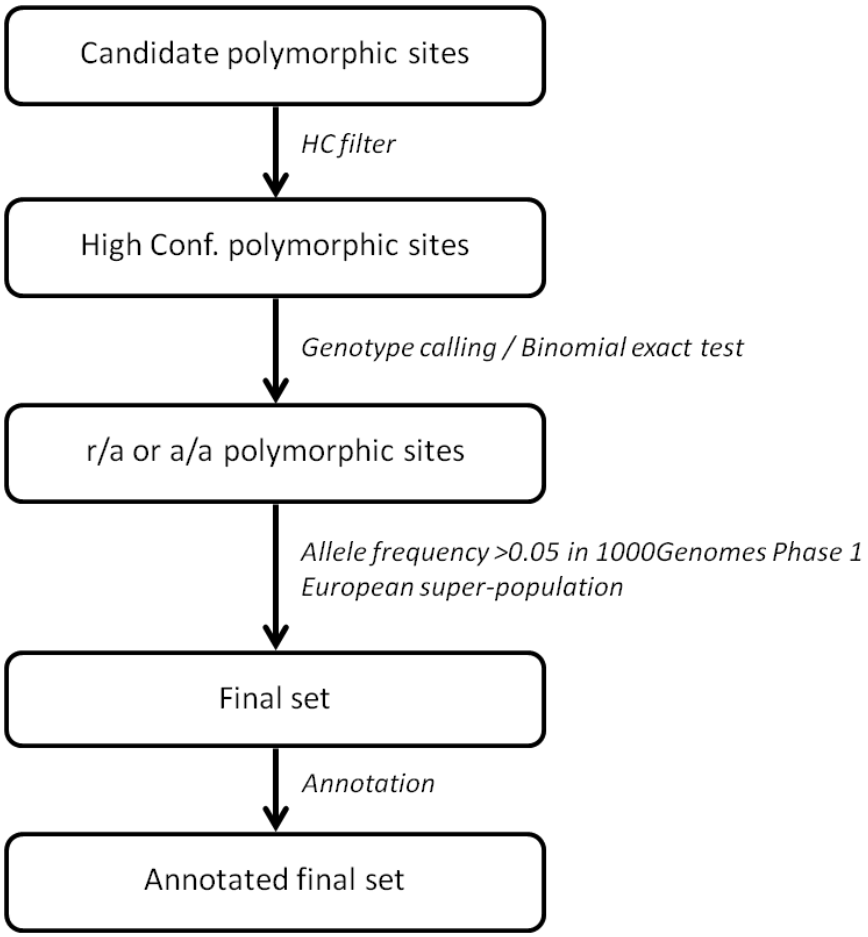


Figure 10. Variants prioritization pipeline.

The above mentioned pipeline produces a prioritized short-list of well-annotated variants that may be used to search for pathogenetic changes related to HCM onset and/or clinical features (Table 10).

Table 10. Number of variants in the final set annotated according to their predicted features.

Patients		P1	P2	P3
Number of variants in the final set		3,350	2,721	2,902
Protein coding	Synonymous	123	128	131
	Non-synonymous (possibly deleterious)*	60 (11)	61 (8)	73 (7)
	Frameshift, Loss/Gain of stop codon	6	3	6
Transcript	Intron/UTR	1,638	1,518	1,550
	Splice site	15	15	18
	Complex In/del	0	1	0
Genome	Intergenic	1,560	1,014	1,140
	Upstream/Downstream	479	422	429
	Regulatory region[†]	516	313	386
	Within nc-transcript[±]	674	633	668

*Non-synonymous is referred to a variant resulting in an amino acid change in the encoded peptide sequence; it is classified as being possibly deleterious when SIFT (http://sift.jcvi.org/) and Polyphen (http://genetics.bwh.harvard.edu/pph2/) predictions on the protein function are concordant. [†]Regulatory region corresponds to those annotated as such in ENSEMBL, being inferred from experimental data through the Regulatory Build process (http://www.ensembl.org/info/docs/funcgen/regulatory_build.html).
[±]nc-transcript is a non coding transcript annotated in ENSEMBL (http://www.ensembl.org) that does not contain an open reading frame.

Reliability of the NGS-based procedure. The reliability of the NGS-based procedure was assessed by validating the variant bases identified within 8 sarcomeric genes against a "best estimate" sequence obtained for each patient by comparative assessment of all the experimental data as described under Materials and Methods. The results of all parameters measured with both methodologies are reported in Table 11.

Table 11. Performance Indexes of DHPLC/Sanger and NGS methods in nucleotide sequence-variants detection.

	P1			P2		
	True nucleotide change: n=26*	Wild –Type sequence: n= 34,041*		True nucleotide change: n=23*	Wild –Type sequence: n=34,044*	
Method						
DHPLC/Sanger						
Positive	TP: 14	FP: 3	PPV: 82.4%	TP: 14	FP: 2	PPV: 87.5%
Negative	FN: 12	TN: 34,038	NPV: 99.9%	FN: 9	TN: 34,042	NPV: 99.9%
Sensitivity % (95% CI)	53.8 (33.4-73.4)			60.9 (38.5-80.3)		
Specificity % (95% CI)	99.9 (99.974-99.998)			99.9 (99.979-99.999)		
NGS						
Positive	TP: 25	FP: 1	PPV: 96.2%	TP: 22	FP: 1	PPV: 95.7%
Negative	FN: 1	TN: 34,040	NPV: 99.9%	FN: 1	TN: 34,043	NPV: 99.9%
Sensitivity % (95% CI)	96.2 (80.4-99.9)			95.7 (78.05-99.89)		
Specificity % (95% CI)	99.9 (99.997-99.999)			99.9 (99.984-99.999)		

P1, P2, patient identification code; TP, true positive; FN, false negative; FP, false positive; TN, true negative; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.
*Numbers refer to bp evaluate with best estimate method, as described under Materials and Methods.

The sensitivity of the NGS-based procedure was 96.2% in patient 1 and 95.7% in patient 2; specificity was very close to 100% in both patients (Table 11). Also predictive values are very satisfactory, being PPV (positive predictive value) 96.2% and 95.7% in the two patients, and NPV (negative predictive value) 99.9% in both patients. The single false negative result obtained in both patients corresponds to an undetected deletion of one G in a four-G strand. Similarly, the single false positive result obtained in the two patients is the same and is localized within a 20-fold CA repeat. No mistakes occurred in single-base variants. The two mistakes reflect well

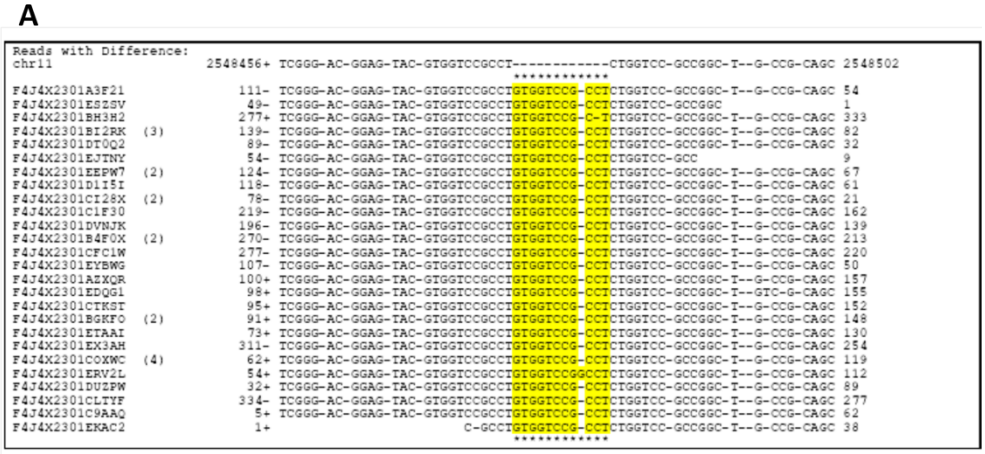
known shortcomings of the NGS sequencing technique [140]. The previously used combined DHPLC/Sanger method showed a very high specificity (practically 100%) comparable to that of the NGS method, but its sensitivity was significantly lower (53.8 and 60.9%); in fact, 12 and 9 variants were not indentified in patient 1 and 2 respectively (Table 11). Further analysis of the DHPLC/Sanger false negatives revealed that 8 and 6 of them respectively are located in a small subset of non coding genomic segments (half of them in a single very long 3‘UTR of the *ACTC1* gene) that, although included in the initial target list, were outside the amplicons obtained for the DHPLC/Sanger analysis, although they were immediately upstream of their 5' UTR or downstream of their 3' UTR. Therefore, these nucleotide position failed to produce an unequivocal positive or negative result but they were not retested due to their non-coding nature. Two additional false negatives per patient are homozygous variations, not detectable by the DHPLC procedure used, which was originally designed to identify rare heterozygous changes. If these variants are not taken into account, the sensitivity of the DHPLC/Sanger method increases to 87.5 and 93% in patient 1 and 2 respectively; these revised values are close to those obtained by others using the same technique [145], but are still lower than those obtained with the NGS-based procedure described herein (Table 11).

Mutation identification and genotype/phenotype correlations. The downstream analysis of the final set of variants revealed a small number of variants that produced potentially “deleterious” effects on protein function (number of variants equal to 11, 8 and 7 in patient 1, 2 and 3, respectively). In patient 1 we detected a variant in the *MYH7* gene (c.976G>C; p.A326P) that produces an alanine-proline change. Interestingly, we failed to identify this mutation in our previous DHPLC analysis of this region [115]. This

mutation, inherited from the patient's healthy father, has been previously described in HCM patients [146]. Consequently, we suggest p.A326P is the disease-causing mutation in patient 1. This patient also carried a rare variant very close to the 3' end of the coding region, at the 20th nucleotide in the 3' UTR (c.5808+20G>A) of the *MYH7* gene. This variant is reported in the 1000-Genome Phase 1 data set with a frequency of 0.8% in the EUR super-population. We found c.5808+20G>A in only two alleles from among 520 chromosomes of healthy individuals (a frequency of only 0.38%), and hence it does not fall within the definition of a single nucleotide polymorphism. This variant was absent in the proband's father, the only parent available for genetic testing. Patient 1 carried a third variant, a heterozygous nonsense mutation, in the conjoined *INS-IGF2* gene (namely c.575C>T; p.Q172X) that occurs 28 amino acids before the 3' end of the coding region. Also this variation was not found in the patient's father. The functions of *INS-IGF2* are not completely known, although it seems to be involved in the regulation of insulin expression [147,148], and thus in diabetes [148,149]. Glucose homeostasis is not altered in patient 1, but periodic screening of blood glucose level would seem warranted in this case.

In patient 2, the expected *MYBPC3* c.3627+2 T>A variant previously identified with the DHPLC/Sanger approach [115] was readily identified as a splicing-related variant. As we previously reported [115], this mutation causes exon 32 skipping and produces a shorter mRNA than the wild-type mRNA. We identified and confirmed a second mutation in patient 2, namely an indel/frameshift, in the *KCNQ1* gene (g. 2548481_2548491dup, Figure 11).

A



B

Chr	11
UCSC Start (nt)	2548481
UCSC End (nt)	2548492
Reference (nt)	-
Variant (nt)	GTGGTCCGCCT
Reference (aa)	LWSAG
Variant (aa)	LWSAS…
Total reads	49
Variant reads	26
Variant reads (%)	53
HC variant	Yes
Variant status	Heterozygous
Gene	KCNQ1
Localization	Exonic
Classification (nt)	Indel- Frameshift

C

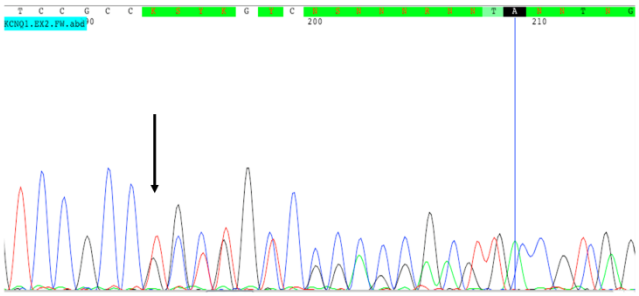


Figure 11. Detection pattern of the KCNQ1 sequence duplication. KCNQ1 c. 524_534dup was identified in patient 2. A) Sequence alignment of the reads in the allele carrying the variant against the reference sequence (first line). The duplicated nucleotides are marked in yellow. B) List of all variant features, including genomic coordinates and the number of total and variant reads (coverage). C) Duplication was confirmed by Sanger sequencing. Arrow indicates the duplication start points. HC, high confidence; nt, nucleotide.

This mutation consists of a duplication of nucleotides GTGGTCCGCCT at position c.524-534, previously reported in a patient with long-QT syndrome [150], which results in a premature stop codon (p.W176Lfs*65). The retrospective examination of the patient's ECGs showed a long QT trait that was interpreted simply as a secondary effect of

hypertrophy. Examination of patient 2's family revealed the independent inheritance of these two variants (Figure 12).

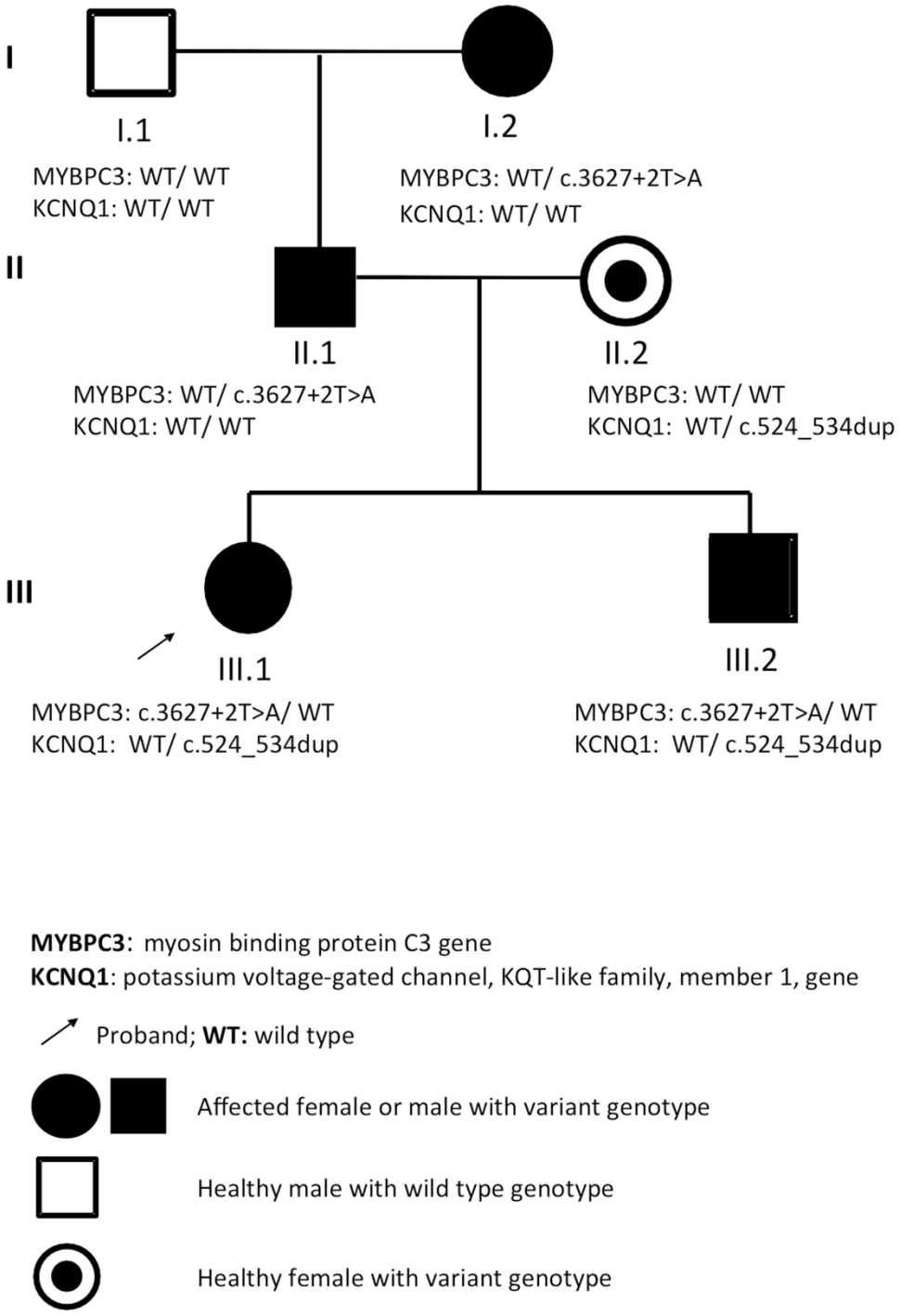


Figure 12. Pedigree of patient 2's family. Familial segregation shows the MYBPC3 and KCNQ1 genotypes. The two variants were independently inherited by the patient and his HCM-affected brother. The arrow indicates the proband.

The patient's mother carries the *KCNQ1* variant. Her ECG showed no sign of LVH or of a long QT trait, but a marked sinus bradycardia (50 bpm) without a history of athleticism. Holter ECG showed marked sinusal bradycardia (medium 53 bpm) without signs of a long QT syndrome, conduction disorders and/or arrhythmias; therefore she was enrolled in a cardiomyopathy monitoring program.

We did not find variations in the conventional sarcomeric genes in patient 3, but we did find and confirm by Sanger sequencing a mutation previously associated with the short QT and Brugada syndromes [151], and identified in 2 of 6,752 Caucasians (rs121912775) in the ESP exome sequencing project (http://evs.gs.washington.edu/EVS/), the *CACNA1C* p.G490R, g.2529447G>A (Figure 13).

A

```
Reads with Difference:
chr12                    2529425+ TGAC-ATCG-AGGG--AGA-AAA--CTGC-GGGG-CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 2529477
                                                            ******
F4J4X2302I0GRH                435+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGC                    470
F4J4X2302JXZNL   (2)          364+ TGACGATCG-AGG---AGA-AAA--CTGCAGG---CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 415
F4J4X2302J4V1J                343+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 395
F4J4X2302GMGOM   (2)          256+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 308
F4J4X2302IA6JV                241+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 293
F4J4X2302HO8NJ                249+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 301
F4J4X2302GD63N                159- TGAC-ATCG-AGGG--AGA-AAA-ACTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 106
F4J4X2302H4WLE                201- TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 149
F4J4X2302J0V7B                219- TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 167
F4J4X2302GQAZL                229- TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 177
F4J4X2302FUGAR                204+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 256
F4J4X2302HI9V3                267- TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 215
F4J4X2302HXYD6                209- TGAC-ATCG-AGGGN--GA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 157
F4J4X2302I8H30                152+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 204
F4J4X2302H783W                140+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 192
F4J4X2302I6X51                 43- TGAC-ATCG-AGGG--AGA-AA-G-CTGCAGGG--CCA-GGCTGGCGTGAGTA              1
F4J4X2302FSEMX                106+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 158
F4J4X2302FP5FQ                101+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 153
F4J4X2302HA2HW   (3)          356- TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 304
F4J4X2302FHEGO                 80+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 132
F4J4X2302IC2E8   (2)           73+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 125
F4J4X2302G98Y1                 69+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGCG--CGGCG  120
F4J4X2302IFGX0                 50+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 102
F4J4X2302HRFDM                375- TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 323
F4J4X2302I2TMC                 34+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCAAG-CTGGCGTGAGTAGGC-A-CGGCGA 86
F4J4X2302G50SQ   (3)           27+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 79
F4J4X2302G27EC                 20+ TGAC-ATCG-AGGG--AGA-AAA--CTGCAGGG--CCA-GGCTGGCGTGAGTAGGC-A-CGGCGA 72
                                                            ******
```

B

Chr	12
UCSC Start (nt)	2529447
UCSC End (nt)	2529447
Reference (nt)	G
Variant (nt)	A
Reference (aa)	G
Variant (aa)	R
Total reads	49
Variant reads	27
Variant reads (%)	55
HC variant	Yes
Variant status	Heterozygous
Gene	CACNA1C
Localization	Exonic
Classification (nt)	Substitution
Classification (aa)	Missense

C

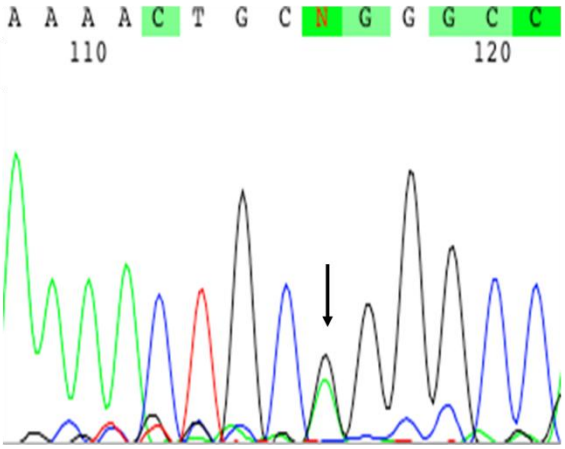


Figure 13. Detection of a missense mutation in the CACNA1C gene. The CACNA1C p.G490R variant was identified in patient 3. A) Alignments of all the reads in the allele carrying the variant against the CACNA1C reference sequence (first line). The yellow lettering indicates the variant nucleotide. B) List of all variant features, including genomic coordinates and the number of total and variant reads (coverage). C) Sanger electropherogram from the same patient confirming the variant. Arrow indicates the variant nucleotide. HC, high confidence; nt, nucleotide.

Patient 3 carried a second missense variant in the *CACNA1C* gene: the novel p.E771G variant, located in the second membrane-spanning domain that contributes to pore formation, which is essential for the channelling function of the protein. *CACNA1C* variants have been linked to a number of cardiac phenotypes, including the Brugada and Timothy

syndromes [151-154]. Patient 3 showed a complex systemic phenotype with features that resembled both syndromes. Several *CACNA1C* polymorphisms are significant risk factors for bipolar disorders, schizophrenia and major depressive disorders [155]. With the caveat that the role of these two *CACNA1C* should be verified in functional assays, it is conceivable that, by modifying intracellular calcium concentration [151], the variant protein could impair myocyte contractility, which in turn could lead to ventricular hypertrophy. We were able to extend variant analysis to the patient's two daughters, who were tested by Sanger sequencing; the younger one (24 years old) was identified as a carrier of the two variations. Electrocardiographic and echocardiographic evaluation showed no signs of cardiac disease; however, like her father, she has a mild depressive disorder. Therefore, this daughter has been enrolled in a clinical follow-up program that includes identification of early signs of HCM.

Table 12 summaryze the most interesting variants identified in the three analyzed subjects.

Table 12. Mutations most likely to exert a pathogenic role in the patients analyzed.

Pts	Chr	Gene	Mutation type	Variation* (inheritance)	Amino acid change
P 1 †	14	MYH7	Missense	c.976G>C (paternal)	p.A326P
	11	INS-IGF2±	Nonsense	c.573C>T	p.Q172X
P 2	11	MYBPC3	Splicing site	c.3627+2T>A (paternal)	/
	11	KCNQ1	Indel-Frameshift	c.524-534dup (maternal)	p.W176Lfs *65
P 3 †	12	CACNA1C	Missense	c.1781G>A	p.G490R
	12	CACNA1C	Missense	c.2625A>G	p.E771G

Pts, patient; Chr, chromosome. *All mutations were at heterozygous state. †Patient 1's biological mother and patient's 3 parents were not available for

molecular testing. [±]Mutation included because it produces a truncated protein.

Analysis of the pooled DNA sample. The same individually sequenced DNAs from the three patients were also sequenced as a pooled DNA sample, which also contained a fourth control sample DNA. In this experiment, the set up produced an expected depth about double that of each single patient. As shown in Table 7, the fraction of covered targets is similar to that observed for each separate subject. The number of variants identified for each subject is shown in Table 13.

Table 13. 'HC' and 'Final' variants identified for each subject in the pool. The obtained values are compared to those expected on the basis of the results of the single subject sequencing experiment.

Dataset	Expected	Found	Found (%)
P1 HC	6,894	5,194	75.3
P1 Final	3,350	2,113	63.1
P2 HC	6,157	4,899	79.6
P2 Final	2,721	1,930	70.1
P3 HC	6,371	4,974	78.1
P3 Final	2,902	1,941	66.9

Comparative analysis of the identified variants (Table 14) shows that most allelic variants common to more than one patient were also identified in the pool. However, more variants were missed as their presence in the 8 analyzed alleles decreased.

Table 14. Analysis of non-reference (variant) alleles found in single subject sequencing experiments and also identified in the pool. Variants missed in the pool (last column) are increasing when they are read on a rarer number of alleles.

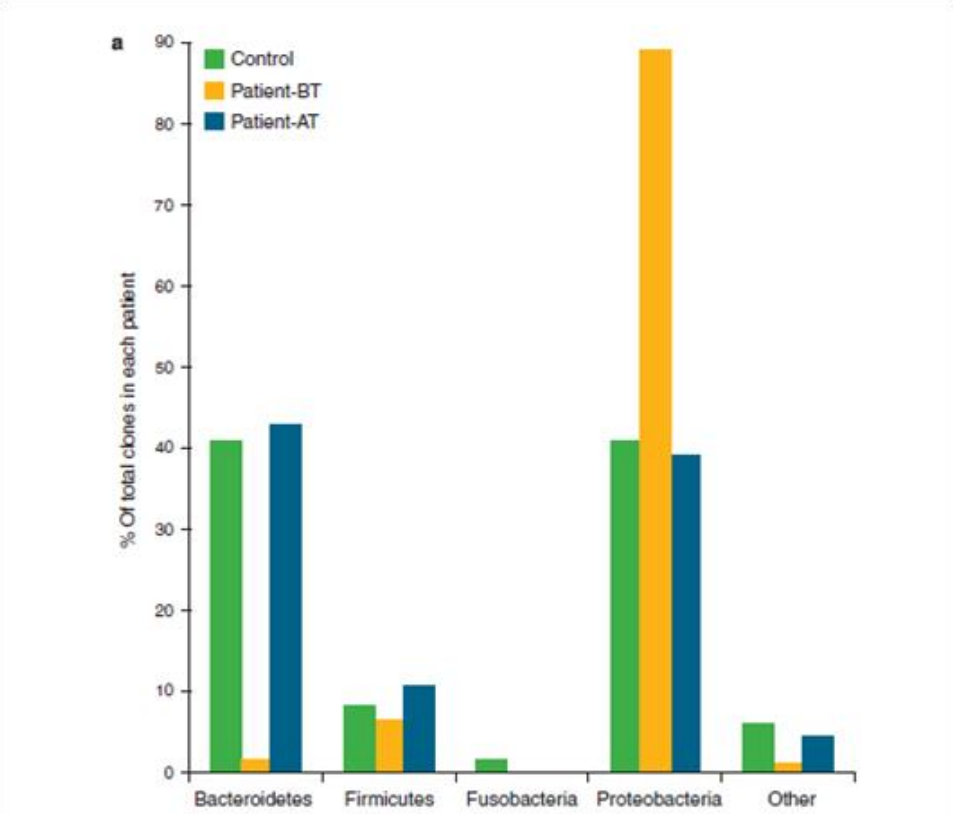
	Pool Variants		
Alleles	**Expected***	**Found**	**Missed**
8/8	683	677	6
7/8	353	353	0
6/8	669	659	10
5/8	717	695	22
4/8	1,182	1,077	105
3/8	1,090	966	124
2/8	2,976	1,756	1,220
1/8	4,414	1,705	2,709

*Expected in the pool from the analysis performed in the individual patient samples

4.2 Metagenomics

4.2.1 Crohn disease

A 16S rRNA NGS-based strategy was used to investigate the gut microbiome composition of the three collected ileum samples. High-quality filtered sequences were used to identify the OTUs: 705, 1,328 and 2,171 different OTUs were obtained in the patient before (BT) and after therapy (AT), and in the control subject, respectively. Based on the taxonomic assignment of the OTUs, we characterized the ileum microbiome of our samples at five phylogenetic levels (from phylum to genus). As shown in Figure 14, *Proteobacteria* were more abundant, and *Bacteroidetes* less abundant, in our CD patient BT than in the control. Interestingly, the composition of the ileum microbiome in the patient AT was virtually the same as in the control (Figure 14).



b

Consensus Lineage from phylum to family	Control	Patient-BT	Patient-AT
P->Bacteroidetes;C->Bacteroidia;O->Bacteroidales;F->Bacteroidaceae	2449	69	2462
P->Bacteroidetes;C->Bacteroidia;O->Bacteroidales;F->Porphyromonadaceae	10		367
P->Firmicutes;C->Clostridia;O->Clostridiales;F->Lachnospiraceae	5	434	579
P->Proteobacteria;C->Betaproteobacteria;O->Burkholderiales;F->Alcaligenaceae	158		
P->Proteobacteria;C->Gammaproteobacteria;O->Enterobacteriales;F->Enterobacteriaceae	1661	7610	2639
P->Proteobacteria;C->Gammaproteobacteria;O->Pasteurellales;F->Pasteurellaceae	7	441	6
P->Proteobacteria;C->Gammaproteobacteria;O->Pseudomonadales;F->Pseudomonadaceae	106		61

Figure 14. Composition of the ileum microbiome characterized in the control subject and in the Crohn patient BT and AT by NGS. (a) Phylum-level classification shows the reduction of *Bacteroidetes* and the significant prevalence of *Proteobacteria* in the patient-BT *vs* the control and the patient-AT. **(b)** The Heatmap table highlights in brownish red the most significant alterations in the bacterial composition of gut microbiome detected in the CD patient BT; the numerical figures indicate the number of bacterial families sequenced in each patient.

Moreover, our data support the reduced bacterial diversity of the CD microbiome. In fact, the Shannon Diversity Index Score was significantly lower ($p < 0.05$) in our CD patient before therapy than after therapy and also when compared with the control (Figure 15).

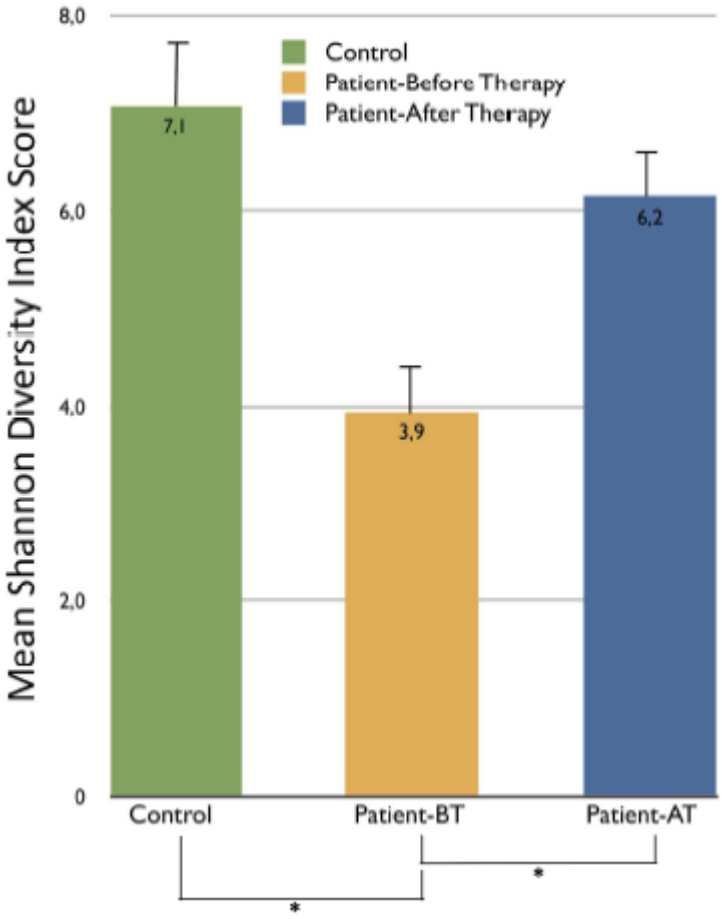


Figure 15. The mean Shannon Diversity Index Score. Values were significantly lower in the patient before therapy (patient-BT) than in the control subject and in the patient after therapy (patient-AT) (* $p < 0.05$).

4.2.2 Celiac disease

By 16S bacterial rRNA sequencing we preliminarily obtained 214,999 post quality filtered sequences. After the OTUs picking procedure, we identified a total of 155,281 sequences corresponding to 6,624 OTUs, of which 1% resulted unclassified and 99% were assigned to known bacteria after the taxonomic classification. Globally, including singleton and doubleton sequences, the non-filtered per frequency taxonomic assignment reported a total of 19 Phyla, 40 Classes, 81 Orders, 159 Families and 328

Genera of which only 22 species were identified. After filtering per frequency higher than 1% a total of 5 Phyla, 6 Classes, 7 Orders, 10 Families, and 10 different genera resulted (Table 15).

Table 15. 16S bacterial RNA samples metadata, globally considered in the study population.

	NGS-based duodenal microbiome profiles
N QF reads	214,999
N post OTU picking reads	155,281
N OTUs	6,624 (1% unclassified; 99% assigned)
Non-filtered-per-frequency taxonomic assignment	• 19 phyla • 40 classes • 81 orders • 159 families • 328 genera
Filtered-per-frequency >1% taxonomic assignment	• 5 phyla • 6 classes • 7 orders • 10 families • 10 genera

NGS: next generation sequencing; N: number; QF: quality filtered; OTU: operational taxonomic units.

The 5 identified phyla, globally considered, were: *Proteobacteria* (44.2%), *Actinobacteria* (15.8%), *Bacteroidetes* (15.8%), *Firmicutes* (14.9%), and *Fusobacteria* (9.3%) (Figure 16A). In particular, the majority of known bacteria sequences (>10% of mean frequency among the three groups) were classified within six genera: *Acinetobacter* (16.5%), *Neisseria*

(15.5%), *Streptococcus* (14.9%), Propionibacterium (13.6%) *Haemophilus* (12.3%), and *Prevotella* (*Prevotellaceae* family, 10.2%). The less abundant genera (2-10% of frequency) identified were: *Fusobacterium* (9.3%), *Prevotella* (*Paraprevotellaceae* family, 2.9%), *Porphyromonas* (2.6%), *Rothia* (2.2%) (Figure 16E).

The microbiome profiles, as analyzed in the 3 study groups, at the phylum level show to be quiet similar for *Actinobacteria* and *Bacteroidetes*, while in the active CD patients a trend in the increase of *Proteobacteria* and *Fusobacteria* and in the decrease of *Firmicutes* was observed, as compared with the other 2 groups (Figure 16A). Among the *Proteobacteria* phylum , bacteria of the *Betaproteobacteria* class (p= 0.01), the *Neisseriales* order (p= 0.02), the *Neisseriaceae* family (p= 0.03) and the *Neisseria* genus (p= 0.03) were significantly more abundant in the active CD patients (Figure 16B−E). In particular, the latter genus was more abundant in the active CD patients (32%) than in the GFD patients (4%) and controls (10%) (Figure 16E). The *Neisseria* genus was the most represented genus (99.8%) associated to the *Neisseriaceae* family. Within the *Proteobacteria* phylum, species of the *Gammaproteobacteria* class were significantly (p=0.03) less abundant in active-CD patients than in the other groups (Figure 16C).

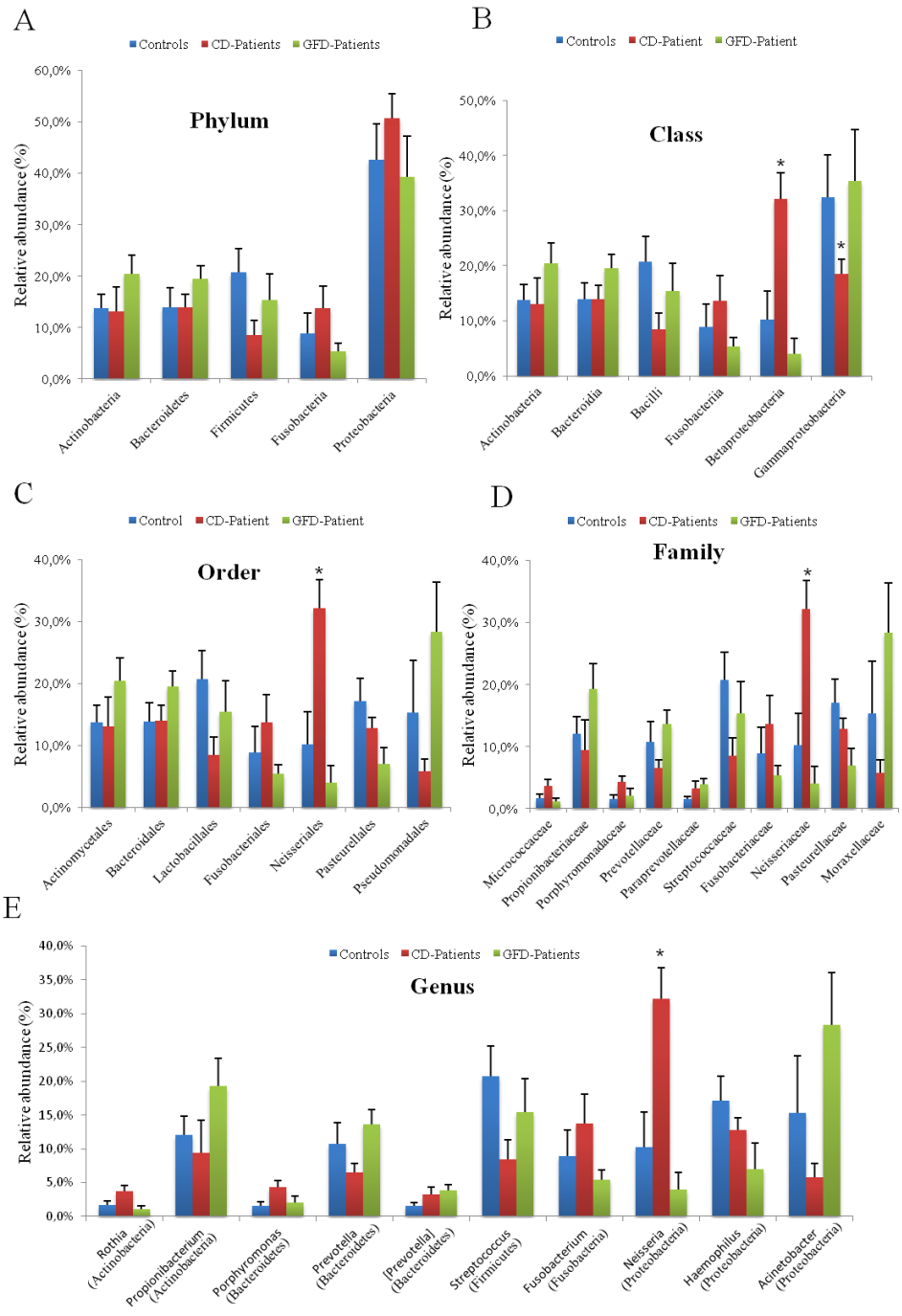


Figure 16. Duodenal microbiome taxonomic composition (from phylum to genus level) in controls, active and GFD CD patients. (A) Phylum level classification among the tested groups (controls, CD patients, GFD patients) reporting the relative abundance. *Proteobacteria* was the most represented phylum in all the groups with an average of 44.2%. No significant differences were found in each of the identified phyla among groups. (B) Class level classification among the three groups showed a trend in reduction in the *Betaproteobacteria* (p= 0.01) class in GFD patients (4%) and controls (10%) compared to CD patients (32%). The *Gammaproteobacteria* class was observed to be decreased (p= 0.03) in CD patients (18%), compared with controls (32%) and GFD patients (35%). (C) Order level classification reported a significant difference (p= 0.02) in the

Neisseriales order in the three groups, controls (~10%), GFD patients (~4%), CD patients (~32%). (D) Family level classification also showed a significant difference (p= 0.03) in the *Neisseriace* family, observed to be decreased in controls (~10%) and GFD patients (4%) groups, compared to CD patients (~32%). (E) The genus level comparison among groups highlighted a statistical significance difference (p= 0.03) in the genus *Neisseria* which abounded in the active CD patients (32%) respect to controls (10%) and GFD patients (4%). Taxa, in parenthesis, refer to phylum which genus belongs to. Error bars indicate standard error. Asterisks refer to taxa that reported a statistical significance difference among the three groups ($p \leq 0.05$, ANOVA).

Alpha diversity, or the within sample diversity, was computed using two different metrics: the Faith's Phylogenetic Diversity (PD) richness estimator [156], and the *observed species* metric, which reports the number of different bacterial OTUs at a rarefaction depth of 854 sequences/sample. The obtained alpha rarefaction curves showed that bacterial community richness did not differ between CD-patients and controls (Figures 17A and B). Beta diversity of bacterial communities is presented in the unweighted and weighted UniFrac PCoA plots (Figures 17C and D). The beta diversity computed with unweighted Unifrac method, was statistically significant within the three groups (p= 0.015, R^2= 0.08; ADONIS). Also, the beta diversity computed with weighted Unifrac method presented a significant difference among the three groups of study (p= 0.029, R= 0.14; ANOSIM). Interestingly, this analysis suggests that active and GFD CD gut communities are more similar to each other than control communities are to each other, indicated by the tighter clustering of active and GFD CD points than control points in Figure 17, with the only exception of a 14 years old active-CD patient (indicated by a black arrow). The significant differences found suggest that there are distinct community types associated with active CD and controls. This may mean that active CD is associated with a

particular overall gut microbial community signature. This signature varies across individuals, but degree of variability is less than that across healthy individuals.

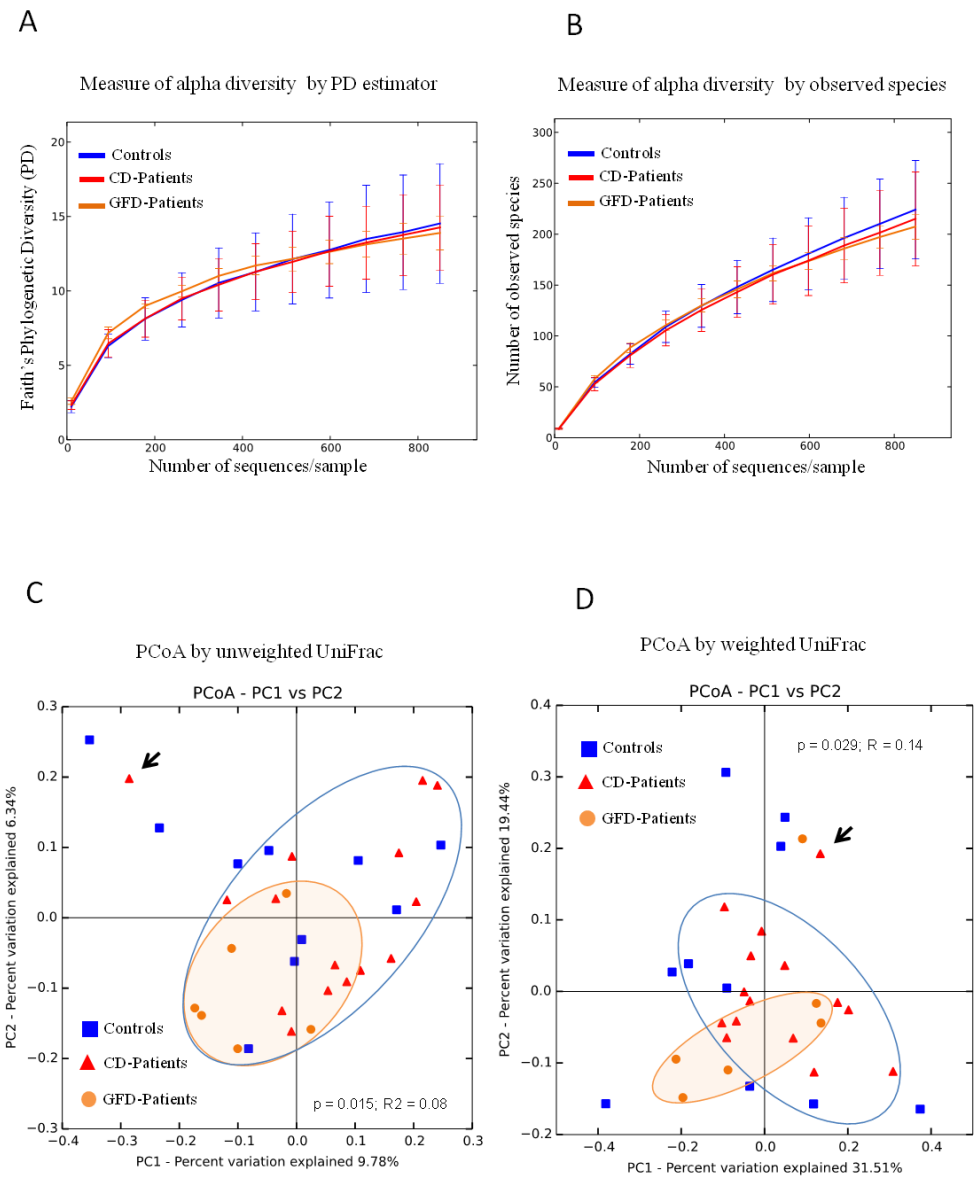


Figure 17. Bacterial diversity analysis. Alpha diversity, measured using Faith's Phylogenetic Diversity estimator (A) or the "observed species" method (B), shows no difference between the 3 groups. Error bars represent standard error of the mean. (C)-(D) Beta diversity computed with the unweighted (C) and weighted (D) UniFrac distances shows a phylogenetic relationship between active and GFD patients, both of which are distinct from the controls which reported a random distribution [(p= 0.015, R2= 0.08) ADONIS)] (C), [(p= 0.029, R= 0.14) ANOSIM] (D). Black arrows indicate the only 14 years old female in the active CD patients.

V. DISCUSSION

5.1 Targeted DNA Sequence Capture

A DNA sequence capture approach followed by NGS was successfully applied to the simultaneous analysis of a large panel of genes possibly related to cardiomyopathies. In particular, a panel of 202 candidate genes, encoding proteins involved in heart functions or that might be related to the development of cardiomyopathies, was used to design a custom array for targeted enrichment before NGS analysis, thus increasing the effectiveness of HCM molecular diagnosis over currently available procedures.

Limiting the analysis to the 8 sarcomeric genes previously analyzed by a combined DHPLC/Sanger approach [115], all the Sanger-confirmed variants previously identified in patients 1 and 2 were confirmed by the NGS procedure described herein except one. With the described procedure, we also detected 21 variants that were missed by the previous DHPLC screening. To evaluate the reliability of the new NGS-based method in detecting variants, we compared its sensitivity and specificity to those of the previously used DHPLC/Sanger method. As shown in Table 11, NGS was very sensitive and specific in detecting mutations, while the sensitivity of the DHPLC/Sanger procedure was much lower (about 50-60%). These low sensitivity values depend on the experimental design used for the DHPLC/Sanger approach, which was designed to detect variants within the exon sequences, and is expected to be less sensitive, for example, in highlighting changes in the untranslated regions or close to the ends of the analyzed segments – a limitation that does not apply to the NGS approach. In fact, when variants expected to be undetectable by DHPLC/Sanger

sequencing, i.e. variants outside amplicons and homozygous variants, are not taken into account in calculating sensitivity, the results become comparable to those of other reports [145], although still remaining worse than the new NGS-based procedure.

The above considerations only refer to a small proportion (1%) of the entire target region analyzed by the NGS-based method. Unlike the DHPLC method, the NGS procedure produces results on a set of 202 genes, corresponding to about 4M bases of the coding genomic sequence. It is noteworthy that even on the limited 8-gene set the performance of the NGS-based method is better. As stated elsewhere [157], it remains in fact to be established whether the diagnostic output of a test that detects a large number of variants in a relatively large number of genes might be compared with a test that detects almost all variants, but in a low number of genes.

In diagnostic terms, the high throughput sequencing procedure enabled us: (i) to identify a known HCM-causative mutation in patient 1, who went previously undiagnosed; (ii) to better delineate the molecular alteration underlying the phenotype of patient 2 by detecting the presence in this HCM subject of a mutation causative of long QT syndrome [150,158]; and (iii) to postulate that the HCM observed in patient 3 was due to *CACNA1C*-related mutations (Timothy syndrome) [152-154].

Although the bioinformatic tools we used eliminate variants that are most likely bereft of clinical significance, the real pathogenic role of a given variant can only be determined with functional studies and/or family segregation analysis to evaluate genotype-phenotype correlations within each family. The variants we identified in the present work (see Table 12) should be classified as “possibly pathogenic” because they fulfil at least one of the following conditions: (i) previous findings showing that they are

disease-causing (4 of them); (ii) their presence on genes known to cause HCM (2); and (iii) the presence of stop codons indicating a truncated protein (1) (Table 12). Further procedures are of course necessary to confirm the causative origin of mutations in order to better rationalize the genotype-phenotype relationship within each family and to better understand the molecular basis of the alteration induced by each mutational event. The annotations and the prioritization strategy described herein could help to simplify these steps.

Sequencing of several patients in a single run is an additional improvement over the conventional Sanger procedure in terms of time and sequencing throughput. We applied our procedure to a pool containing DNA from the three patients to evaluate the feasibility of sequencing several patients simultaneously. Target enrichment and sequencing coverage were highly consistent with the values found in the three separate samples. However, not all variants detected in the individual DNA samples were identified in the pool (Tables 13 and 14), as expected considering the reduced overall sequence depth. The sensitivity of the pool-based approach may easily be enhanced, by increasing the sequencing depth while still saving on the costs and workload associated with library preparation and titration. The use of tags to selectively label the DNA of single patients in multiplexing experiments would have the advantage of univocally assigning the sequencing reads to each patient at the cost of a slightly more complex experiment.

5.2 Metagenomics

5.2.1 Crohn disease

Inflammatory bowel diseases are chronic recurrent diseases of the gastrointestinal tract, and are caused by a combination of genetic and environmental factors, including the gut microbiota. A number of studies have reported significant alterations of gut microbial composition in IBD patients compared with not affected individuals [104,159-161]. Not only IBD patients have an altered rate of *Bacteroidetes*, *Proteobacteria* and *Firmicutes* colonization, but the bacterial diversity of their microbiome is generally lower than that of controls. In this context, we characterized the 16S rRNA ileum-associated microbiome of a Crohn-affected patient at diagnosis and after nutritional therapy. This therapy is known to be effective in the reduction of gut inflammation; we found that it was effective also in the restoring of the gut microbial balance. In fact, at follow-up, the microbial composition did not differ between the patient and control.

Although our findings were obtained in one case-control study, and therefore may be considered preliminary, they strongly suggest that nutritional therapy can improve the inflammatory status of Crohn disease by restoring the composition of the mucosal microbiome. This case of Crohn disease gut microbiome dysbiosis restored by nutritional therapy can be considered proof-of-concept to test a similar approach in several laboratories and also as clinical preliminary assessment for future research and possibly clinical trials.

5.2.2 Celiac disease

Alterations in the gut microbiota and such other factors as haplotypes, breastfeeding, type of delivery, clinical manifestations, time of gluten exposure, and age of onset, have been implicated in CD, but the results reported so far are inconclusive particularly due to differences in the populations studied, and in the samples and techniques used [87,162-165]. In the attempt to shed light on the link between the gut mucosal microbial community and CD onset, we exploited the NGS technique to examine the gut microbiome composition in an Italian adult cohort of CD patients, either active or on a GFD, and in control subjects. We found significant differences between the three groups studied, which indicates that the microbiota distinguishes these groups. Notably, diversity analysis revealed a tighter clustering of active and GFD CD, suggesting that active and GFD CD microbiomes are more similar to each other than the control microbiome is to each of them. Interestingly, the *Neisseria* genus, which belongs to the *Proteobacteria* phylum, was more abundant in active CD patients than in the other two groups. This confirm the previous finding that species of the *Neisseria* genus were more abundant in the duodenum of adult active CD patients than in controls [164,166].

Even if the significance of the observed differences has to be assessed by further evaluations to define their potential pathogenetic role in celiac disease development, we were able to define, using a NGS-based method, the whole duodenal microbiome signature of adult patients affected by active and GFD CD.

V. CONCLUSIONS

The main aim of this PhD project was to use NGS-based strategies to study the molecular basis of human diseases. Two different approaches, target DNA sequence capture and metagenomics, were used to assess the above mentioned issue and were successfully applied to the study of inherited cardiomyopathies and Crohn and celiac diseases, respectively. The data shown here strongly reinforce the concept that NGS techniques open a new era in the search for new disease-causing genes and/or novel modifier genes and in the study of disease pathogenesis.

Taken together, all the above results indicate that the NGS-based procedures described here can be easily applied to increase our understanding of the molecular basis of human diseases and that they can be useful also for routine diagnostic purposes.

VI. REFERENCES

1. Sanger F, Nicklen S, Coulson AR 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463–7.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304–51.
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45.
4. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24:133-41.
5. von Bubnoff A. Next-generation sequencing: the race is on. Cell 2008;132:721-3.
6. Hayden EC. Technology: The $1,000 genome. Nature 2014;507:294-5.
7. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform 2012;3:40.
8. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11:31-46.
9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev Genet 2009:10;57–63.
10. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.
11. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53-9.
12. Valouev A, Ichikawa J, Tonthat T, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 2008;18:1051-63.
13. Harris TD, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. Science 2008;320:106-9.
14. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science 2009;323:133-8.
15. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 2008;26:602-11.

16. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods 2015. doi: 10.1038/nmeth.3290. [Epub ahead of print].
17. Thomas RK, Nickerson E, Simons JF, et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nat Medicine 2006;12:852-5.
18. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903-5.
19. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 2009;3:315-6.
20. Ball MP, Li JB, Gao Y, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 2009;27:361-8.
21. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207-14.
22. ten Bosch JR, Grody W. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn 2008;10:484-92.
23. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics 2011;38:95-109.
24. Smith AM, Heisler LE, St Onge RP, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res 2010;38:e142.
25. Yeager M, Xiao N, Hayes RB, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 2008;124:161-70.
26. Hernan I, Borràs E, de Sousa Dias M, et al. Detection of genomic variations in BRCA1 and BRCA2 genes by long-range PCR and next-generation sequencing. J Mol Diagn 2012;14:286-93.
27. Morinière V, Dahan K, Hilbert P, et al. Improving mutation screening in familial hematuric nephropathies through next generation sequencing. J Am Soc Nephrol 2014; pii:ASN.2013080912.

28. Kohlmann A, Grossmann V, Nadarajah N, Haferlach T. Next-generation sequencing-feasibility and practicality in haematology. Br J Haematol 2013;160:736-53.
29. Trujillano D, Ramos MD, González J, et al. Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. J Med Genet 2013;50:455-462.
30. .Vaca-Paniagua F, Alvarez-Gomez RM, Fragoso-Ontiveros V, et al. Full-exon pyrosequencing screening of BRCA germline mutations in Mexican women with inherited breast and ovarian cancer. PLoS One 2012;7:e37432.
31. Pern F, Bogdanova N, Schürmann P, et al. Mutation analysis of BRCA1, BRCA2, PALB2 and BRD7 in a hospital-based series of German patients with triple-negative breast cancer. PLoS One 2012;7:e47993.
32. Watson CM, Crinnion LA, Morgan JE, et al. Robust diagnostic genetic testing using solution capture enrichment and a novel variant-filtering interface. Hum Mutat 2014;35:434-41.
33. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007;39:1522-7.
34. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum Cet al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:182-9.
35. Liu ZJ, Li HF, Tan GH, et al. Identify mutation in amyotrophic lateral sclerosis cases using HaloPlex target enrichment system. Neurobiol Aging 2014. pii:S0197-4580(14)00472-2.
36. Schaefer E, Helms P, Marcellin L, et al. Next-generation sequencing (NGS) as a fast molecular diagnosis tool for left ventricular noncompaction in an infant with compound mutations in the MYBPC3 gene. Eur J Med Genet 2014;57:129-32.
37. 1000 Genomes Project Consortium, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.

38. Fu W, O'Connor TD, Jun G, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 2013;493:216-20.
39. Woo HM, Park HJ, Park MH, et al. Identification of CDH23 mutations in Korean families with hearing loss by whole-exome sequencing. BMC Med Genet 2014;15:46.
40. Hillman S, Smart M, Bacchelli C, Ocaka L, Williams DJ. 2.4 Whole exome sequencing of growth restricted offspring identifies gene variants implicated in maturity onset diabetes of the young. Arch Dis Child Fetal Neonatal 2014;99:A2.
41. Prada CE, Gonzaga-Jauregui C, Tannenbaum R, et al. Clinical utility of whole-exome sequencing in rare diseases: Galactosialidosis. Eur J Med Genet 2014;57:339-44.
42. Iglesias A, Anyane-Yeboa K, Wynn J, et al. The usefulness of whole-exome sequencing in routine clinical practice. Genet Med 2014; doi:10.1038/gim.2014.58.
43. Shearer AE, Deluca AP, Hildebrand MS, et al. Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proc Natl Acad Sci USA 2010;107:21104-9.
44. Brownstein Z, Friedman LM, Shahin H, et al. Targeted genomic capture and massively parallel sequencing to identify genes for hereditary hearing loss in middle eastern families. Genome Biology 2011;12:R89.
45. Németh AH, Kwasniewska AC, Lise S, et al. Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. Brain 2013;136:3106-18.
46. Shanks ME, Downes SM, Copley RR, et al. Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease. Eur J Hum Genet 2013;21:274-80.
47. Nijman IJ, van Montfrans JM, Hoogstraat M, et al. Targeted next-generation sequencing: a novel diagnostic tool for primary immunodeficiencies. J Allergy Clin Immunol 2014;133:529-34.
48. Hariani GD, Lam EJ, Havener T, Kwok et al. Application of next generation sequencing to CEPH cell lines to discover variants

associated with FDA approved chemotherapeutics. BMC Res Notes 2014;7:360.

49. Johnson DB, Dahlman KH, Knol J, et al. Enabling a genetically informed approach to cancer medicine: a retrospective evaluation of the impact of comprehensive tumor profiling using a targeted next-generation sequencing panel. Oncologist 2014;19:616-22.
50. Ross JS, Wang K, Rand JV, et al. Comprehensive genomic profiling of relapsed and metastatic adenoid cystic carcinomas by next-generation sequencing reveals potential new routes to targeted therapies. Am J Surg Pathol 2014;38:235-8.
51. Chong HK, Wang T, Lu HM, et al. The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. PLoS One 2014;9:e97408.
52. Teekakirikul P, Kelly MA, Rehm HL, Lakdawala NK, Funke BH. Inherited cardiomyopathies: molecular genetics and clinical genetic testing in the postgenomic era. J Mol Diagn 2013;15:158-70.
53. .Mazzanti A, O'Rourke S, Ng K, et al. The usual suspects in sudden cardiac death of the young: a focus on inherited arrhythmogenic diseases. Expert Rev Cardiovasc Ther 2014;12:499-519.
54. van de Meerakker JB, Christiaans I, Barnett P, et al. A novel alpha-tropomyosin mutation associates with dilated and non-compaction cardiomyopathy and diminishes actin binding. Biochim Biophys Acta 2013;1833:833-9.
55. Schaefer E, Helms P, Marcellin L, et al .Next-generation sequencing (NGS) as a fast molecular diagnosis tool for left ventricular noncompaction in an infant with compound mutations in the MYBPC3 gene. Eur J Med Genet 2014;57:129-32.
56. Meder B, Haas J, Keller A, et al. Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies. Circ Cardiovasc Genet. 2011;4:110-22.
57. Mook OR, Haagmans MA, Soucy JF, et al. Targeted sequence capture and GS-FLX Titanium sequencing of 23 hypertrophic and dilated cardiomyopathy genes: implementation into diagnostics. J Med Genet 2013;50:614-26.

58. Lopes LR, Zekavati A, Syrris P, et al. Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. J Med Genet. 2013;50:228-39.
59. Li X, Buckton AJ, Wilkinson SL, et al. Towards clinical molecular diagnosis of inherited cardiac conditions: a comparison of bench-top genome DNA sequencers. PLoS One 2013;8:e67744.
60. Sikkema-Raddatz B, Johansson LF, de Boer EN, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 2013;34:1035-42.
61. van Spaendonck-Zwarts KY, Posafalvi A, van den Berg MP, et al.Titin gene mutations are common in families with both peripartum cardiomyopathy and dilated cardiomyopathy. Eur Heart J. 2014;35:2165-73.
62. Haas J, Frese KS, Peil B, et al. Atlas of the clinical genetics of human dilated cardiomyopathy. Eur Heart J 2014. pii: ehu301.
63. Loporcaro CG, Tester DJ, Maleszewski JJ, Kruisselbrink T, Ackerman MJ. Confirmation of cause and manner of death via a comprehensive cardiac autopsy including whole exome next-generation sequencing. Arch Pathol Lab Med 2014;138:1083-9.
64. D'Argenio V, Salvatore F. The role of the gut microbiome in the healthy adult status. Clin Chim Acta 2015. doi: 10.1016/j.cca.2015.01.003.
65. Dewhirst FE, Chen T, Izard J, et al. The human oral microbiome. J Bacteriol 2010;192:5002-17.
66. Grice EA, Kong HH, Conlan S, et al. Topographical and Temporal Diversity of the Human Skin Microbiome. Science 2006;324:1190-2.
67. González A, Vázquez-Baeza Y, Knight R. SnapShot: The Human Microbiome. Cell 2014;158:690-690.e1.
68. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174-80.
69. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. Microbial degradation of complex carbohydrates in the gut.Gut microbes 2012;3:289-306.
70. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59-65.

71. Walsh CJ, Guinane CM, O'Toole PW, Cotter PD. Beneficial modulation of the gut microbiota. FEBS Lett 2014; doi: 10.1016/j.febslet.2014.03.035.
72. Gajer P, Brotman RM, Bai G, et al. Temporal dynamics of the human vaginal microbiota. Sci Transl Med 2012;4:132ra152.
73. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505:559-63.
74. Wu GD, Chen J, Hoffmann C, et al. Linking longterm dietary patterns with gut microbial enterotypes. Science 2011;334:105-8.
75. Koren O, Goodrich JK, Cullender TC, et al. Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 2012;150:470-80.
76. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. Nature 2012;486:222-7.
77. Perez-Cobas AE, Gosalbes MJ, Friedrichs A, et al. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. Gut 2013;62:1591-601.
78. Walker AW, Duncan SH, Louis P, Flint HJ. Phylogeny, culturing, and metagenomics of the human gut microbiota. Trends Microbiol 2014;22:267-74.
79. Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;486:215-21.
80. Sekirov I, Russell SL, Antunes LCM, Finlay BB. Gut microbiota in health and disease. Physiol Rev 2010;90:859-904.
81. Cryan JF, O'Mahony SM. The microbiome-gut-brain axis: from bowel to behavior. Neurogastroenterol. Motil. 2011;23:187-92.
82. Woese CR, Fox GE, Zablen L, et al. Conservation of primary structure in 16S ribosomal RNA. Nature 1975;254:83-6.
83. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA based community profiling for human microbiome research. PLoS One 2012;7:e39315.
84. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology 2014;146:1489-99.

85. Schnabl B, Brenner DA. Interactions between the intestinal microbiome and liver diseases. Gastroenterology 2014;146:1513-24.
86. Severance EG, Yolken RH, Eaton WW. Autoimmune diseases, gastrointestinal disorders and the microbiome in schizophrenia: more than a gut feeling. Schizophr Res. 2014;doi: 10.1016/j.schres.2014.06.027.
87. Olivares M, Neef A, Castillejo G, et al. The HLA-DQ2 genotype selects for early intestinal microbiota composition in infants at high risk of developing coeliac disease. Gut. 2014;doi: 10.1136/gutjnl-2014-306931.
88. Le Chatelier E, Nielsen T, Qin J, et al. Richness of human gut microbiome correlates with metabolic markers. Nature 2013;500:541-6.
89. Wang Y, Hoenig JD, Malin KJ, et al. 16S rRNA gene based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. ISME J 2009;3:944-54.
90. Bisgaard H, Li N, Bonnelykke K, et al. Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. J Allergy Clin Immunol 2011;128:646-52.
91. Cani, P.D. Gut microbiota and obesity: lessons from the microbiome. Brief. Funct. Genomics 2013;12:381-7.
92. Clarke SF, Murphy EF, Nilaweera K, et al. The gut microbiota and its relationship to diet and obesity: new insights. Gut Microbes 2012;3:186-202.
93. Cox AJ, West NP, Cripps AW. Obesity, inflammation, and the gut microbiota. Lancet Diabetes Endocrinol 2014;doi: 10.1016/S2213-8587(14)70134-2.
94. Ponder A, Long MD. A clinical review of recent findings in the epidemiology of inflammatory bowel disease. Clin Epidemiol 2013;5:237-47.
95. Pizzi LT, Weston CM, Goldfarb NI, et al. Impact of chronic conditions on quality of life in patients with inflammatory bowel disease. Inflamm Bowel Dis 2006;12:47-52.

96. Dyson JK, Rutter MD. Colorectal cancer in inflammatory bowel disease: what is the real magnitude of the risk? World J Gastroenterol 2012;18:3839-48.
97. Neuman MG, Nanau RM. Inflammatory bowel disease: role of diet, microbiota, life style. Transl Res 2012;160:29-44.
98. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955-62.
99. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007;448:427-34.
100. Cho JH. The genetics and immunopathogenesis of inflammatory bowel disease. Nat Rev Immunol 2008;8:458-66.
101. Hooper LV, Midtvedt T, Gordon JI. How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr 2002;22:283-307.
102. Björkstén B. The gut microbiota: a complex ecosystem. Clin Exp Allergy 2006;36:1215-7.
103. Willing BP, Dicksved J, Halfvarson J, et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology 2010;139:1844-54.
104. Walker AW, Sanderson JD, Churcher C et al. High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. BMC Microbiol 2011;11:7.
105. Lionetti P, Callegari M, Cavicchi M et al. Enteral nutrition-induced remission is associated with profound modification of the intestinal microflora in Crohn’s disease. J Pediatr Gastroenterol Nutr 2004;39:S106.
106. Green PH, Cellier C. Celiac disease. N Engl J Med 2007;357:1731-43.
107. Schuppan D, Junker Y, Barisani D. Celiac disease: from pathogenesis to novel therapies. Gastroenterology 2009;137:1912-33.
108. Sollid LM. Coeliac disease: dissecting a complex inflammatory disorder. Nat Rev Immunol 2002;2:647-55.

109. Stene LC, Honeyman MC, Hoffenberg EJ, et al. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am J Gastroenterol 2006;101:2333-40.
110. Akobeng AK, Ramanan AV, Buchan I, et al. Effect of breast feeding on risk of coeliac disease: a systematic review and meta-analysis of observational studies. Arch Dis Child 2006;91:39-43.
111. Vivas S, Ruiz de Morales JM, Fernandez M, et al. Age-related clinical, serological, and histopathological features of celiac disease. Am J Gastroenterol 2008;103:2360-5.
112. Lohi S, Mustalahti K, Kaukinen K, et al. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther 2007;26:1217-25.
113. Sanz Y, Sánchez E, Marzotto M, et al. Differences in faecal bacterial communities in coeliac and healthy children as detected by PCR and denaturing gradient gel electrophoresis. FEMS Immunol Med Microbiol 2007;51:562-8.
114. Collado MC, Calabuig M, Sanz Y. Differences between the fecal microbiota of celiac infants and healthy controls. Curr Issues Intest Microbiol 2007;8:9-14.
115. Frisso G, Limongelli G, Pacileo G, et al. A child cohort study from southern Italy enlarges the genetic spectrum of hypertrophic cardiomyopathy. Clin Genet 2009;76:91-101.
116. Lin AE, Alexander ME, Colan SD, et al. Clinical, pathological, and molecular analyses of cardiovascular abnormalities in Costello syndrome: a Ras/MAPK pathway syndrome. Am J Med Genet 2011;155A:486-507.
117. Christensen G, Chen J, Ross J, Chien KR. Mouse models of human cardiovascular disease. In: Chien KR, ed. Molecular basis of cardiovascular disease. Philadelphia PA USA, Elsevier, 2004;pp.72-106.
118. Barrans JD, Allen PD, Stamatiou D, Dzau VJ, Liew CC. Global gene expression profiling of end-stage dilated cardiomyopathy using a human cardiovascular-based cDNA microarray. Am J Pathol 2002;160:2035-43.
119. Frey N, Olson EN. Cardiac hypertrophy: the good, the bad, and the ugly. Annu Rev Physiol 2003;65:45-79.

120. Carreno JE, Apablaza F, Ocaranza MP, Jalil JE. Cardiac Hypertrophy: Molecular and Cellular Events. Rev Esp Cardiol 2006;59:473-86.
121. Detta N, Frisso G, Zullo A, et al. Novel deletion mutation in the cardiac sodium channel inactivation gate causes long QT syndrome. Int J Cardiol 2013;165:362-5.
122. Sarubbi B, Frisso G, Romeo E, et al. Efficacy of pharmacological treatment and genetic characterization in early diagnosed patients affected by long QT syndrome with impaired AV conduction. Int J Cardiol 2011;149:109-13.
123. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res 2001;11:1725-9.
124. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 2003;55:541-55.
125. Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, Morris RW. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 2010;26:2803-10.
126. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 2011;12:245.
127. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. BMC Bioinformatics 2010;26:2069-70.
128. The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73.
129. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335-6.
130. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460-1.
131. McDonald D, Price MN, Goodrich J, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012;6:610–8.
132. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microb 2007;73:5261-7.

133. Caporaso JG, Bittinger K, Bushman FD, et al. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 2010;26:266-7.
134. De Santis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microb 2006;72:5069-72.
135. Price MN, Dehal PS, Arkin AP. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. PlosOne 2010;5:e9490.
136. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71:8228-35.
137. Clarke KR. Nonparametric multivariate analyses of changes in community structure. Aust J Ecol 1993;18:117–43.
138. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol 2001;26:32–46.
139. Hedges DJ, Burges D, Powell E, et al. Exome sequencing of a multigenerational human pedigree. PLoS One 2009;4:e8232.
140. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA. Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab 2010;100:184-92.
141. Wei Z, Wang W, Hu P, Lyon GJ and Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Research 2011;39:e132.
142. Amstutz U, Andrey-Zürcher G, Suciu D, Jaggi R, Häberle J, Largiadèr CR. Sequence capture and next-generation resequencing of multiple tagged nucleic acid samples for mutation screening of urea cycle disorders. Clin Chem 2011;57:102-11.
143. Ghosh S, Krux F, Binder V, Gombert M, Niehues T, Feyen O, Laws HJ, Borkhardt A. PID-NET: German Network on Primary Immunodeficiency Diseases. Array-based sequence capture and next-generation sequencing for the identification of primary immunodeficiencies. Scand J Immunol 2012;75:350-4.
144. Neveling K, Collin RW, Gilissen C, et al. Next-generation genetic testing for retinitis pigmentosa. Hum Mutat 2012;33:963-72.

145. Chou LS, Lyon E, Wittwer CT. A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductance regulator gene as a model. Am J Clin Pathol 2005;124:330-8.
146. Michels M, Soliman OI, Phefferkorn J, et al. Disease penetrance and risk stratification for sudden cardiac death in asymptomatic hypertrophic cardiomyopathy mutation carriers. Eur Heart J 2009;30:2593-8.
147. Tait KF, Collins JE, Heward JM, et al. Evidence for a Type 1 diabetes-specific mechanism for the insulin gene-associated IDDM2 locus rather than a general influence on autoimmunity. Diabet Med 2004;21:267-70.
148. Prakash T, Sharma VK, Adati N, et al. Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. PLoS One 2010;5:e13284.
149. Fendler W, Klich I, Cieślik-Heinrich A, Wyka K, Szadkowska A, Młynarski W. Increased risk of type 1 diabetes in Polish children - association with INS-IGF2 5'VNTR and lack of association with HLA haplotype. Endokrynol Pol 2011;62:436-42.
150. Lupoglazoff JM, Denjoy I, Villain E, et al. Long QT syndrome in neonates: conduction disorders associated with HERG mutations and sinus bradycardia with KCNQ1 mutations. J Am Coll Cardiol 2004;43:826-30.
151. Antzelevitch C, Pollevick GD, Cordeiro JM, et al. Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death. Circulation 2007;115:442-9.
152. Bidaud I, Lory P. Hallmarks of the channelopathies associated with L-type calcium channels: A focus on the Timothy mutations in Ca(v)1.2 channels. Biochimie 2011;93:2080-6.
153. Splawski I, Timothy KW, Decher N, et al. Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. Proc Natl Acad Sci U S A 2005;102:8089-96.

154. Splawski I, Timothy KW, Sharpe LM, et al. Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell 2004,;119:19-31.
155. Roussos P, Giakoumaki SG, Georgakopoulos A, Robakis NK, Bitsios P. The CACNA1C and ANK3 risk alleles impact on affective personality traits and startle reactivity but not on cognition or gating in healthy males. Bipolar Disord 2011;13:250-9.
156. Faith DP, Baker AM. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evolutionary Bioinformatics 2007;2:121-8.
157. Ware JS, John S, Roberts AM, et al. Next generation diagnostics in inherited arrhythmia syndromes: a comparison of two approaches. J Cardiovasc Trans Res 2013;6:94-103.
158. Chiang CE, Roden DM. The long QT syndromes: genetic basis and clinical implications. J Am Coll Cardiol 2000;36:1-12.
159. Frank DN, Amand ALS, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A 2007;104:13780-5.
160. Sokol H, Seksik P. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. Curr Opin Gastroenterol 2010;26:327-31.
161. Elson CO, Cong Y. Host-microbiota interactions in inflammatory bowel disease. Gut Microbes 2012;3:332-44.
162. Nistal E, Caminero A, Vivas S, et al. Differences in faecal bacteria populations and faecal bacteria metabolism in healthy adults and celiac disease patients. Biochimie 2012;94:1724-9.
163. Sellitto M, Bai G, Serena G, et al. Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. PLoSOne 2012;7:e33387.
164. Nistal E, Caminero A, Herrán AR, et al. Differences of small intestinal bacteria populations in adults and children with/without celiac disease: effect of age, gluten diet, and disease. Inflamm Bowel Dis. 2012;18:649-56.

165. Kennedy NA, Walker AW, Berry SH, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoSOne 2014;9:e88982.
166. Wacklin P, Kaukinen K, Tuovinen E, et al. The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease. Inflamm Bowel Dis 2013;19:934-41.

APPENDIX 1

List of the genes selected to perform the custom array.

Gene Description (Functional classes)	Gene	trIDs*	Chr†	Segs (n)‡	Segs (bp)§	Transc (n)‖
CARDIOMYOPATHY -ASSOCIATED						
actin, alpha 1, skeletal muscle	ACTA1	NM_001100	1	1	3852	1
cardiac muscle alpha actin 1 proprotein	ACTC1	NM_005159	15	2	8523	1
calreticulins Ca-binding chaperons	CALR3	NM_145046	19	7	8574	1
calsequestrin 2 (cardiac muscle)	CASQ2	NM_001232	1	11	13711	1
caveolin 3≠	CAV3	NM_001234,NM_033337	3	2	3423	2
cytochrome c oxidase assembly protein≠	COX15	NM_078470,NM_004376	10	8	15983	2
cysteine and glycine-rich protein 3≠	CSRP3	NM_003476,NM_001127656	11	6	7347	2
Desmin	DES	NM_001927	2	4	7255	1
frataxin≠	FXN	NM_000144,NM_181425	9	5	7275	2
junctophilin≠	JPH2	NM_020433,NM_175913	20	6	12735	2
lysosomal-associated membrane protein≠	LAMP2	NM_002294,NM_001122606,NM_013995	X	8	19044	3
LIM domain binding 3	LDB3	NM_001080115,NM_001080114,NM_001080116,NM_007078	10	13	21033	4
myosin binding protein C, cardiac	MYBPC3	NM_000256	11	7	19751	1
myosin, heavy chain 6, cardiac muscle, alpha	MYH6	NM_002471	14	8	24026	1
myosin, heavy chain 7, cardiac muscle, beta	MYH7	NM_000257	14	4	23206	1
slow cardiac myosin regulatory light chain 2	MYL2	NM_000432	12	5	6796	1
myosin, light chain 3, alkali; ventricular, skeletal, slow	MYL3	NM_000258	3	4	5307	1
myosin light chain kinase 2	MYLK2	NM_033118	20	5	10877	1
myosin VI	MYO6	NM_004999	6	28	40652	1
myozenin 2	MYOZ2	NM_016599	4	5	8124	1
NADH dehydrogenase (ubiquinone) flavoprotein 2	NDUFV2	NM_021074	18	7	7918	1
obscurin, cytoskeletal calmodulin≠	OBSCN	NM_052843,NM_001098623	1	23	82957	2
phospholamban	PLN	NM_002667	6	2	3716	1
AMP-activated protein kinase gamma2 subunit≠	PRKAG2	NM_024429,NM_001040633,NM_01620	7	16	20335	3

solute carrier family 25	SLC25A4	NM_001151	4	2	4750	1
sorcin≠	SRI	NM_198901,NM_003130	7	6	10155	2
titin-cap (telethonin)	TCAP	NM_003673	17	1	2210	1
troponin C type 1	TNNC1	NM_003280	3	2	3474	1
troponin I, cardiac	TNNI3	NM_000363	19	3	5486	1
troponin T type 2, cardiac≠	TNNT2	NM_001001430,NM_000364,NM_001001431,NM_001001432	1	8	14014	4
tropomyosin 1 alpha chain≠	TPM1	NM_001018005,NM_001018004,NM_001018008,NM_000366,NM_001018020,NM_001018007,NM_001018006	15	8	14192	7
titina≠	TTN	NM_003319,NM_133378,NM_133432,NM_133437,NM_133379	2	44	229584	5
vinculin isoform meta-VCL≠	VCL	NM_014000,NM_003373	10	18	25559	2
integrin, alpha 8	ITGA8	NM_003638	10	26	31822	1
cardiac ankyrin repeat protein	CARP	NM_014391	10	4	8118	1
v-raf-1 murine leukemia viral oncogene homolog 1	RAF1	NM_002880	3	10	15534	1
METABOLIC HCM						
acid-alpha glucosidase	GAA	NM_000152	17	5	14287	1
amilo-1-6-glucosidase	AGL	NM_000642	1	19	31359	1
acid-beta glucosidase	GBA	NM_001005741	1	2	8923	1
ION CHANNELS						
potassium voltage-gated channel, KQT-like≠	KCNQ1	NM_181798,NM_000218	11	12	19308	2
potassium channel voltage-gated, subfamily H≠	KCNH2	NM_172056,NM_000238,NM_172057	7	7	15667	3
potassium channel voltage-gated ISK-related subfamily member 1≠	KCNE1	NM_000219,NM_001127670,NM_001127669,NM_001127668	21	4	8545	4
potassium channel voltage-gated ISK-related subfamily member 2	KCNE2	NM_172201	21	2	2805	1
sodium channel, voltage-gated, type V, alpha subunit≠	SCN5A	NM_198056,NM_001099405,NM_001099404,NM_000335	3	23	35206	4
calcium channel, voltage-dependent, L-type, alpha-1C subunit ≠	CACNA1C	NM_001129830,NM_001129827,NM_001129829,NM_000719,NM_001129831,NM_001129839,NM_001129836,NM_001129833,NM_001129832,NM_001129834,NM_001129835,NM_001129837,NM_0011	12	37	58658	20

		29838,NM_0011298 40,NM_001129841,N M_001129842,NM_0 01129843,NM_0011 29844,NM_0011298 46,NM_199460				
sodium channel, voltage-gated, type IV, beta subunit	SCN4B	NM_174934	11	4	9483	1
potassium channel, inwardly rectifying, subfamily J, member2	KCNJ2	NM_000891	17	2	7388	1
MEMBRANE CHANNELS						
solute carrier family 9, isoform A3	SLC9A3	NM_004174	5	8	14281	1
solute carrier family 9, isoform A2	SLC9A2	NM_003048	2	10	16646	1
solute carrier family 9, isoform A4	SLC9A4	NM_001011552	2	11	15714	1
ATPase Ca(2+)-transporting, slow-twitch≠	ATP2A2	NM_001681,NM_17 0665	12	10	23130	2
solute carrier family 6, member 4	SLC64A (SLC6A 4)	NM_001045	17	9	15765	1
GROWTH FACTORS						
transforming growth factor, beta 1	TGF-b (TGFB1)	NM_000660	19	5	8407	1
fibroblast growth factor 1 (acidic) ≠	FGFa (FGF1)	NM_000800,NM_03 3137,NM_033136	5	4	6346	3
insulin-like growth factor 1≠	IGF1	NM_001111285,NM _001111284,NM_00 1111283,NM_00061 8	12	6	13820	4
early growth response 1	EGR1	NM_001964	5	1	4825	1
insulin like growth factor 2≠	IGF2	NM_001127598,NM _000612,NM_00100 7139	11	6	13161	3
FACTORS INVOLVED IN CARDIOMYOCYTE GROWTH AND CONTRACTILITY						
beta1 adrenergic receptor 1	ADRB1	NM_000684	10	1	3863	1
angiotensin receptor 1≠	AGTR1	NM_004835,NM_00 0685,NM_032049,N M_009585,NM_0318 50	3	5	7576	5
endothelin 1	EDN1	NM_001955	6	4	6268	1
endothelin receptor, type A	EDNRA	NM_001957	4	7	11745	1
angiotensinogen	AGT	NM_000029	1	4	7451	1
endothelin receptor. Type B≠	EDNRB	NM_000115,NM_00 3991,NM_00112265 9	13	5	10584	3
Protein kinase C, alpha	PRKCA	NM_002737	17	15	25199	1
INFLAMMATION						
interleukin 6	IL6	NM_000600	7	3	5003	1
interleukin 6 receptor≠	IL6R	NM_000565,NM_18 1359	1	8	13285	2

phospholipasi A2	PLA2G1B	NM_000928	12	3	4367	1
phospholipase D1, phosphatidylcholine-specific≠	PLD1	NM_002662,NM_001130081	3	22	30391	2
phospholipase D1, glycosylphosphatidylinositol-specific≠	GPLD1	NM_177483,NM_001503	6	16	23339	2
nuclear factor kappa-B, subunit 1	NFKB1	NM_003998	4	20	26029	1
interleukin 13	IL13	NM_002188	5	2	3882	1
toll-like receptor 4 precursor	TLR4	NM_138554	9	3	8504	1
inhibitor of kappa light polypeptide	IKBKB	NM_001556	8	14	21503	1
interleukin 23, alpha subunit	IL23A	NM_016584	12	1	2533	1
toll-like receptor 2	TLR2	NM_003264	4	3	6405	1
interleukin 27	IL27	NM_145659	16	4	5132	1
prostaglandin-endoperoxide synthase 2	PTGS2	NM_000963	1	1	9589	1
macrophage migration inhibitory factor	MIF	NM_002415	22	1	1846	1
interleukin 17A	IL17A	NM_002190	6	3	4862	1
interleukin 12B	IL12B	NM_002187	5	6	9479	1
C-reactive protein, pentraxin-related	CRP	NM_000567	1	1	3302	1
arachidonate lipoxygenase 12	ALOX12	NM_000697	17	5	10490	1
arachidonate lipoxygenase 15	ALOX15	NM_001140	17	4	9462	1
arachidonate lipoxygenase 15, type b≠	ALOX15B	NM_001039130,NM_001039131,NM_001141	17	3	8768	3
arachidonate lipoxygenase 5 activating protein	ALOX5AP	NM_001629	13	5	5878	1
arachidonate 5-lipoxygenase	ALOX5	NM_000698	10	9	13759	1
interleukin 1, alpha proprotein	IL1A	NM_000575	2	4	9444	1
interleukin 10	IL10	NM_000572	1	3	5783	1
chemokine (C-C motif) receptor 2≠	CCR2	NM_000647,NM_000648	3	2	5769	2
chemokine (C-X3-C motif)	CX3CR1	NM_001337	3	2	5110	1
chemokine binding protein 2	CCBP (CCBP2)	NM_001296	3	3	5963	1
phospholipase A2	PLA2G7	NM_005084	6	8	12235	1
small inducible cytokine A2 precursor	CCL2	NM_002982	17	1	2926	1
prostaglandin-endoperoxide synthase 1≠	PTGS1	NM_080591,NM_000962	9	7	12923	2
signal transducer and activator of transcription≠	STAT3	NM_213662,NM_139276,NM_003150	17	11	20992	3
selenoprotein S	SEPS1	NM_000808	X	10	12795	1
tumor necrosis factor (ligand) superfamily, member 10	TNFSF10	NM_003810	3	5	6764	1

chemokine (C-X-C motif) ligand 12≠	CXCL12	NM_199168,NM_000609,NM_001033886	10	5	10238	3
interleukin 1, beta	IL1B	NM_000576	2	3	6794	1
chemokine (C-C motif) receptor 6≠	CCR6	NM_031409,NM_004367	6	3	6786	2
interleukin 12A	IL12A	NM_000882	3	3	5923	1
interleukin 11	IL11	NM_000641	19	3	5667	1
interleukin 8 receptor, beta	IL8RB	NM_001557	2	3	5862	1
interleukin 18	IL18	NM_001562	11	6	7151	1
interleukin 1 receptor, type I	IL1R	NM_000877	2	6	14049	1
tumor necrosis factor alpha	TNF	NM_000594	6	1	3764	1
TNF receptor-associated factor 4	TRAF4	NM_004295	17	2	5999	1
interferon, alpha 1	IFNA1	NM_024013	9	1	1877	1
interferon, beta 1	IFNB1	NM_002176	9	1	1841	1
desmoplakin≠	DSP	NM_004415,NM_001008844	6	8	28150	2
complement component 5 preproprotein	C5	NM_001735	9	32	43248	1
small inducible cytokine A27 precursor	CCL27	NM_006664	9	1	1798	1
interleukin 11 receptor, alpha≠	IL11RA	NM_004512,NM_147162	9	4	8591	2
chemokine (C-C motif) ligand 7	CCL7	NM_006273	17	1	3018	1
chemokine (C-X-C motif) ligand 2	CXCL2	NM_002089	4	1	3245	1
interleukin 17F precursor	IL17F	NM_052872	6	3	3811	1
interleukin 27 receptor, alpha	IL27RA	NM_004843	19	6	11056	1
chemokine (C-X-C motif) ligand 13	CXCL13	NM_006419	4	4	5502	1
suppressor of cytokine signaling 3	SOCS3	NM_003955	17	1	4298	1
leukotriene B4 receptor	LTB4R	NM_181657	14	1	4637	1
chemokine (C-C motif) receptor 1	CCR1	NM_001295	3	2	4678	1
interleukin-1 receptor-associated kinase 3	IRAK3	NM_007199	12	9	12273	1
chemokine-like receptor 1 isoform b	CMKLR1	NM_004072	12	3	4889	1
vascular cell adhesion molecule 1≠	VCAM1	NM_001078,NM_080682	1	7	11559	2
MAP KINASES						
mitogen-activated protein kinase 1≠	MAPK1	NM_002745,NM_138957	22	8	14393	2
mitogen-activated protein kinase 3≠	ERK1 (MAPK3)	NM_001040056,NM_001109891,NM_002746	16	4	6964	3
mitogen-activated protein kinase 14 ≠	MAPK14	NM_139014,NM_139013,NM_139012,NM_001315	6	9	15910	4
mitogen-activated protein kinase 8 ≠	MAPK8	NM_002750,NM_139049,NM_139047,NM_139046	10	8	10827	4
mitogen-activated protein kinase	MAP3K1	NM_005921	5	11	22936	1

TRANSCRIPTIONAL FACTORS						
nuclear factor of activated T cells 5 ≠	NFAT5	NM_173214,NM_138713,NM_006599,NM_001113178,NM_138714	16	11	28339	5
peroxisome proliferator-activated receptor-alpha ≠	PPARA	NM_001001928,NM_005036	22	9	19058	2
peroxisome proliferator-activated receptor-delta	PPARD	NM_006238	6	7	10921	1
peroxisome proliferator-activated receptor-gamma≠	PPARG	NM_005037,NM_138712,NM_138711,NM_015869	3	9	12188	4
MADS box transcription enhancer factor 2, polypeptide A≠	MEF2A	NM_001130927,NM_001130928,NM_001130926,NM_005587	15	11	17570	4
MADS box transcription enhancer factor 2, polypeptide B≠	MEF2B	NM_001134794,NM_001134795,NM_005919	19	7	11212	3
MADS box transcription enhancer factor 2, polypeptide C≠	MEF2C	NM_002397,NM_001131005	5	11	19351	2
GATA-binding protein 4	GATA4	NM_002052	8	7	10415	1
GATA-binding protein 5	GATA5	NM_080473	20	3	7206	1
GATA-binding protein 6	GATA6	NM_005257	18	6	9595	1
forkhead box J1	HFH4 (FOXJ1)	NM_001454	17	2	4377	1
MADS box transcription enhancer factor 2, polypeptide D	MEF2D	NM_005920	1	7	15031	1
V-FOS FBJ murine osteosarcoma viral oncogene homolog	FOS	NM_005252	14	1	4383	1
heart and neural crest derivatives-expressed 1	HAND1	NM_004821	5	2	3739	1
heart and neural crest derivatives-expressed 2	HAND2	NM_021973	4	2	4370	1
MK2 homeobox 5	NKX2-5	NM_004387	5	2	3587	1
NOTCH, Drosophila, homolog of, 1	NOTCH 1	NM_017617	9	9	28501	1
ROS PRODUCTION						
NADPH oxidase 1≠	NOX1	NM_007052,NM_013955	X	5	10228	2
xanthine dehydrogenase	XDH	NM_000379	2	23	36677	1
nitric oxide synthase 1	NOS1	NM_000620	12	25	34743	1
nitric oxide synthase 2A	NOS2A	NM_000625	17	14	26423	1
nitric oxide synthase 3	NOS3	NM_000603	7	9	18054	1
HISTONE (DE)ACETYLASES						
histone acetyltransferase 1≠	HAT1	NM_001033085,NM_003642	2	8	10753	2
histone deacetylase 1	HDAC1	NM_004964	1	7	11303	1
histone deacetylase 2	HDAC2	NM_001527	6	8	18079	1
histone deacetylase 3	HDAC3	NM_003883	5	5	9428	1
histone deacetylase 4	HDAC4	NM_006037	2	24	34315	1
histone deacetylase 5≠	HDAC5	NM_005474,NM_001015053	17	7	18729	2
histone deacetylase 6	HDAC6	NM_006044	X	6	15180	1
histone deacetylase 7A≠	HDAC7	NM_001098416,NM	12	9	18377	2

	A	_015401				
histone deacetylase 8	HDAC8	NM_018486	X	9	11987	1
histone deacetylase 9≠	HDAC9	NM_058176,NM_178425,NM_178423,NM_014707,NM_058177	7	22	30572	5
FKBP12-rapamycin complex-associated protein 1	FRAP1	NM_004958	1	34	52998	1
OTHERS						
RAS homolog gene family, member A	RHOA	NM_001664	3	5	6926	1
RAS-related C3 botulinum toxin substrate 1≠	RAC1	NM_006908,NM_198829,NM_018890	7	7	9751	3
sperm associated antigen 6	SPAG6	NM_012443	10	9	12552	1
sperm associated antigen 16	SPAG16	NM_024532	2	16	18190	1
cell division cycle 42≠	CDC42	NM_044472,NM_001791,NM_001039802	1	6	9905	3
Coiled-coil domain-containing protein 63	FLJ35843 (CCDC63)	NM_152591	12	11	12214	1
Coiled-coil domain-containing protein 114	FLJ32926 (CCDC114)	NM_144577	19	2	5262	1
actin-binding RHO-activating protein	ABRA	NM_139166	8	2	4756	1
RHO-associated coiled-coil-containing protein kinase1	ROCK1	NM_005406	18	24	34738	1
ubiquitin-specific protease 7	USP7	NM_003470	16	13	26583	1
Myocardin	MYOCD	NM_153604	17	13	15830	1
serum response factor	SRF	NM_003131	6	4	9434	1
homeodomain-only protein≠	HOP (HOPX)	NM_139212,NM_139211,NM_032495	4	4	5703	3
lymphocyte antigen CD5-like	CD5L	NM_005894	1	5	8136	1
oral-facial-digital syndrome 1	OFD1	NM_003611	X	14	21584	1
WD repeat domain 78	WDR78	NM_024763	1	15	19908	1
WD repeat domain 63	WDR63	NM_145172	1	20	25551	1
Tctex1 domain containing 1	FLJ40873 (TCTEX1D1)	NM_152665	1	4	6853	1
Tctex1 domain containing 4	LOC343521 (TCTEX1D4)	NM_001013632	1	1	2373	1
phosphatidylinositol 3-kinase, catalytic, alpha	PIK3CA	NM_006218	3	10	18997	1
3-a phosphatidylinositide-dependent protein kinase 1≠	PDPK1	NM_031268,NM_002613	16	10	18770	2

retinitis pigmentosa GTPase regulator≠	RPGR	NM_001034853,NM_000328	X	16	23170	2
V-AKT murine thymoma viral oncogene homolog 1≠	AKT1	NM_001014432,NM_005163,NM_001014431	14	5	12302	3
adenylate kinase 1	AK1	NM_000476	9	4	7436	1
adenylate kinase 5	AK5	NM_174858	1	13	16363	1
protein phosphatase 3, catalytic subunit, alpha isoform≠	PPP3CA	NM_000944,NM_001130691,NM_001130692	4	14	18690	3
protein phosphatase 3, catalytic subunit, beta isoform	PPP3CB	NM_021132	10	10	14241	1
protein phosphatase 3, catalytic subunit, gamma isoform	PPP3CC	NM_005605	8	10	14234	1
galactosidase alpha	GLA	NM_000169	X	4	6783	1
sphingomyelin phosphodiesterase 1, acid lysosomal≠	SMPD1	NM_001007593,NM_000543	11	2	5555	2
Radial spoke head 1 homolog	RSPH1	NM_080860	21	7	9041	1
Radial spoke head 3 homolog	RSPH3	NM_031924	6	6	9567	1
flap structure-specific endonuclease1	FEN1	NM_004111	11	2	4251	1
formyl peptide receptor-like 1≠	FPR2	NM_001005738,NM_001462	19	3	5045	2
selectin P precursor	SELP	NM_003005	1	11	18226	1
selectin E	SELE	NM_000450	1	3	12007	1
paraoxonase 1	PON1	NM_000446	7	9	11442	1
paraoxonase 3	PON3	NM_000940	7	8	9468	1
paraoxonase 2≠	PON2	NM_001018161,NM_000305	7	6	9561	2
prosaposin≠	PSAP	NM_002778,NM_001042465,NM_001042466	10	8	13885	3

*trIDs: transcripts identification; †chr: chromosome; ‡Segs (n): number of segments; §Segs (bp): length (bp) of the screened gene regions; ||Transc (n): number of transcripts captured; ≠: more than one transcript was captured.

APPENDIX 2

Publications

1. Errico F, **D'Argenio V**, Sforazzini F, Iasevoli F, Squillace M, Guerri G, Napolitano F, Angrisano T, Di Maio A, Keller S, Vitucci D, Galbusera A, Chiariotti L, Bertolino A, de Bartolomeis A, Salvatore F, Gozzi A, Usiello A. A role for D-aspartate oxidase in schizophrenia and in schizophrenia-related symptoms induced by phencyclidine in mice. Transl Psychiatry 2015;5:e512.
2. **D'Argenio V**, Salvatore F. The role of the gut microbiome in the healthy adult status. Clin Chim Acta 2015. doi:10.1016/j.cca.2015.01.003.
3. **V D'Argenio**, F Salvatore. Psoriasis genetics: State of the art. G Ital Dermatol Venereol 2014;149 (suppl 5):39-41.
4. **Valeria D'Argenio**, Eugenio Notomista, Mauro Petrillo, Piergiuseppe Cantiello, Valeria Cafaro, Viviana Izzo, Barbara Naso, Luca Cozzuto, Lorenzo Durante, Luca Troncone, Giovanni Paolella, Francesco Salvatore, Alberto Di Donato. Complete sequencing of Novosphingobium sp. PP1Y reveals a biotechnologically meaningful metabolic pattern. BMC Genomics 2014. 15:384.
5. Aceto S, Sica M, De Paolo S, **D'Argenio V**, Cantiello P, Salvatore F, Gaudio L. The Analysis of the Inflorescence miRNome of the Orchid Orchis italica Reveals a DEF-Like MADS-Box Gene as a New miRNA Target. PLoS One 2014; 9:e97839.
6. **Valeria D'Argenio**, Giorgio Casaburi, Vincenza Precone, Francesco Salvatore. Comparative Metagenomic Analysis of Human Gut

Microbiome Composition Using Two Different Bioinformatic Pipelines. Biomed Res Int 2014; 2014:325340.

7. **Valeria D'Argenio**, Maria Valeria Esposito, Jean Ann Gilder, Giulia Frisso, Francesco Salvatore. Should a BRCA2 Stop Codon Human Variant, Usually Considered a Polymorphism, Be Classified as a Predisposing Mutation? Cancer 2014;120:1594-5.
8. **V D'Argenio**, G Frisso, V Precone, A Boccia, A Fienga, G Pacileo, G Limongelli, G Paolella, Raffaele Calabrò, F Salvatore. DNA sequence capture and next generation sequencing for the molecular diagnosis of genetic cardiomyopathies. J Mol Diagn 2014; 16:32-44.
9. **V D'Argenio**, V Precone, G Casaburi, E Miele, M Martinelli, A Staiano, F Salvatore, L Sacchetti. An Altered Gut Microbiome Profile in a Child Affected by Crohn's Disease Normalized After Nutritional Therapy. American Journal of Gastroenterology 2013;108(5):851-2. doi: 10.1038/ajg.2013.46. IF=9.2; Citazioni=7
10. **D'Argenio V**, Petrillo M, Cantiello P, Naso B, Cozzuto L, Notomista E, Paolella G, Di Donato A, Salvatore F. De novo sequencing and assembly of the whole genome of Novosphingobium sp.PP1Y. J Bacteriol 2011; 193: 4296.

Oral Communications

1. Analisi del Microbiota intestinale. Corso Precongressuale "Le tecnologie di sequenziamento massivo parallelo applicate alla diagnostica molecolare clinica". 46° Congresso Nazionale SIBioc, Roma, 13 Ottobre 2014.
2. Tecniche di Next Generation Sequencing per lo studio del microbioma: applicazioni in patologia umana. I Workshop

ARFACID “Le frontiere della microbiologia nella moderna pratica clinica”. Napoli, 3 Ottobre 2014.

3. Next generation sequencing as a tool for data acquisition at genomic level: examples in prokaryotes and eukaryotes. EMBO workshop “The Genome: Structure, Expression And Evolution”. Napoli, 22 Settembre 2014.
4. Ruolo del microbioma nella sarcoidosi polmonare. XII Corso Nazionale di Biologia Cellulare e Molecolare in Pneumologia – BIOCEP. Napoli, Ospedale Monaldi, 23 Giugno 2014.
5. Implementing CFTR diagnostic testing. Multiplicom Corporate Satellite Meeting “Advances of MASTR™ in routine clinical diagnostics”. ESHG2014, Milano, 1 Giugno 2014.
6. Le basi molecolari per un biorisanamento avanzato: tecnologie genomiche per lo studio dei microrganismi. Giornata di Studio: biotecnologie e risanamento dei suoli. Roma, Senato della Repubblica, Sala Santa Maria in Aquiro, 7 Febbraio 2014.
7. La genetica della psoriasi: stato dell’arte. Le Psoriasi – Convegno multidisciplinare. Napoli, Centro Congressi Federico II, 28-30 Novembre 2013.
8. Il ruolo del microbioma nelle malattie infiammatorie croniche intestinali. Tecnologia 454: una finestra sulla biodiversità microbica. Milano, Museo della scienza e della tecnologia, 21 Novembre 2013.
9. Analisi di miRNA attraverso Next Generation Sequencing. Le Giornate Mediterranee di Medicina di Laboratorio. IV Congresso Interregionale SIBioC. Sorrento (NA), Hilton Sorrento Palace, 10 Ottobre 2013.

10. BRCA1 and BRCA2 mutations through Next Generation Sequencing. Breast Cancer – Progress and Controversies. Napoli, Hotel Royal Continental, 14 Giugno 2013.
11. Next generation sequencing in research and diagnostics of genetic cardiomyopathies. EUROMEDLAB Milano 2013. Milano, 22 Maggio 2013.
12. NGS in the Study of Human Diseases: the Examples of Cardiomyopathies and Ocular Diseases. The Translational Science of Mendelian Disorders from Transomics to Dally Life. BGI Next Generation Sequencing Workshop. Milano, 5 Dicembre 2012.
13. NGS e Medicina: esempi nello studio di cardiomiopatie e patologie oculari. Corso di Aggiornamento professionale F.I.Bio.: Next Generation Sequencing applications and future perspectives. Napoli, CEINGE Biotecnologie Avanzate, 27 Aprile 2012.
14. Target enrichment strategies for next generation sequencing technologies for the study of human diseases: the example of hypertrophic cardiomyopathies. Cambridge Healthtech Institute: Innovative Sample Prep & Target Enrichment in Clinical Diagnostics. Newport Beach (CA, USA), Hyatt Regency Hotel, 18-19 Aprile 2012.
15. Next Generation Sequencing in cardiomyopathies. Mediterranean school in cardiovascular sciences. Vietri sul Mare (SA), LIoyd's Baia Hotel, 20 Ottobre 2011.
16. Analisi del DNA attraverso sequenziamento High Throughput. Corso PFA n°261-1821: La genetica nella Pratica Clinica III. San Giovanni Rotondo (Fg), Casa Sollievo della Sofferenza, 30 Settembre 2011.

17. L’analisi del genoma attraverso il sequenziamento degli acidi nucleici. Aggiornamenti in Medicina e Tecnologia Molecolare. Caserta, Complesso Monumentale di San Leucio, 11 Febbraio 2011.

Proceedings

1. **Valeria D’Argenio**, Maria Valeria Esposito, Massimiliano D’Aiuto, Antonella Telese, Marcella Nunziato, Flavio Starnone, Alessandra Calabrese, Giulia Frisso, Giuseppe D’Aiuto, Francesco Salvatore. Next generation sequencing screening of the BRCA1 and BRCA2 genes. SIGU 2014, Bologna 30-31 Ottobre 2014.
2. Valentina del Monaco, **Valeria D’Argenio**, Massimiliano D’Aiuto, Fatima De Palma, Donatella Montanaro, Giuseppina Liguori, Giuseppe D’Aiuto, Gerardo Botti, Alfonso Baldi, Raffaele Calogero, Francesco Salvatore. Comprehensive transcriptome profiling of breast cancers. ESHG2014, Milan 31Maggio-3 Giugno 2014.
3. **Maria Valeria Esposito**, Massimiliano D’Aiuto, Antonella Telese, Vincenza Precone, Marcella Nunziato, Alessandra Calabrese, Giulia Frisso, Giuseppe D’Aiuto, Valeria D’Argenio, Francesco Salvatore. BRCA1 and BRCA2 mutation detection by a Next Generation Sequencing approach: an epidemiological study conducted in Southern Italy. ESHG2014, Milan 31Maggio-3 Giugno 2014.
4. **Valeria D’Argenio**, Maria Valeria Esposito, Massimiliano D’Aiuto, Alessandra Calabrese, Giuseppe D’Aiuto, Francesco Salvatore. Analysis of a novel BRCA1 splicing mutation in hereditary breast and ovarian cancer woman. ESHG2014, Milan 31Maggio-3 Giugno 2014.

5. Antonella Telese, **Valeria D'Argenio**, Irene Postiglione, Paola Nardiello, Giuseppe Castaldo, Francesco Salvatore. Validation of a next generation sequencing approach for rapid and accurate CFTR mutations screening. ESHG2014, Milan 31Maggio-3 Giugno 2014.
6. **V. D'Argenio**, M.V. Esposito, M. D'Aiuto, V. Precone, P. Cantiello, A. Calabrese, G. Frisso, G. D'Aiuto, F. Salvatore. BRCA1 and BRCA2 mutation detection by a next generation sequencing approach: an epidemiological study in Southern Italy. SIC2013.
7. **V. D'Argenio**, G. Casaburi, V. Precone, C. Ciacci, J.C. Caporaso, L. Sacchetti, F. Salvatore. Characterization of the entire celiac disease intestinal microbiome by Next Generation Sequencing. EUROMEDLAB Milano 2013. Milano, 19-23 Maggio 2013.
8. **V. D'Argenio**, M.V. Esposito, M. D'Aiuto, V. Precone, P. Cantiello, A. Calabrese, G. Frisso, G. D'Aiuto, F. Salvatore. BRCA1 and BRCA2 rapid germline mutations screening by Next Generation Sequencing approach. EUROMEDLAB Milano 2013. Milano, 19-23 Maggio 2013.
9. G. Esposito, **V. D'Argenio**, G. Guerri, G. Sauchelli, A. Boccia, I.C.M. Tandurella, M. D'Antonio, F. De Falco, G. Paolella, F. Salvatore. A novel mutation in RP1 is a major cause of autosomal dominant retinitis pigmentosa in Southern Italy. EUROMEDLAB Milano 2013. Milano, 19-23 Maggio 2013.
10. **V. D'Argenio**, G. Guerri, A. Telese, A. Palmieri, A. Daniele, F. Salvatore. Long-range PCR and Next Generation Sequencing for the identification of PAH mutation status in HPA italian patients. EUROMEDLAB Milano 2013. Milano, 19-23 Maggio 2013.

11. **D'Argenio V**, Petrillo M, Naso B, Cantiello PG, Pagliarulo C, Cozzuto L, Salvatore P, Alifano P, Paolella G, Salvatore F. New insights about size (12 Mb) and evolution of a "rare actinomycete" by whole genome sequence of Nonomuraea sp. ATCC 39727. 29th Congresso Nazionale SIMGBM. Pisa, 21-23 Settembre 2011.
12. Carata E, Colicchio R, Talà A, Pagliuca C, Pasanisi D, **D'Argenio V**, Paolella G, Salvatore F, Salvatore P, Alifano P. Searching for novel secondary metabolites by genome data mining in Nonomuraea sp. ATCC 39727. 29th Congresso Nazionale SIMGBM. Pisa, 21-23 Settembre 2011.
13. **D'Argenio V**, Frisso G, Boccia A, Fienga A, Limongelli G, Precone V, Pacileo G, Calabrò R, Paolella G, Salvatore F. DNA sequence capture array and next generation sequencing to identify new disease-causing genes: the case of hypertrophic cardiomyopathy. 36th FEBS Congress. Torino, 25-30 Giugno 2011. FEBS JOURNAL 2011; 278SI (Suppl 1): 283.