J Biochem Tech (2016) 7(3): 1093-1101
ISSN: 0974-2328

Comparison of two NGS platforms for metagenomic analysis of clinical samples

Tommaso Biagini, Barbara Bartolini, Emanuela Giombini, Fabrizio Ferrè, Marina Selleri, Gabriella Rozera, Maria R. Capobianchi, Giovanni Chillemi, Alessandro Desideri

Received: 17 September 2015 / Received in revised form: 06 April 2016, Accepted: 03 January 2017, Published online: 18 February 2017

Abstract

The advent of next generation sequencing platforms has greatly improved the direct assessment of human metagenomes permitting to obtain results in a time-frame compatible with diagnostic needs. The performances of two platforms, namely Roche 454 GS FLX+ and Illumina MiSeq, have been compared using a metagenomic approach aimed at the identification of the pathogenic agent, an RNA virus in this case, in a clinical sample. Both platforms resulted able to correctly identify the H1N1 virus in the sample, but also provided a similar overview of the microbial community. The detailed analysis of two very different but clinically relevant microorganisms that often co-infect patients, H1N1 and Streptococcus pneumoniae, showed differences in terms of depth of coverage and genome coverage, and showed that some genomic regions are more frequently represented in Illumina sequencing reads, while others in the 454 reads. These findings indicate that the platforms and the corresponding analysis procedures permit the high quality assembly of the H1N1 virus and the identification of the complex microbial community in the sample, supporting the usage of these approaches in a clincal setting.

Tommaso Biagini, Alessandro Desideri

Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica, 00133, Rome, Italy and Molecular Digital Diagnostics (MDD), Via S. Camillo de Lellis 01100 Viterbo Italy

Barbara Bartolini, Emanuela Giombini, Marina Selleri, Gabriella Rozera, Maria R. Capobianchi

"L. Spallanzani" National Institute for Infectious Diseases, Via Portuense 292, 00149 Rome, Italy

Fabrizio Ferrè

Department of Pharmacy and Biotechnology (FaBiT), University of Bologna Alma Mater, Via Belmeloro 6, 40126 Bologna, Italy

Giovanni Chillemi

CINECA, SCAI SuperComputing Applications and Innovation Department, Via dei Tizii 6, Rome 00185, Italy

Email: giochillemi@gmail.com

Keywords: de novo assembly; metagenomics; pathogen detection; Roche 454 sequencing; Illumina sequencing; next generation sequencing

Introduction

Massive parallel sequencing (MPS) represents a powerful methodology for the direct assessment of human metagenomes, i.e. all the nucleic acid sequences belonging to the entire microbial community present in different body sites. The computational handling of the large amount of sequence data generated in a high-throughput sequencing run, is an area of necessary research focus.

Metagenomic approaches through MPS have many advantages over traditional methods of pathogen detection, such as PCR, ELISA, culture, etc. (Gardner et al., 2003; Lemmon and Gardner, 2008). In fact, they allow the detection of non-cultivable microorganisms, are not dependent on a priori knowledge of the microorganism under investigation, and do not necessarily rely on predefined target genomes (Moore et al., 2011). The advent of next generation sequencing (NGS) platforms has greatly improved the metagenomic analysis, thanks to the high-content sequence output. In particular, the relative speed of some NGS platforms allows obtaining results in a time-frame compatible with diagnostic needs. In fact our time of analysis ranges from a minimum of 24 hours to a maximum of 48 hours, that added to the processing/sequencing time, allowing to obtain a complete and accurate unbiased analysis of clinical samples within a few days overall.

Despite their huge potential in metagenomic analysis, NGS technologies are affected by significant unresolved challenges, such as difficulties in sample preparation, insufficient depth and breadth of sequencing coverage for the detection of pathogens present at very low levels within the sample, and lack of appropriate reference genomes for sequencing reads alignment. Furthermore, at present time there are no standard criteria in terms of what is intended with "identification" of a pathogen in a sample and what is the minimum number of organism-specific reads necessary to make a true positive call. In fact, criteria to invoke confidence may vary per sample

type/complexity or may be organism-dependent (Moore et al., 2011; Yang et al., 2011).

A large depth of coverage increases the confidence of microbial strain calls, but again, there is no agreed standard regarding the minimum depth of coverage for metagenome samples and, in many cases, it would be reasonable to expect that a region of a pathogen genome, detected within a metagenomic sample, may be only present within the reads at 1x coverage (i.e. represented by one read only). Just as there is currently no agreed-upon standard to indicate what breadth or depth of coverage would be required to reach a "microorganism identification," currently there is paucity of knowledge regarding the actual limits of detection for each sequencing platform and protocol (Frey et al., 2014; Luo et al., 2012). These problems notwithstanding, successful examples of NGS metagenomic applications for the detection and/or characterization of causal agents in diseases of animals and other eukaryotes can be found in the literature (Biagini et al., 2014; Greninger et al., 2010; Palacios et al., 2011; Towner et al., 2008). Most of these studies are directed to the identification of bacterial genomes using amplicon-based approaches targeting the 16S rDNA as a universal bacterial target, while studies based on MPS, more suitable for bacterial genome reconstruction and pathogen discovery, are less standardized. A critical limitation in the use of metagenomics for the identification of viral pathogens is the lack of a universal viral genome marker, analogous to bacterial 16S rDNA (Sergeant et al., 2014; Ugalde et al., 2013). Despite these issues, metagenomic analysis based on NGS has been useful in discovering novel pathogenic viruses (Cheval et al., 2011; Lu et al., 2014). Moreover, the complete characterization of the microbial community in samples from patients with viral infections allows the investigation of the role of co-infections, as it has been the case of pandemics and seasonal influenza (Nisii et al., 2010; Smith et al., 2013).

In the studies of viral and eventually bacterial metagenomes in biological samples for diagnostic purpose, the development of bioinformatic procedures able to produce fast and accurate identification of pathogen(s) is mandatory (Barzon et al., 2013; Capobianchi et al., 2013). Roche 454 and Illumina/Solexa platforms produce millions of short sequence reads, which vary in length from tens of base pairs to ~800 nt. A first step in the analysis of metagenomic data can be the assembly of sequences into longer contigs, since this procedure might improve functional annotation, increases accuracy and decreases ambiguity of assignment (Howe et al., 2014; Palacios et al., 2011). The success is limited by the formation of chimeric contigs, i.e. the assembly of reads originating from different taxonomic groups due to homology and random similarities, occurring at all taxonomic levels, hence the selection of the assembly algorithm and the setting of its parameters become crucial and might depend on the focus of each study. It is critical to assess the quality of the generated contigs, and several studies have attempted to evaluate the sequencing errors and artifacts specific for each NGS platform (Luo et al., 2012; Quince et al., 2009). These issues are particularly relevant when microorganism identification and characterization must occur in real clinical samples (Petty et al., 2014). In fact, a A small number of studies concerning the efficiency of different platforms in metagenomic applications have been reported (Jia et al., 2013; Mitra et al., 2010), but up to now there is no comparative analysis aimed at real clinical samples. The identification of a solid procedure is necessary to detect virus such as the H1N1 infuenza virus that displays a large genetic diversity due to high mutation rate and multiple reassortment and that in 2009 has rapidly spread worldwide by human to human transmission (Dawood et al., 2009).

In this study, the performances of two NGS platforms, namely the Roche 454 GS FLX+ and the Illumina MiSeq, were compared for the ability in identifying pathogenic agents in a metagenomic clinical sample, i.e. a nasopharyngeal swab submitted to the laboratory for the diagnosis of influenza. Although obtained on a single clinical sample, the results indicate that both platforms, when coupled to the appropriate bioinformatic pipelines, are able to efficiently detect both RNA and DNA pathogens, although with interesting differences in the extent and depth of genomic coverage.

Materials and Methods

Sample Preparation and quality control

A nasopharyngeal swab, randomly selected from samples submitted to the Laboratory of Virology and resulted positive to 2009 pandemic influenza A H1N1 strain by means of conventional real time PCR, was analyzed (influenza H1N1pdm09 M gene load: 3.3x107 cp/ml). The amplification products, generated with a prior semi-random RT-PCR (Bartolini et al., 2011), after purification (AMPure beads) and quantification (TBS 380 Fluorometer) were split in two aliquots to be sequenced with two platforms: Roche 454 GS FLX+ and Illumina MiSeq.

The RNA was extracted by using a QIAmp viral RNA minikit (Qiagen) and eluted in 60 µl of RNase-free water. Reverse transcription (RT) was performed with a random tagged primer as previously described (Allander et al., 2005; Bartolini et al., 2011). More in detail, 10 µl of the extracted RNA was mixed with primer FR26RV-N (5'GCCGGAGCTCTGCAGATA TCNNNNNN3') at 10 µM, incubated at 65°C for 5 min and chilled on ice. A mix containing 4 µl of 5×First-Strand buffer (Invitrogen), 2 µl of DTT (100 mM), 1µl dNTPs (10 mM), 0.2 µl (8U) of recombinant RNase inhibitor (Promega), and 0.5 µl (100U) of SuperScript III reverse transcriptase (Invitrogen) was incubated at 25°C for 10 min and at 42°C for 50 min. After denaturation (94°C for 3 min and chilling on ice), 2.5U of 3′-5′ exo- Klenow DNA polymerase (New England Biolabs) were added. The reaction was incubated at 37°C for 1 h, followed by 75°C for 10 min for enzyme inactivation.

Five µl of cDNA from each sample were then used as a template for the PCR amplification, that was performed with High Fidelity Taq Gold DNA polymerase (Applied Biosystems) with the FR20RV primer (5'GCCGGAGCT CTGCAGATATC3'); cycling conditions were: 10 min at 94°C, 40 cycles of amplification (94°C for 1 min, 65°C for 1 min, and 72°C for 2 min).

Roche 454 GS FLX+ sequencing

The amplicons were linked to adapter molecules to allow library preparation (rapid library preparation kit). The size of the amplicons after the PCR amplification of the cDNA ranged between 100 and 1800 bp. Before starting the sequencing procedures, the PCR products were purified with AMPure beads to remove small fragments. All emPCR reactions were performed using GS FLX + Titanium Lib-L-LV kits. Template-to-bead ratios were optimized via titration. The subsequent sequencing run was performed using a pico-titre plate with two-region gasket following the manufacturer's

instructions (Roche, Titanium version, GS FLX + platform) at the "L. Spallanzani" National Institute for Infectious Diseases. The starting 454 dataset was obtained after filtering the PCR primers from the obtained sequencing reads using in-house python script. When the primer sequence position was found internally in the read, the read was split into two parts and the read sequence primer was eliminated. The resulting dataset were composed of 453,050 reads ranging in size from 50 to 800 nt (average length: 340 nt).

Illumina MiSeq sequencing

Genomic DNA was sequenced at the University of Salerno, dept. Of medicine and Surgery, Lab of Molecular Medicine and Genomics, using a MiSeq Sequencer. Sequencing produced 9,640,522 paired-end reads of 250+250 bases. The library was produced using the Nextera XT kit, and the DNA insert size was between 400bp and 1.5 kbp. Sequence quality control was performed using FastQC Version 0.11.2, downloaded from the Babraham Bioinformatics Institute (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and trimming and adaptor removal using Trimmomatic (Bolger et al., 2014) with default parameters.

De novo assembly of 454 data

For this analysis we followed the protocol fully described in Biagini et al (Biagini et al., 2014). In particular, the de novo assembly of reads into contigs was performed on the 454 dataset using an Overlap Layout Consensus (OLC) algorithm as implemented in newbler (Version 2.6) using a minimum overlap cut-off of 50 bp with 90% sequence identity. The output of newbler was subject of a second round of assembly and refinement using CAP3 (Huang and Madan, 1999), with default parameters. A contig is defined as the set of overlapping DNA reads representing a consensus region of genome and a singleton as a read that did not show any significant overlap with any othe read but that, even alone, it represents a significant genome region. Both are considered in the Megan and statistical analysis). The minimal considered contig length is 100 bp.

De novo assembly of Illumina data

For the MiSeq data, the de novo assembly of reads into contigs was done using different programs (Meta-Velvet version 1.2.01 (Namiki et al., 2012), Ray Meta (version v2.3.1) (Boisvert et al., 2012), Abyss (version 1.5.2) (Simpson et al., 2009)) based on the De Brujin Graph algorithm (Miller et al., 2010). Each program was run varying the k-mer length value between 21 and 64. A contig is defined as the set of overlapping DNA reads representing a consensus region of genome and a singleton as a read that did not show any significant overlap with any other read but that, even alone, it represents a significant genome region. Both are considered in the Megan and statistical analysis. The minimal considered contig length is 100 bp.

Taxonomic classification and pathogen genome analysis

The reads belonging to the starting datasets, as well as the resulting assembled contigs, were aligned using nucleotide BLAST, independently for the two sequencing technologies, on the NCBI nucleotide database using the following cut-offs: identity > 85%; overlap > 70%; and E-value < 10-5, low complexity filter = default DUST approach and word size=11 (default). The resulting BLAST file has been imported into MEGAN (Huson and Weber, 2013) in order to obtain a taxonomic classification of the contigs. A detailed mapping of reads from the Influenza A (H1N1)pdm09 genome was obtained by BLASTing the 454 data or Illumina data with the California 2009A/California/07/2009 genomic sequence (NCBI TaxID: 641809); a similar treatment was employed for reads from the Streptococcus pneumoniae genome (NCBI TaxID: 170187).

MEGAN analysis

The MEGAN analysis has been carried out, for both data sets, on both the assembled contigs and the singletons. The *min-score* filter has been set to 150 and retaining only the *top-percent* filter to 5% to increase accuracy and decrease ambiguity of assignment.

Depth of coverage estimation

The coverage depth was calculated as the total number of bases from all the overlapped reads in the assembly procedure used to generate the reconstructed genome portion, divided over the length of the reconstructed genome. This value was calculated for the individual gene segments for Influenza A (H1N1)pdm09, and for the total reconstructed genome for both Influenza A (H1N1)pdm09 and Streptococcus pneumoniae genomes.

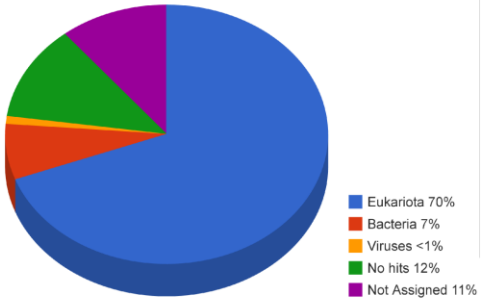
Data availability

Sequencing data were submitted to the Sequence Read Archive (SRA) database, and are publicly available with BioProject identifier SRP060334, BioSample ID SRS978813, and Experiment ID SRX1081190 for the Roche 454 data and SRX1081191 for the Illumina MiSeq data.

Results

As an example of real life clinical samples, a nasopharyngeal swab, randomly selected from samples submitted to the Laboratory for influenza diagnosis during the 2009 pandemic influenza outbreak (Huson and Weber, 2013), was analyzed (influenza H1N1pdm09 M gene load: 3.3x107 cp/ml). For NGS analysis, semi-random RT-PCR was performed on the total nucleic extract from this clinical sample, as previously reported (Allander et al., 2005; Bartolini et al., 2011), then amplification products were split in two aliquots and sequenced with Roche 454 GS FLX+ and Illumina MiSeq platforms, respectively.

Figure 1
A Percentage of Blast hits of 454 reads



B Percentage of Blast hits of Illumina reads

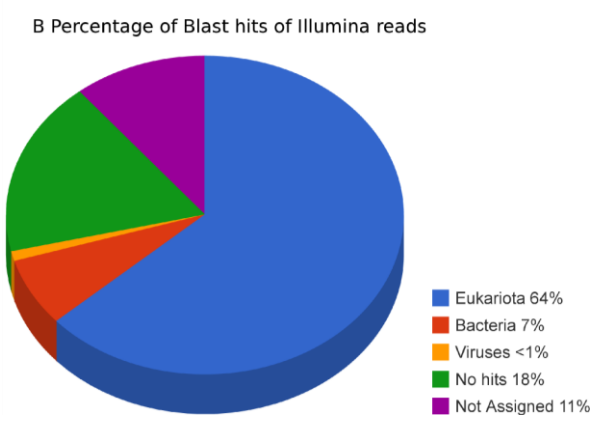


Figure 1: Percentage of BLAST hits

Analysis of Roche 454 sequencing data

Percentage of Blast hits obtained for each taxon using the NCBI nucleotide database sequence classes (BLAST parameters: > 90% identity, > 70% identity, E-value $<10^{-5}$) for 454 data (A) and Illumina data (B).

The 454 FLX+ Titanium instrument generates a number considerably lower than that generated by the Illumina MiSeq instrument, but with the advantage of producing longer reads (up to 800 nucleotides). In our study, after trimming of the PCR primers and filtering of the 454 reads shorter than 50 nt (Chou and Holmes, 2001; Schmieder and Edwards, 2011), the resulting dataset consisted of 453,050 reads ranging in size from 50 to 800 (average length: 340 nt). BLASTn assignment was performed on the selected 454 reads against the NCBI nucleotide database to estimate the microorganism content in the sample. Using stringent cut-offs of >90% identity, >70% overlap and E-value <10-5, 318,654 (70%) reads were classified as eukaryotic, 32,401 (7%) as bacterial, 50,074 (11%) as not assigned (not shown) and 93 (<1%) as viral (Fig. 1a), most of which (83) identified as influenza A (H1N1)pdm09 virus. The alignment of the 83 reads belonging to influenza A (H1N1)pdm09 to its reference genome (California 2009A/California/07/2009; NCBI TaxID: 641809) indicated that the reads are distributed along all the 8 viral genes, covering 65% of the reference genome, although the 8 genes are not equally represented (PB2 87% sequence covered , PB1 67%, PA 72% , HA 72% , NP 73%, NA 65%, M 47%, 42%) as shown in figure S1 and Table 1.

Moreover, the distribution of the reads within each gene segment is not uniform, being larger in the central part than in the ends. This analysis, based on a comparison of the reads with a reference genome, provided a relatively wide percentage of genomic coverage (65%), but it is biased by the possible presence of genomic rearrangements (i.e. inversions, deletions and insertions) respect to the used reference.

An alternative de novo analysis has been also applied, in order to obtain a genome reconstruction independent from any reference genome. This approach eliminates the possible problems due to genomic rearrangements, but may lead to a reduction of genomic coverage in the presence of low quality reads. Read assembly was performed through a de novo assembly of the reads without any host filtering using a two-step strategy based on two Overlap Layout Consensus (OLC) algorithms (Simpson et al., 2009), newbler and CAP3, that has been recently demonstrated to be more efficient than using newbler alone in a de novo assembly study (Biagini et al., 2014). The host (i.e. human) reads were not filtered since this procedure is quite expensive in terms of computational time and provides a limited improvements in terms of quality of the pathogen's reconstructed genome, quantifiable in the order of 0.5-1% (Biagini et al., 2014).

Using an overlap cut-off of 50 bp with 90% identity, the assembly of all the reads using the OLC approach, implemented in the newbler/CAP3 pipeline, produced 10,453 contigs and 77,234 singletons (Table 2).

Table 1. Percentage of reconstructed genes for the influenza obtained by directly mapping the reads over the California influenza A(H1N1)pdm09 genome or by the de novo assembly approach.

Influenza virus A gene	Mapping on reference genome (454)	*De novo* assembly (454)	Mapping on reference genome (Illumina)	*De novo* assembly (Illumina)
PB2	87%	43%	24%	24%
PB1	67%	52%	9%	9%
PA	72%	64%	36%	36%
HA	72%	72%	36%	36%
NP	73%	67%	46%	36%
NA	65%	43%	0%	0
M	47%	35%	0%	0
NS	42%	17%	50%	38%
TOTAL GENOME	65%	52%	25%	23%

Table 2. Assembly results using newbler/CAP3 or Ray Meta on the 454 and Illumina datasets, respectively.fT

	newbler/CAP3 (454)	Ray Meta (Illumina)
Number of contigs	10,457	25,289
Average length of contigs	439 bp	716 bp
Length of smallest contig	100 bp	240 bp
Length of largest contig	4,485 bp	3,316 bp
Number of singletons	77,234	8,116

The alignment of the assembled contigs on the NCBI nucleotide database unambiguously confirmed the presence of the influenza A (H1N1)pdm09 virus. More in detail, 7 contigs and 6 singletons identified the 8 viral genes, with percentage of identity with the reference sequence ranging from 98% to 99%, and a percentage of covered genome of 52% (PB2 43%, PB1 52%, PA 64% , HA 72% , NP 67%, NA 43%, M 35%, 17%) (Table 1).

These results indicate that the de novo assembly covers a lower percentage of genome compared to the reference genome-guided reconstruction of the experimental reads (Table1), however this procedure is more reliable since it is not biased by the reference comparison. In the case here described, only one out of the 7 H1N1 contigs contains a single read not belonging to the virus but to the host genome, classifying the resulting contig as chimeric; this read contains only 1.2% of

the total number of nucleotides identified as influenza virus. The 7 contigs unambiguously identify 5 of the 8 influenza A (H1N1)pdm09 virus genes, namely PB2, PB1, PA, HA, NP, whilst the NA gene is identified by 2 singletons and the NS and M genes are identified by one singleton. The relative coverage depth for each reconstructed gene, reported in Table 4, indicates that the best depth is obtained for the PB2, PB1, NP and HA genes.

In order to verify the ability of the de novo procedure to characterize also the bacterial microorganisms present in the nasopharyngeal swab the MEGAN tool has been used to represent the alignment of the 10,457 contigs and 77,234 singletons (Fig. 2a). The MEGAN taxon tree highlights the presence of several co-infecting bacteria, among which Streptococcus pneumoniae. This bacterium was investigated in detail to assess how a metagenomic approach, initially developed to identify an RNA virus, is able to identify and characterize a pathogen completely different in nature such as a bacterium. The clinical relevance of such comparison relies in the fact that the influenza virus and Streptococcus pneumoniae are two respiratory tract pathogens, responsible for exacerbated disease in co-infected individuals (Smith et al., 2013).

Our assembly produced 6 contigs and 4 singletons that, from the alignment to the Streptococcus pneumoniae genome (NCBI TaxID: 170187), display a sequence identity ranging from 80% to 99%, and a total genome coverage of 0.2% (Table 5). The covered percentage genome is very low, but this was expected since the reads are coming from a procedure focused to the identification of an RNA virus. However, the identification is unambiguous. In detail, the identified contigs (median length of 513.5 bp, range 122-747 bp) are all chimeric, but the content of nucleotides coming from reads assigned to other organisms is lower than 1% of the total number of nucleotides belonging to the Streptococcus pneumoniae, permitting a definite identification of the bacterium, as also supported by the high genome coverage depth (Table 5).

Analysis of Illumina sequencing data

Illumina technology generates shorter sequences than 454, but in a larger number. In particular, MiSeq generates up to several million paired-end reads per instrument-run with a maximum length of 2X250 nt. MiSeq sequencing of our sample produces a dataset formed by 9,640,522 paired reads with an insert-size between 400bp and 1.5kbp.

The full reads dataset has been aligned with BLASTn on the NCBI nucleotide database, using cut-off values identical to those used for the 454 reads analysis. As shown in Fig. 1b, the procedure identifies 5,487,231 reads (64%) as eukaryotic, 876,947 (7%) as bacterial, 54,768 (11%) as not assigned (not shown) and 1257 (<1%) as viral. About 40% of the viral reads matched influenza A(H1N1)pdm09 virus; the alignment of these reads to the influenza California influenza A (H1N1)pdm09 reference genome permitted the identification of 6 of the 8 viral genes. Because of the lack of identification of two genes, the percentage of reference genome covered is only 25% and the 6 genes are not equally covered (PB1 24%, PB2 9%, PA 36%, HA 36%, NP 46% and NS 50%) (Table1) and do not display a preferred covered region, being for instance the 3' and the 5' the preferred covered region for the HA and PA genes, respectively. Moreover, they also display a different depth of coverage (Table 4) (Fig.S1).
As already stated, the direct alignment of the reads may be affected by genomic rearrangements with respect to the genome used as reference, that can be overcome with a de novo assembly. We carried out the assembly of the Illumina reads without any previous host reads filtering, and using a De Brujin graph approach (Miller et al., 2010) as implemented in several tools, namely Ray Meta (Boisvert et al., 2012), Meta-Velvet (Namiki et al., 2012) and Abyss (Simpson et al., 2009), with a k-mer size ranging from 21 to 64bp. The best result in terms of N50, maximum scaffold size, and number of scaffolds was obtained using Ray Meta with a k-mer size of 41, which produced 33,405 contigs. The comparison of the 33,405 classified contigs with the original reads indicated that 25,289 are composed by at least two reads, the remaining 8,116 being singletons (Table 2). Alignment of the assembled contigs on the NCBI nucleotide database unambiguously confirmed the presence of the influenza A(H1N1)pdm09 virus. In particular, 7 contigs identified 6 of the 8 viral genes PB2 (total contig length 699 nt), PB1 (total contig length 220 nt), PA (total contig length 811 nt), HA (total contig length 645 nt), NP (total contig length 556 nt), NS (total contig length 329 nt) with percentages of identity ranging from 99% to 100% and a percentage of covered genome of 23% (PB2 24%, PB1 9%, PA 36%, HA36%, NP 36%, NS 38%) as shown in Table 1. In this case, the coverage values obtained with the reference genome-guided reconstruction, and with the de novo assembly are similar, being 25% and 23% respectively (Table 1). We obtained identical coverage results for PB2, PB1, PA and HA , and a decrease in the gene coverage for the NP (from 46% to 36%) and NS genes (from 50% to 38%) (Table 1). None of the 7 influenza A(H1N1)pdm09 contigs are chimeric (Table 3).

Also in this case we have verified the ability of the de novo procedure to characterize the bacterial microorganisms present in the nasopharyngeal swab by representing the alignment of the obtained 25,289 assembled contigs and 8,116 singletons with the MEGAN tool. The taxonomic distribution of microorganisms is similar to that obtained with the 454 sequencing (Fig. 2b). A total of 102 contigs align with the Streptococcus pneumoniae genome, with sequence identity ranging from 79% to 99%, a percentage of covered genome of 2% and a coverage depth of 26X (Table 5). The total coverage of the Streptococcus pneumoniae genome is still relatively low if compared to the total influenza virus genome coverage, as expected for reads that are coming from a semi-random retro-transcribed PCR (RT-PCR), but it still permits an unambiguous identification.

Table 3. Characteristics of influenza A(H1N1)pdm09 contigs and singletons obtained with the newbler/CAP3 assembler and with the Ray Meta assembler

	newbler/CAP3 (454)	**Ray Meta (Illumina)**
Assembly results	7 Contigs+6 singletons	7 Contigs
Identity	98 to 99%	99% to 100%
Genome covered	52%	23%
Chimeric contigs (number of reads derived from other organism; %)	1 (1; 1.2%)	0
Coverage depth	4.1	41.6
Contig length (median; range)	467.0; 170-1037	387.0; 221-824

Megan taxon tree of the most represented bacterial and viral species identified in the nasopharyngeal swab identified by alignment of 454 reads (A) and Illumina reads (B) on the NCBI nucleotide database. The two investigated pathogens are underlined in red.

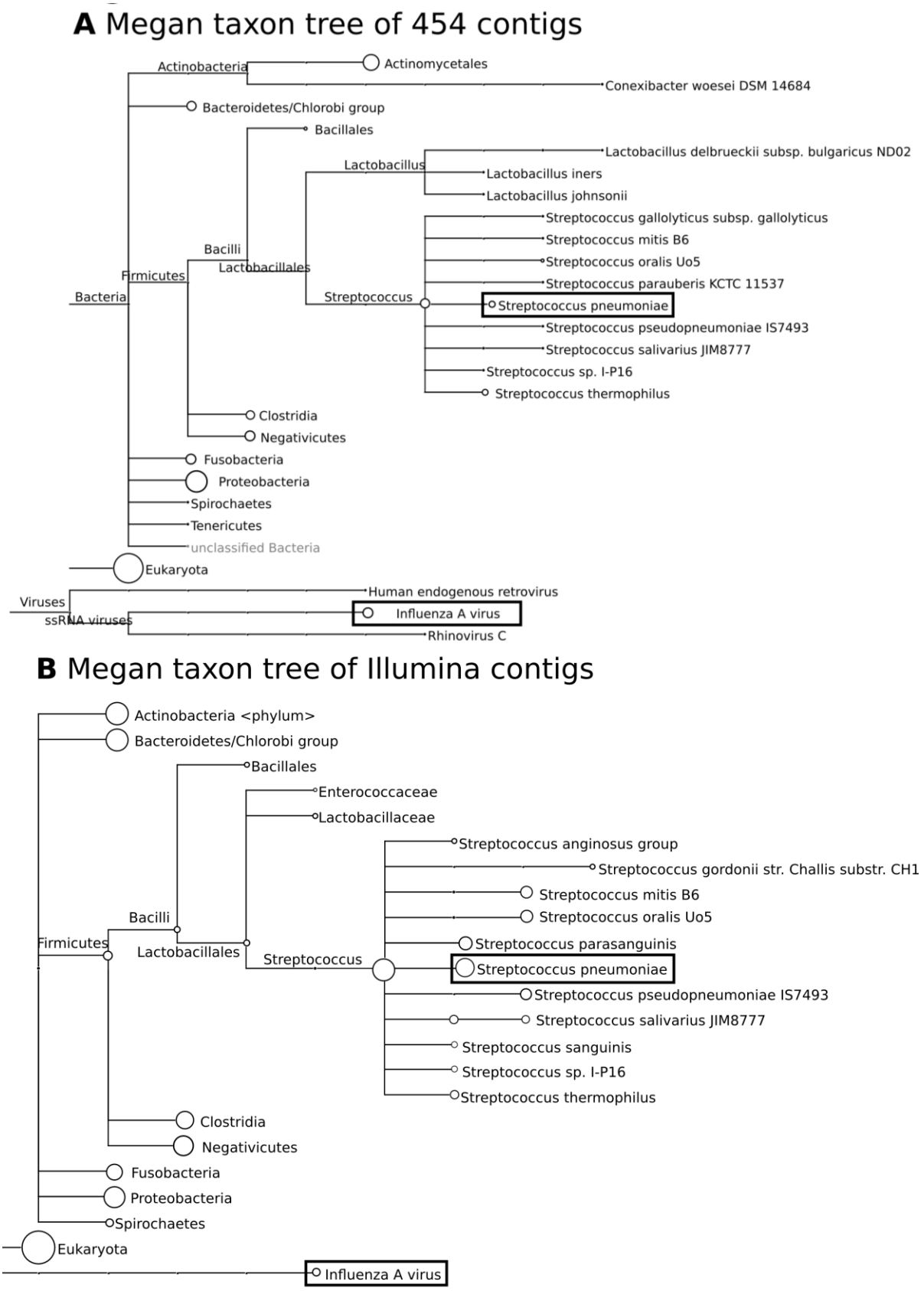


Figure 2: Megan taxon tree

Table 4. Coverage depth for the influenza A(H1N1)pdm09 genes obtained with the newbler/CAP3 and Ray Meta assemblers

GENE	454	Illumina
PB2	6	5.2
PB1	6.8	2
PA	2	20.1
HA	4.5	151.3
NP	4.8	4.3
NA	1.3	NA*
M	1	NA
NS	1	5.2

*NA: not applicable, no contigs

Discussion

In this study, the performance of the Roche 454 GS FLX+ and Illumina MiSeq platforms for metagenomic studies have been directly compared in their ability of identifying the microorganisms present in a clinical sample consisting of a nasopharyngeal swab from a patient diagnosed for influenza A(H1N1)pdm09 infection. The amplification products were generated with a sequence-independent approach and analyzed in parallel with both NGS platforms using two different strategies: (i) direct reads comparison to a reference genome; (ii) a de novo, reference genome-independent assembly. The results showed that the H1N1 virus is unambiguously identified by both platforms and by both analyses, which also described an almost identical overall microbial community in the clinical sample. Although these results could not be generalized, since they are based on a single comparative experiment, they provide some interesting information. Namely, the results indicate that even using a metagenomic RNA-based approach aimed at the identification of RNA viruses. The ability of both platforms to identify influenza in clinical samples have been reported (Bialasiewicz et al., 2014; Nakajima et al., 2010) as well as a transcriptome analysis of bacterial communities in clinical samples (Giannoukos et al., 2012). In this work a detailed analysis has been conducted on two organisms of clinical relevance: the influenza A(H1N1)pdm09 (an RNA virus) and Streptococcus pneumoniae (a bacterium), showing some interesting platform-specific differences.

The influenza virus analysis conducted with the 454 platform was able to identify the 8 viral genes, whilst in the case of the Illumina sequencing, only 6 of the 8 viral genes were detected. Mapping the reads coming from 454 platform on the California influenza A(H1N1)pdm09 permitted to cover 65% of the viral genome, while the de novo assembly procedure reconstructed a smaller fraction (52%) (Table 1). The analysis based on the Illumina platform covered 25% of the reference genome, and 23% was covered with the de novo assembled contigs (Table1). The relatively larger decrease in coverage observed for the 454 platform data, going from the reference-guided mapping to de novo assembly, compared to that resulting from the Illumina reads, suggests that in this case the reads obtained from the 454 sequencing are likely of lower quality with respect to the ones coming from Illumina. The coverage, and the coverage depth, for each gene are not uniform and do not seem to follow a specific trend. Coverage and depth are also not correlated with the percentage of G/C content, that was considered relevant in some previous studies (Chen et al., 2013) (Fig. S1). The reconstruction extent of individual gene segments obtained with the 454 sequencing in the present study is also in agreement with previous results from our group (Bartolini et al., 2011; Biagini et al., 2014).

For both platforms, the de novo assembly has been carried out without any filtering of the reads from the host genome, to reduce the required computational time. This procedure could be thought to introduce biases in the reconstructed contigs, however our results showed that the contigs almost fully overlap with the influenza A(H1N1)pdm09 and only one chimeric contig was found. The described procedure is fast and efficient, which can be of utmost importance in a clinical setting. The highest percentage of overall covered genome was reached with the 454 sequencing, whilst the coverage depth is generally higher for Illumina data (Table 4).

These differences may be important in selecting a sequencing platform for specific aims since, for instance, a greater coverage depth increases the confidence in the species and strain identification. Nevertheless, coverage depth is not uniform, being for example the PA and HA genes the most represented ones in the Illumina data, while PB2 and PB1 resulted the most abundant in the 454 reads (Table 4). It can be noticed that in our experiment, the total number of sequenced bases is 1,810 Mbp and 154 Mbp for Illumina and 454, respectively, and that the number of bases assigned to H1N1 is 111 kbp and 36 kbp for Illumina and 454, respectively. Therefore, although the ratio of total sequenced bases is more than one order of magnitude in favor of Illumina, the ratio in the H1N1 virus bases is only three, suggesting that the longer length of 454 sequences partially compensates its reduced coverage.

In Table 5 we report the results of the genome reconstruction of Streptococcus pneumoniae. Also for Streptococcus pneumonia in this case both technologies were able to detect the pathogen, but with some differences. The Illumina technology produced a percentage of genome reconstruction larger than that obtained with the 454 (2% vs 0.2%), in contrast to what observed for the influenza virus.

T**able 5.** Characteristics of the *Streptococcus pneumoniae* contigs obtained with the newbler/CAP3 and Ray Meta assemblers.

	newbler/CAP3 (454)	Ray Meta (Illumina)
Assembly results	7Contigs+6 singletons	7 Contigs
Identity	80 to 99% (average 93.8%)	99% to 100% (average 93.1%)
Genome covered	0.2%	2%
Chimeric contigs (number of reads derived from other organism; %)	6 (7; <1%)	90 (900; 6.2%)
Coverage depth	55.8	26.2
Contig length (median; range)	513.5; 122-747	623.0; 163-1556

Some parts of the reconstructed genome have low identity, but the values of coverage depth are good for both technologies, and the weight of ectopic reads in contigs is low.

Conclusions

In this work we show that both 454 and Illumina technologies are efficient for the detection of pathogens in the analyzed clinical sample, allowing the correct identification of the virus, consistent with the diagnostic results obtained with a conventional real time PCR approach. In addition, a significant proportion of the viral genome can be reconstructed with both methods, the 454 platform providing a wider and the Illumina platform a deeper coverage. Streptococcus pneumoniae, a microorganism frequently contributing to the disease

pathogenesis, was clearly identified with both platforms, even though the bacterial genome length and deep coverage are quite lower than that of the viral pathogen. This result is due, on one hand, to the large difference in the dimension of the two genomes, and to the fact that the analysis was conducted on retrotrascribed RNA, that favors the amplification of the viral genome. However this procedure, which starts from the extracted RNA from clinical samples, can be considered the most reliable for a metagenomic approach, since it permits to identify both DNA and RNA pathogens, as it was here demonstrated.

Our data indicate that the application of NGS to the metagenomic analysis of clinical samples, in particular those obtained from patients with unknown infectious diseases, is feasible, not only with respect to the pathogen identification, but also for the genetic characterization of viral and bacterial pathogens without the need of culture amplification. We also showed that the performances are not identical for the two platforms: differences include original length of the reads, length of contigs, depth of coverage and extension of the pathogen genome reconstruction. These differences may be exploited to obtain the most suitable information according to the diagnostic needs. For instance, considering the viral pathogen in our analysis, the high number of influenza virus-specific reads obtained with the Illumina platform can have an important influence on the confidence on pathogen identification and on the quality of the reconstructed genomic regions, due to the high coverage depth, which is, however, high only in limited regions. On the other hand, 454 sequences are characterized by a relatively low coverage depth but are more uniformly distributed along all the genomic regions, thus covering a greater portion of the viral genome. Concerning the selected bacterial pathogen, the opposite scenario occurred, i.e. the greater number of Illumina reads allowed a more extensive reconstruction. Also in this case, the level of coverage depth is sufficient for the unambiguous identification of the pathogen with both technologies. More extended studies are necessary to fully highlight the differences in the performance of the available NGS platforms in metagenomic studies and to identify performance gaps and advantages to be exploited for their application in the clinical diagnostic setting.

References

Allander, T., Tammi, M. T., Eriksson, M., Bjerkner, A., Tiveljung-Lindell, A., and Andersson, B. (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proc. Natl. Acad. Sci. U. S. A. 102, 12891–12896. doi:10.1073/pnas.0504666102.

Bartolini, B., Chillemi, G., Abbate, I., Bruselles, A., Rozera, G., Castrignanò, T., et al. (2011). Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing. New Microbiol. 34, 391–397.

Barzon, L., Lavezzo, E., Costanzi, G., Franchin, E., Toppo, S., and Palù, G. (2013). Next-generation sequencing technologies in diagnostic virology. J. Clin. Virol. 58, 346–350. doi:10.1016/j.jcv.2013.03.003.

Biagini, T., Bartolini, B., Giombini, E., Capobianchi, M. R., Ferrè, F., Chillemi, G., et al. (2014). Performances of Bioinformatics Pipelines for the Identification of Pathogens in Clinical Samples with the De Novo Assembly Approaches: Focus on 2009 Pandemic Influenza A (H1N1). Open Bioinforma. J. 8. Available at: http://benthamopen.com/ABSTRACT/TOBIOIJ-8-1 [Accessed September 17, 2015].

Bialasiewicz, S., McVernon, J., Nolan, T., Lambert, S. B., Zhao, G., Wang, D., et al. (2014). Detection of a divergent Parainfluenza 4 virus in an adult patient with influenza like illness using next-generation sequencing. BMC Infect. Dis. 14, 275. doi:10.1186/1471-2334-14-275.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13, R122. doi:10.1186/gb-2012-13-12-r122.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Capobianchi, M. R., Giombini, E., and Rozera, G. (2013). Next-generation sequencing technology in clinical virology. Clin. Microbiol. Infect. 19, 15–22. doi:10.1111/1469-0691.12056.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., and Hwang, C. C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS One 8, e62856. doi:10.1371/journal.pone.0062856.

Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., et al. (2011). Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J. Clin. Microbiol. 49, 3268–3275. doi:10.1128/JCM.00850-11.

Chou, H. H., and Holmes, M. H. (2001). DNA sequence quality trimming and vector removal. Bioinformatics 17, 1093–1104. doi:10.1093/bioinformatics/17.12.1093.

Dawood, F. S., Jain, S., Finelli, L., Shaw, M. W., Lindstrom, S., Garten, R. J., et al. (2009). Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. N. Engl. J. Med. 360, 2605–2615. doi:10.1056/NEJMoa1404595.

Frey, K. G., Herrera-Galeano, J. E., Redden, C. L., Luu, T. V, Servetas, S. L., Mateczun, A. J., et al. (2014). Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. BMC Genomics 15, 96. doi:10.1186/1471-2164-15-96.

Gardner, S. N., Kuczmarski, T. a., Vitalis, E. a., and Slezak, T. R. (2003). Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. J. Clin. Microbiol. 41, 2417–2427. doi:10.1128/JCM.41.6.2417-2427.2003.

Giannoukos, G., Ciulla, D. M., Huang, K., Haas, B. J., Izard, J., Levin, J. Z., et al. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. 13, R23. doi:10.1186/gb-2012-13-3-r23.

Greninger, A. L., Chen, E. C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., et al. (2010). A metagenomic analysis of pandemic influenza a (2009 H1N1) infection in patients from North America. PLoS One 5, e13381. doi:10.1371/journal.pone.0013381.

Howe, A. C., Jansson, J. K., Malfatti, S. a, Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. Proc. Natl. Acad. Sci. U. S. A. 111, 4904–9. doi:10.1073/pnas.1402564111.

Huang, X., and Madan, a (1999). CAP 3: A DNA sequence assembly program. Genome Res. 9, 868–877. doi:10.1101/gr.9.9.868.

Huson, D. H., and Weber, N. (2013). "Microbial community analysis using MEGAN.," in Methods Enzymol (Elsevier), 465–485. doi:10.1016/B978-0-12-407863-5.00021-6.

Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., and Wei, C. (2013).

NeSSM: A Next-Generation Sequencing Simulator for Metagenomics. PLoS One 8, e75448. doi:10.1371/journal.pone.0075448.

Lemmon, G. H., and Gardner, S. N. (2008). Predicting the sensitivity and specificity of published real-time PCR assays. Ann. Clin. Microbiol. Antimicrob. 7, 18. doi:10.1186/1476-0711-7-18.

Lu, J., Wu, J., Zeng, X., Guan, D., Zou, L., Yi, L., et al. (2014). Continuing Reassortment Leads to the Genetic Diversity of Influenza Virus H7N9 in Guangdong, China. J. Virol. 88, 8297–8306. doi:10.1128/JVI.00630-14.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7, e30087. doi:10.1371/journal.pone.0030087.

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327. doi:10.1016/j.ygeno.2010.03.001.

Mitra, S., Schubach, M., and Huson, D. H. (2010). Short clones or long clones? A simulation study on the use of paired reads in metagenomics. BMC Bioinformatics 11 Suppl 1, S12. doi:10.1186/1471-2105-11-S1-S12.

Moore, R. a., Warren, R. L., Freeman, J. D., Gustavsen, J. a., Chénard, C., Friedman, J. M., et al. (2011). The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PLoS One 6, e19838. doi:10.1371/journal.pone.0019838.

Nakajima, N., Hata, S., Sato, Y., Tobiume, M., Katano, H., Kaneko, K., et al. (2010). The first autopsy case of pandemic influenza (A/H1N1pdm) virus infection in Japan: Detection of a high copy number of the virus in type II alveolar epithelial cells by pathological and virological examination. Jpn. J. Infect. Dis. 63, 67–71. doi:10.1371/journal.pone.0010256.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 40, e155–e155. doi:10.1093/nar/gks678.

Nisii, C., Meschi, S., Selleri, M., Bordi, L., Castilletti, C., Valli, M. B., et al. (2010). Frequency of detection of upper respiratory tract viruses in patients tested for pandemic H1N1/09 viral infection. J. Clin. Microbiol. 48, 3383–3385. doi:10.1128/JCM.01179-10.

Palacios, G., Lowenstine, L. J., Cranfield, M. R., Gilardi, K. V. K., Lukasik-braum, M., Kinani, J., et al. (2011). Metapneumovirus Infection in Wild Mountain Gorillas ,. Emerg. Infect. Dis. 17, 711–714. doi:10.3201/eid1704100883.

Petty, T. J., Cordey, S., Padioleau, I., Docquier, M., Turin, L., Preynat-Seauve, O., et al. (2014). Comprehensive human virus screening using high-throughput sequencing with a user-friendly representation of bioinformatics analysis: a pilot study. J. Clin. Microbiol. 52, JCM.01389–14–. doi:10.1128/JCM.01389-14.

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., et al. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods 6, 639–641. doi:10.1038/nmeth.1361.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. doi:10.1093/bioinformatics/btr026.

Sergeant, M. J., Constantinidou, C., Cogan, T. a, Bedford, M. R., Penn, C. W., and Pallen, M. J. (2014). Extensive microbial and functional diversity within the chicken cecal microbiome. PLoS One 9, e91941. doi:10.1371/journal.pone.0091941.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. Genome Res. 19, 1117–1123. doi:10.1101/gr.089532.108.

Smith, A. M., Adler, F. R., Ribeiro, R. M., Gutenkunst, R. N., McAuley, J. L., McCullers, J. A., et al. (2013). Kinetics of Coinfection with Influenza A Virus and Streptococcus pneumoniae. PLoS Pathog. 9, e1003238. doi:10.1371/journal.ppat.1003238.

Towner, J. S., Sealy, T. K., Khristova, M. L., Albariño, C. G., Conlan, S., Reeder, S. a., et al. (2008). Newly discovered Ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog. 4, e1000212. doi:10.1371/journal.ppat.1000212.

Ugalde, J. a., Gallardo, M. J., Belmar, C., Muñoz, P., Ruiz-Tagle, N., Ferrada-Fuentes, S., et al. (2013). Microbial Life in a Fjord: Metagenomic Analysis of a Microbial Mat in Chilean Patagonia. PLoS One 8, e71952. doi:10.1371/journal.pone.0071952.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., et al. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach