Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing

Barbara Bartolini[1], Giovanni Chillemi[2], Isabella Abbate[1], Alessandro Bruselles[1], Gabriella Rozera[1], Tiziana Castrignanò[2], Daniele Paoletti[2], Ernesto Picardi[4], Alessandro Desideri[2,3], Graziano Pesole[4,5], Maria R. Capobianchi[1]

[1]National Institute for Infectious Diseases (INMI) L. Spallanzani, Rome, Italy;
[2]CASPUR Rome, Italy;
[3]Dipartimento di Biologia, University of Rome "Tor Vergata", Rome, Italy;
[4]Dipartimento di Biochimica e Biologia Molecolare "E. Quagliariello", University of Bari, Italy;
[5]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

SUMMARY

De novo high-throughput pyrosequencing was used to detect and characterize 2009 pandemic influenza A (H1N1) virus directly in nasopharyngeal swabs in the context of the microbial community. Data were generated with a prior sequence-independent amplification by 454 pyrosequencing on GS-FLX platform (Roche). Influenza A assembled reads allowed near full-length genome reconstruction with the simultaneous analysis of site-specific heterogeneity. The molecular approach applied proved to be a powerful tool to characterize the new pandemic H1N1 influenza virus in clinical samples. This approach could be of great value in identifying possibly new reassortants that may occur in the near future.

KEY WORDS: Influenza, High-throughput pyrosequencing, Nasopharyngeal swabs

Received March 26, 2011 *Accepted* May 17, 2011

High-throughput sequencing technologies have proved to be a powerful tool for in-depth analysis of viral quasispecies in HIV and HBV infected patients (Bruselles *et al.*, 2009; Rozera *et al.*, 2009; Solmone *et al.*, 2009). More in general, massive parallel pyrosequencing seems to represent a useful methodology for direct study of the metagenome i.e. all the nucleic acid sequences belonging to the entire microbial community present in a human body district, also in view of possible identification of new pathogens (Petrosino *et al.*, 2009).

The present study applied DNA shotgun high-throughput pyrosequencing to nasopharyngeal fluids from patients infected with H1N1 pandemic influenza A, to analyze influenza genome heterogeneity, without *in vitro* replication biases, in the context of the microbial community present in the clinical samples.

Nasopharyngeal swabs collected during an early pandemic event from 2 symptomatic patients positive for 2009 pandemic influenza A H1N1 were analyzed. Data were generated by 454 Life technology with a prior semi-random RT-PCR as previously reported (Allander *et al.*, 2005).

The amplicons purified and quantified were subsequently linked to adapter molecules to allow library preparation and subsequent pyrosequencing steps following the manufacturer's instructions (Roche, Titanium version, GS FLX platform). Bioinformatic analysis was performed to categorize each read obtained through a BLAST search against a predefined series of nucleotide and protein databases at NCBI. A suitably devised assembling workflow was used to reconstruct the

Corresponding author
Maria Rosaria Capobianchi
Laboratory of Virology
National Institute for Infectious
Diseases "L. Spallanzani"
Via Portuense, 292 - 00149 Rome, Italy
E-mail: maria.capobianchi@inmi.it

entire influenza genome (see below). A clonal plasmid encompassing a region starting from 589 nt to 908 nt of 4 segment (NCBI TaxID: 641809, GenBank: FJ7966974) was used to assess the pyrosequencing error rate with respect to standard Sanger sequencing. In the studied region, considering Sanger sequencing true, the error rate of 454 technology for mismacthes was 0.035%, 0.069% for insertions and 0.032% for deletions.

A total of 502,056 reads (average length 233 bp) were obtained from the two nasopharyngeal samples analyzed.

Using stringent cut offs of >90% identity, >70% overlapping and E-value <10^{-5}, BLASTN analysis classified 206,469 reads (41%) as human, 76,794 (15 %) as bacteria, 1415 (0.3 %) as eukaryotes, 488 (0.1 %) as viruses other than influenza A. Furthermore, a total of 80,596 reads (16%) were identified as influenza A sequences. For a complete list of all the mapped reads Table 1. In both samples, >99.5% of the influenza A reads show the highest sequence similarity with the 2009 pandemic influenza A (H1N1). All such reads were used to reconstruct a full genome using a new computational strategy implemented in ad hoc custom scripts written in python language. Blast mapping results against the entire reference genome of A/California/07/2009 (H1N1) (NCBI TaxID: 641809) were parsed to detect all supporting reads for each reference position. Next, we called the consensus nucleotide sequence the most frequent base at each position after a quality score filtering.

In spite of the different total number of influenza sequences, we observed in both samples a remarkable heterogeneity in the representation of the eight influenza A genome segments, with segments 1 (PB2) and 2 (PB1) largely over-represented, with about 27% and 58% of all the influenza sequences, respectively (Figure 1), while a significantly lower number of reads, ranging from 0.3 to 8 %, mapped on the remaining six segments (Figure 1).

In patient 1, displaying a high viral load (2x10^7 cp/ml) of influenza A (H1N1), the sequences obtained were numerous enough to allow the reconstruction of a nearly complete H1N1 genome (95% of the genome assembled) (Figure 2). In sample 2, with 7x10^5 cp/ml viral load, the reconstructed genome fragments accounted for about 30% of the entire influenza virus genome (not shown).

TABLE 1 - *Number of BLAST hits obtained against different sequence classes (BLAST parameters: >90% identity, >70% overlapping, E-value<10-5). The accession numbers for human mtDNA and rRNAs are: AC_000021, NR_003287, NR_003286, NR_003285, NR_023363. Human ncRNAs and mRNAs have been extracted from fRNAdb (Kin et al., 2007) and RefSeq (Pruitt et al., 2007), respectively.*

BLAST hits	*Patient 1*	*Patient 2*	*Total*
human mtDNA	60428	7598	68026
human rRNAs (28SrRNA, 18SrRNA, 5.8SrRNA, 5S5rRNA)	13837	19979	33816
human ncRNAs	64	188	252
human mRNAs	4188	4240	8428
human genome (NCBI36/hg18) and other human sequences	48320	47627	95947
Influenza A (H1N1)	80417	179	80596
Other viruses	336	152	488
Bacteria	85	76709	76794
Eukaryotes	998	417	1415
Unmapped reads	34928	101366	136294
Total	243601	258455	502056

References: Kin T., Yamada K., Terai G., Okida H., Yoshinari Y., Ono Y., Kojima A., Kimura Y., Komori T., Asai K. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucleic Acids Res. 35, D145-8.

Pruitt K.D., Tatusova T., Maglott D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins Nucleic Acids Res. 35, D61-5.

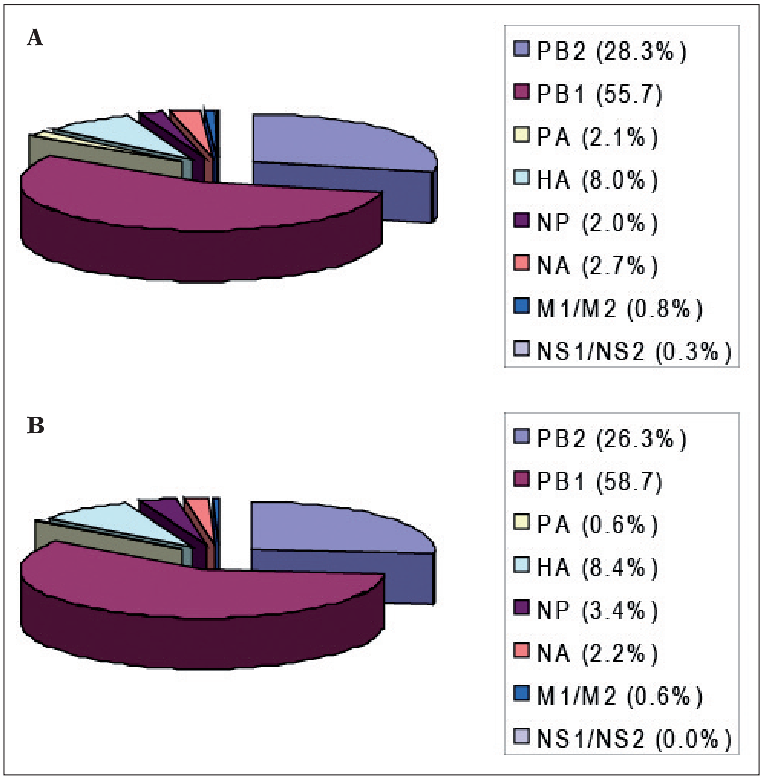

FIGURE 1 - *% of reads mapped on the reference genome of influenza A (A/California/07/2009 (H1N1), NCBI TaxID: 641809), for the nasopharyngeal swab of patient 1 on a total of 80417 of influenza A reads (A) and the nasopharyngeal swab of patient 2 on a total of 179 of influenza A reads (B).*

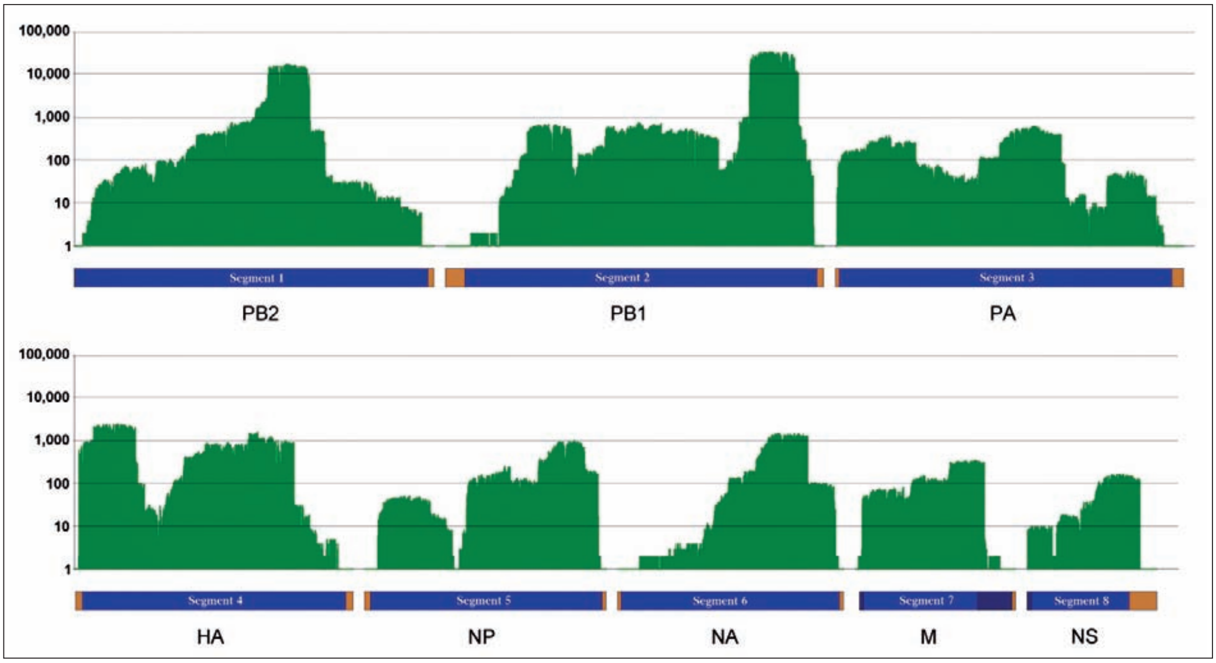

FIGURE 2 - *Near full genome characterization of 2009 pandemic influenza A (H1N1) from the nasopharyngeal swab of patient 1. Coverage depth (in logarithmic scale) across the eight fragments of the viral genome. The assembled contigs have been obtained by using as reference genome the influenza A/California/7/2009 (H1N1) (see Supplementary Table 2). Only positions with ≥Q20 quality and coverage ≥ 20 have been considered (in blue bar), whereas other positions are indicated with orange bar.*

TABLE 2 - *Nucleotide changes observed in the consensus sequence from patient 1 with respect to the reference genome (A/California/07/2009 (H1N1), NCBI TaxID: 641809).*

Segment	*Gene*	*Reference position*	*Reference nt*	*Consensus nt*	*Coverage*	*Consensus support*	*AA change*
1	PB2	1223	G	A	2254	2253	-
1	PB2	1441	G	A	10109	10093	R->K
1	PB2	1582	A	G	490	490	K->R
1	PB2	1877	C	T	30	30	-
2	PB1	426	A	G	23	23	-
2	PB1	532	A	T	448	329	N->Y
2	PB1	587	G	A	592	564	G->D
2	PB1	1184	A	G	371	357	K->R
3	PA	32	A	G	93	93	G->D
3	PA	694	C	T	49	49	-
3	PA	1208	G	A	552	552	S->N
3	PA	1318	G	A	511	511	-
3	PA	2010	T	G	16	16	-
4	HA	190	C	A	2279	2279	-
4	HA	210	C	T	2147	2143	-
4	HA	343	C	T	2236	2236	-
4	HA	366	T	C	2100	2096	-
4	HA	375	A	G	1379	1375	-
4	HA	936	A	G	833	833	-
4	HA	1057	A	G	785	785	I->V
4	HA	1250	A	T	1187	1182	K->M
5	NP	171	C	T	43	43	-
5	NP	404	T	A	39	39	L->Q
5	NP	1174	C	T	366	366	-
6	NA	1067	G	A	1082	1082	-

The near full-length genome from patient 1 (GenBank accession numbers of the 8 segments: from HM992562 to HM992569) showed 25 nucleotide changes (10 non-synonymous) with respect to the reference genome (Table 2).

The intra-host nucleotide variability against the autologous consensus sequence for each coding segment is reported in Table 3 (patient 1). The genomic region showing the highest variability was segment 4, encoding for hemagglutinin (HA, mean frequency of variation/site: 67 x 10^{-4}). The genomic region with the lowest heterogeneity was segment 8 (mean frequency of variation/site: 6.7 x 10^{-4}), encoding for non structural protein 1 (NS1) and nuclear export protein (NEP). The ratio of non synonymous vs synonymous substitutions was <1 for all gene products, with the exception of NEP (1.0) indicating some positive force for selection, according to previous findings (Valli *et al.,* 2010).

Overall, amino acid substitutions were found to reach a frequency higher than 5% in a few sites (Table 4), not involving known Influenza A patho-

TABLE 3 - *Nucleotide variations/site with respect to the consensus sequence measured for all gene products (patient 1).*

Segment	*Product*	*Variation/site* ($\cdot 10^{-4}$)	*N/S ratio*
1	PB2	35.9	0.35
2	PB1	9.5	0.35
3	PA	56.8	0.38
4	HA	67.0	0.41
5	NP	24.5	0.37
6	NA	15.4	0.36
7	M1	10.2	0.35
7	M2	14.5	0.62
8	NS1	10.4	0.58
8	NEP	3. 1	1.00

TABLE 4 - *List of aminoacid positions showing substitutions with respect to the consensus sequence of patient 1, detected at frequency higher that 5% overall.*

Protein	*Position (AA)*	*Mutation*	*%*
PB2	383	Q→L,P,H	7.0
	384	L→S	9.8
	504	V→L,A,G	10.1
PB1	75	E→G	16.7
	136	Y→N	26.5
	225	N→Y,I	5.6
	229	K→Q,I,T	18.0
	246	M→V	25.0
	305	D→G	7.0
	593	D→G	7.3
	614	E→G,V,A,D	5.0
PA	110	Y→H,C,S	18.0
	402	S→Y	6.8
HA	182	Y→C	10.2
NP		None >5%	
NA	401	G→R, stop	9.5
M1	7	V→G	6.5
	197	E→G	8.0
M2	7	V→G	6.5
NS1	66	E→G	20.0
	122	A→V	6.45
NEP		None >5%	

genetic determinants, including those related to binding to sialic acid receptors (Bautista *et al.*, 2010; Hale *et al.*, 2010; Kuroda *et al.*, 2010). In conclusion, high-throughput sequencing may be a powerful tool to identify a novel pathogen in clinical samples, even without prior *in vitro* propagation, since the application of random PCR strategies directly on clinical samples permits to obtain enough nucleic acid material to undergo whole genome sequencing.

More importantly, no advance genetic information is needed at variance with microarray technologies (Berthet *et al.*, 2010; Dawood *et al.*, 2009). In addition, since high throughput sequencing may provide as output the whole genome sequences, especially when a high viral load is present in the clinical sample, this makes possible to perform phylogenetic analysis of the novel agent discovered, showing the relationships between the viral variants present in the outbreak with known ancestor viruses.

Due to the small number of samples analyzed, the present data cannot yield precise information on the extent of viral load required to detect virus-specific sequences in a given clinical sample. However, considering the yield of influenza-specific sequences obtained, it is reasonable that, in the present experimental conditions, with an average yield of 250,000 reads per sample, influenza-specific sequences may be detected down to 10^3 cp/ml.

It should also be emphasized that shotgun high-throughput sequencing, simultaneously studying the microbial community present in a clinical sample, may allow the co-detection of different pathogens that may be responsible for co-infections (Koon *et al.*, 2010). Our results are in line with a recent paper (Kuroda *et al.*, 2010) with some important improvements. The coverage obtained in our study enabled us to perform a detailed analysis of viral heterogeneity throughout the viral genome.

In patient 1 a tenfold difference between the least variable (segment 8, encoding for NS1 and NEP proteins, mean frequency of variation/site: 6.7 x 10^{-4}) and the most variable (segment 4 encoding for HA protein, mean frequency of variation/site: 67 x 10^{-4}) gene segments was observed. Amino acid changes were identified in almost all genome fragments, with few substitutions present at a frequency higher than 5%. By contrast, in the study by Kuroda *et al.* only major components of the viral quasispecies were identified in the HA gene.

The sensitivity in detecting specific mutations depends on both their absolute concentration and on the coverage obtained in a given experiment. However, it is not possible to increase the sensitivity of the assay indefinitely because of the threshold of the intrinsic error rate of pyrosequencing (for Titanium version on the average 0.2%) (Abbate *et al.*, 2010) without introducing correction algorithms that depend on the specific genome region of interest.

Taking into account all these considerations, an intriguing aspect of our findings is the uneven representation of influenza virus nucleic acid segments in clinical samples with the PB1 fragment being always the most represented. Even though the present data do not discriminate between cRNA, vRNA or mRNA origin of the sequences (Kuroda *et al.,* 2010), the evidence of uncoordinated levels of representation of virus-specific nucleic acid segments in clinical material, supported by sequence analysis of the amplicons, is clearcut, and may have pathogenetic significance. Interestingly, the present study established site-specific viral heterogeneity directly *ex vivo*, without the introduction of *in vitro* replication bias (Ramakrishnan *et al.,* 2009). Possible mutations that may confer to the virus the ability to escape both host immunity and antiviral drugs may be highlighted without the selection bias determined by the *in vitro* isolation procedures. This possibility appeared very promising in a recent study addressing the intra-host variability of HA and NS by the classical cloning approach to samples from birds infected with avian influenza viruses (Iqbal *et al.,* 2009). Unbiased high-throughput sequencing may help in the future to identify possible new reassortants. It is reasonable to think that the rapidly evolving techniques of next-generation sequencing, providing more affordable, higher throughput sequence-based data, will become the gold standard approach in the near future for the discovery of new microbial pathogens and for the differential diagnosis of unknown infectious diseases (Harris *et al.,* 2009).

ACKNOWLEDGMENTS

This study was supported in part by grants from the Italian Ministry of Health (Ricerca corrente e finalizzata), by Fondo per gli Investimenti della Ricerca di Base (FIRB) 2003 "Laboratorio Italiano di Bioinformatica", by "Laboratorio di Bioinformatica per la Biodiversità Molecolare" (Progetto FAR, DM19410). The authors acknowledge CASPUR HPC facility for the resource time granted.

REFERENCES

Abbate I., Rozera G., Tommasi C., Bruselles A., Bartolini B., Chillemi G., Nicastri E., Narciso P., Ippolito G., Capobianchi M.R. (2010). Analysis of co-receptor usage of circulating viral and proviral HIV genome quasispecies by ultra-deep pyrosequencing in patients who are candidates for CCR5 antagonist treatment. *Clin. Microbiol. Infect.* doi: 10.1111/j.1469-0691.2010.03350.x.

Allander T., Tammi M.T., Eriksson M., Bjerkner A., Tiveljung-Lindell A., Andersson B. (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples PNAS. **102**, 12891-12896

Bautista E., Chotpitayasunondh T., Gao Z., Harper S.A., Shaw M., Uyeki T.M., Zaki S.R., Hayden F.G., Hui D.S., Kettner J.D., Kumar A., Lim M., Shindo N., Penn C., Nicholson K.G. (2010). Clinical aspects of pandemic 2009 influenza A (H1N1) virus infection. *N. Engl. J. Med.* **362**, 1708-1719.

Berthet N., Leclercq I., Dublineau A., Shigematsu S., Burguière A.M., Filippone C., Gessain A., Manuguerra J.C. (2010). High-density resequencing DNA microarrays in public health emergencies. *Nat. Biotechnol.* **28**, 25-27.

Bruselles A., Rozera G., Bartolini B., Prosperi M., Del Nonno F., Narciso P., Capobianchi M.R., Abbate I. (2009). Use of massive parallel pyrosequencing for near full-length characterization of a unique HIV Type 1 BF recombinant associated with a fatal primary infection. *AIDS Res. Hum. Retroviruses.* **25**, 937-942.

Dawood F.S., Jain S., Finelli L., Shaw M.W., Lindstrom S., Garten R.J., Gubareva L.V., Xu X., Bridges C.B., Uyeki T.M. (2009). Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N. Engl. J. Med.* **360**, 2605-2615.

Hale B.G., Steel J., Manicassamy B., Medina R.A., Ye J., Hickman D., Lowen A.C., Perez D.R., García-Sastre A. (2010). Mutations in the NS1 C-terminal tail do not enhance replication or virulence of the 2009 pandemic H1N1 influenza A virus. *J. Gen. Virol.* **91**, 1737-1742.

Harris T.D., Buzby P.R., Babcock H., Beer E., Bowers J., Braslavsky I., Causey M., Colonell J., Dimeo J., Efcavitch J.W., Giladi E., Gill J., Healy J., Jarosz M., Lapen D., Moulton K., Quake S.R., Steinmann K., Thayer E., Tyurina A., Ward R., Weiss H., Xie Z. (2008) Single-molecule DNA sequencing of a viral genome. *Science.* **320**, 106-109.

Iqbal M., Xiao H., Baillie G., Warry A., Essen S.C., Londt B., Brookes S.M., Brown I.H., Mccauley J.W. (2009). Within-host variation of avian influenza viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**, 2739-2747.

Koon K., Sanders C.M., Green J., Malone L., White H., Zayas D., Miller R., Lu S., Han J. (2010). Co-detection of pandemic (H1N1) 2009 virus and other respiratory pathogens. *Emerging Infectious Diseases.* **16**, 1976-1978.

Kuroda M., Katano H., Nakajima N., Tobiume M., Ainai A., Sekizuka T., Hasegawa H., Tashiro M., Sasaki Y., Arakawa Y., Hata S., Watanabe M., Sata T. (2010). Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. *PLoS One.* **5**, e10256.

Petrosino J.F., Highlander S., Luna R.A., Gibbs R.A., And Versalovic J. (2009). Metagenomic Pyrosequencing and Microbial Identification. *Clinical Chemistry*. **55**, 856-66.

Ramakrishnan M.A., Tu Z.J., Singh S., Chockalingam A.K., Gramer M.R., Wang P., Goyal S.M., Yang M., Halvorson D.A., Sreevatsan S. (2009). The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. *PLoS One*. **4**, e7105.

Rozera G., Abbate I., Bruselles A., Vlassi C., D'offizi G., Narciso P., Chillemi G., Prosperi M., Ippolito G., Capobianchi M.R. (2009). Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology*. **12**, 6-15.

Solmone M., Vincenti D., Prosperi M.C., Bruselles A., Ippolito G., Capobianchi M.R. (2009). Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J. Virol*. **83**, 1718-26.

Valli M.B., Meschi S., Selleri M., Zaccaro P., Ippolito G., Capobianchi M.R., Menzo S. (2010). Evolutionary pattern of pandemic influenza (H1N1) 2009 virus in the late phases of the 2009 pandemic. *PLoS Curr. Influenza*. **3**, RRN1149.