Each new emerging infection may become a big challenge to the medical community. Changing environment, tropical deforestation, melting of the Antarctic ice, growing population density and uncontrolled use of antibiotics provoke emergence and evolution of pathogens. Epidemics caused by new strains of the influenza virus, respiratory syndromes associated with coronaviruses, outbreaks of hemolytic Escherichia coli infections and antibiotic-resistant superbacteria are hazards to humans. Among high-priority measures for pathogen control that are yet to be taken is development of fast and accurate techniques for pathogen identification. Our review looks at the cases of new infections registered in the 21st century and explores feasibility of next generation sequencing for the detection and identification of new pathogens.

Keywords: infectious agents, pathogens, techniques for identification of new pathogens, next generation sequencing, coronavirus, influenza virus, horizontal gene transfer, resistance to antibiotics

Funding: this work was supported by the Russian Foundation for Basic Research, grant no. 15-54-04004.

Correspondence should be addressed: Valentin Makarov
ul. Leninskie gory, d. 1., str. 40, Moscow, Russia, 119991; makarovvalentine@gmail.com

Received: 03.02.2017 Accepted: 08.02.2017

ВОЗНИКНОВЕНИЕ НОВЫХ ИНФЕКЦИЙ В XXI ВЕКЕ И СПОСОБЫ ИХ ИДЕНТИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ (NGS)

Каждое новое инфекционное заболевание может быть серьезным вызовом для современной медицины. Изменение окружающей среды, вырубка тропических лесов, таяние ледников в Антарктике, увеличение плотности населения и повсеместное использование антибиотиков — все это факторы, провоцирующие появление новых патогенов. Эпидемии гриппа, вызываемые новыми штаммами вируса; респираторные синдромы, вызываемые новыми коронавирусами; вспышки инфекций, причиной которых является гемолитическая кишечная палочка; возникновение резистентных к антибиотикам супербактерий — примеры того, какие опасности могут поджидать человечество.

Ключевые слова: инфекционные агенты, патогенные организмы, методы выявления новых патогенов, высокопроизводительное секвенирование (next generation sequence).

Финансирование: работа поддержана Российским фондом фундаментальных исследований, грант № 15-54-04004.

Для корреспондентов: Макаров Валентин Владимирович
ул. Ленинские горы, д. 1, стр. 40, г. Москва, 119991; makarovvalentine@gmail.com

Статья получена: 03.02.2017 Статья принята к печати: 08.02.2017
Outbreaks of infectious diseases are a continuous threat to global health. A lot of effort is being put into the identification and study of new pathogens, among which are Middle East respiratory syndrome coronavirus, Zaire ebolavirus, and South American Zika virus. The table below lists factors that contribute to the emergence of new pathogens. However, a considerable proportion of epidemics are caused by known pathogens, such as poliovirus, influenza virus, or vibrio cholerae.

Risk factors contributing to the emergence of new diseases

<table>
<thead>
<tr>
<th>Pathogen-associated factors</th>
<th>Accumulation of mutations in the genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal gene transfer</td>
<td></td>
</tr>
<tr>
<td>Multipathogen infections</td>
<td></td>
</tr>
<tr>
<td>Changing antigenic determinants</td>
<td></td>
</tr>
<tr>
<td>Host-associated factors</td>
<td></td>
</tr>
<tr>
<td>Susceptibility to infections</td>
<td></td>
</tr>
<tr>
<td>Global transport networks</td>
<td></td>
</tr>
<tr>
<td>Increasing population density</td>
<td></td>
</tr>
<tr>
<td>Religious rituals and national traditions</td>
<td></td>
</tr>
<tr>
<td>Bioterrorism</td>
<td></td>
</tr>
<tr>
<td>Environment-associated factors</td>
<td></td>
</tr>
<tr>
<td>Changing environmental profile</td>
<td></td>
</tr>
<tr>
<td>Industrial development</td>
<td></td>
</tr>
<tr>
<td>Climate changes</td>
<td></td>
</tr>
<tr>
<td>Errors in healthcare</td>
<td></td>
</tr>
<tr>
<td>Animal epidemics</td>
<td></td>
</tr>
<tr>
<td>Wars and famine</td>
<td></td>
</tr>
</tbody>
</table>

Most outbreaks are caused by purely environmental factors, such as climate-related or geographical. However, human impact on the environment may also be a contributing factor. For example, some zoonotic diseases find their way into human communities because a natural habitat of their hosts has been destroyed. Aggravated by deforestation of mountain slopes, flooding causes outbreaks of cholera and other infectious diseases in populated areas. Some “anthropogenic” epidemics are directly linked to purposeful manipulations of pathogens. Modified in a lab, bioagents may be infectious or capable of acquiring virulence genes horizontally and therefore pose a serious biological threat. Mechanisms of new pathogens emergence are shown on Fig. 1.

Unfortunately, there are no thoroughly elaborated algorithms and ready commercial solutions to identify previously unknown pathogens. Techniques used to study their properties will vary in each individual case. The following review provides a detailed description of cases of emerging infectious agents of the 21st century and prompts a discussion about a possibility of elaborating a universal approach to pathogen detection using novel sequencing technologies.

New pathogens of the 21st century: examples and mechanisms of emergence

New coronaviruses

The 21st century has already seen the emergence of at least 9 new pathogens (Fig. 2). In 2002 the global healthcare was challenged by a previously unknown atypical pneumonia agent that came from China. In November 2002 a farmer died in the city of Foshan (Guangdong Province). Although the cause of death was inconclusive, it was clear that the patient had been afflicted with an unknown dangerous disease. On November 27, 2002 the Global Public Health Intelligence Network, a warning system developed by Health Canada in collaboration with the World Health Organization (WHO), picked up reports of an infection outbreak in China. Following a short investigation, WHO requested further information from China’s authorities. However, it was only after the epidemic crossed Chinese borders that details became available to the public. In February 2003 an American businessman died in Hanoi hospital after contracting pneumonia in China. The rate of disease progression was shocking. By March 15, the term “severe acute respiratory syndrome” (SARS) had been coined [1–3]; by March 27, its causative agent had been identified as a new coronavirus referred to as SARS-CoV [4–6]. From November 2002 to July 2003, a total of 8 098 patients in 25 countries contracted SARS; 774 patients died. In some populations [7] and age groups [8] mortality was as high as 40–55 %. Further scattered outbreaks of the infection were reported late in 2003 and early in 2004 in Singapore, Taiwan, Beijing and Guangzhou. All of them were linked to the cases of laboratory contamination and virus transmission from animals to humans [9], after the ban was lifted to sell palm civets in wet markets and serve palm civet dishes in restaurants imposed during the atypical pneumonia outbreak [10].

No effective antiviral agents were available at the time of the SARS outbreak [11], so basically, the treatment plan included supportive care and antibiotics to fight a secondary bacterial infection [12]. But due to the unprecedented international response, the outbreak was successfully contained [13]. Among the measures taken were contact tracing and isolation of people with suspected or confirmed SARS [14]. At present, SARS-CoV no longer circulates in the human population; however, a chance of a new epidemic remains as there are natural reservoirs of SARS ancestors, such as bats or other mammals [15].

Challenging as it was, researchers managed to identify the virus. Clinical specimens collected from patients with SARS were studied using cell cultures and molecular techniques. The virus was isolated in cell culture and then its 300-nucleotide-long RNA was detected by “random” polymerase chain reaction (PCR). Genetic characteristics of the virus revealed a very distant kinship to known coronaviruses (50 to 60 % similarity of nucleotide sequences). Based on the identified sequences, high sensitivity PCR- and real-time PCR-based assays were designed for virus detection. The virus was found in the clinical specimens of patients with SARS while the control samples came out negative. The sputum of infected patients was also found to contain high concentrations of viral RNA (up to 100 million molecules per 1 mL). Very low RNA concentrations were detected in blood plasma of infected patients in the acute phase of the disease and in their excrements by the end of treatment [4]. In spite of the fact that SARS outbreak was contained, SARS-CoV was not the only pathogen to threaten humans in the 21st century.

In 2003 a 7-month old baby presented to a hospital with obstructive bronchitis and conjunctivitis. A few tests were run to establish the presence of respiratory viruses, but all of them came out negative. A group of researchers headed by Lia van der Hoek proposed a modified technique for virus discovery based on cDNA-amplified fragment length polymorphisms (Virus-Discovery-cDNA-AFLP, VIDISCA). This method employs reverse transcription PCR (RT-PCR) with subsequent partial
Rapid accumulation of mutations in the genome of coronaviruses

Adaptation to the new host (human bocavirus, human metapneumovirus) and reassortment (influenza virus)

Recombination and reassortment (influenza virus)

Horizontal gene transfer (Enterohemorrhagic Escherichia coli, bacteria with multidrug resistance)

Evolution of gene functions by human-induced selection pressure (bacteria resistant to “last-resort” antibiotics)

Fig. 1. Genetic mechanisms of emergence of new pathogens among viruses (A) and bacteria (B)

cDNA digestion by frequently cutting restriction enzymes. The assay results revealed a certain similarity of the discovered sequences to the sequences of the already known coronaviruses; however the difference between them was still sufficient to classify the studied coronavirus as new. Later, the virus was termed “human coronavirus NL63” [16].

In January 2004, a 71-year old patient from China presented to a hospital with pneumonia. Attempts to replicate the virus in cell cultures, RT-PCR and direct antigen tests of nasopharyngeal aspirates showed the absence of known respiratory viruses in the patient. RT-PCR performed to target a conserved region of the coronavirus polymeric gene confirmed the presence of a coronavirus but attempts to culture it failed. Partial sequencing of the viral genome showed that its sequence was highly homologous to the sequences of other βCoV viruses including HCoV-OC43, but had a different origin. This human coronavirus referred to as HCoV-HKU1 was later isolated from the aspirate of another female patient [17]. Shortly thereafter, the virus was cultured in human ciliated respiratory epithelial cells, but on the whole its replication in cell culture still remains a difficult task. Since its discovery, HCoV-HKU1 has been proved to occur worldwide, and the retrospective analysis of stored nasopharyngeal swabs confirms that it can be traced back at least to 1995 [18].

In June 2012 the world became aware of the existence of a new strain of a human coronavirus. A 60-year old patient was suffering from a severe respiratory infection at Dr. Soliman Fakeeh Hospital in Jeddah, Saudi Arabia. Standard tests could not identify the pathogen. Patient’s sputum samples were sent to Rotterdam (Netherlands) where the virus was identified as a new coronavirus and termed HCoV-EMC (human coronavirus from Erasmus Medical College). The patient died later from acute pneumonia followed by kidney failure [19]. Since the discovery of the pathogen, a few of its isolates have been reported in the literature, various databases or mass media under different names. To study the virus, a research group was formed consisting of virologists whose major interest was in coronaviruses. To avoid confusion, the virus was given another name: the Middle East Respiratory Syndrome Coronavirus (MERS-CoR), which was approved by its discoverers, WHO and Ministry of Health of Saudi Arabia [19].

From June 2012 to February 7, 2014 there were 182 cases of MERS registered, of which 79 were lethal. According to WHO, by June 11 2014 there had been 699 laboratory confirmed cases; 209 people died [20]. Statistical reports reveal a 3-fold increase in disease prevalence over 4 months meaning that the epidemic is still raging. Mortality rates of up to 30 % are especially high in patients with comorbidities; patients with immunodeficiency or other primary diseases are also susceptible to the infection [21, 22]. There is also a serious risk of nosocomial transmission [23].

Clinical manifestations of MERS are similar to those of the acute viral respiratory infection and include such common respiratory symptoms as cough, fever and gastrointestinal dysfunction [24] before the onset of pneumonia [21]. Patients with MERS also tend to develop acute respiratory syndrome (ARS), renal failure, pericarditis and disseminated intravascular coagulation [24]. A risk of a pandemic is low since the virus is unlikely to effectively transmit between humans [24] and is transmitted only through close contact [25], between family members [26] or medical workers [27]; nosocomial transmission is also possible [28]. Patients with compromised immunity are especially susceptible.

The origin of MERS is not fully understood. Perhaps, the first transmission was from the camel to the human.

Over the past decade 4 new coronaviruses have been discovered, of which 2 are extremely dangerous; the other 2 were discovered accidentally and their signs are hard to distinguish from the signs of common acute viral respiratory infections. Our brief review shows that emergence of new highly virulent strains is very probable and only requires a couple of nucleotide polymorphisms in the viral genome to happen.

**Human metapneumovirus**

A new virus was isolated from the samples of 28 patients in the Netherlands in 2001. The symptoms of the infection were similar to those caused by the respiratory syncytial virus (RSV). A few patients were hospitalized; some required mechanical ventilation. Viral isolates were cultured in tertiary monkey kidney cells. Their cytopathic effect was pretty much identical to that of RSV. Electron microscopy of the supernatant of the infected...
H1N1 (Mexico, 2009)
HMPV (Netherlands, 2001)
E. coli 0104:H4 (Germany, 2011)
NDM-1 (India, 2009)
HCoV-HKU1 (China, 2004)

Fig. 2. Geographic distributions of emerging pathogens in the 21st century

cells detected paramyxovirus-like particles. But the use of real-time PCR primer sets for paramyxovirus detection yielded no results. Then a decision came to use RT-PCR assays with random primers to obtain information on the sequence of the unknown virus. Based on the similarity of sequences and genomic organization, it was concluded that the studied virus was a close relative of the avian pneumovirus. The virus was identified as a new member of the *Metapneumovirus* genus and called human metapneumovirus (HMPV) [29]. It was the first metapneumovirus capable of infecting humans. Although HMPV was discovered in 2001, phylogenetic analysis showed that the virus had been circulating in the human population for the last 50 years or so [30, 31]. From 7 to 19 % of respiratory infections in children who received either inpatient or outpatient care were caused by HMPV [32–34]. The literature reports that this virus ranks second in frequency among the respiratory viruses [35].

**Human bocavirus**

The first human bocavirus (hBoV) was discovered in 2005 in nasopharyngeal aspirates of 282 Swedish patients with the unknown infection of the lower respiratory tract. To remove all contaminating RNA from the samples, the latter were treated with DNasease prior to conducting RT-PCR with random primers. Bioinformatic analysis of obtained sequences revealed the presence of a new parvovirus in the samples that was highly homologous to bovine and canine parvoviruses (hence the name *Bocavirus*). The new virus was given a name of hBoV1 [36]. Three other strains of hBoV were discovered in 2010 and are now referred to as hBoV2, hBoV3 and hBoV4 [37–39].

HBoV1 causes respiratory diseases and is present everywhere across the globe accounting for about 19 % of all viral infections of the upper and lower respiratory tract in humans [40–42]. HBoV1 effectively infects epithelial cells of human airways and induces their cytolysis [43–45]. These data are confirmed by clinical observations indicating that the infection manifests as a respiratory condition. In contrast, hBoV2, hBoV3 and hBoV4 colonize the gastrointestinal tract; hBoV2 and possibly hBoV3 are associated with gastroenteritis [46, 47]. Interestingly, hBoV2 is the only intestinal bocavirus isolated from a nasal swab; therefore it may be associated with respiratory diseases [48, 49]. Though hBoV1 is found in all age groups, it is prevalent in infants of 6 to 24 months old [50, 51] and rare in adults [52–56]. Generally, transmission and infection occur throughout the year but are more often in winter and spring [55, 57–59].

**Influenza virus**

Another mechanism of pathogen evolution is genome recombination. A typical example here is a highly variable human influenza virus (IV) with a segmented RNA-genome. When several strains invade a host, their RNA segments may reassort to produce new pathogenic strains. Adaptive changes occurring in two surface proteins (hemagglutinin and neuraminidase) of the virus determine its ability to cause pandemics.

Water birds are a natural reservoir of IV in which the virus has evolved into its current state through several adaptation stages. Incredible diversity of IV strains is found in anseriformes and charadriiformes, including 17 hemagglutinin and 9 neuraminidase subtypes [60]. Transmission of the virus to land birds and mammals has triggered its rapid evolution [61]. Some strains of IV circulate in human populations (H1N1, H3N2), pigs (H1N1, H1N2), horses (H3N8, H7N7) and dogs (H3N8) [62]. Pigs have become a major reservoir for the pandemic strains
of the virus because they have receptors for both avian and human IV (2,3-sialic acids and 2,6-sialic acids, respectively) [63, 64]. Pigs are effective "mixing tanks" for the virus, a source of new reassortants that have mixed (recombinant) genomes and can cause another pandemic [61].

Pandemics are the most severe manifestation of the infection, with a 20–40 % global prevalence rate. One of the first documented IV pandemics occurred in 1918 when the deadly Spanish influenza took lives of 25 million people worldwide [60]. It was followed by the Asian flu (H2N2) in February 1957, Hong-Kong flu (H3N2) in 1968, Russian flu (H1N1) in 1977 and swine flu (H1N1) in 2009. The latter became the first and so far the last pandemic of the 21st century. H1N1 emerged through reassortment between the Eurasian swine influenza strain and North American triple reassortant H1N2 [65, 66]. In comparison with its "evil" ancestor, it is less virulent; however, it still caused 200,000 and 83,000 deaths by respiratory and cardiovascular complications, respectively [67].

Since the discovery of a new H7N9 strain of avian influenza on March 30, 2013, China’s authorities have reported 135 laboratory confirmed cases of infection, with 45 deaths in Shanghai, Anhui, Jiangsu and Zhejiang [68]. The only case registered outside China was in Taiwan; however, the patient contracted the virus in China [69]. Those were the first cases of transmission of H7N9 avian influenza to humans [70, 71]. Initially, nonfatal viral infections caused by H7 strains (H7N2, H7N3, H7N5) were observed across Europe and in the USA [72]. The only exception in terms of fatality was a death case of H7N7 infection reported in 2003 in the Netherlands [73, 74]. Interestingly, those outbreaks occurred at the time of the flu outbreaks in poultry, but no such pattern was observed for H7N9. Cases of H7N9 infection seem to be epidemiologically unrelated, but the possibility of virus transmission between humans remains [75]. Delayed serologic response in patients infected with H7N9 complicates detection of the virus by serologic tests [76]. Besides, unlike H5N1, H7N9 infection in poultry tends to be latent, which makes identification of its source and a route of transmission much harder and increases a risk of a pandemic.

Shiga toxin-producing Escherichia coli

Another mechanism contributing to the emergence of new pathogens relies on the acquisition of new properties by an organism, such as an ability to produce toxins or resistance to antibiotics. A consequence of such genetic transformation was an epidemic caused by the O104:H4 strain of the enterohemorrhagic Escherichia coli in 2011 in Germany. It was the most severe outbreak ever registered caused by shiga toxin-producing E. coli (STEC); in total, 3 842 cases were reported including 2 987 cases of laboratory confirmed gastroenteritis (with 18 deaths) and 855 cases of hemolytic uremic syndrome (with 35 deaths) [77]. The outbreak started on May 8, reached its peak on May 22 and was over on July 4. The outbreak may have been halted because people had been warned against using contaminated food; however, delivery of contaminated products to markets may have also stopped. Allegations about the source of the infection were publicly debunked (at first cucumbers and cabbages were thought to be contaminated, but that was not true) [77]. On June 10, German authorities announced that infection had come from Egyptian sprouts of fenugreek [78].

Epidemiologic analysis of the infection initially transmitted through food is hard to perform once a pathogen learns to transmit between humans. Human-to-human transmission of enterohemorrhagic Escherichia coli O157:H7 was observed in about 20 % of households with an infected patient who had contracted the virus through food [79]. Secondary household transmission of the O104:H4 strain between adults was also observed in France [80] and the Netherlands [81]; it became possible due to the delayed onset of the infection compared to the standard incubation time (7 to 9 days for O104:H4). Secondary transmissions were observed in Hessen (Germany) that lied outside of the epidemic area in the North [82]. Investigations proved the facts of household and nosocomial transmissions; there was also a case of transmission between laboratory staff.

Within a very short time, the O104:H4 strain isolated in Germany was sequenced by a few groups of researchers. The first sequence was obtained in the Beijing Institute of Genomics from a sample provided by the University of Hamburg. Expeditied by the use of the Ion Torrent platform, sequencing of the bacterial genome only took 3 days. The first annotated sequence was published by researchers from the University of Goettingen who used the following genome sequencers: Flex [83], Ion Torrent [84] and PacBio RS [85]. A combination approach based on the used of several next generation sequencing techniques yielded higher assembly quality (longer read lengths, fewer errors and missed regions, etc.). Sequence mapping revealed a similarity between the studied strain and 4 other strains of enterohemorrhagic Escherichia coli that had also caused infection outbreaks, including enteroaggregative E. coli (EAEC) isolated from AIDS-stricken patients with chronic diarrhea in the 1990s in Central Africa [86]. However, the African strain did not contain the Stx2 prophage [84]. Mellmann et al. proposed a model of O104:H4 evolution according to which the progenitor strain had transformed into O104:H4 by removing or acquiring mobile DNA elements through horizontal transfer [83]; a German variant of the pathogen had acquired plasmids that carried fimbriae/pili genes (AAP/I) and lost plasmids that carried the genes of TEM-1 and CTX-M-15 enzymes responsible for developing resistance to antibiotics. Comparison of the epidemic strains also revealed extensive rearrangements in the isolates, including deletions, insertions and inversions, which indicated considerable genomic mobility. Researchers also found that it was those structurally different regions that contained fragments encoding virulence factors.

Why was strain O104:H4 so virulent? The study of genome and virulence genes showed that this strain had an unusual combination of SPEC virulence genes (prophage Stx2, long polar fimbriae, tellurite resistance, iron metabolism) and EAEC virulence genes (AAP/I, transcriptional regulator AggR, dispersin Aap and shigella enterotoxin Set1) [87]. The latter are localized to the pAA virulence plasmid [83]. Thus, the virulence of the O104:H4 strain is ensured by two different mobile elements, prophage Stx2 and plasmid pAA, which is quite unusual. It may have been the combination of SPEC and EAEC virulence factors that shaped this new extremely dangerous pathogen. It causes cytotoxic damage to the intestinal epithelium facilitating systemic absorption of shiga-toxin, which may explain the high prevalence of hemolytic uremic syndrome in Germany. But in spite of 2 antibiotic-resistance genes in O104:H4, the epidemiologic situation, in particular mortality rates, could have been worse if the virus had had resistance to a broader range of antibiotics.

Antibiotic resistance and superbacteria

Since the first cases reported in the 1980s, strains with multidrug resistance (MDR) have become common sources of
nosocomial infections [88]. Many countries, including Russia, have increasingly witnessed infections resistant to traditional antibiotic treatments. It should be noted that major sources of infections caused by such pathogens as methicillin-resistant staphylococcus aureus, vancomycin-resistant enterococcus and other gram-negative bacteria with MDR are intensive care units [89].

Carbapenem resistance of gram-positive bacteria poses a particular problem. Carbapenems are drugs of choice used to treat many infections caused by gram-negative bacteria [90]. Extensive use of carbapenems promoted antibiotic resistance in bacteria. The most common carbapenem-resistant superorganisms are Pseudomonas aeruginosa, Acinetobacter baumannii and enterobacteria [91].

Pseudomonas aeruginosa causes acute invasive infections in patients with compromised immunity or in critical condition. Isolates of P. aeruginosa obtained from patients of intensive care units demonstrated resistance to carbapenems in 28–37 % of cases [92, 93]. A. baumannii is also one of the major sources of nosocomial infections. Initially this pathogen was sensitive to imipenem treatment in most medical institutions. But soon its strains were rapidly evolving to develop carbapenem resistance. At the moment 50–60 % of nosocomial infections associated with A. baumannii do not respond to imipenem treatment [94, 95]. Many enterobacteria, a broad range of beta-lactamase-producing E. coli and strains of Klebsiella pneumoniae resistant to carbapenems pose a serious threat to patients in intensive care because carbapenems are used as last resort antibiotics [96].

The driving force of carbapenem resistance is thought to be the extensive use of the third generation cephalosporins, aztreonam and ipinem. Emerged in the 21st century, superorganisms are totally resistant to any known antibiotics and are a serious challenge to modern medicine. Emerging pathogens are a product of both acquired resistance genes and the activation of “hidden” resistance genes resulting from a few significant nucleotide polymorphisms. Such genetic modifications are typical for microorganisms. In this light, a focus on the bacterial resistome — a sum of all resistance genes in the entire microbial community — is a prerequisite for effective identification and elimination of pathogens.

The resistome concept is based on the fact that soil actinobacteria and many other microorganisms actively produce antimicrobial compounds. It seems obvious that in order to survive, a microorganism not only has to develop defense against antibiotics: it also needs an ability to produce them. As proved by some studied, many resistome components emerged long before antibiotics were introduced into clinical use [97]. Metagenomic analysis of ancient DNA samples collected in permafrost zones revealed the presence of beta-lactam-, tetracycline-, and glycopeptide-resistance genes [98]. It was shown that modern glycopeptide-producing organisms harbor ancient glycopeptide resistance genes (vanHAX). Moreover, the VanA protein, one the most important products of glycopeptide resistance genes, has preserved its function and 3D structure over centuries [99]. In another study, bacteria found in caves that had had no contact with the surface for over 4 million years proved to be resistant to 14 different antibiotics [100]. Genotyping and biochemical assays show that resistance genes are present in the microbial pangenome regardless of the human-induced selective pressure [100].

Although the independent ancient origin of antibiotic resistance genes is evident, humans have largely contributed to the formation and transformation of the resistome in its current state. Resistance proteognes do not form a stable phenotype but are capable of transforming into resistance genes when undergoing a mutation or due to contextual changes. Mutations of the enzyme facilitating its transition from one functional class to another are highly unlikely to occur while the expansion of the substrate specificity range in the enzymes with retained function is very probable. Structural studies demonstrated the evolutionary proximity between lincomycolides and aminoglycoside nucleotide transferases and polymerases, and this allows for a supposition that progenitor polymerases were resistance proteoges that later evolved into antibiotic-modifying genes [101].

Conserved structural elements and biochemical mechanisms detected in a similar way indicate that protein kinases and protein acetyltransferases share common ancestors with resistance proteoges from which aminoglycoside resistance genes were derived [102, 103]. Moreover, resistance genes themselves can function as resistance proteoges. For example, aminoglycoside acetyltransferase aac(60)-i-cr ensures resistance to quinolones [102]. The ancestral enzyme aac(60)-i-la ensures resistance to kanamycin (which is an aminoglycoside); mutations of its two amino acid residues Trp102Arg and Asp179Tyr turned to be sufficient to extend its substrate specificity to include a number of quinolone antibiotics, such as ciprofloxacin, without losing aminoglycoside acetyltransferase activity.

The frequency of resistance proteoges in the resistome is unknown. To be considered clinically significant, these proteoges have to undergo a series of important evolutionary events. However, the examples above show that enzymes have a potential to include more substrates in their “profile” and might contribute to the emergence of new resistance genes.

Similar to resistance proteoges, silent resistance genes cannot form a resistant phenotype in their current structural state. Unlike proteoges, these genes can be detected in the resistome based on the homology between their sequences and the sequences of known resistance genes. For example, two antibiotic-sensitive strains of Citrobacter freundii isolated before antibiotics entered the clinical setting contain AmpC beta-lactamase genes [103]. Mutations that trigger AmpC expression in these strains induce resistance to broad-spectrum cephalosporins. The wild type of Salmonella enterica cultured in the enriched growth-supporting medium is sensitive to streptomycin and spectinomycin. However, the same strain is resistant to both drugs when cultured in nutrient-poor medium due to the activation of aminoglycoside adenylytransferase gene aadA [104]. Overexpression of aadA from a plasmid resulted in streptomycin resistance (the minimum inhibitory concentration of streptomycin increased). Thus, a total expression level of a resistance gene may be critical in the formation of a resistant phenotype.

If a mutation is seen as a driving force of evolution, then horizontal gene transfer is a magic wand that can transform the inactive resistance gene into a fully functional one by increasing the number of gene copies or changing the context that ensures gene expression under a strong promoter. Having become a component of a mobile element, resistance genes discover an opportunity to spread throughout the entire microbial pangenome where they can pick up further mutations reinforcing their function and expanding the range of possible enzyme substrates in response to the environmental selection pressure.

Staphylococcus aureus with its variety of genes capable of horizontal transfer in human pangenome is a perfect illustration of their role in antibiotic resistance: mobile elements account for 15–20 % of its genome, including
bacteriophages, pathogenicity islands, plasmids, transposons, and staphylococcal cassette chromosome mec [105]. Accumulation of these mobile elements is a result of selection pressure, but the element source is bacteria that once co-existed with Staphylococcus aureus. While details of interactions between pathogens and commensal bacteria remain largely unclear, we are coming to realize that major reservoirs of resistance genes available to pathogens are harbored by the human microbiome [106]. Thus, metagenomic libraries that include samples of intestinal microorganisms of infants, children and teenagers report resistance to 14 antibiotics [107]. Moreover, all libraries report resistance to tetracycline, trimethoprim, trimethoprim-sulfamethoxazole, D-cycloserine, chloramphenicol, and penicillin, and some of them report resistance to aminoglycosides, glycolcyclines and beta-lactams. About 3 % of all antibiotic resistance genes listed in those libraries are associated with mobile elements, such as transposons or integrons [108]. The effect of antibiotics on intestinal microbial communities is actively studied. For example, some antibiotics, especially metronidazole and beta-lactam, negatively affect the variety of microorganisms in the gastrointestinal tract [108]. If any bacterial taxon starts to dominate the gut flora, it increases a risk of bacteremia [109].

Members of the human gut flora can acquire resistance genes horizontally (from farm animals to humans through food). A group of researchers discovered that 42 unique resistance genes had been transmitted to the human microbiome by agricultural isolates, which allows for the assumption that the normal flora of farm animals, as well as waste, may contribute to the development of drug resistance in human pathogens [109]. Mobile elements that carry antibiotic resistance genes are widely spread in the microorganisms we consume with food [109]. Mobile resistance determinants, i.e. transposons, plasmids, integrons, and phages, are beneficial for the human microbiome. Unfortunately, overuse of antibiotics on farms is not rare. Monitoring on Chinese pig farms [114] showed that antibiotic resistance genes were found almost everywhere in the soil, as the latter was fertilized with the manure of pigs who had received antibiotic-containing food. Tests of pathogen-containing agricultural samples revealed a 3-fold increase in the number of unique resistance genes compared to the controls, including resistance to clinically significant antibiotics, such as macrolides (mPnA and erm), cephalosporins (bla-TEM and blaCTX-M), aminoglycosides (aph and aad) and tetracycline (tet). The number of transposases in the genomes of pathogens found in pig manure and soil samples was 90 000 and 1 000 times higher, respectively, than in the controls. The number of transposases positively correlates with the frequency of resistance genes (especially tetracycline resistance genes) in the microbiome of agricultural products.

To sum up, all mechanisms of emergence of new pathogens can fall into two categories: 1. host-to-host transmission of the known pathogen accompanied by an acute infection in the new host due to the lack of adaptation of the latter (a good example here is a cytokine storm); 2. emergence of new pathogenic properties in the known biological agents usually acquired through horizontal gene transfer.

Identification of new pathogens using traditional methods. Difficulties

So far, a lot of technologies and commercial applications have been developed for pathogen detection and identification. They can “spot” nucleic acids and antigens typical for a pathogen. Although many of those methods are claimed to meet the strictest requirements for sample preparation, processing rate, accuracy and reliability, only a few of them can be used in real life circumstances, especially in the field [115]. Biohazard detection systems must ensure timely identification and confirmation of biological risk factors straight in the sample yielding as few false positive or false negative results as possible. Such systems must be able to detect a modified or an unknown pathogen. Devices for biohazard detection must be portable, easy to use and capable of detecting several or even dozens or hundreds of factors simultaneously [115].

Currently there are a few diagnostic methods that meet most of the listed requirements but there is not a single tool that would meet all of them. Unlike chemical detectors capable of scanning a sample for health-threatening amounts of chemical compounds, low-sensitive biological detectors rarely “spot” potentially hazardous amounts of pathogens straight in the sample; what is more, the sample must be preprocessed before the test. Diagnostic systems based on nucleic acid amplification are generally more sensitive than antibody-based systems [115]. For example, PCR assays can detect individual molecules of microbial nucleic acids within a relatively short time [116–118]. However, this technique still requires thorough preparation of the sample and cannot directly detect toxins or infectious agents deprived of nucleic acids (such as prions) [115].

Specificity is a no less important parameter of a diagnostic method, as there is always a need to minimize background signals or false positive results when processing a complex mix of organic and inorganic compounds. High levels of competitor antigens or DNA fragments in the sample may render the test nonspecific. High sensitivity of PCR-based assays may actually be their drawback in the case of contaminated samples yielding false positive results due to the presence of various substances, including humic acids and heme, that inhibit polymerase activity.

Another important requirement for a diagnostic method is its reproducibility, which may be influenced by a number of factors, including reagent stability or varying test conditions. The impact of these factors may be reduced by introducing standards for sample collection and subsequent analysis.

In addition to the requirements listed above, diagnostic methods must be capable of performing a multiplex analysis, i.e. detect more than one bioagent in a sample. Samples often contain a mix of toxins, bacteria, viruses, etc. Besides, there may also be genetically or antigen-modified elements, previously unknown microorganisms or emerging strains of well-known pathogens, all of which are extremely difficult to detect. It should be noted that even regular bioagents are hard to detect in contaminated samples. Human specimens (blood or excrements), food, water, or air samples are “difficult” objects to detect in contaminated samples; what is more, the sample must be preprocessed before the analysis, which means longer tests and renders field diagnostics impossible.

Sample composition determines conditions for its storage and transportation. Air and water samples must be brought to concentrations allowing preliminary detection of target molecules. Air samples must be transformed to a liquid state because the majority of diagnostic tests work with liquids. Sample volumes and transportation are also important especially when it comes to living organisms. Sometimes to
assess the risk, the viability of a pathogen must be confirmed; in this case standard genetic or immunological assays will be of no use.

Over the past years, methods for detection and identification of unknown pathogens have been actively developed and profusely funded [115]. The most promising technology among them is next generation sequencing.

**Next generation sequencing, Basic principles**

The term “next generation sequencing” (NGS) is used to describe a group of methods for parallel sequencing of multiple fragments that unlike Sanger sequencing allow reading massive volumes of primary DNA sequences in one go. NGS has become a truly universal method of describing genomes of living organisms. Currently NGS-based applications are actively used in scientific research, molecular systematics, bioengineering, cellular and molecular biology, and in routine human activities: medical practice, criminology, selection, etc.

There are two major groups of NGS types: sequencing of multiple preamplified DNA fragments and single-molecule sequencing.

All sequencing methods based on template amplification share the same principle regardless of the reagents or devices used. First, a library is prepared by DNA fragmentation and adapter ligation. Then, library fragments are immobilized on beads or flow cell surface; each fragment is amplified by emulsion bead PCR or bridge PCR, respectively. Specific primers are then hybridized to adapter sites and sequencing is performed. This process is accompanied by signal emission. Signal type depends on the platform used. The signal is registered by the device that converts it into a nucleotide sequence.

Pyrosequencing or 454-sequencing was a pioneer NGS variation. The idea behind it is as follows: when a nucleotide is added to an elongating complementary sequence, light is emitted. [121–123]. Another NGS type, semiconductor sequencing, is based on measuring changes in pH values caused by H\(^+\) release that occurs during formation of phosphodiester bonds as nucleotides are added to a complimentary strand [124–127]. Another NGS variation is sequencing by ligation: a sequencer captures a fluorescent signal emitted during complimentary strand synthesis in a flow cell into which a mixture of fluorescently labeled nucleotide probes (octamers) and a DNA ligase are pumped [128]. The most common type of NGS is sequencing by synthesis which employs fluorohore labeled reversible terminator nucleotides. Amplification is performed inside a porous flow cell into which reagents for DNA synthesis are pumped [129]. After cluster PCR amplification, clusters of clonal DNA copies are generated to the cell surface, with each cluster corresponding to one read. High cluster density (up to 800–900 thousand per mm\(^2\)) provides sufficient throughput in terms of the obtained data. Clusters of DNA molecules are then sequenced according to the principle similar to the Sanger method [130, 131].

Among the drawbacks of NGS based on the use of preamplified DNA fragments are sequencing errors in homopolymer regions or regions that contain single nucleotide polymorphisms; problems related to repeat resolution; dependence of read accuracy on GC-content of DNA fragments, etc. [132–134]. All of these factors dictate the need for alternative sequencing techniques, such as single-molecule sequencing.

One of its types is based on the use of DNA polymerase to catalyze incorporation of a fluorescently labeled nucleotide into the elongating strand. Incorporation is captured by a highly sensitive CCD-camera. Once the nucleotide is incorporated, the fluorescent label is removed and fluorescence goes back to normal values. Then another nucleotide enters the DNA polymerase active site and the cycle is repeated [135]. Phage φ29 DNA polymerase used in this technique can process up to 10 nucleotides per second. This technique can be used to sequence long DNA molecules — up to 10 000–20 000 base pairs and already has a number of practical applications [136–142].

Another type of single-molecule sequencing uses electrophoretic cells equipped with a nanopore membrane. Single stranded DNA molecules are threaded into the pore; as the molecule enters the pore, the amount of current that passes through it changes [113]. Based on the properties of this change, such as duration and amplitude, it is possible to accurately identify the nucleotide that enters the pore at a particular time point. So far, this approach has been implemented in one commercial sequencer (MinION by Oxford Nanopore Technologies, UK) distributed under the early access program [143]. The advantage of this sequencing type is a possibility to run long reads without having to use expensive equipment. Its major drawback is high error rates (12–20 %) [144]. However, it is becoming clear that nanopore sequencing is an increasingly promising technique for metagenomic studies, sequencing of short genomes, identification of viral and bacterial agents. Nanopore sequencing has been successfully used as a diagnostic test to detect Ebola and Chikungunya viruses. It is also a good technique for conducting metagenomic research of bacterial resistome and sequencing large-scale genomes [145–147].

**NGS-based strategies for pathogen identification**

The main group of pathogens that pose the biggest threat to humans includes bacteria and viruses. Other pathogens such as fungi or protists are no less dangerous but do not normally require a genetic analysis to be identified. At present, the major technique for describing a diversity of microorganisms in the sample is metagenomic analysis. It has become possible and even routine due to NGS. The diversity of microorganisms found in the sample can be described using two different strategies: targeted sequencing of selected marker regions and large-scale (whole) metagenome sequencing.

The first method is simple, cheap and takes less time for sample preparation, sequencing and data processing. However, it has its limitations and can only be used to detect the presence of different organisms in the sample. In contrast, the second method yields a full profile of the microbial community, including the description of its genetic properties. Usually regions of the 16S rRNA gene of prokaryotes and 18S rRNA gene of eukaryotes are recommended as marker regions for processing metagenomic samples; for fungal samples ITS regions are recommended [148–150]. However, the task may dictate the use of other markers. For example, for generating a resistome profile, regions of antibiotic resistance genes should be selected.

The second method is costly and time consuming. However, whole genome sequencing provides a basis for further assembly of a reference genome [148]. These are three ways to analyze the obtained data (fragments of microbial DNA contained in the studied metagenomic sample). The first method involves comparison of marker sequences with known sequences obtained from databases that describe genomes of similar organisms [151–153]. The second method involves...
clustering of all reads into taxon groups (based on their similarity to known whole genome sequences, etc.) [154–157]. The third method is based on the assembly of the obtained contigs into genes or even genomes de novo [158, 159]. Whole genome sequencing and methods of data interpretation are highly useful tools for pathogen identification as they help to identify individual genes in the sample.

Both approaches have their own drawbacks and advantages. Sequencing of individual regions is fast and cost-effective and gives a general idea of the genetic diversity of the sample, while whole metagenome sequencing provides full information on pathogen determinants (metaresistome etc.). Most of the obtained data will not be of any particular value, but what is important is that sequencing will yield a comprehensive list of genetic elements that determine epidemiologic properties of pathogens contained in the sample. It might be possible to use a reagent kit instead of conducting whole genome sequencing that consists of several hundreds of oligonucleotide sequences complementary to important epidemiologically significant determinants. The use of the kit would speed up metagenomic processing and cut its costs. Then total/whole genome sequencing could be used in some difficult cases following preliminary pathogen detection to provide a detailed genetic profile of a mixed sample or isolated pathogen.

Specific aspects of NGS-based identification of viruses and bacteria

State-of-art NGS techniques are ideal when there is a need to analyze, identify and describe genomic sequences of isolated prokaryotic organisms. NGS certainly holds promise as an effective tool for identification of unknown pathogens in mixed samples. However, there may be a difficulty in detecting horizontal transfer factors in samples containing prokaryotic organisms; drawing up accurate genomic profiles for individual members of such microbial communities may also be an issue. To minimize these issues when working with chromosome-bearing genetic elements, a better coverage of chromosome sequences of individual metagenomic components by single reads is required.

Quality of data obtained through sequencing is largely determined by sample preparation. Its significance becomes obvious once we take a closer look at the aspects of virus identification. Identification of new viruses is a challenging task: viral nucleic acids are very hard to isolate from junk nucleic acids. Extraction of nucleic acid from virus particles obtained through ultrafiltration of large DNA viruses results in sample contamination by the so-called gene transfer agents: nonviral DNA packed in viral capsids [160]. Identification of small highly variable RNA viruses is complicated by the presence of contaminating amounts of rRNA in nucleic acid samples. There are certain difficulties with primer selection; primers need to be universal and allow amplification of at least genus-specific viral cDNA [161]. One of the popular techniques used to identify emerging viruses relies on a modified PCR assay (VIDISCA) followed by NGS (Fig. 3) [162]. Below is a brief description of the technique.

First, the sample is selectively enriched with viral nucleic acid; as part of the procedure, the sample is centrifuged to remove residual cells and mitochondria. The sample is also treated with nucleases to remove interfering chromosomal and mitochondrial DNA and RNA from lysed cells. Adding RNAsa to the sample causes degradation of cellular RNA, but the viral nucleic acid remains intact, because it is packed inside a capsid. Then nucleases are inactivated and viral nucleic acids are extracted from the sample. RNA is reverse transcribed into cDNA and a complementary strand is synthesized from viral RNA or genomic DNA [163]. Double stranded DNA is then digested by frequently cutting restriction enzymes (HinP1-I and Mse-I). The cleaved DNA is then ligated to HinP1-I and Mse-I adaptors with complementary overhangs. Target molecules are amplified using primers specific to each adaptor. For further selective amplification primers with a supplementary base (G, A, T or C) are used. In total, 16 combinations of primers are used; each sample is compared to the negative control (uncontaminated serum or plasma and supernatant of noninfected cultures). PCR products specific for infected samples are then cloned and Sanger-sequenced.

This technique is quite difficult to perform and its throughput is relatively low; reproducibility may also be an issue [163]. Currently, a modification of the method is attempted based on a combination of PCR with NGS. The amplified fragments are conjugated to nanoparticles and sequenced by massive parallel sequencing. The original method was based on the pyrosequencing technology; the license for it was acquired by Roche. But a serious problem arose related to a low number of clean reads due to the presence of ribosomal RNA (rRNA) in the sample. Therefore, the method yielded poor results. There are a few approaches that can help to reduce the amount of...
contaminating rRNA in the samples such as the use of specially designed primers that do not anneal to rRNA, low-frequency-cleavage restriction enzymes and specific oligonucleotides for blocking cDNA synthesis on rRNA [163]. Although these “patches” significantly reduce the number of amplified rRNA fragments, the obtained result is still far from being perfect, as viruses are detected in only 50% of contaminated samples. However, if the problem of rRNA removal from the samples is fully solved, the technique will certainly be one of the most time-saving and accurate tools ever used for the detection of previously unknown viruses.

CONCLUSION

Emergence of new bacteria and viruses that pose a serious threat to global health is inevitable and dictated by evolution. Viruses and bacteria are highly adaptive due to a number of molecular mechanisms at their disposal, such as recombination, reassortment and horizontal gene transfer. Coupled with a capacity to produce abundant progeny and human-induced selection pressure, these mechanisms expedite emergence of new pathogens. Considering close international contacts among humans, pathogen spread to new areas aggravating the risk of epidemics. However, this risk may be reduced by the development of new methods for infection control (vaccination, medications, new sterilization technologies), and techniques for pathogen identification that must take into account the genetic adaptive capacity of pathogens. Literature review revealed that there are no ready commercial solutions for identification of organisms with new pathogenic properties. Traditional PCR and immunoassays have a number of limitations. One of the most promising methods used to identify a broad range of pathogens is next generation sequencing.

Next generation sequencing is one of the few available methods that can detect a pathogen, generate its genetic and epigenetic profile, and provide information on the microbial community inhabiting the sample. Rapid evolution of sequencing techniques makes the analysis easier, cheaper and faster. Enhanced with a variety of software applications, next generation sequencing becomes an effective tool for identification of previously unknown pathogens.

References


Kapoor A, Simmonds P, Slikas E, Li L, Bodhidatta L, Coleman CM, Frieman MB. Coronaviruses: Important Emerging
Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, et al. van den Hoogen BG, de Jong JC, Groen J, Kuiken T, 
Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y, Boucher D, Weibel C, Ferguson D, Kapoor A, Slikas E, Simmonds P, 
Chieochansin T, Naeem A, Cleiri DJ, Ricketti AJ, Vernaleo JR. Severe acute respiratory
de Sousa R, Reusken C, Koopmans M. MERS coronavirus: Data 
Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of 
Memish ZA, Zumla A, Al-Hakeem RF, Al-Rabeeaah AA, Stephens GM. Family cluster of Middle East respiratory syndrome 
Arthur JL, Higgins GD, Davidson GP, Ginvey RC, Ratcliffe RM. A novel bocavirus associated with acute gastroenteritis in 
Kapoor A, Simmonds P, Slikas E, Li L, Bodhidatta L, Sethabutr O, et al. Human bocaviruses are highly diverse, 
dispersed, recombination prone, and prevalent in enteric 
Sethabutr O, et al. Human bocaviruses are highly diverse, 
dispersed, recombination prone, and prevalent in enteric 
Sethabutr O, et al. Human bocaviruses are highly diverse, 
dispersed, recombination prone, and prevalent in enteric 
Sethabutr O, et al. Human bocaviruses are highly diverse, 
dispersed, recombination prone, and prevalent in enteric


BULLETIN OF RSMU | 1, 2017 | VESTNIKRGMU.RU | REVIEW - EVOLUTION OF INFECTIONS


review of infections


