OPEN ACCESS

Applied genomics for identification of virulent biothreats and for disease outbreak surveillance

Martin C Nwadiugwu,[1] Nelson Monteiro[2,3]

[1]Department of Biomedical Informatics, University of Nebraska Omaha, Omaha, Nebraska, USA
[2]The Forsyth Institute, Cambridge, MA, USA
[3]Department of Developmental Biology, Harvard School of Dental Medicine, Boston, MA, USA

Correspondence to
Martin C Nwadiugwu, Department of Biomedical Informatics, University of Nebraska Omaha, Omaha, Nebraska, USA; mnwadiugwu@unomaha.edu

Received 8 February 2021
Accepted 18 December 2021

To cite: Nwadiugwu MC, Monteiro N. *Postgrad Med J* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/postgradmedj-2021-139916

ABSTRACT

Fortifying our preparedness to cope with biological threats by identifying and targeting virulence factors may be a preventative strategy for curtailing infectious disease outbreak. Virulence factors evoke successful pathogenic invasion, and the science and technology of genomics offers a way of identifying them, their agents and evolutionary ancestor. Genomics offers the possibility of deciphering if the release of a pathogen was intentional or natural by observing sequence and annotated data of the causative agent, and evidence of genetic engineering such as cloned vectors at restriction sites. However, to leverage and maximise the application of genomics to strengthen global interception system for real-time biothreat diagnostics, a complete genomic library of pathogenic and non-pathogenic agents will create a robust reference assembly that can be used to screen, characterise, track and trace new and existing strains. Encouraging ethical research sequencing pathogens found in animals and the environment, as well as creating a global space for collaboration will lead to effective global regulation and biosurveillance.

BACKGROUND

In recent years, research has been directed towards strengthening our biodefense mechanisms to cope with disease outbreak and threats from harmful biological agents. However, proactive and preventative measures such as securing human, agriculture and food systems from biological contamination, stemming the misuse and trafficking of dangerous biomaterials, are potential solutions that can increase our preparedness and forestall harmful pathogenic invasion. Virulence is the ability of pathogens to enter, persist and replicate in a host and virulence factors allow for successful invasion of a host without much resistance because they elude and subvert host defenses,[1] and enable the effectiveness of pathogenic infections. Virulence factors can be genetically identified experimentally and altered from pathogenic to non-pathogenic forms[2]; they include adherence materials that stick to host cells which may be located on the surface protein of pathogens, invasion factors such as capsules that shields pathogens from host phagocytosis, endotoxins such as lipopolysaccharide, exotoxins such as neurotoxins, cytotoxins, and enterotoxins etcetera, and siderophores (iron-binding factors that compete with the host for iron).[3]

Global genome-based biosurveillance system of zoonotic pathogens[4 5] could be a cornerstone for biodefense. The US Institute of Medicine suggested that a major future outbreak of infectious disease may be from a pathogen in a populated area that was relatively unknown; and metagenomics (complete genomic analysis of all organisms present in an acquired environmental sample) have been touted as a potential means to discover virulence factors and identify unrecognised pathogens[4] to distinguish between intentional or accidental release, and advance our knowledge of the pathogenesis of infections and the role played by virulence factors.[1 4] Genomics can be applied to diagnose, manage and predict diseases based on sequencing and gene rearrangement,[6] as well as to compare and uncover interactions between animal species and classify the diversity of infectious agents.[5] When this technique is applied in public health response to an outbreak, insight from the sequenced genome could be relied on to predict possible infection locations and strengthen traceback capability.

Furthermore, when horizontal gene transfer (HGT) between microorganisms is detected using genomics, pathogenic and non-pathogenic changes in their virulence phenotype and their interactions within symbiomes may be identified, including dispersal of genes encoding for virulence factors. HGT plays a role in sustaining pathogenic evolution and this knowledge is beneficial in deciphering the origin of a new pathogenic strain and in biosurveillance. Ongoing international genomics projects geared at strengthening biosurveillance and biodefense capabilities include major consortiums such as the Global Virome Project (GVP), Earth BioGenome Project (EBP) and the international barcode of life (iBOL).

The main objective of this paper is to explore how genomics can be applied to proactively prevent an outbreak by studying chromosomally encoded virulence factors. It aims to explain how virulence factors can and have been identified, how changes in pathogenicity can be determined using genomic techniques, and how genomic techniques and global collaboration can be used to prevent disease outbreak. What is more, it briefly explains how genomic knowledge and technology could be manipulated for disruptive purposes. The paper concludes with a recommendation on approaches that could strengthen ongoing biodefense efforts.

DETECTING VIRULENCE FACTORS USING GENOMIC TECHNIQUES

One of the main goals of agencies such as the US National Strategy for Countering Bio-threats is to build a global capacity for disease detection surveillance to ensure safety and public health.[4] However, the ability to differentiate pathogens and use this information as countermeasures to defend

biological threats is still limited. Nonetheless, because high throughput genome sequencing can identify DNA assortment and rearrangement, as well as HGT, quasi-species distribution and virulence evolution, it has emerged as a technique for identifying and differentiating pathogens.

Pathogens have virulence factors that exist in avirulent and attenuated strains due to mutations that attenuate their virulence, however, this may only occur in one of several factors.[2] Virulence factors existing in virulent strain of pathogens are more of a concern because they are pathogenic. Even so, they can be predicted using similar approaches in reverse vaccinology—a vaccine target prediction method conceptualised by the advent of whole genome sequencing (WGS). Reverse vaccinology is the application of bioinformatics method to rationally design vaccines using genomic data without the need for isolating, inactivating and growing the specific microorganism as proposed by Louis Pasteur in the twentieth century.[7] This method has been applied to identify new functional meningococcal antigens,[8] and to predict genes that encode for virulence in the genome of existing pathogens. For example, virulence factors on the surface of *Pseudomonas aeruginosa*, a bacterium spread to humans through improper hygiene have been analysed by excluding proteins that are not present on its surface.[9] Pan-genome analysis is another method that can be used to identify genes that encode for virulence in pathogens and differentiate their strains from avirulent type because they can be applied to analyse the genome of all strains of species, while also capturing subsets of strains shared by close members.[2] Therefore, reverse vaccinology and pan-genome analysis are two techniques that employs genomic science and data to identify genes of a pathogen encoding for virulence.

An application that has been deployed in identifying virulence factors is the Vaxign programme. The Vaxign programme is a web-based vaccine design platform used to query protein sequences to find possible targets[10]; it has been used in genome prediction of virulence factors in pathogens such as *Corynebacterium diphtheriae* and *Staphylococcus aureus*.[11] Using the Vaxign programme, gene encoding virulent factors of pathogens can be predicted after genome sequence annotation by querying the data in the seed strain of the pathogen. Moreover, their conservation, homology, and location on the pathogen can also be identified.[2] The secreted factors found on the surface of virulent microorganisms are possible virulence adherence materials that may share similar strain with other pathogens, allowing for the detection of diverse strain in multiple species. Therefore, the Vaxign programme is a web-based application that can allow for the prediction and gene annotation of virulent factors.

WGS-based bioinformatics methods have been used by Mason *et al*[12] to predict virulence genes and antimicrobial susceptibility from 1379 isolates of *S. aureus* that were previously characterised using PCR, disc diffusion, broth and agar dilution methods. The bioinformatics methods employed namely, de Bruijn, read-based and Basic Local Alignment Search Tool (BLAST), predicted the absence or presence of 83 virulence genes and antimicrobial resistance determinants. The prediction for antimicrobial susceptibility matched 98.3% cases of the laboratory phenotype.[12] Similarly, microarrays—a high throughput transcriptomic profiling tool, have been used to distinguish between *Escherichia coli* and other pathogenic strains containing virulence factors.[13] According to Chizhikov *et al*[13] the efficiency of the discrimination among foodborne pathogens when microarray was used in combination with PCR technique was able to detect 15 virulence factors in strains of *Shigella, Salmonella and E. Coli* which unlike gel electrophoretic analysis of multiplexed PCR products, produced results that were indistinct. While microarray technology was quite popular and useful for automated detection of pathogens in the last two decade, the science and technology of genomics have since advanced and there are now more powerful next-generation tools with lower sequencing cost.

DETECTING HGT TO IDENTIFY PATHOGENIC AND NON-PATHOGENIC CHANGES IN THE VIRULENCE PHENOTYPE

HGT is the transmission of fragments of genetic information from diverse evolutionary history between organisms through horizontal inheritance.[14] It can be detected via computational analysis and it provide clues to identifying pathogenic and non-pathogenic changes in virulent phenotype. HGT involves the transfer of genetic material either between related (eg, eukaryote to eukaryote) or different organisms or species (eg, prokaryote to eukaryote)[15]; it is a process that can bring new genes and genotypes of different lineages into a genome leading to the emergence of new pathogenic clones. While HGT is rare in eukaryotes, it is however a mechanism of phenotypic innovation[14] that leads to variations in environmental adaptation, antibiotic resistance and the spread of virulence factor encoding genes.[15 16] Before now, genes were thought to be nontransferable, but it has been discovered that with appropriate signal and host environment they can be transmissible in vivo,[16] which points to possible mechanism avirulent strains acquire pathogenicity. A study by Memisević *et al*[17] to identify genes that could differentiate between non-pathogenic and pathogenic strains, and to search for proteins with virulence factors in glanders using comparative genomics found that three proteins that were previously not considered virulent had in-vivo virulence phenotype. Also, de novo microbial genome assembly has been used to characterise exchange of virulence factors between different species and strain, such as the exchange of cytolytic toxin between *S. aureus* sequence type 22 and *E. coli* serotype O104:H4.[17] Consequently, these examples show genomics as an important tool in detecting changes in pathogenicity, HGT, and identifying virulence phenotype.

The major driver for pathogenic emergence is anthropogenic alteration that occurs when human host and animals move into unnatural areas, increasing the probability of contact with pathogens shared by both humans and other mammals.[5 18] The pathogens are transmitted to the host by a vector (eg, an arthropod) usually through host contact with the vector. RNA viruses and bacteria are the popular vectors because they make up approximately 70% of about 300 diseases that have afflicted humans over the last 50 years and are easily transmissible via air, food, and close contact.[5] Bacterial pathogens harbour chromosomal pathogenicity islands or mobile and integrative genetic elements such as virulence plasmids that encode many virulence determinants.[19] While the common origin of these gene products suggest horizontal gene acquisition for adaptation to new niches, they are not self-transmissible under lab conditions as their transfers are aided by local host environment.[16] Moreover, their transformation, conjugation or transduction between strains may only be observed in vivo.[16] Understanding the conditions that necessitate HGT and virulence phenotype will foster opportunities for therapeutics, surveillance of pathogenic changes, as well as interventions that silences virulence expression, and solutions that potentially foster non-pathogenic changes.[16]

Determining the degree of virulence of a pathogen can be useful in estimating the lethality of its disease-causing agent and in comparing different bacteria and their relative virulence[3] in a bid to finding protective intervention strategies. For example, the growth of some bacteria is linked with competition for iron

in human host environment, which is not readily available, so they must strip iron from the host by synthesising siderophores to compete for transferrin-bound iron.[3] Consequently, interventions that deny available iron to these bacteria may be a biodefense approach that halts their multiplicity.

HOW GENOMICS CAN BE APPLIED TO PROACTIVELY PREVENT AN OUTBREAK?

A proactive strategy for preventing infectious disease outbreak is understanding the processes and causes of pathogen transfers by implementing genomic surveillance interception system.[5] An interception system would consist of a catalogue of diagnostic information on pathogens, their environment, their diverse communities and interactions, and the degree of human susceptibility to their attacks. This system will be strengthened by genome sequencing, which can generate and characterise complete libraries of primary pathogenic agents that are harmful to humans, thus providing necessary diagnostic data that enables rapid screening for current and emerging pathogens, and in identifying their primary host. A complete genomic library of sequenced pathogens of many microorganisms will provide a comprehensive database that can be queried when there are indications of a disease outbreak.[5] This will invariably reduce the downtime needed to identify and curtail future similar and unexpected outbreak. For example, real-time genomic analysis of Ebola virus which was made available just around the peak outbreak suggested that it arose from singular insertion into humans, and that there were survivor and sexual transmissions which accelerated the spread.[20]

Current genomic approach geared towards characterising and preventing disease outbreak includes ongoing research by the iBOL investigating DNA barcoding as standard for specie identification and eukaryote diversity using deep sequencing for PCR products to analyse genetic variation (amplicon sequencing). BIOSCAN a new iBOL research programme is aiming to discover and reveal species interaction using sequence data and multiple gene scans to identify and assemble specimens for analysis.[21] The EBP is another ongoing project that involves a consortium of 32 institutions using advanced technology to sequence, characterise and catalogue all known eukaryote genome including the basis of their pathogenic susceptibility.[22] These large-scale genomic projects are vital because it is suggested that only 0.2% of the genome of eukaryote species have been sequenced,[5] and also believed that by sequencing and analysing genomes across eukaryotic tree, the origins and evolution of host-pathogen associations will be clarified. Additionally, uniting all large-scale genomic projects and consortiums into an international genomic collaborative effort will help coordinate a global interception system which currently do not exist. This will necessitate systems of protection such as the restriction of wildlife farming and trafficking, and monitoring of new and re-emerging pathogens and their persistence in the environment, to prevent new pathogenic strain that could be harmful to humans.[5 23]

Since pathogens can be transmitted to humans via food, WGS can be leveraged in food industrial processing and delivery outlets to obtain detailed data of possible pathogens that may be detected from routine supplies; so that by comparing and screening genome data of new and emerging pathogens with publicly available data using bioinformatics applications such as GenomeTrakr, known pathogenic strains are likely to be unpacked before they spread and overwhelm public health measures.[23] An example of previous successful implementation was the recurring salmonella outbreak linked with tomatoes in 2009 that lead to subsequent sequencing and archiving of the pathogens detected from environmental samples by the US Food and Drug Administration (FDA) in the GenomeTrakr database. When there was another salmonella outbreak in some restaurants in Washington in 2010, WGS of the pathogen showed close similarity with those previously archived in GenomeTrakr database. This information was used in investigations that matched shipping records to the city's restaurant chains leading to successful public health curtailing of the foodborne pathogen.[23] Additionally, retrospective studies conducted in Germany and United Kingdom using WGS to identify virulent tuberculosis transmission was able to reassemble the infection timeline, identify superspreaders and determine transmission direction.[24 25] These examples to mention a few, show genomics as an invaluable resource in curtailing biothreats because it can help in identifying the path of disease transmission, the expression, interaction, characteristics, and homology of a particular strain,[26] and provide useful clues to investigation seeking to discover if a pathogen was deliberately engineered.

Sequencing non-pathogenic microorganisms especially viruses and bacteria and archiving them in a publicly searchable database is another way of proactively strengthening public health response to infectious disease outbreak. This information can be handy in potentially determining lineage in the event of the emergence of an unknown pathogenic strain, since it has been reported that non-pathogenic strains may evolve and acquire pathogenicity.[14 16] Pathogens collected from the environment and from food are being sequenced and stored in GenomeTrakr by the FDA to aid traceback capabilities in the event of a disease outbreak. Besides, this genome-driven biosurveillance effort also provides data that can be harnessed to predict genes of pathogenic stains that may be resistant to antibiotics, as research by the FDA National Antimicrobial Resistance Monitoring System have shown a relationship between the presence of known resistance gene and antibiotic resistance.[23]

Genomics technology can be used in tandem with other biosurveillance initiatives either as a primary or secondary tool. The Rapid Syndromic Validation Project by the Sandia National Laboratory which transitioned into the Syndrome Reporting Information System (SYRIS) is a database that facilitates rapid communication and early warning of disease outbreaks between healthcare providers, and epidemiologist in the United States. This population health surveillance tool allows physicians to enter demographic information of patients exhibiting signs and symptoms of interest to enable the US Department of Health determine how infectious an emerging disease is, and if it arose from a deliberate or natural attack.[27] Similarly, informational and environmental surveillance are two other popular surveillance systems that have been deployed in the United States to detect accidental or deliberate emergence of biological agents.[28] The Bioagent Autonomous Networked Detector project (BAND) is an example of an environmental surveillance system deployed in high-risk areas to sample the air every 3 hours for pathogens classified by the United States Centers for Disease Control and Prevention (CDC) as Category A bioterrorism agent.[17 28] The CDC classifies pathogens as category A if they can be easily transmitted, or if they result in high mortality and causes public panic that requires special public health action, while those categorised as B results in moderate mortality and are moderately infectious. Pathogens in category C are emerging but could have high mortality rate and be easily bioengineered for mass transmission.[29] Table 1 shows pathogens and their category according to the CDC, as well as their transmission vectors and virulence factors.[29–35] The SYRIS, BAND and CDC system for pathogen

Table 1 Pathogens and their classification

Pathogens	Category	Causative agent	Vector	Some popular virulence factor
Plague (Yersinia pestis)	A	Bacterium	Rodent flea	Endotoxin—lipopolysaccharide (LPS), lipid A with poor Toll-like receptor 4 (TLR4)
Smallpox (Variola major)	A	Virus	Infected human	Endotoxin—SPICE, CKBP-II
Botulism (clostridium botulinum)	A	Bacteria	toxin, bacteria spore	Exotoxin—neurotoxin
Anthrax (Bacillus anthracis)	A	Bacteria	Infected animal	Exotoxin—capsule, lethal toxin, and edema toxin
Tularaemia (Francisella tularensis)	A	Bacteria	Skin contact, tick, contaminated water or aerosols, lab exposure, deer fly bites	Unknown, but proteins such as MglA, IglC, AcpA and MinD have been implicated.
Viral haemorrhagic fever Filovirus (Ebola) and Arenavirus (Lassa virus)	A	Virus	Infected animal—Ebola, multimammate rat—Lassa virus	Ebola: adherence factor (glycoprotein), VP35 Lassa: Glocoprotein, nucleoprotein
Brucellosis (Brucella species)	B	Bacteria	Infected animal	Endotoxin—LPS, T4SS secretion and BvrR/BvrS.
Glanders (Burkholderia mallei)	B	Bacteria	affected animals(eg, horses)	Quorum sensing, adhesion, capsular polysaccharide, actin-based motility, LPS
melioidosis (Burkholderia pseudomallei)	B	Bacteria	Contaminated water	Endotoxin—LPS
Psittacosis (Chlamydia psittaci)	B	Bacteria	Contact with infected birds	Unknown
Q Fever (Coxiella burnetti)	B	Bacteria	Contact with animal faeces, milk, products	Endotoxin—LPS
Ricin toxin (Ricinus communis)	B	Poison, warfare agent	Castor beans	Toxin
Epsilon toxin (clostridium perfringens)	B	Bacteria	Raw meat and poultry	Alpha toxin, kappa toxin
Salmonella	B	Bacteria	Food	Invasion factors—type III secretion systems, T3SS1 and T3SS2
Typhus fever (Rickettsia prowazekii)	B	Bacteria	fleas, lice, chiggers	T4SS secretion, Phospholipase A2
Staphylococcal enterotoxin B	B	Bacteria	Food poisoning	TSST-1
Vibrio cholera	B	Bacteria	Contaminated food or water	Cholera toxin
Eastern equine encephalitis—Alphaviruses	B	Virus	Infected mosquito	Glocoprotein
Nipah virus	C	Virus	Infected animals (bats, pigs) and infected humans.	V and C proteins
Hantavirus	C	Virus	Rodents urine, faeces, saliva	Surface glycoproteins G1 and G2, G1 protein

classification are initiatives that may be integrated with genomics databases to facilitate integrated diagnosis.

As seen in table 1, the most common causative pathogenic agent is bacteria, and the popular transmission vector is via contact with an infected animal. Apart from the pathogens listed in the CDC categories, it is vital that poorly annotated genomes of relatively unknown pathogens that are yet to be included in the bioterrorism agent lists are constantly monitored and assessed for potential risks[4] in association with future outbreaks. This is a reason why the study to sequence coronaviruses found in bats in China and the ongoing study by Brazilian researchers to collect and study viruses present in wild animals at the Fiocruz Institute should be encouraged.[36 37] The study by Latinne *et al*[36] to sequence coronaviruses in bats in China found new 781 coronaviruses after genetic sequence analysis. Some notable investigations where WGS has been applied to characterise a disease outbreak include the study by Gardy *et al*[38] that employed social network analysis and WGS of 32 *Mycobacterium tuberculosis* isolates to decipher differences in isolated bacteria. The study reported that a rise in tuberculosis outbreak in British Columbia, Canada, was caused by two simultaneous and independent events. Similarly, Harris *et al*,[39] Köser *et al*,[40] and Nübel *et al*[41] are other researchers that have described how WGS application was able to distinguish closely related strains of a bacteria (methicillin-resistant staphylococcus aureus) within neonatal intensive care units.

In addition, the Los Alamos National Laboratory in the US developed a tool that uses DNA fragments to fingerprint strains added to the anthrax database for rapid identification of drug resistant strains and for comparison in the event of an attack from new and emerging strains.[27] The large curation and validation of bioterrorism agents by the US Department of Energy Chem-Bio Non-proliferation Programme and the Lawrence Livermore Lab in 2002 for detecting pathogens applied WGS. The result of the biocuration was used in providing biosecurity at the 2002 winter Olympics.[42] At that time, the study encountered and reported problems relating to gene prediction, annotation and alignment. Although bioinformatics methods and databases have since advanced[43] leading to improvements in multiple sequence alignment, determination of gene regulatory networks and prediction of subcellular localisation, many genes are still poorly annotated which represents a challenge for pathogen detection when implementing comparative genomics. However, these examples show how genomics can and have been proactively applied in the identification of virulent bio-threats.

DISRUPTIVE APPLICATION OF GENOMIC TECHNIQUES

Genome engineering and sequencing can be deliberately misused and manipulated to cause higher mortality and infection either by governments as part of their strategy on biological warfare or by criminals seeking to create massive destruction motivated by abstract belief and ideology (bioterrorism), or to seek revenge

and extort money from people in what is termed 'biocrime'.[44] In these biological attacks, virulence factors of pathogenic organisms may be sequenced, engineered and isolated for the ulterior motive of causing harm and spreading infectious diseases. Nevertheless, the successful implementation of biological attacks are rare and this is partly because of prohibitions by the Biological and Toxic weapons convention (BTCW), and the Geneva Protocol.[45] Surprisingly, there have been a significant increase in biological attack since the discovery of virulence abilities of microorganisms and DNA sequencing in the 19th and 20th century. For instance, 74 cases involving the use of biological toxins were listed by the National Consortium for the study of Terrorism between 1990 and 2011.[46] This is a staggering figure compared with the combined total of 27 attacks between 1900 and 1989,[44] showing an uptick in biological attacks and the need for more intentional global regulation and biosurveillance.

Despite the potential duality in the application of genome engineering and sequencing techniques, an interesting angle in its application would be the ability to decipher if an attack was deliberate or accidental.[47] With the current state of the art of genomic sequencing, it is possible to demonstrate or detect whether an organism involved in such event has been genetically modified (GM) from its genetic sequence.[47–50] The creation of a GM organism (GMO) for example involves several steps, including, the selection of the gene/s to be edited, isolated, packaged (suitable vector such as viruses), and inserted into the host genome.[49] Other genetic elements, such as promoter and terminator, are added to select the marker gene. Possible identification of these elements may also confirm the genetic engineering technique used to modify a genome with and without the sequence information.[50] A deliberate attack may be spotted by cloned vectors at restriction sites and other evidence of genetic engineering.[17] WGS can unearth intentional bio-attacks by analysing the homology and phylogenetic tree of a strain to show if there are related strains that suggest the likelihood of natural evolution from a particular ancestor, and if there are evidence to suggest vertical gene transfer or HGT. This technique uses the whole DNA library database, and the generated reads are analysed with bioinformatics tools for the purpose of GM correlation with already available data. Indeed, clustered regularly interspaced short palindromic repeats (CRISPR)-based genome-editing technologies have revolutionised molecular biology by vastly simplifying the process of creating modified organisms which interestingly, can be detected and sequenced from the DNA left behind in soil, or water, in the form of faeces, urine or saliva.[49] Thus, non-invasive environmental DNAs can be used to trace GM sequences and for biosurveillance and biodefence efforts to prevent an outbreak.

Nonetheless, a concerted and harmonised global effort is necessary for effective vigilance. Surprisingly, not all nations are signatories to the BTCW and the Geneva protocol; and while it is expected that countries that are signatories will keep to the terms of agreement, there are no guarantees that the dynamics of international politics will not jeopardise it. Moreover, a BTCW inspection unit would improve monitoring and regulate research, funding, and stockpiling of engineered biological agents that have no acceptable medical, scientific and peaceful application.[44] In addition, international collaborations would ensure that toxins and agents capable of being used for bioattacks are not easily accessible anywhere on the globe to groups known to act without regards for laid down principles. This includes groups like the Aum Shinrikyo sect and Al-Qaida that have been documented to be reportedly seeking to acquire biological agents.[44 51] What is more, the falling cost of genome sequencing in recent years could mean that any skilled individual may carry out sequencing experiments without being attached to a responsible lab. In addition, countries yet to sign the BTCW and Geneva Protocol, as well as countries funding research in biological warfare could be potential targets for illegitimate activities which calls for more international vigilance and consensus support for a global watchdog. One suggestion is the creation of an international body charged with coordinating peaceful use of new and advanced genomic technology for global biosurveillance, compliance and knowledge sharing.

LIMITATIONS OF GENOMICS APPLICATIONS

Although WGS is no longer a time-consuming process, there are other bottlenecks that could limit maximising its usage. The absence of a standard scientific and legal protocol for the acquisition of sample, the critical evaluation of sequenced data quality, as well as accurate data analysis to gain insight from genomic data are some of the limitations.[17] Due to the differing quality assessment in next-generation sequencing technologies, the potential for recording different annotation and result is real, therefore, harmonisation of results from different centres is important. A good example is the Human Microbiome Project that harmonises results of artificial DNA sample from different sequencing machines.[52] More so, while it is suggested that de novo assembly is preferred to reference genome assembly for WGS because new sequence might be divergent from the chosen reference, there is no unique de novo assembler that performs better in all assembly metrics.[52] It is vital that the results from genome sequencing experiments are interpreted in relation with the limitations of the selected sequencing and bioinformatics method.

CONCLUSION

The application of the science and technology of genomics can be used to measure virulence, differentiate changes in pathogenicity, detect HGT and virulence evolution. It can also be used to identify new, emerging and genetically engineered pathogenic strains by observing genomic sequence data and evidence of genetic engineering such as cloned vectors at restriction sites. To leverage and maximise the application of genomics for real-time biothreat diagnostics, a complete genomic library of pathogenic and non-pathogenic agents will create a robust reference assembly that can be used to screen, characterise, track and trace new and existing strains. Despite the lack of consensus on quality control bioinformatics metrics for accurate and harmonised analysis of sequence data, and the lack of standard legal and scientific sample acquisition, the EBP, GVP, iBOL and BIOSCAN are various large-scale research projects leveraging advanced genomic techniques for improved diagnostics. In all, creating a global space where these and other similar projects can be coordinated to work together to lead an international interception system will ensure more timely, proactive, and effective global biosurveillance and regulation.

Acknowledgements The authors would like to acknowledge Michelle Ross (PhD), Senior Partner at Martin, Blanck and Associates for providing useful suggestions to strengthen the manuscript; and to Dario Ghersi (PhD) and Ketewambi Yves Shamavu for reviewing aspects the paper. Finally the authors acknowledges Godwin I Nwadiugwu (late) for supporting this endeavor.

Contributors MN devised the question and concepts, and wrote the manuscript. NM contributed to the bio-engineered organism concepts and reviewed the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Main messages

- Virulent protein targets have been predicted using reverse vaccinology by excluding proteins that are not present on the surface of pathogens. The Vaxign programme is an application that has been deployed in identifying virulence factors by querying protein sequences to find possible targets.
- A complete genomic library of pathogenic and non-pathogenic agents will create a reference assembly to strengthen global interception system for real-time biothreats and bioterrorism diagnostics.
- Distinguishing deliberate and genetically engineered strains from new naturally emerging pathogenic strains can be done by observing annotated and genomic sequence data and cloned vectors at restriction sites.
- Encouraging research sequencing pathogens found in animals and the environment, creating a global space for collaboration will lead to effective regulation, interception and policing of virulent agents.

Key references

- Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* 2018;19:9–20. doi:10.1038/nrg.2017.88
- Gilchrist CA, Turner SD, Riley MF, *et al*. Whole-genome sequencing in outbreak analysis. *Clin Microbiol Rev* 2015;28:541–63. doi:10.1128/CMR.00075-13
- Ervin A, Hultgren A, Rhyne E, *et al*. Sensing dispersal of chemical and biological agents in urban environments. In Voeller JG. eds. Wiley handbook of science and technology for homeland security. Hoboken, NJ, Wiley 2010; 423–34
- FDA. Proactive application of whole genome sequencing technology. 2017. Available: https://www.fda.gov/food/whole-genome-sequencing-wgs-program/proactive-applications-whole-genome-sequencing-technology.
- Valdivia-Granda WA. (2012). Biodefense oriented genomic-based pathogen classification systems: challenges and opportunities. *J Bioterror Biodef* 2012;3:1000113. doi: 10.4172/2157-2526.1000113.

Self assessment questions

1. How can genomics be used to uncover biothreats?
   a. Identification of virulence factors and emerging pathogens.
   b. Promote wildlife trafficking.
   c. Congest biobanks.
2. What can strengthen bioterrorism diagnostics?
   a. Redlining.
   b. Travel bans.
   c. A global interception system.
3. What are the limitations of applying genomic techniques in coordinating international interception systems?
   a. Lack of consensus bioinformatics quality control metrics.
   b. No sequencing technologies.
   c. Government interference.
4. A database previously used in surveilling biothreats?
   a. PubMed.
   b. GenomeTrakr.
   c. Embase.
5. An example of an environmental surveillance system?
   a. FDA.
   b. BAND.
   c. CDC.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

14 Cross AS, Valdivia-Granda WA, Granda WAV, *et al*. What is a virulence factor? Biodefense oriented genomic-based pathogen classification systems: challenges and opportunities. *J Bioterror Biodef* 2008;12:1000113.
2 He Y. Bacterial Whole-Genome Determination and Applications. In: *Molecular medical microbiology*. Elsevier, 2015: 357–68.
3 Peterson JW. Bacterial Pathogenesis. In: Baron S, ed. *Medical microbiology*. 4th ed. Galveston: University of Texas Medical Branch at, 1996. https://www.ncbi.nlm.nih.gov/books/NBK8526/;
4 Valdivia-Granda WA, Granda WAV. Biodefense oriented genomic-based pathogen classification systems: challenges and opportunities. *J Bioterror Biodef* 2012;3:1000113.
5 Kress WJ, Mazet JAK, Hebert PDN. Opinion: intercepting pandemics through genomics. *Proc Natl Acad Sci U S A* 2020;117:13852–5.
6 Medical dictionary. Applied genomics. Available: https://medical-dictionary.thefreedictionary.com/applied+genomics
7 Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. *Immunity* 2010;33:530–41.
8 Bidmos FA, Siris S, Gladstone CA, *et al*. Bacterial vaccine antigen discovery in the reverse vaccinology 2.0 era: progress and challenges. *Front Immunol* 2018;9:2315.
9 Bianconi I, Alcalá-Franco B, Scarselli M, *et al*. Genome-Based Approach Delivers Vaccine Candidates Against *Pseudomonas aeruginosa*. *Front Immunol* 2018;9:3021.
10 He Y, Xiang Z, Mobley HLT. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *Journal of Biomedicine and Biotechnology* 2010;2010:1–15.
11 Azevedo V, D'Afonseca V, *et al*. Reannotation of the Corynebacterium diphtheriae NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria. *Open Access Bioinformatics* 2012;1:1.
12 Mason A, Foster D, Bradley P, *et al*. Accuracy of different bioinformatics methods in detecting antibiotic resistance and virulence factors from Staphylococcus aureus whole-genome sequences. *J Clin Microbiol* 2018;56:e01815–7.
13 Chizhikov V, Rasooly A, Chumakov K, *et al*. Microarray analysis of microbial virulence factors. *Appl Environ Microbiol* 2001;67:3258–63.
14 Deng Y, Xu H, Su Y, *et al*. Horizontal gene transfer contributes to virulence and antibiotic resistance of Vibrio harveyi 345 based on complete genome sequence analysis. *BMC Genomics* 2019;20:761.
15 Sulaiman S, Yusoff NS, Mun NS. Inference of Horizontal Gene Transfer: Gaining Insights Into Evolution via Lateral Acquisition of Genetic Material. In: *Encyclopedia of bioinformatics and computational biology*. Elsevier, 2019: 953–64.
16 Mel SF, Mekalanos JJ. Modulation of horizontal gene transfer in pathogenic bacteria by in vivo signals. *Cell* 1996;87:795–8.
17 Memisević V, Zavaljevski N, Pieper R, *et al*. Novel Burkholderia mallei virulence factors linked to specific host-pathogen protein interactions. *Mol Cell Proteomics* 2013;12:3036–51.
18 Ashford RW, Crewe W. *The parasites of Homo sapiens: an annotated checklist of the protozoa, helminths and arthropods for which we are home*. 2nd ed. CRC Press, 1998.
19 Kado CI. Horizontal gene transfer: sustaining pathogenicity and optimizing host-pathogen interactions. *Mol Plant Pathol* 2009;10:143–50.
20 Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* 2018;19:9–20.
21 Hebert PDN, Hollingsworth PM, Hajibabaei M. From writing to reading the encyclopedia of life. *Philos Trans R Soc Lond B Biol Sci* 2016;371:20150321.
22 Lewin HA, Robinson GE, Kress WJ, *et al*. Earth BioGenome project: sequencing life for the future of life. *Proc Natl Acad Sci U S A* 2018;115:4325–33.
23 FDA. Proactive application of whole genome sequencing technology. Available: https://www.fda.gov/food/whole-genome-sequencing-wgs-program/proactive-applications-whole-genome-sequencing-technology

24 Walker TM, Ip CLC, Harrell RH, *et al*. Whole-Genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013;13:137–46.
25 Roetzer A, Diel R, Kohl TA, *et al*. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 2013;10:e1001387.
26 Nwadiugwu MC. Expression, interaction, and role of pseudogene Adh6-ps1 in cancer and other disease phenotypes. *Bioinform Biol Insights* 2021;15:117793222110405.
27 U.S. Department of Energy (DOE). Human genome program. human genome news. Available: https://web.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v12n1/HGN121_2.pdf
28 edErvin A, Hultgren A, Rhyne E. *Sensing dispersal of chemical and biological agents in urban environments*. Wiley, 2010.
29 CDC. Cdc bioterrorism agents/diseases by category. Available: https://emergency.cdc.gov/agent/agentlist-category.asp
30 Montminy SW, Khan N, McGrath S, *et al*. Virulence factors of Yersinia pestis are overcome by a strong lipopolysaccharide response. *Nat Immunol* 2006;7:1066–73.
31 GŁOWACKA P, ŻAKOWSKA D, NAYLOR K, *et al*. Brucella – virulence factors, pathogenesis and treatment. *Pol J Microbiol* 2018;67:151–61.
32 Webb JR, Sarovich DS, Price EP, *et al*. *Burkholderia pseudomallei* Lipopolysaccharide Genotype Does Not Correlate With Severity or Outcome in Melioidosis: Host Risk Factors Remain the Critical Determinant. *Open Forum Infect Dis* 2019;6:ofz091.
33 VetBact. Chlamydophila psittaci. Available: https://www.vetbact.org/?artid=96
34 Institute of Pathogen Biology. Virulence factors of pathogenic bacteria. Available: http://www.mgc.ac.cn/VFs/contact.htm
35 Muyangwa M, Martynova EV, Khaiboullina SF, *et al*. Hantaviral proteins: structure, functions, and role in hantavirus infection. *Front Microbiol* 2015;6:1326.
36 Latinne A, Hu B, Olival KJ, *et al*. Origin and cross-species transmission of bat coronaviruses in China. *Nat Commun* 2020;11:4235.
37 Larson C, Ghosal A, Sousa SM. Scientist focus on bats for clues to prevent next pandemic. AP news. Available: https://apnews.com/article/pandemics-brazil-rio-de-janeiro-animals-forests-5a7dff4d7ad18209edf4e35e62607087
38 Gardy JL, Johnston JC, Sui SJH, *et al*. Whole-Genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011;364:730–9.
39 Harris SR, Cartwright EJP, Török ME, *et al*. Whole-Genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *Lancet Infect Dis* 2013;13:130–6.
40 Köser CU, Holden MTG, Ellington MJ, *et al*. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012;366:2267–75.
41 Nübel U, Nachtnebel M, Falkenhorst G, *et al*. Mrsa transmission on a neonatal intensive care unit: epidemiological and genome-based phylogenetic analyses. *PLoS One* 2013;8:e54898.
42 Slezak T, Kuczmarski T, Ott L, *et al*. Comparative genomics tools applied to bioterrorism defence. *Brief Bioinform* 2003;4:133–49.
43 Nwadiugwu MC. Gene-Based clustering algorithms: comparison between Denclue, Fuzzy-C, and birch. *Bioinform Biol Insights* 2020;14:117793222090985.
44 Jansen HJ, Breeveld FJ, Stijnis C, *et al*. Biological warfare, bioterrorism, and biocrime. *Clin Microbiol Infect* 2014;20:488–96.
45 Organisation for the Prohibition of Chemical Weapons (OPCW). Protocol for the Prohibition of the use in war of Asphyiating poisonous or other gases and of bacteriological methods of warfare, 1925. Available: http://www.opcw.org
46 JMAG PL. Ricin letter mailed to Presidient and Senator. National Consortium for the study of terrorism and responses to terrorism. Available: https://www.start.umd.edu/sites/default/files/files/publications/br/STARTBackgroundReport_RicinMailings_April2013.pdf
47 Lewis G, Jordan JL, Relman DA, *et al*. The biosecurity benefits of genetic engineering attribution. *Nat Commun* 2020;11:6294.
48 Xu CCY, Ramsay C, Cowan M, *et al*. Transgenes of genetically modified animals detected non-invasively via environmental DNA. *PLoS One* 2021;16:e0249439.
49 Sharma A, Gupta G, Ahmad T. Next generation agents (synthetic agents): Emerging threats and challenges in detection, protection, and decontamination. In: *Handbook on biological warfare preparedness*. Elsevier, 2020: 217–56.
50 Salisu IB, Shahid AA, Yaqoob A, *et al*. Molecular approaches for high throughput detection and quantification of genetically modified crops: a review. *Front Plant Sci* 2017;8:1670.
51 Milton L. Assessing the biological weapons and bioterrorism threat. Diane publishing, 2005. Available: https://armscontrolcenter.org/wp-content/uploads/2016/02/M-Leitenbergs-full-Army-War-College-Book.pdf;
52 Gilchrist CA, Turner SD, Riley MF, *et al*. Whole-Genome sequencing in outbreak analysis. *Clin Microbiol Rev* 2015;28:541–63.

Answers

1. a
2. c
3. a
4. b
5. b