IntechOpen

Genotyping

Edited by Ibrokhim Abdurakhmonov

GENOTYPING

Edited by **Ibrokhim Abdurakhmonov**

Genotyping
http://dx.doi.org/10.5772/intechopen.71242
Edited by Ibrokhim Abdurakhmonov

Contributors

Peng Li, Keisuke Tanaka, Rumi Ohtake, Saki Yoshida, Takashi Shinohara, Peter Hristov, Rositsa Shumkova, Ani Georgieva, Daniela Sirakova, Boyko Neov, Georgi Radoslavov, Yuny Erwanto, Jean Bernard Lekana-Douki, Larson Amédée Boundenga, Imran Shahid, Munjed Ibrahim, Muhammad Usman Nawaz, Mohammad Imam, Waleed Almalki, Mohammed AlRabia, Yu Yong-Xin, Mohd Nasir Mohd Desa, Zarizal Suhaili, Mazen Al-Obaidi, Zhengfeng Wang, Se-Ping Dai, Ju-Yu Lian, Hong-Feng Chen, Wan-Hui Ye, Hong-Lin Cao

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2018 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, The Shard, 25th floor, 32 London Bridge Street
London, SE19SG – United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard copies can be obtained from orders@intechopen.com

Genotyping, Edited by Ibrokhim Abdurakhmonov
p. cm.
Print ISBN 978-1-78923-280-6
Online ISBN 978-1-78923-281-3

Meet the editor

Ibrokhim Y. Abdurakhmonov received his B.Sc degree (1997) in *Biotechnology* from the National University of Uzbekistan, M.Sc degree in *Plant Breeding* (2001) from Texas A&M University of the USA, PhD degree (2002) in *Molecular Genetics*, Doctor of Science degree (2009) *in Genetics*, and full professorship (2011) in *Molecular Genetics and Molecular Biotechnology* from the Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan. He founded (in 2012) and is currently leading the Center of Genomics and Bioinformatics of Uzbekistan. He serves as an associate editor/editorial board member of several international and national journals on plant sciences. He received Government award 2010 chest badge "Sign of Uzbekistan" 2010 TWAS Prize and "ICAC Cotton Researcher of the Year 2013" for his outstanding contribution to cotton genomics and biotechnology. He was elected as The World Academy of Sciences (TWAS) Fellow (2014) on *Agricultural Science* and as a co-chair/chair of "Comparative Genomics and Bioinformatics" workgroup (2015) of International Cotton Genome Initiative (ICGI). He was elected (2017) as a member and vice-president of the Academy of Sciences of Uzbekistan. He was appointed (2017) as the Minister of Innovational Development of Uzbekistan.

Contents

Preface

Genotyping is the methodological process of detecting the allelic content of loci in a given genome, so-called genotype, which helps to reveal differences between two individuals of comparisons or allelic frequencies among individuals of a population. Genotyping is overwhelmingly based on comparative assessment of DNA sequences of targeted individuals using a variety of existing molecular marker technologies based on size differentiation and fluorescence detection in single or multiple marker combinations.

Any individual, tissue or fossil material with available DNA samples can be genotyped and differentiated. Genotyping is the key step toward the analyses of genetics, evolution, relatedness, diversity, differentiation and divergence of individuals or populations, tagging of important loci/genes to distinct phenotypes or disease, helping in diagnostics and treatment of diseases in medicine and veterinary or molecular breeding of agricultural crops, barcoding of unique biological materials, solving forensics issues, and/or help in controlling the spread of pathogens by tracing the origin of outbreaks.

Genotyping procedures and tools have been evolved in concordance with scientific discoveries in understanding of the structure and function of DNA molecules and technological advances on analyzing and screening a genetic material. Their history include the early-time blot-hybridization and restriction enzyme-based fragment size simplex loci assays and revolutionizing polymerase chain reaction-based detection tools with its extension to high-throughput fluorescent-dye-based multiplexing methods. The emergence of next generation sequencing and novel chemistry technologies further advanced genotyping methods from partial to whole-genome genotyping in the era of genomics and bioinformatics science development.

This *Genotyping* book intends to provide the reader with an overview of the basics of genotyping process, available approaches and protocols, as well as novel, low-cost, high-throughput whole-genome genotyping tools and genotype data handling with the examples of genotyping applications in some organisms.

Here we successfully compiled nine chapters covering updated topics of genotyping in many organisms using various molecular markers helping to distinguish genotypes, viral or bacterial strains, pathological or infected conditions, etc. These chapters, describing the lasted advancements in this area, should be a useful addendum to current literature on genotyping, which is helpful for scientists, students and readers of life science direction.

I greatly acknowledge the efforts of all the authors of the book chapters and thank them for their valuable contributions. I also appreciate the IntechOpen book department for giving me the opportunity to work on this book project and Ms. Danijela Sakic, Author Service Manager at IntechOpen, for her help and support during editorial process.

Ibrokhim Y. Abdurakhmonov
Center of Genomics and Bioinformatics
Academy of Sciences of Uzbekistan
Tashkent, Uzbekistan

Section 1

Genotyping Markers and Methods

Chapter 1

Explore the Novel Biomarkers through Next-Generation Sequencing

Peng Li

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73205

Abstract

Next-generation sequencing is being a robust technology for the practice of clinical diagnosis. The reason is that this technology offers the advantage of higher sensitivity and the potential to detect the full genome sequence of pathogens, including unknown pathogen species. In view of the exceptional advantages of next-generation sequencing, the technology can be used to improve and revolutionize conventional pathogenic detection methodologies. The technological result holds great possibilities in helping to support clinicians with richer insights into host's genomic features, including the appropriateness to drug resistance based on the sequencing mapping of the microorganism. Besides, the technology will help discover the source of infection and insights into treatment directions, furthermore lead to the advancements in diagnosis. Eventually, this technology will benefit the clinical community in infectious disease prediction and prevention.

Keywords: pathogens, clinical disease, diagnosis, next-generation sequencing, disease biomarker

1. Introduction

Bacteria and virus consist of the most of the microbial pathogens. The microbial pathogens are responsible for infectious diseases causes in human hosts [1]. The microbial infections can cause the serious clinical symptoms in the human host, such as the inflammation, fever, pain and septic shock. They can even lead to the host death if the patient is treated with a delay. Thus, the early and accurate identification of pathogen is very important in clinical practice [2], as the proper antimicrobial treatment can be used to prevent the infection effectively. However, the conventional diagnostics, like polymerase china reaction (PCR), enzyme-linked immunosorbent assay (ELISA) and microbial cell culture, are lack of the ability for the

Application	References
Biodiversity	[7]
Disease diagnostics	[8]
Nutrition	[9]
Environment	[10]
Aging disease	[11]
Forensic biology	[12]
Agriculture	[13]

Table 1. The application of NGS in biotechnology.

unknown or the high-mutated pathogens detection. Therefore, the novel pathogen diagnostics is necessary for the clinical healthcare.

Next-generation sequencing (NGS) is the latest scientific technology till date to sequence the target gene or genome [3]. NGS technology refers to one high-throughput DNA sequencing method. In a single experiment, it can determine the sequence of the target gene or full genome with a total size of larger than millions of base pairs (bp) [4]. Sequencing thousands of genes or even genomes in one experiment is consequently made possible using this NGS technology.

Because of its robustness, NGS is widely being applied in biotechnologies, such as in forensic biology, plant science and environmental contamination, etc., (**Table 1**). For example, the genome of *Mycobacterium tuberculosis* had been determined by Genome Analyzer to study the pathogen epidemiology [5]. In 2013, US Food and Drug Administration (FDA) has cleared Illumina MiseqDx to be the first *in vitro* NGS diagnosis platform [6]. In the recent years, with more development in this technology, NGS will provide more comprehensive information for the clinicians in clinical studies. In particular, there is more potential to translate currently available NGS technology into the pathogen detection.

2. The pathogen mutation is a challenge for infectious disease diagnosis

The microorganism pathogens are mostly responsible for the infection disease in the human body. PCR, ELISA and cell culture are the conventional methods for the pathogens detection [14]. However in PCR, one pair of primers, including forwards and reverse primers are compulsory to design according to the target gene or genome spectrum. Meanwhile in ELISA, the functional antibodies are indispensable for the microorganism antigen detection. Finally for cell culture, the related culture medium is also required for the according microorganisms. Thus, these conventional methods will not be working for some unknown or high-mutated microorganism, as the species sequencing information is deficient.

Furthermore, the pathogens, like virus and bacteria, have a high mutation rate which makes microorganism gene or proteins easily mutated [15]. The mutated gene would lead to the dramatic change of the protein structure [16]. The changed protein may be not be detected by the original antibody, as a result, it will be challenging for the identification of these pathogens using the conventional methods. Therefore, there is a limitation for the conventional methods to detect the unknown and high-mutated pathogenic species. These mutated species will lead to the drug resistance and other clinical problems. The problems will also compromise the success of current antimicrobial treatment, leading to further increase in pathogen infection incidence and host mortality. Hence, the more powerful sequencing technology is urgently needed in the clinic community.

3. The development of next-generation sequencing technology

The first method of sequencing DNA was developed by Sanger [17]. He first devised a method that allowed for the determination of small sequences in ribonucleic acids. The Sanger sequencing method was developed here using the chain termination technology. Because of his development, Sanger was awarded the Nobel Prize in 1980 [18]. At the beginning stage, the 3′ end of the primer anneals closely to the target DNA sequence. The addition of nucleotides on the 3′ end of the template could either be a usual unlabeled nucleotide or a fluorescently labeled nucleotide. After that, no more nucleotides are able to bind to that particular strand of DNA sequence. This happens to all the templates that have been loaded into the Sanger sequencing platform and denatured into single strand DNA, leading to various single strands with different lengths and with diverse labeled nucleotides. As soon as the sequencing has ended, the finished sample would be denatured once again to remove the sequenced strands from the original loaded DNA. The sequenced strands would then be inputted into an Agarose gel for observing under a UV light. While the sample is undergoing Agarose gel running, it can be noted that the varying lengths of sequenced DNA would run at diverse times across the Agarose electrophoresis. The longer fragments are able to travel much slower than their shorter counterparts. The bands that will be then analyzed at the end of the gel. The result will be arranged according to their fragment size and labeling, therefore giving a visual image of the base sequence in the inputted sample. Sanger sequencing had been regarded as a gold standard metric because of its high accuracy. The method of the Sanger sequencing had been applied to sequence the first *Homo sapiens* genome [19].

However, even there are technological advances in the Sanger sequencing method such as the automation, the Sanger sequencing was still time-consuming and very costly. The growing interest in the sequencing of the personal genomes fueled the development of new robust technology. The NGS technology has been introduced currently. NGS has many functions, such as the sequencing of an entire genome, deep sequencing for a target region of the genome, or even multiplex sequencing which allows many samples to be sequenced at one time. In the NGS workflow (**Figure 1**), the multiplex sequencing function is utilized so that various samples can be sequenced simultaneously. The principle behind NGS platform is alike to that of Sanger sequencing. The signals produced from fluorescently or radioactively

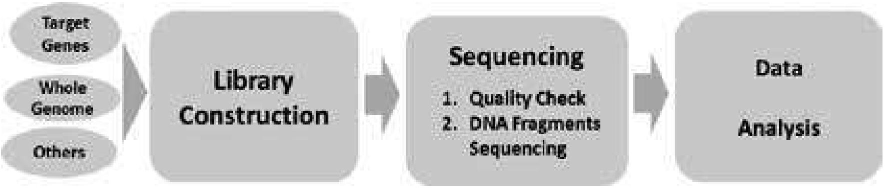


Figure 1. The NGS workflow consists of library construction, sequencing and data analysis.

labeled nucleotides are received, allowing the bases of the template DNA sequence to be read in order. But the difference between the two sequencing technologies is that NGS has the competence to handle numerous sequencing reactions simultaneously. Many templates of DNA are able to be processed at the same time, which makes the whole genome sequencing time to be much more rapid with this robustness. Thus, NGS will be the next important sequencing tool used in biological and clinical samples as it offers a super speed with a higher accuracy.

Over these years, NGS technologies have matured and thus lowered cost and dramatically increased throughput. This technology eliminates the time-consuming and labor-intensive step to generate single clones via bacterial cloning and gel electrophoresis, and using the parallel processing to simultaneously sequence a large number of DNA sample. Thus, instead of generating hundreds of longer reads (more than 1000 bp), NGS technology produces millions of shorter reads (100 ~ 600 bp) ranging on the order of gigabases per run (**Table 2**). Consequently, the major work of sequencing has shifted from the benchtop to the desktop. For example, a nanopore NGS instrument MinION had been developed and applied for pathogens detection [20, 21]. Moreover the analysis of the NGS data presents new advancements mainly due to the short read lengths and require significant investment in big data processing, including hardware, software and bioinformatics. Therefore, the high-throughput data is easily to be handled. As a result, what once took 1 week to sequence using Sanger sequencing can now be accomplished in a matter of days in a desktop NGS platform.

Superior to the traditional sequencing method, NGS is also able to sequence unknown DNA sequence. The unknown genome sequence can be *de novo* assembled. Depending on the platforms

	Sanger sequencing	NGS
Experimental time	~1 week	~2 days
Cost per sample	Expensive	More cost-effective
Reads per sample	One read	Up to Millions
Reading length	Long reading length	Short reading length
Cloning vector required	Yes	No
Specific primer design	Yes	No
Gel electrophorese required	Yes	No

Table 2. Comparison of Sanger sequencing and NGS.

used, NGS can sequence from tens of thousands to more than a billion molecules in a single sequencing running. And it is independent of the known microorganism sequence. In addition, NGS obtains this feature because of its ability to sequence thousands of the inputted sample, in a parallel style, rather than sequencing a single DNA template (**Figure 1**). This particularly parallel analysis is achieved by the miniaturization of the volume of the individual sequencing reaction, which limits the size of the instrument and reduces the cost of reagents per reaction.

4. Identification of the novel biomarkers through NGS

NGS technology proves to be a cost-effective, rapid, yet a highly sensitive method to sequence large amounts of DNA at once. This can enhance infectious disease research which may eventually lead to new biomarkers discoveries. These discoveries can be translated into new diagnostic, prognostic and therapeutic targets. In previous studies, NGS technology was used to detect the common mutation in viral samples from infected patients. Across the various samples analyzed, two common viral mutations were identified in all of the samples. A silent mutation and a missense mutation was detected. These common mutations identified code for viral reverse transcriptase subunits p66 and p51. As reverse transcriptase is extremely important for the survival of the virus, the common mutations identified are possible novel biomarkers for virus among local strains. With the identification of these novel biomarkers, it would serve to improve diagnosis as well as treatment.

In the particular study for pathogen diagnosis, NGS had allowed for the sequencing and identification of even the smallest variants of the viral or bacterial genome. The NGS results were able to illustrate the various base mutations that occurred across all multiple samples provided. Therefore, NGS has the robust advantage over conventional diagnostics of having higher sensitivity, especially about low-frequent mutation or variants. This could be the reason that the traditional sequencing method counts on a given position when sequencing a determined DNA base. A minor mutation or variants will possibly have a low signal-to-noise ratio that is unclear from the background noise. But for NGS, it makes use of complete coverage over the full gene or the whole genome which provides a much higher sensitivity regarding minor mutant or variant. From these results, perhaps these mutations would prove to have a rather significant impact on the drug-resistant capabilities of the virus. Further studies would be also extended to other viral or bacterial genome research. These will provide the characteristics of the microbial pathogens and the disease transmission pathways.

The common mutations identified at virus are located on the pol gene that codes for reverse transcriptase subunits [22]. Unlike other silent mutations, the missense mutation has more implications. As a result, there is an alteration in the corresponding amino acid from leucine to phenylalanine. Aligned with the reference genome, the protein function could be altered because of the missense mutation. The missense mutation would result in changes to the protein structure of reverse transcriptase, causing the conformational changes. These changes to the protein have potential to cause resistance to antimicrobial drugs, allowing the virus to continue developing in the host.

Consequently drug resistance remains a challenge for the treatment of pathogen infection. It arises from the pathogen's ability to mutate rapidly. The infected patients can initially be infected with a drug-resistant virus or develop drug resistance after starting therapy. Studies have been conducted to identify the mutations due to the resistance to antiviral medicines. More than 50 and 40 reverse transcriptase mutations have been found to be associated with nucleoside reverse transcriptase inhibitor (NRTI) and non-nucleoside reverse transcriptase inhibitor (NNRTI) drugs respectively [23]. The viral reverse transcriptase is highly important as it catalyzes the conversion of single-stranded RNA to double-stranded viral DNA for integration into the host genome. This enzyme plays an important role in the life cycle of the virus and has been a good target for the development of antiretroviral drugs for the treatment of pathogens. Currently, there are two broad classes of drugs that target the viral reverse transcriptase; NRTI (like stavudine and emtricitabine) and NNRTI (like nevirapine and etravirine) [24].

The ability of the virus to mutate at the specific position may be due to the selective pressure. There is a high possibility that these mutations can be commonly observed in viral strains. The development of such mutations would allow for the survival and continuity of the local subtype virus. Hence, the identification of these common mutations could be used as novel biomarkers for the diagnosis. Studies have shown similar mutations identified in local viral strains. This further certifies that the mutation found is likely to be common in viral strains. Thus, the common mutations found through NGS technology can be applied as diagnostic biomarkers. Furthermore, it can be developed into the potential marker for the future drugs to defense against pathogens. This can significantly help the microbiological laboratories in large-scale studies of the virus, which aims to aid in the clinical management of pathogen infection [25]. Nevertheless, further studies should be done with a larger sample size to confirm if the identified common mutation is still observed in a larger population.

Compared with microbial culture, NGS technology is a culture-free detection methodology. Metagenomics sequencing will provide a fast, reliable tool for a rapid microbial diagnosis. The conservation region of 16 s ribosomal ribonucleic acid (rRNA) amplicons can be sequenced by a standard workflow. Thus, this methodology can identify hundreds or even thousands species at one time. The microbial system biology can be also investigated at the same time. The automated software or pipeline will help the metagenomics to be a standard microbial detection method.

The molecular epidemiology studies can be investigated deeply, using NGS results. And the investigation of subtype differences in clinical phenotypes and treatment outcomes can be achieved. It is fully predictable that NGS will be much better than the often used conventional diagnosis methods, for providing sequence information for the genomic, microbiological and clinical studies. The current standard of diagnosis methods, commercially available PCR, ELISA, cell culture or other assays target short sub-gene fragments for drug resistance determination. Furthermore, it is interesting to highlight that mutations outside these regions can influence drug resistance [26]. Additionally, NGS is more powerful to study of drug resistance, and disease transmission [27]. There are many emerging multi-drug resistant organisms globally, thus it is significant to investigate the molecular microbiology of the new pathogens.

Antimicrobial therapy is widely performed as a form of treatment in fighting against pathogen infection. Unfortunately, some microorganism has the ability to rapidly mutate which results in changes in its genetic or protein structure. This provides pathogens with the potential to develop resistance to existing antimicrobial drug or treatment. NGS can provide the solution of sequencing a pathogenic gene/genome and the identification of a common mutation in the targeted region. The decrease in cost and increase in accuracy, resolution and reproducibility of NGS allows large-scale sequencing of the virus to be performed efficiently. The study made use of NGS platform together with the usage of the library preparation to sequencing the mutated pathogenic samples. The advancement of NGS has brought about many benefits to the field of biological sciences and will continue to play a big role in the disease diagnostics.

5. Future perspectives

By using NGS, the advantages of this technology was able to be observed throughout the duration of the experiments. One of the really good advantages includes the deep sequencing protocol that occurs during NGS. Deep sequencing is the process of sequencing the same region several times, from hundreds to ten of thousands times coverage. When amplicons are able to be sequenced at a really high depth of coverage, the sequence mutations can be highlighted. It allows for the detection of multiple variants that are really low in number within its population. Somatic mutations that cannot be identified easily using the Sanger sequencing can be easily sequenced, making rare infectious diseases easier to study in a clinical environment.

However, the main NGS platforms used has a limited length of the sequence generated in individual reactions. The read length for the majority of platforms is in the range of hundreds of base pairs. In order to sequence DNA longer than the feasible read length, the material need fragmented before analysis. Following sequencing, the reads are analyzed through informatics to provide the information on the sequence of the whole target molecule. The short read, particularly the high-throughput sequencing methodology used in NGS was a different solution that developed sequencing competence. It enables population-scale sequencing and establishes a foundation for the novel genomic medicine as part of healthcare. NGS technologies are increasingly used for diagnosis and monitoring of infectious diseases such as virus infection. NGS is more powerful than other methods, such as Sanger sequencing, especially in the improved accuracy of the unkown region. The new technology is less costly and is more capable to detect the repeated fragments. Although the usage of NGS still has its setbacks, such as the relatively expensive price of the sequencing consumables required to conduct one sequencing run, perhaps in the near future the overall cost of NGS will be reduced; with increased popularity of this sensational sequencing method, NGS will eventually be as a cheap and standard method in biomarker discovery. Also, an adapted genotyping prediction informatics could be developed based on data acquired from whole-genome sequences of drug-resistant isolates. The predicted novel treatment resistance conferring mutations would be validated against phenotypic assay as well as clinical data acquired from the patients. The correlative application between the new solution and the conventional methods, such as PCR would be also determined together to assess the performance of the drugs.

Abbreviations

bp	Base pairs
ELISA	Enzyme-linked immunosorbent assay
NGS	Next-generation sequencing
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	Nucleoside reverse transcriptase inhibitor
PCR	Polymerase chain reaction
rRNA	ribosomal ribonucleic acid

Author details

Peng Li

Address all correspondence to: li_peng@sp.edu.sg

Singapore Polytechnic, Singapore

References

[1] Gomez-Diaz E, Jorda M, Peinado MA, Rivero A. Epigenetics of host-pathogen interactions: The road ahead and the road behind. PLoS Pathogens. 2012;**8**:e1003007. DOI: 10.1371/journal.ppat.1003007

[2] Dark PM, Dean P, Warhurst G. Bench-to-bedside review: The promise of rapid infection diagnosis during sepsis using polymerase chain reaction-based pathogen detection. Critical Care. 2009;**13**:217. DOI: 10.1186/cc7886

[3] van Vliet AH. Next generation sequencing of microbial transcriptomes: Challenges and opportunities. FEMS Microbiology Letters. 2010;**302**:1-7. DOI: 10.1111/j.1574-6968.2009.01767.x

[4] Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: From basic research to diagnostics. Clinical Chemistry. 2009;**55**:641-658. DOI: 10.1373/clinchem.2008.112789

[5] Kato-Maeda M et al. Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS One. 2013;**8**:e58235. DOI: 10.1371/journal.pone.0058235

[6] FDA-approved next-generation sequencing system could expand clinical genomic testing: Experts predict MiSeqDx system will make genetic testing more affordable for smaller labs. American Journal of Medical Genetics: Part A. 2014;**164A**:x-xi. DOI: 10.1002/ajmg.a.36461

[7] Huete-Perez JA, Quezada F. Genomic approaches in marine biodiversity and aquaculture. Biological Research. 2013;**46**:353-361. DOI: 10.4067/S0716-97602013000400007

[8] Lecuit M, Eloit M. The potential of whole genome NGS for infectious disease diagnosis. Expert Review of Molecular Diagnostics. 2015;**15**:1517-1519. DOI: 10.1586/14737159.2015.1111140

[9] Liu GE. Applications and case studies of the next-generation sequencing technologies in food, nutrition and agriculture. Recent Patents on Food, Nutrition & Agriculture. 2009;**1**:75-79. DOI: 10.2174/2212798410901010075

[10] Wong K, Fong TT, Bibby K, Molina M. Application of enteric viruses for fecal pollution source tracking in environmental waters. Environment International. 2013;**45**:151-164. DOI: 10.1016/j.envint.2012.02.009

[11] Yang HJ, Ratnapriya R, Cogliati T, Kim JW, Swaroop A. Vision from next generation sequencing: Multi-dimensional genome-wide analysis for producing gene regulatory networks underlying retinal development, aging and disease. Progress in Retinal and Eye Research. 2015;**46**:1-30. DOI: 10.1016/j.preteyeres.2015.01.005

[12] Yang Y, Xie B, Yan J. Application of next-generation sequencing technology in forensic science. Genomics, Proteomics & Bioinformatics. 2014;**12**:190-197. DOI: 10.1016/j.gpb.2014.09.001

[13] Bansal U, Bariana H. Advances in identification and mapping of rust resistance genes in wheat. Methods in Molecular Biology. 2017;**1659**:151-162. DOI: 10.1007/978-1-4939-7249-4_13

[14] Priyanka B, Patil RK, Dwarakanath S. A review on detection methods used for foodborne pathogens. The Indian Journal of Medical Research. 2016;**144**:327-338. DOI: 10.4103/0971-5916.198677

[15] Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. Journal of Virology. 2010;**84**:9733-9748. DOI: 10.1128/JVI.00694-10

[16] Hanna N. Advances in the treatment of second-line non-small-cell lung cancer. Lung Cancer. 2005;**50**(Suppl 1):S15-S17. DOI: 10.1016/S0169-5002(05)81554-5

[17] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. 1975;**94**:441-448. DOI: 10.1016/0022-2836(75)90213-2

[18] Sanger F. The early days of DNA sequences. Nature Medicine. 2001;**7**:267-268. DOI: 10.1038/85389

[19] Tsiatis AC et al. Comparison of sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. Journal of Molecular Diagnostics. 2010;**12**:425-432. DOI: 10.2353/jmoldx.2010.090188

[20] Kilianski A et al. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. Gigascience. 2015;**4**:12. DOI: 10.1186/s13742-015-0051-z

[21] Bates M, Polepole P, Kapata N, Loose M, O'Grady J. Application of highly portable MinION nanopore sequencing technology for the monitoring of nosocomial tuberculosis infection. International Journal of Mycobacteriology. 2016;**5**(Suppl 1):S24. DOI: 10.1016/j.ijmyco.2016.10.035

[22] Shaw WH et al. Identification of HIV mutation as diagnostic biomarker through next generation sequencing. Journal of Clinical and Diagnostic Research. 2016;**10**:DC04-DC08. DOI: 10.7860/JCDR/2016/19760.8140

[23] Shafer RW, Schapiro JM. HIV-1 drug resistance mutations: An updated framework for the second decade of HAART. AIDS Reviews. 2008;**102**:67-84

[24] Arts EJ, Hazuda DJ. HIV-1 antiretroviral drug therapy. Cold Spring Harbor Perspectives in Medicine. 2012;**2**:a007161. DOI: 10.1101/cshperspect.a007161

[25] Gall A et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. Journal of Clinical Microbiology. 2012;**50**:3838-3844. DOI: 10.1128/JCM.01516-12

[26] Delviks-Frankenberry KA, Nikolenko GN, Pathak VK. The "connection" between HIV drug resistance and RNase H. Virus. 2010;**2**:1476-1503. DOI: 10.3390/v2071476

[27] Islam MA et al. Emergence of multidrug-resistant NDM-1-producing Gram-negative bacteria in Bangladesh. European Journal of Clinical Microbiology & Infectious Diseases. 2012;**31**:2593-2600. DOI: 10.1007/s10096-012-1601-2

Chapter 2

Microsatellite Capture Sequencing

Keisuke Tanaka, Rumi Ohtake, Saki Yoshida and
Takashi Shinohara

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72629

Abstract

Microsatellites (simple sequence repeats, SSRs) that consist of repetitive sequences of one to six bases are ubiquitous in most eukaryotic genomes. The use of molecular markers for this region is efficacious in molecular-assisted breeding, molecular phylogenetics, and population genetics. Recently, the detection of a number of SSRs using a high-throughput DNA sequencing assay has become possible. Particularly, microsatellite capture sequencing using our developed protocol can detect SSRs more effectively by enriching the DNA library using an SSR probe. Our protocol used in this study demonstrates the possibility of using low-input DNA (≥1 ng), and while the use of restriction enzymes was more suitable for identifying the heterozygous genotype than sonication was, sonication facilitated the detection of various SSR flanking regions with both species-specific and common characteristics more than restriction enzyme digestion did. Moreover, a simulation analysis using various scale reads estimated that a few thousand SSRs could be detected from 50 K reads per sample. Furthermore, we described an *in silico* polymorphic detection and phylogenetic analysis method based on microsatellite capture sequencing data.

Keywords: microsatellite, SSR, capture sequencing, molecular marker, non-model organisms, Myrtaceae

1. Introduction

Molecular markers for DNA were developed in the 1980s and have been used in a wide variety of research fields as a tool for detecting sequence polymorphism between individuals, cultivars, and lineages. In addition, various polymorphic detection methods using molecular markers have been devised based on the structural characteristics of DNA and molecular biological techniques. Among them, the molecular markers based on microsatellite (simple sequence repeat, SSR) regions enabled the development of robust assays with higher resolution and reliability than those of conventional methods. Generally, SSR constructs consist of a repeat motif with one to six nucleotides, and SSR markers are useful for marker-assisted selection and construction of linkage maps as well as molecular phylogenetics and population genetics because they have various advantages such as high polymorphism, genomic specificity, abundance, and codominance [1, 2]. While SSR markers are very effective, the SSR detection required to construct the marker is often time-consuming. Each SSR detection technique uses colony hybridization, microsatellite enrichment, or both based on the biotin-streptavidin interaction [3–5]. Moreover, another technique was recently reported in which markers were developed in parallel with SSR detection using dual-suppression PCR [6]. However, these approaches have low throughput (a few samples or several tens of samples) since they depend on capillary sequencing.

A current high-throughput DNA sequencing technology, known as next-generation sequencing (NGS), allows the acquisition of huge amounts of data in a single assay. This technology facilitates exhaustive analyses such as whole-genome and RNA sequencing. Additionally, multiple samples can be analyzed at the same time since a specific sequence tag that identifies individuals is added to each library. Thus, the time-consuming assays required for traditional sequencing could be avoided by using such high-throughput DNA sequencing methods. Moreover, high-throughput DNA sequencing was recently used for SSR detection (**Table 1**) [7–42]. These previous studies report that high-throughput DNA sequencing can sufficiently analyze even non-model organisms. In agricultural research field, numerous global major crops such as rice, grapes, and poplar have provided abundant genomic information, whereas little has been reported on regional minor crops have including molecular

Target species	**Library style**	**NGS platform**	**Number of reads**	**Sequences including SSR**
Agkistrodon contortrix (Copperhead snake)	WG	GS-FLX	128,773	14,612
Anisogramma anomala	WG	GAIIX	26,036,313	44,247
Aristeus antennatus (Red shrimp)	WG	GS-FLX	165,507	247
Aristotelia chilensis (Maqui)	WG	GS-FLX	165,043	24,494
Artocarpus altilis	WG	MiSeq	2,341,465	47,607
Aspidistra saxicola	cDNA	HiSeq2000	13,133,336	4764
Brachiaria ruziziensis (Ruzigrass)	WG	GAII	186,764,108	139,098
Callosobruchus chinensis (Adzuki bean weevil)	WG	HiSeq2500	106,888,024	6593
Camelina sativa	cDNA	GAIIX	10,830,000	14,140
Camellia sinensis (Tea plant)	cDNA	HiSeq2000	26,874,116	5649
Carthamus tinctorius (Safflower)	WG	HiSeq2000	48,502,680	23,067
Catha edulis (Khat)	WG	GS-FLX	65,401	11,678
Catla catla (Catla)	WG	PGM	29,794	21,477
Chrysanthemum nankingense	cDNA	GAII	53,720,166	2813
Daphne kiusiana	WG	MiSeq	4,936,656	28,495
Handroanthus billbergii	WG	MiSeq	2,169,901	61,074
Hydropotes inermis (Water Deer)	WG	GS-FLX	260,467	20,101

Target species	Library style	NGS platform	Number of reads	Sequences including SSR
Hymenolaimus malacorhynchos (Blue duck)	WG	GS-FLX	17,215	231
Ipomoea batatas (Sweetpotato)	cDNA	GAII	59,233,468	4114
Lathyrus sativus (Grasspea)	MCS	GS-FLX	493,364	129,886
Mangifera indica (Mango)	WG	HiSeq2000	90,323,371	106,049
Miscanthus sinensis	cDNA	GS-FLX	241,051	381
Moa fossil	WG	GS-FLX	79,796	195
Panicum miliaceum (Broomcorn millet)	MCS	GS-FLX	1,087,428	223,894
Panicum virgatum (Switchgrass)	cDNA	GS-FLX	979,903	21,437
Pilosocereus genus	RAD	GS-FLX	2,282,266	54,420
Pisum sativum (Pea)	WG	HiSeq2500	173,245,234	8899
Prunus virginiana (Chokecherry)	WG	GS-FLX	145,094	405
Pseudosciaena crocea (Large yellow croaker)	WG	GS-FLX	207,246	2535
Python molurus bivittatus (Burmese python)	WG	GS-FLX	117,515	6616
Raja pulchra (Skate)	WG	GS-FLX	453,549	19,658
Raphanus sativus (Radish)	cDNA	HiSeq2000	71,950,000	11,928
Scabiosa columbaria	cDNA	GS-FLX, GAII	29,522,184	4320
Sesamum indicum (sesame)	cDNA	HiSeq2000	26,266,670	6276
Vicia faba (Faba bean)	MCS	GS-FLX	532,599	125,559
Viola mirabilis	WG	GS-FLX	443,935	36,670

WG: whole genome; MCS: microsatellite capture sequencing; RAD: restriction site associated DNA.

Table 1. Examples of simple sequence repeat (SSR) detection with high-throughput sequencing.

markers. Particularly, the development of genomic resources for tropical and subtropical fruits or underutilized fruit crops is limited [43]. We carried out microsatellite capture sequencing using a high-throughput DNA sequencing technology to obtain sequences with SSR regions of candidate SSR markers for five Myrtaceae plants (*Feijoa sellowiana*, *Myrciaria dubia*, *Psidium guajava*, *Psidium littorale*, and *Syzygium samarangense*) that are tropical and subtropical fruits [44].

In this chapter, we expound on the microsatellite capture sequencing method for detecting SSR regions using high-throughput DNA sequencing based on our developed protocol.

2. Overview of microsatellite capture sequencing method

The microsatellite capture sequencing method based on our protocol is explained in this section (**Figure 1**). Some procedures in this protocol are optional fragmentation, SSR enrichment, and data analysis using merge paired-end read with a short-read sequencer.

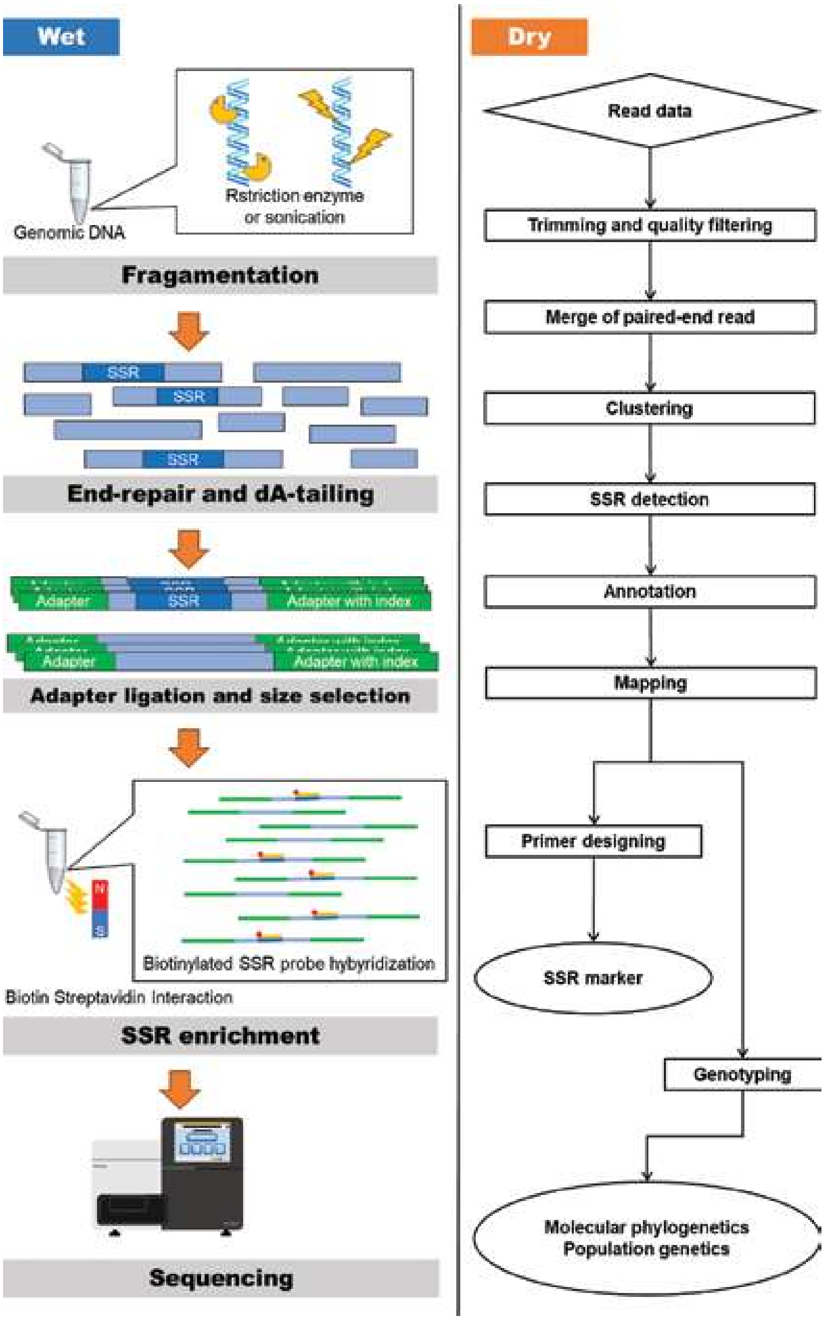


Figure 1. Workflow of our simple sequence repeat (SSR) detection method using microsatellite capture sequencing.

The purity and yield of extracted DNA are determined based on the absorbance ratios of 230/260 and 260/280 nm detected using a spectrometer and should be >1.5. In addition, a single band without smear should be obtained in 1% agarose gel electrophoresis. Subsequently, 1000 ng of the DNA is fragmented by digestion with the appropriate restriction enzyme or by shearing to an average fragment size of 500 bp using an adaptive focused acoustics sonicator (Covaris, Woburn, MA, USA). The fragmented DNA is purified using a QIAquick PCR purification kit (Qiagen, Hilden, Germany), and then a standard Illumina NGS library is constructed using end-repair, dA-tailing, adaptor ligation, size selection, and PCR. We suggest using the NEBNext Ultra DNA Library Prep kit for Illumina (New England Biolabs, Ipswich, MA, USA) with DNA samples ≥10 ng, while the KAPA Hyper Prep kit for Illumina (Kapa Biosystems, Woburn, MA, USA) is recommended for DNA sample < 10 ng. Size selection is conducted using AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) with the approximate insert size set to 400–600 bp. The adaptor-ligated DNA is amplified through 15 high-fidelity PCR cycles. Subsequently, the PCR product is purified in a 20-μL volume via a cleanup stem using AMPure XP magnet beads (Beckman Coulter).

The purified product is mixed with 1 μL of a customized biotinylated SSR probe $(GA)_{10}$ from a 100 μM stock in TE buffer (one of the probes is typically used for SSR enrichment), incubated at 95°C for 10 min, and then placed on ice. The mixture is hybridized by incubating at 60°C for 60 min. After washing 20 μL of the Dynabeads MyOne streptavidin C1 beads (Life Technologies, Carlsbad, CA, USA), they are resuspended in 29 μL of 6× SSC buffer, added to each hybridized mixture, and incubated at 25°C for 30 min. The mixture is washed with 2× SSC buffer (once) and 1× SSC buffer (twice). Next, 23 μL of the SSR-enriched library is amplified using 15 high-fidelity PCR cycles with index primers. Subsequently, the amplified product is purified to a 20-μL volume via a cleanup step using AMPure XP magnetic beads (Beckman Coulter). The library quality and concentration are assessed using an Agilent Bioanalyzer 2100 (Agilent Technologies, Waldbronn, Germany) and Agilent DNA 1000 kit (Agilent Technologies). The specific concentration of each library is determined using quantitative real-time PCR using a KAPA library quantification kit (Kapa Biosystems). The library is first diluted to a concentration of 10 nM and then mixed in equal amounts. After denaturation with 0.2 N NaOH, the final concentration of the library mixture is diluted to 15 pM, including the 1% PhiX library (Illumina, CA, USA). The library mixture is sequenced using 2 × 300 bp paired-end sequencing using a MiSeq (Illumina). Reads in the FASTQ format were generated using a pipeline MiSeq reporter (version 2.5.1.3, Illumina).

Raw reads containing adaptors are removed using Trimmomatic version 0.32 [45]. Additionally, the FASTX-Toolkit version 0.0.13.2 [46] is used to clip uncertain bases called "N" and filter reads based on the quality score. The parameters of the quality filtering are as follows: (1) required minimum quality score is 20 and (2) minimum percentage of bases that must have [−q] quality is 80. The unpaired reads are then removed from the total remaining using a custom Perl script, and the preprocessed paired reads are integrated using the FLASh version 1.2.11 [47]. Furthermore, the integrated reads with similar sequences are clustered using the CD-HIT-EST version 4.6 [48], and the clustered reads including SSR regions are searched using the SSRIT for a stand-alone version (ftp://ftp.gramene.org/pub/gramene/archives/software/scripts/ssr.pl) [49]. The search parameters are (1) unit size, 2 and (2) minimum repeats, 10. Subsequently, sequences with a length of more than 20 bp flanking the SSR region are listed to enable the design of the primer set. The listed sequences are annotated with the top-hit description using the local BLAST program with the following settings: (1) execution program, BLASTn; (2) database, the NCBI nonredundant nucleotide database nt; and (3) e-value, 1e−4. Additionally, consensus sequences from these sequences are constructed using read mapping in CLC Genomics Workbench 9.5 (CLC Bio-Qiagen, Aarhus, Denmark) with the following settings: (1) mismatch cost, 2; (2) insertion cost, 3; (3) deletion cost, 3; (4) length fraction, 0.5; and (5) similarity fraction, 0.8.

3. Available amount of DNA sample

Our protocol recommends using 1000 ng of the DNA sample. However, occasionally, only low amounts of DNA are obtained depending on the experiment design. Therefore, we adjusted the amounts to 1000, 100, 10, and 1 ng with rice (*Oryza sativa*) DNA as a test sample and investigated the feasibility of using these amount of DNA samples in our protocol (**Figure 2**). The above amounts (1000, 100, and 10 ng) of DNA samples were used with the NEBNext Ultra DNA Library Prep kit for Illumina (New England Biolabs), and 1 ng of the DNA sample was used with the KAPA Hyper Prep kit for Illumina (Kapa Biosystems). The processed DNA was quantified using an Agilent Bioanalyzer 2100 (Agilent Technologies) and Agilent DNA 1000 kit (Agilent Technologies) before and after SSR enrichment. As a result, the concentration of the processed DNA before SSR enrichment (after standard NGS library construction) was determined to be in the range of 226.9–14.7 nM in proportion to the input DNA, while that after SSR enrichment was gradually detected in the 23.5–3.4 nM range, although the concentration was reduced compared to the input DNA. Although the peaks after SSR enrichment with 10 and 1 ng input DNA could not be confirmed, they were checked using the Agilent high sensitivity DNA kit (Agilent Technologies). Therefore, our protocol can construct the SSR enriched library for high-throughput DNA sequencing if the prepared input DNA is ≥1 ng. If the constructed library does not meet the ≥1 nM required concentration for high-throughput DNA sequencing,

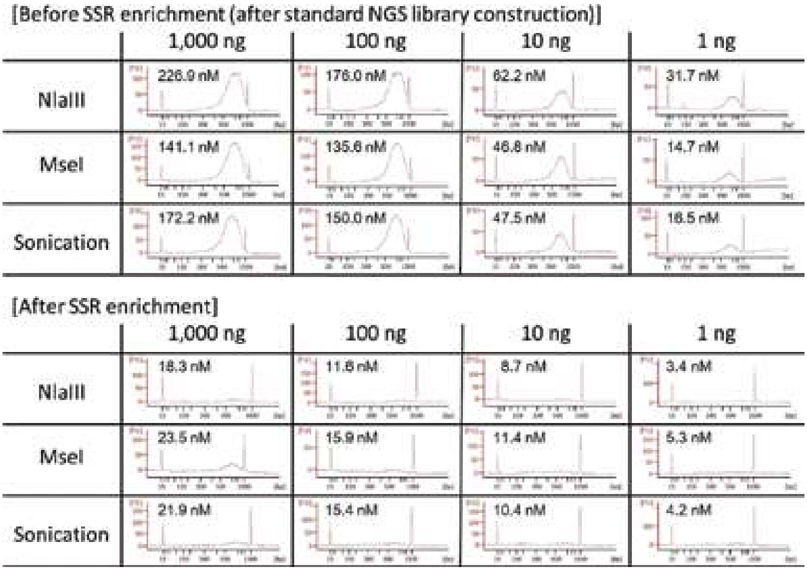


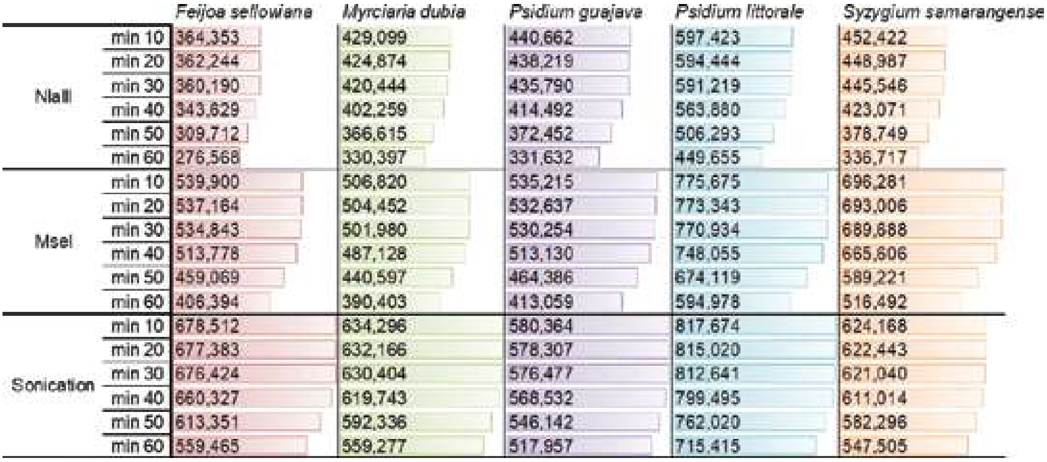
Figure 2. DNA profiling before and after simple sequence repeat (SSR) enrichment using Agilent Bioanalyzer 2100.

we suggest the following approaches: (1) elution with less buffer volume by re-performing the cleanup, (2) enrichment using an evaporator, and (3) several additional PCR cycles. The library constructed on a 1-ng scale may be useful for analyzing a few valuable samples such as cell masses with microsatellite instability and herbarium specimens.

4. Effect of fragmentation

Our protocol chooses between restriction enzyme digestion and sonication for the DNA fragmentation. In this experiment, the effect of data analysis on these different fragmentation methods was investigated using the sequence data (accession number DRA004725) for the Myrtaceae plants with the microsatellite capture sequencing.

During the integration process, the minimum overlapping length parameter was appropriated at 10-base intervals (min 10, 20, 30, 40, 50, and 60; **Figure 3**). The result showed that after integration and clustering, the integrated reads tended to be higher after fragmentation by sonication. Of the two restriction enzymes, MseI yielded more integrated reads than NlaIII did. The recognition site of MseI consists of only adenine and thymine (5′-T|TAA-3′), whereas that of NlaIII consists of all nucleotides (5′-CATG|-3′). Although the GC content in the whole genome is lower than 50% for many plants [50], Myrtaceae species also tend to exhibit low GC content [51, 52]. Thus, these results showed that the varying number of integrated reads obtained when different restriction enzymes were used was reasonable. Additionally, we compared variations in the integrated reads between the minimum overlapping parameters in the integration process. At min 10, the integrated reads constructed from the original paired-end reads were 25–43, 33–52, and 44–61% for NlaIII, MseI, and sonication, respectively while the corresponding values at min 60 were 19–32, 25–40, and 37–54%, respectively. Therefore, integrated reads ranging from several percentage points to approximately 10% at most could be varied by setting an arbitrary



		Feijoa sellowiana	*Myrciaria dubia*	*Psidium guajava*	*Psidium littorale*	*Syzygium samarangense*
NlaII	min 10	364,353	429,099	440,662	597,423	452,422
	min 20	362,244	424,874	438,219	594,444	448,987
	min 30	360,190	420,444	435,790	591,219	445,546
	min 40	343,629	402,259	414,492	563,880	423,071
	min 50	309,712	366,615	372,452	506,293	378,749
	min 60	276,568	330,397	331,632	449,655	336,717
MseI	min 10	539,900	506,820	535,215	775,675	696,281
	min 20	537,164	504,452	532,637	773,343	693,006
	min 30	534,843	501,980	530,254	770,934	689,688
	min 40	513,778	487,128	513,130	748,055	665,606
	min 50	459,069	440,597	464,386	674,119	589,221
	min 60	406,394	390,403	413,059	594,978	516,492
Sonication	min 10	678,512	634,296	580,364	817,674	624,168
	min 20	677,383	632,166	578,307	815,020	622,443
	min 30	676,424	630,404	576,477	812,641	621,040
	min 40	660,327	619,743	568,532	799,495	611,014
	min 50	613,351	592,336	546,142	762,020	582,296
	min 60	559,465	559,277	517,957	715,415	547,505

Figure 3. Number of contigs after integration of paired-end reads and deletions of duplicated integrated-reads.

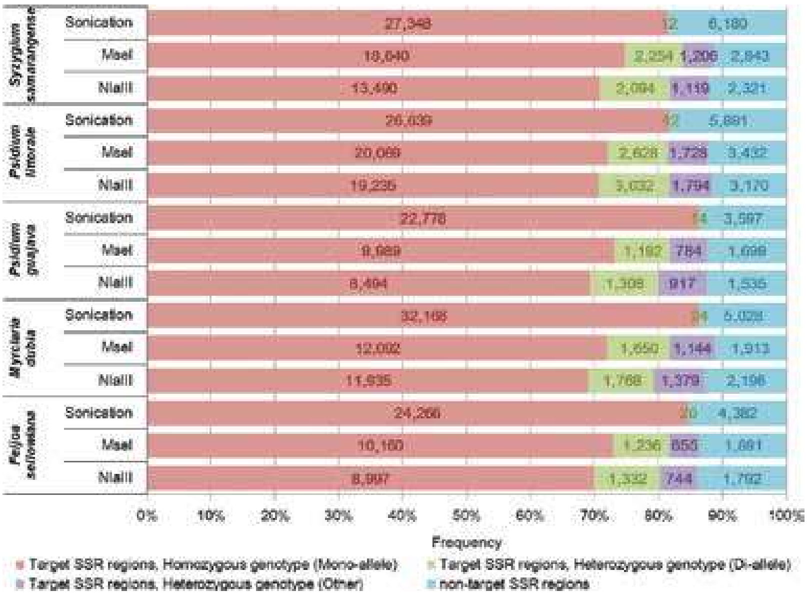


Figure 4. Simple sequence repeat (SSR) proportions of target and other regions.

parameter for the minimum overlapping length. We recommend using an overlapping length parameter of min 60 to select integrated reads with higher reliability.

After searching SSR regions of the clustered reads using the overlapping length parameter of min 60, most motifs (82–89%) were target SSR regions [$(CT)_n$, $(TC)_n$, $(GA)_n$, or $(AG)_n$; **Figure 4**]. This result could be attributed to the biotinylated probe used to enrich the SSR region. The various probe conditions required for capturing SSR regions have been reported previously [53–55]. Our protocol also showed that the target SSR region could be captured efficiently. Among the SSRs shown in **Figure 4**, probe-related SSR regions were characterized based on genotypic frequency. All species and the fragmentation conditions showed high and low rates of homozygous and heterozygous genotypes, respectively. The heterozygous genotype rate for all species was substantially higher after fragmentation using restriction enzymes than it was after fragmentation using sonication (18.75–20.86, 15.66–18.77, and 0.04–0.16% for NlaIII, MseI, and sonication, respectively). Additionally, fragmentation by NlaIII was more likely to detect the heterozygous genotype than that by MseI was and, thus, for our approach we recommend fragmentation using restriction enzyme digestion. Although the five Myrtaceae plants analyzed in this study are diploid, approximately one-third of heterozygous genotypes were more than tri-allelic in all species. This factor may be associated with the occurrence of multiple homologous copies or PCR error during library construction.

We confirmed the unique and common genes with SSR flanking regions in a family based on annotations (**Figure 5**). In *F. sellowiana*, 372 (NlaIII), 337 (MseI), and 887 (sonication) SSR regions were fragmentation-specific, whereas 440 (NlaIII), 474 (MseI), and 624 (sonication) SSR regions were common among other fragmentations. In *M. dubia*, 338 (NlaIII), 320 (MseI), and 1065 (sonication) SSR regions were fragmentation-specific, whereas 481 (NlaIII), 482 (MseI), and 724 (sonication) SSR regions were common among other fragmentations. In *P. guajava*, 227 (NlaIII),

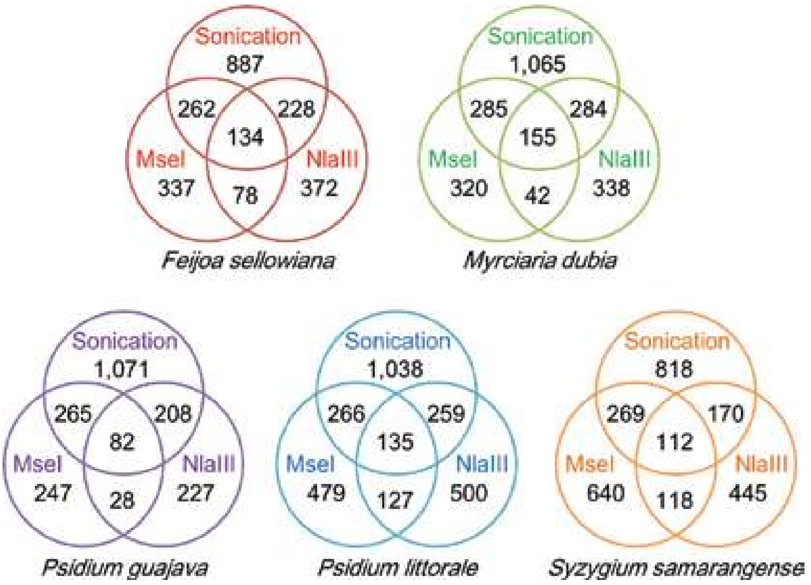


Figure 5. Venn diagram based on annotation within three fragmentations for each Myrtaceae plant.

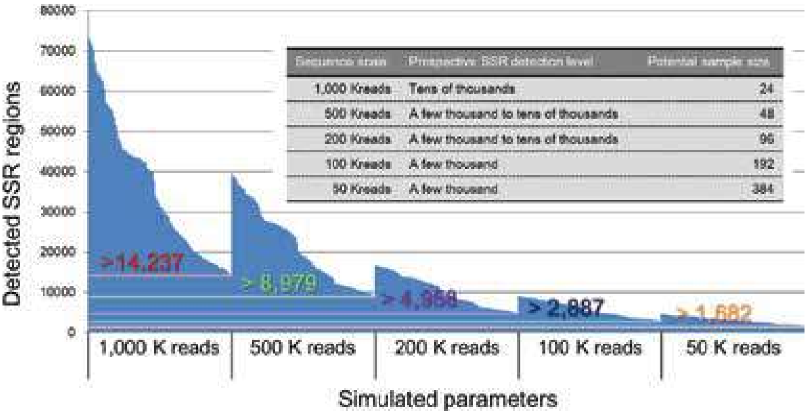
247 (MseI), and 1071 (sonication) SSR regions were fragmentation-specific, whereas 318 (NlaIII), 375 (MseI), and 555 (sonication) SSR regions were common among other fragmentations. In *P. littorale*, 500 (NlaIII), 479 (MseI), and 1038 (sonication) SSR regions were fragmentation-specific, whereas 521 (NlaIII), 528 (MseI), and 660 (sonication) SSR regions were common among other fragmentations. In *S. samarangense*, 445 (NlaIII), 640 (MseI), and 818 (sonication) SSR regions were fragmentation-specific, whereas 400 (NlaIII), 499 (MseI), and 551 (sonication) SSR regions were common among other fragmentations. Therefore, the detected SSR flanking region had both fragmentation-specific and common characteristics. Notably, sonication yielded the most characteristics for all groups.

We constructed consensus sequences from the listed sequences including SSRs based on the read mapping. The percentage values of unsuited consensus sequences for molecular marker development determined by including the unknown nucleotide "N" were 4.4% (NlaIII), 5.3% (MseI), and 1.4% (sonication) in *F. sellowiana*; 5.7% (NlaIII), 5.4% (MseI), and 3.5% (sonication) in *M. dubia*; 9.7% (NlaIII), 10.7% (MseI), and 5.3% (sonication) in *P. guajava*; 6.5% (NlaIII), 5.7% (MseI), and 1.8% (sonication) in *P. littorale*; and 7.0% (NlaIII), 7.9% (MseI), and 1.3% (sonication) in *S. samarangense*. Conversely, approximately 90% of the consensus sequences could be candidate molecular markers for some gene and trait.

Fragmentation by restriction enzyme is limited in the restriction site flanking region, whereas fragmentation by sonication targets the whole genome. The comparison of different fragmentation methods for genomic DNA revealed that restriction enzymes were more suitable for identifying the heterozygous genotype than sonication was, whereas sonication facilitated the detection of various SSR flanking regions with both species-specific and common characteristics more than restriction enzyme digestion did. Therefore, the choice of the DNA fragmentation approach appears to depend on the ultimate research purpose. In particular, the effective detection of a heterozygous genotype using DNA fragmentation by restriction enzymes is expected to contribute to the development of molecular markers for molecular-assisted breeding and population genetics that require the clear distinction of alleles.

5. Simulation analysis using various sequence data scales

The number of paired reads for the Myrtaceae plants ranged from 2 × 1,296,448 to 2 × 1,708,634. Here, various sequence data scales are simulated to demonstrate how SSRs can be detected (**Figure 6**). For the simulation, different scale data consisting of 1000, 500, 200, 100, and 50K reads were prepared by sampling without replacement of the original sequence data of the five Myrtaceae plants and three fragmentation approaches. Additionally, paired reads of each scale dataset were integrated using the minimum overlapping parameter of 10-base intervals (min 10, 20, 30, 40, 50, and 60), and sequences with SSR regions were detected from the integrated reads after clustering the same sequences. The simulation result showed that the 1000K reads scale detected tens of thousands of SSRs; 500 and 200K reads detected a few thousand to tens of thousands of SSRs; and 100 and 50K reads detected a few thousand SSRs. Thus, the 50K reads detected approximately 1/10 of the SSRs detected by the 1000K reads. Moreover, 384 samples could be used in a single assay when a sequence of 50K reads per sample is assumed, and the sequence cost is very reasonable.



Sequence scale	Prospective SSR detection level	Potential sample size
1,000 Kreads	Tens of thousands	24
500 Kreads	A few thousand to tens of thousands	48
200 Kreads	A few thousand to tens of thousands	96
100 Kreads	A few thousand	192
50 Kreads	A few thousand	384

Figure 6. Simulation result based on assuming various read scales.

6. *In silico* polymorphic detection and phylogenetic analysis

SSRs detected using microsatellite capture sequencing are available as SSR markers by designing primer sets from the SSR flanking region. On the other hand, when common sequences with SSR regions among samples are prepared as reference sequences, the sequence data of each sample can be mapped to the reference, SSR polymorphisms can be detected among the samples, and phylogenetic analysis is possible based on the polymorphic data. Here, we explain an *in silico* polymorphic detection and phylogenetic analysis method based on microsatellite capture sequencing data.

According to the protocol described above, the sequence set with the SSR region is prepared from analyzed data using paired-read integration, clustering of same sequences, and SSR detection. The sequence sets of each sample are merged into one file, and the merged sequence sets are re-clustered using CD-HIT-EST version 4.6 [48]. The sequence data of each sample are mapped to the clustered sequence as a reference by using the CLC Genomics Workbench 9.5 (CLC Bio-Qiagen, Aarhus, Denmark). The consensus sequences of each sample are created from the mapped data. The SSR repeat data of the consensus sequences are detected using SSRIT for a stand-alone version [49]. A polymorphic table merging the SSR detected data of each sample is constructed using the following script (merge_SSR.pl):

```
#!/usr/bin/perl.

use strict;

our @hashlist = ();

our @fnlist = ();

our %keyhash = ();

eval {.

        if($#ARGV <0) {

                  print “usage: mergeSSR.pl [sample tsv files (ex. mergeSSR.pl *.txt)]\n”;

                  exit −1;

          }

          ### header ########################################

          print “Locus”;

          for (my $i = 0; $i < = $#ARGV; $i++){

                    my $filename = $ARGV[$i];

                    print “\t”.$filename;

                    push(@fnlist, $filename);

          }

          print “\n”;

          ### file to hash Locus ########################################

          for (my $i = 0; $i < = $#fnlist; $i++){

                    my $filename = $fnlist[$i];
                    my %hash = ();
                     open (IN,” < $filename”) | | die “cannot open $filename: $!”;
                     while (my $line = <IN>) {
                              chomp $line;
                              my @dt = split/\t/,$line;
                              my $key = $dt[0].”_”.$dt[1];
                              my $val = $dt[4];
                              $hash{$key} = $val;
                              $keyhash{$key} = $key;
                     }
                     close (IN1);
                     push(@hashlist, \%hash);
  }
  #### data #################################################
  foreach my $key (keys %keyhash){
            print $key;
            foreach my $row (@hashlist) {
                      if(exists $row- > {$key}){
                              print “\t”.$row- > {$key};
                      } else {.
                              print “\t”;
                     }
           }
          print “\n”;
  }
   exit 0;
};
```

The polymorphic table is edited to the input data of the Populations format. The genetic distance between samples is calculated using a distance matrix method using the Populations

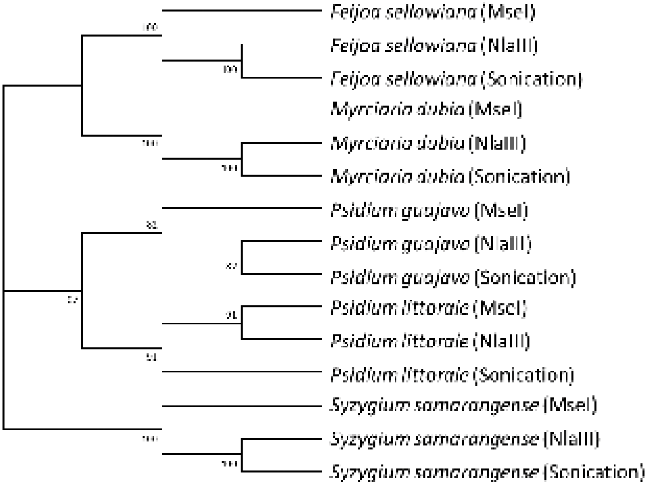

Feijoa sellowiana (MseI)
Feijoa sellowiana (NlaIII)
Feijoa sellowiana (Sonication)
Myrciaria dubia (MseI)
Myrciaria dubia (NlaIII)
Myrciaria dubia (Sonication)
Psidium guajava (MseI)
Psidium guajava (NlaIII)
Psidium guajava (Sonication)
Psidium littorale (MseI)
Psidium littorale (NlaIII)
Psidium littorale (Sonication)
Syzygium samarangense (MseI)
Syzygium samarangense (NlaIII)
Syzygium samarangense (Sonication)

Figure 7. Neighbor-joining dendrogram (only topology) constructed from *in silico* polymorphic data of five Myrtaceae plants.

version 1.2.30 [56]. A dendrogram is drawn using the MEGA version 7 [57]. For example, we have shown the result of the phylogenetic analysis of the Myrtaceae plants (**Figure 7**). SSR polymorphisms in 38,636 loci were compared between samples, and the result showed that all organisms had a single clade even if the fragmentation differed.

7. Conclusion

Recently, the SSR detection approach using high-throughput DNA sequencing has been performed on various organisms (**Table 1**). Although SSRs are detected from the whole genome as a part of the data analyses in many cases, the microsatellite capture sequencing approach included in our protocol can detect numerous SSRs more effectively by enriching NGS library using an SSR probe than conventional approaches. Detected SSR data will considerably increase the spread of NGS in the future. Therefore, the construction of a database will be required to manage the massive amount of SSR data.

Acknowledgements

This work was supported by the Ministry of Education, Culture, Sports, Science, and Technology-Supported Program for the Strategic Research Foundation at Private Universities (S1311017). We thank Hironobu Uchiyama and Eri Kubota for their invaluable advice and Satoshi Sano, Hiroto Sekitoh, and Yu Hamaguchi for technical support.

Author details

Keisuke Tanaka[1], Rumi Ohtake[1], Saki Yoshida[2] and Takashi Shinohara[3]*

*Address all correspondence to: t3shinoh@nodai.ac.jp

1 NODAI Genome Research Center, Tokyo University of Agriculture, Tokyo, Japan

2 International Agricultural Development, Graduate School of Agriculture, Tokyo University of Agriculture, Tokyo, Japan

3 Junior College of Tokyo University of Agriculture, Tokyo University of Agriculture, Tokyo, Japan

References

[1] Ijaz S. Microsatellite markers: An important fingerprinting tool for characterization of crop plants. African Journal of Biotechnology. 2011;**10**:7723-7726

[2] Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. Microsatellite markers: An overview of the recent progress in plants. Euphytica. 2011;**177**:309-334. DOI: 10.1007/s10681-010-0286-9

[3] Rassmann K, Schlötterer C, Tautz D. Isolation of simple-sequence loci for use in polymerase chain reaction-based DNA fingerprinting. Electrophoresis. 1991;**12**:113-118. DOI: 10.1002/elps.1150120205

[4] Karagyozov L, Kalcheva ID, Chapman VM. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research. 1993;**21**:3911-3912. DOI: 10.1093/nar/21.16.3911

[5] Connell JP, Pammi S, Iqbal MJ, Huizinga T, Reddy AS. A high through-put procedure for capturing microsatellites from complex plant genomes. Plant Molecular Biology Reporter. 1998;**16**:341-349. DOI: 10.1023/A:1007536421700

[6] Lian C, Hogetsu T. Development of microsatellite markers in black locust (*Robinia pseudoacacia*) using a dual-suppression-PCR technique. Molecular Ecology Resources. 2002;**2**:211-213. DOI: 10.1046/j.1471-8286.2002.00213.x-i2

[7] Castoe TA, Poole AW, Gu W, Jason de Koning AP, Daza JM, Smith EN, Pollock DD. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Molecular Ecology Resources. 2010;**10**:341-347. DOI: 10.1111/j.1755-0998.2009.02750.x

[8] Cai G, Leadbetter CW, Muehlbauer MF, Molnar TJ, Hillman BI. Genome-wide microsatellite identification in the fungus *Anisogramma anomala* using Illumina sequencing and genome assembly. PLoS One. 2013;**8**:e82408. DOI: 10.1371/journal.pone.0082408

[9] Heras S, Planella L, Caldarazzo I, Vera M, García-Marín JL, Roldán MI. Development and characterization of novel microsatellite markers by next generation sequencing for the blue and red shrimp *Aristeus antennatus*. PeerJ. 2016;**4**:e2200. DOI: 10.7717/peerj.2200

[10] Bastías A, Correa F, Rojas P, Almada R, Muñoz C, Sagredo B. Identification and characterization of microsatellite loci in maqui (*Aristotelia chilensis* [Molina] Stunz) using next-generation sequencing (NGS). PLoS One. 2016;**11**:e0159825. DOI: 10.1371/journal.pone.0159825

[11] De Bellis F, Malapa R, Kagy V, Lebegin S, Billot C, Labouisse JP. New development and validation of 50 SSR markers in breadfruit (*Artocarpus altilis*, Moraceae) by next-generation sequencing. Applications in Plant Sciences. 2016;**4**:e1600021. DOI: 10.3732/apps.1600021

[12] Huang D, Zhang Y, Jin M, Li H, Song Z, Wang Y, Chen J. Characterization and high cross-species transferability of microsatellite markers from the floral transcriptome of *Aspidistra saxicola* (Asparagaceae). Molecular Ecology Resources. 2014;**14**:569-577. DOI: 10.1111/1755-0998.12197

[13] Silva PI, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira ME. Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics. 2013;**14**:e17. DOI: 10.1186/1471-2164-14-17

[14] Duan CX, Li DD, Sun SL, Wang XM, Zhu ZD. Rapid development of microsatellite markers for *Callosobruchus chinensis* using illumina paired-end sequencing. PLoS One. 2014;**9**:e95458. DOI: 10.1371/journal.pone.0095458

[15] Mudalkar S, Golla R, Ghatty S, Reddy AR. *De novo* transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. Plant Molecular Biology. 2014;**84**:159-171. DOI: 10.1007/s11103-013-0125-1

[16] Tan LQ, Wang LY, Wei K, Zhang CC, LY W, Qi GN, Cheng H, Zhang Q, Cui QM, Liang JB. Floral transcriptome sequencing for SSR marker development and linkage map construction in the tea plant (*Camellia sinensis*). PLoS One. 2013;**8**:e81611. DOI: 10.1371/journal.pone.0081611

[17] Ambreen H, Kumar S, Variath MT, Joshi G, Bali S, Agarwal M, Kumar A, Jagannath A, Goel S. Development of genomic microsatellite markers in *Carthamus tinctorius* L. (safflower) using next generation sequencing and assessment of their cross-species transferability and utility for diversity analysis. PLoS One. 2015;**10**:e0135443. DOI: 10.1371/journal.pone.0135443

[18] Curto MA, Tembrock LR, Puppo P, Nogueira M, Simmons MP, Meimberg H. Evaluation of microsatellites of *Catha edulis* (qat; Celastraceae) identified using pyrosequencing. Biochemical Systematics and Ecology. 2013;**49**:1-9. DOI: 10.1016/j.bse.2013.02.002

[19] Sahu BP, Sahoo L, Joshi CG, Mohanty P, Sundaray JK, Jayasankar P, Das P. Isolation and characterization of polymorphic microsatellite loci in Indian major carp, *Catla catla* using next-generation sequencing platform. Biochemical Systematics and Ecology. 2014;**57**:357-362. DOI: 10.1016/j.bse.2014.09.010

[20] Wang H, Jiang J, Chen S, Qi X, Peng H, Li P, Song A, Guan Z, Fang W, Liao Y, Chen F. Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome permits large-scale unigene assembly and SSR marker discovery. PLoS One. 2013;**8**:e62293. DOI: 10.1371/journal.pone.0062293

[21] Lee JH, Cho WB, Yang S, Han EK, Lyu ES, Kim WJ, Moon BC, Choi G. Development and characterization of 21 microsatellite markers in *Daphne kiusiana*, an evergreen broad-leaved

shrub endemic to Korea and Japan. Korean J Pl Taxon. 2017;**47**:6-10. DOI: 10.11110/kjpt.2017.47.1.6

[22] Morillo E, Buitron J, Limongi R, Vignes H, Argout X. Characterization of microsatellites identified by next-generation sequencing in the Neotropical tree *Handroanthus billbergii* (Bignoniaceae). Applications in Plant Sciences. 2016;**4**:e1500135. DOI: 10.3732/apps.1500135

[23] Yu JN, Won C, Jun J, Lim Y, Kwak M. Fast and cost-effective mining of microsatellite markers using NGS technology: An example of a Korean water deer *Hydropotes inermis argyropus*. PLoS One. 2011;**6**:e26933. DOI: 10.1371/journal.pone.0026933

[24] Abdelkrim J, Robertson B, Stanton JA, Gemmell N. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques. 2009;**46**: 185-192. DOI: 10.2144/000113084

[25] Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). BMC Genomics. 2010;**11**:e726. DOI: 10.1186/1471-2164-11-726

[26] Yang T, Jiang J, Burlyaeva M, Hu J, Coyne CJ, Kumar S, Redden R, Sun X, Wang F, Chang J, Hao X, Guan J, Zong X. Large-scale microsatellite development in grasspea (*Lathyrus sativus* L.), an orphan legume of the arid areas. BMC Plant Biology. 2014;**14**:e65. DOI: 10.1186/1471-2229-14-65

[27] Ravishankar KV, Dinesh MR, Nischita P, Sandya BS. Development and characterization of microsatellite markers in mango (*Mangifera indica*) using next-generation sequencing technology and their transferability across species. Molecular Breeding. 2015;**35**:e93. DOI: 10.1007/s11032-015-0289-2

[28] Ho CW, TH W, Hsu TW, Huang JC, Huang CC, Chiang TY. Development of 12 genic microsatellite loci for a biofuel grass, *Miscanthus sinensis* (Poaceae). American Journal of Botany. 2011;**98**:e201-e203. DOI: 10.3732/ajb.1100071

[29] Allentoft M, Schuster SC, Holdaway R, Hale M, McLay E, Oskam C, Gilbert MT, Spencer P, Willerslev E, Bunce M. Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. BioTechniques. 2009;**46**:195-200. DOI: 10.2144/000113086

[30] Liu MX, Xu Y, Yang TY, Qiao ZJ, Wang RY, Wang YY, Lu P. Development of species-specific microsatellite markers for broomcorn millet (*Panicum miliaceum* L.) via high-throughput sequencing. Advances in Crop Science and Technology. 2017;**5**:e297. DOI: 10.4172/2329-8863.1000297

[31] Wang Y, Zeng X, Iyer NJ, Bryant DW, Mockler TC, Mahalingam R. Exploring the switchgrass transcriptome using second-generation sequencing technology. PLoS One. 2012;**7**: e34225. DOI: 10.1371/journal.pone.0034225

[32] Bonatelli IA, Carstens BC, Moraes EM. Using next generation RAD sequencing to isolate multispecies microsatellites for *Pilosocereus* (Cactaceae). PLoS One. 2015;**10**:e0142602. DOI: 10.1371/journal.pone.0142602

[33] Yang T, Fang L, Zhang X, Hu J, Bao S, Hao J, Li L, He Y, Jiang J, Wang F, Tian S, Zong X. High-throughput development of SSR markers from pea (*Pisum sativum* L.) based on next generation sequencing of a purified Chinese commercial variety. PLoS One. 2015;**10**: e0139775. DOI: 10.1371/journal.pone.0139775

[34] Wang H, Walla JA, Zhong S, Huang D, Dai W. Development and cross-species/genera transferability of microsatellite markers discovered using 454 genome sequencing in chokecherry (*Prunus virginiana* L.). Plant Cell Reports. 2012;**31**:2047-2055. DOI: 10.1007/s00299-012-1315-z

[35] Lü Z, Li H, Liu L, Cui W, Hu X, Wang C. Rapid development of microsatellite markers from the large yellow croaker (*Pseudosciaena crocea*) using next generation DNA sequencing technology. Biochemical Systematics and Ecology. 2013;**51**:314-319. DOI: 10.1016/j.bse.2013.09.019

[36] Hunter ME, Hart KM. Rapid microsatellite marker development using next generation pyrosequencing to inform invasive Burmese python—*Python molurus bivittatus*—Management. International Journal of Molecular Sciences. 2013;**14**:4793-4804. DOI: 10.3390/ijms14034793

[37] Kang JH, Park JY, Jo HS. Rapid development of microsatellite markers with 454 pyrosequencing in a vulnerable fish, the mottled skate, *Raja pulchra*. International Journal of Molecular Sciences. 2012;**13**:7199-7211. DOI: 10.3390/ijms13067199

[38] Zhai L, Xu L, Wang Y, Cheng H, Chen Y, Gong Y, Liu L. Novel and useful genic-SSR markers from *de novo* transcriptome sequencing of radish (*Raphanus sativus* L.). Molecular Breeding. 2014;**33**:611-624. DOI: 10.1007/s11032-013-9978-x

[39] Angeloni F, Wagemaker CA, Jetten MS, Op den Camp HJ, Janssen-Megens EM, Francoijs KJ, Stunnenberg HG, Ouborg NJ. *De novo* transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. Molecular Ecology Resources. 2011;**11**:662-674. DOI: 10.1111/j.1755-0998.2011.02990.x

[40] Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics. 2011;**12**:e451. DOI: 10.1186/1471-2164-12-451

[41] Yang T, Bao SY, Ford R, Jia TJ, Guan JP, He YH, Sun XL, Jiang JY, Hao JJ, Zhang XY, Zong XX. High-throughput novel microsatellite marker of faba bean via next generation sequencing. BMC Genomics. 2012;**13**:e602. DOI: 10.1186/1471-2164-13-602

[42] Kang JS, Lee B, Kwak M. Isolation and characterization of microsatellite markers for *Viola mirabilis* (Violaceae) using a next-generation sequencing platform. Plant Species Biology. 2017;**32**:448-454. DOI: 10.1111/1442-1984.12159

[43] Rai MK, Shekhawat NS. Genomic resources in fruit plants: An assessment of current status. Critical Reviews in Biotechnology. 2015;**35**:438-447. DOI: 10.3109/07388551.2014.898127

[44] Tanaka K, Ohtake R, Yoshida S, Shinohara T. Effective DNA fragmentation technique for simple sequence repeat detection with a microsatellite-enriched library and high-throughput sequencing. BioTechniques. 2017;**62**:180-182. DOI: 10.2144/000114536

[45] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;**30**:2114-2120. DOI: 10.1093/bioinformatics/btu170

[46] Gordon A, Hannon GJ. FASTX-Toolkit. In: FASTQ/A Short-Reads Pre-Processing Tools. 2010. Available from: http://hannonlab.cshl.edu/fastx_toolkit/ Accessed: 2017–8-31

[47] Magoč T, Salzberg SL. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;**27**:2957-2963. DOI: 10.1093/bioinformatics/btr507

[48] Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;**22**:1658-1659. DOI: 10.1093/bioinformatics/btl158

[49] Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Research. 2001;**11**:1441-1452. DOI: 10.1101/gr.184001

[50] Meister A, Barow M. DNA base composition of plant genomes. In: Doležel J, Greilhuber J, Suda J, editors. Flow Cytometry with Plant Cells: Analysis of Genes, Chromosomes and Genomes. Weinheim: Wiley; 2007. pp. 177-215. DOI: 10.1002/9783527610921.ch8

[51] Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, Bossinger G, Merchant A, Udovicic F, Woodrow IE, Tibbits J. Chloroplast genome analysis of Australian eucalypts—Eucalyptus, corymbia, angophora, allosyncarpia, and stockwellia (Myrtaceae). Molecular Phylogenetics and Evolution. 2013;**69**:704-716. DOI: 10.1016/j.ympev.2013.07.006

[52] Izuno A, Hatakeyama M, Nishiyama T, Tamaki I, Shimizu-Inatsugi R, Sasaki R, Shimizu KK, Isagi Y. Genome sequencing of *Metrosideros polymorpha* (Myrtaceae), a dominant species in various habitats in the Hawaiian islands with remarkable phenotypic variations. Journal of Plant Research. 2016;**129**:727-736. DOI: 10.1007/s10265-016-0822-3

[53] Acquadro A, Portis E, Lanteri S. Isolation of microsatellite loci in artichoke (*Cynara cardunculus* L. Var. *scolymus*). Molecular Ecology Resources. 2003;**3**:37-39. DOI: 10.1046/j.1471-8286.2003.00343.x

[54] He G, Meng R, Newman M, Gao G, Pittman RN, Prakash CS. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biology. 2003;**3**:e3. DOI: 10.1186/1471-2229-3-3

[55] Stajner N, Jakse J, Kozjak P, Javornik B. The isolation and characterisation of microsatellites in hop (*Humulus lupulus* L.). Plant Science. 2005;**168**:213-221. DOI: 10.1016/j.plantsci.2004.07.031

[56] Langella O. Populations, 1.2.28. 1999. Available from: http://bioinformatics.org/populations/ [Accessed: 2017-8-31]

[57] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;**33**:1870-1874. DOI: 10.1093/molbev/msw054

Chapter 3

Allele Size Miscalling due to the Pull-Up Effect Influencing Size Standard Calibration in Capillary Electrophoresis: A Case Study Using HEX Fluorescent Dye in Microsatellites

Zheng-Feng Wang, Se-Ping Dai, Ju-Yu Lian,
Hong-Feng Chen, Wan-Hui Ye and Hong-Lin Cao

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73028

Abstract

Microsatellites are important genetic markers and have been broadly employed in many genetic studies. Currently, polymorphisms in microsatellites are often detected by an automated system of capillary electrophoresis with fluorescent dyes. In this situation, different dye combinations may cause pull-up/bleed-through problems, which introduce noise signals from one dye channel into another, causing genotyping errors. Here, we report the detection of such a problem at two microsatellite loci that used the HEX dye. Using three datasets, we tested for noise effects in four allele-scoring programmes: Genemapper, Genemarker, Gelquest and Fragman. We found that, because some allele sizes were identical or close to the size of one of the internal size standards, all four programmes gave allele size calling errors due to wrongly identifying pull-up signals as the internal size standard. In addition, because allele miscalling in this study was caused by the fluorescent dye that the microsatellites used introducing noise of the same colour as the internal size standard used, the pull-up correction function in Genemapper, Genemarker and Fragman failed to deal with this. Considering that pull-up peak scoring errors can occur with any dye colour, the phenomenon is not limited to the current HEX dye. Using different software and visual scoring of each result will allow accurate sizing of microsatellite alleles.

Keywords: Fragman, Gelquest, Genemapper, Genemarker, genetic markers, allele size scoring

1. Introduction

Microsatellites are important genetic markers that are widely used in evolutionary, ecological, animal and plant breeding and forensic studies [1, 2]. Because their polymorphisms are generally caused by the gain or loss of a repeat unit, they can be easily detected using a gel separation method to detect length variations. Therefore, using fluorescent dye-labelled primers with automatic capillary electrophoresis is one of the most popular methods for high-throughput assessment of their polymorphisms. Microsatellite allele fragments are then estimated or calculated by comparison with a co-migrated internal size standard, which uses a different fluorescent dye that displays a different colour from that used for the microsatellite.

In this process, incorrect microsatellite genotype scoring can occur at many steps [3–7]. Stutter, null alleles and allelic dropout are the three major problems, which have been discussed extensively [5, 7]. These problems were generally related to sample quality, polymorphisms in microsatellite priming sites, PCR amplification procedures and others. Unlike errors relating to DNA templates or PCR procedures, pull-up/bleed-through ("pull-up" hereafter) in capillary electrophoresis per se has also been an important problem causing allele miscalling.

The pull-up effect is due to spectral overlap of the fluorescent dyes in capillary electrophoresis producing more than one peak colour for the allele with one colour dominant and the others minor [7]. Pull-up peaks of the minor colours occur when a peak has reached intensity saturation such that the major peak cannot increase in signal intensity due to saturation, but minor peaks that are normally of very low intensity (background) reach appreciable signal intensity. In addition, because allele calling from capillary electrophoresis depends on detecting the emission spectra of fluorescent dyes, a spectral standard is needed to compensate for the emission spectrum overlap between dyes. Therefore, an incorrect spectral standard will also cause the pull-up effect [8].

Because most DNA fragment analysis software (aiding allele size calling, see below) provides functions to deal with the pull-up effect, such a problem can be easily overlooked. Such functions, called pull-up correction, work by removing the extra noise colours and only keeping the dye colour that the microsatellite uses for the allele peak [9]. However, because the dye colour used by the internal size standard will be performed to calibrate allele sizes, the function will not correct this dye colour. Therefore, if the dye that the microsatellite uses leads to noise colour that is the same as the colour the internal size standard uses, the pull-up correction function does not correct such noise colour. In this case, errors can still occur if researchers are unaware of the problem.

Here, we report the cause of such errors that arose when we used the HEX fluorescent dye for microsatellite loci and the ROX dye for the internal size standard and used electrophoresis in an ABI 3730 automated sequencer. Because some allele fragments happened to overlap with or were close to one of the size standard fragments, a pull-up effect caused miscalling of alleles.

2. Materials and methods

We used two microsatellite loci to reveal size scoring errors. One was locus HQ-53 (EMBL Accession No. HG421133) in *Engelhardia roxburghiana*, a diploid species belonging to the family Juglandaceae [10]. The other was locus WJ-39 (GenBank Accession No. KY428838) in Chinese tallow tree (*Triadica sebifera*), a tetraploid species belonging to the family Euphorbiaceae [11]. HQ-53 is a dinucleotide microsatellite locus, and WJ-39 is a trinucleotide microsatellite locus.

Such errors were first found at locus HQ-53 when we used it to genotype 522 *E. roxburghiana* samples in a 20-ha (400 × 500 m) DHS plot in the 1155-ha DHS National Nature Reserve on the southern verge of the Tropic of Cancer in the subtropical part of South China [12]. For this allele size calling procedure, we used HEX dye for the HQ-53 locus and ROX dye for the internal size standard. For each sample, when we double-checked the Genemapper (see below) scored electrophoresis results with other scoring programmes, some results were inconsistent. Then, when we isolated and characterised microsatellites in *T. sebifera*, a similar problem occurred at locus WJ-39 for which we used the same HEX and ROX dye combination. Therefore, we think that such allele miscalling could be a common problem if ignored.

The primers for the two loci were designed by Primer3 software [13, 14]. Therefore, for HQ-53, it occurred by chance that one amplified allele was 200 bp and the other was 198 bp after PCR amplification, sizes that are identical or close to the 200 bp size standard fragment. For WJ-39, instead of directly using the designed primers, the forward primer was 5′-tailed with the 15 bp 5′-CAGTCGGGCGTCATC-3′ sequence (CAG-tagged sequence) to decrease the cost at the microsatellite screening stage [15]. Two PCR amplification steps were then employed for this locus. For the first step, PCR amplification was performed with the CAG-tagged forward primer plus the reverse primer using 12 reaction cycles. One microlitre of amplification product was then used for the second 35-cycle PCR amplification but with the fluorescently labelled CAG-tagged sequence as the forward primer. The allele sizes after PCR amplification were 250 and 253 bp, while a 250 bp fragment also occurred in the internal size standard used.

We provided three datasets (**Table 1**). Datasets 1 and 2 are for locus HQ-53, and they include 48 and 96 samples, respectively. These datasets contain the results using the HEX dye (producing green peaks in the electropherograms; **Figures 1** and **2**), and they were electrophoresed on an ABI 3730 sequencer in 2013 and 2012. Here, we have provided two datasets in HQ-53 just to illustrate that such errors were not once-only electrophoresis problems (in fact, such errors occurred frequently in allele size scoring in locus HQ-53, and we just chose two datasets as examples). Dataset 3 is for locus WJ-39 and contains results electrophoresed on the same sequencer in 2016. Dataset 3 includes only six samples. The reason Dataset 3 only contains six samples was because we used these six samples to identify polymorphisms at locus WJ-39 before deciding to use this locus or not for large-scale genotyping in *T. sebifera*. In addition, small sample sizes were cost-saving and facilitated our use of different treatments (different dye combinations; see **Table 1**) to reveal the way to avoid such allele size miscalling. It is worth mentioning that for Treatment-5 and Treatment-6 in Dataset 3, we run each experiment (including PCR amplification and electrophoresis) twice for them on different days to check the consistency between experiments. This was because we found that the results in

Dataset	Species and its ploidy level	Sample size	Microsatellite locus and its type tested in samples	Fluorescent dye for locus	Fluorescent dye for internal size standard	Treatment
Dataset 1	*Engelhardia roxburghiana* (diploid)	48	HQ-53 (dinucleotide)	HEX	GeneScan™ 500 ROX	
Dataset 2	*Engelhardia roxburghiana* (diploid)	96	HQ-53 (dinucleotide)	HEX	GeneScan™ 500 ROX	
Dataset 3	*Triadica sebifera* (tetraploid)	6	WJ-39 (trinucleotide)	HEX	GeneScan™ 500 ROX	Treatment-1
				FAM	GeneScan™ 500 ROX	Treatment-2
				HEX	GeneScan™ 500 ROX	Treatment-3 (PCR products in Treatment-1 diluted 20-fold)
				HEX	GeneScan™ 500 ROX	Treatment-4 (PCR products in Treatment-1 diluted 50-fold)
				HEX	GeneScan™ 500 LIZ	Treatment-5 (repeated twice, named Treatment-5-1 and Treatment-5-2)
				FAM	GeneScan™ 500 LIZ	Treatment-6 (repeated twice, named Treatment-6-1 and Treatment-6-2)

Table 1. Summary of datasets used to illustrate allele size calling errors.

Treatment-5 gave many size calling errors for the first experiment (see results and **Table 2**). We then ran the second experiments to confirm that. All electrophoreses and data analyses were performed by Thermo Fisher Scientific, Inc. in Guangzhou branch, China.

Datasets 1 and 2 were analysed by Genemapper ID v3.2 software previously, while Dataset 3 was analysed by Genemapper 4.1. To make the results comparable among datasets, Datasets 1 and 2 were re-analysed with Genemapper 4.1. After checking the results in Datasets 1 and 2 with the two software versions, they were found to be identical.

Because Genemapper is expensive, most researchers cannot afford it. However, when samples are sent to companies or laboratories that have ABI sequencers, they will provide microsatellite allele size calling after electrophoresis. Therefore, for most researchers, these results

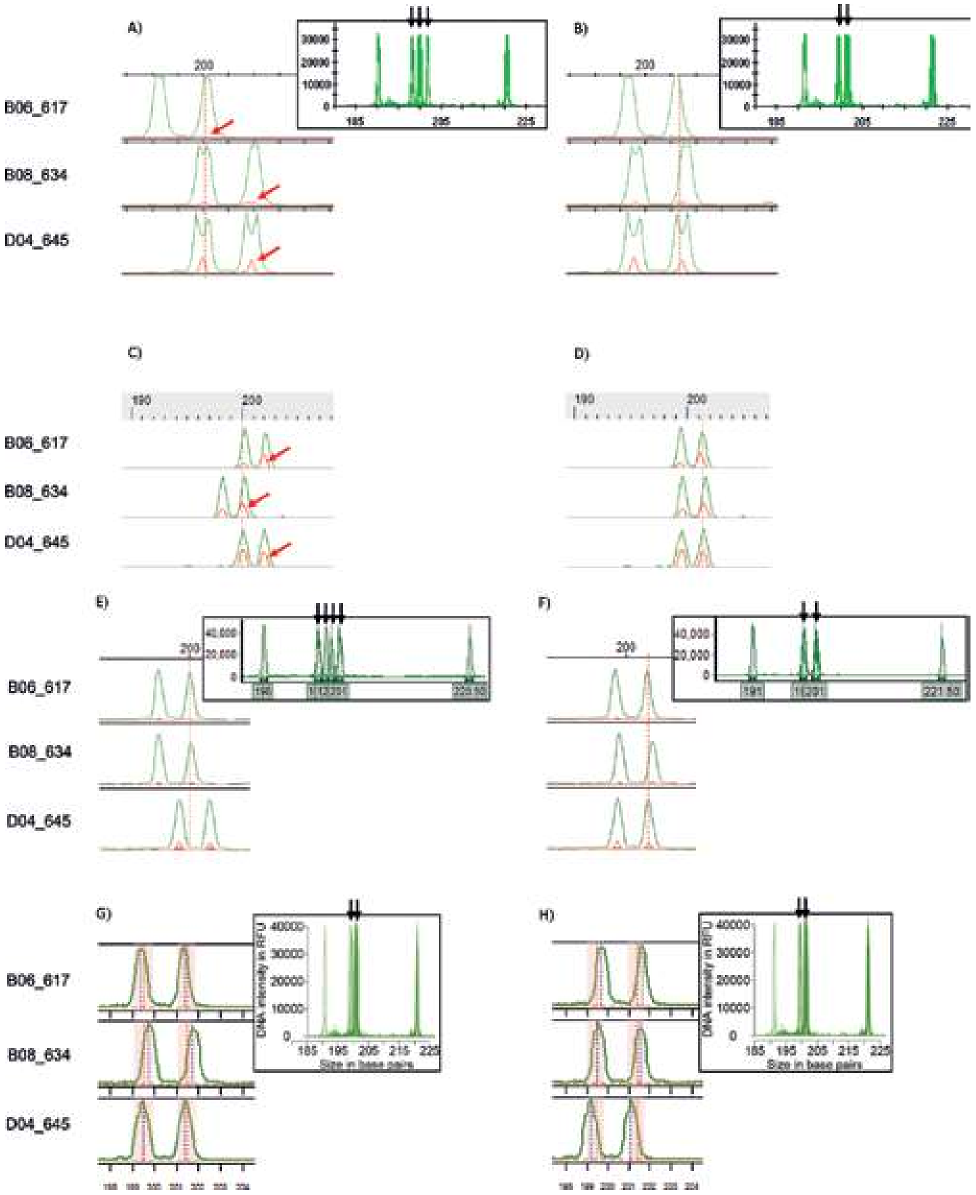


Figure 1. Electropherograms showing alleles (green peaks with HEX fluorescent dye) in example samples from Dataset 1 using Genemapper 5.0 (A–B), Gelquest 3.1.3 (C–D), Genemarker 2.7 (E–F) and Fragman 1.0.7 (G–H). The results in A, C, E and F were derived using the full set of internal size standard fragments (red peaks with ROX fluorescent dye), and B, D, F and H were scored with the 200 bp fragment omitted from the internal size standard. For each electropherogram, except those generated by Gelquest, the upper or lower right is the allele panel constructed by overlapping allele peaks. The black arrows on each allele panel correspond to the alleles shown in each electropherogram. Red-dashed lines in the electropherograms indicate the positions where the 200 bp internal size standard should appear, and the red arrows show the actual position of the 200 bp internal size standard. Sample names and electrophoresis names (in parentheses) are indicated on the left side of each electropherogram. For Genemarker, its allele panels (E and F) were constructed using only six samples because it was a demo version. [Colour figure can be downloaded and viewed at http://molecular-ecologist.com/pd.jsp?id=1#_jcp=2].

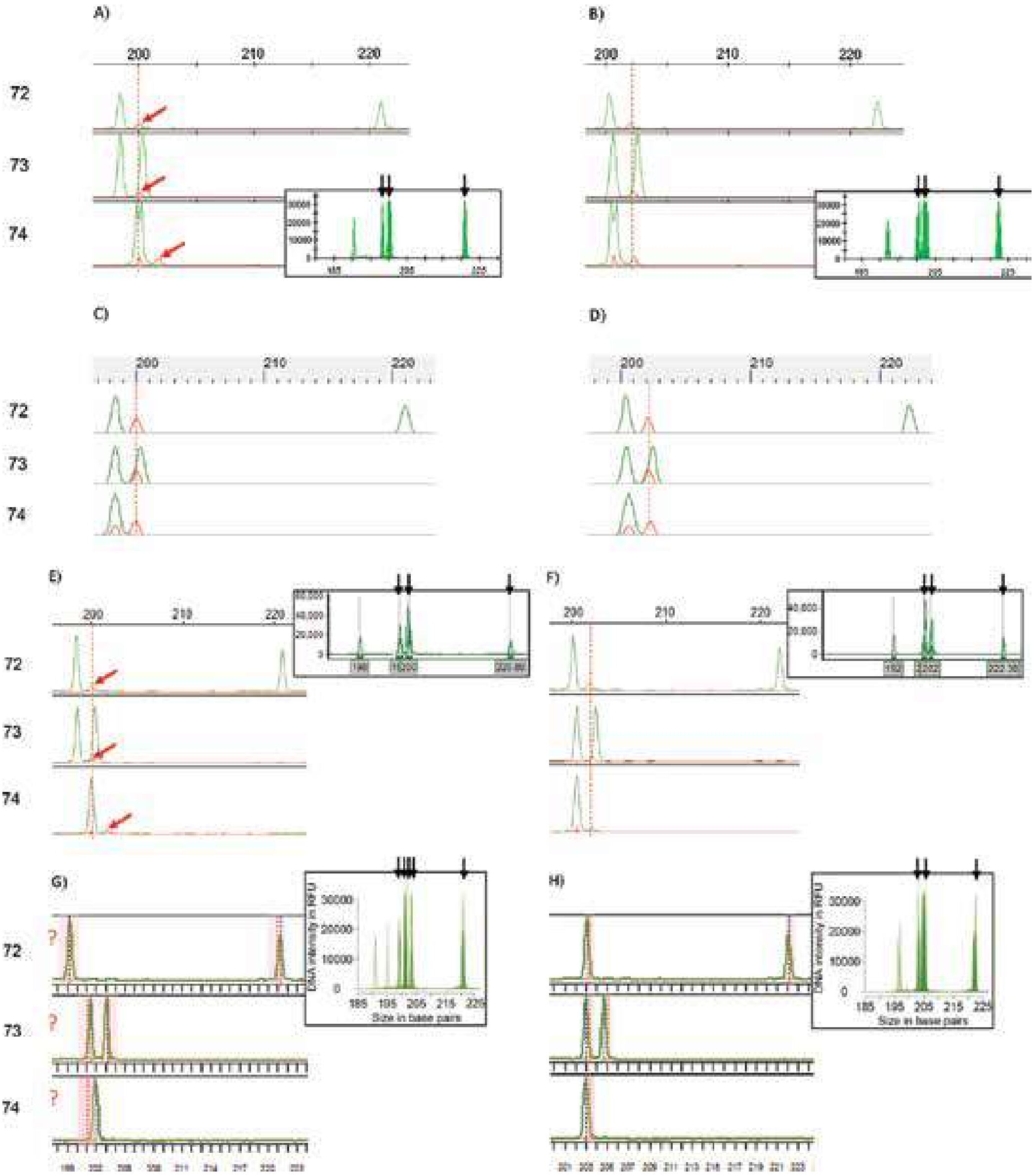


Figure 2. Electropherograms showing alleles (green peaks with HEX fluorescent dye) in example samples from Dataset 2 using Genemapper 5.0 (A–B), Gelquest 3.1.3 (C–D), Genemarker 2.7 (E–F) or Fragman 1.0.7 (G–H). The results in A, C, E and F were derived using the full internal size standard fragment set (red peaks with ROX fluorescent dye), and B, D, F and H were scored with the 200 bp fragment omitted from the internal size standard. For each electropherogram, except those generated by Gelquest, the upper or lower right was the allele panel constructed by overlapping allele peaks. The black arrows on each allele panel correspond to the alleles shown in each electropherogram. Red-dashed lines in the electropherograms indicate the positions where the 200 bp internal size standard should appear, and the red arrows show the actual position of the 200 bp internal size standard. Question marks indicate allele sizes that were not scored because the allele panel was generated improperly. Sample names and electrophoresis names (in the parentheses) are indicated on the left side of each electropherogram. For Genemarker, its allele panels (E and F) were constructed using only six samples because it was a demo version. [Colour figure can be downloaded and viewed at http://molecular-ecologist.com/pd.jsp?id=1#_jcp=2].

Dataset (treatment)	Programme and its version used to call allele size	Allele numbers generated by allele panel with full internal size standard	Allele numbers generated by allele panel with a particular internal size standard fragment omitted[a]	No. of samples with wrong allele calling[b]	Proportion of incorrect calls among samples
Dataset 1	Genemapper 5.0	5	4	4	8.3% (4/48)
	Gelquest 3.1.3	—	—	8	16.7% (8/48)
	Genemarker 2.7	6	4	4	8.3% (4/48)
	Fragman 1.0.7	4	4	0	0.0% (0/48)
Dataset 2	Genemapper 5.0	4	4	1	1.0% (1/96)
	Gelquest 3.1.3	—	—	0	0.0% (0/96)
	Genemarker 2.7	4	4	1	1.0% (1/96)
	Fragman 1.0.7	7	4	Not count	Not count
Dataset 3 (Treatment 1)	Genemapper 5.0	4	3	2	33.3% (2/6)
	Gelquest 3.1.3	—	—	0	0.0% (0/6)
	Genemarker 2.7	4	3	1	16.7% (1/6)
	Fragman 1.0.7	5	3	Not count	Not count
Dataset 3 (Treatment 2)	Genemapper 5.0	3	3	0	0% (0/6)
	Gelquest 3.1.3	—	—	1	16.7% (1/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)
Dataset 3 (Treatment 3)	Genemapper 5.0	3	3	0	0% (0/6)
	Gelquest 3.1.3	—	—	0	0% (0/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)
Dataset 3 (Treatment 4)	Genemapper 5.0	3	3	0	0% (0/6)
	Gelquest 3.1.3	—	—	0	0% (0/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)
Dataset 3 (Treatment-5-1)	Genemapper 5.0	3	3	6	100% (6/6)[c]
	Gelquest 3.1.3	—	—	0	0% (0/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)

Dataset (treatment)	Programme and its version used to call allele size	Allele numbers generated by allele panel with full internal size standard	Allele numbers generated by allele panel with a particular internal size standard fragment omitted[a]	No. of samples with wrong allele calling[b]	Proportion of incorrect calls among samples
Dataset 3 (Treatment-5-2)	Genemapper 5.0	3	3	0	0% (0/6)
	Gelquest 3.1.3	—	—	0	0% (0/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)
Dataset 3 (Treatment-6-1)	Genemapper 5.0	3	3	6	0% (0/6)[c]
	Gelquest 3.1.3	—	—	0	0% (0/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)
Dataset 3 (Treatment-6-2)	Genemapper 5.0	3	3	0	0% (0/6)
	Gelquest 3.1.3	—	—	0	0% (0/6)
	Genemarker 2.7	3	3	0	0% (0/6)
	Fragman 1.0.7	3	3	0	0% (0/6)

[a] Fragments omitted from internal size standard were 200 bp for Datasets 1 and 2 and 250 bp for Dataset 3.

[b] Wrong allele calling refers to the allele size called with full internal size standard in the programme being different from the size called by all the programmes after omitting a particular internal size standard fragment. Because the allele panel generated by the Fragman programme was doubtful, we did not score it, and the error rates were not counted in Dataset 2 and Dataset 3 (Treatment 1).

[c] There were five calling errors using Genemapper 4.1.

Table 2. Genotyping results in different datasets.

are their final "standard" results and are generally not critically assessed. Because of a version update, we could only find a Genemapper 5.0 trial version for comparing to the other allele calling software (see below). It is worth noting that Genemapper 5.0 also gave identical results to Genemapper ID v3.2 and Genemapper 4.1 using the same internal size standard except some inconsistency in Treatment-5 in Dataset 3 (see **Table 2** footnote).

We found size calling problems when we checked the consistency of the company-provided results with the calling results from Gelquest (http://www.sequentix.de/gelquest/), another DNA fragment analysis programme. In addition to these two programmes, we also used two other third-party software programmes to compare the results among them. These were Genemarker 2.7 demo version (http://www.softgenetics.com/GeneMarker.php) and the newly developed R software Fragman 1.0.7 [9].

Because of a significant bug in the latest Gelquest version 3.4.3, which meant that it could not use different size standards or adjust size standards, the previous version 3.1.3 was used. By comparing the results from the same dataset under the same internal size standard, these two versions produced the same results. The Genemarker 2.7 demo version could only allow inputting six samples at a time, so for datasets that included more than six samples, the samples were broken into several portions and analysed one portion at a time.

All datasets were analysed by the four above-mentioned software programmes, both with the full set of fragments for the internal size standard and with particular fragments (200 bp for HG-53, 250 bp for WJ-39) omitted from the internal size standard. For simplification, we omitted the version numbers of the four programmes. Therefore, if not specifically noted, they were Genemapper 5.0, Gelquest 3.1.3, Genemarker 2.7 and Fragman 1.0.7.

3. Results

For Dataset 1, using the full set of fragments in the internal size standard, Genemapper, Genemarker and Fragman produced different allele panels (**Figure 1A**, **E** and **G** and **Table 2**). Genemarker indicated that there were six alleles, Genemapper five and Fragman four. By omitting the 200 bp fragment from the internal size standard, these three programmes generated the same allele panel (**Figure 1B**, **F** and **H** and **Table 2**). However, unlike Genemapper and Genemarker, Fragman generated roughly the same size panel before and after omitting the 200 bp fragment. Without the 200 bp fragment in the internal size standard, all four programmes (including Gelquest) gave the same allele calling results. Then, considering the allele size calling results using the full set of internal size standard fragments, Genemapper generated four calling errors, Gelquest eight, Genemarker four and Fragman zero (**Table 2**).

For Dataset 2, only Fragman generated very different size panels before and after the 200 bp fragment was omitted (**Figure 2G** and **H** and **Table 2**). The allele panel generated by Fragman, with the full set of internal size standard fragments indicating that seven alleles existed, was also highly different from those generated from the other two programmes, Genemapper and Genemarker. Therefore, the allele size results called for the panel by Fragman with full fragments in the internal size standard were doubtful, and we did not score them. After comparing the results produced without the 200 bp internal size standard fragment by the four programmes, we found that they were consistent. Thus, considering the allele size calling results using the full set of internal size standard fragments, no errors occurred using Gelquest, while both Genemapper and Genemarker gave one size calling error each (**Table 2**) even though their size panels were the same before and after omitting the 200 bp fragment of the internal size standard.

For Treatment-1 in Dataset 3, comparing the size panels generated using the full set of fragments of the internal size standard with those generated using the internal size standard with the 250 bp fragment omitted, Genemapper, Genemarker and Fragman each had different panels (**Figure 3** and **Table 2**). The allele panel generated by Fragman with the full set of fragments of the internal size standard was also different from those generated by the other two programmes, making its size calling results doubtful. After omitting the 250 bp fragment, all four programmes gave consistent results. Therefore, particular size calling errors using the full set of fragments of the internal size standard were two for Genemapper, one for Genemarker and zero for Gelquest (**Table 2**).

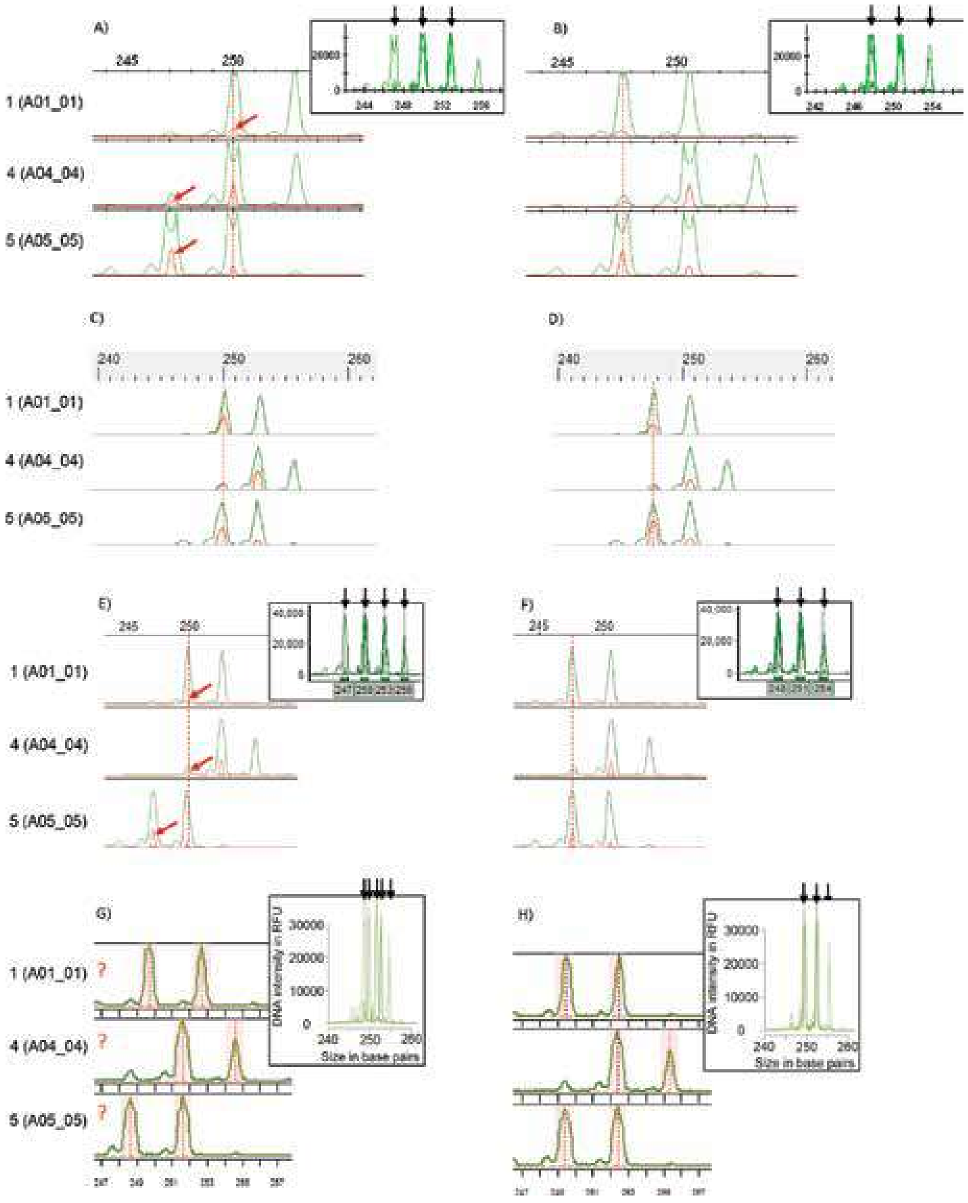


Figure 3. Electropherograms showing alleles (green peaks with HEX fluorescent dye) in example samples from Treatment-1 in Dataset 3 using Genemapper 5.0 (A–B), Gelquest 3.1.3 (C–D), Genemarker 2.7 (E–F) or Fragman 1.0.7 (G–H). The results in A, C, E and F were derived using the full internal size standard fragment set (red peaks with ROX fluorescent dye), and B, D, F and H were scored with the 250 bp fragment omitted from the internal size standard. For each electropherogram, except those generated by Gelquest, the upper or lower right was the allele panel constructed by overlapping allele peaks. The black arrows on each allele panel correspond to the alleles shown in each electropherogram.Red-dashed lines in the electropherograms indicate the position where the 250 bp internal size standard should appear, and the red arrows show the actual position of the 250 bp internal size standard. Question marks indicate allele sizes that were not scored because the allele panel was generated improperly. Sample names and electrophoresis names (in the parentheses) are indicated on the left side of each electropherogram. For Genemarker, its allele panels (E and F) were constructed using only six samples because it was a demo version. [Colour figure can be downloaded and viewed at http://molecular-ecologist.com/pd.jsp?id=1#_jcp=2].

For Treatment-2 in Dataset 3, all four programmes, regardless of whether the 250 bp fragment was omitted from the internal size standard or not, produced consistent results (see allele panels and sample examples in S1 Figure and **Table 2**).

For Treatment-3 and Treatment-4 in Dataset 3, which are the results from diluting the PCR products of Treatment-1 in Dataset 3 20- and 50-fold, consistent results were found for all four programmes regardless of whether the 250 bp fragment was omitted from the internal size standard or not (**Table 2** and S2 and S3 Figures).

For Treatment-5-1 in Dataset 3, which used HEX and LIZ in combination (**Table 1**), Genemapper, Genemarker and Fragman all produced the same allele panel pattern, indicating that three alleles existed whether the 250 bp internal size standard was omitted or not (S4 Figure and **Table 2**). However, close examination indicated that, without omitting the 250 bp internal size standard for Genemapper, its 250 bp internal size standards in all the samples were located in the wrong position, the pull-up peak position (see examples in S4 Figure), causing allele size calling errors in all of the samples (**Table 2**). The other three programmes, Gelquest, Genemarker and Fragman, scored consistent results with or without omitting the 250 bp internal size standard fragment.

For Treatment-5-2 in Dataset 3, which was the second experiment for Dataset 3 using the HEX and LIZ combination, all four programmes obtained consistent results regardless of whether the 250 bp fragment was omitted from the internal size standard or not (**Table 2** and S5 Figure).

For Treatment-6-1 and Treatment-6-2 in Dataset 3, similarly to Treatment-2, all four programmes, regardless of whether the 250 bp fragment was omitted from the internal size standard, produced consistent results (**Table 2** and S6 and S7 Figures).

4. Discussion

We report here one kind of microsatellite genotyping error caused by pull-up in capillary electrophoresis. Certainly, this is not new in microsatellite genotyping errors [7, 16]. However, to our knowledge, this is the first report that microsatellite genotyping is prone to such an effect when pull-up peaks influence the size standard match. In this case, the HEX fluorophore introduced extra signals (peaks) in the ROX fluorophore channel, causing size calibration collapse and allele miscalling. However, such a combination was not deemed to cause improper allele calling. In our case, this only occurred when the allele peak overlapped or was close to one of the internal size standard peaks.

Pull-up problems have received attention in previous studies [7, 16], and functions to diminish their influence on allele calling have been integrated into three programmes used in this study, i.e., Genemapper, Genemarker and Fragman. However, as the allele panels were constructed by these three programmes with the full set of fragments of the internal size standard (**Figures 1–3**), none of them dealt with this problem effectively. Therefore, extra signals caused by pull-up effects (extra red peaks in the internal size standard channel) were still strong and influenced the sizing calibration. For example, for sample 4(A04_04) in Treatment-1 in Dataset 3, because of the false pull-up, the 250 bp peak was much higher than the true 250 bp peak in the internal size standard channel (**Figure 4**), and Genemapper wrongly identified the false peak as the 250 bp standard and even indicated "Sizing Quality = 1.0", meaning a complete match. Therefore, the true allele pattern in this sample was shifted left due to wrong standard matching (**Figure 3A**). Similarly, for sample 74 in Dataset 2, there were two red peaks side by side around the 200 bp standard position, and Genemarker incorrectly chose the left peak (**Figure 5**). In this situation, Genemarker still gave a calibration score of 95, the highest among six samples. However, for sample 74 (**Figure 2E**, also A and C), compared to the right red peak that was independent, the left red peak occurred under the allele peak and was clearly caused by the pull-up effect.

According to the Genemapper user guide (DNA Fragment Analysis by Capillary Electrophoresis, Publication Number 4474504, Revision B, https://tools.thermofisher.com/content/sfs/manuals/4474504.pdf), for internal size matching, Genemapper used the ratio-matching method, which is based on relative height and distance of neighbouring peaks. This algorithm theoretically ignores anomalous peaks that occur between two size standard peaks (page 99 in the user guide). However, in this case, because the pull-up peaks only had a 2 or 3 bp difference from the particular size standard peak, the ratio-matching method could not differentiate them very effectively.

To address the pull-up effect that occurred above, we provide four solutions:

1. Choosing a different fluorescent dye for the internal size standard or microsatellite loci. For example, from Treatment-2 and Treatment-6 in Dataset 3 (**Table 2** and S1, S6 and S7 Figures), when we used FAM dye for the WJ-39 locus, whether ROX or LIZ dye was used in the internal size standard, the pull-up effect was not problematic, and all software produced consistent results. However, because it was not known in advance that the HEX and ROX combination or HEX and LIZ combination (as Treatment-5-1 displayed; **Table 2** and S4 Figure) would lead to a problem, and this might only have been apparent at the end of the experiment, changing dyes was therefore not cost-efficient.

2. Redesigning the primers to avoid allele sizes identical or close to the internal standard sizes. However, because we generally only used one microsatellite sequence as a template to design the primers, the whole allele pattern in the population or species remains unknown. Therefore, some unknown alleles could still have fragment lengths identical or close to the lengths of the internal size standard.

3. Avoiding overloading the PCR products in capillary electrophoresis. Overloading was the major reason for the pull-up effect. It is generally suggested that the fluorescence signal should be approximately 150–4000 relative fluorescence units (RFU) [7]. In our case, the pull-up effect was clearly caused by overloading PCR products. From the allele panels in **Figures 1–3**, the RFU values in our samples were generally higher than 20,000. Therefore, to meet the instrument requirements, it is necessary to optimise the final PCR product concentration for each locus by, for example, adjusting the DNA template concentration, PCR cycling or using post-PCR dilution (as Treatment-3 and Treatment-4 display; **Table 2** and S2 and S3 Figures). However, these steps certainly increase the cost and time of the analysis [16], especially for a laboratory without an automated sequencer instrument. Indeed, high RFU values are not uncommon. For example, in the literature of Fragman software [9], many samples displayed high RFU values (see **Figures 1**, **4** and **5** in their literature).

4. Omitting particular size standard peaks that overlap or are close to the allele peaks. Compared to the above solutions, this was both cost- and time-efficient. In Genemapper, Genemarker and Gelquest, all provided size standards with a particular size fragment are excluded (such as 250 bp and/or 340 bp) because of their abnormal migration behaviour. In this study, after omitting the 200 or 250 bp fragment, all four programmes resulted in identical allele patterns for Datasets 1–3. Furthermore, for Dataset 3, the allele pattern with 250 bp omitted using HEX dye was identical to the pattern results derived from Treatment-2 in Dataset 3 with the full set of internal size standard fragments using FAM dye. Therefore, creating a custom size standard by omitting particular fragments from the internal size standard could circumvent the pull-up problem.

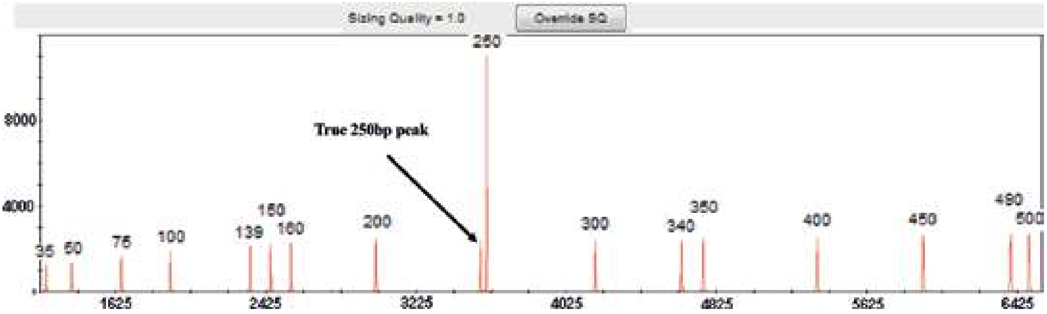


Figure 4. Size standard calibration for sample 4(A04_04) in Treatment-1 in Dataset 3 using Genemapper 5.0. [Colour figure can be downloaded and viewed at http://molecular-ecologist.com/pd.jsp?id=1#_jcp=2].

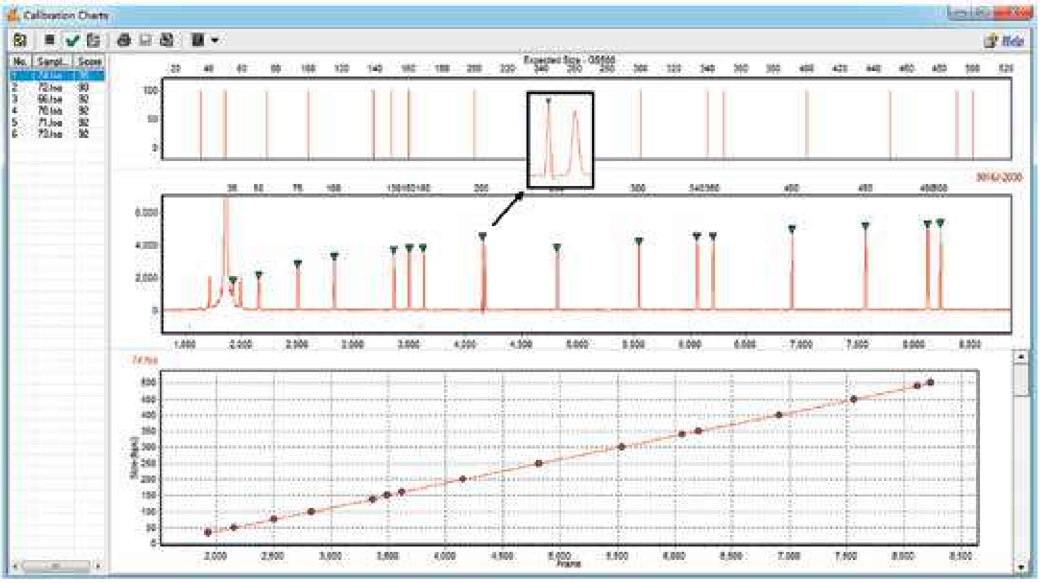


Figure 5. Size standard calibration for sample 74 in Dataset 2 using Genemarker 2.7. [Colour figure can be downloaded and viewed at http://molecular-ecologist.com/pd.jsp?id=1#_jcp=2].

5. Conclusions

This study was the first to describe one particular microsatellite allele size calling error attributed to using HEX dye and ROX dye in capillary electrophoresis. This cautions researchers to carefully assess the results of automatic allele calling. Using different software and visually scoring, each result will allow accurate sizing of microsatellite alleles. Of the four software programmes used here, both Genemapper and Genemarker are commercial, and most laboratories cannot afford them. However, compared to Genemapper, Genemarker is easier to use. Of the two free programmes, Gelquest has a graphical user interface that is easy to use and provides many user-friendly functions to help display sample alleles, while Fragman did not. Since ABI sequencers are most commonly used for analysing microsatellites, polymorphisms are generally identified by Genemapper; thus, we recommend that researchers use Gelquest as an alternative tool to check the consistency of the results.

Acknowledgements

We thank Dan-Dan Zhang, Ying Chen, Lin-Fang Wu, Yi-Fan Zhuang and others who have helped in the field and in the laboratory. This study is supported by the National Natural Science Foundation of China (Grant No. 41371078 and No. 31170196); the Science and Technology Program of Guangzhou, China (Grant No. 201707010339 and No. 201705040006); the Science and Technology Program of Guangdong Province, China (Grant No. 2016A030303003 and Grant No. 2017A030303060); and the Chinese Forest Biodiversity Monitoring Network.

Conflict of interest

The authors declare no conflict of interest.

A. Appendices and nomenclature

S1 Figure Electropherograms showing alleles in example samples of Treatment-2 in Dataset 3.

S2 Figure Electropherograms showing alleles in example samples of Treatment-3 in Dataset 3.

S3 Figure Electropherograms showing alleles in example samples of Treatment-4 in Dataset 3.

S4 Figure Electropherograms showing alleles in example samples of Treatment-5-1 in Dataset 3.

S5 Figure Electropherograms showing alleles in example samples of Treatment-5-2 in Dataset 3.

S6 Figure Electropherograms showing alleles in example samples of Treatment-6-1 in Dataset 3.

S7 Figure Electropherograms showing alleles in example samples of Treatment-6-2 in Dataset 3.

S8 Datasets Datasets 1–3 used for this study.

Appendices (S1–S8) can be downloaded from the website http://www.molecular-ecologist.com/pd.jsp?id=1#_jcp=2.

Author details

Zheng-Feng Wang[1]*, Se-Ping Dai[2], Ju-Yu Lian[1], Hong-Feng Chen[1], Wan-Hui Ye[1] and Hong-Lin Cao[1]

*Address all correspondence to: wzf@scib.ac.cn

1 Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China

2 Guangzhou Institute of Forestry and Landscape Architecture, Guangzhou, China

References

[1] Jarne P, Lagoda PJL. Microsatellites, from molecules to populations and back. Trend in Ecology & Evolution. 1996;**11**(10):424-429. DOI: 10.1016/0169-5347(96)10049-5

[2] Kantartzi SK. Micorsatellite: Methods and Protocols. New York: Humana Press; 2013. DOI: 10.1007/978-1-62703-389-3

[3] Fernando P, Evans BJ, Morales JC, Melnick DJ. Electrophoresis artefacts—A previously unrecognised cause of error in microsatellite analysis. Molecular Ecology Notes. 2001;**1**:325-328. DOI: 10.1046/j.1471-8278.2001.00083.x

[4] Hoffman JI, Amos W. Microsatellite genotyping errors: Detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology. 2005;**14**:599-612. DOI: 10.1111/j.1365-294X.2004.02419.x

[5] Dewoody J, Nason JD, Hipkins VD. Mitigating scoring errors in microsatellite data from wild populations. Molecular Ecology Notes. 2006;**6**:951-957. DOI: 10.1111/j.1471-8286.2006.01449.x

[6] Hess MA, Rhydderch JG, LeClair LL, Buckley RM, Kawase M, Hauser L. Estimation of genotyping error rate from repeat genotyping, unintentional recaptures and known parent–offspring comparisons in 16 microsatellite loci for brown rockfish (*Sebastes auriculatus*). Molecular Ecology Resources. 2012;**12**:1-10. DOI: 10.1111/1755-0998.12002

[7] Flores-Rentería L, Krohn A. Scoring microsatellite loci. In: Kantartzi SK, editor. Micorsatellite: Methods and Protocols. New York: Humana Press; 2013. pp. 319-336. DOI: 10.1007/978-1-62703-389-3.ch21

[8] Cloete K, Ristow WPG, Kasu M, D'Amato ME. Design, installation, and performance evaluation of a custom dye matrix standard for automated capillary. Electrophoresis. 2017;**38**:617-623. DOI: 10.1002/elps.201600257

[9] Covarrubias-Pazaran G, Diaz-Garcia L, Schlautman B, Salazar W, Zalapa J. Fragman: An R package for fragment analysis. BMC Genetics. 2016;**17**:62. DOI: 10.1186/s12863-016-0365-6

[10] Zhang DD, Luo P, Chen Y, Wang ZF, Ye WH, Cao HL. Isolation and characterization of 12 polymorphic microsatellite markers in *Engelhardia roxburghiana* (Juglandaceae). Silvae Genetica. 2014;**63**(3):109-112

[11] Zhuang YF, Wang ZF, Wu LF. New set of microsatellites for Chinese tallow tree, *Triadica sebifera*. Genetics and Molecular Research. 2017;**16**(2):gmr16029624. DOI: http://dx.doi.org/10.4238/gmr16029624

[12] Wang ZF, Lian JY, Ye WH, Cao HL, Zhang QM, Wang ZM. Pollen and seed flow under different predominant winds in wind-pollinated and wind-dispersed species *Engelhardia roxburghiana*. Tree genetics & Genomes. 2016;**12**:19. DOI: 10.1007/s11295-016-0973-3

[13] Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007;**23**(10):1289-1291. DOI: 10.1093/bioinformatics/btm091

[14] Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—New capabilities and interfaces. Nucleic Acids Research. 2012;**40**(15):e115

[15] Schuelke M. An economic method for the fluorescent labeling of PCR fragments. Nature Biotechnology. 2000;**18**:233-234. DOI: 10.1038/72708

[16] Dimsoski P. Development of a 17-plex microsatellite polymerase chain reaction kit for genotyping horses. Croatian Medical Journal. 2003;**44**(3):332-335

Section 2

Examples of Genotyping Applications

Chapter 4

Methods for Genotyping of the Honey Bee (*Apis mellifera* L.: Hymenoptera: Apidae) in Bulgaria

Peter Hristov, Rositsa Shumkova, Ani Georgieva,
Daniela Sirakova, Boyko Neov, Gyulnas Dzhebir and
Georgi Radoslavov

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73306

Abstract

Honey bees are insects of great biological, ecological, and economic importance. They are the subject of a variety of scientific studies. As social insects, they are a preferred and widely used model for clarification of the evolution of social behavior. Because of the haplodiploidy as a mechanism for determining gender, differentiation in the functions of individuals in the bee family, and for its economic importance, *Apis mellifera* is a significant subject of research in the fields of ontogeny, population genetics, and selective breeding. The biological significance of bees is rooted in the fact that they are main pollinators in the natural environment. About 80% of the pollination of entomophilous plants is carried out by *Apis mellifera.* In all crops, active pollination significantly increases their yields. Honey bees are a valuable economic asset due to the ensemble of their products, which include honey, bee pollen, propolis, royal jelly, and bee venom, used by humans for food and treatment. The main objective of this chapter was to describe the basic methods used for genotyping of *Apis mellifera* in Bulgaria. These techniques have been useful to produce a system of population criteria, and taxonomically important molecular markers are applicable in future activities related to the preservation and selection of the Bulgarian honey bee.

Keywords: *Apis mellifera*, Bulgarian honey bee, genotyping methods

1. Introduction

The geographic distribution of honey bee *Apis mellifera* L. includes Africa, Europe, and the Near East regions. At present, 29 subspecies of *A. mellifera* are recognized based on morphometric characters [1–4]. These subspecies are also described as "geographic races" because their distributions correspond to distinct geographic areas.

Five evolutionary lineages have been characterized based on morphometric, molecular, ecological, ethological, and physiological traits [5]. Four of those occur naturally in the Mediterranean Basin: African lineage (A), West and North European lineage (M), Southeast Europe lineage (C), and Near and Middle East lineage (O) [6–11].

On the basis of morphometric analysis, Ruttner [2] showed that *A. m. macedonica* existed as a native honey bee in Bulgaria. According to Petrov [12], in Bulgaria, there exists a local honey bee, which is called *A. m. rodopica*. This local bee inhabits the highest and the central parts of the Rhodope Mountains. Moreover, *A. m. rodopica* cannot be found in any other regions. Because of this natural isolation, these honey bees are preserved from genetic admixture and influence of other bee races. The Bulgarian honey bee has been adapted to the territory of the country's landscape and conditions, in which bee families are developing normally and do not show inclinable swarming. Due to the proven biological and productive qualities of the populations and their adaptation to the conditions specific to Bulgaria [19], a gene pool of honey bees needs to be thoroughly studied and preserved.

At present, the subjects of professional and unprofessional beekeeping in Bulgaria are over 800,000 bee colonies. About 13,000 of them are under the control of the National Beekeeping Association. The structure of the sector shows that beekeeping in the country is still extensive and scattered. About 71% of the beekeepers have up to 50 bee families, approximately 24% care for apiaries with 50 to 149 families, and only 5% have apiaries with over 150 families [20].

The genetic diversity of honey bee populations in Europe is threatened by (1) uncontrolled introgression with other subspecies into adapted local populations, (2) stresses from the changing environment due to global climate change, (3) environmental pollution, and (4) the emergence of new pathogens. To respond to these threats and be in full compliance with the requirements of the Convention on Biological Diversity, specific tasks and activities should be organized. This is to preserve the subspecies of honey bees and their ecotypes as a genetic reserve for future selective breeding activities. A lack of efforts and targeted activities for that purpose will have a devastating impact on honey bees and living nature, including the pollination of wild flora and agricultural plants. On a global scale, this would lead to a widespread decline in biodiversity and agriculture yield and would reflect on the planet's food sources, as 1/3 of human food depends on bee pollination [21].

The Bulgarian local honey bee was studied, in the past, mainly for morphoethological traits [13–15]. These investigations indicated that in almost all low-elevation parts of the country, bees mainly crossbred with *A. m. ligustica*. Besides, for more than three decades in the last century with a purpose of selective breeding, *A. m. ligustica, A. m. carnica, and A. m. caucasica* were introduced to Bulgaria. A comparative analysis established that the Bulgarian honey bee is the most productive in the particular conditions of the environment which it inhabits. With the appearance of new methods, mainly molecular markers, a population structure of local honey bees was studied [16–18]. The main techniques for elucidation of the population structure of Bulgarian honey bee are divided according to the genetic markers used:

1) allozymes or protein markers; 2) microsatellite analysis; and 3) genetic markers—derived from mitochondrial genome such as 16s rDNA, *COI*, and *ND5* gene regions.

In this paper, we will try to summarize the basic results from genotyping of the Bulgarian *Apis mellifera* and compare them with other bee populations to show that Bulgarian honey bees possess unique genetic features.

2. Allozyme analysis

The allozyme assay is based on gel electrophoresis results for the allelic variants of the respective enzyme loci. This is one of the first methods for characterization and application of molecular markers to study population genetics of honey bees [22, 23]. Comparative allozyme analysis showed [18] that a small number of loci were characterized for polymorphism in different honey bee populations and there were no fixed significant allelic differences between the subspecies. The importance of allozyme markers for population genetic characterization is the difference in allelic frequencies. An advantage of the allozyme genotyping method is that it is economical and easy to apply, and the results achieved are easy for interpretation.

Allozyme analysis is the most widely used approach to study the population structure of the Bulgarian honey bee. Many authors studied mainly six enzymic systems: malate dehydrogenase (MDH); malic enzyme (ME); esterase (EST); alkaline phosphatase (ALP); phosphoglucomutase (PGM) and hexokinase (HK) for genotyping of Bulgarian honey bees [24–27]. The honey bee samples from Bulgaria were recognized according to morphometric analysis [2] as *A. m. carnica, A. m. caucasica, A. m. ligustica*, and *A. m. macedonica*. The summarized results of reported investigations showed the presence of the following taxonomic markers that could be used in population genetic diagnosis of honey bees subspecies: MDH-1^{125} for subspecies *A. m. carnica*; EST-3^{88} for *A. m. macedonica* (of non-Bulgarian origin); and PGM^{80} for *A. m. caucasica*; HK^{87} for *A. m. macedonica* (type rodopica). The absence of ALP^{90} was specific to the bee populations of *A. m. macedonica*. On the other hand, the absence of the MDH-1^{80} allele, a high frequency of MDH-1^{65} and PGM1^{14}, and a low frequency of EST-3^{94} in the gene pool allowed researchers to make genetic comparisons within the subspecies *A. m. macedonica* and demonstrated specific characteristics of the Bulgarian honey bee. Moreover, the authors of reports [18, 19] observed a clear subdivision between *A. m. macedonica* populations from Bulgaria and the surrounding Balkan countries [28]. All honey bee populations from Bulgaria were clustered separately from the *A. m. macedonica* from other countries.

The local Bulgarian honey bee named by Petrov [22] as "rodopica" could be a different ecotype of *A. m. macedonica* (**Figure 1**); however, as mentioned by Ruttner [2], there is no indication of geographical variations within the *A. m. macedonica* subspecies.

Isozyme analyses were also used in other studies to compare honey bee population from Bulgaria and other countries of the Balkan Peninsula [17, 29–31]. The most important results of these investigations could be summarized as follows:

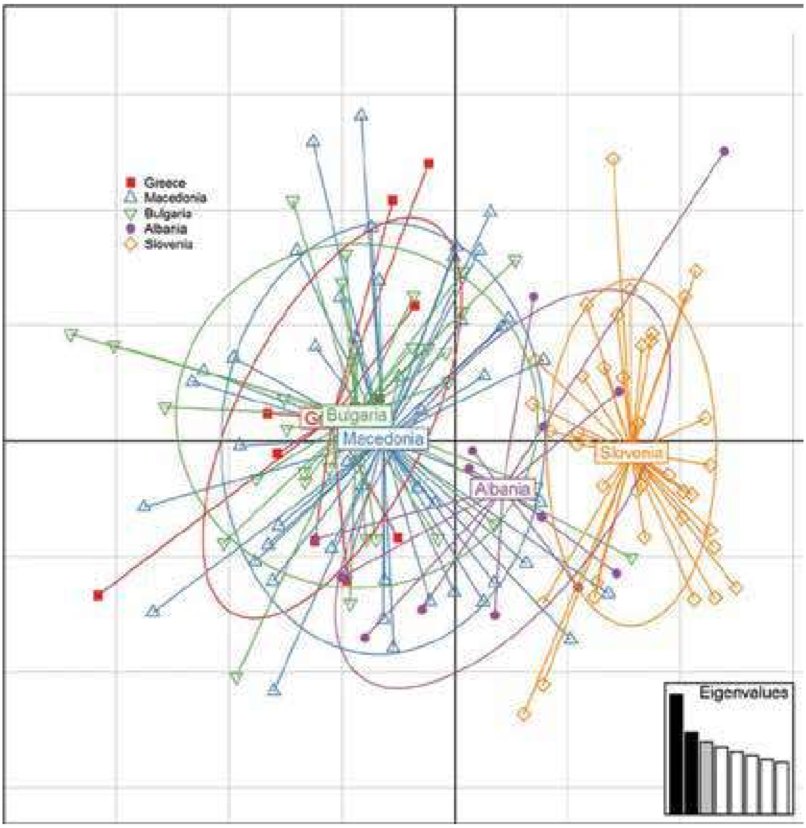


Figure 1. Principal component analysis (PCA) of the *A. m. macedonica* and *A. m. carnica*. PCA shows that *A. m. carnica* is situated in separate cluster [31].

1. The absence of the MDH-1^{80} allele from the gene pool of Bulgarian populations is an important distinctive feature, which can be used to distinguish Bulgarian honey bees from other *A. m. macedonica* populations of the Balkan Peninsula.
2. It is noteworthy that the EST-3^{88} allele was found in the gene pool only in two of the *A. m. macedonica* populations and was absent in the gene pool of *A. m. carnica*.

In this sense, this allele could be used as a taxonomic marker for both sublevel characterization and interpopulation comparisons of the subspecies *A. m. macedonica*;

1. The presence of the HK87 allele was detected only in the gene pool of the Bulgarian honey bee populations and the allele HK121 was found only in *A. m. carnica* populations of Serbia with a low occurrence rate.

For this reason, both the HK87 and the HK121 alleles could be used for comparison between the subspecies *A. m. macedonica* and *A. m. carnica* and for interpopulation comparisons at the subspecies level.

Thus, application of allozyme markers in Bulgarian honey bees supported the differences between the populations of the two subspecies *A. m. carnica* and *A. m. macedonica* at the biochemical level; therefore, they provided a clear opportunity to distinguish between Bulgarian honey bee from the Greek and Macedonian populations of *A. m. macedonica* at both taxonomic and interpopulation level within the given subspecies.

3. Microsatellite analysis

Microsatellite DNA analysis is among the most preferred and currently applied genotyping approaches for studying genetic variability within populations, population structure, as well as characterization of the phylogenetic relationships. As preferred genetic markers, microsatellites (or simple sequence repeats (SSRs)) are also successfully used for characterization of specific genetic features, comparative population-genetic analyses, such as genetic mapping, as well as DNA barcoding. Bulgarian honey bees were studied using nine microsatellite loci (**Table 1**). The results showed that all microsatellite loci were polymorphic.

On the basis of different frequencies of the alleles of the given microsatellite loci, the author concluded that these markers can be used as a genetic marker to explore the genetic diversity of local Bulgarian honey bee.

Other investigation has explored the genetic characterization of Bulgarian honey bees as well as their phylogenetic relationships with all European subspecies *Apis mellifera* using 24 microsatellite loci. The results showed that all microsatellite loci used in the study were polymorphic with a total of 260 allelic variants (**Table 2**).

Name of the locus	**Am, unified nomenclature for *A. mellifera* microsatellite markers Am001 to Am552**	**Accession no. in EBI**	**Size of the sequenced allele (bp)**	**Motifs, repeats between primers**	**Annealing temperature (°C)**	**MgCl$_2$ (mM)**
Ac011	406	AJ509637	127	(CT)19	50	1.5
A024	10	AJ509241	100	(CT)11	55	1.2
A043	25	AJ509256	140	(CT)12	55	1.5
A088	52	AJ509283	150	(CT)10(GGA)7	55	1.2
Ap226	224	AJ509455	231	(CT)8	50	1.5
Ap238	232	AJ509463	251	(AT)6(GT)3(AT)7(GA)8	50	1.5
Ap243	235	AJ509466	260	(TCC)9	50	1.5
Ap249	239	AJ509470	221	(GA)6(GA)8	50	1.5
Ap256	242	AJ509473	162	(GA)12AT(GA)3	50	1.5

Table 1. Loci, size of allele (bp), indication of the optimal annealing temperature (°C), and concentration of MgCl$_2$ (mM) in local Bulgarian honey bee populations (*A. m. macedonica*) [16].

LocusA79	Allele	*A. m. mac.* Bg	*A. m. mac.* Gr	*A. m. mac.* Mac	*A. m. car.*	*A. m. lig.*	*A. m. mel.*	*A. m. anatol.*
1	91	0.000	0.000	0.011	0.190	0.081	0.673	0.000
2	97	0.188	0.167	0.091	0.024	0.000	0.010	0.379
3	99	0.000	0.000	0.034	0.048	0.000	0.010	0.000
4	101	0.104	0.208	0.159	0.048	0.068	0.010	0.229
5	103	0.250	0.500	0.295	0.143	0.324	0.077	0.271
6	105	0.188	0.042	0.091	0.357	0.203	0.019	0.057
7	107	0.104	0.042	0.159	0.143	0.149	0.048	0.050
8	109	0.083	0.042	0.114	0.024	0.054	0.038	0.007
9	111	0.021	0.000	0.023	0.024	0.054	0.058	0.000
10	113	0.000	0.000	0.011	0.000	0.027	0.000	0.007
11	115	0.042	0.000	0.011	0.000	0.014	0.038	0.000
12	117	0.021	0.000	0.000	0.000	0.000	0.010	0.000
13	119	0.000	0.000	0.000	0.000	0.000	0.010	0.000
14	121	0.000	0.000	0.000	0.000	0.027	0.000	0.000

Table 2. Polymorphism of A79 locus: Alleles and allelic frequencies [30].

Results indicated the following: (1) low levels of genetic differentiation between the Bulgarian *A. m. macedonica* and the populations of *A. m. macedonica* originating from Greece and the Republic of Macedonia (0.009–0.017) as well as between *A. m. macedonica* and *A. m. carnica* (0.032–0.051); (2) moderate levels of genetic differentiation between population of the subspecies *A. m. macedonica* (0.071 for the Bulgarian population) and *A. m. ligustica* (0.059–0.081) and between *A. m. macedonica* and *A. m. anatoliaca* (0.087–0.110); and (3) significantly high levels of genetic differentiation between *A. m. macedonica* and *A. m. mellifera* (0.290 and 0.307, respectively).

In addition, the results are of exclusive importance to perform genetic analysis of the population of *A. m. macedonica.* Results help to finalize the comparison of the Bulgarian honey bee with the other populations of *A. m. macedonica* (Greece, Macedonia, and other Balkan countries), as well as those of the subspecies *A. m. carnica*, *A. m. ligustica*, *A. m. mellifera*, and *A. m. anatoliaca*. A comparison of honey bee species is important due to given the fact that in the modern course of development of beekeeping in Bulgaria, at certain periods (the 1960s and 1970s of the twentieth century), the introduction of bee queens of the *A. m. carnica* and *A. m. ligustica* was carried out. In that, different crossbreeding schemes between these subspecies were applied.

In this regard, Uzunov [31] investigated honey bee populations from Albania, Bulgaria, Greece, and the Republic of Macedonia in South Eastern Europe using 25 microsatellite loci.

The PCA plot showed that the Slovenian bees (or known as carnica bees) were clearly distinct from all the others *A. m. macedonica* bees (**Figure 1**).

Moreover, the results based on microsatellite analyses indeed showed a certain degree of differentiation between the bees of Bulgaria and *A. m. macedonica* from other regions. The results strongly pointed out that the Bulgarian honey bee, belonging to the *A. m. macedonica* based on classical morphometry description [31], possessed some specific genetic features, which differs from the other populations of this subspecies. This significantly justifies Bulgarian honey bees as a different ecotype adapted to the specific conditions in the territory of our country. To verify this, detailed analyses of the honey bee populations in Bulgaria with all neighboring geographical regions will be needed.

The introgression of *A. m. carnica* alleles into Bulgarian bees has been detected. The influence of the type bees could be explained by past imports of *carnica* honey bee queens, given its wide commercial spread across Europe. Thus, imports of honey bee queen from other subspecies inevitably influenced genetic diversity of native Bulgarian honey bees.

In summary, results from microsatellite DNA analysis in complex with classic morphometric analysis could be used as evidence in support of an earlier hypothesis. The local honey bee of Bulgaria belongs to subspecies *A. m. macedonica,* but it represents a different ecotype (ecotype rodopica), adapted to the specific conditions of our country. The local honey bee of Bulgaria differs from the other populations of the same subspecies by a set of population genetic indices.

4. Mitochondrial genetic markers

4.1. PCR-RFLP analysis of 16s rDNA, *COI*, and *ND5* gene regions

Mitochondrial DNA (mtDNA) analysis is also used for characterization of genetic diversity among *A. mellifera* subspecies and its ecotypes [6, 32–36]. The mitochondrial genome is determined as extremely suitable for population genetic studies of *A. mellifera,* as well as to conduct phylogenetic studies, comparisons and analyses of the entire *Apis* family. Because of maternal heredity manifested by mtDNA in honey bees, all bee workers and drones in the bee colony share similar mitochondrial haplotype [37].

The mtDNA variants, as well as the morphometric characteristics of honey bees, are currently associated with the characterization of their biogeographic zoning. It contributes for the discrimination of previously described evolutionary groups of *Apis mellifera* [38, 39].

Methods of hybridization in selective breeding may affect the distribution of mtDNA variants of the gene pool of local populations. Such changes affecting population structure (especially introgression) cannot be detected by morphometric analysis [34]. Because all the individuals of the bee family have the same haplotype, a number of authors assumed mtDNA as the suitable genetic marker for population genetic research. In the application of mitochondrial DNA analysis, however, it should be borne in mind that the results can be significantly influenced by imports of bee queens of different origins [40].

In applying mtDNA analysis, it is important to note the possibility that different haplotypes are distinguished and grouped according to classical morphological analyses [39, 41] and the geographical distribution of the subspecies *A. mellifera* [42, 43].

In this aspect, a PCR-RLFP analysis of three gene regions (16s rDNA, *COI*, and *ND5*) was applied in native Bulgarian honey bee *A. m. macedonica* on the territory of the whole country [22]. The results from this study showed that the size of PCR-amplified mtDNA products for all population of the Bulgarian honey bee is of the order of 964 bp, 1028 bp, and 822 bp for the 16s rDNA, *COI*, and *ND5* gene, respectively.

With regard to the restriction profile of 16s rDNA gene segment, it was found that Sau3A I, Ssp I, Hinc II, and EcoR I, are useful for recognize fragment sites (**Table 3**).

With respect to the *COI* gene segment, Sau3A I has three restriction sites, but Fok I and Bcl I have two restriction sites (**Table 4**).

With respect to the *ND5* gene segment, Dra I and Taq I identified two sites of recognition, Nla III—one and Ssp I—three restriction sites. The Bulgarian populations of *A. m. macedonica* included in this analysis did not have intra- and interpopulation variability using *ND5* polymorphisms (**Table 5**).

The results from this study [22] were compared with other similar investigations for native honey bee populations from Greece, Cyprus [24] and Crete Island [42]. It can be concluded

Sau3A **I**	***Ssp*** **I**	***Dra*** **I**	***Hinc*** **II**	***EcoR*** **I**
B	A	A	A	A
548	628	964	598	492
416	336		366	472

Table 3. Size (in base pairs—bp) determined by the 16s rDNA gene in the analyzed Bulgarian populations of *A. m. macedonica* [22].

Nco **I**	***Sau3A*** **I**	***Fok*** **I**	***Bcl*** **I**	***Ssp*** **I**	***Sty*** **I**	***BstU*** **I**	***Xho*** **I**
B	A	A	A	A	A	B	A
1028	371	476	465	1028	1028	1028	1028
	349	425	326				
	280	127	237				
	28						

Table 4. Fragment size estimates (in base pairs) of all fragment patterns observed on *COI* gene region among the Bulgarian populations of *A. m. macedonica* [22].

Dra I	*Taq* I	*Nla* III	*Hinc* II	*Fok* I	*Ssp* I
A	A	A	A	A	A
429	375	585	822	822	385
285	258	237			206
108	189				124
					107

Table 5. Fragment size estimates (in base pairs) of all fragment patterns observed on *ND5* gene region among the Bulgarian populations of *A. m. macedonica* [22].

that the studied Bulgarian bee populations of honey bees of the *A. m. macedonica* differ from Greek ones. The mtDNA profile of local populations of *A. mellifera macedonica* from Greece with regard to the *COI* segment after digestion with endonucleases Sty I and Nco I [36] was not found in mitochondrial profile of Bulgarian populations.

Comparing the results of Ivanova [20] with results from the studies of Bouga et al. [42], the difference between the Bulgarian and Greek populations of *A. m. macedonica* was reported. In that, a restriction profile of *COI* after Ssp I digestion, as well as *ND5* the gene segment after digestion with Hinc II and Fok I restriction enzymes, was studied.

Thus, the results from PCR-RFLP analysis of three mitochondrial gene regions such as16s rDNA, *COI*, and *ND5* did not reveal any different restriction profile in Bulgarian native honey bee *A. m. macedonica* populations. The difference existed in comparison with other honey bee populations of *A. m. macedonica* from other Balkan countries [44].

4.2. Single nucleotide polymorphism (SNP) assay followed by direct sequencing of mitochondrial *COI* and *ND5* gene regions

Single nucleotide polymorphism (SNP) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a single nucleotide (A, T, C or G) in the genome differs between members of a biological species or paired chromosomes. SNPs within a coding sequence do not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code.

One of the widely used methods for searching of unknown SNPs is PCR analysis followed by direct sequencing [37]. This technique is based on the amplification of specific parts of the genome in the PCR reaction and sequencing of the obtained sequence amplification product. The great advantage of SNP polymorphism is its universality of the genome between particular species and highly efficient identification of polymorphism within the tested sequence [37]. The high density of SNPs markers in the genome leads to their extensive utilization in genetic and population analyses.

In this aspect, the first investigation examined the phylogenetic relationships and gene flow as a result of migratory beekeeping among honey bee populations from various areas of Albania, Bulgaria, Cyprus, Greece, Italy, Slovenia, and Turkey using sequencing analysis of two mitochondrial regions. In that, the *ND5* and the *COI* gene segments were studied [43].

The results showed that among the 10 races and commercial strains studied, 7 different haplotypes were revealed for *COI*, 8 for *ND5*, and 12 for the combined dataset (**Table 6**). It should also be noted that the most common combined haplotype (haplotype 1) included bees from nearly all the different races examined (including Bulgarian honey bee) except *A. m. ligustica*. These data showed that the different races could not be discriminated, as it was known that all belonged to the East Mediterranean C lineage [8, 9, 46, 47].

Another investigation provided additional information about genetic profile of local Bulgarian honey bee [46]. This research presented, for the first time, SNP analysis of a *COI* mitochondrial DNA (mtDNA) gene segment of the local Rhodope Mountains honey bee (*Apis mellifera rodopica*). The results showed, unexpectedly, that 14 haplotypes possess a second DNA fragment (named as D1) which is not part of *COI* gene. Its length is 768 bps from the beginning of 1343 bps of reference sequence (*A. m. ligustica*; Acc. No. L06178) [48]. This fragment possibly affects translation of cytochrome c oxidase I protein due to disruption of open reading frame at the C-end of protein by multiple stop codons (**Figure 2**).

Investigation of homology of mutated D1 fragment in GenBank [48] database did not show any homology, except this with a fragment of *A. mellifera COI* gene starting from position 264 bps to 346 bps (82 bps length). Such type of gene reorganization is a possible explanation

Haplotype 1	*A. m. cypria* (PYR, KOR, DAL1 and 2), Aegean race near to *A. m. adami* (LMN1, 2, 3, 4, NIS 1, 2), *A. m. adami* (CRE4), *A .m. macedonica* (MAC3, 4, 5, 6 and 7, THR1, 2, 3 and 4, JIJ), commercial strain (genetically improved *A. m. cypria* (SB), *A. m. cecropia* (SAR3, MES1, 3, 4), *A. m. anatoliaca* (BAR), *A. m. meda* (OSM), *A. m. carnica* (LUB1 and 2)
Haplotype 2	*A. m. ligustica* (PER1 and 2, RAV1 and 2)
Haplotype 3	*A. m. cecropia* (SAR1), *A. m. carnica* (KEF2)
Haplotype 4	*A. m. cecropia* (SAR2)
Haplotype 5	*A. m. macedonica* (MAC1 and 2)
Haplotype 6	Aegean race near to *A. m. adami* (CHI1 and 2 and RHD1)
Haplotype 7	*A. m. adami* (CRE2 and 3)
Haplotype 8	*A. m. cecropia* (MES2), Aegean race near to *A. m. adami* (RHD2, 3, 4), *A. m. ligustica* (FOR)
Haplotype 9	*A. m. carnica* (LIT)
Haplotype 10	*A. m. adami* (CRE1)
Haplotype 11	*A. m. carnica* (GOR)
Haplotype 12	*A. m. cecropia* (LAR1 and 2)

Abbreviations: Turkey—Osmaniye (OSM); Bartin (BAR); Albania Sarandë (SAR); Slovenia—Litija (LIT), Gorenjska (GOR), Ljubljana (LUB); Bulgaria—Jijevo (JIJ), Plovdiv (PLV); Cyprus—Pyrgas (PYR), Kornos (KOR), Dali (DAL); Italy—Ravenna (RAV), Perugia (PER), Forli (FOR); Greece–Macedonia (MAC) (including Chalkidiki), Thrace (THR), Crete island (CRE), Larissa (Central Greece) (LAR), Messinia (Peloponnese) (MES), Kefalonia (Ionian islands) (KEF), Chios (Aegean Sea Island) (CHI), Rhodes (Aegean Sea Island) (RHD), Nisyros (Aegean Sea Island) (NIS), Limnos (Aegean Sea Island) (LMN).

Table 6. Honey bee populations studied grouped in 12 haplotypes [43].

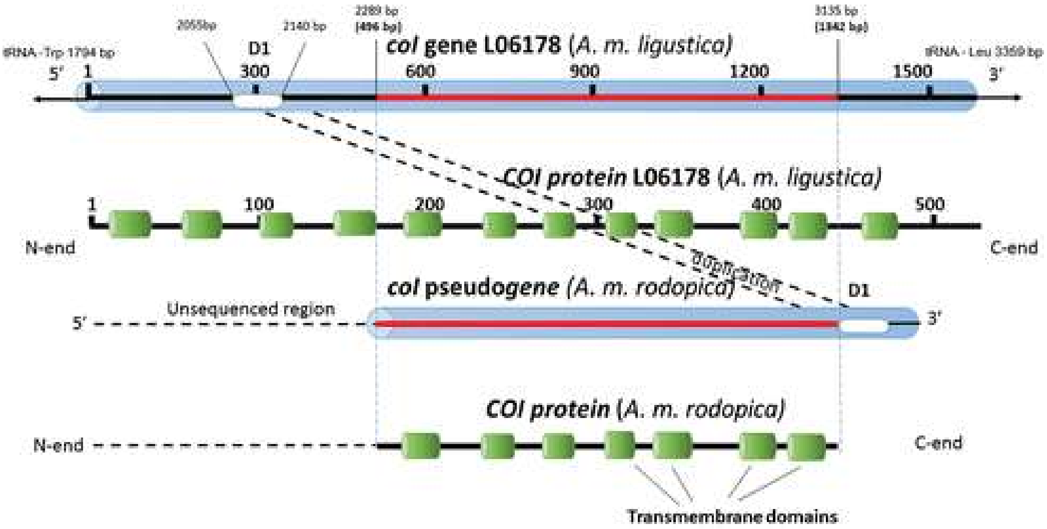


Figure 2. *COI* pseudogene organization and effects of potential translation of the COI protein on *Apis mellifera rodopica* [45].

of the duplication of this D1 fragment of *A. mellifera* originated from Rhodope Mountains, but the mechanism of this duplication event needs further investigation. The authors of this study [46] concluded that this specific characteristic duplication of the *COI* gene for the Rhodopes honey bee (*Apis mellifera rodopica*) can be used as a genetic marker to discriminate the local Bulgarian honey bees and to support the related conservation activities.

5. Conclusion

The conducted experiments and obtained results on the basis of various genetic methods for genotyping enabled the detailed characterization of genetic diversity among the studied bee populations of Bulgaria. The application of complex comparative genetic analysis provided genetic evidence on belonging of the Bulgarian honey bee to *Apis mellifera macedonica* as well as the presence of specific genetic characteristics that define them as a different local ecotype for Bulgaria, named previously as "rodopica" [26].

The genetic diversity studies provided summarized, systematized, and enriched information on genetic polymorphism in Bulgarian honey bee populations. Scientific efforts characterized the level of genetic differentiation relative to other populations of the *Apis mellifera macedonica* on the Balkan Peninsula and to other subspecies of *Apis mellifera*, in the territory of Europe.

The summarized population genetic studies from genotyping of Bulgarian honey bee also provided significant information on the availability of reliable genetic markers, applicable to the formation of a system of conservation activities for the national genetic resources of Bulgarian honey bee population.

Acknowledgements

The authors express gratitude to the National Science Fund of the Bulgarian Ministry of Education and Science, Sofia, for the financial support (grant number 06/10 17.12.2016).

Conflict of interest

No potential conflict of interest was reported by the authors.

Author details

Peter Hristov[3]*, Rositsa Shumkova[1], Ani Georgieva[2], Daniela Sirakova[3], Boyko Neov[3], Gyulnas Dzhebir[4] and Georgi Radoslavov[3]

*Address all correspondence to: peter_hristoff@abv.bg

1 Agricultural and Stockbreeding Experimental Station, Agricultural Academy, Smolyan, Bulgaria

2 Department of Pathology, Institute of Experimental Morphology, Pathology and Morphology and Anthropology with Museum, Bulgarian Academy of Sciences, Sofia, Bulgaria

3 Department of Animal Diversity and Resources, Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria

4 Department of Structure and Function of Chromatin, Institute of Molecular Biology, Bulgarian Academy of Sciences, Sofia, Bulgaria

References

[1] Engel M. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; Apis). Journal of Hymenoptera Research. 1999;**8**:165-196

[2] Ruttner F. Biogeography and Taxonomy of Honey Bees. 1st ed. Berlin: Springer-Verlag; 1988. p. 284

[3] Sheppard WS, Arias MC, Greech A, Meixner MD. *Apis mellifera ruttneri*, a new honey bee sub-species from Malta. Apidologie. 1997;**28**:287-293. DOI: 10.1051/apido:19970505

[4] Sheppard WS, Meixner MD. *Apis mellifera pomonella*, a new honey bee sub-species from Central Asia. Apidologie. 2003;**34**:367-375. DOI: 10.1051/apido:2003037

[5] De La Rúa P, Fuchs S, Serrano J. Biogeography of European honey bees. In: Lodesani M, Costa C, editors. Beekeeping and Conserving Biodiversity of Honey Bees. Sustainable

Bee Breeding. Theoretical and Practical Guide. Hebden Bridge, UK: Northern Bee Books; 2005. pp. 15-52. DOI: 10.1051/apido/2009027

[6] Arias MC, Sheppard WS. Molecular phylogenetics of honey bee subspecies (*Apis mellifer*a L.) inferred from mitochondrial DNA sequences. Molecular Phylogenetic and Evolution. 1996;**5**:557-566. DOI: 10.1006/mpev.1996.0050

[7] Cánovas F, De La Rúa P, Serrano J, Galián J. Geographical patterns of mitochondrial DNA variation in *Apis mellifera iberiensis* (Hymenoptera: Apidae). Journal of Zoological Systematics and Evolutionary Research. 2008;**46**:24-30. DOI: 10.1111/j.1439-0469.2007.00435.x

[8] Franck P, Garnery L, Celebrano G, Solignac M, Cornuet JM. Hybrid origin of honey bees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). Molecular Ecology. 2000;**9**:907-921. DOI: 10.1046/j.1365-294x.2000.00945.x

[9] Franck P, Garnery L, Loiseau A, Oldroyd BP, Hepburn HR, Solignac M, Cornuet JM. Genetic diversity of the honeybee in Africa: Microsatellite and mitochondrial data. Heredity. 2001;**86**:420-430. DOI: 10.1046/j.1365-2540.2001.00842.x

[10] Garnery L, Solignac M, Celebrano G, Cornuet JM. A simple test using restricted PCR-amplified mitochondrial DNA to study the genetic structure of *Apis Mellifera* L. Experientia. 1993;**49**:1016-1021. DOI: 10.1007/BF02125651

[11] Miguel I, Iriondo M, Garnery L, Sheppard WS, Estonba A. Gene flow within the M evolutionary lineage of *Apis mellifera*: Role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the Western Europe. Apidologie. 2007;**38**:141-155. DOI: 10.1051/apido:2007007

[12] Petrov P. Bulgarian honey bee *Apis mellifica rodopica* and it race standard. Agrarian University of Plovdiv, Scientific works. 1995;**XL**(3):317-319

[13] Bouga M, Alaux C, Bienkowska M, Büchler R, Carreck NL, Cauia E, Wilde J. A review of methods for discrimination of honey bee populations as applied to European beekeeping. Journal of Apicultural Research. 2011;**50**:51-84. DOI: 10.3896/IBRA.1.50.1.06

[14] Tzonev I. Biometric investigations on the *A. mellifera* (methods, season dynamics and family changeability). News of the Student Scientist Society. 1960:157-191. There is no DOI

[15] Velichkov V. Honey bee races in Bulgaria. Beekeeping. 1970;**10**:7-11

[16] Nikolova S. Genetic variability of local Bulgarian honey bees *Apis mellifera macedonica* (*rodopica*) based on microsatellite DNA analysis. Journal of Apicultural Sciences. 2011; **55**:117-129

[17] Ivanova E, Staykova T, Bouga M. Allozyme variability in honey bee populations from some mountainous regions in southwest of Bulgaria. Journal of Apicultural Research. 2007;**46**:3-8. DOI: 10.3896/IBRA.1.46.1.02

[18] Ivanova E, Staykova T, Petrov P. Allozyme variability in populations of local Bulgarian honey bee. Biotechnology and Biotechnological Equipment. 2010;**24**:371-374. DOI: 10.1080/13102818.2010.10817868

[19] Ivanova E, Staykova T, Stoyanov I, Petrov P. Allozyme genetic polymorphism in Bulgarian honey bee (*Apis mellifera* L.) populations from the south-eastern part of the Rhodopes. Journal of BioScience and Biotechnology. 2012;**1**:45-49

[20] Ivanova E. Population-genetic variability of *Apis mellifera* L. in Bulgaria [thesis]. Plovdiv: Plovdiv University; 2017

[21] Aizen MA, Garibaldi LA, Cunningham SA, Klein AM. How much does agriculture depend on pollinators? Lessons from long-term trends in crop production. Annals of Botany. 2009;**103**:1579-1588. DOI: 10.1093/aob/mcp076

[22] Petrov P, Ganev G. Breeding Program for Preserving Local Bulgarian Honey Bee. Plovdiv; 2013. p. 54

[23] Mestriner MA. Biochemical polymorphisms in bees (*Apis mellifera ligustica*). Nature. 1969;**223**:188-189. DOI: 10.1038/223188a0

[24] Mestriner RA, Contel EPB. The P-3 and Est loci in the honey bee *Apis mellifera*. Genetics. 1972;**72**:733-738 There is no DOI

[25] Ivanova E. Additional information on allozyme variability of honey bees, *Apis mellifera* Linnaeus, 1758, from Bulgaria. Acta Zoologica Bulgarica. 2015;**67**:573-578

[26] Georgieva V, Ivanova E, Petrov P, Petkov N. Genetic characterization of *Apis mellifera macedonica* (type *rodopica*) populations selectively controlled in Bulgaria. Journal of Central European Agriculture. 2016;**17**:620-628. DOI: 10.5513/JCEA01/17.3.1753

[27] Bouga M, Kilias G, Harizanis PC, Papasotiropoulos V, Alahiotis S. Allozyme variability and phylogenetic relationships in honey bee (Hymenoptera: Apidae: *A. mellifera*) populations from Greece and Cyprus. Biochemical Genetics. 2005;**43**:471-484. DOI: 10.1007/s10528-005-8163-2

[28] Kandemir I, Kence A. Allozyme variation in a central Anatolian honey bee (*Apis mellifera* L.) population. Apidologie. 1995;**26**:503-510. DOI: 10.1051/apido:19950607

[29] Ivanova EN, Bienkowska M, Petrov PP. Allozyme polymorphism and phylogenetic relationships in Apis Mellifera subspecies selectively reared in Poland and Bulgaria. Folia Biologica (Krakow). 2011;**59**:121-126. DOI: 10.3409/fb59_3-4.121-126

[30] Petrov P. Possibilities for using some quantitative characteristics in the taxonomy of Bulgarian honey bees *Apis mellifera rodopica*. Size of the forewing. Journal of Animal Sciences. 1996;**4**:75-77

[31] Uzunov A, Meixner M, Kiprijanovska H, Andonov S, Gregorc A, Ivanova E, Bouga M, Dobi P, Büchler R, Francis R, Kryger P. Genetic structure of *Apis mellifera macedonica* in the Balkan Peninsula based on microsatellite DNA polymorphism. Journal of Apicultural Research. 2014;**53**:288-295. DOI: 10.3896/IBRA.1.53.2.10

[32] Smith DR, Brown WM. Restriction endonuclease cleavage site and length polymorphisms in mitochondrial DNA of *Apis mellifera mellifera* and *A. m. carnica*(Hymenoptera: Apidae). Annals of the Entomological Society of America. 1990;**83**:81-88. DOI: 10.1093/aesa/83.1.81

[33] De La Rúa P, Simon UE, Tide AC, Moritz RFA, Fuchs S. MtDNA variation in *Apis cerana* populations from the Philippines. Heredity. 2000;**84**:124-130. DOI: 10.1046/j.1365-2540.2000.00646.x

[34] Cornuet JM, Garnery L. Mitochondrial DNA variability in honeybees and its phylogeographic implications. Apidologie. 1991;**22**:627-642. DOI: 10.1051/apido:19910606

[35] Meusel MS, Moritz RFA. Transfer of paternal mitochondrial DNA in fertilization of honeybees (*Apis mellifera* L.) eggs. Current Genetics. 1993;**24**:539-543. DOI: 10.1007/BF00351719

[36] Garnery L, Franck P, Baudry E, Vautrin D, Cornuet JM, Solignac M. Genetic diversity of the west European honey bee (*Apis mellifera mellifera* and *Apis m. iberica*), II microsatellite loci. Genetics Selection Evolution. 1998;**30**:49-74. DOI: 10.1186/1297-9686-30-S1-S49

[37] Jiang Z, Wang H, Michal JJ, Zhou X, Liu B, Woods LCS, Fuchs RA. Genome wide sampling sequencing for SNP genotyping: Methods, challenges and future development. International Journal of Biological Sciences. 2016;**12**(1):100-108. DOI: 10.7150/ijbs.13498

[38] Pedersen BV. On the phylogenetic position of the Danish strain of the black honeybee (the Laeso bee), *Apis mellifera mellifera* L. (Hymenoptera: Apidae) inferred from mitochondrial DNA sequences. Entomologica Scandinavica. 1996;**27**:241-250. DOI: 10.1163/187631296X00070

[39] Smith DR, Palapoli MF, Taylor BR, Garnery L, Cornuet JM, Solignac M, Brown WM. Geographical overlap of two mitochondrial genomes in Spanish honeybee (*Apis mellifera iberica*). Journal of Heredity. 1991;**82**:96-100. DOI: 10.1093/oxfordjournals.jhered.a111062

[40] Garnery L, Cornuet JM, Solignac M. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. Molecular Ecology. 1992;**1**:145-154. DOI: 10.1111/j.1365-294X.1992.tb00170.x

[41] Ruttner F, Tassencourt L, Louveaux J. Biometrical-statistical analysis of the geographic variability of *Apis mellifera* L. I. Material and methods. Apidologie. 1978;**9**:363-381. DOI: 10.1051/apido:19780408

[42] Bouga M, Harizanis PB, Kilias G, Alahiotis S. Genetic divergence and phylogenetic relationships of honey bee *Apis mellifera* (Hymenoptera: Apidae) populations from Greece and Cyprus using PCR-RFLP analysis of three mtDNA segments. Apidologie. 2005;**36**:335-344. DOI: 10.1051/apido:2005021

[43] Harizanis PC, Nielsen DI, Bouga M. Diagnostic molecular markers discriminating Africanized honey bee from Greek and Cypriot honey bees (*Apis mellifera*, Hymenoptera: Apidae). Journal of Apicultural Research. 2006;**45**:197-202. DOI: 10.3896/IBRA.1.45.4.05

[44] Martimianakis S, Klossa-Kilia E, Bouga M, Kilias G. Phylogenetic relationships of Greek *Apis mellifera* subspecies based on sequencing of mtDNA segments (*COI* and *ND5*). Journal of Apicultural Research. 2011;**50**:42-50. DOI: 10.3896/IBRA.1.50.1.05

[45] Kandemir I, Kence M, Sheppard WS, Kence A. Mitochondrial DNA variation in honey bee (*Apis mellifera* L.) populations from Turkey. Journal of Apicultural Research. 2006;**45**: 33-38. DOI: 10.3896/IBRA.1.45.1.08

[46] Ozdil LF, Yildiz MA, Hall HG. Molecular characterization of Turkish honey bee populations (*Apis mellifera* L.) inferred from mitochondrial DNA RFLP and sequence results. Apidologie. 2009;**40**:570-576. DOI: 10.1051/apido/2009032

[47] Radoslavov G, Hristov P, Shumkova R, Mitkov I, Sirakova D, Bouga M. A specific genetic marker for the discrimination of native Bulgarian honey bees (*Apis mellifera rodopica*): Duplication of *COI* gene fragment. Journal of Apicultural Research. 2017;**56**(3):196-202. DOI: 10.1080/00218839.2017.1307713

[48] Crozier RH, Crozier YC. The mitochondrial genome of the honey bee *Apis mellifera*: Complete sequence and genome organization. Genetics. 1993;**133**:97-117

Chapter 5

Molecular Based Method Using PCR Technology on Porcine Derivative Detection for Halal Authentication

Yuny Erwanto

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76071

Abstract

Halal food and pharmaceuticals were taken into account as food and pharmaceutical products permitted to be consumed by Muslim according to Sharia law. Due to the development of science and technology, especially in food and pharmaceutical industry, some industries use non-halal components such as pig derivatives in food and pharmaceutical products to reduce the production cost. Non-halal components added in food and pharmaceutical products are difficult to detect visually due to close similarity between non-halal and halal components present in food and pharmaceutical products. Some of the methods already developed in the laboratory include spectroscopic methods using infrared radiation, Fourier-transform infrared spectroscopy (FTIR) and nuclear magnetic resonance spectroscopy, chromatography-based methods, electronic nose, DNA-based and differential scanning calorimetric for pig component analysis. Food and pharmaceutical matrix is very complex to be analyzed; therefore, the signals obtained during chemical and biological analyses are very complex and are difficult to interpret. Even though the spectroscopy- and chromatography-based methods are able to determine the pig derivative component, there are some difficulties for the application in the field of blind samples, caused by the complex matrix of food or pharmaceutical. Among analytical methods, the polymerase chain reaction (PCR) based on the molecular genetic analysis was believed to be highly sensitive, valid and judgable as well as reliable for the analytical instrument. Therefore, this chapter describes some methodologies based on DNA technology such as conventional PCR using universal primer through restriction fragment length polymorphism or specific primer. This chapter also gives detailed information on the application of the real-time PCR using species-specific primer for porcine determination as well as for halal authentication.

Keywords: food, halal, molecular methods, pharmaceutical, pig derivative detection

1. Introduction

Today most of the Moslem countries are still growing in terms of life style, traveling, food and so on. Food is the primary need for humankind and cannot be separated from life; consequently, everywhere human beings live, including Moslems, they need food for living. In addition, awareness among Muslims on the need and necessity to consume only halal food is annually increasing, so the global market of halal food will be predominant in the future. It is understood that the production of halal food is not only beneficial to Muslims but also to food producers, since it will increase market space for Muslims especially and non-Muslims also who are interested in consuming halal food for health reasons. Human food production starts from the ingredients chosen, preparing the equipment, processing methods, packaging and labeling.

Human mobility which is increasing requires the availability of food according to the needs of each individual. Food security extends not only to safety and health issues but also to religious considerations. In the global realm, the need of tools to ensure that food is consumed safely and is a guarantee that it is free from non-halal materials requires a technology and non-halal detection application. This chapter is discussed in relation to some reliable methods of DNA-based analysis that is needed in justifying the presence of non-halal materials for Muslim consumers, in particular, the ingredients derived from pigs which are materials prohibited in Islam for consumption.

Identification of the presence of non-halal compounds becomes a very important problem in human food products, especially those containing prohibited substances such as ingredients derived from pigs. In the concept of food safety assurance, raw material data used is very important to be included. The world with a Muslim population of about 1.3 billion makes food security globally not only toward healthy and safe food products but also must show the halal aspect of the food product [1].

Authentication is the most important criteria and is always an issue in the field of food. The one that encourages the importance of authentication is the raw material, because it determines whether the food is deserved to be consumed. In Islam, "Halal" means that it is allowed or permitted and haram refers to anything that is not allowed and includes sin when violated. If used in relation with food and drink, halal means allowed to be eaten or drunk. It is mandatory for Muslims to eat halal food and consume halal products. For a Muslim, food should not only be healthy and of good quality, but more importantly, the food must be halal. Nowadays, many kinds of animal products have been commonly consumed and spread across countries without limitations. The relatively new products such as nuggets, sausage, burger, hot dog, meat ball and so on are widely accepted by consumers regardless of gender, ethnics and age. Further processing of animal products has given more economic advantage for the producers and convenience for the consumers.

In recent years, there has been an increasing trend in some countries for mixing prohibited substances, especially pigs in food products for the purpose of falsification to obtain economic benefits [2]. Identification becomes very important so that the status of halal food becomes clear. Two approaches that can be made to determine whether pig substances are present are analysis of the pig meat elements or pig material in food products.

Various methods of detection of non-halal materials have developed rapidly. The method that has significantly developed is *FTIR spectrophotometry* [3] differential scanning calorimetry [4], liquid chromatography [5], immunoelectrophoresis [6], polymerase chain reaction (PCR)-RFLP DNA based methods [7] and real-time PCR [8].

Some of these methods are too laborious and time-consuming, consequently, an analytical technique offering rapid and reliable methods must be used. One of the promising methods suitable for routine analysis is polymerase chain reaction (PCR) [9]. The DNA-based analysis method has several advantages: DNA can be found in all cell types in an individual with identical genetic information. DNA is a stable molecule in the extraction process of several different types of samples.

Deoxyribonucleic acid (DNA) is a nucleic acid that contains genetic material and serves to regulate the biological development of all cellular life forms [10]. DNA contains genetic information stored in a nucleotide sequence and is used to synthesize all cell protein molecules and organisms. Another function is to provide information that is passed on to the cells or offspring of a child. Both of these functions require DNA molecules that serve as templates or models [11]. DNA has a double structure with components of pentose sugars (deoxyribose), phosphate groups and base pairs. DNA base pairs consist of purine and pyrimidine bases. The purine base consists of adenine (A) and guanine (G). The pyrimidine base is cytosine (C) and thymine (T) [12]. DNA is very important and widely used in analysis. DNA isolation is the key to successful identification of livestock derived from raw materials; therefore, the different methods of DNA isolation in each food stuff must be adapted to the conditions of the food type and food used in the diet.

2. DNA isolation stage

Identification of species based on DNA technology depends on the success of DNA isolation. The success of DNA isolation is the basis for continuing with the next test stage. Therefore, the isolation method should be known by any researcher who wishes to detect the species or foodstuffs. Work experience in the DNA isolation laboratory is a basic skill that needs to be known in order to isolate the various food matrixes. Some of the important factors that influence the success of DNA isolation are:

1. type of food
2. food matrix
3. cooking time
4. cooking temperature

DNA isolation has several stages, that is, (1) isolation of cells; (2) lysis of walls and cell membranes; (3) extraction in solution; (4) purification; and (5) precipitation. There are two principles in performing DNA isolation, that is, centrifugation and precipitation. The main principle of centrifugation is to separate the substance based on molecular weight by providing a centrifugal force so that the heavier substances will be at the bottom, while the lighter substances will be located at the top. The centrifugation technique is carried out in a machine called centrifugation machine with varying speeds, for example, 2500 rpm (rotation per minute) or 3000 rpm [10].

Chemically, the destruction of cells is performed by utilizing chemical compounds such as ethylenediamine tetra acetic (EDTA) and sodium dodecyl sulfate (SDS). EDTA functions to destruct cells by binding magnesium ions (this ion serves to maintain cell integration and activity of nuclease enzyme which damages the nucleic acids). SDS is a kind of detergent that serves to destruct cell membranes. The proteinase K enzyme can be used to destruct proteins. Impurities due to cell lysis are separated by centrifugation. Then, the nucleotide molecules (DNA and RNA), which have been separated, are cleared of the residual proteins using phenol. In this process, a small part of RNA can also be cleaned. The RNAase enzyme is used to clean the residual proteins and polysaccharides from the solution. Purification of DNA can be done by mixing the DNA solution with NaCl, which serves to concentrate, separate DNA from the solution and precipitate DNA when mixed with ethanol [13].

The measurement of DNA concentration was done in two ways, that is, by spectrophotometer and by ethidium bromide fluorescence technique (EtBr). If the sample is pure, without large amounts of contamination such as protein, agarose, phenol or other nuclei, then the use of spectrophotometer that calculates the amount of UV irradiation absorbed by the bases is the proper way. The fluorescence technique of EtBr is used when the sample is contaminated [14].

DNA of ingredients in the form of fresh ingredients can generally be more easily isolated because most of the tissues have not been damaged. DNA of animal-derived food with cell structures easy to lysis is more easily isolated compared to a mixture of plant matrix ingredients which have stronger cell structures. Cooking duration will affect the quality and quantity of DNA that can be isolated, longer cooking duration will degrade the DNA; consequently, a lot of DNA fragments into smaller sizes, and as a result, at the time of amplification, there will occur difficulty with big-size amplicon target compared to small amplicon. The cooking temperature also has effect, the higher cooking temperature so the DNA in the food tends to be fragmented as a result for the size of the DNA with high fragment length target is difficult to amplify at PCR and difficult when identify the origin species of ingredients.

3. DNA isolation of animal tissue genome

The basic principle of total DNA isolation from the tissue is to break down and extract the tissue, so cell extracts composed of tissue cells, DNA and RNA cells are formed. Then, the cell extract was purified to produce a cell pellet containing total DNA [10].

DNA is usually isolated from animal tissue cells using methods that lyse cells but prevent DNA fragmentation. This step usually involves EDTA (ethylenediamine tetra acetic) which in the process will bind magnesium ions (the cofactors required by the DNase enzyme). Furthermore, the cell membrane is preferably solubilized with a detergent. If physical disruption is required, it should be done as minimum as possible. In this disruption process, the nuclease enzyme released from the cellular component can efficiently digest the nucleic acids, so the work of the nuclease enzyme must be inhibited. Cell disruption and most of the following steps should be performed at 4°C, using a glassware and an autoclave solution (the autoclave function is to destruct DNase activity in the apparatus or solution). After removing the nucleic acid from the cell, RNA can be removed by the addition of heat-treated RNase to inactivate the DNase of the contaminant (RNase is relatively stable to heat because of the presence of disulfide bonds which will cause the renaturation process when cooled).

Other major contaminants, that is, proteins, are eliminated with phenol solution or phenol-chloroform mixture (both will denature the proteins but do not denature nucleic acids). After the mixture is made into an emulsion, centrifugation is conducted. After centrifying, an organic part will be formed in the lower part and in the top aqueous layer, which is a layer composed of a denatured protein and inside it is contained DNA. The aqueous fluid is taken and deproteinized several times until no material is visible in the middle layer. Then, the DNA that does not contain protein is mixed with two parts of ethanol. DNA will be a precipitate, separate from the sample solution. After being centrifuged, the DNA pellets are then dissolved again [15].

4. DNA isolation of plant genome

Samples from plants are mostly found in food products both individually and food matrix such as sausage, meatballs and nuggets in which there are materials from animals and plants. The most commonly used method for plant DNA isolation is the *Cetyl Trimethyl Ammonium Bromide* (CTAB) method. CTAB is a common method used in plant DNA extraction because plant DNA contains polysaccharides and polyphenol compounds [16]. There are three major steps in the extraction of plant DNA, that is, the destruction of cell walls (lysis), separation of DNA from solid materials such as cellulose and proteins and DNA purification [17].

CTAB has the same function as SDS, which acts as a lipid solvent of the cell membrane. SDS is a kind of detergent that can emulsify lipids. After overnight incubation with ethanol absolute and ammonium nitrate, the DNA filaments are known and then centrifuged to obtain the DNA pellets. After the supernatant was removed, 70% of ethanol was added for next precipitation. Then, the pellet is dried and stored at −20°C [18].

5. DNA isolation of food products

Food products are a complex food mixture. Food products may contain PCR inhibitors such as polysaccharides, polyphenols and proteins [19]. In addition, food products have undergone several stages of processing, such as mechanical treatment, heating, chemical and enzymatic. As a result, DNA isolation from food products often has difficulties [20].

Some authors have compared methods of DNA extraction from food. Zimmerman et al. [21] compared nine different extraction methods of soybean; food samples showed that five of the methods had the extracted DNA damaged, while in four other methods (CTAB, Wizard, DNeasy and nucleon Phytopure), the amount of DNA produced was relatively low but had good quality. Previously, Olexová et al. [22] have found good DNA extraction results using CTAB on soybean, corn and wheat products (flour, biscuits and instant porridge). Each sample in the food matrix needs the specific DNA isolation methods, as described by researcher [20] and they tried to compare four methods of DNA extraction (NucleoSpin kit, commercial and GeneSpin, CTAB, and Wizard method) applied to soybean processed products; CTAB method extraction showed the best results. Further, Sambrook [23] used CTAB modification by adding salt (NaCl) and twice as much CTAB volume in extracting food product samples for the identification of pork content.

6. DNA isolation in research practice

Today, the DNA isolation commonly uses DNA isolation kit as it is faster and is a simple method than the manual; consequently, the price is more expensive. The commercial kit has been available in varying amounts and benefits. However, the use of kits in DNA isolation is not suitable for student learning in college; therefore, the use of manual methods in student research is highly recommended. The use of kits in DNA isolation is usually simple and follows the procedures of the company, while manual DNA isolation requires the preparation of chemicals independently. The following table illustrates some basic methods and basic principles of DNA isolation (**Table 1**).

The results of DNA isolation research using the modified [14] method for sample of meatball and sausage as illustrated in **Figure 1** is as follows.

Food products are complex food mixtures which may contain inhibitors such as polysaccharides, polyphenols and proteins [19]. In addition, food products have undergone several

Method	The principle of DNA separation
TNES [24]	• Destruction of cell walls using EDTA, SDS and NaCl • Removal of protein with proteinase K • Removal of polysaccharide protein residues by using phenol, chloroform and isoamyl alcohol • Precipitation of DNA with saline solution (NaCl)
[14] modified	• Destruction of cell walls using EDTA and SDS • Removal of protein with proteinase K • Removal of polysaccharide protein residues by using phenol, chloroform and isoamyl alcohol • Precipitation of DNA with saline solution (NaCl)
CTAB	• Cell wall barrier and protein removal by extracting buffer (EDTA, Tris–HCl, NaCl, and K proteinase) • Precipitation of DNA with isoproponal

Table 1. The basic principle of DNA isolation in the DNA extraction method performed.

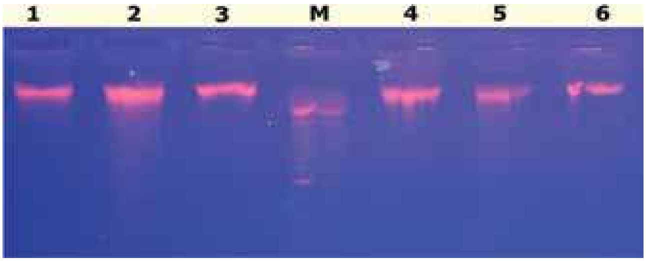


Figure 1. Result of DNA isolation of meatball and sausage of various species on 1% agarose gel: (1) 100% cow meatballs, (2) 100% chicken meatballs, (3) 100% pork meatballs, (M) 100 bp marker, (4) 100% cow sausage, (5) 100% chicken sausage, (6) 100% pork sausage.

stages of processing, such as mechanical treatment, heating, chemical and enzymatic treatments. As a result, DNA isolation from food products often has difficulties [20]. The complicities of sample affected the DNA quality [25] and the presence of additives on foods also affected the results of DNA isolation. Therefore, optimization and modification of the basic method to obtain DNA with good quality and quantity is required.

As this study applied the Sambrook method [14] without overnight incubation, the results showed that the quality of the DNA is not good enough and is marked by the appearance of DNA smear and some samples were not isolated (data not shown); therefore, the modification to prolong the lysis time and cell digestion by adding the incubation time to ±15 h (overnight) at a temperature of 55°C was made. The modification by adding the incubation time was intended so that the non-DNA molecules can be digested perfectly, thus minimizing the contaminants in the resulting DNA, since the sample was a processed meat consisting of a mixture of complex foods. The sample used in this study was a processed food product consisting of a mixture of meat (animal tissue) and non-meat (flour and spices) products that were considered as an inhibitor in the activity of DNA isolation.

The K proteinase was capable of digesting the cell completely after incubation of at least 10 h [23]. Meanwhile, previous research [26] founded that the duration of cell lysis depends on the shape of the extracted product. The more diverse the mixture and texture of the product, longer the time required. The optimization of ASL (Qiagien)-based extraction and purification of DNA by applying overnight to the cell lysis stage had been successfully performed for the identification of pork in processed meat products in Saudi Arabia [27]. The same method [28] was also used for *overnight* incubation at the lysis stage and digestion for the identification of pig DNA in food products. Chapela et al. [29] also used *overnight* incubation at the cell lysis stage in DNA extraction from canned tuna to identify species of animal origin.

DNA extraction of whole tissue is easier to do than processed meat such as meatballs, sausages and abon, where DNA has been degraded. The DNA obtained for analysis must be pure and intact to obtain good analysis.

The contaminants that are often present in the process of DNA extraction and purification are proteins [30]. DNA measurements were performed by comparing absorbance ratios at wavelengths of 260 and 280 nm (OD $_{260}$: OD $_{280}$) [15].

Repetition	Methods		
	Wasko [24]	Sambrook [14]	CTAB
1	1.64	1.91	1.70
2	1.65	1.91	1.70
3	1.63	1.91	1.70
Mean	1.64 ± 0.01	1.91 ± 0.00	1.70 ± 0.00

Table 2. Purity comparison of DNA isolation result of meatball sample with various methods [59].

The quality and purity of DNA is influenced by various factors including the isolation method and the type of food samples being prepared. The purity of DNA isolates from meatball samples with various methods of DNA isolation is described in **Table 2**.

Good DNA extraction methods not only provided good DNA results but also high DNA purity [30]. In the study, DNA purity measurements obtained from several DNA extraction methods had been performed. From **Table 2**, it was known that the purity of DNA obtained from the modified Sambrook method gave best results on meatball samples, (1.91 ± 0.00; 1.92 ± 0.00; and 1.87 ± 0.00). This showed that the method used could produce DNA with good purity. The value of DNA purity ranges from 1.8 to 2.0. The value of purity would be lower if there was contamination of protein or phenol [15]. Method of [23] obtained DNA purity value of 1.64 ± 0.01 (meatballs), while DNA purity value of the CTAB method was 1.7 ± 0.00 (meatballs). This value may indicate the presence of protein contamination in isolated DNA. Furthermore, in the practice of food, DNA isolation with a varying food matrix requires choice of method, method modification and development of the most suitable method according to food matrix condition whose DNA will be isolated.

7. Confirmation of pig species with conventional PCR

The wide variety of food products available on the market in the world seems favorable but leads to several fears for the Muslim community because the consumption of pork in food products is prohibited; therefore, some analytical methods offering fast and reliable results are continuously developed by some researchers. PCR-based methods have been applied through the PCR-RFLP method for the authentication of animal products especially meatballs in the market [9].

Polymerase chain reaction (PCR) is one of the most widely studied and widely used techniques of nucleic acid amplification in vitro. PCR is used to multiply the number of DNA molecules on a particular target by synthesizing new DNA molecules that complement the target DNA molecules through the aid of enzymes and oligonucleotides as primers in a thermocycle. The target length of DNA ranges from tenths to thousands of nucleotides whose positions are flanked by a pair of primers. The primer that is located before the target is called the forward primer and which is located after the target is called the reverse primer. The enzyme used as a novel printer of DNA molecules is known as the polymerase enzyme [31]. The length of the

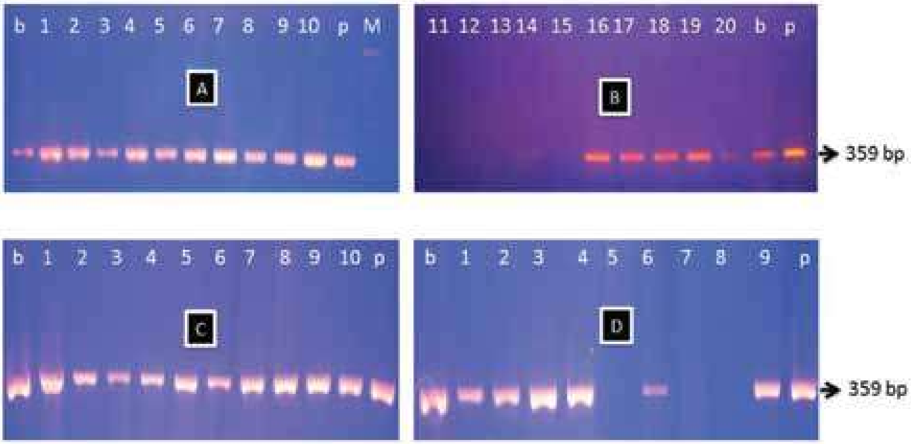


Figure 2. Amplification of mitochondrial cytochrome *b* DNA gene fragments 359 bp long in samples from meatballs products separated by 2% high-resolution agarose gel electrophoresis, M: Marker 100 bp DNA ladder (Invitrogen), A-D: PCR products of cytochrome *b* gene from 20 of commercial meatball samples from various regent in Indonesia b: Amplicon of beef DNA, and p: Amplicon product of raw pork DNA on [34]. Source: Aust-Asian J. Anim. Sci Vol. 27 No. 10.

DNA gene is 1.140 bp and has some stable sequences used for universal primers and several sequences of variables normally used for animal identification by the PCR-RFLP method [32].

A pair of primers was employed in PCR reaction (9), the PCR primers used were cytochrome b gen as follows: (5′-CCA TCC AAC ATC TCA GCA TGA TGA AA-3′) and CYTb2 (5′-GCC CCT CAG AAT GAT ATT TGT CCT CA-3′), as reported by Kocher et al. [33]. The PCR-RFLP was applied using mitochondrial cyt b gene in a final volume of 25 μl containing 250 ng of extracted DNA in order for the porcine identification with mega mix royal PCR buffer. Amplification was performed in PCR system 2400 (Perkin Elmer), programmed to perform the pre-denaturation step of 94°C for 2 min, followed by 35 cycles which is carried out with following steps: denaturation at 95°C for 36 s, annealing at 51°C for 73 s, and extension at 72°C for 84 s. Final extension at 72°C was conducted for 3 min for complete synthesis of elongated DNA molecules. The DNA amplicon then cleaved using BseDI restriction enzyme and results could able to differentiate porcine among other meat [9].

PCR has the potential sensitivity and specificity required to achieve detection of a target sequence from template DNA. Gene of cytochrome *b* was used for the amplification of PCR and resulted in DNA fragmentation of approximately 360 bp for bovine, chicken and porcine. These small amplicons are ideal for use with processed foods where DNA commonly was degraded. This result indicated that pig, beef and chicken DNA in processed meat was successfully amplified in PCR reaction (**Figure 2**). The result of PCR amplification was similar to that [35] which reported the presence of the 360 bp fragment. This result also indicated that the frying of meat products at 100°C for 15 min did not result in the degradation of DNA into small size (<200 bp). They studied the feasibility of using mitochondrial cyt b gene to detect porcine material in commercial food products from various markets and evaluated them for the presence of porcine DNA. Visualization through electrophoresis of PCR products clearly

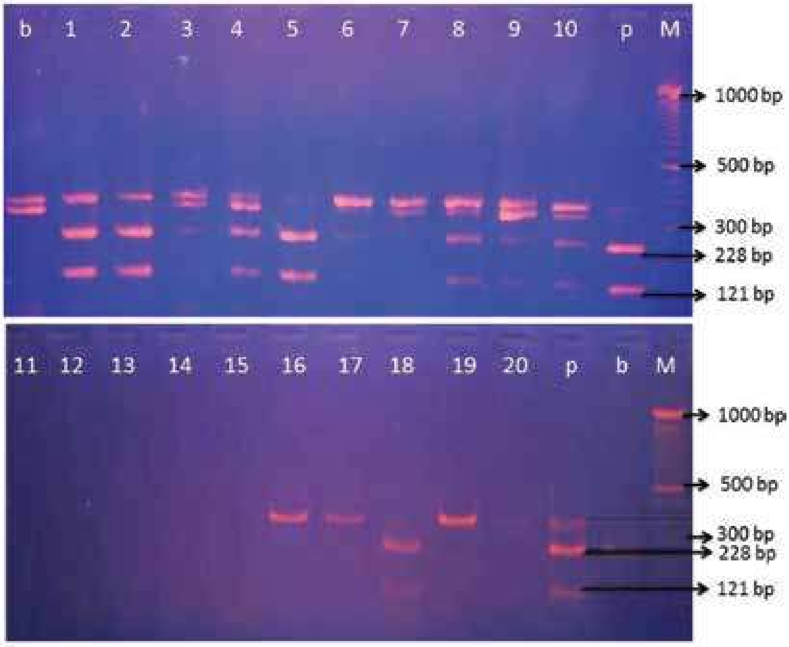


Figure 3. Restriction fragment length polymorphism using *BseD*I restriction enzymes. M: Marker 100 bp DNA ladder, lane 1–20: DNA fragment of different meatball samples from 20 commercial meatballs b: DNA fragment raw beef cytochrome *b* gene cleaved into different pattern, and p: DNA fragment of PCR product of raw pork cytochrome *b* gene cleaved into 228 and 121 bp [34]. Source: Aust-Asian J. Anim. Sci Vol. 27 no. 10.

detected porcine DNA, while no amplification occurred in others meat sources such as cattle, chicken, sheep and horse [36].

The result of the amplicon cleavage (**Figure 3**) showed that the enzyme of BseDI could differentiate pork meat among chicken, bovine or goat.

PCR product was digested using the *BseDI* restriction enzyme at 2 U/μl for 3-h incubation time. The result of analysis from various levels of pork in beef sausage and chicken nugget indicated that mixture of pork could be digested by the *BseDI* restriction enzyme and was detected until level 1%. *BseD*I cleavage bands visualized in the gel were enough and suitable for the discrimination of pork in processed food. *BseD*I endonuclease cleaved the cytochrome b gene products of pig species into two DNA fragments of 228 and 131 bp and did not cleave cytochrome b gene of beef and chicken [33].

Recent research [37] founded that pork DNA fragment of mitochondrial cytochrome b gen could be amplified well after cooking for 0–120 min (**Figure 4**), while cooking more than 120 min caused DNA degradation and failed to amplify mitochondrial DNA fragment, even for the PCR using more than 500 bp primer lengths. In addition, the amplification of 300 bp DNA fragment using specific primer showed DNA amplicon fragment in slight appearance. The band density decreased to less than 50% in sausage after cooking for more than 120 min. The 200 bp primer fragment could amplify all DNA samples from cooking for 120 min. Longer cooking time showed slighter appearance in gel electrophoresis and lower band intensity. Sausages were cooked for less than 120 min showed almost the same intensity, then decreased until about 60% band intensity. It

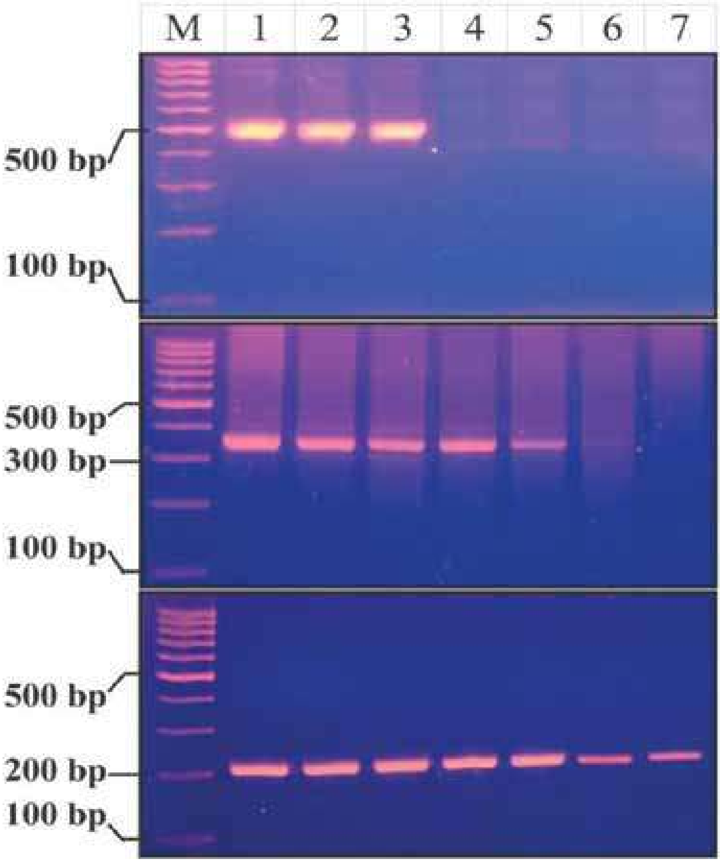


Figure 4. Amplification results of cooking time treatment. M: Marker, 1: no cooking (control), 2: 15 min, 3: 30 min, 4: 60 min, 5: 120 min, 6: 240 min, 7: 480 min. Above: 500 bp primer, middle: 300 bp primer, bottom: 200 bp primer.

indicated that the DNA from pork sausages after cooking for 120 min did not cause denaturation process, and the molecular weight did not decrease. The charged molecules can be separated in agarose gel electrophoresis according to smaller molecules migrating faster than larger molecules.

While cooking for more than 120 min, damage was caused on DNA fragment and could be amplified using species specific primer only above 200-bp length fragment (**Figure 4**). The DNA fragmentation occurred when meat was processed and cooked. The other researches [38] stated that the damaging of the DNA, as well as other meat structural changes (shrinkage of meat fibers, the aggregation and gel formation of myofibril and sarcoplasmic proteins, shrinkage and solubilization of the connective tissue, etc.), is due to the denaturation during heating.

8. Confirmation of pig species using real-time PCR

Compared to conventional end-point PCR, the main advantage of a real-time method is the possibility of performing quantitative measurements. The meat species identification using conventional PCR assays require the visualization using gel electrophoresis. The mistake analysis could increase with the chance of the cross-contamination or sample shifting; in addition, the process

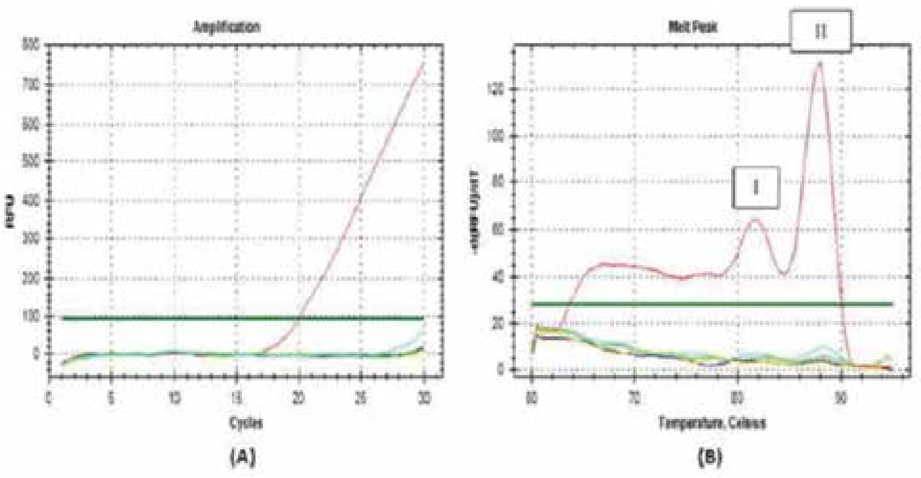


Figure 5. Amplification curves A: and melting peak; B: specificity of pork mitochondrial D-Loop686 primer on DNA from various raw meat. Red line: pork; yellow line: beef; dark blue line: chicken; green lines: goat; blue line: horse [42].

could not be automated. This condition may be a consequence because the PCR tubes must be opened after amplification. An alternative, shorter time-consuming analysis to differentiate the PCR products of various meat species would improve the assay. Melting curve analysis with an intercalating dye using a real-time PCR machine could remove the gel electrophoresis step, offered that the PCR products of the different species melt at different temperatures [39]. Basic principle of real-time PCR is the development of conventional PCR methods, in which amplification products can be monitored directly during each amplification cycle and can be measured. Real-time PCR testing allows for the identification of the early stages of the PCR process, which is more accurate than the endpoint analysis associated with agarose gel electrophoresis or polyacrylamide. The collecting data of real-time PCR was obtained using fluorescence molecules that showed a correlation between fluorescence intensity with PCR product abundance [40]. Real-time PCR is widely accepted as a powerful assay for the species identification and quantification of nucleic acid molecules. This procedure has higher sensitivity and specificity, larger dynamic range of detection, and less carry-over contamination risk. In the quantitative real-time PCR (qPCR) technique, amplification of the target gene is monitored by an increased fluorescence signal which enables direct assessment of the results after the PCR application without additional detection steps [41].

The amplification process of the PCR product can be observed from the beginning of the reaction to completion as shown in **Figure 5**. The number of PCR cycles is seen on the X-axis and the fluorescence of the amplification reaction on the Y axis. The amplification plot shows two phases, that is, the exponential phase followed by the non-exponential plateau phase. During the exponential phase, the amount of PCR products is approximately twice that of each cycle. In the plateau phase the reaction slows down (28–40 cycles) [43].

Threshold (initial limit) real-time PCR is the signal level which reflects statistically significant increases in baseline signal. The threshold is adjusted to differentiate the relevant amplification signal among the background. Typically, the real-time PCR instrument software automatically adjusts 10 times the standard deviation of the baseline fluorescence value [44]. Threshold cycle or commonly abbreviated as Ct is the number of cycles during the fluorescence signal, that is, the reaction cuts the threshold. The Ct value is used to calculate the number of initial DNA copies because the value of Ct is the inverse of the initial number of targets. If the number of templates decreases, then the number of cycles required for amplification will increase [45].

In the initial reaction, the fluorescence on the base level and the increase in fluorescence is not detected at 1–18 cycles, although the product accumulates exponentially (**Figure 5**) until finally the amplification products sufficient to emit fluorescence signal that can be detected. The number of cycles when amplification occurs is called early cycle (threshold cycle or Ct). If the Ct value is measured in exponential phase, as long as the reagent is not limited to, *real-time* PCR can accurately and reliably calculate the amount of DNA present in the reaction. There are several types of fluorescence based on chemicals used for the detection of *real-time* PCR, which can be classified into four types: probe hydrolysis as TaqMan®, probe hairpin as molecule Beacon molecule, probe hybridization labeled with fluorescence (FRET) and the DNA intercalation dye as SYBR® Green and EvaGreen® [46, 47].

Fluorescent compounds *SYBR Green* are asymmetric cyanine compounds that have two aromatic rings connected by metena bridges, and each ring contains a nitrogen atom, and one of which is positively charged [48]. When SYBR Green is in the solution containing double-stranded DNA (ds-DNA), then the SYBR Green can be bound to ds-DNA and will emit a fluorescent signal at a wavelength of 254 (excitation) and 497 (emission) nm [49]. Characterization of fluorescence level that is obtained by SYBR Green can be seen through the Ct value, that is, at the DNA fragment amplification cycle, which was first detected. SYBR Green is widely used for the purpose of amplification optimization using qPCR, because it is easier and cheaper than probe [50].

The previous research [42] performed the real-time polymerase chain reaction using mitochondrial displacement loop686 and cytochrome b (cytb) gene primers for the identification specific pork DNA among other four types of DNA species: beef, chicken, goat and horse. The annealing temperature was at 62°C; however, the mitochondrial D-Loop686 showed high specificity to porcine DNA detection, as indicated by no amplification results that appeared on other species (beef, chicken, horse and goat meat) (**Figure 5**). It was proven that this additional peak does not result from primer dimer, since no amplification product on sample without DNA template is present.

The implementation of d-Loop 686 DNA and cyt b gene primers was carried out in order to identify pork DNA on laboratory prepared "dendeng" (Indonesian traditional beef Jerky). The results indicated that mitochondrial d-Loop686 primer had high specificity on pork DNA target. Results of experiment indicated that the primer has ability to detect pork DNA until the concentration of pork was 0.5% in the beef dendeng product with the consistent curve profile. The cytb gene primers also had a similar ability with d-Loop686 primer to amplify pork DNA until the pork concentration was of 0.5% in beef "dendeng". The results showed that these two primers, d-Loop region and the cyt b gene, had the capability to identify pork adulteration on the mixed pork-beef "dendeng" product at all variations of tested concentrations [42].

Soares et al. [8] applied a real-time PCR approach based on SYBR Green dye for the quantitative detection of pork meat in processed meat products. For the development of the method

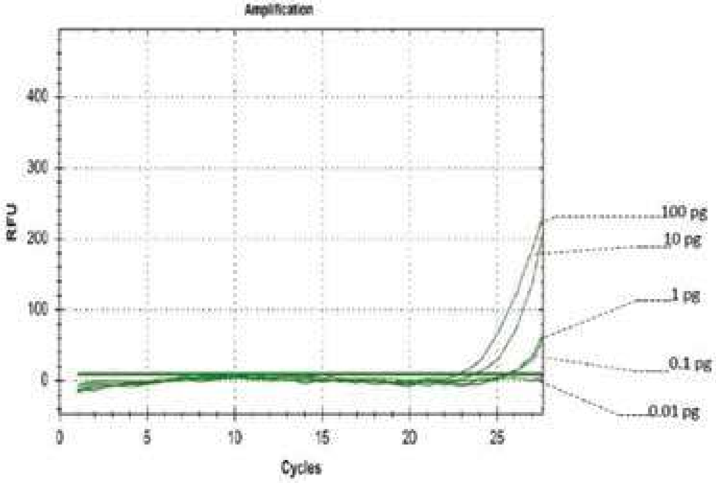


Figure 6. Amplification curves of binary reference mixtures (0.01–100 pg. of pork DNA) by real-time PCR with SYBR green I dye for pork detection and the detection limit was 0.1 pg.

analysis, binary meat mixtures containing known amounts of pork meat in poultry meat were used to obtain a normalized calibration model from 0.1 to 25% with high linear correlation and PCR efficiency. The method revealed high specificity by melting curve analysis and was successfully validated through its application to blind meat mixtures, which confirmed its adequacy for pork meat determination. The full applicability of the method was further demonstrated in commercial meat products, allowing verification of labeling compliance and identification of meat species in processed foods. **Figure 6** showed limit of detection of the real-time PCR which was able to detect the porcine DNA until 0.1 pg. contamination.

DNA degradation, inhibition or differences in the amount and quality of the DNA obtained from the food matrix sample was found. It meant that the processed meat products generally had several ingredients, including those from vegetable sources, and different processing treatments that might affect the target-gene amplification, the use of an endogenous control, which enables these variations to be controlled. Therefore, the structure of calibration curve based on real-time PCR normalization is essential for reliable quantitative analysis because this process controls variations in extraction yields and efficiency of amplification. The endogenous control allows verifying if amplification variations found with the species-specific primers were due to differences in target species content or to other factors such as (8).

Kesmen et al. [41] conducted the real-time PCR investigation and they concluded that TaqMan probe real-time PCR-based assay can be recommended for the detection of animal tissue by food control agencies or laboratories, and it might be a reliable and practical method for the determination of technically predictable contamination and/or intentional admixtures in complex processed meat products. **Table 3** described the application of the PCR-based technology on porcine determination of the commercial sample and **Table 4** showed the type of gene, fragment length and sequence of the primer already applied by some researchers.

Based on **Table 4**, the primer design for the application of the specific primer length target depends on the matrix sample. The sample with long period preparation and processing commonly

Species analysis	Issues	References
Animal Species	PCR assay for the identification of animal species in feedstuffs	[51]
Analysis of rat's meat DNA	Authentication of beef meatball from rat's meat	[52]
Analysis of pork DNA in dendeng	Authentication of "dendeng" from pork	[42]
Analysis of porcine gelatin	Analysis of porcine gelatin in capsule shell	[53]
Analysis of pork in meatball	Authentication of beef meatball from pork	[34]
Analysis of pork DNA	Identification of pig species in food product	[1]
Quantification of bovine, porcine, chicken and turkey species	Contamination and adulteration prohibited component in food and feed	[54]
Porcine Gelatin	Using gelatin in pharmacy and confectionery	[55]

Table 3. Analysis of porcine species-based PCR technology on commercial samples.

Genealogy	Specific primer sequence (5′–3′)	Products size (bp)	References
Mitochondrial Cyt b	5′ GCCTAAATCTCCCCTCAATGGTA 3′ 5′ ATGAAAGAGGCAAATAGATTTTCG 3′	212	[55, 56]
12S rRNA-tRNA	5′ CTA AAT ATC AAG CAC CAT CAC A 3′ 5′ ACA TTG TGG GAT CTA CTT GG 3′	290	[51]
cytochrome b	5′ CGC CTT ACG TTC TAA TGA CAT 3′ 5′ ATC CTA CTC T AG CC C ACC CA 3′	500	[9]
ND5 gene	5′ CCATCCCAATTATAATATCCAACTC-3′ 5′ TGATTATTTCTTGGCCTGTGT GT 3′	141	[57]
12S rRNA	5′ TGCAGTCTGTCTCCTCCAAA 3′ 5′ CGATAATTGGATCACATTTCTG 3′	152	[58]
Mitochondrial Dloop 686	5′ GTTACGGGACATAACGTGCG –3′ 5′ GGCAAGGCGTTATAGGGTGT –3′	114	[8]

Table 4. Various length of specific target DNA primers for porcine detection.

the DNA degraded into 100–200 bp, consequently if the PCR method applied using more than 300 bp DNA target. Some food matrix need long time preparation, conse.

9. Conclusion

It is common that food should consider an individual with health, culture, life style, religious faith, budget and choice. Food products without pork samples on the ingredients list, or even labeled 'pork-free', obtained from the commercial market and retail stores that could be confirmed by PCR-based methods. Many experiments showed the detection limit of less than 0.1% porcine material level. This data showed that some samples could be considered contaminants of porcine materials. Reviews of these findings showed the value of such an analytical approach to guarantee the commercial products in world free trade. PCR-based method is proposed to be a reliable and sensitive protocol for routine analysis.

Author details

Yuny Erwanto[1,2]*

*Address all correspondence to: yunyer@ugm.ac.id

1 Department of Animal Products Technology, Faculty of Animal Science, Gadjah Mada University, Yogyakarta, Indonesia

2 Halal Research Center, Gadjah Mada University, Yogyakarta, Indonesia

References

[1] Erwanto Y, Abidin MZ, Sismindari X, Rohman A. Pig species identification in meatballs using polymerase chain reaction-restriction fragment length polymorphism for halal authentication. International Food Research Journal. 2012;**19**:901-906

[2] Aida AA, Che Man YB, Raha AR, Son R. Detection of pig derivatives in food products for halal authentication by polymerase chain reaction-restriction fragment length polymorphism. Journal of the Science of Food and Agriculture. 2007;**87**:569-572. DOI: 10.1002/jsfa.2699

[3] Rohman A, Sismindari X, Erwanto Y, Che Man YB. Analysis of pork adulteration in beef meatball using Fourier Transform Infrared (FTIR) spectroscopy. Journal of Meat Science. 2011;**88**:91-95. DOI: 10.1016/j.meatsci.2010.12.007

[4] Marikkar JMN, Lai OM, Ghozali HM, Che Man YB. Detection of lard and randomized lard as adulterants in refined-bleached-deodorized palm oil by differential scanning calorimetry. Journal of the American Oil Chemists' Society. 2001;**78**:1113-1119. DOI: 10.1007/s11746-001-0398-5

[5] Marikkar JMN, Ghazali HM, Che Man YB, Peiris TSG, Lai OM. Distinguishing lard from other animal fats in vegetable oils admixtures of some liquid chromatographic using the data coupled with multivariate analysis of data. Journal of Food Chemistry. 2005;**91**:5-14. DOI: 10.1016/jfoodchem.2004.01.080

[6] Necidová L, Renčová E, Svoboda I. Counter immunoelectrophoresis: a simple method for the detection of species-specific muscle proteins in heat-processed products. Veterinární Medicína – Czech. 2002;**47**(5):143-147. DOI: 10.17221/5818-VETMED

[7] Erwanto Y, Abidin MZ, Rohman A, Sismindari X. Using PCR-RFLP. BseDI enzyme for authentication in pork sausage and nugget products. Media Peternakan. 2011;**34**(1): 14-18. Bogor: EISSN

[8] Soares S, Amaral JS, MBPP O, Mafra I. A SYBR green real-time PCR assay to detect and quantify the pork meat in processed poultry meat products. Meat Science. 2013;**94**:115-120. DOI: 10.1016/j.meatsci.2012.12.012

[9] Erwanto Y, Yuliatmo Y, Sugiyono, Rohman A, Sismindari X. Species specific polymerase chain reaction (PCR) assay for identification of pig (*Sus domesticus*) skin in "Rambak" crackers. In: Proceedings of The 1st International Conference on Tropical Animal Science and Production, July 26-29, 2016; Bangkok

[10] Faatih M. Isolation and digestion chromosomal DNA isolation and digestion of chromosomal DNA. Research Journal of Science & Technology. 2009;**10**:61-67

[11] Martin DW. Biokimia (Harper's Review of Biochemistry). 20th ed. Jakarta: EGC; 1987

[12] Yuwono T. Biologi Molekular. Jakarta: Penerbit Erlangga; 2010

[13] Sulandari S, Zein MSA. Free Practice Laboratory DNA. Zoology Field. Bogor: Biology Research Center, Indonesian Institute of Sciences; 2003

[14] Sambrook J, Fritch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 1989. pp. 16-17

[15] Sambrook J, Russel DW. Molecular Cloning: A Laboratory Manual. 3rd ed. Vol. 1. New York: Cold Spring Harbor Laboratory Press; 2001

[16] Jose J, Usha R. Geminiviral. Extraction of DNA from a highly mucilaginous plant (*Abelmoschus esculentus*). Plant Molecular Biology Reporter. 2000;**18**:349-355

[17] Surzycki S. Basic Techniques in Molecular Biology. Berlin/Heidelberg/New York: Springer; 2000

[18] Tan SC, Yiap BC. DNA, RNA, and protein extraction: The past and the present. Journal of Biomedicine and Biotechnology. 2009;**2009**:10. DOI: 10.1155/2009/574398. Article ID 574398

[19] Terry P, Jain M, Miller AB, Howe GR, Rohan TE. Dietary intake of folic acid and colorectal cancer risk in a cohort of women. International Journal of Cancer. 2002;**97**:864-867

[20] Mafra I, Silva SA, EJMO M, CSF S, Beatriz M, Olivera PP. Comparative study of DNA extraction methods for soybean derived food products. Food Control. 2008;**19**:1183-1190. DOI: 10.1016/j.foodcont.2008.01.004

[21] Zimmerman A, Lüthy J, Pauli U. Quantitative and qualitative evaluation of nine in different methods for nucleic acids extraction on soya bean food samples. Zeitschrift fuer Forschung und Lebensmittel Untersuchyng A. 1998;**207**:81-90. DOI: 10.1007/s00217005 0299

[22] Olexová L, Dovičovičová Ľ, Kuchta T. Comparison of three types of methods for the isolation of DNA from fl ours, biscuits and instant paps. European Food Research and Technology. 2004;**218**:390-393. DOI: 10.1007/s00217040872

[23] Sambrook J. Mini-preparation (isolation and purification of plasmid). 2009. Available from: http://www.all-about-molecullar-wordpress.com. [Accessed: January 25, 2012]

[24] Wasko AP, Martins C, Oliveira C, Foresti F. Nondestructive genetic sampling in fish. An improved method for DNA extraction from fish fins and scales. Hereditas. 2003;**138**:161-165. DOI: 10.1034/j.1601-5223.2003.01503.x

[25] Nuraini H. Development of repetitive sequence element 1 (PRE-1) as a molecular marker for detecting material pig on processed meat products [dissertation]. Bogor: Graduate Program in Bogor Agricultural University; 2004

[26] Primasari A. Sensitivitas gen sitokrom B (cyt B) sebagai marka spesifik pada genus rattus dan mus untuk menjamin keamanan pangan produk asal daging [thesis]. Bogor: Institut Pertanian Bogor; 2011

[27] Alaraidh IA. Improved DNA extraction method for porcine contaminants, detection in imported meat to the Saudi market. Saudi Journal of Biological Sciences. 2008;**15**:225-229

[28] Che Man YB, Aida AA, Raha AR, Son R. Identification of pork derivatives in food products by species-specific polymerase chain reaction (PCR) for halal verification. Food Control. 2007;**18**:885-889. DOI: 10.1016/j.foodcont.2006.05.004

[29] Chapela MJ, Sotelo CG, Pérez-Martín RI, Pardo MA, Pérez-Villareal B, Gilardi P, Riese J. Comparison of DNA extraction methods from muscle of canned tuna for species identification. Food Control. 2007;**18**:1211-1215. DOI: 10.1016/j.foodcont.2006.07.016

[30] Tun-Nguyen CT, Son R, Raha AR, Lai OM, Clemente Micheal WVL. Comparison of DNA extraction efficiences using various methods for the detection of genetically modified organisms (GMOs). International Research Journal. 2009;**16**:21-30

[31] Muladno. Teknologi Rekayasa Genetika, Edisi Kedua. Bogor: IPB Press; 2010

[32] Minarovič T, Trakovická A, Rafayová A, Lieskovská Z. Animal species identification by PCR-RFLP of cytochrome b. Scientific Paper: Journal of Animal Science and Biotechnology. 2010;**43**:296-299

[33] Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Wilson AC. Dynamic of mitochondrial DNA evolution in animal: Amplifications and sequencing with conversed primers. Proceedings of the National Academy of science of the USA. 1989;**86**:6169-6200

[34] Erwanto Y, Abidin MZ, Muslim EYP, Sugiyono S, Rohman A. Identification of pork contamination in meatballs of indonesia local market using polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) analysis. Asian Australasian Journal of Animal Science. 2014;**27**:1487-1492

[35] Aida AA, Che Man YB, Wong CMVL, Raha AR, Son R. Analysis of raw meats and fats of pigs using polymerase chain reaction for halal authentication. Meat Science. 2005;**69**:47-52. DOI: 10.1016/j.meatsci.2004.06.020

[36] Tanabe S, Miyauchi E, Muneshie A, Mio K, Sato C, Sato M. PCR method of detecting the pork in foods for verifying allergen labeling and for identifying hidden pork ingredients in processed foods. Bioscience, Biotechnology, and Biochemistry. 2007;**71**:1663-1667. DOI: 10.1271/bbb.70075

[37] Erwanto Y, Yuliatmo R, Fitriyanto NA, Abidin MZ, Sugiyono S, Rohman A. DNA isolation and pig species detection on sausage with various cooking temperatures and times. In: Proceedings of the 2nd International Conference on Tropical Agriculture; October 26-27, 2017; Yogyakarta

[38] RWD R. Electron microscopy of bovine muscle: II—The effects of heat denaturation on post rigor sarcolemma and endomysium. Meat Science. Published by Faculty of Animal Science, Gadjah Mada University, Yogyakarta; 1989;**26**:281-294. DOI: 10.1016/0309-1740(89)90013-2

[39] Thanakiatkrai P, Kitpipit T. Meat species identification by two direct-triplex real time PCR assays using low resolution melting. Food Chemistry. 2017;**233**:144-150. DOI: 10.1016/j.foodchem.2017.04.090

[40] López-Andreo M, Lugo L, Garrido-Pertierra A, Prieto MI, Puyet A. Identification and quantitation of species in complex DNA mixtures by real-time polymerase chain reaction. Analytical Biochemstry. 2005;**339**:73-82. DOI: 10.1016/j.ab.2004.11.045

[41] Kesmen Z, Sahin F, Yetim H. PCR assay for the identification of animal species in cooked sausages. Meat Science. 2007;**77**:649-653. DOI: 10.1016/j.meatsci.2007.05.018

[42] Maryam S, Sismindari Raharjo TJ, Sudjadi RA. Determination of porcine laboratory contamination in jerky prepared using mitochondrial D-Loop686 and CYT b gene primers by real time polymerase chain reaction. International Journal of Food Properties. 2016;**19**:187-195. DOI: 10.1080/10942912.2015.1020434

[43] Real-Time PCR Applications Guide. Bio-Rad laboratories, Inc., USA [Internet]. 2006. Available from: http://www.gene-quantification.com/real-time-pcr-guide-bio-rad.pdf

[44] Anonymous. Basic of Real-Time PCR, Life Technologies Corporation Inc. C232085 0812. 2012

[45] Jiang J, Chan TC, Temenak JJ, Dasch GA, Ching WM, Richards AL. Development of a quantitative real-time polymerase chain reaction assay specific for Orientia tsutsugamushi. The American Journal of Tropical Medicine and Hygene. 2004;**70**:351-356

[46] Laube I, Zagon J, Broll H. Quantitative determination of commercially relevant species in foods by real-time PCR. International Journal of Food Science and Technology. 2007;**42**:336-341. DOI: 10.1111/j.1365-2621.2006.01249.x

[47] Jonker KM, Tilburg JJ, Hagele GH, de Boer E. Species identification in meat products using real-time PCR. Food Additives Contaminants Part A Chemical Analysis Control Exposure and Risk Assessment. 2008;**25**:527-533. DOI: 10.1080/02652030701584041.

[48] Nygren J, Svanvik N, Kubista M. The interactions between the fluorescentdyethiazole orange and DNA. Biopolymers. 1998;**46**:39-51. DOI: 10.1002/(SICI)1097-0282(199807) 46:1<39:AID-BIP4>3.0.CO;2-Z

[49] Kumar A, Kumar R, Sharma B, Gokulakrishnan P, Mendiratta S, Sharma D. Identification of species origin of meat and meat products on the DNA basis: A review. Critical Review in Food Science and Nutrtion. 2013;**5**:1340-1351. DOI: 10.1080/10408398.2012.693978

[50] Fraga D, Meulia T, Fenster S. Real-Time PCR, Dalam. In: Current Protocols Essential Laboratory Techniques. Hoboken: John Wiley & Sons; 2008. p. 75

[51] Dalmasso A, Fontanella E, Piatti P, Civera T, Rosati S, Bottero MT. A multiplex PCR assay for the identification of animal species in feedstuffs. Molecular and Cellular Probes. 2004;**18**:81-87. DOI: 10.1016/j.mcp.2003.09.006

[52] Yanita IW, Sudjadi Rohman A. Detection of rat meat adulteration in meat ball formulations employing real time PCR. Asian Journal of Animal Sciences. 2015;**9**:460-465. DOI: 10.3923/ajas.2015.460.465

[53] Sudjadi Wardani HS, Sepminarti T, Rohman A. Analysis of porcine gelatin DNA in commercial capsule shell using real-time polymerase chain reaction for halal authentication. International Journal of Food Properties. 2016;**19**:2127-2134. DOI: 10.1080/ 10942912.2015.1110164

[54] Shehata HR, Li J, Chen S, Redda H, Cheng S, Tabujara N, Li H, Warriner K, Hanner R. Droplet digital polymerase chain reaction (ddPCR) assays integrated with an internal control for quantification of bovine, porcine, chicken and turkey species in food and feed. PLoS One. 2017;**12**:e0182872. DOI: 10.1371/journal.pone.0182872

[55] Shabani H, Mehdizadeh M, Mousavi SM, Dezfouli EA, Solgi T, Khodaverdi M, Rabiei M, Rastegar H, Alebouyeh M. Halal authenticity of gelatin using species-specific PCR. Food Chemistry. 2015;**184**:203-206. DOI: 10.1016/j.foodchem.2015.02.140

[56] Lahiff S, Glennon M, O'Brien L, Lyng J, Smith T, Maher M, Shilton N. Species specific PCR for the identification of ovine, porcine and chicken species in meat and bone meal (MBM). Molecular and Cellular Probes. 2001;**15**:27-35

[57] Ali ME, Razzak MA, Hamid SB, Rahman MM, Amin MA, Rashid NR, Asing. Multiplex PCR assay for the detection of five meat species forbidden in Islamic foods. Food Chemistry. 2015;**177**:214-224. DOI: 10.1016/j.foodchem.2014.12.098

[58] Farouk AE, Batcha MF, Greiner R, Salleh HM, Salleh MR, Sirajudin AR. The use of a molecular technique for the detection of porcine ingredients in the Malaysian food market. Saudi Medical Journal. 2016;**27**:1397-1400

[59] Fitrianingsih. Optimization of DNA isolation method from processed meat as a basic step for pork contamination detection. [Thesis]. Faculty of Animal Science, Gadjah Mada University; 2013

Chapter 6

Genotyping Approaches for Identification and Characterization of *Staphylococcus aureus*

Mazen M. Jamil Al-Obaidi, Zarizal Suhaili and
Mohd Nasir Mohd Desa

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75969

Abstract

Genotyping methods are vital epidemiological tools for discriminating different bacterial isolates within same species, which in turn provide useful data in tracing source of infection and disease management. There have been a revolutionary efforts in ways to distinguish between bacterial types and subtypes at molecular level utilizing DNA in the genomes. Notably, the growth of various DNA typing methods has provided innovative apparatuses for improved surveillance and outbreak investigation. Thus, early identification and genotyping are indispensable as resources for managing therapeutic treatment and the control of rapid expansion of clinically important bacteria. Methicillin-resistant *Staphylococcus aureus* (MRSA) has been in a great attention due to its contagious nature and subjected to various typing analyses. Thus, in this chapter, we aimed to review the contribution of various genotyping methods of commonly used as well as those unique to *staphylococci* in understanding its epidemiology, infection and dissemination pattern, and to provide an impression of specific advantages and disadvantages of each tool.

Keywords: genotyping, MLST, RFLP, *Staphylococcus aureus*, WGS

1. Introduction

Typing is a process to characterize the species and properties of organisms, in particular the discrimination at the strain level both phenotypically and genetically. Conventional typing such as serotypes, biotype, and phage type has been in practice for many years. Nevertheless, typing at molecular level is nowadays very essential due to its specificity, which is often used to support the associated phenotypic characteristics. For example, one species may comprise many subtypes or a subpopulation which one might be more pathogenic than the others. Thus, genotyping plays an important role to identify potential differences at genetic level as well as for epidemiological traceability of all the presented isolates [1].

A good typing method must have the discriminatory power to differentiate all unrelated isolates epidemiologically to facilitate any outbreak investigation. This will allow investigation to demonstrate person-to-person strain transmission, subsequently, allowing preventive measures to be designed to inhibit further dispersion of the pathogens. Additionally, genotyping method must be inexpensive, rapid, easy to interpret and highly reproducible [1]. For a continuous surveillance, genotyping methods must produce results with a sufficient stability over time. Also, it should produce portable data and can be easily accessed through open source web-based database or a client-server database connected via the internet, facilitating global comparison of the isolates.

Genotyping methods are basically based on phenotyping, PCR/sequence typing and genome typing approaches. Remarkably, a great effort has been put up in epidemiological investigations of *Staphylococcus aureus* due to its role as a leading nosocomial, community and livestock-acquired bacterial pathogen. Globally, dynamic spread of methicillin-resistant *Staphylococcus aureus* (MRSA) strains stimulates the increasing rates of these strains in several regions rapidly. Additionally, the global emergence of MRSA had influenced significantly the health care systems all around the worlds over the past 50 years. The epidemiological changes of *S. aureus* infection in human beings and animals are being focused for two main reasons: (i) to understand the evolution and dissemination pattern of the species and (ii) to find a proper antimicrobial treatment strategy and effective infection management. Therefore, epidemiological studies utilizing various typing techniques are continuously on the go in various parts of the world especially those with increasing rates of MRSA infections. In this chapter, we aim to review the contribution of various genotyping methods commonly in use as well as those unique to *Staphylococcal* particularly MRSA, that might assist to detect the outbreak infections, conduct epidemiological surveillance by means of rapid typing and to provide an impression of specific advantages and disadvantages of each typing tool.

2. Phenotypic detection and identification of MRSA

Numerous conventional or molecular methods can be applied for detection and identification of *S. aureus* including colony morphology, production of coagulase activity and by various enzyme activity. Also, commercial latex agglutination tests and the API Staph system (bioMerieux) are examples of assays available for identification of *S. aureus,* which remains the methods of choice due to their feasibility and low cost. Additionally, there are other phenotypic methods such as biotyping and immunoblotting, serotyping, phage-typing and multilocus enzyme electrophoresis (MLEE) that have been frequently applied previously.

Upon identification of *S. aureus*, antimicrobial susceptibility profile is always performed so that the choice of antimicrobial treatment can be formulated. The antimicrobial testing procedures to a broad range of antimicrobial agents have been standardized and improved for the accuracy of reporting, following guidelines either by the Clinical & Laboratory Standards Institute (CLSI), The European Committee on Antimicrobial Susceptibility Testing (EUCAST) and British Society for Antimicrobial Chemotherapy (BSAC) guidelines. The methods are usually carried out to determine which antibiotic shall be the most effective in treating bacterial infection *in vivo*. This simple and useful method offer relatively inexpensive and is usually applied for clinical investigation of pathogen isolated from various types of infection. Examples of the antimicrobial susceptibility tests are disc diffusion, agar/tube agar dilution, CHROMagar oxacillin resistance screening as well as agar-based and Epsilometer test (E-Test).

Currently, these tests are considered as the most popular methods of choice that can support genotyping data [2]. Nevertheless, previous reports suggested that phenotypic methods of identification have drawbacks due to the variability in expression of phenotypic characterization by isolates belonging to the same species and their reliance on subjective interpretation of test results [2]. As a result, phenotypic detection and identification was heavily burdened with several issues; including low reproducibility, reliability, sensitivity and specificity as well as lack of resolution in epidemiology investigation. Therefore, several reports have suggested that genotypic identification and detection methods offered a higher discriminatory power, reliability, reproducibility and typeablity [3]. Genotypic identification can be done with the phenotypic approach together for a better comparative analysis. Furthermore, the introduction of molecular screening for MRSA detection as well as identification directly from clinical specimens has been developed to enhance and identify common *Staphylococcal* spp., as well as to speed up the detection methods especially in clinical research [4].

3. Genotyping of MRSA

The ultimate goals for bacterial typing are to further clarify the population dynamicity and also to track the spread of the microorganisms. As mentioned earlier, traditional bacterial typing of phenotypic-based alone, does not provide the prudent resolution for identifying and tracking an infection-causing pathogen, and also does not clearly describe the transmission pattern of an outbreak. However, molecular typing has been an invaluable tool for molecular epidemiologist as well as clinical researchers for tracing the spread of particular strains, discovering the route of dissemination and the potential reservoirs. Usually, the outcomes of epidemiological investigations are often used to guide and assist the clinical treatment of the patients by selecting the appropriate antimicrobial agents. Furthermore, molecular typing contributes to the comprehensive understanding of the epidemiology of infection and facilitates infection control measures as well as management [5].

It is well known that *S. aureus* is frequently associated with clonal spread as reported by many studies utilizing various typing methods on huge numbers of *S. aureus* strains. For example, molecular strain typing of MRSA is implemented in order to elucidate genetic variation to guide in outbreak investigation as well as to characterize genetic macroevolution for spatial-temporal and evolutionary studies [6]. In those studies, PCR-based methods are commonly used for typing as they are easy and fast technique. Other methods such as pulsed-field gel electrophoresis (PFGE), coagulase gene PCR-restriction fragment length polymorphism

(RFLP) and *Staphylococcal* cassette chromosome mec typing (SCC*mec*) also play similar important roles in molecular typing of both MSSA and MRSA [7]. Additionally, sequence-based techniques also play an vital role in genotyping, including *spa* gene-typing and multilocus sequence typing (MLST), that have been considered as a very useful tool for epidemiological studies, particularly MRSA [8]. Several reports suggested two methods known as PFGE and MLST that are considered as 'gold standard' in typing of both MSSA and MRSA, although these typing methods are often time-consuming, costly and laborious [8]. In the subsequent sections, various genotyping methods are presented to elaborate each extent to establish molecular epidemiology studies.

3.1. Polymerase chain reaction (PCR)-based identification

Polymerase chain reaction (PCR) is known as enzymatic method used to exponentially amplify a specific preselected fragment of DNA. It is well known that PCR uses a thermostable DNA polymerase *in-vitro* to multiply copies of a specific nucleic acid region exponentially. The procedures require DNA template from the organisms being typed, thermostable DNA polymerase, two synthetic oligonucleotide primers and four standard deoxyribonuclease triphosphate that are incorporated into newly synthesized DNA. There are numerous PCR-based amplification methods that have been applied widely in the subtyping of various microorganisms, including *S. aureus* especially MRSA as stated in this chapter. Various phenotypic tests have been used for identification of MRSA from other *Staphylococal* spp., such as screening for production of protein A, cell-bound clumping factor, extracellular coagulase and heat-stable nuclease [9]. However, a good package of rapid molecular detection is also required for screening and identification of certain antibiotic resistance determinants as well as virulence factors of *S. aureus* [3, 10]. To date, most of the molecular approaches for the identification of MRSA have been a PCR-based method with a range of primers designed to amplify specific targeted markers encompassing species-specific, antibiotics resistance as well as virulence determinant [11]. Other PCR sequencing-based methods have been developed for the identification of *S. aureus* from other coagulase-negative *staphylococci* targeting 16S rRNA, RNA polymerase B (*rpoB*), *femA*, *tuf* and *gap* genes. However, these approaches have their own limitations as it is not sufficiently discriminatory to differentiate closely related *staphylococcal* spp., where database of these genes only include a limited number of *Staphylococcal* spp. [12].

3.1.1. Amplified fragment length polymorphism (AFLP)

AFLP is a PCR-based method applied in DNA fingerprinting and genetic research. In this method, restriction enzymes (e.g. endonuclease) are used to cut the genomic DNA of the typed species, subsequently, double-stranded oligonucleotide adaptors which are comprised of a core sequence and an enzyme-specific sequence, are bound to one of the sticky ends of the restricted fragments. After that, a PCR thermocycler is used to amplify those restricted fragments ending with the adapter selectively, using primers complementary to the adapter sequence. Then, the restriction site sequence and a number of additional nucleotides from the end of the unknown DNA are designed. Usually, the restriction fragments (50–100) are amplified using florescent dye-labeled PCR primers, to detect those separated fragments by size using automated DNA sequencer. Likewise, gel electrophoresis can also be used to visualize and analyze the amplified fragments of typed DNA. Upon analysis, a high-resolution banding is generated via computer reflecting the genetic relatedness among bacterial isolates [13].

This kind of technique has a higher discriminatory power in comparison to PFGE, where it was shown in a study that AFLP analysis provided greater genetic resolution and was less sensitive to DNA quality during genetic typing of bacterial pathogens *E.coli* O157:H7 in epidemiological investigation [14]. Additionally, like other DNA banding pattern-based method, AFLP can be automated and has portable results, as well as reproducible approach to facilitate the analysis [14]. Previous study has been conducted by Fossum and Bukholm [15] reported MRSA population was revolutionized from hospital-acquired MRSA (HA-MRSA) to community-acquired MRSA CA-MRSA in the south-eastern part of Norway through increase in MRSA clones harboring SCC*mec* IV as shown by AFLP, MLST and spa typing methods. AFLP analysis grouped the MRSA isolates into clusters according to the clonal complexes (CCs), but did not discriminate among the different sequence types (STs) or spa types inside each CC [15]. In the United Kingdom, there are 16 phage types of epidemic MRSA (eMRSA) strains that have been identified, of which eMRSA-3, -15 and -16 now predominate [16]. Through this approach specifically fluorescent AFLP (FAFLP), it was able to classify eMRSA phage type from 1 to 16 by identifying eMRSA phage type of *S. aureus* (eMRSA-15) from UK [17] and into 9 clone clusters in European isolates [18]. Thus, AFLP is considered as a tool with highly discriminatory power against these strains of MRSA. As a result, this technique is considered suitable for MRSA epidemics surveillance at national and international levels as well as reproducible approach. Additionally, it is found that AFLP approach is more reproducible than PFGE and MLST, and it is more suitable for inter-laboratory data exchange using sequence-based data [19]. The main drawbacks of this method are labor-intensive and expensive.

3.1.2. 16s ribosomal RNA (16s rRNA)

16S rRNA comprises ~1500 pair nucleotide sequence coding for catalytic RNA that is part of the 30S ribosomal subunit. 16S rRNA gene is comprised of nine variable regions (V1–V9/30–100 base pairs long), that show sequence diversity among different bacterial species, subsequently enable for identification purposes. V1–2-3 regions are located at the 50 end of the 16s rRNA gene which is shown to be appropriate and more sensitive than other regions for identification of different types of bacteria [20]. This gene is constant in function, promising a valid molecular chronometer, where it exists in all prokaryotic cells. Therefore, it is used to elucidate both close and distant phylogenetic relationships at the genus and at the species level [21] based on the differences in the nucleotide sequence of 16s RNA gene. Additionally, dedicated 16S databases [22] that include near full-length sequences for a large number of strains and their taxonomic placements are available. The sequence from an unknown strain can be compared against these available sequences in the database. This approach is considered as a common substitute for traditional methods using (rRNA) gene sequencing [23]. It is less time-consuming and labor intensive, where DNA sequencing can offer more absolute taxonomic classification than culture-based approaches for numerous organisms [23]. However, there are also limitations with this approach associated with the short read lengths, variances ascending from the diverse regions selected, sequencing errors and difficulties in evaluating operational taxonomic units (OTUs) [23, 24]. Additionally, single marker (16S rRNA) usage is considered challenging to assess the bacterial diversity, subsequently difficult to identify bacterial species [25], as well as the resolution of 16S rRNA that is very limited among closely related species. A previous study has shown that 16S rRNA combined with *mec*A and *nuc* using multiplex PCR, is considered as useful tool for rapid characterization of MRSA [26]. Thus, this multi-gene technique is considered a better discrimination tool among unrelated isolates, particularly in *S. aureus* [27].

3.1.3. Staphylococcal cassette chromosome mec typing (SCCmec)

SCC*mec* complex is a mobile genetic element that confers the methicillin resistance profile in *S. aureus*. MSSA may emerge to become MRSA upon acquiring this genetic complex. SCC*mec* contains essential elements which can be detected by regular PCR; (i) *ccr* genes which are constituted by ccrA and ccrB and (ii) mec gene complex which is composed of *mec*A gene and its regulatory genes, mecI and mecRI. Currently, 11 major types of SCC*mec* elements (I–XI) have been identified based on the organization of the *mec* gene complex, *ccr* gene complex and integrated plasmids (http://www.SCC*mec*.org/). To date, there are four allotypes (types 1, 2, 3 and 5) of ccr complex and three classes (A, B and C) of *mec* complex [28]. Different combinations of these complex classes and allotypes generate various SCC*mec* types. SCC*mec* elements are currently classified into types I–V based on the nature of the *mec* gene complex and *ccr* allotypes [26]. At present time, multiplex PCR is used for the characterization of SCC*mec* types. For example, Okuma et al. [29] developed primers that were specific for SCC*mec* IVa and SCC*mec* type IVb, meanwhile, Hisata et al. [30] developed multiplex PCR for the specific identifications of SCC*mec* type IIa, IIb, IVc and IVd.

Also, two different multiplex PCR methods were developed by Zhang et al. [28] and Milheirico et al. [31] for specific characterization of SCC*mec* type I to SCC*mec* type V. Likewise, 9 pairs of primers were used by Zhang et al. [28] for identification of SCC*mec* type I, II, III, IV (a, b, d) and V, whereas 10 pairs of primers were used as described by Milheirico et al. [31]. Interestingly, Boye et al. [32] developed an easy screening of MRSA SCC*mec* typing only by using multiplex PCR with a combination of four pairs of primers, where clear and easily discriminated band pattern was obtained for all major five types of SCC*mec*. These characterization methods could be used to distinguish HA-MRSA and CA-MRSA typing, where SCC*mec* types I, II, III and VIII are usually acquired by HA-MRSA, while SCC*mec* types IV, V, VI and VII are acquired in CA-MRSA [33]. Thus, it is very useful and important molecular tool in understanding the potential epidemiological background of the strains.

3.1.4. Multiple-locus variable-number tandem repeat assay (MLVA)

Multiple-locus variable-number tandem repeat analysis (MLVA), was previously known as a variable-number tandem repeat (VNTR) [34] by making use the VNTR polymorphism. In 2008, after the introduction of spa typing as a standard molecular typing method in the Germany MRSA surveillance, MLVA was added as a supportive typing technique. This method involves PCR amplification of five specific loci (*sdr, clfA, clfB, ssp* and *spa*) of *S. aureus* which is composed of seven individual genes (*sdr*CDE, *clf*A, *clf*B, *ssp*A, *spa*, *mec*A and *fnb*P) [35] using multiplex PCR mixture followed by separation of the amplified bands on agarose gel and comparison of the band patterns between strains to identify genetic clusters or clones [36]. This genotyping method showed a successful typing of MRSA isolates in many studies [37], in term of determining the genetic diversity and evolutionary lineage with discriminatory power.

It was shown that this approach has a reproducibility as good as PFGE technique [34]. The main drawback of this approach is that in highly conserved genomes, there may not be sufficient DNA polymorphisms in these limited sequence targets to exhibit alleles. Another limitation was, small deletions and insertion in the regions flanking the repeat units may lead to misinterpretations, making the MLVA results slightly more ambiguous than sequenced-based methods. However, to overcome this limitation, the DNA sequence of each new allele is determined to confirm the deduced number of repeats [38]. However, the level of discrimination can be increased by adding more loci and repeating the assay with different restriction enzymes [39].

3.1.5. Repetitive element polymerase chain reaction (rep-PCR)

rep-PCR is a DNA-based technique that discriminates microbes at subspecies or strain level by observing genomic DNA fingerprint patterns [40]. In this approach, the hybridization of primers to noncoding intergenic repetitive sequences takes place across the genome. The amplicons are produced during DNA amplification of the repetitive elements. Depending on the distribution of the repeat elements across the genome, the genetic relatedness between the bacterial isolates can be inferred by comparing the banding pattern of the amplicons. Enterobacterial repetitive intergenic consensus' (ERIC 124–127 bp), 'the repetitive extragenic palindromic' (REP 35–40 bp), and the 'BOX 154 bp' sequences are examples of conserved repeat sequences that have been used successfully in rep-PCR typing [40]. This kind of approach is considered as highly discriminatory tool for different bacterial organisms such as *S. aureus* and *Campylobacter jejuni* [41, 42]. However, there was one drawback for this method which was low rate of reproducibility, due to the uses of traditional agarose gels for electrophoresis, which might result in a discrepancy in relation to the use of different reagents and gel electrophoresis systems.

Alternatively, rep-PCR approach is developed and used by a semi-automated method using DiversiLab system (bioMérieux, Marcy l'Etoile, France), where clinically important organisms can be detected by commercial PCR kits [43]. Then, high-resolution chip-based microfluidic capillary electrophoresis is used to separate amplified genomic DNA within repetitive elements, where chip-based microfluidic capillary electrophoresis can increase the determination and reproducibility of the rep-PCR method compared to traditional gel. In the next step, DiversiLab software is used to normalize and analyze the data automatically. Several reports have evaluated the usefulness of this approach (DiversiLab) in outbreak-related and epidemiology unrelated bacterial isolates [44]. It was shown that this approach is rapid, reproducible and easy for typing microorganisms. Hospital outbreaks of MRSA have been identified using this useful DiversiLab tool by Fluit et al., [45]. In contrast, other study found that this tool is not highly discriminative tool for MRSA typing particularly in outbreak setting [46]. The main limitation of this approach is the DiversiLab databases are stored only on manufacturer server, resulting in some users not allowed to use this typing system due to security purposes.

3.1.6. Restriction fragment length polymorphism (RFLP)

In PCR-RFLP approach, restricted enzymes are used to detect the variations in homologous DNA fragments. Then, the DNA fragments are amplified using regular PCR, subsequently these fragments are separated by gel electrophoresis based on length of the fragments. For example, coagulase gene (coa) and *Staphylococcal* protein A (spa) gene RFLP amplified fragment of DNA could be identified through this technique. Previous studies have shown PCR-RFLP typing of coa gene as useful tool to discriminate *S. aureus* strains on the basis of sequence variation within the 3′ end coding region of the gene [47]. The amplification discriminatory power of coa gene depends on the heterogeneity of the region containing 81 bp tandem repeats at the 3′ coding region of the coa, where this region is different in the number of tandem repeats and the location of AluI and HaeIII restriction sites among different isolates [48]. AluI is better than HaeIII in *S. aureus* typing, but both can be used to be more reliable and sufficient power in discrimination issues. It is found that coa-RFLP typing has discriminatory power for *S. aureus* strains particularly in MRSA strains [49]. On the other hand, the repeated part of spa is located at 3′end and identified as X region; the repetitive part of region X comprises of up to 12 elements each with a length of 24 nucleotides. High polymorphic is defined by this 24-nucleotide region with respect to the number and sequence of repeats. Variety of X region causes protein A variation [50]. Thus, the potential dissemination of MRSA can be detected by the number of repeats in the region X of spa [51]. As a result, the PCR-RFLP assays (coa and spa RFLPs) are useful molecular markers for a rapid, and initial study of MRSA outbreaks [51]. Wichelhaus et al. [52] reported that this method is proven as to have a good discriminatory power, typeability and reproducibility in MRSA typing. Moreover, this technique can be used in routine infection control program in health care systems as well as epidemiological investigations [48].

3.2. Sequence typing method for bacterial identification

3.2.1. *Staphylococcal protein A (spa) typing*

Staphylococcal protein A (spa) typing is a sequence-based method that targets VNTR of the spa gene region encoding protein A [53]. The spa gene region is polymorphic as a result of spontaneous mutations and loss or gain of repeat. Besides, spa gene is reported to be a highly effective tool in subtyping both MSSA and MRSA [54]. As mentioned above, the region X of *spa* gene consists of 24 bp repeats sequences, and the diversity of the strains is recognized by duplications and deletions of the sequence in this region of the gene. The variation in the sequences is used to assign repeats numbers [55]. Sequenced data can be analyzed using free accessible offline bioinformatics tool Ridom Bioinformatics (Ridom, GmBH, Germany) (http://spaserver.ridom.de) [56]. This spa server database also provides global frequencies information related to the mapping of the spa with the MLST *S. aureus* database. To date, 748 diverse repeats with more than 17,416 spa types have been described from 131 countries with total strains 384,806 (http://spaserver.ridom.de). Sequences of perfect quality are synchronized with spa server (Ridom server) [57] specifically for spa typing, providing a typical worldwide nomenclature together with integral quality control.

Subsequently, Based Upon Repeat Pattern (BURP) algorithm is used to analyze the diverse spa types associated to each other. The analysis shows a good consistency with MLST-CCs, where ST that shares at least five of seven identical alleles are grouped into a single CC [58]. The advantages of this typing method are the results generated are easy to interpret, less time-consuming, highly reproducible, less laborious and highly comparable between laboratories via ridom.spa.server compared with PFGE. Besides, spa typing is impressive for its ease of interpretation and suitability for international comparison. It is also able to detect both slowly and rapidly accumulated molecular variations as well as to investigate outbreaks in epidemiological studies and molecular evolutions of population structure [59]. However, non-typeable (NT) isolates are increasingly found in the Dutch MRSA surveillance as well as globally. Thus, to overcome the issue of NT strains, other typing method should be concurrently used to be a supportive method [60]. Malachowa et al. [61] found that spa typing was more approximate to MLST approach upon comparing four genotyping methods (PFGE, MLST, MLVA and spa typing) in 59 *S. aureus* strains. Additionally, HA-MRSA, CA-MRSA and livestock MRSA (LA-MRSA) dissemination can be monitored by a combination of these analyses together with spa typing in epidemiological studies at a global level [62].

3.2.2. Multilocus sequences typing (MLST)

MLST has been invented to overcome the poor or insufficient portability of traditional and older molecular typing application. The main idea of this tool is based on MLEE [63] which depends on the differences in electrophoretic mobility of various enzymes exist in a bacterial species. *Neisseria meningitidis* was the first species subjected to MLST analysis in 1998 [64]. After that, this tool was developed to detect other type of bacterial species, where it became a widely accepted tool for molecular epidemiological studies as well as evolutionary studies of pathogen at the molecular level [65]. This molecular subtyping method was developed for bacterial characterization to facilitate rapid and global comparisons among species [66]. In term of MRSA, seven housekeeping genes are amplified and sequenced for internal sequences [67].

In the subsequent analysis, MRSA isolates are grouped within a single CC when five out of seven housekeeping genes (400–500 bases) in that particular MRSA isolates having identical sequences and isolates with the seven same allelic profiles may be descended from a common ancestor [66, 67]. If there are various alleles at each of the seven loci, the isolates are unlikely to have the identical allelic profiles by chance, while isolates that have similar allelic profile can be considered members of the same clone [66]. The variations found among these genes are mostly synonymous and neutral. Since these genes accumulate variations in a slow manner, they are considered to be reliable indicator of evolutionary history [68]. The main advantage of this tool is that whole produced data are obvious due to standardized nomenclature internationally and reproducible. Additionally, ST profile as well as alleles sequences are available in huge central databases (http://pubmlst.organd www.mlst.net) [69] that are freely accessible online. Moreover, the genetic relatedness between bacterial strains within a species can also be identified via the databases.

Thus, it is a useful tool to compare the data with other laboratories via web-based electronic data. Furthermore, it allows the exchange of data collected over internet through the MLST database. BURST software package can analyze the evolutionary events within *S. aureus* population [67]. For instance, MRSA-ST239 was found to disperse in different countries although carrying a similar ST [70]. The drawback of the technique is the high cost, time-consuming, labor-intensive, and also has no discrimination power for cases related to short-term outbreak. For the later, this technique may not discriminate well the epidemic spread of bacterial strains within a limited time frame [19]. Nevertheless, MLST is still considered as the rapid method for subtyping for MRSA in clinical research, and has been shown to be useful in global epidemiological studies of *S. aureus* [67].

3.3. Genomics-based typing tool

3.3.1. Pulsed-field gel electrophoresis (PFGE)

PFGE is an approach used to detect the dispersion of large segments of DNA using gel with high electrical fields that facilitate changing in DNA direction periodically [71]. In brief, molecular sieve of gel is used to transform DNA from cathode to anode using common electrophoresis method. Two electric fields are used in PFGE technique, where it allows to change the directions of the DNA as mentioned above. Subsequently, ethidium bromide dye is used to differentiate the DNA band spectrum as a typing result. Clinically, various types of bacteria can be genotyped by PFGE which is considered as the "gold standard" genotyping method. It is assumed as an epidemiological tool for most bacterial species since 1990s [71]. Currently, PFGE is used worldwide to identify and characterize isolates of bacteria in outbreak investigations [19, 71, 72]. It is also considered as prototype tool to analyze center to center transmission events [73].

In term of *S. aureus*, isolation of intact bacterial chromosomes are required prior to PFGE procedures, where these isolated chromosomes subsequently is broken down into large DNA fragments using cutting restriction endonuclease such as SmaI. Subsequently, the restriction fragments can be separated via agarose gel "pulse-field" electrophoresis, where those separated DNA fragment could be monitored as a banding pattern in the gel. For easier analysis, large restriction fragments (30 kb–1 Mb) are separated based on their size in a dependent manner, yielding few bands on the gel [74]. It is well known that traditional electrophoresis is able to separate DNA fragments up to 20–50 kb only. Thus, this method has been invented to overcome this weakness through modifying the direction of the electrical field to mobilize DNA fragments of up to −2 Mb [75]. Subsequently, the gels are dyed and captured by an imaging system and analyzed using BioNumerices software programs with the Dice coefficient and un-weight pair group matching analysis (UPGMA) setting according to the criteria as described by Fred et al. [72]. After that, graphical dendrogram may be generated by DNA fingerprinting software.

PFGE has been found to show a higher discriminatory power than PCR-RFLP of coa gene and other PCR-based fingerprinting methods as it enables the entire chromosome to be analyzed, whereas the PCR-based fingerprinting methods explore only selected (random) portions of it [76]. Previous studies stated the reproducibility of PFGE is considered high due to the standardization of protocols [77], allowing national and international surveillance systems [78], and standard interpretation guidelines to investigate the emergence of bacterial species particularly *S. aureus*. Previous study had been done to compare different tools such as MLST, PFGE and AFLP for genetic typing of *S. aureus*. It was found that PFGE is less reproducible, and less useful for long-term epidemiolgical investigations or phylogenetic relationships evolution in *S. aurues* strains [19]. Thus, this method is found extremely helpful in the short-term investigation and identification of MRSA outbreaks in hospital, community and livestock-associated [79]. The solid advantage of this application is the ability to address a large number of an investigated genome (>90%). However, there are certain disadvantages of this application such as time-consuming and labor-intensive, as well as insufficient resolution power to differentiate bands of identical size. It also requires highly skilled operators and there are no standardized reagents with technically laborious and lack of centralized criteria for interpreting the banding patterns [80].

3.3.2. DNA microarray

DNA microarray typing method uses a collection of DNA probes that are attached to a solid surface in ordered manner. Ideally, complementary nucleotide sequences for specific bacterial isolates are detected by DNA probes. This approach is specific tools to identify several genes for specific bacterial strains. It can also be used to identify allelic variations of a gene which exists in all strains for particular species. Usually, target DNA could be labeled by chemical, enzymatic reaction and DNA microarray hybridization. Then, labeled target DNA and an immobilized probe create signal due to successful hybridization, giving measurement automatically using scanner. Currently, this approach is extensively used to analyze genomic mutations such as single-nucleotide polymorphisms (SNPs). It is also found that this approach is an excellent application to identify exceptional antibiotic resistance and virulence genes simultaneously to represent epidemiological markers of certain isolates of interest [81]. Whole genome microarray approach is the alternative tool for whole genome sequencing (WGS) for saving time, expenses and efforts, where it has ability to investigate genetic features of isolates involved in outbreak. For example, 31 chromosomes and 46 plasmids were identified from a various set of *E. coli* isolates, subsequently, the presence or absence of genes were detected in very recently emerged *E. coli* O104:H4 using microarray system [82]. Interestingly, more than 3000 clinical and veterinary isolates of MRSA were characterized epidemiologically through Alere StaphyType DNA microarray system, covering 334 target sequences, including 170 distinct genes and their allelic variants [83], showing a high level of biodiversity among MRSA, especially among strains harboring SCC*mec* IV and V elements. Overall, this technology is highly accurate, but the reproducibility data needs to be established to the broad application to be shared globally. Additionally, this approach is considered not practical if the target of typing is SNPs of highly clonal species. Another disadvantage of this approach that the detection is limited only to sequences that is included in the array.

3.4. Whole genome sequencing (WGS)

To investigate genome variations, cost-effective way has been invented for genetic investigations, which is second generation sequencing (NGS) or high-throughput sequencing. This technique is named second generation to differentiate it from first generation sequencing based on the Sanger method. The main advantage of this approach over several traditional sequencing methods is the ability to create millions of reads (35–700 bp length) in one shot, which also leads to a reduction in cost. The nucleotide sequence of the genome is constructed by gathering numerous short sequences reads from overlapping regions, or comparison with previous reference sequences genomes (re-sequencing).

Currently, WGS is considered as a high attractive tool for epidemiological studies [84], and it is believed that this method has the potential as routine tool for bacterial identification and characterization in the near future. Nevertheless, the main challenge for this approach is the interpretation and computation of the huge set of data. This approach is currently used to determine the genetic relatedness between bacterial isolates based on sequence analysis of the whole genome. Additionally, WGS has the ability to distinguish various genomes within an SNP, which cannot be achieved in conventional molecular typing approaches. Thus, characterization of transmission events and outbreaks will be accurate. However, extensive studies must be conducted to translate this prospective tool into a routine practice. It is well known that the methods based on SNPs permit detailed and targeted analysis of variations among related organisms. Thus, WGS using SNPs analysis can identify the isolates related to an outbreak from non-outbreak isolates. Moreover, various phenotypic characteristics such as virulence and antibiotic resistance of particular pathogen can also be inferred by WGS technique. Finally, this approach enables the search for genetic markers, such as the presence or absence of a gene or an amino acid substitution in a protein, facilitating the linkage with the occurrence, severity and virulence of the disease.

Clinically, SNPs analysis on MRSA isolates recovered from an outbreak in a unit care for neonates using WGS sequencing approach was able to offer relevant data within a time frame that can stimulate patient care [85]. Additionally, through WGS, data can also reliably predict antibiotic susceptibility phenotype of MRSA from an outbreak scene [86], leading to development of hospital infection management and patient outcomes in routine clinical practice. Some previous studies took benefits from WGS by investigating CA-MRSA in USA including USA300-0114 [86], where genetic variation was found. Considering the fact that the isolates were recovered and originated from a confined geographical area, the WGS analysis suggested the continuous evolution of this clone within the limited region. These results offer additional support for the use of WGS as a first-line screening method, which is comparable with those gained by phenotypic methods [87]. Furthermore, additional genes may be added to the panel to increase the coverage and sensitivity, where sequenced isolates can be screened to recognize new resistance genes.

As a result, WGS is considered as a rapid prediction of resistance which contributes to effective clinical management, particularly for *S. aureus*. Subsequently, this approach permits the characterization of transmission routes to improve infection control strategies and manage the outbreak. Once the genetic basis of virulence is understood, WGS could permit determination of emerging infectious strains (and new virulence genes) locally and globally. Furthermore, if genetic diversity is characterized over time, it will provide new knowledge of the *S. aureus* population structure, subsequently leading us to obtain extra information and understand the genetic basis of the disseminating strains. As a whole, it is suggested that WGS is considered as a very useful tool in epidemiological investigations to discriminate MRSA, and it may assist to trace person-to-person transmission in health care systems [88].

4. Conclusions

Higher rates of morbidity and excessive healthcare costs are the two main reasons that can be caused by the growing number of HA, CA and LA-MRSA infections. The management of these infections must be conducted through the screening of individuals as well as infection control program. Currently, MRSA can be reliably detected within hours using rapid screening methods. However, the continuous evolution of SCC*mec* MRSA strains requires frequent monitoring of the strains. Therefore, genotyping techniques must be sufficient with internal and external quality control and standardized internationally for MRSA diagnostics. For that reason, to reduce the severe clinical and economical effects of MRSA, rapid and accurate typing is required especially for epidemiological investigations. Currently, conventional and molecular methods are used in combination for MRSA typing. Nevertheless, controversy is still on-going to choose which molecular typing methods will suit every requirement to ascertain molecular epidemiology studies.

For instance, AFLP has a higher discriminatory power in comparison to PFGE, where it provided greater genetic resolution and is less sensitive to DNA quality during genetic typing of bacterial pathogens in epidemiological investigation. Additionally, AFLP can be automated and has portable results, as well as reproducible approach to facilitate the analysis. Moreover, AFLP it is more suitable for inter-laboratory data exchange using sequence-based data. 16s rRNA analysis is considered a good discrimination approach among unrelated isolates, particularly in *S. aureus* only if it is combined with other gene identification such as *nuc* and *mec*A. SCC*mec* is also very useful and important molecular tool in understanding the epidemiology of methicillin resistance as well as supporting the clonal strain relatedness. DiversiLab rep-PCR tool is very useful to identify MRSA in the hospital outbreaks. In contrast, it is reported that rep-PCR is not highly discriminative tool for MRSA typing particularly in outbreak setting. RFLP can be used in routine infection control program in health care systems as well as epidemiological investigation. It has a good discriminatory power, typeability and reproducibility in MRSA typing. Spa typing is a based sequence typing method, where its results are easy to interpret, less time-consuming, highly reproducible, less laborious and highly comparable between laboratories via ridom.spa.server. MLST is still considered as the rapid method for subtyping for MRSA in clinical research, and has been shown to be useful in global epidemiological studies of *S. aureus,* and the results are comparable between laboratories using MLST server for interpretation. PFGE is found extremely helpful in the short-term investigation and identification of MRSA outbreaks in hospital, community and livestock-associated; however, this method has insufficient resolution power to differentiate bands in identical size as the main drawback of this method. DNA microarray technology is highly accurate, but the reproducibility data needs to be established to the broad application of this technology, and it is not practical if the target of typing is SNPs of highly clonal species. Also, this approach is difficult to identify sequences not included in the array.

Finally, WGS is considered as a very useful tool in epidemiological investigations to discriminate MRSA, and it may assist to trace person-to-person transmission in health care systems. Additionally, this technique is considered suitable for MRSA epidemics surveillance at national and international levels as well as reproducible approach, which is essential as baseline resources for managing therapeutic treatment and the control of rapid expansion of these strains.

Author details

Mazen M. Jamil Al-Obaidi[1], Zarizal Suhaili[2] and Mohd Nasir Mohd Desa[1]*

*Address all correspondence to: mnasir@upm.edu.my

1 Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia

2 Faculty of Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

References

[1] van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clinical Microbiology and Infection. 2007;**13**:1-46

[2] Corrente M, Normanno G, Martella V, Bellacicco AL, Quaglia NC, Dambrosio A, Buonavoglia D, D'Abramo M, Buonavoglia C. Comparison of methods for the detection of methicillin resistance in *Staphylococcus aureus* isolates from food products. Letters in Applied Microbiology. 2007;**45**:535-539

[3] Suhaili Z, Johari SA, Mohtar M, Abdullah ART, et al. Detection of Malaysian methicillin-resistant *Staphylococcus aureus* (MRSA) clinical isolates using simplex and duplex real-time PCR. World Journal of Microbiology and Biotechnology. 2009;**25**:253-258

[4] Perez LRR, Antunes ALS, Bonfanti JW, Pinto JB, et al. Detection of methicillin-resistant *Staphylococcus aureus* in clinical specimens from cystic fibrosis patients by use of chromogenic selective agar. Journal of Clinical Microbiology. 2012;**50**:2506-2508

[5] Van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clinical Microbiology Reviews. 2001;**14**(3):547-560

[6] Trindade PA, McCulloch JA, Oliveira GA, Mamizuka EM. Molecular techniques for MRSA typing: Current issues and perspectives. The Brazilian Journal of Infectious Diseases. 2003;**7**:32-43

[7] Yamamoto T, Nishiyama A, Takano T, Yabe S, et al. Community-acquired methicillin-resistant *Staphylococcus aureus*: Community transmission, pathogenesis, and drug resistance. Journal of Infection and Chemotherapy. 2010;**16**:225-254

[8] Li V, Chui L, Louie L, Simor A, et al. Cost-effectiveness and efficacy of *spa*, SCC*mec*, and PVL genotyping of methicillin-resistant *Staphylococcus aureus* as compared to pulsed-field gel electrophoresis. PLoS One. 2013;**8**:e79149

[9] Saiful AJ, Mastura M, Zarizal S, Mazurah MI, et al. Detection of methicillin-resistant *Staphylococcus aureus* using *mec*A/*nuc* genes and antibiotic susceptibility profile of Malaysian clinical isolates. World Journal of Microbiology and Biotechnology. 2006; **22**:1289-1294

[10] Thong KL, Lai MY, Teh CSJ, Chua KH. Simultaneous detection of methicillin-resistant *Staphylococcus aureus*, *Acinetobacter baumannii*, *Escherichia coli*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* by multiplex PCR. Tropical Biomedicine. 2011;**28**:21-31

[11] Sabet NS, Subramaniam GM, Navartnam P, Sekaran S. Detection of methicillin-and aminoglycoside-resistant genes and simultaneous identification of *S. aureus* using triplex real-time PCR TaqMan assay. Journal of Microbiological Methods. 2007;**68**:157-162

[12] Bergeron M, Dauwalder O, Gouy M, Freydiere AM, Bes M, Meugnier H, Benito Y, Etienne J, Lina G, Vandenesch F, Boisset S. Species identification of staphylococci by amplification and sequencing of the tuf gene compared to the gap gene and by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. European Journal of Clinical Microbiology & Infectious Diseases. 2011;**30**:343-354

[13] Mortimer P, Arnold C. FAFLP: Last word in microbial genotyping? Journal of Medical Microbiology. 2001;**50**:393-395

[14] Zhao S, Mitchell SE, Meng J, Kresovich S, et al. Genomic typing of Escherichia coli O157:H7 by semi-automated fluorescent AFLP analysis. Microbes and Infection. 2000;**2**:107-113

[15] Fossum AE, Bukholm G. Increased incidence of methicillin-resistant *Staphylococcus aureus* ST80, novel ST125 and SCC*mec*IV in the south-eastern part of Norway during a 12-year period. Clinical Microbiology and Infection. 2006;**12**:627-633

[16] Report P. Revised guidelines for the control of methicillin-resistant *Staphylococcus aureus* infection in hospitals. British Society for Antimicrobial Chemotherapy, Hospital Infection Society and the Infection Control Nurses Association. The Journal of Hospital Infection. 1998;**39**:253-290

[17] Grady R, O'Neill G, Cookson B, Stanley J. Fluorescent amplified-fragment length polymorphism analysis of the MRSA epidemic. FEMS Microbiology Letters. 2000;**187**:27-30

[18] Grady R, Blanc D, Hauser P, Stanley J. Genotyping of European isolates of methicillin-resistant *Staphylococcus aureus* by fluorescent amplified-fragment length polymorphism analysis (FAFLP) and pulsed-field gel electrophoresis (PFGE) typing. Journal of Medical Microbiology. 2001;**50**:588-593

[19] Melles DC, van Leeuwen WB, Snijders SV, Horst-Kreft D, et al. Comparison of multilocus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), and amplified fragment length polymorphism (AFLP) for genetic typing of *Staphylococcus aureus*. Journal of Microbiological Methods. 2007;**69**:371-375

[20] Benga L, Benten WPM, Engelhardt E, Köhrer K, et al. 16S ribosomal DNA sequence-based identification of bacteria in laboratory rodents: A practical approach in laboratory animal bacteriology diagnostics. Laboratory Animals. 2014;**48**:305-312

[21] Conlan S, Kong HH, Segre JA. Species-level analysis of DNA sequence data from the NIH human microbiome project. PLoS One. 2012;**7**(10):e47075

[22] Cole JR, Wang Q, Cardenas E, Fish J, et al. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Research. 2009;**37**

[23] Petti CA, Polage CR, Schreckenberger P. The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. Journal of Clinical Microbiology. 2005;**43**:6123-6125

[24] Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental Microbiology. 2010; **12**:1889-1898

[25] Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B: Biological Sciences. 2006;**361**:1929-1940

[26] Montazeri EA, Khosravi AD, Jolodar A, Ghaderpanah M, Azarpira S. Identification of methicillin-resistant *Staphylococcus aureus* (MRSA) strains isolated from burn patients by multiplex PCR. Burns. 2015;**41**:590-594

[27] Dolzani L, Tonin E, Lagatolla C, Monti-Bragadin C. Typing of *Staphylococcus aureus* by amplification of the 16S-23S rRNA intergenic spacer sequences. FEMS Microbiology Letters. 1994;**119**:167-173

[28] Zhang K, McClure J-A, Elsayed S, Louie T, Conly JM. Novel multiplex PCR assay for characterization and concomitant subtyping of staphylococcal cassette chromosome mec types I to V in methicillin-resistant *Staphylococcus aureus*. Journal of clinical microbiology. 2005;**43**(10):5026-5033

[29] Okuma K, Iwakawa K, Turnidge JD, Grubb WB, et al. Dissemination of new methicillin-resistant *Staphylococcus aureus* clones in the community. Journal of Clinical Microbiology. 2002;**40**:4289-4294

[30] Hisata K, Kuwahara-Arai K, Yamanoto M, Ito T, et al. Dissemination of methicillin-resistant staphylococci among healthy Japanese children. Journal of Clinical Microbiology. 2005;**43**:3364-3372

[31] Milheiriço C, Oliveira DC, De Lencastre H. Update to the multiplex PCR strategy for assignment of mec element types in *Staphylococcus aureus*. Antimicrobial Agents and Chemotherapy. 2007;**51**:3374-3377

[32] Boye K, Bartels MD, Andersen IS, Møller JA, Westh H. A new multiplex PCR for easy screening of methicillin-resistant *Staphylococcus aureus* SCC mec types I-V. Clinical Microbiology and Infection. 2007;**13**:725-727

[33] Asghar AH. Molecular characterization of methicillin-resistant Staphylococcus aureus isolated from tertiary care hospitals. Pakistan journal of medical sciences. 2014;**30**(4):698

[34] Sabat A, Krzyszton-Russjan J, Strzalka W, Filipek R, Kosowska K, Hryniewicz W, Travis J, Potempa J. New method for typing *Staphylococcus aureus* strains: Multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. Journal of Clinical Microbiology. 2003;**41**:1801-1804

[35] Francois P, Huyghe A, Charbonnier Y, Bento M, et al. Use of an automated multiple-locus, variable-number tandem repeat-based method for rapid and high-throughput genotyping of *Staphylococcus aureus* isolates. Journal of Clinical Microbiology. 2005;**43**:3346-3355

[36] Roussel S, Felix B, Vingadassalon N, Grout J, et al. *Staphylococcus aureus* strains associated with food poisoning outbreaks in France: Comparison of different molecular typing methods, including MLVA. Frontiers in Microbiology. 2015;**6**:1-12

[37] Loncaric I, Künzel F, Licka T, Simhofer H, et al. Identification and characterization of methicillin-resistant *Staphylococcus aureus* (MRSA) from Austrian companion animals and horses. Veterinary Microbiology. 2014;**168**:381-387

[38] Bosch T, Schouls LM. Livestock-associated MRSA: Innocent or serious health threat? Future Microbiology. 2015;**10**:445-447

[39] De La Puente Redondo VA, Navas Méndez J, García Del Blanco N, Ladrón Boronat N, et al. Typing of *Haemophilus parasuis* strains by PCR-RFLP analysis of the tbpA gene. Veterinary Microbiology. 2003;**92**:253-262

[40] Versalovic J, Schneider M, De Bruijn F. Genomic fingerprinting of bacteria using repetitive sequence-based polymerase chain reaction. Methods in Molecular and Cellular Biology. 1994:25-40

[41] Sabat A, Malachowa N, Miedzobrodzki J, Hryniewicz W. Comparison of PCR-based methods for typing *Staphylococcus aureus* isolates. Journal of Clinical Microbiology. 2006; **44**:3804-3807

[42] Wilson MK, Lane AB, Law BF, Miller WG, et al. Analysis of the pan genome of *Campylobacter jejuni* isolates recovered from poultry by pulsed-field gel electrophoresis, multilocus sequence typing (MLST), and repetitive sequence polymerase chain reaction (rep-PCR) reveals different discriminatory Capabil. Microbial Ecology. 2009; **58**:843-855

[43] Healy M, Huong J, Bittner T, Lising M, et al. Microbial DNA typing by automated repetitive-sequence-based PCR. Journal of Clinical Microbiology. 2005;**43**:199-207

[44] Deplano A, Denis O, Rodriguez-Villalobos H, De Ryck R, et al. Controlled performance evaluation of the diversilab repetitive-sequence- based genotyping system for typing multidrug-resistant health care-associated bacterial pathogens. Journal of Clinical Microbiology. 2011;**49**:3616-3620

[45] Fluit AC, Terlingen AM, Andriessen L, Ikawaty R, et al. Evaluation of the DiversiLab system for detection of hospital outbreaks of infections by different bacterial species. Journal of Clinical Microbiology. 2010;**48**:3979-3989

[46] Babouee B, Frei R, Schultheiss E, Widmer AF, Goldenberger D. Comparison of the DiversiLab repetitive element PCR system with spa typing and pulsed-field gel electrophoresis for clonal characterization of methicillin-resistant *Staphylococcus aureus*. Journal of Clinical Microbiology. 2011;**49**:1549-1555

[47] Hookey JV, Richardson JF, Cookson BD. Molecular typing of *Staphylococcus aureus* based on PCR restriction fragment length polymorphism and DNA sequence analysis of the coagulase gene molecular typing of *Staphylococcus aureus* based on PCR restriction fragment length polymorphism and DNA Sequenc. Journal of Clinical Microbiology. 1998;**36**:1083

[48] Himabindu M, Muthamilselvan DS, Bishi DK, Verma RS. Molecular analysis of coagulase gene polymorphism in clinical isolates of methicilin resistant *Staphylococcus aureus* by restriction fragment length polymorphism based genotyping. American Journal of Infectious Diseases. 2009;**5**:163-169

[49] Ishino K, Tsuchizaki N, Ishikawa J, Hotta K. Usefulness of PCR-restriction fragment length polymorphism typing of the coagulase gene to discriminate arbekacin-resistant methicillin-resistant *Staphylococcus aureus* strains. Journal of Clinical Microbiology. 2007;**45**:607-609

[50] Shakeri F, Shojai A, Golalipour M, Alang SR, et al. Spa diversity among MRSA and MSSA strains of *Staphylococcus aureus* in north of Iran. International Journal of Microbiology. 2010

[51] Montesinos I, Salido E, Delgado T, Cuervo M, Sierra A. Epidemiologic genotyping of methicillin-resistant *Staphylococcus aureus* by pulsed-field gel electrophoresis at a university hospital and comparison with antibiotyping and protein A and coagulase gene polymorphisms. Journal of Clinical Microbiology. 2002;**40**:2119-2125

[52] Wichelhaus TA, Hunfeld KP, Böddinghaus B, Kraiczy P, et al. Rapid molecular typing of methicillin-resistant *Staphylococcus aureus* by PCR-RFLP. Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America. 2001;22:294-298

[53] Strommenger B, Kettlitz C, Weniger T, Harmsen D, et al. Assignment of Staphylococcus isolates to groups by spa typing, SmaI macrorestriction analysis, and multilocus sequence typing. Journal of Clinical Microbiology. 2006;**44**:2533-2540

[54] Ruppitsch W, Indra A, Sto A, Mayer B, et al. Classifying spa types in complexes improves interpretation of typing results for methicillin-resistant *Staphylococcus aureus*. Journal of Clinical Microbiology. 2006;**44**:2442-2448

[55] Deurenberg RH, Beisser PS, Visschers MJ, Driessen C, Stobberingh EE. Molecular typing of methicillin-susceptible *Staphylococcus aureus* isolates collected in the Yogyakarta area in Indonesia, 2006. Clinical Microbiology and Infection. 2010;**16**:92-94

[56] Deurenberg RH, Vink C, Kalenic S, Friedrich AW, et al. The molecular evolution of methicillin-resistant *Staphylococcus aureus*. Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases. 2007;**13**:222-235

[57] Collery MM, Smyth DS, Twohig JM, Shore AC, et al. Molecular typing of nasal carriage isolates of *Staphylococcus aureus* from an Irish university student population based on toxin gene PCR, agr locus types and multiple locus, variable number tandem repeat analysis. Journal of Medical Microbiology. 2008;**57**:348-358

[58] Mellmann A, Weniger T, Berssenbrügge C, Keckevoet U, et al. Characterization of clonal relatedness among the natural population of *Staphylococcus aureus* strains by using spa sequence typing and the BURP (based upon repeat patterns) algorithm. Journal of Clinical Microbiology. 2008;**46**:2805-2808

[59] Rodriguez M, Hogan PG, Satola SW, Crispell E, et al. Discriminatory indices of typing methods for epidemiologic analysis of contemporary *Staphylococcus aureus* strains. Medicine. 2015;**94**:1-8

[60] Bosch T, Pluister GN, van Luit M, Landman F, et al. Multiple-locus variable number tandem repeat analysis is superior to spa typing and sufficient to characterize MRSA for surveillance purposes. Future Microbiology. 2015;**10**:1155-1162

[61] Malachowa N, Sabat A, Gniadkowski M, Krzyszton-Russjan J, et al. Comparison of multiple-locus variable-number tandem-repeat analysis with pulsed-field gel electrophoresis, spa typing, and multilocus sequence typing for clonal characterization of *Staphylococcus aureus* isolates. Journal of Clinical Microbiology. 2005;**43**:3095-3100

[62] Kck R, Brakensiek L, Mellmann A, Kipp F, et al. Cross-border comparison of the admission prevalence and clonal structure of meticillin-resistant *Staphylococcus aureus*. Journal of Hospital Infection. 2009;**71**:320-326

[63] Selander RK, Caugant DA, Ochman H, Musser JM, et al. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Applied and Environmental Microbiology. 1986;**51**:873-884

[64] Maiden MC, Bygraves JA, Feil E, Morelli G, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America. 1998;**95**:3140-3145

[65] Dingle KE, Colles FM, Wareing DR, Ure R, et al. Multilocus sequence typing system for *Campylobacter jejuni*. Journal of Clinical Microbiology. 2001;**39**:14-23

[66] Enright MC, Day NPJ, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. Journal of Clinical Microbiology. 2000;**38**:1008-1015

[67] Urwin R, Maiden MCJ. Multi-locus sequence typing: A tool for global epidemiology. Trends in Microbiology. 2003;**11**:479-487

[68] Feil EJ, Cooper JE, Grundmann H, Robinson DA, et al. How clonal is *Staphylococcus aureus*? Journal of Bacteriology. 2003;**185**:3307-3316

[69] Database P. No Title. PubMed database. 2015

[70] Neetu TJP, Murugan S. Genotyping of methicillin resistant *Staphylococcus aureus* from tertiary care hospitals in Coimbatore, South India. Journal of Global Infectious Diseases. 2016;**8**:68-74

[71] Herschleb J, Ananiev G, Schwartz DC. Pulsed-field gel electrophoresis. Nature Protocols. 2007;**2**:677-684

[72] Tenover FC, Arbeit R, Archer G, Biddle J, Byrne S, Goering R, Hebert GHGA, Hill B, Hollis R, McDougal LK, Micheal Miller J, Maury Mulligan MAP. Comparison of traditional and molecular methods of typing isolates of *Staphylococcus aureus*. Journal of Clinical Microbiology. 1994;**32**:407-415

[73] McDougal LK, Steward CD, Killgore GE, Chaitram JM, et al. Pulsed-field gel electrophoresis typing of oxacillin-resistant *Staphylococcus aureus* isolates from the United States: Establishing a national database. Journal of Clinical Microbiology. 2003;**41**:5113-5120

[74] Goering RV. Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease. Infection, Genetics and Evolution. 2010;**10**:866-875

[75] Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell. 1984;**37**:67-75

[76] Murchan S, Kaufmann ME, Deplano A, de Ryck R, Struelens M, Zinn CE, et al. Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: A single approach developed by consensus in 10 European laboratories and its application for tracing the spre. Journal of Clinical Microbiology. 2003;**41**:1574-1585

[77] Mulvey MR, Chui L, Ismail J, Louie L, et al. Development of a Canadian standardized protocol for subtyping methicillin-resistant *Staphylococcus aureus* using pulsed-field gel electrophoresis. Journal of Clinical Microbiology. 2001;**39**:3481-3485

[78] Simor AE, Gilbert NL, Gravel D, Mulvey MR, et al. Methicillin-resistant *Staphylococcus aureus* colonization or infection in Canada: National Surveillance and changing epidemiology, 1995-2007. Infection Control and Hospital Epidemiology. 2010;**31**:348-356

[79] Lim KT, Yeo CC, Suhaili Z, Thong KL. Comparison of methicillin-resistant and methicillin-sensitive *Staphylococcus aureus* strains isolated from a tertiary hospital in Terengganu, Malaysia. Japanese Journal of Infectious Diseases. 2012;**65**:502-509

[80] Yu F, Ying Q, Chen C, Li T, et al. Outbreak of pulmonary infection caused by Klebsiella pneumoniae isolates harbouring blaIMP-4 and blaDHA-1 in a neonatal intensive care unit in China. Journal of Medical Microbiology. 2012;**61**:984-989

[81] Miao J, Chen L, Wang J, Wang W, et al. Current methodologies on genotyping for nosocomial pathogen methicillin-resistant *Staphylococcus aureus* (MRSA). Microbial Pathogenesis. 2017;**107**:17-28

[82] Jackson SA, Kotewicz ML, Patel IR, Lacher DW, et al. Rapid genomic-scale analysis of *Escherichia coli* O104:H4 by using high-resolution alternative methods to next-generation sequencing. Applied and Environmental Microbiology. 2012;**78**:1601-1605

[83] Monecke S, Coombs G, Shore AC, Coleman DC, Akpaka P, Borg M, et al. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *staphylococcus aureus*. PLoS One. 2011;**6**(4):e17936

[84] Ben Zakour NL, Venturini C, Beatson SA, Walker MJ. Analysis of a *Streptococcus pyogenes* puerperal sepsis cluster by use of whole-genome sequencing. Journal of Clinical Microbiology. 2012;**50**:2224-2228

[85] Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. New England Journal of Medicine. 2012;**366**:2267-2275

[86] Eyre DW, Golubchik T, Gordon NC, Bowden R, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open. 2012;**2:**e001124–e001124

[87] Nonhoff C, Rottiers S, Struelens MJ. Evaluation of the Vitek 2 system for identification and antimicrobial susceptibility testing of *Staphylococcus* spp. Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases. 2005;**11**:150-153

[88] Harris SR, Feil EJ, Holden MTG, Quail MA, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;**327**:469-474

Chapter 7

Genotyping for *Plasmodium* spp.: Diagnosis and Monitoring of Antimalarial Drug Resistance

Jean Bernard Lekana-Douki and Larson Boundenga

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76601

Abstract

Malaria is one the world's most widespread lethal diseases. *Plasmodium falciparum*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi* induce human pathology. These species could be differentially diagnosed using the genotyping of cytochrome b, *Pf*dhfr and RNA 18S. The persistence of *P. falciparum*, the most lethal parasite, is mainly due to antimalarial drug resistance. Indeed, a few years after the start of the ambitious malaria eradication program in 1960, chloroquine resistance emerged in Asia and spread widely in all the endemic areas. It was associated with genotypes in *P. falciparum* chloroquine resistance transporter (CVIET, SVMNT, CVMNT, CVIDT, SVIET and CVMET). The use of new drugs such as sulfadoxine-pyrimethamine (SP) leads quickly to SP-resistant parasites associated with genotypes on *P. falciparum* DiHydroFolate reductase (I51-R59-N108-I164) and *P. falciparum* DiHydroPteroate synthetase (436-437-580-613). Recently, the delay of parasite clearance has been described with artemisinine (the most efficacious antimalarial drug). This resistance was associated with the K13 propeller genotype. Since malaria species and antimalarial drug resistance markers could be characterized using nucleic acid sequences, genotyping is needed for malarial monitoring of species distribution and antimalarial drug resistance.

Keywords: *Plasmodium* parasites, drug resistance, diagnostic, genotyping

1. Introduction

Malaria remains a major public health problem. More than 40% of the world's population (3.3 billion) live malarial endemic areas in varying degrees. Despite tremendous efforts in the fight, and though this strategy or plan resulted in significantly decreasing the burden in the last 20 years, Malaria still is persistent in nearly 91 countries (**Figure 1**). In 2016, the overall incidence was 216 million cases among these and 445,000 deaths were recorded [1]. Africa continues to account for 90% of malaria burden. African children under 5 years of age are the most affected. This infectious disease is due to the invasion of *Plasmodium* spp. Currently, the five species are known to infect and to induce significantly malaria in humans are *Plasmodium ovale, Plasmodium malariae, Plasmodium knowlesi, Plasmodium vivax and Plasmodium falciparum*.

P. falciparum is the most virulent species as it is responsible for more than 90% of malaria deaths [2, 3]. Its specific biology with antigenic variation, sequestration of infected blood cells and interactions with host cells leads to severe malaria [4, 5]. Also, it is the most recent human infection with limited adaptation in the host [6]. *P. knowlesi*, which is a specific Asian monkey parasite (*Macaca* genus), was recently transferred to humans, causing high mortality in the south of Asia [7–9]. One of its differences from other species is the time of its life cycle which is 24 h whereas 40–48 h for other human *Plasmodium* spp. *P. vivax* is the most prevalent in Asia and South America. Due to the Duffy-negative statue of Black people, it is the rarest in Black Africa. However, recent studies reported the presence of *P. vivax* in Blacks with Duffy-negative from some countries of central Africa such as Equatorial Guinea, Congo Republic *and so on* [10, 11]. *P. malariae* is less prevalent in Asia, while it is most common in sub-Saharan Africa and southwest Pacific [12]. It often finds minor prevalence compared to *P. falciparum*. This parasite is thought to be a zoonotic infection because is genetically close to *P. brasilianum* which infects monkeys of South America [12, 13]. *P. ovale* is prevalent in Sub-Saharan Africa, South-East Asia, India, Papua New Guinea, Timor and Indonesia [14]. It is the less-prevalent human malarial parasite. However, in most places where *P. ovale* is observed, it is relatively uncommon and its prevalence rarely exceeds 5% [12].

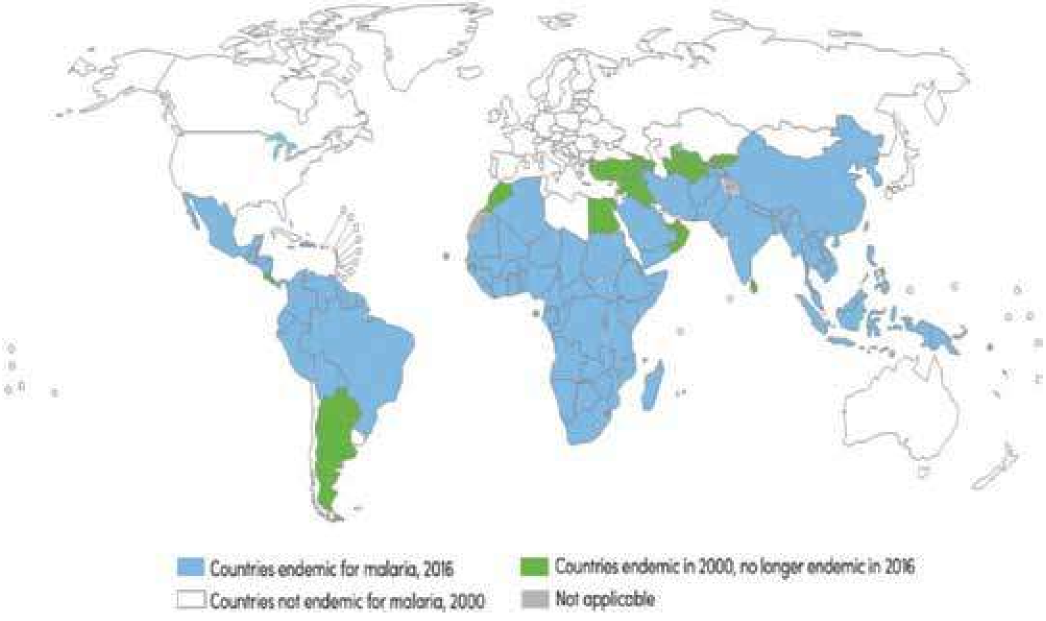

Countries endemic for malaria, 2016
Countries not endemic for malaria, 2000
Countries endemic in 2000, no longer endemic in 2016
Not applicable

Figure 1. Countries endemic for malaria in 2000 and 2016. From WHO [1].

Decisions concerning malaria treatment depend on the identification of the species causing the disease. Traditionally, this diagnosis was based on the microscopic detection of *Plasmodium* parasites in Giemsa-stained blood slides. In recent decades, antigen detection assays and molecular detection assays were introduced as alternatives to microscopy [15]. These approaches were very useful; however, they are not very reliable. Indeed, the morphological features and life history traits of a parasite species can vary from one host species to another [16, 17]. For antigen detection assays are mainly aimed at the identification of *P. falciparum*. Indeed [18], only a few assays are able to identify infections caused by other human malaria parasites [15, 19]. However, the development of molecular tools for the identification of species in diagnosis and genotyping permits a better reading of plasmodial diversity circulating among the human population and allows best highlighting the phenomenon of resistance of *Plasmodium* than microscopic tools.

Plasmodium species like several other genera have specific genetic markers such as *18S rRNA, ITS, cytochrome b* and so on, used in studying speciation. These markers play an important role in molecular analysis of genotyping and monitoring antimalarial drug resistance. The persistence of malaria burden is partly due to the emerging and widespread nature of the antimalarial drug resistance. Indeed, in the 1950s, WHO launched the malaria eradication program, with chloroquine and *dichloro-diphenyl-trichloroethan* (DDT). But in the early 1960s, chloroquine resistance emerged [20, 21]. Antimalarial drug resistance is defined as the parasite's ability to survive after the absorption of drug doses is greater than patient-tolerated doses. This chapter described the methods of genotyping the malaria for species diagnosis that helps to monitor drug resistance.

2. Genotyping for *Plasmodium* spp. diagnosis

The genotyping of *Plasmodium* spp. infections allows for the characterization of distinct species and subpopulations present in hosts. The genotyping techniques presented in the following allow for the characterization of different *Plasmodium* species by sizing the polymerase chain reaction (PCR) product of the polymorphic marker gene merozoite surface protein.

2.1. PCR using the specificity of 18S RNA

rRNA is one of the ribosome components. Among rRNA, the *18S rRNA* gene is the most frequently cited marker for malaria detection. It is composed of highly conserved regions which can be targeted for a qualitative detection of *Plasmodium* spp. and of variable zones allowing species identification. Early in the 1990s, Snounou and colleagues reported high-sensitivity methods for the detection of plasmodium species using nested PCR [22]. This method was based on the conversion of 18S rRNA among human plasmodium and the specificity of 18S rRNA from each parasite. So, the method includes the first PCR with primers named rPLU1/5 and rPLU3/4 that match with human plasmodium. This step is followed by the amplification of the product of primary PCR separately with the four species-specific primer pairs: rFAL1 and rFAL2, rVIV1 and rVIV2, rMAL1 and rMAL2 and rOVA1 and rOVA2 to identify the species *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale*, respectively. The primary PCR with rPLU1/5 and rPLU3/4 gives a 1100 bp band in a 2% agarose gel electrophoresis (**Figure 2**), in the presence of *Plasmodium*.

The nested PCR rFAL1 and rFAL2 generate a 205 bp in presence of the *P. falciparum* parasite. The nested PCR rVIV1 and rVIV2 generate a 120 bp in presence of the *P. vivax* parasite. The nested PCR rMAL1 and rMAL2 generate a 144 bp in presence of the *P. malaria* parasite. The nested PCR rOVA1 and rOVA2 generate an 800 bp in presence of the *P. ovale* parasite [22, 23].

2.2. RT-PCR NASBA 18S rRNA

Nucleic acid sequence-based amplification (NASBA) is a method in molecular biology, which is used to amplify RNA sequences. This novel approach of genotyping, based on the amplification of nucleic acid sequence (real-time QT-NASBA), was developed by Compton [24]. Immediately after its discovery, the NASBA method was used for the rapid diagnosis and quantification of HIV-1 in patients [25]. Some years later, Schooner et al. [26] developed a real-time quantitative nucleic acid sequence-based amplification (real time QT-NASBA) for the detection of *Plasmodium falciparum* 18S rRNA with a sensitivity of 10–50 parasites/ml [26, 27]. Thus, NASBA method that uses primers and probes were selected on the basis of the sequence of the small subunit 18S rRNA gene [26, 28], to characterize or identify the different human *Plasmodium* species [18]. For NASBA, Schooner et al. have defined primers Plas-1F (59-TCAGATACCGTCGTAATCTTA-39) and Plas-2R T7 (59-AATTCTAATACGACTCACTATAGGGAGAGAACTTTCTCGCTTGCGCGAA-39) which were used [26]. The RNAs from *P. falciparum, P. malariae, P. vivax* and *P. ovale* are specifics

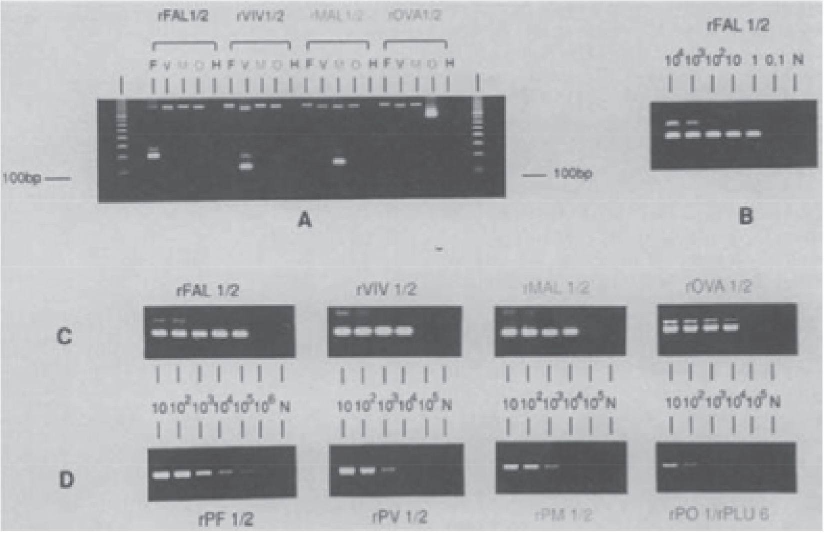

rFAL1/2
rVIV1/2
rMAL1/2
rOVA1/2
F V M O H F V M O H F V M O H F V M O H
100bp
100bp
A
rFAL 1/2
10^4 10^3 10^2 10 1 0.1 N
B
C
rFAL 1/2
rVIV 1/2
rMAL 1/2
rOVA 1/2
10 10^2 10^3 10^4 10^5 10^6 N
10 10^2 10^3 10^4 10^5 N
10 10^2 10^3 10^4 10^5 N
10 10^2 10^3 10^4 10^5 N
D
rPF 1/2
rPV 1/2
rPM 1/2
rPO 1/rPLU 6

Figure 2. Specificity and sensitivity of the PCR detection assay. (a) Nested PCR amplification for the demonstration of the specificity of the primers employed. Control genomic DNA from *P. falciparum* (F), *P. vivax* (V), *P. malariae* (M), *P. ovale* (O) and human blood (H); (b) nested PCR assay for the detection of in vitro cultured *P. falciparum* ring-stage parasites; (c) nested PCR amplification using diluted control DNAs; (d) product of amplification of diluted control DNAs, using the PCR assay (Source: Snounou et al. [22]).

and they should all be amplified by the NASBA isolates with these primers. Therefore, detection is done by capture probe (59-ACCATAAACTATG CCGACTAGG-39) which is bound to magnetic beads. Finally, samples are hybridized separately to ruthenium-labeled WT (59-CCTTATGAGAAATCAAAGTC 39) and Q (59-AATAACTGCACCAGTGTATA-39) detection probes, followed by ECL detection in a NucliSens ECL reader [26, 29]. The NASBA method is very sensitive and specific. It can be used for the detection, identification and quantitative measurement of low parasitaemia of *Plasmodium species*, that is, the lower detection limit of the assay is 100–1000 molecules in vitro RNA for all human malaria parasites.

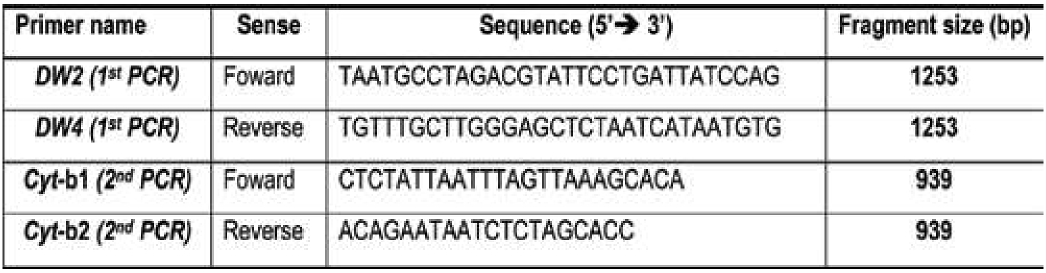
2.3. PCR sequencing cytochrome b

Cytochrome b (*Cyt*-b) is one of the respiratory chain systems present in mitochondria. It is highly conserved in *Plasmodium* species. But some of its region shows diversities and species specificities. It is also in multiple copies (20–100 copies) in the haploid genome, suggesting the increase of sensitivity. So, the alignment of *Cyt*-b sequences obtained after amplification by the nested PCR portion of mitochondrial DNA using couples of primers (**Table 1**) allows distinguishing the *Plasmodium* species present in the host (**Figure 3**).

The first round of PCR (PCR_1) and use of the primers DW2/F and DW4/R produce the fragments of 1253 bp. In the second round of PCR (PCR_2), primers *Cyt*-b1F and *Cyt*-b2R are used and this generates PCR fragments of 939 bp [30]. This approach was used more for the characterization of malaria parasites in primates (human and non-human) [31, 32] because it allows to identify all species known or unknown circulating in vertebrates [33].

2.4. PCR-RFLP from cytochrome b

Due to the specie specificity and diversity of Cytochrome b, it could be digested with the r estriction enzyme *Alu I* leading to species-specific patterns [34]. In this case, a nested PCR is used. Primers Plas 1 (5′-GAGAATTATGGAGTGGATGGTG-3′) and Plas 2 (5′-GTGGTAA TTGACATCCWATCC-3′) are used for primary PCR producing an 816 bp fragment, following by a nested with Plas 3 and Plas 4 primers. The n-PCR produces a 787 bp fragment. The digestion with *Alu I* gives 159 and 640 bp fragments; 187, 249 and 381 bp fragments; 111, 169, 187, 249 and 270 bp; and 224 and 584 bp fragments for *P. falciparum*, *P. malariae*, *P. vivax* and *P. ovale*, respectively, on agarose gel electrophoresis (**Figure 4**).



Primer name	Sense	Sequence (5'→ 3')	Fragment size (bp)
DW2 (1st PCR)	Foward	TAATGCCTAGACGTATTCCTGATTATCCAG	1253
DW4 (1st PCR)	Reverse	TGTTTGCTTGGGAGCTCTAATCATAATGTG	1253
Cyt-b1 (2nd PCR)	Foward	CTCTATTAATTTAGTTAAAGCACA	939
Cyt-b2 (2nd PCR)	Reverse	ACAGAATAATCTCTAGCACC	939

Table 1. Amplification primers of the cytochrome b gene (*Cyt*-b). *Cyt*-b is amplified by nested PCR.

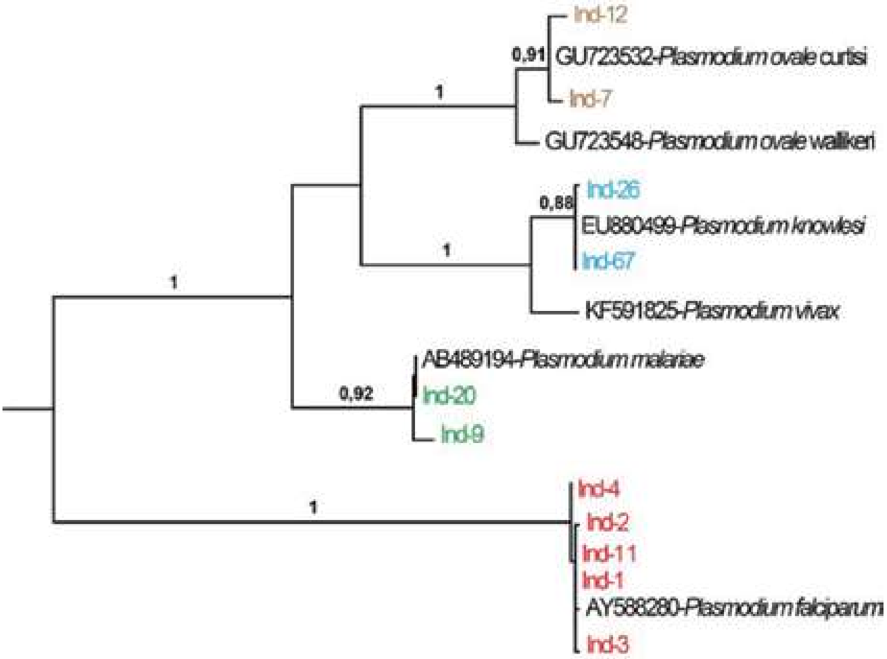

Ind-12
0,91
GU723532-Plasmodium ovale curtisi
1
Ind-7
GU723548-Plasmodium ovale wallikeri
0,88
Ind-26
EU880499-Plasmodium knowlesi
1
Ind-67
1
KF591825-Plasmodium vivax
AB489194-Plasmodium malariae
0,92
Ind-20
Ind-9
Ind-4
1
Ind-2
Ind-11
Ind-1
AY588280-Plasmodium falciparum
Ind-3

Figure 3. The tree of relationship of *Plasmodium* species. This tree was built with a portion of *Cyt*-b sequences characterized among some patients and other *Cyt*-b sequences from Genbank (accession numbers AB489194, EU880499, GU723548 and GU723532). The tree was built using maximum likelihood with Cyt-b sequences of 925pb sizes.

2.5. PCR-DHFR

Figure 5 shows the equation of transformation of dihydrofolate to tetrahydrofolate. The dihdrofolate reductase is one of the important malaria proteins involved in the plasmodium folate synthesis. This gene is coded in chromosome 4 and is highly conserved between

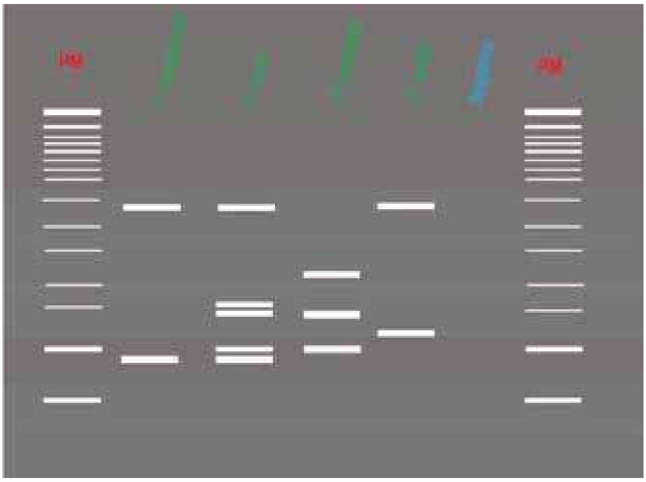


Figure 4. Illustration of a digestion cytochrome fragment using *Alu* enzyme.

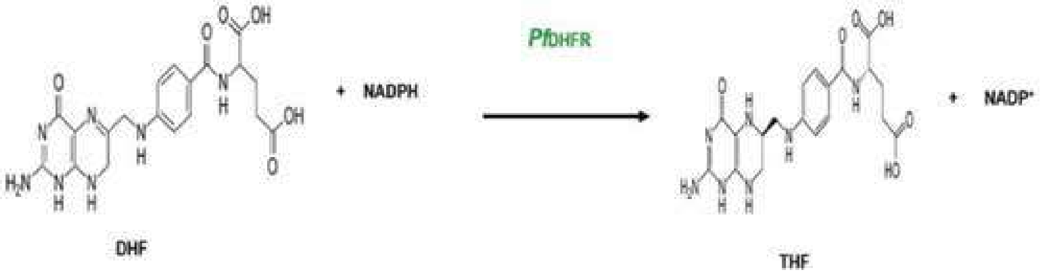


Figure 5. The equation of the DHF-THF reaction. Dihydrofolate reductase is an enzyme that catalyzes the NADPH-dependent reduction of dihydrofolate to tetrahydrofolate. This reaction is essential for the de novo synthesis of purines and certain amino acids. This enzyme is essential for rapid growth and is the target for the action of the important antimalarial drugs pyrimethamine and proguanil.

distantly related species, like plasmodium species. Its linker region revealed significantly a higher sequence diversity than the relatively conserved enzymatic diversity. Different species of *Plasmodium* are characterized by a unique linker sequence. So, it has been used to identify human plasmodium species [35]. Nonsynonymous mutations on DHFR are associated with pyrimethanime resistance. Using primers ATGGARSAMSTYTSMGABGTWTTYGA and AAATATTGRTAYTCTGGRTG for primary PCR gives a 1000 bp fragment. The nested PCR species specific with primers shown in **Table 2** allow to amplify specifically 160, 177, 144, 231/237/243 and 134 bp fragments for *P. falciparum, P. malariae, P. vivax, P. ovale and P. knowlesi.*

Primer	Species	Sequence (5′–3′)	Annealing (°C)	No. of cycles	Product (bp)
Pla-DHFR-F Pla-TS-R	*Plasmodium sp.*	ATGGARSAMSTYTSMGABGTWTTYGA AAATATTGRTAYTCTGGRTG	50	30	1000
Pla-DHFR-NF Pla-TS-R	*Plasmodium sp.*	AAATGYTTYATYATWGGDGG AAATATTGRTAYTCTGGRTG	55	35	509–587
PF-Lin-F PF-Lin-R	*P. falciparum*	AAAAGGAGAAGAAAAAAATAA AAAATAAACAAAATCATC	50	35	160
PM-Lin-F PM-Lin-R	*P. malariae*	GACCCAAGAATCCCTCCC CCCATGAAGTTATATTCC	50	35	177
PV-Lin-F PV-Lin-R	*P. vivax*	CGGGAGCACTGCGGACAGCG CACGGGCACGCGGCGGGGC	55	35	144
PO-Lin-F PO-Lin-R	*P. ovale*	GACACACAAAATGATGGGGA ATTGTCCTTTCCTTGACTCG	55	35	231, 237 or 243
PK-Lin-F PK-Lin-R	*P. knowlesi*	CGATGGATATGGATAGTGG CGCGGGAGAGCATTTCCTC	58	35	134

Table 2. Primer sequences and the PCR condition for detection of ***Plasmodium spp***. that infect humans.

3. Genotyping for the monitoring of antimalarial drug resistance

3.1. Antimalarial drug resistance

Antimalarial drug resistance is the ability of *P. falciparum* to survive after absorption of the drug at concentrations greater than concentrations tolerated by the patient. Drug resistance arises rarely because it is the result of some non-lethal mutations, but it spreads relatively quickly. The clinical failure treatment is the lasted event in the long way of changes in parasites genes. Antimalarial drug resistance markers are genes associated with antimalarial drug resistance. Among them, the most characterized are *Pf*crt (*P. falciparum chloroquine resistance transporter*), *Pf*mdr1 (*P. falciparum multidrug resistance 1*), *Pfdhfr* (*P. falciparum dihydrofolate reductase*), *Pfdhps* (*P. falciparum dihydropteroate synthase*), *Pfmrp* (*P. falciparum multidrug resistance protein*), *Pfatpase 6* (*P. falciparum atpase 6*) and *Pf*K13 (*P. falciparum kelch 13*).

3.2. The genotyping of markers associated with antimalarial drug resistance

3.2.1. P. falciparum chloroquine resistance transporter (Pfcrt)

*Pf*crt is an ATP binding cassette (ABC) protein, able to fixe and hydrolyze ATP. The general structure of the ABC transporter contains cytosolic nucleotide binding domains (NBDs), a nucleotide hydrolysis site and transmembrane segments. These transporters extrude effective drugs from the digestive vacuole and function as an efflux pump leading to the decrease of intracellular accumulation of drugs in the parasite. The genetic cross-experiment between chloroquine-sensitive and chloroquine-resistance strains allowed the identification of *P. falciparum* chloroquine-resistance transporter (*Pf*crt) in 2000 [20, 36]. It is a 45 kDa protein coding in chromosome 7, containing 10 predicted transmembrane domains located on the membrane of the digestive vacuole (**Figure 6**).

Several mutations have been identified in this transporter. The main mutation T76 allows for the abolition of accumulation of the drug chloroquine in the digestive vacuole. The association of mutations in codons 72, 73, 74, 75 and 76 defined different haplotypes. These haplotypes show a spatio-temporal specificity. CVMNK is the wild-type haplotype that is found in chloroquine-sensitive parasites. In Africa, the most prevalent chloroquine-resistance haplotype is the CVIET. It was also found with less prevalence in South America and in Southeast of Asia. Another *Pf*crt resistance haplotype, named South American haplotype, is the SVMNT. That is rarely found in Africa and Asia, has relatively little fitness cost and was associated with the emergence of amodiaquine resistance too [37, 38]. The other main *Pf*crt-resistance haplotypes were CVMNT, CVMET, SVIET and CVIDT. CVMNT is most prevalent in South America and in Asia but rarely found in Africa [39, 40]. CVMET is the rarest haplotype found in Asia whereas SVIET is the South American haplotype. CVIDT is the specific haplotype from Asia. Haplotypes CVIET and SVMNT were also associated with the plasmodial resistance against amodiaquine and lumefantrine [41]. In our recent study from Gabon and Congo we found new haplotypes. But the involvement of these in antimalarial resistance is needed yet (unpublished yet). The withdrawal of chloroquine has led to the decrease of the T76 genotype.

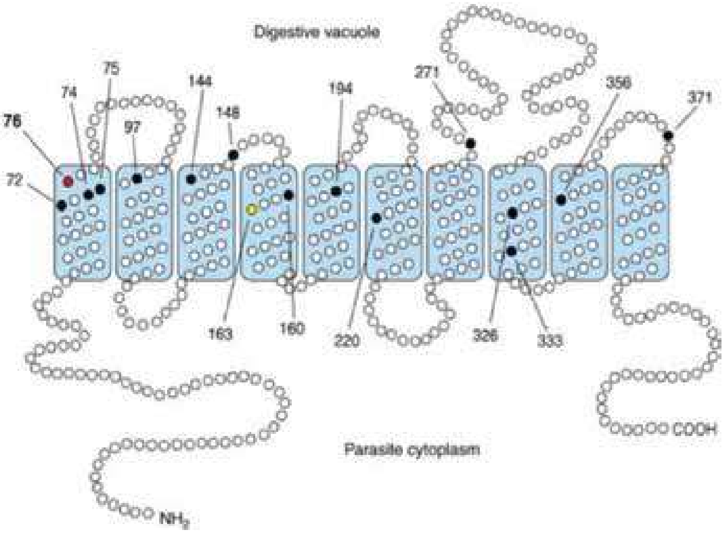


Figure 6. The predicted structure and representative haplotypes of *P. falciparum* chloroquine-resistance transporter. *Pf*CRT is predicted to have 10 transmembrane domains, with its N and C terminals located on the cytoplasmic side of the digestive vacuole membrane. Mutations identified in the *Pf*crt cDNA sequences from CQR lines (black circles), the crucial K76T mutation common to all CQR strains (red) and the S163R mutation identified in the amantadine- and halofantrine-resistant parasites (yellow circle) are indicated (this figure was pulled in [40]).

Furthermore, recent data showed that the use of artemisinine-based combination therapies, mainly artemether-lumefantrine, selected the wild-type K76 genotype. Other polymorphisms in *Pf*crt were described including A144F, I194T, A220S, Q271E and N326S [40].

For the genotype haplotype 72–76 of *Pf*crt, PCR sequencing and PCR-RFLP are used. That is based on nested PCR with several primers (**Table 3**). The RFLP pattern indicates a specific genotype for each codon.

3.2.2. *Plasmodium falciparum multidrug resistance 1 (Pfmdr1)*

P. falciparum multidrug resistance 1 is a homologous of the bacterial multidrug resistance 1 protein. It is one of the main antimalarial drug resistance markers. Like *Pf*crt, *Pf*mdr1 is also an ABC transporter, located in the membrane of a digestive vacuole, coded in chromosome 5 by the 4758 bp gene. Its structure shows two homologous parts containing six transmembrane domains plus a nucleotide binding domain (**Figure 7**).

Some isolates exhibit multicopies of *Pfmdr1*. Polymorphisms in codons 82, 184, 1034, 1042 and 1246 (N86Y, Y184F, S1034C, N1042D and D1246Y) were associated with antimalarial resistance against mefloquine, lumefantrine, artemether, halofantrine, quinine and chloroquine [36, 42]. Copy number and polymorphisms of the *pfmdr1* gene have been investigated as molecular markers of mefloquine resistance. With the treatment of artemether-lumefantrine, a selection of N86 genotype was reported [43]. That was confirmed in areas where this ACT was implemented [44]. In same way, it was shown that ACT led to the selection of haplotypes NFD, NYD in codons 86, 184, 1246.

Genes, codons	Primer names	Primers	T°C	Restriction enzyme	Sizes (bp)*
Pfcrt, C72S	CRT72MS	TTTATATTTTAAGTATTATTTATTTAAGTGGA	55	*Mbo I*	55 + 38
	76-D2	CAAAACTATAGTTACCAATTTTG			
Pfcrt, M74I	CRT745MS	TAAGTATTATTTATTTAAGTGTATGTGTCAT	55	*Nla III*	53 + 31
	76-D2	CAAAACTATAGTTACCAATTTTG			
Pfcrt, N75M	CRT745MS	TAAGTATTATTTATTTAAGTGTATGTGTCAT	50	*BspHI*	53 + 31
	76-D2	CAAAACTATAGTTACCAATTTTG			
Pfcrt, 76	Pfcrt-76A	GCGCGCGCATGGCTCACGTTTAGGTGGAG	55	*Apo I*	136 + 56
	Pfcrt-76B	GGGCCCGGCGGATGTTACAAAACTATAGTTACC			

Sizes* indicate the sizes of fragments generated after restriction enzyme digestions. T°C= hybridization temperature during PCR program.

Table 3. Sequences of primer sets and restriction enzymes used to characterize polymorphisms.

To the genotype Pfmdr1, followed primers and restriction enzymes are used for PCR-RFLP or for PCR followed by sequencing gene. Digestion of PCR products gives fragments of 126 and 165 bp for mutant 86Y whereas wild type N86 is not digested by restriction enzyme *Afl III*. In codon 1246, wild-type D1246 is digested by restriction enzyme *Bgl II,* giving fragments of 113 and 90 bp, whereas the mutant is not digested (**Table 4**). For codon 184, the genotype is assessed using the PCR followed by sequencing.

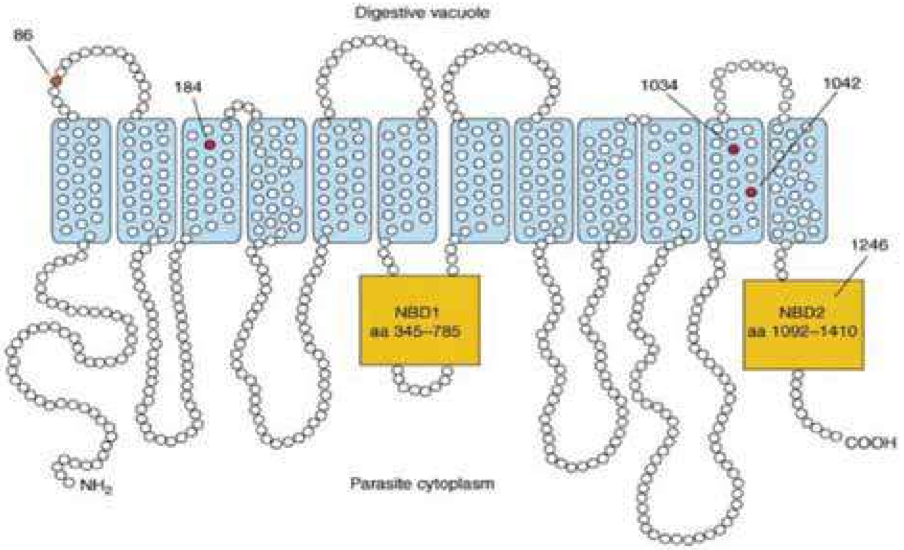


Figure 7. Genetic polymorphisms in *P. falciparum* multidrug resistance-1. *Pf*MDR1 has two homologous halves, each with six predicted transmembrane domains and a nucleotide-binding pocket. The nucleotide-binding domains (NBD1 and NBD2; orange boxes) are each formed by large cytoplasmic domains. Polymorphic amino acids found in the K1 allele (N86Y) and the 7G8 allele (Y184F, S1034C, N1042D and D1246Y) are indicated. The D1246Y mutation is located in the predicted NBD2 (this figure was pulled, from Ref. [40]).

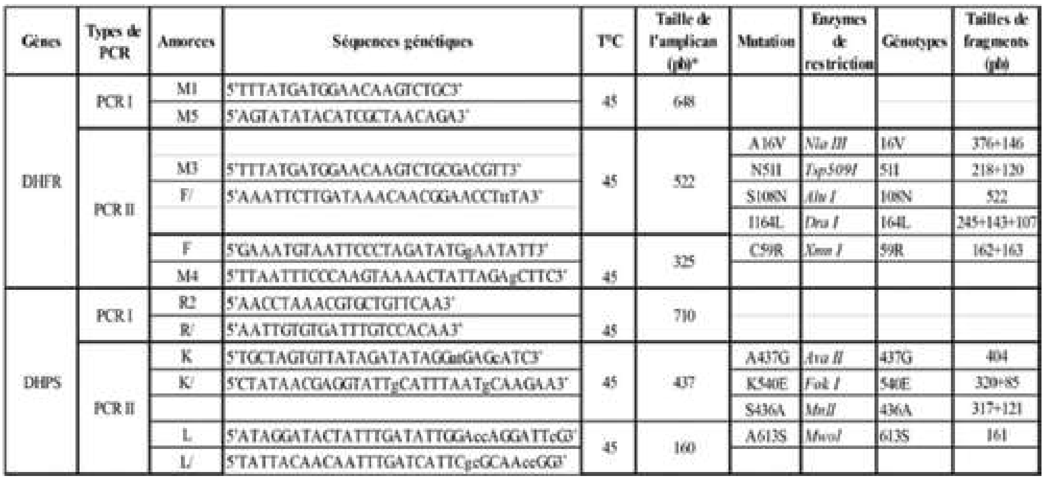
Genes, Codons	Primer names	Primers	T°C	Restriction enzyme	Sizes (bp)*
Pfmdr1, N86Y	mdr86D1	TTTACCGTTTAAATGTTTACCTGC	45	*Afl III*	126 + 165
	mdr86D2	CCATCTTGATAAAAAACACTTCTT			
Pfmdr1, D1246Y	mdr1246D1	AATGTAAATGAATTTTCAAACC	45	*Bgl II*	113 + 90
	mdr1246D2	CATCTTCTCTTCCAAATTTGATA			

Sizes* indicate the sizes of fragments generated after restriction enzyme digestions. T°C = hybridization temperature during PCR program.

Table 4. Sequences of primer sets and restriction enzymes used to characterize ***Pf***mdr1 polymorphisms.

3.2.3. *Plasmodium falciparum dihydrofolate reductase (PfDHFR)*

P. falciparum dihydrofolate reductase is mainly involved in the synthesis of the thymidine base. This enzyme is the target of pyrimethamine which blocked the synthesis of DNA and lead to the death of the parasite. After implementation of this drug, isolates with resistance against it were described. This parasite carried the mutations in the *Pf*dhfr gene. Among these, the mutation of serine to asparagine in codon 108 is the main resistance mutation. Additional mutations in codons 16, 51, 59 and 164 contribute to the increase of the resistance level of the parasite against pyrimethamine. The double mutant 108N + 51I and 108N + 59R increased the IC50 of parasites from 2 to 16 times, compared to the simple mutant N108 [45, 46]. In this way the triple mutant 51I + 59R + 108N or 59R + 108N + 164R shows the highest resistance level compared to the double mutant. Quadruple mutants exhibit a higher level of resistance compared to the triple mutant [47]. In Africa, the AIRNI haplotype is the most prevalent. PCR-RFLP was developed to genotype *Pf*dhr codons associated with pyrimethamine resistance. For these, primers and restriction enzyme are contained in **Table 5**..



Gènes	Types de PCR	Amorces	Séquences génétiques	T°C	Taille de l'amplican (pb)*	Mutation	Enzymes de restriction	Génotypes	Tailles de fragments (pb)
DHFR	PCR I	M1	5'TTTATGATGGAACAAGTCTGC3'	45	648				
		M5	5'AGTATATACATCGCTAACAGA3'						
	PCR II			45	522	A16V	*Nla III*	16V	376+146
		M3	5'TTTATGATGGAACAAGTCTGCGACGTT3'			N51I	*Tsp509I*	51I	218+120
		F/	5'AAATTCTTGATAAACAACGGAACCTttTA3'			S108N	*Alu I*	108N	522
						I164L	*Dra I*	164L	245+143+107
		F	5'GAAATGTAATTCCCTAGATATGgAATATT3'		325	C59R	*Xmn I*	59R	162+163
		M4	5'TTAATTTCCCAAGTAAAACTATTAGAgCTTC3'	45					
DHPS	PCR I	R2	5'AACCTAAACGTGCTGTTCAA3'		710				
		R/	5'AATTGTGTGATTTGTCCACAA3'	45					
	PCR II	K	5'TGCTAGTGTTATAGATATAGGatGAGcATC3'	45	437	A437G	*Ava II*	437G	404
		K/	5'CTATAACGAGGTATTgCATTTAATgCAAGAA3'			K540E	*Fok I*	540E	320+85
						S436A	*MnlI*	436A	317+121
		L	5'ATAGGATACTATTTGATATTGGAccAGGATTcG3'	45	160	A613S	*MwoI*	613S	161
		L/	5'TATTACAACAATTTGATCATTCgcGCAAccGG3'						

Table 5. Primer sets and RFLP conditions to genotypes ***Pf***DHFR and ***Pf***DHPS.

3.2.4. Plasmodium falciparum dihydropteroate synthase (PfDHPS)

P. falciparum dihydropteroate synthase (*Pf*dhps) is one of the enzymes involved in the line of thymidine synthesis. It transforms pteridine to dihydropteroate in the presence of pABA. It is coded by the gene located in chromosome 8 of *P. falciparum*. The implementation of sulfadoxine to treat malaria led to *P. falciparum* isolates' resistance against this drug. Genotype analysis reported the association of *Pf*dhps mutations with sulfadoxine resistance [48]. So, mutations S436A, A437G, K540E, G581A and A613S were reported [49]. The mutation A437G increased the IC_{50} of isolates by five times [50]. Isolates with triple mutations S436A + A437G + K540E and S436A + A437G + A613S showed the level of sulfadoxine resistance around 9–24 times higher than the single mutant.

Triple mutant I51R59N108 in *Pf*dhfr and double mutants G437E540 and G437S581S in Pfdhps increase the risk of failure when treated with sulfadoxine-pyrimethamine [51]. The drug resistance against sulfadoxine and pyrimethamine could be monitored by genotyping *Pf*dhfr and *Pf*dhps according to the PCR-RFLP conditions shown in **Table 5**.

3.2.5. Plasmodium falciparum ATPase 6 (PfATPASE6)

The sarcoplasmic/endoplasmic reticulum Ca^{2+}-ATPase, ortholog of *P. falciparum* (*PfSERCA* or *PfATPase6*), is an active Ca^{2+} protein transporter. This is a part of P-type ATPases enzymes that transport ions across biological membranes with the energy provided by the ATP hydrolysis. Several P-type ATPases, that were reported in *P. falciparum* and *Pf*serca, correspond to type 6. The gene coding this protein is located in chromosome 1 and contains 3687 bp. The protein allows the trafficking of calcium though the sarcoplasmic-endoplasmic membrane. Several polymorphisms such as S769N, Y243H, K431E, G110A, A2694T or E623A were reported in *Pf*ATPase 6.

*Pf*ATPase 6 has been associated with antimalarial drug resistance [52]. The most important paper on this protein was the report of its involvement in artemisinin resistance [53]. But, the role of *Pf*ATPase 6 in artemisinin resistance was strongly disputed. S769 N was associated with artemether resistance in French Guinea [53]. However, several articles investigating *Pf*ATAPase 6 in artemisinin drug resistance failed to confirm this point [54–56]. Genotyping using PCR followed by sequencing is usually used to monitor antimalarial drug resistance. Primers used are previously described by Tahar et al. [57].

3.2.5.1. P. falciparum multidrug resistance protein (PfMRP)

P. falciparum multidrug resistance protein is another marker of antimalarial drug resistance. It belongs to the C subfamily of ABC transporters containing two NBDs, two membrane-spanning domains and six transmembrane domains. *Pf*mrp can transport glutathione, glucuronate, as well as glucuronide and sulfate-conjugated compounds that increased susceptibility to several antimalarial drugs like chloroquine, quinine or artemisinin [58]. *Pf*mrp1 and *Pf*mrp2 are expressed in the plasma membrane in all asexual stages of the parasite. Their expression is upregulated by mefloquine and chloroquine. *Pf*mrp1 is a 215 kDa protein coding on chromosome 1. Mutations on codons H191Y and S437A were associated with quinoline resistance [59]. But no association was found between these polymorphisms and resistance against pyronaridine [60]. The hypothetic mechanism of involving of *Pf*mrp on antimalarial drug resistance is the pump efflux mechanism associated with the extrusion of glutathione. The genotyping of *Pf*mrp 1 and Pfmrp2 is achieved using PCR followed by sequencing. The primers used for amplification and sequencing are pfmrp-501F 5′-TTT CAA AGT ATT CAG TGG GT-3′ and pfmrp-1409R 5′-GGC ATA ATA ATT GAT GTA AA-3′.

3.2.5.2. P. falciparum Kelch 13 (PfK13)

Chinese populations used *Artemisia annua* to treat malaria for a long time. Over the years, artemisinin was extracted from this plant. That is a sesquiterpene lactone. Now, artemisinin is one of the best antimalarial drugs. So, WHO recommends the use of this drug in association with other antimalarial drugs, artemisinin combination therapy (ACT) for the treatment of uncomplicated malaria. Since the 1990s, the use of artemisinin was highly intensified in Asia. Consequently, the delay of *P. falciparum* clearance was reported after treatment with artemisinin which translates to artemisinin resistance [56]. These slow clearance rates are associated with enhanced survival rates of ring-stage parasites briefly exposed in vitro to dihydroartemisinin. Recently, a large-scale study identified molecular markers of this resistance: polymorphisms on propeller kelch 13 (K13) [61]. That is a region on chromosome 13 (**Figure 8**) [62].

Several investigations on K13 genotyping reported that mutations M476I, Y493H, R539T, I543T, P553L and C580Y conferred a greater artemisinin resistance [63]. Other mutations F446I and A578S were described in PfK13. A578S, widespread in Africa, is not associated with artemisinin resistance. These genotypes are investigated by PCR sequencing.

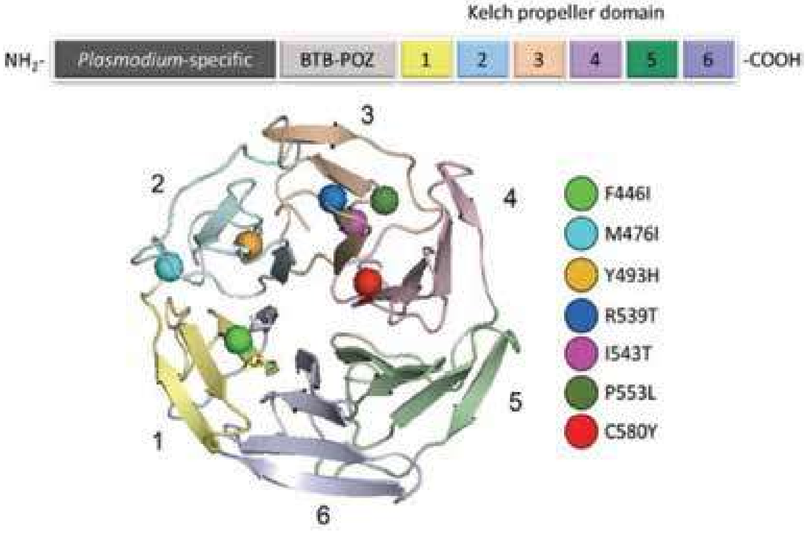


Figure 8. *P. falciparum* kelch13 (K13) protein. The parasite K13 protein consists of plasmodium-specific sequences, a BTB-POZ domain and six kelch domains that are predicted to form a six-blade propeller. In the structural model, the original M476I mutation discovered by Ariey et al. [61] and six other mutations associated with artemisinin resistance (all details of this figure was pulled of Fairhurst and Dondrop [62]) are shown.

4. Conclusions

Due to the nucleotide specificities of each *Plasmodium* species and the molecular changes associated with antimalarial drug resistance, genotyping is used for *Plasmodium* species diagnosis and monitoring antimalarial drug resistance. This genotyping could be achieved by specific PCR, PCR-RFLP, PCR-sequencing or RT-PCR of some molecular markers.

Author details

Jean Bernard Lekana-Douki[1,2]* and Larson Boundenga[1]

*Address all correspondence to: lekana_jb@yahoo.fr

1 Unité d'Evolution, Epidémiologie et Résistances Parasitaires (UNEEREP), Centre International de Recherches Médicales de Franceville (CIRMF), Franceville, Gabon

2 Département de Parasitologie-Mycologie Médecine Tropicale, Faculté de Médecine, Université des Sciences de la Santé, Libreville, Gabon

References

[1] WHO. World Malaria Report 2016. 2017. p. 196. ISBN: 978-92-4-156552-3

[2] Hawley WA, Phillips-Howard PA, ter Kuile FO, Terlouw DJ, Vulule JM, Ombok M, Nahlen BL, Gimnig JE, Kariuki SK, Kolczak MS. Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. The American Journal of Tropical Medicine and Hygiene. 2003;**68**:121-127

[3] Bull PC, Marsh K. The role of antibodies to *Plasmodium falciparum*-infected-erythrocyte surface antigens in naturally acquired immunity to malaria. Trends in Microbiology. 2002;**10**:55-58

[4] Berendt A, Ferguson D, Newbold C. Sequestration in *Plasmodium falciparum* malaria: Sticky cells and sticky problems. Parasitology Today. 1990;**6**:247-254

[5] Kyes S, Horrocks P, Newbold C. Antigenic variation at the infected red cell surface in malaria. Annual Review of Microbiology. 2001;**55**:673-707

[6] Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH. Out of Africa: Origins and evolution of the human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. International Journal for Parasitology. 2017;**47**:87-97

[7] Cox-Singh J, Singh B. Knowlesi malaria: Newly emergent and of public health importance? Trends in Parasitology. 2008;**24**:406-410

[8] Lee KS, Divis PC, Zakaria SK, Matusop A, Julin RA, Conway DJ, Cox-Singh J, Singh B. *Plasmodium knowlesi*: Reservoir hosts and tracking the emergence in humans and macaques. PLoS Pathogens. 2011;**7**:e1002015

[9] Cox-Singh J, Davis TM, Lee K-S, Shamsul SS, Matusop A, Ratnam S, Rahman HA, Conway DJ, Singh B. *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. Clinical Infectious Diseases. 2008;**46**:165-171

[10] Ngassa Mbenda HG, Das A. Molecular evidence of *Plasmodium vivax* mono and mixed malaria parasite infections in Duffy-negative native Cameroonians. PLoS One. 2014;**9**:e103262

[11] Russo G, Faggioni G, Paganotti GM, Djeunang Dongho GB, Pomponi A, De Santis R, Tebano G, Mbida M, Sanou Sobze M, Vullo V, Rezza G, Lista FR. Molecular evidence of *Plasmodium vivax* infection in Duffy negative symptomatic individuals from Dschang, West Cameroon. Malaria Journal. 2017;**16**:74

[12] Mueller I, Zimmerman PA, Reeder JC. *Plasmodium malariae* and *Plasmodium ovale*—The 'bashful'malaria parasites. Trends in Parasitology. 2007;**23**:278-283

[13] Ayala F, Escalante A, Rich S. Evolution of *Plasmodium* and the recent origin of the world populations of *Plasmodium falciparum*. Parassitologia. 1999;**41**:55-68

[14] Collins WE, Jeffery GM. *Plasmodium ovale*: Parasite and disease. Clinical Microbiology Reviews. 2005;**18**:570-581

[15] Makler MT, Palmer CJ, Ager AL. A review of practical techniques for the diagnosis of malaria. Annals of Tropical Medicine and Parasitology. 1998;**92**:419-433

[16] Manwell RD. Malaria infections by four species of *Plasmodium* in the duck and chicken, and resulting parasite modifications. American Journal of Hygiene. 1943;**38**:211-222

[17] Jordan HB. The effect of host constitution on the development of *Plasmodium floridense*. The Journal of Eukaryotic Microbiology. 1975;**22**:241-244

[18] Mens PF, Schoone GJ, Kager PA, Schallig HD. Detection and identification of human *Plasmodium* species with real-time quantitative nucleic acid sequence-based amplification. Malaria Journal. 2006;**5**:80

[19] Moody A. Rapid diagnostic tests for malaria parasites. Clinical Microbiology Reviews. 2002;**15**:66-78

[20] Trape JF. The public health impact of chloroquine resistance in Africa. The American Journal of Tropical Medicine and Hygiene. 2001;**64**:12-17

[21] Payne D. Spread of chloroquine resistance in *Plasmodium falciparum*. Parasitology Today. 1987;**3**:241-246

[22] Snounou G, Viriyakosol S, Zhu XP, Jarra W, Pinheiro L, do Rosario VE, Thaithong S, Brown KN. High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Molecular and Biochemical Parasitology. 1993;**61**:315-320

[23] Johnston SP, Pieniazek NJ, Xayavong MV, Slemenda SB, Wilkins PP, da Silva AJ. PCR as a confirmatory technique for laboratory diagnosis of malaria. Journal of Clinical Microbiology. 2006;**44**:1087-1089

[24] Compton J. Nucleic acid sequence-based amplification. Nature. 1991;**350**:91-92

[25] Kievits T, van Gemen B, van Strijp D, Schukkink R, Dircks M, Adriaanse H, Malek L, Sooknanan R, Lens P. NASBATM isothermal enzymatic in vitro nucleic acid amplification optimized for the diagnosis of HIV-1 infection. Journal of Virological Methods. 1991;**35**:273-286

[26] Schoone GJ, Oskam L, Kroon NC, Schallig HD, Omar SA. Detection and quantification of *Plasmodium falciparum* in blood samples using quantitative nucleic acid sequence-based amplification. Journal of Clinical Microbiology. 2000;**38**:4072-4075

[27] Schneider P, Schoone G, Schallig H, Verhage D, Telgt D, Eling W, Sauerwein R. Quantification of *Plasmodium falciparum* gametocytes in differential stages of development by quantitative nucleic acid sequence-based amplification. Molecular and Biochemical Parasitology. 2004;**137**:35-41

[28] Schneider P, Wolters L, Schoone G, Schallig H, Sillekens P, Hermsen R, Sauerwein R. Real-time nucleic acid sequence-based amplification is more convenient than real-time PCR for quantification of *Plasmodium falciparum*. Journal of Clinical Microbiology. 2005;**43**:402-405

[29] Van Gemen B, van Beuningen R, Nabbe A, van Strijp D, Jurriaans S, Lens P, Kievits T. A one-tube quantitative HIV-1 RNA NASBA nucleic acid amplification assay using electrochemiluminescent (ECL) labelled probes. Journal of Virological Methods. 1994; **49**:157-167

[30] Ollomo B, Durand P, Prugnolle F, Douzery E, Arnathau C, Nkoghe D, Leroy E, Renaud FA. New malaria agent in African hominids. PLoS Pathogens. 2009;**5**:e1000446

[31] Prugnolle F, Durand P, Neel C, Ollomo B, Ayala FJ, Arnathau C, Etienne L, Mpoudi-Ngole E, Nkoghe D, Leroy E, Delaporte E, Peeters M, Renaud F. African great apes are natural hosts of multiple related malaria species, including *Plasmodium falciparum*. Proceedings of the National Academy of Sciences of the United States of America. 2010;**107**:1458-1463

[32] Boundenga L, Ollomo B, Rougeron V, Mouele LY, Mve-Ondo B, Delicat-Loembet LM, Moukodoum ND, Okouga AP, Arnathau C, Elguero E, Durand P, Liégeois F, Boué V, Motsch P, Le Flohic G, Ndoungouet A, Paupy C, Ba CT, Renaud F, Prugnolle F. Diversity of malaria parasites in great apes in Gabon. Malaria Journal. 2015;**14**:111

[33] Boundenga L, Makanga B, Ollomo B, Gilabert A, Rougeron V, Mve-Ondo B, Arnathau C, Durand P, Moukodoum ND, Okouga AP, Delicat-Loembet L, Yacka-Mouele L, Rahola N, Leroy E, Ba CT, Renaud F, Prugnolle F, Paupy C. Haemosporidian parasites of antelopes and other vertebrates from Gabon, Central Africa. PLoS One. 2016;**11**:e0148958

[34] Steenkeste N, Incardona S, Chy S, Duval L, Ekala M-T, Lim P, Hewitt S, Sochantha T, Socheat D, Rogier C, Mercereau-Puijalon O, Fandeur T, Ariey F. Towards high-throughput molecular detection of *Plasmodium*. New approaches and molecular markers. Malaria Journal. 2009;**8**:86

[35] Tanomsing N, Imwong M, Theppabutr S, Pukrittayakamee S, Day NP, White NJ, Snounou G. Accurate and sensitive detection of *Plasmodium* species in humans by use

of the dihydrofolate reductase-thymidylate synthase linker region. Journal of Clinical Microbiology. 2010;**48**:3735-3737

[36] Basco LK, Ringwald P, Thor R, Doury JC, Le Bras J. Activity in vitro of chloroquine, cycloguanil, and mefloquine against African isolates of *Plasmodium falciparum*: Presumptive evidence for chemoprophylactic efficacy in central and West Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1995;**89**:657-658

[37] Beshir K, Sutherland CJ, Merinopoulos I, Durrani N, Leslie T, Rowland M, Hallett RL. Amodiaquine resistance in *Plasmodium falciparum* malaria in Afghanistan is associated with the pfcrt SVMNT allele at codons 72 to 76. Antimicrobial Agents and Chemotherapy. 2010;**54**:3714-3716

[38] Warhurst DC, Steele JC, Adagu IS, Craig JC, Cullander C. Hydroxychloroquine is much less active than chloroquine against chloroquine-resistant *Plasmodium falciparum*, in agreement with its physicochemical properties. The Journal of Antimicrobial Chemotherapy. 2003;**52**:188-193

[39] Sa JM, Twu O. Protecting the malaria drug arsenal: Halting the rise and spread of amodiaquine resistance by monitoring the *Pf*CRT SVMNT type. Malaria Journal. 2010;**9**:374

[40] Valderramos SG, Fidock DA. Transporters involved in resistance to antimalarial drugs. Trends in Pharmacological Sciences. 2006;**27**:594-601

[41] Mwai L, Kiara SM, Abdirahman A, Pole L, Rippert A, Diriye A, Bull P, Marsh K, Borrmann S, Nzila A. In vitro activities of piperaquine, lumefantrine, and dihydroartemisinin in Kenyan *Plasmodium falciparum* isolates and polymorphisms in pfcrt and pfmdr1. Antimicrobial Agents and Chemotherapy. 2009;**53**:5069-5073

[42] Bray PG, Ward SA. A comparison of the phenomenology and genetics of multidrug resistance in cancer cells and quinoline resistance in *Plasmodium falciparum*. Pharmacology & Therapeutics. 1998;**77**:1-28

[43] Sisowath C, Stromberg J, Martensson A, Msellem M, Obondo C, Bjorkman A, Gil JP. In vivo selection of *Plasmodium falciparum* pfmdr1 86N coding alleles by artemether-lumefantrine (Coartem). The Journal of Infectious Diseases. 2005;**191**:1014-1017

[44] Lekana-Douki JB, Dinzouna Boutamba SD, Zatra R, Zang Edou SE, Ekomy H, Bisvigou U, Toure-Ndouo FS. Increased prevalence of the *Plasmodium falciparum* Pfmdr1 86N genotype among field isolates from Franceville, Gabon after replacement of chloroquine by artemether-lumefantrine and artesunate-mefloquine. Infection, Genetics and Evolution. 2011;**11**:512-517

[45] Peterson DS, Walliker D, Wellems TE. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. Proceedings of the National Academy of Sciences of the United States of America. 1988;**85**:9114-9118

[46] Triglia T, Cowman AF. Primary structure and expression of the dihydropteroate synthetase gene of *Plasmodium falciparum*. Proceedings of the National Academy of Sciences of the United States of America. 1994;**91**:7149-7153

[47] Nzila-Mounda A, Mberu EK, Sibley CH, Plowe CV, Winstanley PA, Watkins WM. Kenyan *Plasmodium falciparum* field isolates: Correlation between pyrimethamine and chlorcycloguanil activity in vitro and point mutations in the dihydrofolate reductase domain. Antimicrobial Agents and Chemotherapy. 1998;**42**:164-169

[48] Osman ME, Mockenhaupt FP, Bienzle U, Elbashir MI, Giha HA. Field-based evidence for linkage of mutations associated with chloroquine (*pf*crt/*pf*mdr1) and sulfadoxine–pyrimethamine (pfdhfr/pfdhps) resistance and for the fitness cost of multiple mutations in *P. falciparum*. Infection, Genetics and Evolution. 2007;**7**:52-59

[49] Triglia T, Menting JG, Wilson C, Cowman AF. Mutations in dihydropteroate synthase are responsible for sulfone and sulfonamide resistance in *Plasmodium falciparum*. Proceedings of the National Academy of Sciences of the United States of America. 1997; **94**:13944-13949

[50] Triglia T, Wang P, Sims PF, Hyde JE, Cowman AF. Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. The EMBO Journal. 1998;**17**:3807-3815

[51] Picot S, Bienvenu AL, Konate S, Sissoko S, Barry A, Diarra E, Bamba K, Djimde A, Doumbo OK. Safety of epoietin beta-quinine drug combination in children with cerebral malaria in Mali. Malaria Journal. 2009;**8**:169

[52] Eckstein-Ludwig U, Webb R, Van Goethem I, East J, Lee A, Kimura M, O'neill P, Bray P, Ward S, Krishna S. Artemisinins target the SERCA of *Plasmodium falciparum*. Nature. 2003;**424**:957-961

[53] Jambou R, Legrand E, Niang M, Khim N, Lim P, Volney B, Ekala MT, Bouchier C, Esterre P, Fandeur T, Mercereau-Puijalon O. Resistance of *Plasmodium falciparum* field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6. Lancet. 2005;**366**:1960-1963

[54] Fairhurst RM, Nayyar GM, Breman JG, Hallett R, Vennerstrom JL, Duong S, Ringwald P, Wellems TE, Plowe CV, Dondorp AM. Artemisinin-resistant malaria: Research challenges, opportunities, and public health implications. The American Journal of Tropical Medicine and Hygiene. 2012;**87**:231-241

[55] Kiaco K, Teixeira J, Machado M, do Rosario V, Lopes D. Evaluation of artemether-lumefantrine efficacy in the treatment of uncomplicated malaria and its association with pfmdr1, pfatpase6 and K13-propeller polymorphisms in Luanda, Angola. Malaria Journal. 2015;**14**:504

[56] Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, Tarning J, Lwin KM, Ariey F, Hanpithakpong W, Lee SJ, Ringwald P, Silamut K, Imwong M, Chotivanich K, Lim P, Herdman T, An SS, Yeung S, Singhasivanon P, Day NP, Lindegardh N, Socheat D, White NJ. Artemisinin resistance in *Plasmodium falciparum* malaria. The New England Journal of Medicine. 2009;**361**:455-467

[57] Tahar R, Ringwald P, Basco LK. Molecular epidemiology of malaria in Cameroon. XXVIII. In vitro activity of dihydroartemisinin against clinical isolates of *Plasmodium falciparum* and sequence analysis of the *P. falciparum* ATPase 6 gene. The American Journal of Tropical Medicine and Hygiene. 2009;**81**:13-18

[58] Sanchez CP, Dave A, Stein WD, Lanzer M. Transporters as mediators of drug resistance in *Plasmodium falciparum*. International Journal for Parasitology. 2010;**40**:1109-1118

[59] Pirahmadi S, Zakeri S, Afsharpad M, Djadid ND. Mutation analysis in pfmdr1 and pfmrp1 as potential candidate genes for artemisinin resistance in *Plasmodium falciparum* clinical isolates 4years after implementation of artemisinin combination therapy in Iran. Infection, Genetics and Evolution. 2013;**14**:327-334

[60] Pradines B, Briolant S, Henry M, Oeuvray C, Baret E, Amalvict R, Didillon E, Rogier C. Absence of association between pyronaridine in vitro responses and polymorphisms in genes involved in quinoline resistance in *Plasmodium falciparum*. Malaria Journal. 2010;**9**:339

[61] Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, Khim N, Kim S, Duru V, Bouchier C, Ma L, Lim P, Leang R, Duong S, Sreng S, Suon S, Chuor CM, Bout DM, Ménard S, Rogers WO, Genton B, Fandeur T, Miotto O, Ringwald P, Le Bras J, Berry A, Barale JC, Fairhurst RM, Benoit-Vical F, Mercereau-Puijalon O, Ménard D. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. Nature. 2014;**505**:50-55

[62] Fairhurst RM, Dondorp AM. Artemisinin-resistant Plasmodium falciparum malaria. Microbiology Spectrum. 2016;**4:**EI10-0013-2016

[63] Straimer J, Gnädig NF, Witkowski B, Amaratunga C, Duru V, Ramadani AP, Dacheux M, Khim N, Zhang L, Lam S, Gregory PD, Urnov FD, Mercereau-Puijalon O, Benoit-Vical F, Fairhurst RM, Ménard D, Fidock DA. Drug resistance. K13-propeller mutations confer artemisinin resistance in *Plasmodium falciparum* clinical isolates. Science. 2015;**347**:428-431

Chapter 8

Resistance-Associated Substitutions/Variants Correlate to Therapeutic Outcomes of Novel Direct-Acting Antivirals in Different HCV Genotype Treated Individuals

Imran Shahid, Munjed Mahmoud Ibrahim,
Muhammad Usman Nawaz,
Mohammad Tarque Imam and Waleed H. AlMalki

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76395

Abstract

The expanded classification of hepatitis C virus (HCV) genome into various genotypes and numerous subtypes significantly correlates to therapeutic outcomes of interferon-free direct-acting antivirals (DAAs) in HCV treated patients. In particular, genotypes 3 and 4 are still harder to treat, and higher sustained virologic response (SVR) rates are not achieved in some difficult-to-treat specific populations (i.e., HCV subtype 1a patients, compensated and decompensated cirrhotic patients, HCV/HIV co-infection, and prior treatment failure with pegylated interferon plus ribavirin and first-generation protease inhibitor based therapeutic regimens). Furthermore, the pre-existing and treatment-emergent resistance associated substitutions (RAS) at specific amino acid positions within the viral quasispecies may increase the chances of viral breakthrough (HCV RNA remains lower limit of quantification, but increased to 100 IU/mL or $1.\log_{10}$ during DAAs therapy), viral relapse (undetectable viral load at the end of treatment but positive within the follow-up of 6 months), and discontinuation of therapy in treated individuals. Although the clinical importance of RAS is not entirely elucidated, it is believed that such substitutions decrease the therapeutic efficacy of DAAs in treated individuals. Similarly, the emergence of multiclass hepatitis C virus resistance to interferon-free DAAs failure in real-world experiences demands eagerly tailored second-line anti-hepatitis C therapies. This book chapter comprehensively overviews the clinical correlation of HCV genotypes, viral quasispecies and harboring RAS to treatment outcomes of revolutionary interferon-free DAAs in hepatitis C-treated patients.

IntechOpen

Keywords: hepatitis C virus, quasispecies, direct-acting antivirals, NS3/4A serine protease inhibitors, NS5A inhibitors, NS5B inhibitors, antiviral drug resistance, resistance-associated amino acid variants, resistance-associated substitutions, viral breakthrough, viral relapse, virologic failure

1. Introduction

The diverse genetic heterogeneity of hepatitis C virus genome, poor fidelity of virus replication enzyme (an RNA-dependent RNA polymerase enzyme (RdRp) encoded by NS5B protein in hepatitis C viral genome) and rapid HCV genome replication rate classify hepatitis C genome into various genotypes (GT) or clades (seven genotypes)) and numerous subtypes (at least 67 subtypes) [1, 2]. Such type of huge genetic diversity, a hallmark of single strand RNA viruses is amazing because of the discovery of the virus by molecular cloning methods and further nucleotide sequences from the plasma of a chimpanzee as compared to the isolation/characterization of other human RNA viruses [3, 4]. Afterwards, complete HCV genome were isolated and sequenced from different HCV isolates from various parts of the world [5]. The polymerase enzyme lacks proofreading mechanism of viral genome which generates closely related but diverse population of viral variants known as viral quasispecies even within the infected individuals (at a rate of approximately 1 mutation/replication cycle) [6, 7]. The propagation of HCV infection is a highly dynamic process due to a few hours of viral half-life, rapid replication rate *in vivo* and an error-prone nature of NS5B encoded viral replication enzyme [8, 9]. The viral progeny is produced by a rate of an estimated 10 trillion copies per day which exist as quasispecies of numerous closely related viral variants within a single patient [10]. Although, HCV based acquired immunity is developed after primary hepatitis C infection by constant mutation; however; HCV intends to escape such natural/acquired host immune barriers of viral detection/elimination and propagates/maintain persistent infection [10].

2. Hepatitis C virus genome heterogeneity

Hepatitis C viral genome varies 30–50% at genotype level and 15–30% among different subtypes [11]. However, this variation also exists within a specific genotype at nucleotide sequence level where a difference of 1–5% is reported in a single infected patient [12]. These nucleotide variants may be a possible cause of the origination of pre-existing or treatment emergent resistance-associated variants or substitutions (RAV or RAS) in treated subjects. The sequence variability is uniformly and equally distributed throughout the viral genome; however, not reported in highly conserved genome region (e.g., 5′UTR, 3′UTR, and core region) and some hyper variable (HVR) region in E2 protein [13]. HVR1 in E2 protein is also demonstrated a predisposing factor for persistent viral infection [14]. Geographical distribution of hepatitis C genotypes also varies where genotype 1 (subtype 1a/1b) is frequently prevailed in the United States and Western Europe, followed by genotype 2 and 3 infection [15]. However, the other genotypes are found in distinct regions, where genotype 3 is the most common in South Asia, genotype 4 in Central Africa (almost endemic in Egypt), genotype 5 in South Africa and genotype 6 in Southeast Asia [15].

2.1. HCV genotype testing

HCV genotype testing is very important to predict the overall treatment duration as well as the outcome of direct-acting antivirals in treated individuals. For this reason, it is performed at baseline to identify patients to initiate therapy and select appropriate regimens. In principle, the nucleotide variations to certain targeted genes (e.g., core, E1, NS5A, and NS5B) of viral genome as well as untranslated regions (e.g., 5′ UTR) are performed by sequencing reaction [12, 45]. An ideal approach to perform HCV genotyping includes polymerase chain reaction (PCR) amplification of targeted gene and sequencing, or PCR amplification and hybridization with genotype-specific probes, or real-time reverse transcription PCR (RT-PCR) approach. No food and drug administration (FDA) approved methods exist to determine HCV genotypes and various institutes and laboratories have developed their own specific protocols.

Some reference methods demonstrate the amplification and direct sequencing of NS5B or 5′UTR regions, their alignments and phylogenetic analysis. However, the methods are time consuming, expensive and require equipment/software usually used in research laboratories. Similarly, those are used to epidemiological studies where exact genotype is needed. HCV genotype testing by such methods is advantageous because it reveals genomic variability, and the presence of quasispecies during the natural progression of the disease and overall response to antiviral therapy. Commercially available kits are used to perform HCV genotyping in clinical practice which employ PCR amplification and hybridization with genotype-specific probes.

Currently and most widely used methods include reverse-hybridization line probe INNO-LiPA HCV II assay® (Innogenetics, Ghent, Belgium), simplified direct sequencing Trugene 5′NC HCV Genotyping assays® (Siemens AG), and the Abbott Real-Time HCV Genotype II Assay® (Abbott Laboratories). These assays are generally very reliable with high degree of concordance and the margin of incorrect typing is rare (i.e., <3%). However, the mixed genotypes are detected but uncommon and 5% specimens cannot be genotyped due to low viral load, PCR amplification threshold and very high genome sequence variations.

In principle, INNO-LiPA HCV II is a reverse hybridization line probe assay which uses specific oligonucleotide probes to capture 5′ UTR of hepatitis C genome. The current version of the assay (i.e., INNOLiPA version 2.0) is a next-generation line probe assay which detects 5′UTR and core region of viral genome while INNO-LiPA HCV version 2.0 Siemens AG® identifies HCV GT1 subtypes (1a, 1b, 1c etc.,) in clinical and commercial studies. The Trugene 5′C HCV Genotyping Kit (Siemens AG®) analyzes 5′ UTR and compare with the genomic libraries of HCV genotypes. The Abbott Genotype II Assay (Abbott Laboratories®) is based on Real-Time PCR method which quantifies viral mRNA and identify hepatitis C GT 1 (subtype 1a, 1b) GT 2 (subtype 2a, 2b) 3, 4, 5, and 6 [45].

2.2. Therapeutic outcomes of DAAs against various HCV genotypes

Since 2010, the treatment strategies for HCV infection have been revolutionized after the advent of interferon-free direct-acting antivirals (**Table 1** enlists the FDA approved and recommended interferon-free direct-acting antivirals for hepatitis C infected patients. The table also concisely demonstrates the patient category, recommended dose and treatment duration with special recommendations) [2, 16, 40]. Such therapeutic regimens achieve higher sustained virologic response rates in treated individuals along with favor tolerability, fewer side effects

Treatment regimens	**Dose (mg/day)**	**Treatment duration (weeks)**	**Treatment recommendations**
Daclatasvir [DCV] (Daklinza®)	60	12 8 (when HCV RNA level is <6 million IU/mL. 24 (for compensated or decompensated cirrhosis with or without RBV, liver transplant, HCV/HIV co-infection and no baseline NS5A mutations) 8 (for acute hepatitis C patients)	**a.** GT1 (Subtype 1a & 1b), GT2, GT3, GT4, and GT5/6 treatment naïve, treatment experienced, without or with compensated or decompensated cirrhosis patients. **b.** PEG-IFN/RBV, PEG-IFN/RBV plus SOF and DAAs experienced patients. **c.** NS3 PIs inhibitor + PEG-IFN/RBV experienced patients. **d.** GT1, GT2, GT3, GT4, GT5, and GT6 treatment naïve/experienced kidney or liver transplant recipients with or without compensated cirrhosis. **e.** Acute hepatitis C patients.
Sofosbuvir [SOF] (Sovaldi®)	400	12 24 (without RBV for compensated cirrhosis, liver transplant, HCV/HIV co-infection and no basal Q80K mutations) 8 (for acute hepatitis C patients)	**a.** GT1 (Subtype 1a & 1b), GT2, GT3, GT4, and GT5/6 treatment naïve, treatment experienced, without or with compensated cirrhosis patients. **b.** PEG-IFN/RBV treatment experienced patients. **c.** NS3 PIs inhibitor + PEG-IFN/RBV experienced patients. **d.** GT1, GT2, GT3, GT4, GT5, and GT6 treatment naïve/experienced kidney or liver transplant recipients with or without compensated cirrhosis. **e.** Acute hepatitis C patients.
Simeprevir [SMV] (Olysio®)	150	12 24 (without RBV for compensated cirrhosis or no basal Q80K mutations)	**a.** GT1 (Subtype 1a & 1b), and GT4 treatment naïve, treatment experienced, without or with compensated cirrhosis patients. **b.** PEG-IFN/RBV treatment experienced patients. **c.** No basal Q80K mutations.
Ledipasvir/sofosbuvir [LDV/SOF] (Harvoni®)	90/400	12 24 (with or without RBV for compensated or decompensated	**a.** GT1 (Subtype 1a & 1b), GT4, GT5/6 treatment naïve, treatment experienced, without or with

Treatment regimens	Dose (mg/ day)	Treatment duration (weeks)	Treatment recommendations
		cirrhosis, liver transplant, HCV/HIV co-infection, and SOF/NS5A-based treatment failed) 8 (for acute hepatitis C patients)	compensated cirrhosis and decompensated cirrhotic patients. **b.** SOF or NS5A- based treatment failure. **c.** PEG-IFN/RBV treatment experienced patients. **d.** NS3 PIs inhibitor + PEG-IFN/RBV experienced patients. **e.** GT1, GT4, GT5, and GT6 treatment naïve/experienced kidney or liver transplant recipients with or without compensated cirrhosis. **f.** Acute hepatitis C patients.
Dasabuvir, ombitasvir, paritaprevir, ritonavir [DSV/OMV/PTV/r] (Viekira Pak®)	500/25/ 150/100	12 24 (with weight-based RBV for compensated cirrhosis)	**a.** GT1 (Subtype 1a & 1b), and GT4 (without dasabuvir) treatment naïve, treatment experienced, without or with compensated cirrhotic patients. **b.** PEG-IFN/RBV treatment experienced patients. **c.** HCV along with chronic kidney disease patients.
Sofosbuvir/velpatasvir [SOF/VEL] (Epclusa®)	400/100	12 24 (without RBV for decompensated cirrhosis, liver transplant, HCV/HIV co-infection, and SOF/NS5A-based treatment failure) 8 (for acute hepatitis C patients)	**a.** GT1 (Subtype 1a & 1b), GT2, GT3, GT4, and GT5/6, treatment naïve, treatment experienced, without or with compensated cirrhosis and decompensated cirrhotic patients. **b.** SOF or NS5A- based treatment failure. **c.** PEG-IFN/RBV treatment experienced patients. **d.** NS3 PIs inhibitor + PEG-IFN/RBV experienced patients. **e.** GT1, GT2, GT3, GT4, GT5, and GT6 treatment naïve/experienced kidney or liver transplant recipients with or without compensated cirrhosis. **f.** Acute hepatitis C patients.
Elbasvir/grazoprevir [EBR/GZR] (Zepatier®)	50/100	12 16 (for baseline NS5A RASs for elbasvir)	**a.** GT1 (Subtype 1a & 1b), and GT4 treatment naïve, treatment experienced, without or with compensated cirrhosis patients. **b.** PEG-IFN/RBV treatment experienced patients. **c.** HCV along with chronic kidney disease patients. **d.** No baseline NS5A RAS for elbasvir.
Sofosbuvir/velpatasvir/ voxilaprevir [SOF/VEL/ VOX] (Vosevi®)	400/ 100/100	12	**a.** GT1 (Subtype 1a & 1b), GT2, GT3, GT4, and GT5/6 treatment naïve, treatment experienced, without or

Treatment regimens	Dose (mg/ day)	Treatment duration (weeks)	Treatment recommendations
			with compensated cirrhosis patients **b.** NS5A alone or SOF/NS5A-based treatment failure.
Glecaprevir/pibrentasvir [GLE + PIB] (Mavyret®)	300/120	12 8 (without cirrhosis) 16 (NS5A-based treatment failure without prior treatment of NS3 PIs inhibitor)	**a.** GT1 (Subtype 1a & 1b), GT2, GT3, GT4, and GT5/6 treatment naïve, treatment experienced, without or with compensated cirrhosis patients. NS5A alone or NS3-based treatment failure but not both.

[1] The data shown in the table was derived from phase III clinical trials of IFN-free DAA regimens approved and recommended by the US FDA for the treatment of hepatitis C. RBV = ribavirin, PEG-IFN = pegylated interferon, GT = genotype, RAS = resistance-associated substitutions, and PIs = protease inhibitors.

Table 1. Recommended therapeutic regimens for hepatitis C virus infection[1] [34].

and fewer drug-drug interactions [16]. However, there are certain challenges to meet while achieving the global goal of HCV eradication soon. In parallel to that high therapy costs, treatment access to poor countries, real-world clinical data, and the emergence of resistance-associated variants are big challenges to coup [2, 16].

Interestingly, the new DAA regimens attain higher SVR rates in all genotypes patients (i.e., genotype 1–6) but still the therapeutic efficacy varies at genotypes level as well as subtypes level and even in harder to treat specific populations (e.g., HCV GT1 subtype 1a, genotype 3 & 4 patients with compensated and decompensated cirrhosis, chronic kidney disease and severe liver-impairment patients and HCV/HIV coinfected patients) [16]. DAA regimens alone, in combination (e.g., Olysio®, Sovaldi®, Daklinza® with or without ribavirin) or as a fixed-dose combination (Harvoni®, Viekira Pak®, Epclusa®, Zepatier®, Vosevi®, Mavyret®) achieve higher SVR rates (>95%) in GT1, 2, 5 and 6 treated patients. However, the GT 3 patients exhibited SVR rates ≤90–95% as most of the clinical studies performed for the approval of DAA regimens [16]. Similarly, the viral relapse, virologic breakthrough and treatment discontinuation were prominent in cirrhotic patients.

It was also demonstrated that single or dual DAA regimens could not achieve higher SVR rates in HCV genotype 3 patients and addition of another DAAs (i.e., triple DAA regimens) is highly recommended to achieve higher SVR rates for this genotype. HCV genotype 4 patients with or without cirrhosis also achieved compromised SVR rates (≤85–95%) in clinical studies of approved regimens [16]. Due to this reason, the newly approved regimens are cautiously recommended in compensated or decompensated cirrhotic patients. These mechanisms or phenomena are involved for the variable therapeutic response of all oral DAAs to various HCV genotypes or subtypes are not fully elucidated. However, the remarkable viral genome heterogeneity, high viral load, disease progression and in particular the emergence of viral escape mutants are considered the predisposing factors in this prospect [16–18]. The incoming sections pragmatically overview the molecular kinetics of the emergence of RAS, their effect on treatment response and possible ways to prevent them.

3. The clinical dynamics of RAS for various HCV genotypes

The antiviral drug resistance is a commonly observed phenomenon in chronically infected HCV patients who are recommended to take telaprevir, boceprevir (first-generation NS3/4A protease inhibitors (PIs)) or simeprevir (second-generation PIs) as therapeutic regimens to treat the infection [19]. This problem may also arise in HCV-infected individuals during or after the treatment completion when administered to telaprevir, boceprevir or simeprevir as monotherapy or in combination with pegylated interferon (PEG-IFN) and ribavirin (RBV) [20]. Interestingly, it is rarely reported in those infected patients who are administered to asunaprevir, paritaprevir, grazoprevir (second-generation PIs) and non-nucleoside polymerase inhibitors (NNIs, e.g., dasabuvir) and nucleotide RNA polymerase inhibitors (e.g., sofosbuvir) [17, 18]. Similarly, the development and approval of next-wave interferon-free DAA regimens (e.g., ledipasvir, daclatasvir, ombitasvir, elbasvir, velpatasvir, voxilaprevir, glecaprevir, and pibrentasvir) for chronic hepatitis C and difficult to treat specific populations have shown promise in clinical trials while achieving higher SVR rates, improved adverse event profile, fewer drug-drug interactions and a strong barrier to antiviral drug resistance [18]. Nevertheless, the viral escape mutants are often emerged against one particular drug in interferon-free DAA combination regimens, although the frequency of emergence is lower (**Table 2**).

Numerous genetic variants or different HCV isolates (termed as quasispecies) are persistently produced in HCV-infected individuals due to the high mutation rate of the viral genome (10^{-5}–10^{-4} nucleotide per replication cycle) and poor fidelity of the virus replication enzyme (i.e., RNA-dependent RNA polymerase) during HCV replication [21, 22]. Some variants develop sophisticated mutations which may have the tendency to alter the conformation of the binding sites of NS3/4A serine protease, NS5A, and NS5B inhibitors in their targeted active sites and ultimately decrease their therapeutic efficacy [23, 24]. These pre-existing genome variants have a fitness advantage with specific antivirals and may become the dominant viral quasispecies during or after the treatment completion [23, 24]. HCV quasispecies mostly exhibit an attenuated replication and usually displaced by the wild-type HCV genome after stopping the exposure to direct-acting antivirals [23, 24].

At HCV genotypes level, the genotype 1 is the most studied GT regarding the DAAs resistance profile [18]. Genotype 1 infected patients are more prone to develop RAS during or after the treatment completion or exist with pre-existing RAS before the start of therapy [18]. At subtype levels, subtype 1a demonstrates the least genetic barrier to drug resistance than 1b [18]. Genotype 3 and to somehow genotype 4 are still harder to treat and SVR rates are not achieved very significantly in some specific populations (compensated cirrhotic or decompensated cirrhotic patients, treatment experienced patients with first-generation PIs, HCV/HIV co-infection, liver transplant, renal impairment and dialysis patients) as compared

DAAs	RAS^2 (alone or in combination)	RAS effect on treatment response^3
Telaprevir [31] (Incivek®)	R155, A156, V36, T54, D168, R155K/T, V36M, V36M + R155K, T54A/S, V36A/L, A156S/T, V36G/I, I132V, R155G/M, A156V/F/N or D168N	Telaprevir was discontinued by the US FDA after the advent and recommendation of new IFN-free DAA regimens for HCV-infected individuals; however, the treatment experienced patients with first-generation protease inhibitors (telaprevir, boceprevir) still express baseline and treatment emergent RAS and are treated with newer IFN free DAA regimens to achieve higher SVR12 rates.
Boceprevir [31] (Victrelis®)	For HCV subtype 1a: V36M, T54S, R155K, V36A, T54A, V55A, V55I, V107I, R155T, A156S, V158I, D168N, I/V 170T, I/V170F For HCV subtype 1b: T54A, T54S, V55A, A156S, I/V170A, V36A, V36M, T54C, T54G, V107I, R155K, A156T, A156V, V158I, I/V170T, M175L	Boceprevir was discontinued by the US FDA after the advent and reommendation of new IFN-free DAA regimens for HCV-infected individuals; however, the treatment experienced patients with first-generation protease inhibitors (telaprevir, boceprevir) still express baseline and treatment emergent RAS and are treated with newer IFN free DAA regimens to achieve higher SVR12 rates.
Daclatasvir [46] (Daklinza®) [ALLY-1] [ALLY-2] [ALLY-3] [ALLY-3+] [COMMAND-4] [HALLMARK-DUAL] [HALLMARK-QUAD] [UNITY-1] [UNITY-2]	Pre-existing or treatment-emergent substitutions: M28T, Q30H/K/R, L31M/V, H54R, H58D/P, Y93C/N, P32, A30K/S, L31i, S62A/L/P/R/T, Y93H, A112T, L159F, E237G, Q355H, S282T + Q355H	HCV genotype 1a patients with RAS M28, Q30, L31 or Y93: **a.** SVR 12 with NS5A polymorphism 76% (13/17) a.1 without cirrhosis 100%(11/11) a.2 with cirrhosis 33% (2/6) **b.** SVR 12 without polymorphism 95% (142/149) b.1 without cirrhosis 99% (100/101) b.2 with cirrhosis 88% (42/48) HCV genotype 3 patients with RAS Y93H: **a.** SVR 12 with NS5A polymorphism 54% (7/13) a.1 without cirrhosis 67% (6/9) a.2 with cirrhosis 25% (1/4) **b.** SVR 12 without NS5A polymorphism 92% (149/162) b.1 without cirrhosis 98% (125/128) b.2 with cirrhosis 71% (24/34)
Sofosbuvir [47] (Sovaldi®) [BOSON] [FISSION] [FUSION] [NEUTRINO] [PHOTON-1,2] [POSITRON] [VALENCE]	Treatment-emergent substitutions: S282T, L159F, V321A, L320F	The cutoff value was below 1% while detecting treatment-emergent RAS agaisnt sofosbuvir in different clinical trials, so not significant change in SVR12 of different treated groups were demonstrated.
Simeprevir [48] (Olysio®) [ATTAIN] [C212] [OPTIMIST-1] [OPTIMIST-2] [PROMISE] [QUEST-1] [QUEST-2] [RESTORE]	Treatment-emergent substitutions: F43, Q80, S122, R155, A156, D168, D168E, D168V, Q80R, R155K, Q80X + D168X, R155X + D168K, Q80K, S122A/G/I/T, S122R, R155Q, D168A, D168F, D168H, D168T, I170T	HCV genotype 1a patients with any RAS F43, Q80, S122, R155, A156, or D168 95% (110/116) D168E 15% (17/116) D168V 10% (12/116) Q80R 4% (5/116) R155K 77% (89/116) Q80X + D168X 4% (5/116) R155X +D168K 13% (15/116) Q80K, S122A/G/I/T, S122R, R155Q, D168A, D168F, <10%

DAAs	RAS[2] (alone or in combination)	RAS effect on treatment response[3]
		D168H, D168T, I170T HCV genotype 1b patients with any RAS F43, Q80, S122, R155, A156, or D168 86% (70/81) D168E 17% (14/81) D168V 60% (49/81) Q80R 12% (10/81) R155K 0% (0) Q80X+ D168X 14% (11/81) R155X+D168K 4% (3/81) Q80K, S122A/G/I/T, S122R, R155Q, D168A, D168F, <10% D168H, D168T, I170T
Ledipasvir/sofosbuvir [49] (Harvoni®) [ION-1] [ION-2] [ION-3] [ION-4]	NS5A RAS: K24R, M28T/V, Q30R/H/K/L, L31M, Y93H/N, Q30R, Y93H/N, L31M, L31V/M/I, H58D/P, Y93H/C NS5B RAS: L159, V321, D61G, A112T, E237G, S473, M289I, S282T, L32OV/I, V321I + L31M, Y93H, Q30L	Virologic relapse rate with or without baseline NS5A polymorphism: **a.** Treatment naive GT1 patients with baseline NS5A polymorphism = 6% at week 8 and 1% at week 12. **b.** Treatment naive GT1 patients without baseline NS5A polymorphism = 5% at week 8 and 1% at week 12. **c.** Treatment experienced GT1 patients with baseline NS5A polymorphism = 22% at week 12 and 0% at week 24. **d.** Decompensated cirrhotic GT1 patients with baseline NS5A polymorphism = 7% at week 12 and 5% without polymorphism. **e.** Limited data for GT 2, 3, 4, 5 or 6 patients with baseline NS5A polymorphism.
Dasabuvir, ombitasvir, paritaprevir, ritonavir [50] (Viekira Pak®) [PEARL-II] [PEARL-III] [PEARL-IV] [RUBY-I] [SAPPHIRE-I] [SAPPHIRE-II] [TURQUOISE-I] [TURQUOISE-II] [TURQUOISE-III]	Treatment-emergent substitutions: NS3 RAS: V36A/M/T, F43L, V55I, Y56H, Q80K/L, I132V, R155K, A156G, D168, P334S, S342P, E357K, V406A/I, T449I, P470S NS4A RAS: V23A NS5A RAS: K24R, M28A/T/V, Q30E/K/R, H/Q54Y, H58D/P/R, Y93C/H/N NS5B RAS: G307R, C316Y, M414I/T,E446K/Q, A450V, A553I/T/V, G554S, S556G/R, G558R, D559G/I/N/V, Y561H	**a.** Virologic failure with NS3 Q80K polymorphism = 38% **b.** Virologic failure with ombitasvir associated NS5A polymorphism = 22% **c.** Virologic failure with dasabuvir associated NS5B polymorphism = 05%
Sofosbuvir/velpatasvir [51] (Epclusa®) [ASTRAL-1] [ASTRAL-2] [ASTRAL-3] [ASTRAL-4] [ASTRAL-5] [POLARIS-2] [POLARIS-3] [POLARIS-4]	Treatment-emergent substitutions: NS5A RAS: Y93N, K24M/T, L31I/V, Q30R, L31M, H58P NS5B RAS : L314F/I/P	Overall viral relapse rates with pre-exisitng NS5A RAS in patients without cirrhosis/compensated cirrhosis and decompensated cirrhosis = 15% **1.** Viral relapse in GT1 patients with compensated cirrhosis = 1%. **2.** Viral relapse in GT3 patients with compensated cirrhosis = 33%. **3.** No viral relapse in GT2, 4, 5 and 6 compensated cirrhotic patients. **4.** Viral relapse in GT1 patients with decompensated cirrhosis = 2%. **5.** Viral relapse in GT3 patients with decompensated cirrhosis = 15%.

DAAs	RAS[2] (alone or in combination)	RAS effect on treatment response[3]
		6. No viral relapse in GT2, 4, 5 and 6 decompensated cirrhotic patietns.
Elbasvir/grazoprevir (Zepatier®) [41–43, 52] [C-EDGE CO-STAR] [C-EDGE co-infection] [C-EDGE treatment-experienced] [C-EDGE treatment-naive] [C-SURFER]	Treatment-emergent substitutions: NS5A RAS: M28A/G/T, Q30H/K/R/Y, L31F/M/V, H58D, Y93H/N/S, L28M, L28S/T, M31I/V, P58D NS3 RAS: V36L/M, Y56F/H, V107I, RI55I/K, A156G/T/V, V158A, D168A/G/N/V/Y, A156M/T/V, V170I	SVR12 rates in GT1a patients without baseline NS5A polymorphism; With 12 weeks treatment = 98% With 16 weeks treatment = 100% SVR12 rates in GT1a patients with baseline NS5A polymorphism; With 12 weeks treatment = 70% With 16 weeks treatment = 100% No impact on SVR12 in GT1a patients with NS3 Q80K polymorphism SVR12 rates in GT1b patients with baseline NS5A polymorphis = 94% SVR12 rates in GT1b patients without baseline NS5A polymorphis = 99% No impact on SVR12 in GT1b patients with NS3 Q80K polymorphism SVR12 rates in GT4 patients with baseline NS5A polymorphi = 100% SVR12 rates in GT4 patients without baseline NS5A polymorphis = 95% SVR12 rates in GT4 patients with baseline NS3 polymorphism = 100% SVR12 rates in GT4 patients without baseline NS3 polymorphism = 96%
Sofosbuvir/velpatasvir/voxilaprevir (Vosevi®) [53] [POLARIS-1] [POLARIS-2] [POLARIS-3] [POLARIS-4] [POLARIS-5]	NS5A RAS: Q30T, L31M, Y93H/N, V36A, E92K, A30K NS3 RAS: Q41K, V55A, R155M, M28T NS5B RAS: S282T	Overall SVR12 rates in patients with or without baseline NS3 and NS5A polymorphism = 97% Overall SVR12 rates in patients with NS5B polymorphism = 95%
Glecaprevir/pibrentasvir (Mavyret®) [54] [ENDURANCE-1] [ENDURANCE-2] [ENDURANCE-3] [ENDURANCE-4] [EXPEDITION-1] [EXPEDITION-2] [EXPEDITION-4] [MAGELLAN-1 (Part-2) [SURVEYOR-II] (Part-3) [SURVEYOR-II] (Part-4)	Treatment-emergent substitutions: NS3 RAS: Y56H/N, Q80K/R, A156G, Q168L/R, A166S NS5A RAS: M28A/G, A30G/K, L31F, P58T, Y93H, L31M, Q30K/R, H58D, P29Q/R	Baseline NS3 and NS5A polymorphism in GT1, 2, 4, 5 and 6 patients had no impact on treatment resaponse Overall SVR rates in G3 patients without cirrhosis but with NS5A A30K polymorphism = 78% Overall SVR rates in GT3 patients with baseline NS5A Y93H polymorphism = 100%

[2]The data for resistance associated substitutions mentioned against interferon free regimens in table 2 was derived from phase III clinical trials and clinical trials registered to ClinicalTrials.gov.

[3]The treatment outcomes data for interferon-free regimens in RAS detected, pre-existing or treatment-experienced RAS was retrieved from phase III clinical trials and clinical trials registered to ClinicalTrials.gov.

Table 2. Resistance-associated substitutions associated with interferon-free DAAs regimens and their overall impact on treatment response.

to other viral genotypes (**Table 2**) [16, 17]. Interestingly, the RAS associated with GT 3 and 4 patients are not responsible for the failure to achieve higher SVRs in specific populations as the clinical studies demonstrated [17]. However, the limited number of patients in those clinical trials and possible biasness are some major limitations of these studies, which further demands to extensively elucidate in large patient populations [17].

Many studies demonstrate that such variants/substitutions reduce the chances to achieve higher SVR rates as well as are a potential cause of viral relapse, virologic breakthrough and treatment discontinuation in treated individuals [17, 18]. Although the ratio of viral escape mutants (also known as resistance-associated variant (RAV) and RAS)) to emerge is low with the administration of second generation (e.g., simeprevir and sofosbuvir) and next-wave direct-acting antivirals (e.g., daclatasvir, ledipasvir, dasabuvir, ombitasvir, paritaprevir, elbasvir, grazoprevir, velpatasvir, glecaprevir, pibrentasvir, and voxilaprevir) in clinical studies; however, their impact on treatment response is still significant [17, 18]. Similarly, the treatment experienced patients with first-generation DAAs (i.e., telaprevir and boceprevir) having no therapeutic response or with virologic relapse and viral breakthrough exist in real-world clinical settings and when treated with the second and next-wave DAAs still express the pre-existing or treatment emergent RAVs [17, 18]. NS3/4A, NS5A and NS5B baseline polymorphism and pre-existing and treatment-emergent RAS are also big hurdle even the patients are administered with next-wave direct-acting antivirals [17, 18]. The phase III clinical studies explicit the emergence of these RAS with variable SVRs in different genotype treated individuals, although the data are limited (**Table 2** concisely overviewed the baseline and treatment-emergent RAS and their impact on therapeutic outcome of FDA approved anti-hepatitis C regimens in different HCV genotypes) [17, 18].

One good example is the resistance variants of NS5A protein which can pre-exist in the viral quasispecies population (both in treatment-naïve and treatment-experienced patients) as well as emerge during or after treatment completion (i.e., treatment-emergent RAS) [18]. Similarly, the detection of resistance variants with currently available laboratory techniques is difficult as the viral variants usually replicate at low levels; however; the next-generation sequencing (NGS) techniques make it feasible to do at a certain cutoff level [18]. HCV quasispecies can be detected at low levels in approximately 1% patients, which are resistant to protease or non-nucleoside polymerase inhibitors (NNIs) and have never been treated with these specific antivirals before [18]. For this reason, such therapeutic regimens are administered cautiously in patients who are previously resistant (i.e., patients treated with PEG-IFN/RBV and dual therapies based on PEG-IFN/RBV plus first-generation PIs, first-generation NS5A and NS5B inhibitor resistant which could not achieve SVR rates after treatment completion and patients with virologic relapse, virologic breakthrough and treatment discontinuation) and detected with viral escape mutants [17, 18]. First-generation NS5A inhibitors (i.e., daclatasvir and ledipasvir) have low genetic barrier to resistance while the next-wave NS5A-targeting molecules (e.g., elbasvir, grazoprevir) are potent inhibitors with pan-genotypic drug efficacy against HCV genotypes 1 to 6 and various subtypes [17, 18].

4. Viral resistance substitutions against first-generation direct-acting antivirals

The patients who take telaprevir or boceprevir as monotherapy may develop antiviral resistance within a few days during treatment [20, 23]. The minor resistant populations against these drugs exist at baseline in all HCV-infected individuals and are selected rapidly with telaprevir or boceprevir monotherapy [20]. Similarly, notable drug-drug interactions with many human immunodeficiency virus (HIV) antiretrovirals and calcineurin inhibitors also decrease the therapeutic activity of telaprevir and boceprevir monotherapy (due to severe drug adverse events, numerous possible drug-drug interactions and rapid emergence of RAVs, the first-generation PIs have been discontinued by the FDA to treat hepatitis C patients in the US and other parts of the world) [20, 31]. R155 is the most overlapping position in NS3/4A serine protease (a protein involved in HCV translation and also potential drug active site for the design and development of protease inhibitors), where different mutations may produce and confer resistance to nearly all protease inhibitors, (An exception is MK-5172) [25–29]. *In vivo* mutations at four positions (V36A/M/L, T54A, R155K/M/S/T, and A156S/T) and only one *in vitro* (i.e., replicon system) mutation (A156) has detected and characterized against telaprevir [30]. These mutations either alone (V36A/M, T54A, R155K/T, A156S) or as double mutations (A156T/V, V36M + R155K, V36M + 156T) confer low to high resistance barrier against telaprevir by altering the catalytic active sites of NS3/4A serine protease [30]. The pattern of resistance against telaprevir also differs significantly among HCV subtypes. The clinical studies reveal that antiviral resistance occurs much more frequently in HCV genotype 1a infected patients as compared to genotype 1b either using telaprevir alone or in combination with PEG-IFN α plus RBV [31]. It is due to a single nucleotide polymorphism at position R155K in NS3/4A serine protease, where codon AGA encodes R in HCV subtype 1a versus 1b (where codon CGA also encodes R) [30, 31]. In HCV genotype 1a isolates, only a single nucleotide substitution is required to change R to K at position 155 while 2 nucleotide changes require in subtype 1b [50]. Some studies also demonstrate that subtype 1a display higher fitness advantage than genotype 1b isolates, which is a predisposing factor in developing viral escape mutants and viral breakthroughs to other positions within NS3/4A catalytic subunit and other genomic regions of 1a isolates [30, 31].

5. RAS against second-generation direct-acting antivirals

Q80R/K polymorphism is responsible for low-level resistance to a macrocyclic protease inhibitor, simeprevir. The clinical studies predict Q80K variants up to 50% in HCV genotype 1a-infected patients (which is approximately 20% in Europe and 50% in the United States) and almost 1% of 1b isolates [32]. Lower SVR rates and a slow viral decline have reported in HCV genotype 1a patients treated with simeprevir-based triple therapy in phase III clinical studies (20% lower in HCV genotype 1a than 1b) [33]. Q80K polymorphism and NS3 genotype testing prior to therapy is highly recommended for HCV subtype 1a patients to avoid any adverse events, low virologic response and treatment discontinuation during therapy [32, 33].

The viral variants associated with NS3 PIs may detect by first synthesizing cDNA by reverse transcription reaction, followed by performing polymerase chain reaction (PCR) and then sequencing reaction [32, 33]. Q80K polymorphism and viral variants testing against NS3 PIs have been launched in the USA by Quest Diagnostics® and LabCorp® [32, 33].

6. RAS against next-wave interferon-free DAAs

The first-generation NS5A inhibitors (e.g., daclatasvir) lead initially to higher SVR rates in treated patients, but the emergence of viral resistance occurs rapidly indicating its relatively lower genetic barrier to resistance [35]. The viral resistant mutants were found very commonly at amino acid residue Q30E and Y93N of NS5A protein in subtype 1a patients and confer the highest level of drug resistance [18]. Some studies demonstrate that these mutations are responsible for increasing the EC50 (i.e., the concentration of a drug which produces therapeutic response halfway between the baseline and maximum after a certain period of time) of daclatasvir in treated patients [18]. Similarly, L31 and Y93 substitution positions express the greatest aptitude for resistance to daclatasvir, where double mutations sometime increase the EC50 of DCV to a far greater degree. However, viral resistance substitutions were reported less frequently at position L31 and Y93 in HCV subtype 1b patients [18, 44]. From the clinical point of view, these substitutions against DCV are also considered to be responsible for resistance to other NS5A inhibitors as discussed below.

Ledipasvir in combination with sofosbuvir, as a fixed-dose combination was approved for GT1 patients with or without cirrhosis [18]. The fixed-dose combination also demonstrates pangenotypic clinical efficacy in patients with GT3, 4, 5, and 6 patients. The approval was based on the achievement of SVR rates ≥95% in GT 1 treatment-naive and treatment-experienced patients without cirrhosis. SVR rates were achieved >78% in decompensated cirrhotic patients awaiting liver transplant while 100% SVR rates were demonstrated for liver transplant recipients with fixed-dose LDV/SOF plus RBV. The addition of RBV did not significantly impact SVR rates in patients without cirrhosis; however, the addition is mandatory for GT1 and 4 decompensated cirrhotic patients as well as liver transplant recipients for 24-weeks. In phase III clinical trials, Q30R, Y93H/N, and L31M were the most commonly detected RAS in subtype 1a treatment failure patients while only one mutation Y93H was detected for 1b. However, the impact of these baseline RAVs was very limited on the overall therapeutic outcome of the regimens. Similarly, LDV shows strong therapeutic activity against SOF- induced mutants (e.g., S282T) as no drug cross-resistance between these two drugs were reported in clinical studies and vice versa. Another advantage of this fixed-dose combination (FDC) is to confer antiviral activity against RAVs associated with other NS5B NNIs and NS3 PIs [18].

Ombitasvir (OMV) another NS5A inhibitor was approved in combination with paritaprevir (PTV), r (ritonavir) and dasabuvir (DSV) for the treatment of difficult to treat GT1 specific populations as achieved higher SVR rates (~100%) in treatment naive (1a subtype) and treatment-experienced patients (IFN-based 1b subtype) [18]. Similarly, OMV plus PTV/r without DSV were recommended to treat GT4 chronic hepatitis C (CHC) patients as DSV clinically ineffective against GT4 patients. However, the drug combination is strictly prohibited to administer in patients with decompensated or moderate to severe hepatic impairment. Despite being the multiprotein targeting regimens with the chances to develop mutations, the pooled analysis showed high genetic barrier to drug resistance. Both pre-existing and treatment based RAVs were reported in virologic failure experienced patients; however, interestingly baseline RAVs did not impact the overall efficacy of treatment. OMV monotherpay for 12 weeks in treatment-naive GT1 patients also generated variants in both subtypes but without any baseline RAVs. The most surviving variants in GT1 subtype 1a patients were reported at amino acids positions M28, Q30, and Y93; however, only one substitution Y93H was noticed in GT1 subtype 1b patients although with 77-fold more drug resistance. Due to this reason, OMV is always recommended in combination with PTV/r/DSV or PTV/r [18].

A fixed-dose combination (FDC) of elbasvir/grazoprevir (Zepatier®) (50 mg/100 mg) one a day has been approved by the United States Food and Drug Administration (US FDA) for the treatment of HCV GT 1 & 4 infected patients with chronic kidney diseases and HCV/HIV-co-infection. However, the treatment is recommended with cautions in some specific populations including viral subtype 1a, prior treatment experienced with NS3 PIs, and NS5A associated RASs at position M28, Q30, L31, or Y93) [18]. Similarly, the treatment regimen is prescribed with many precautions in subtype 1a patient with prior testing of NS5A associated RAVs, because it determines the overall treatment duration and the inclusion of ribavirin to therapy [18]. This regimen achieved higher SVR rates in all patient arms (~97%) in particular previous non-responders to IFN-based therapies as well as in individuals with severe renal impairment (94%). The low therapeutic outcomes were revealed for GT1 subtype 1a patients with substitutions at positions M28T, Q30, L31, or Y93 after 12-weeks drug administration in clinical studies. Another interesting fact also revealed that those mutations against elbasvir also decrease the therapeutic efficacy of other NS5A inhibitors. However, elbasvir was found fully active against the mutations generated by grazoprevir (NS3/4A PIs) while used in combination. Moreover, the mutations existing against SOF-based therapeutic regimens are harmless to elbasvir [18].

RAVs associated with NS5A inhibitors do not impair replication fitness during the treatment as compared to viral resistant mutants of NS3 PIs and consequently do not disappear during follow-up examinations at the end of therapy [18, 35]. Viral resistance mutants against NS5A inhibitors persist even after 1 year follow-up studies in treated individuals but interestingly no cross-resistance has been reported between DCV and other DAAs as yet [18]. For this reason, the prior testing of NS5A variants in such patients before the treatment initiation is essential to determine overall treatment duration and inclusion of RBV in therapy.

7. Clinical significance of viral escape mutants

The clinical importance of RAVs is still not clear, but some studies have revealed that these mutations are commonly shared between first and second-generation direct-acting antivirals and to less extent for next-wave DAAs [36, 37]. Similarly, the clinical relevance of the viral escape mutants is also not completely understood. However, numerous studies demonstrate that these pre-existing variants may reduce the chances to achieve higher SVR rates with DAA-based triple therapies if the patients are individually less sensitive to PEG-IFN α plus RBV treatment [38, 39]. Due to this overlapping resistance profile, one protease inhibitor cannot be substituted for the other, and even a combination of two protease inhibitors does not make sense to be used in the cases of viral breakthroughs and treatment relapse in infected patients [37]. As a result, if an HCV-infected patient fails to response one PI, the retreatment with other direct-acting antivirals may seem very difficult [38, 39]. PEG-IFNα and RBV are considered an integral part of telaprevir- or boceprevir-based triple therapies, as some studies suggest that RAVs are not associated with less sensitivity to interferon and ribavirin-based combination therapies [40]. Interestingly, if the patient response is weak toward PEG-IFNα/RBV therapeutic regimen, the risks to develop viral resistant mutants are significantly higher [40]. HCV genome sequencing to determine the sequences of RAVs before or during therapy have no rational because it has no practical consequences. The exception is testing for Q80K variants in HCV genotype 1a patients which are recommended before simeprevir administration in the US prescribing information [32, 33]. It is uncertain that the test is cost effective in other parts of the world where genotype 1a is not highly prevalent, and Q80K polymorphism is rear. In QUEST-1 clinical trials, 41% HCV genotype 1a patients had this particular variant and their SVR rates were not significantly increased as compared to placebo when treated with simeprevir [32, 33]. However, the SVR rates were almost similar to HCV genotype 1b patients without Q80K variants in HCV genotype 1a patients [32, 33]. Interestingly, if Q80K variants detect at baseline, even then the chances to achieve optimal SVR rates will be higher provided that simeprevir is a part of the therapeutic regimen [32, 33]. In this scenario, a combination of next-wave DAAs (i.e., sofosbuvir and daclatasvir) with a very high resistance barrier and weak antiviral (e.g., ribavirin) activity may lead to high SVR rates. However, such drugs cannot be combined with first-generation DAAs (telaprevir or boceprevir) due to lack of clinical data and potential drug-drug interactions via the Pgp transporter proteins [18]. If viral escape mutants emerge during or after therapy in treated patients, for how long will they persist and which type of adverse effects would produce is not clearly understood. Some studies have reported that viral escape mutants revert to wild type within 1–2 years after the completion of treatment with first-generation PIs; however, RAS associated with NS5A inhibitors may persist for long time even after the treatment completion [18, 40]. NS5A baseline polymorphism and NS5A RAS detection by cloning sequencing is strongly recommended before the start of treatment in patients with persistent NS5A variants.

8. Prevention to viral escape mutants

The emergence of viral escape mutants against direct-acting antivirals has an adverse impact on treatment failure, when retreated with the same or other DAA-based combination therapies (**Table 2**) [18]. Phase, III follow-up studies of teleprevir and boceprevir-based triple therapies, revealed this fact where a rapid decline of viral escape mutants was detected (below the limit of detection, i.e., >20% of quasispecies) by using population sequencing techniques [31, 40]. However, these resistance mutants were detectable after several years in a single patient treated with telaprevir or boceprevir by using cloning sequencing techniques within smaller phase 1b studies [31]. Similarly, one study related to the retreatment of 5 HCV-infected patients with simeprevir-based triple therapy (who developed early simeprevir resistance during monotherapy and demonstrated SVR rates in only 3 out of 5 patients), also indicated a possible effect of low-level persistence of viral escape mutants [31, 32].

Adherence to the dose of medication (especially for PIs) and compliance with futility rules are two significant ways which may adopt during therapy to avoid viral escape mutants [17, 18]. Similarly, it may be managed by alternative treatment strategies and by improving the pharmacokinetics profile of the newly developed direct-acting antivirals. Currently, the approvals and recommendations of next-wave all oral interferon-free regimens have shifted the treatment paradigms for difficult to treat "specific" populations including the patients found resistant to first- and second-generation PIs and first-generation NS5A and NS5B inhibitors (**Table 1**) [34, 41]. Interferon free combination regimens where, one drug with higher therapeutic activity but lower genetic barrier to drug resistance and other with strong barrier to drug resistance but with lower therapeutic activity may reduce the chances of viral relapse and viral breakthroughs in treated individuals [17, 18, 34]. Furthermore, some non-nucleoside analog inhibitors with low antiviral efficacy but the high barrier to drug resistance are also in investigational trials to be a valuable part of oral interferon-free regimens to treat patients who are previously resistant to first- and second-generation DDA-based triple therapies [17, 18].

The clinical data improvising the failure of IFN free DAAs in treated individuals is still limited from the phase III clinical trials of the regimens and retreatment statistics are not sufficient to accomplish standard recommendations [17]. However, some currently available retreatment data for treatment-failure regimens is briefly mentioned here. For NIs-based (e.g., sofosbuvir) retreatment patient, 24 weeks treatment with addition of RBV is recommended, unless contraindicated. Sofosbuvir based triple or quadruple therapeutic regimens for 12 or 24 weeks along with RBV are also considerable if applicable. Similarly, for treatment failure of SOF and SMV, preferable retreatment options include a combination of LDV/SOF or SOF/DCV for 24 weeks in cirrhotic patients and along with RBV for 12 weeks. For SOF plus RBV failure, the retreatment strategies include SOF-based triple regimens including PEG-IFN and RBV for 12 weeks or alone RBV for 24 weeks. For SOF/LDV failure with NS5B variants, retreatment with PEG-IFN/RBV plus SOF was recommended for 12 weeks. Some retreatment strategies have been reported from real-world clinical practice studies, where the treatment failure of DCV-based regimens was retreated with SOF plus SMV and with or without RBV for 12 weeks. Despite achieving higher SVR rates, the retreatment strategies are still deficient in scientific evidences to support their recommendations [17].

9. Conclusions

The pre-existing or treatment-emergent resistance-associated variants in hepatitis C-treated patients decrease the overall cure rates (i.e., higher SVR rates) of direct-acting antivirals and other anti-hepatitis C regimens. These variants may cause viral relapse, viral breakthrough and treatment failure during or after the completion of therapy. The clinical impact of resistance-associated variants/substitutions is significant on the overall treatment outcome as the clinical studies predict variable SVR rates in different HCV genotype patients. The detection of resistance-associated variants is of utmost importance prior to initiation of therapy, to decide treatment duration as well as to choose retreatment or alternate treatment plan for previously treatment failure patients with first- and second-generation DAAs or to some extent new-wave DAAs. The discovery and development of interferon free combination regimens with pan-genotypic drug efficacy provide optimism to treat such difficult-to-treat populations where one drug with high antiviral efficacy and other one with strong barrier to drug resistance achieves highly significant sustained virologic response rates in treated individuals.

Acknowledgements

The authors would like to thank the Institute of Scientific Research and Revival of Islamic Heritage (ISR) at Umm Al-Qura University (project ID: 43310017) for the financial support.

Conflict of interest

None.

Author details

Imran Shahid[1,2]*, Munjed Mahmoud Ibrahim[3], Muhammad Usman Nawaz[4],
Mohammad Tarque Imam[5] and Waleed H. AlMalki[1]

*Address all correspondence to: iyshahid@uqu.edu.sa

1 Department of Pharmacology and Toxicology, College of Pharmacy, Umm-Al-Qura University, Makkah, Saudi Arabia

2 Applied and Functional Genomics Laboratory, Center of Excellence in Molecular Biology (CEMB), University of the Punjab, Lahore, Pakistan

3 Pharmaceutical Chemistry Department, College of Pharmacy, Umm Al-Qura University, Makkah, Saudi Arabia

4 Department of Pharmacy, King Abdulaziz Medical City, National Guard Health Affairs, Jeddah, Saudi Arabia

5 Department of Clinical Pharmacy, College of Pharmacy, Umm-Al-Qura University, Makkah, Saudi Arabia

References

[1] Smith B, Bukh J, Kuiken C, Muerhoff S, Rice M, Stapleton T, Simmonds P. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: Updated criteria and genotype assignment web resource. Hepatology. 2014;**59**:318-327. DOI: 10.1002/hep.26744

[2] Shahid I, AL WH, Hafeez H, Hassan S. Hepatitis C virus infection treatment: An era of game changer direct acting antivirals and novel treatment strategies. Critical Reviews in Microbiology. 2014;**42**:535-547. DOI: 10.3109/1040841X.2014.970123

[3] Choo L, Kuo G, Weiner J, Overby R, Bradley W, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989;**244**:359-362

[4] Delisse M, Descurieux M, Rutgers T, D'Hondt E, De W, Arima T, Barrera-Sala M, Ercilla G, Ruelle L, Cabezon T. Sequence analysis of the putative structural genes of hepatitis C virus from Japanese and European origin. Journal of Hepatology. 1991;**13**:S20-S23

[5] Li S, Tong P, Vitvitski L, Lepot D, Trepo C. Evidence of two major genotypes of hepatitis C virus in France and close relatedness of the predominant one with the prototype virus. Journal of Hepatology. 1991;**13**:S33-S37

[6] Chen J, Lin H, Tai F, Liu PC, Lin J, Chen S. The Taiwanese hepatitis C virus genome: Sequence determination and mapping the 5′ termini of viral genomic and antigenomic RNA. Virology. 1992;**188**:102-113

[7] Neumann U, Lam P, Dahari H, Gretch R, Wiley E, Layden J, Perelson S. Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. Science. 1998;**282**: 103-107

[8] Lohmann V, Roos A, Korner F, Koch O, Bartenschlager R. Biochemical and structural analysis of the NS5B RNA-dependent RNA polymerase of the hepatitis C virus. Journal of Viral Hepatitis. 2000;**7**:167-174

[9] Gomez J, Martell M, Quer J, Cabot B, Esteban JI. Hepatitis C viral quasispecies. Journal of Viral Hepatitis. 1999;**6**:3-16

[10] Gretch R, Polyak SJ, Wilson J, Carithers L Jr, Perkins D, Corey L. Tracking hepatitis C virus quasispecies major and minor variants in symptomatic and asymptomatic liver transplant recipients. Journal of Virology. 1996;**70**:7622-7631

[11] Ray B, Meyer K, Steele R, Ray R. Transcriptional repression of p53 promoter by hepatitis C virus core protein. The Journal of Biological Chemistry. 1997;**272**:10983-10986

[12] Sheridan I, Pybus G, Holmes C, Klenerman P. High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. Journal of Virology. 2004;**78**:3447-3454

[13] Simmonds P. Genetic diversity and evolution of hepatitis C virus–15 years on. The Journal of General Virology. 2004;**85**:3173-3188. DOI: 10.1099/vir.0.80401-0

[14] Lindenbach D, Evans J, Syder J, Wolk B, Tellinghuisen L, Liu C, Maruyama T, Hynes O, Burton R, McKeating A, Rice M. Complete replication of hepatitis C virus in cell culture. Science. 2005;**309**:623-626. DOI: 10.1126/science.1114016

[15] Bukh J, Miller H, Purcell H. Genetic heterogeneity of hepatitis C virus: Quasispecies and genotypes. Seminars in Liver Disease. 1995;**15**:41-63. DOI: 10.1055/s-2007-1007262

[16] Shahid I, AlMalki WH, Hassan S, Hafeez H. Real-world challenges for hepatitis C virus medications: A critical overview. Critical Reviews in Microbiology. 2018 Mar;**44**(2):143-160. DOI: 10.1080/1040841X.2017.1329277

[17] Buti M, Esteban R. Management of direct antiviral agent failures. Clinical and Molecular Hepatology. 2016 Dec;**22**(4):432-438. DOI: 10.3350/cmh.2016.0107

[18] Gitto S, Gamal N, Andreone P. NS5A inhibitors for the treatment of hepatitis C infection. Journal of Viral Hepatitis. 2017 Mar;**24**(3):180-186. DOI: 10.1111/jvh.12657

[19] Sarrazin C, Kieffer TL, Bartels D, Hanzelka B, Muh U, Welker M, et al. Dynamic hepatitis C virus genotypic and phenotypic changes in patients treated with the protease inhibitor telaprevir. Gastroenterology. 2007;**132**:1767-1777. DOI: 10.1053/j.gastro.2007.02.037

[20] Jacobson IM, Pawlotsky JM, Afdhal NH, Dusheiko GM, Forns X, Jensen DM, et al. A practical guide for the use of boceprevir and telaprevir for the treatment of hepatitis C. Journal of Viral Hepatitis. 2012;**E19**(Suppl 2):1-26. DOI: 10.1111/j.1365-2893.2012.01590.x

[21] Wohnsland A, Hofmann WP, Sarrazin C. Viral determinants of resistance to treatment in patients with hepatitis C. Clinical Microbiology Reviews. 2007;**20**:23-38. DOI: 10.1128/CMR.00010-06

[22] Moradpour D, Penin F, Rice CM. Replication of hepatitis C virus. Nature Reviews. Microbiology. 2007;**5**:453-463. DOI: 10.1038/nrmicro1645

[23] Dvory-Sobol H, Wong KA, Ku KS, Bae A, Lawitz EJ, Pang PS, et al. Characterization of resistance to the protease inhibitor GS-9451 in hepatitis C virus-infected patients. Antimicrobial Agents and Chemotherapy. 2012;**56**:5289-5295. DOI: 10.1128/AAC.00780-12

[24] Kuntzen T, Timm J, Berical A, Lennon N, Berlin AM, Young SK, et al. Naturally occurring dominant resistance mutations to hepatitis C virus protease and polymerase inhibitors in treatment-naive patients. Hepatology. 2008;**48**:1769-1778. DOI: 10.1002/hep.22549

[25] Camma C, Petta S, Cabibbo G, Ruggeri M, Enea M, Bruno R, et al. Cost-effectiveness of boceprevir or telaprevir for previously treated patients with genotype 1 chronic hepatitis C. Journal of Hepatology. 2013;**59**:658-666. DOI: 10.1016/j.jhep.2013.05.019

[26] Biswal BK, Cherney MM, Wang M, Chan L, Yannopoulos CG, Bilimoria D, et al. Crystal structures of the RNA-dependent RNA polymerase genotype 2a of hepatitis C virus reveal two conformations and suggest mechanisms of inhibition by non-nucleoside inhibitors. The Journal of Biological Chemistry. 2005;**280**:18202-18210. DOI: 10.1074/jbc.M413410200

[27] Lesburg CA, Cable MB, Ferrari E, Hong Z, Mannarino AF, Weber PC. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. Nature Structural Biology. 1999;**6**:937-943. DOI: 10.1038/13305

[28] Sarrazin C, Zeuzem S. Resistance to direct antiviral agents in patients with hepatitis C virus infection. Gastroenterology. 2010;**138**:447-462. DOI: 10.1053/j.gastro.2009.11.055

[29] Romano KP, Ali A, Aydin C, Soumana D, Ozen A, Deveau LM, et al. The molecular basis of drug resistance against hepatitis C virus NS3/4A protease inhibitors. PLoS Pathogens. 2012;**8**:e1002832. DOI: 10.1371/journal.ppat.1002832

[30] Welsch C, Domingues FS, Susser S, Antes I, Hartmann C, Mayr G, et al. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of the hepatitis C virus. Genome Biology. 2008;**9**:R16. DOI: 10.1186/gb-2008-9-1-r16

[31] Susser S, Vermehren J, Forestier N, Welker MW, Grigorian N, Fuller C, et al. Analysis of long-term persistence of resistance mutations within the hepatitis C virus NS3 protease after treatment with telaprevir or boceprevir. Journal of Clinical Virology. 2011;**52**:321-327. DOI: 10.1016/j.jcv.2011.08.015

[32] Lenz O, de Bruijne J, Vijgen L, Verbinnen T, Weegink C, Van Marck H, et al. Efficacy of re-treatment with TMC435 as combination therapy in hepatitis C virus-infected patients following TMC435 monotherapy. Gastroenterology. 2012;**143**:1176-1178. e1171-1176. DOI: 10.1053/j.gastro.2012.07.117

[33] Manns M, Marcellin P, Poordad F, de Araujo ES, Buti M, Horsmans Y, et al. Simeprevir with pegylated interferon alfa 2a or 2b plus ribavirin in treatment-naive patients with chronic hepatitis C virus genotype 1 infection (QUEST-2): A randomised, double-blind, placebo-controlled phase 3 trial. Lancet. 2014;**384**:414-426. DOI: 10.1016/S0140-6736(14) 60538-9

[34] Kaymakoğlu S, Köksal İ, Tabak F, Akarca US, Akbulut A, Akyüz F, Bodur H, Çağatay A, Dinçer D, Esen Ş, Güner R, Gürel S, Köse Ş, Şentürk Ö, Şimşek H, Yamazhan T, Yılmaz Y, Idilman R, Guidelines Study Group VH. Recommendation for treatment of hepatitis C virus infection. The Turkish Journal of Gastroenterology. 2017 Dec;**28**(Suppl 2):94-100. DOI: 10.5152/tjg.2017.22

[35] Fridell RA, Wang C, Sun JH, O'Boyle DR, Nower P, Valera L, et al. Genotypic and phenotypic analysis of variants resistant to hepatitis C virus nonstructural protein 5A replication complex inhibitor BMS-790052 in humans: In vitro and in vivo correlations. Hepatology. 2011;**54**:1924-1935. DOI: 10.1002/hep.24594

[36] Lok AS, Gardiner DF, Lawitz E, Martorell C, Everson GT, et al. Preliminary study of two antiviral agents for hepatitis C genotype 1. The New England Journal of Medicine. 2012; **366**:216-224. DOI: 10.1056/NEJMoa1104430

[37] Lin C, Gates CA, Rao BG, Brennan DL, Fulghum JR, Luong YP, et al. In vitro studies of cross-resistance mutations against two hepatitis C virus serine protease inhibitors, VX-950

and BILN 2061. The Journal of Biological Chemistry. 2005;**280**:36784-36791. DOI: 10.1074/jbc.M506462200

[38] McCown MF, Rajyaguru S, Kular S, Cammack N, Najera I. GT-1a or GT-1b subtype-specific resistance profiles for hepatitis C virus inhibitors telaprevir and HCV-796. Antimicrobial Agents and Chemotherapy. 2009;**53**:2129-2132. DOI: 10.1128/AAC.01598-08

[39] Akuta N, Suzuki F, Hirakawa M, Kawamura Y, Yatsuji H, Sezaki H, et al. Amino acid substitution in hepatitis C virus core region and genetic variation near the interleukin 28B gene predict viral response to telaprevir with peginterferon and ribavirin. Hepatology. 2010;**52**:421-429. DOI: 10.1002/hep.23690

[40] Barnard R, Zeuzam S, Vierling J, Sulkowski M, Manns M, Long J. Analysis of resistance associated amino acid variants in non-SVR patients enrolled in a retrospective long-term follow-up analysis of boceprevir phase 3 clinical trials. Hepatology. 2011;**54**:164

[41] Rockstroh JK, Nelson M, Katlama C, Lalezari J, Mallolas J, Bloch M, et al. Efficacy and safety of grazoprevir (MK-5172) and elbasvir (MK-8742) in patients with hepatitis C virus and HIV co-infection (C-EDGE CO-INFECTION): A non-randomised, open-label trial. The Lancet HIV. 2015;**2**(8):e319-e327. DOI: 10.1016/S2352-3018(15)00114-9

[42] Roth D, Nelson DR, Bruchfeld A, Liapakis A, Silva M, Monsour H Jr, et al. Grazoprevir plus elbasvir in treatment-naive and treatment-experienced patients with hepatitis C virus genotype 1 infection and stage 4-5 chronic kidney disease (the C-SURFER study): A combination phase 3 study. Lancet. 2015;**386**(10003):1537-1545. DOI: 10.1016/S0140-6736(15)00349-9

[43] Zeuzem S, Ghalib R, Reddy KR, Pockros PJ, Ben Ari Z, Zhao Y, et al. Grazoprevir-elbasvir combination therapy for treatment-naive cirrhotic and noncirrhotic patients with chronic hepatitis C virus genotype 1, 4, or 6 infection: A randomized trial. Annals of Internal Medicine. 2015;**163**(1):1-13. DOI: 10.7326/M15-0785

[44] Gutierrez A, Lawitz J, Poordad F. Interferon-free, direct-acting antiviral therapy for chronic hepatitis C. Journal of Viral Hepatitis. 2015;**22**:861-870. DOI: 10.1111/jvh.12422

[45] Muhammad A, Varsha M, Muhammad F. Laboratory methods for diagnosis and management of hepatitis C virus infection. Laboratoriums Medizin. 2013;**44**:292-299. DOI: 10.1309/LMASROYD8BRS0GC9

[46] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/206843s006lbl.pdf [Accessed: January 15, 2018]

[47] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2013/205123s001lbl.pdf [Accessed: January 16, 2018]

[48] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2015/204671s002lbl.pdf [Accessed: January 20, 2018]

[49] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/205834s000lbl.pdf [Accessed: January 22, 2018]

[50] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/206619lbl.pdf [Accessed: January 25, 2018]

[51] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/208341s000lbl.pdf [Accessed: January 30, 2018]

[52] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/208261s002lbl.pdf. [Accessed: February 02, 2018]

[53] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/209195s000lbl.pdf. [Accessed: February 07, 2018]

[54] Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/209394s000lbl.pdf. [Accessed: February 10, 2018]

Chapter 9

Norovirus GII.17: The Emergence and Global Prevalence of a Novel Variant

Yongxin Yu, Yingjie Pan, Shuling Yan and
Yongjie Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76256

Abstract

A rare norovirus (NoV) genotype GII.17 has recently emerged and rapidly became predominant in most East Asian countries in the winters of 2014–2015. In this study, we report the diversity of NoV GII.17 in detail; a total of 646 GII.17 sequences obtained during 1978–2015 were analyzed and subjected to meta-analysis. At least five major recombinant GII.17 clusters were identified. Each recombinant variant group appeared to have emerged following the time order: GII.P4-GII.17 (1978–1990), GII.P16-GII.17 (2001–2004), GII.P13-GII.17 (2004–2010), GII.Pe-GII.17 (2012–2015) and GII.P3-GII.17 (2011–2015). The newly emerged GII.P3-GII.17 variant, which exhibited significant sequence and structure variations, is evolving toward a unique lineage. Our results indicate that circulation of GII.17 appears to change every 3–5 years due to replacement by a newly emerged variant and that the evolution of GII.17 is sequentially promoted by inter-genotype recombination, which contributes to the exchange between non-GII.17 and GII.17 RdRp genes and drives the evolution of GII.17 capsid genes.

Keywords: emergence, evolution, genotyping, global prevalence, norovirus, recombination

1. Introduction

Norovirus (NoV) is the predominant etiological viral agent of acute gastroenteritis across all ages; usually old people and young kids are more susceptible to these viruses [1]. Currently, the genus of *norovirus* can be divided into at least seven genogroups (GI–GVII). Of these, GI, GII and

GIV can be detected in the samples from human gastroenteritis [2, 3]. Within each genogroup, NoV strains can be further subdivided into diverse (over 40) genotypes based on sequence similarity of the RNA-dependent RNA polymerase (RdRp) and major capsid genes [2, 3]. Globally, GII strains have already contributed to over 75% of human NoV cases [4, 5], whereas a special genotype GII.4 has been found to be responsible for the majority of outbreaks since 1990s [6, 7], and novel GII.4 variants emerged every 2–4 years [6–8].

In the winter of 2014–2015, a rare NoVs genotype, GII.17, emerged in most of the East Asian regions including China (Guangdong [9–13], Jiangsu [14], Zhejiang [15], Hebei [16], Hong Kong [17], Taiwan [18], Beijing [19], and Shanghai [20], Japan [21] and South Korea [22]). Soon after, the novel GII.17 strain became the most predominant NoV strain, replacing the pandemic strain GII.4 Sydney 2012, responsible for the majority of gastroenteritis outbreaks in this region [23]. A limited number of cases were also reported in other countries (Italy [24], Romania [25] and the USA [26]).

Recombination allows a substantial exchange of genetic materials and is a major driving force of viral evolution [2, 27]. However, although multiple studies have reported on the evolutionary dynamics of NoVs, the role of recombination in shaping NoV evolutionary history is still very significant. Given the importance of GII.17 NoV as a cause of epidemic gastroenteritis in recent years (2014–2015), it is crucial to better understand how this genotype has evolved over time.

This study aimed to determine the mechanisms of evolution in norovirus GII.17 strain from 1978 to 2015, with particular focus on the effects of recombination events on the acquisition of non-GII.17 RdRps.

2. Materials and methods

2.1. NoV GII.17 sequence datasets

Sequence datasets were constructed following the strategy described by Yu et al. [28]. Briefly, NoV GII.17 sequences were all collected from two divided pathways (**Figure 1**): firstly, search results from the GenBank nucleotide database using combination keywords such as "norovirus and GII.17" and "norovirus and 17" in December 2015 and secondly, publications from the PubMed and Google scholar literature databases (papers published between 2003 and 2015) that contain a combination of "norovirus" and "GII.17" in the titles, keywords and abstracts [28]. Non-English literatures were excluded [28].

Subsequently, all the NoV GII.17 sequences were downloaded in FASTA format. Then the corresponding information of each sequence was edited by using Geneious [29] as in a uniform format which included the sequence name (or accession number) and length, sample source, sampling time and sites [28]. Duplicated sequences were removed [28]. To delete the non-GII.17 sequences, all the candidate sequences were analyzed with the online-based genotyping tool [30], which is designed to identify norovirus genotypes based on phylogenetic analysis.

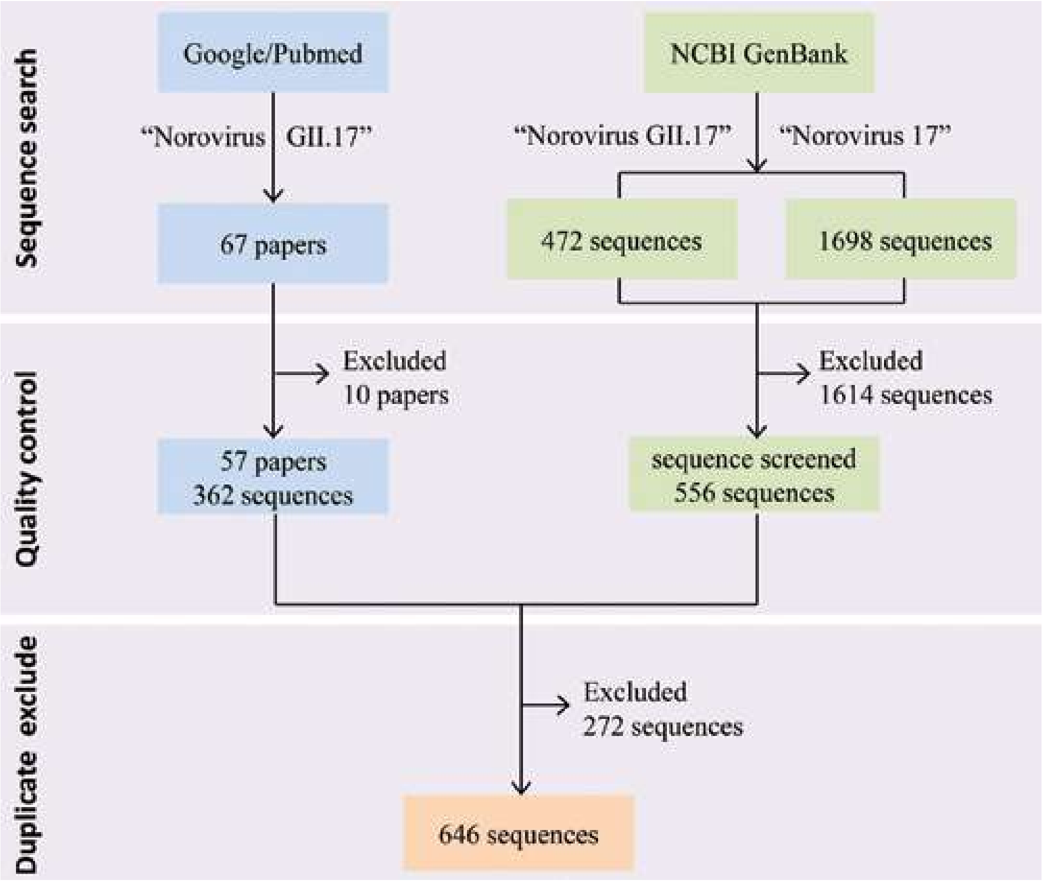


Figure 1. Flow chart of sequence collection strategy.

2.2. Phylogenetic analysis

Nucleotide sequences were aligned using the ClustalW program. Phylogenetic analysis was performed with MEGA 5.1 package [31] based on partial sequences of ORF1 (241 nt) and ORF2 (181 nt). The reference strains were retrieved from NCBI nucleotide database [32]. Phylogenetic trees were reconstructed using the Tamura-Nei model and maximum likelihood methods. Bootstrap was calculated with 1000 pseudo-replicate datasets. The distance scale represents the number of nucleotide substitution per position.

2.3. Recombination analysis

To detect recombination events, sequences were aligned with ClustalW program in the MEGA 5.1 package [31] and then checked manually. The reference strains were retrieved from NCBI nucleotide database. NoV strains were defined as recombinants if they were grouped into different genotypic clusters on the phylogenetic trees, which were reconstructed using full-length sequences of ORF1 (5107 nt) and ORF2 (1623 nt), respectively.

Simplot method [33] was employed to identify the recombination breakpoint site and to further verify the NoV recombinants. The bootstrap values were plotted for a window of 400 nt, moving in increments of 10 nt along the alignment.

2.4. Homology modeling

The tertiary structures of the capsid P domains of NoV GII.17 were modeled using SWISS-MODEL online server [34]. The recently published GII.17 domain dimer X-ray crystal structure (PDB accession: 5F4O) [35] was used as the template for generating homology models. The constructed models were examined and edited using PyMoL [36].

3. Results

3.1. NoV GII.17 sequence dataset

A total of 472 and 1698 sequences were obtained from the GenBank nucleotide database using "norovirus and GII.17" and "norovirus and 17" as search keywords, respectively. After manually checking sequence genotyping results, the non-GII.17 sequences (n = 1614) of the above two sequence datasets were excluded from subsequent analysis (**Figure 1**). A literature search yielded 67 citations, among which 57 of the research articles (85%) reported norovirus GII.17 sequences (n = 362). All screened sequences from these two independent sources were combined, and duplicated sequences were removed. Finally, a total of 646 sequences belonging to NoV GII.17 were obtained (**Figure 1**).

These 646 sequences ranged from 205 to 7570 nt in length. Over 60% of them (n = 427) were shorter than 400 nt, while 57 sequences covered the nearly complete or the complete viral genome of more than 7000 nt. All the NoV GII.17 sequences were located within ORF1, ORF2 or the region overlapping ORF1 and ORF2. Specifically, 3.41% of sequences (n = 22) belonged to ORF1, and 79.26% (n = 512) belonged to ORF2. The remaining 112 sequences contained regions from both ORF1 and ORF2.

3.2. Genetic diversity of NoV GII.17

Genotyping analysis revealed that many of the GII.17 sequences were genetic recombinants since two distinct genotypic regions, ORF1 (RdRp) and ORF2 (VP 1), were identified in the same sequence. All the recombinants contained the same ORF2 genotype of GII.17 while exhibiting varying ORF1 genotypes of GII.P3, GII.P13, GII.P16, GII.Pe and GII.P4. Phylogenetically, all the GII.17 ORF2 sequences were categorized into at least five major clusters. Interestingly, each cluster composed of only one type of GII.17 recombinant, for example, GII.P4-GII.17, GII.P16-GII.17, GII.P13-GII.17, GII.Pe-GII.17 and GII.P3-GII.17 (**Figure 2A**). The collection dates of the corresponding strains within each cluster indicated a temporally sequential clustering and were distributed as follows: GII.P4-GII.17 cluster during 1978–1990, GII.P16-GII.17 cluster during 2001–2004, GII.P13-GII.17 cluster during 2004–2010, GII.Pe-GII.17 cluster during 2012–2015 and GII.P3-GII.17 cluster during 2011–2015 (**Figure 2A**).

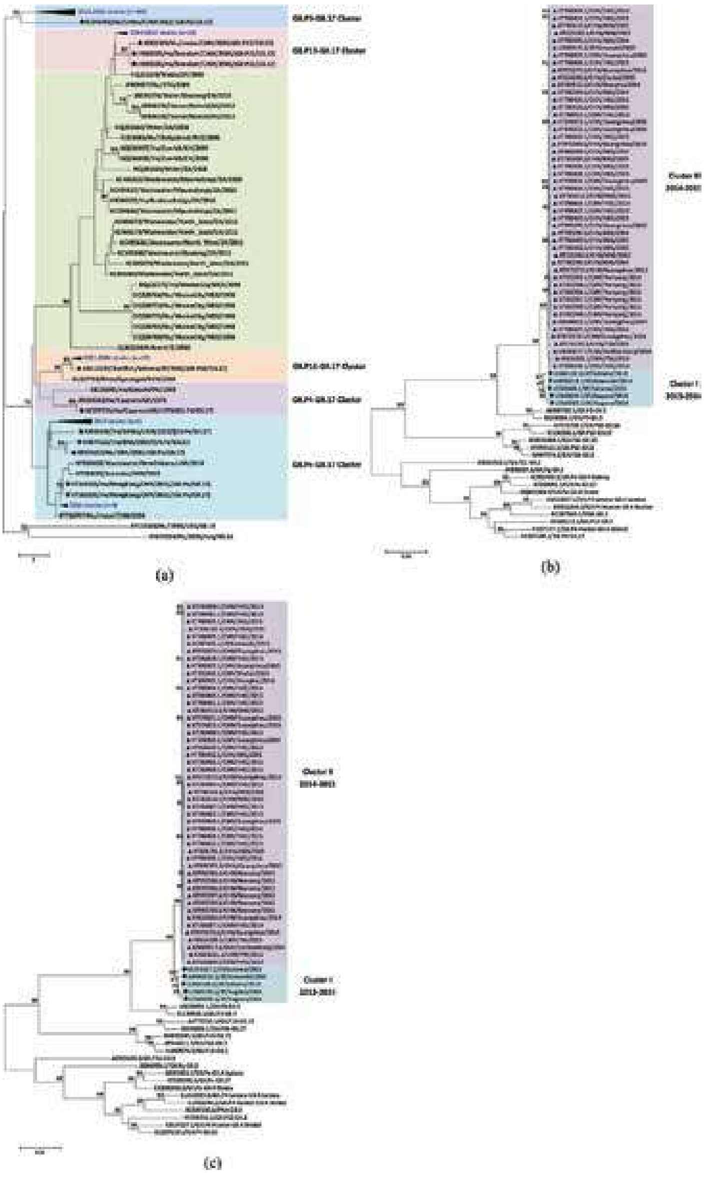


Figure 2. Phylogenetic trees of NoV GII.17. The trees were generated using the maximum likelihood method. Gaps in alignment were ignored in the analysis. Bootstrap (1000 replicates) analysis was used for statistical support; values >70% are shown. Tamura-Nei was used as the model of substitution. The distance scale represents the number of nucleotide substitutions per position. The sequences were obtained from GenBank and named according to the GenBank accession ID, strain name, followed by the year and country of isolation (AUS, Australia; BD, Bangladesh; BRA, Brazil; CAM, Cameroon; CHN, China; ETH, Ethiopia; FRA, France; GF, French Guiana; IE, Ireland; ITA, Italy; JP, Japan; KE, Kenya; KOR, Korea; MEX, Mexico; MOR, Morocco; NIC, Nicaragua; PK, Pakistan; RUS, Russia; SIG, Singapore; ZA, South Africa; CH, Switzerland; TH, Thailand; UY, Uruguay; the USA, United States of America). (A) The rooted tree was reconstructed using the 5′ ends of VP1 nucleotide sequences (189 nt) of all GII.17 NoVs, corresponding to the location of LC037415: 5141–5329 nt. The recombinant strains are indicated by black pentagrams. The recombinant variant clusters are highlighted in different colors. GII.13 (AY113106) and GII.21 (AY675554) were used as the out-groups. Two un-rooted trees were reconstructed using near-complete ORF1 (B) and ORF2 (C) nucleotide sequences (n = 72). The recently emerged NoV GII.17 sequences (2013–2015) are indicated with diamonds (cluster I, 2013–2014) and triangles (Cluster II, 2014–2015).

Notably, 27 sequences did not belong to any of the above-described lineages (**Figure 2A**); however, they were more closely related to the GII.P13-GII.17 cluster. Coincidently, all of these sequences were isolated in a similar time period (1998–2012) as the GII.P13-GII.17 cluster (2004–2010) (**Figure 2A**).

The GII.P3-GII.17 cluster, containing 486 sequences, was the largest recombinant lineage of the five recombinant clusters, with most of the sequences discovered in the past 2 years (2014 and 2015) (**Figure 2A**). In addition, the VP1 sequences (viral protein 1, the major capsid protein) of the recently emerged epidemic strains during 2014 and 2015 were also classified into this GII.P3-GII.17 cluster (**Figure 2A**).

3.3. Recombination confirms the emergence of GII.17

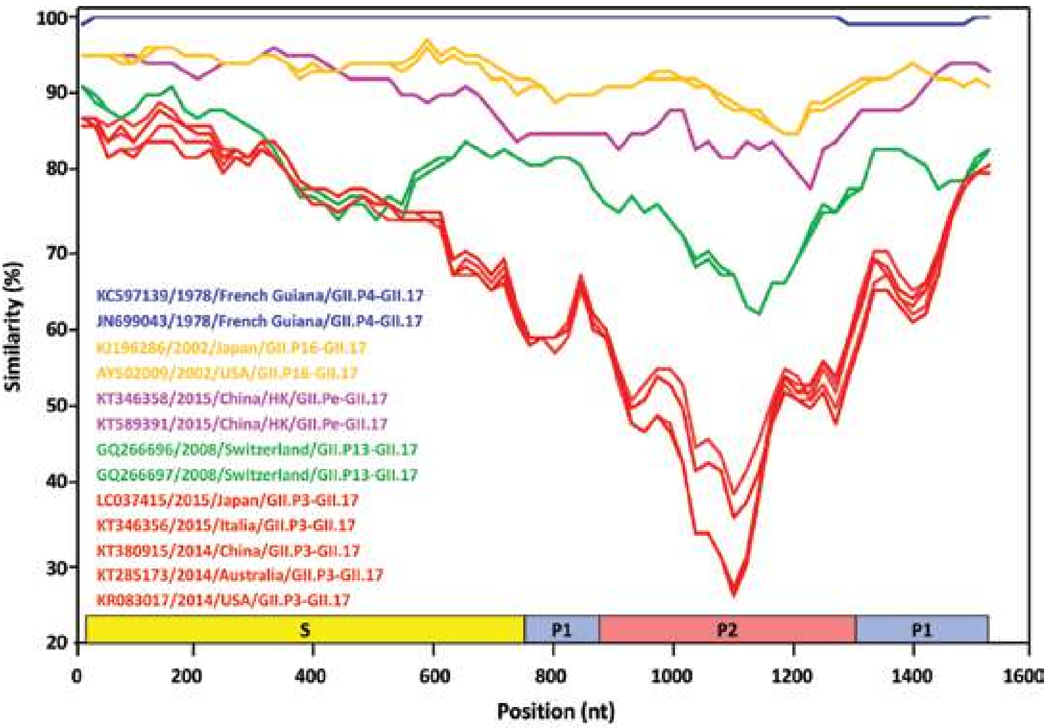
Given that only a small portion of the ORF2 sequences (181 nt) was included in the phylogenetic tree (**Figure 2A**) and that the bootstrap value of the GII.P3-GII.17 cluster was less than 80%, two more phylogenetic trees were constructed using the full-length sequences of ORF1 and ORF2 (n = 54) in order to further verify the recombination in sequences from the GII.P3-GII.17 cluster. On the ORF1 tree (**Figure 2B**), all the sequences were grouped into GII.3 genotype clusters with a high bootstrap value (99%). On the ORF2 tree (**Figure 2C**), sequences were grouped into GII.17 genotype clusters with a high bootstrap value (99%). These results demonstrated that the new epidemic strains were in fact GII.P3-GII.17 intergenic recombinants.

Simplot analysis was also performed to confirm the recombination events. The candidate recombinant sequences shared high similarities with GII.3 in ORF1 and with GII.17 in ORF2, respectively. In addition, recombination breakpoints were identified within the overlapping regions of ORF1 and ORF2.

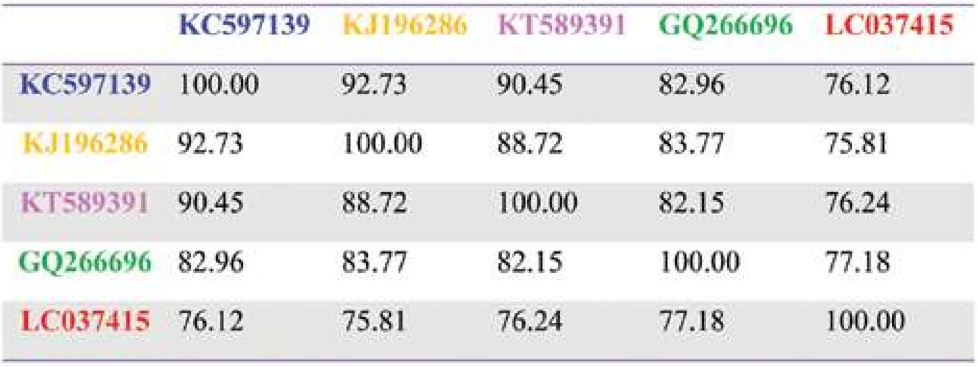
3.4. Sequence variations of the GII.17 P domain

To explore the evolution of NoV GII.17 capsid genes and the importance of intergenic recombination, five representatives of the P domain sequences (GII.P3, GII.P13, GII.P16, GII.Pe and GII.P4) from distinct clusters discovered between 1978 and 2015 were subjected to analysis. Sequence similarity among all complete genomic sequences was analyzed using Simplot. GII.P16-GII.17 and GII.Pe-GII.17 variants shared high similarities with the GII.17 variant that emerged the earliest (GII.P4-GII.17 in 1978) at the identification of 92.73 and 90.45%, respectively (**Figure 3B**). However, the variant GII.P13-GII.17 that emerged after revealed a decline in similarity with GII.P4-GII.17 (82.96% of identity). The recently emerged GII.P3-GII.17 variant showed the largest divergence (76.12% of identity) from the GII.P4-GII.17 variant (**Figure 3B**). Interestingly, a typical "V" shape was observed in our Simplot analysis (**Figure 3A**), suggesting that the P2 subdomain is the most hypervariable region in the GII.17 capsid and continues to evolve (**Figure 3A**).

A structure-based sequence alignment of the P domain (amino acid) was performed to examine any potential differences among these five GII.17 variants. Overall, P1 subdomain sequences were relatively conserved (**Figure 4**); however, significant sequence differences were observed in the outer loop regions in the P2 subdomain that includes the B-loop (aa 291–299), P-loop (aa



(a)



	KC597139	KJ196286	KT589391	GQ266696	LC037415
KC597139	100.00	92.73	90.45	82.96	76.12
KJ196286	92.73	100.00	88.72	83.77	75.81
KT589391	90.45	88.72	100.00	82.15	76.24
GQ266696	82.96	83.77	82.15	100.00	77.18
LC037415	76.12	75.81	76.24	77.18	100.00

(b)

Figure 3. Sequence similarity of P domain (nucleotides). (A) Simplot analysis: The Simplot method was performed to show the similarity between complete NoV GII.17 VP1 nucleotide sequences. Thirteen representative sequences were picked from five recombinant clusters and highlighted in different colors. The horizontal axis represents the nucleotide positions of VP1, and the vertical axis represents the sequence similarity compared to the GII.P4-GII.17 (KC597139) strain. (B) Similarity analysis: The pair-wise identity of five representative GII.17 sequences from each recombinant cluster, GII.P4-GII.17 (KC597139), GII.P16-GII.17 (KJ96286), GII.Pe-GII.17 (KT589391), GII.P13-GII.17 (GQ266696) and GII.P3-GII.17 (LC037415), was analyzed.

339–352), A-loop (aa 375–383) and T-loop (aa 395–400) (**Figure 4**). In addition, slight changes were also found in the S-loop (aa 438–445) located in the P1 subdomain and in the U-loop (aa 408–417) located within the junctional region of the P1 and P2 subdomains (**Figure 4**).

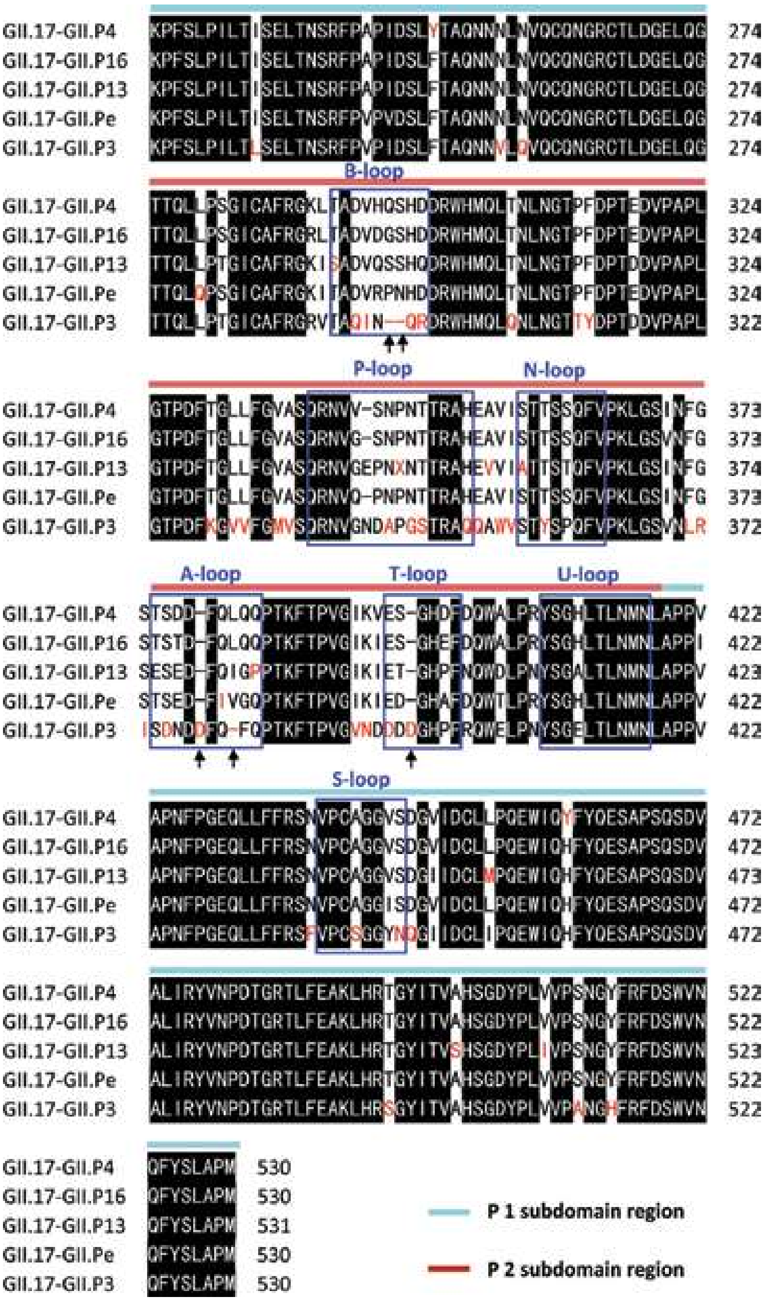

GII.17-GII.P4 KPFSLPILTISELTNSRFPAPIDSLYTAQNNNLNVQCQNGRCTLDGELQG 274
GII.17-GII.P16 KPFSLPILTISELTNSRFPAPIDSLFTAQNNNLNVQCQNGRCTLDGELQG 274
GII.17-GII.P13 KPFSLPILTISELTNSRFPAPIDSLFTAQNNNLNVQCQNGRCTLDGELQG 274
GII.17-GII.Pe KPFSLPILTISELTNSRFPVPVDSLFTAQNNNLNVQCQNGRCTLDGELQG 274
GII.17-GII.P3 KPFSLPILTLSELTNSRFPVPIDSLFTAQNNVLQVQCQNGRCTLDGELQG 274
B-loop
GII.17-GII.P4 TTQLLPSGICAFRGKLTADVHQSHDDRWHMQLTNLNGTPFDPTEDVPAPL 324
GII.17-GII.P16 TTQLLPSGICAFRGRLTADVDGSHDDRWHMQLTNLNGTPFDPTEDVPAPL 324
GII.17-GII.P13 TTQLLPTGICAFRGKISADVQSSHQDRWHMQLTNLNGTPFDPTDDVPAPL 324
GII.17-GII.Pe TTQLQPSGICAFRGKITADVRPNHDDRWHMQLTNLNGTPFDPTEDVPAPL 324
GII.17-GII.P3 TTQLLPTGICAFRGRVTAQIN--QRDRWHMQLQNLNGTTYDPTDDVPAPL 322
P-loop
N-loop
GII.17-GII.P4 GTPDFTGLLFGVASQRNVV-SNPNTTRAHEAVISTTSSQFVPKLGSINFG 373
GII.17-GII.P16 GTPDFTGLLFGVASQRNVG-SNPNTTRAHEAVISTTSSQFVPKLGSVNFG 373
GII.17-GII.P13 GTPDFTGLLFGVASQRNVGEPNXNTTRAHEVVIATTSTQFVPKLGSINFG 374
GII.17-GII.Pe GTPDFTGLLFGVASQRNVQ-PNPNTTRAHEAVISTTSSQFVPKLGSINFG 373
GII.17-GII.P3 GTPDFKGVVFGMVSQRNVGNDAPGSTRAQDAWVSTYSPQFVPKLGSVNLR 372
A-loop
T-loop
U-loop
GII.17-GII.P4 STSDD-FQLQQPTKFTPVGIKVES-GHDFDQWALPRYSGHLTLNMNLAPPV 422
GII.17-GII.P16 STSTD-FQLQQPTKFTPVGIKIES-GHEFDQWALPRYSGHLTLNMNLAPPI 422
GII.17-GII.P13 SESED-FQIGPPTKFTPVGIKIET-GHPFNQWDLPNYSGALTLNMNLAPPV 423
GII.17-GII.Pe STSED-FIVGQPTKFTPVGIKIED-GHAFDQWTLPRYSGHLTLNMNLAPPV 422
GII.17-GII.P3 ISDNDDFQ-FQPTKFTPVGVNDDDDGHPFRQWELPNYSGELTLNMNLAPPV 422
S-loop
GII.17-GII.P4 APNFPGEQLLFFRSNVPCAGGVSDGVIDCLLPQEWIQYFYQESAPSQSDV 472
GII.17-GII.P16 APNFPGEQLLFFRSNVPCAGGVSDGVIDCLLPQEWIQHFYQESAPSQSDV 472
GII.17-GII.P13 APNFPGEQLLFFRSNVPCAGGVSDGIIDCLMPQEWIQHFYQESAPSQSDV 473
GII.17-GII.Pe APNFPGEQLLFFRSNVPCAGGISDGVIDCLLPQEWIQHFYQESAPSQSDV 472
GII.17-GII.P3 APNFPGEQLLFFRSFVPCSGGYNQGIIDCLIPQEWIQHFYQESAPSQSDV 472
GII.17-GII.P4 ALIRYVNPDTGRTLFEAKLHRTGYITVAHSGDYPLVVPSNGYFRFDSWVN 522
GII.17-GII.P16 ALIRYVNPDTGRTLFEAKLHRTGYITVAHSGDYPLVVPSNGYFRFDSWVN 522
GII.17-GII.P13 ALIRYVNPDTGRTLFEAKLHRTGYITVSHSGDYPLIVPSNGYFRFDSWVN 523
GII.17-GII.Pe ALIRYVNPDTGRTLFEAKLHRTGYITVAHSGDYPLVVPSNGYFRFDSWVN 522
GII.17-GII.P3 ALIRYVNPDTGRTLFEAKLHRSGYITVAHSGDYPLVVPANGHFRFDSWVN 522
GII.17-GII.P4 QFYSLAPM 530
GII.17-GII.P16 QFYSLAPM 530
GII.17-GII.P13 QFYSLAPM 531
GII.17-GII.Pe QFYSLAPM 530
GII.17-GII.P3 QFYSLAPM 530
P 1 subdomain region
P 2 subdomain region

Figure 4. Sequence alignment of P domain (amino acids). The structure-based sequence alignment of the P domains of GII.P4-GII.17 (AGI17592), GII.P16-GII.17 (AII73747), GII.Pe-GII.17 (ALD83748), GII.P13-GII.17 (ACT68315) and GII. P3-GII.17 (BAR42289) from five recombinant clusters was performed. Regions spanning the P1 and P2 subdomains are indicated by stripes. Identical residues are highlighted with black background, while distinct residues are shown in red characters. The conventional GII genogroup HBGA surface binding loops (A-, B-, P-, S-, T-, U- and N-) are indicated by blue frames.

Interestingly, various mutations were also identified in the P2 subdomain, especially in the loops on the outer surface of the protein. Comparing to the other four earlier variants, three deletions (aa 349, 350 and 381) and two insertions (aa 379 and 397) were observed in the recently emerged GII.P3-GII.17 variant (**Figure 4**). These mutations, which are located in the B-loop (aa 349 and 350), A-loop (aa 379 and 381) and T-loop (aa 397), might have altered the binding capacity of NoV to host human histo-blood group antigen (HBGA) (**Figure 4**). These results also suggested that the GII.17 P domain might have evolved from a non-prevalent strain into an epidemic GII.17 variant strain (GII.P3-GII.17).

The position of amino acids in VP1 corresponds to AGI17592. Arrows indicate deletions (aa 349, 350 and 381) and insertions (aa 379 and 397) in the recently emerged GII.P3-GII.17 variant, as compared to the other GII.17 variants.

3.5. Structure shift of the GII.17 P domain

To better understand how this novel NoV GII.17 spread so rapidly and widely, the P domain structures of all five GII.17 variants were constructed based on homology modeling. Comparison of the overall structures of the five GII.17 variants revealed that most amino acid substitutions occurred in the P2 subdomain, while the P1 subdomain remained relatively conserved.

Figure 5 shows the surface loops that formed the conventional GII NoV binding interface (B-, T-, N-, P-, U-, S- and A-loops). The front-side view indicates that the outer structures of the N- and U-loops of all five GII.17 variants remained relatively unchanged (**Figure 5**). However, compared to the other four GII.17 variants, distinct structural changes were observed in the A-, B- and T-loops of the GII.P3-GII.17 variant (**Figure 5**). The top-side view revealed that the P-loops of the GII.P13-GII.17 and the GII.P3-GII.17 variants are protruding from the surface of the proteins (**Figure 5**), a feature clearly different from those observed in the other three GII.17 variants.

To clarify the specific changes observed in the loops, superpositions of the P domain structures of the five representative GII.17 variants were constructed (**Figure 6**). The most prominent difference was observed in the P-loop (**Figure 6**). Three distinct structures of the P-loop were observed among the five GII.17 variants, with the GII.P4-GII.17, GII.P16-GII.17 and GII.Pe-GII.17 variants sharing the same structure. In contrast, the GII.P13-GII.17 and GII. P3-GII.17 variants each presented a unique structure (**Figure 6**). In addition, the T-, B-, U- and A-loops also exhibited considerable differences. Notably, the GII.P4-GII.17, GII.P16-GII.17, GII.Pe-GII.17 and GII.P13-GII.17 variants shared similar T-, B-, U- and A-loops; while the GII.P3-GII.17 variant displayed a unique shift in structure in these four loops (**Figure 6**).

3.6. Global distribution of NoV GII.17 variants

Sequences of the five distinct variants of GII.17 (n = 646) were obtained from at least 24 countries and regions from five continents: Asia (China, Hong Kong, Taiwan, Japan, Bangladesh, Singapore, South Korea, Pakistan and Thailand), Europe (Italy, France and Russia), America (Brazil, Mexico, Nicaragua, Uruguay, French Guiana and the USA), Africa (Cameroon,

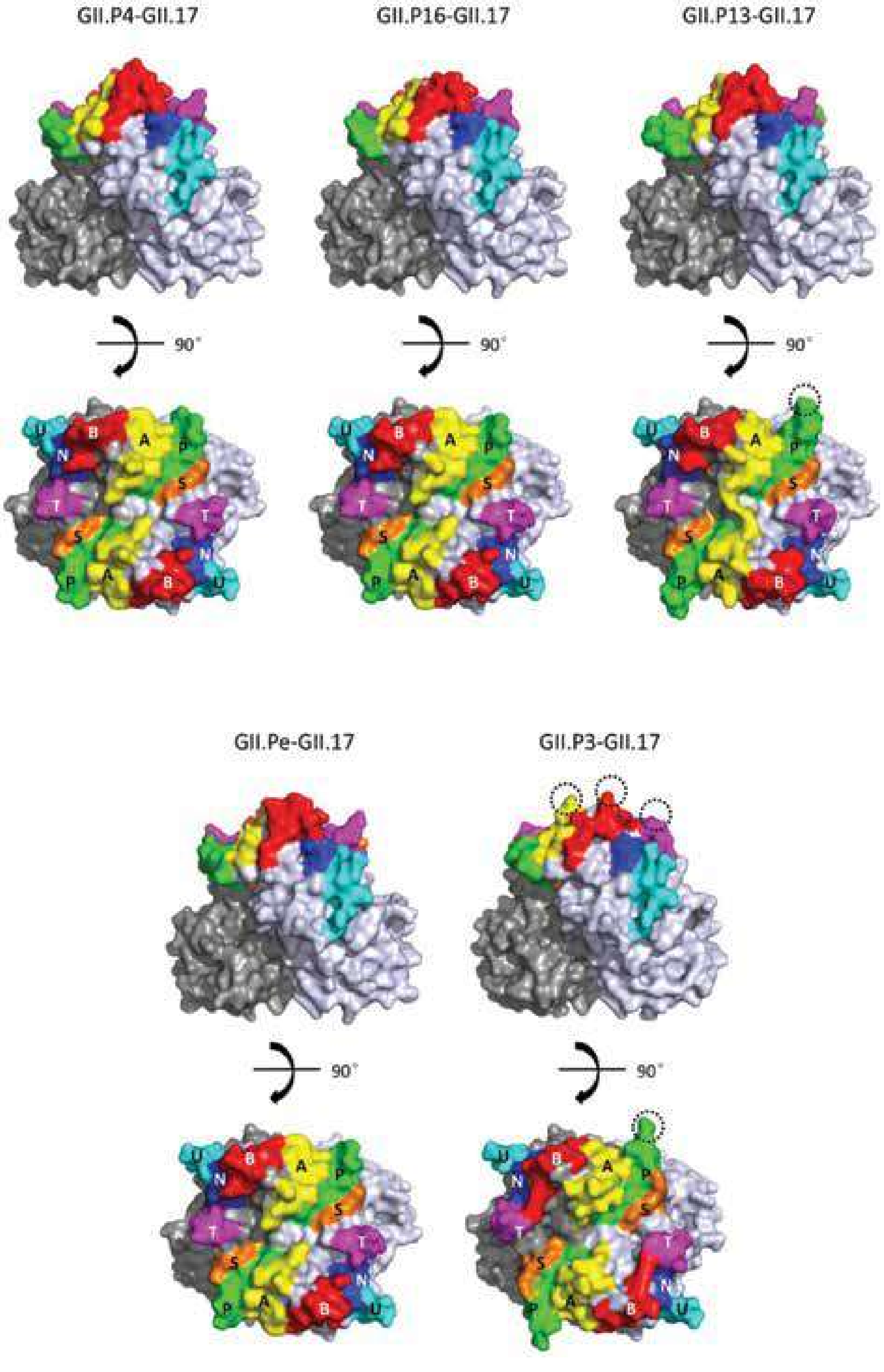


Figure 5. Three-dimensional structure of P domain. Three-dimensional P domain structures of the five GII.17 variants were predicted by homologous modeling. The seven surface-binding loops, that is, A-, B-, P-, S-, T-, U- and N-loops, are indicated in yellow, red, green, orange, pink, sky blue and blue, respectively. The black circles highlight the changes present in the binding loops among different GII.17 variants.

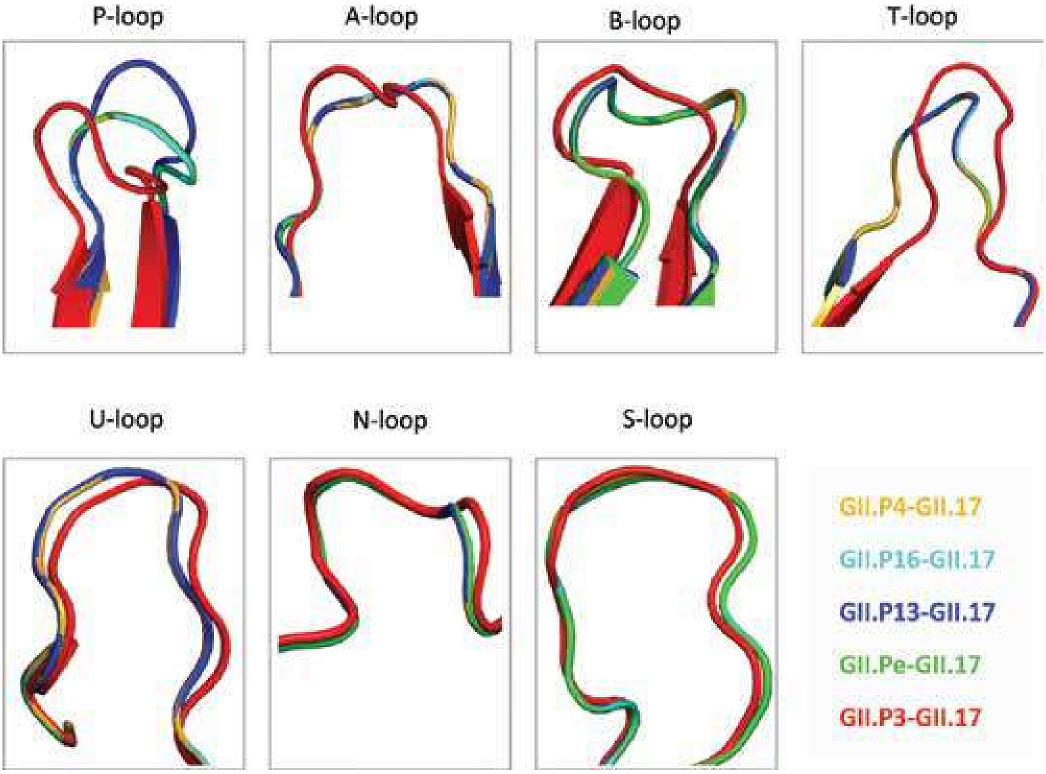


Figure 6. Comparison of the seven major surface loops of the GII.17 variants. Structures of the seven surface-binding loops located on the P domain were compared by superposition. Color schemes: wheat, GII.P4-GII.17 (AGI17592); sky blue, GII.P16-GII.17 (AII73747); green, GII.Pe-GII.17 (ALD83748); blue, GII.P13-GII.17 (ACT68315) and red, GII.P3-GII.17 (BAR42289).

Ethiopia, Kenya, Morocco and South Africa) and Oceania (Australia). The number of sequences obtained in these regions was unevenly distributed, with certain countries being over-represented compared to others. Most of the sequences were obtained in Asia (89.30%) and Europe (7.31%). The number of sequences collected in China was the highest (71.12%), followed by South Korea (9.80%), Kenya (4.99%) and Japan (4.81%).

Since most of these NoV GII.17 sequences were categorized into five recombination variant clusters; it is important to understand how the different GII.17 variants are distributed globally. Only three sequences belonging to the GII.P4-GII.17 variant were collected from French Guiana and Pakistan (**Figure 7A**). Most of the sequences of the GII.P16-GII.17 variant were isolated in Japan (n = 17) (**Figure 7A**). In addition, sequences of the GII.Pe-GII.17 variant were well dispersed among different countries including Cameroon (n = 1), Ethiopia (n = 6), China (n = 7), South Korea (n = 1), Brazil (n = 2) and the USA (n = 1) (**Figure 7A**). Sequences from the GII.P13-GII.17 variant were even more widely distributed: China (n = 3), Japan (n = 1), Bangladesh (n = 10), Singapore (n = 3), South Korea (n = 3), Thailand (n = 2), France (n = 2), Russia (n = 2), Brazil (n = 2), Mexico (n = 1), Nicaragua (n = 1), Uruguay (n = 1), Cameroon (n = 3) and South Africa (n = 1) (**Figure 7A**).

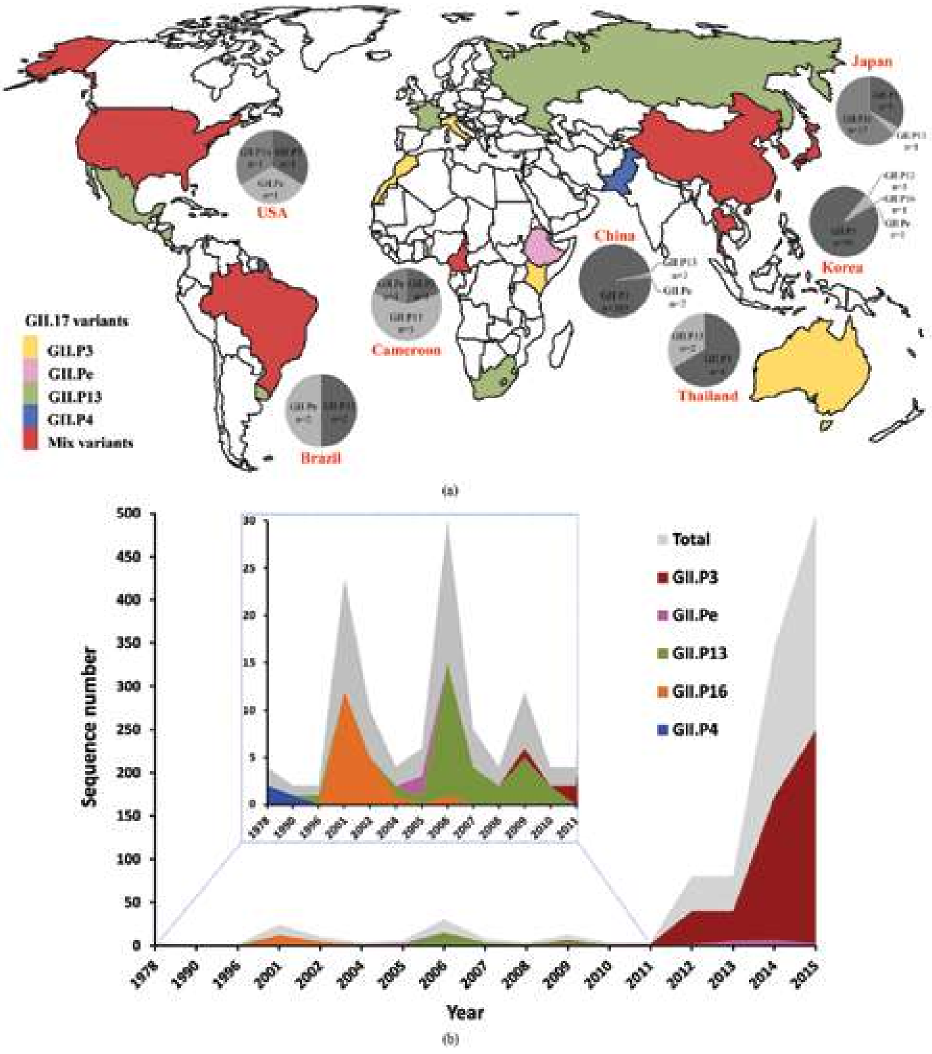


Figure 7. Global distribution and prevalence of GII.17 variants. (A) Geographical distribution of the GII.17 variants from 1978 to 2015: Different GII.17 variants are shown in different colors. The distributions of multiple variants in the same regions are shown in pie-charts (map template from digital vector maps) and (B) Yearly prevalence of GII.17 variants during 1978–2015: Different GII.17 variants are shown in different colors. Part of the graph representing data from the period between 1978 and 2011 was partially enlarged to allow the examination of detailed results.

As for the GII.P3-GII.17 variant, its sequence distribution was also quite scattered, with most of the sequences found in China (n = 389), South Korea (n = 50), Kenya (n = 28), Japan (n = 9) and Thailand (n = 4) (**Figure 7A**).

Interestingly, multiple GII.17 variants were often observed in one country or region (**Figure 7A**). For example, over 70% of the GII.17 sequences (399 in 561) were detected in China, including sequences from the GII.P3-GII.17, GII.P13-GII.17 and GII.Pe-GII.17 variants (**Figure 7A**). Except for GII.P4-GII.17, the other four GII.17 variants were found in South Korea, with GII. P3-GII.17 being the most dominant. Different GII.17 variants were also observed in Japan

(GII.P3-GII.17, GII.P13-GII.17 and GII.P16-GII.17), Cameroon (GII.P3-GII.17, GII.P13-GII.17 and GII.Pe-GII.17), the USA (GII.P3-GII.17, GII.P16-GII.17 and GII.Pe-GII.17), Thailand (GII. P3-GII.17 and GII.P13-GII.17) and Brazil (GII.P13-GII.17 and GII.Pe-GII.17) (**Figure 7A**).

3.7. Yearly prevalence of GII.17 variants

Figure 7B showed the yearly distribution of the GII.17 variants detected worldwide. The GII.17 sequences isolated from 1978 to 2015 were unevenly distributed. The number of isolated sequences peaked in 2001, 2006 and 2009, then continued to increase significantly from 2011 to 2015 and reached the highest peak in 2015 (**Figure 7B**). This result suggested that circulation of GII.17 might change every 3–5 years due to replacement by a newly emerged variant. The GII.P4-GII.17 variant, which emerged the earliest, only appeared in 1978 and 1990 (**Figure 7B**). About a dozen years later, the GII.P16-GII.17 variant was discovered in 2002 and persisted till 2004, when the GII.P13-GII.17 variant emerged and replaced it. A few of the GII. P3-GII.17 variants emerged between 2009 and 2013, and the number of sequences that belonged to this variant sharply increased during 2014 and 2015. Over 70% of the GII.17 sequences belonged to this GII.P3-GII.17 variant and were isolated during this period (**Figure 7B**). It is worth noting that the GII.Pe-GII.17 variant was also observed between 2012 and 2015, although the number of sequences identified during this period was far less than that of the GII.P3-GII.17 variant (**Figure 7B**).

4. Discussion

NoV, one of the leading causes of human acute gastroenteritis, is wildly distributed around the world. In the past decades, NoV GII.4 was identified as the most predominant genotype involved in numerous epidemic outbreaks, for example, in 2002, 2004, 2006, 2009 and 2012 [37]. Recently, a novel NoV GII.17 variant has emerged and is responsible for multiple disease outbreaks mainly in China and Japan [9, 11–22]. GII.17 appeared as a dominant strain replacing the GII.4 strain in these regions during 2014–2015 [23].

NoV GII.17 strains have been circulating among various human populations for over 37 years, with previously emerged GII.17 sequences being sporadically detected in multiple regions in various continents including most of Africa, Asia, Europe, North America and South America (**Figure 7A**). However, it is still unclear as to why this recently emerged GII.17 variants spread so rapidly and widely within such a short time period.

4.1. Genetic diversity and recombination

Previous studies have suggested that GII.17 sequences could be classified into distinct clusters based on the divergence of VP1. However, only a few GII.17 sequences were analyzed in these studies [14, 15, 18, 20, 24, 26]. In order to investigate the diversity of NoV GII.17 in detail, large sequence dataset was constructed and subjected to meta-analysis and genotyping in this study (**Figure 1**). Surprisingly, at least five major recombinant GII.17 clusters, including GII.P4-GII.17, GII.P16-GII.17, GII.P13-GII.17, GII.Pe-GII.17 and GII.P3-GII.17, were identified. In addition, each recombinant group appeared to have emerged following a particular time order. For example, the earliest group of GII.P4-GII.17 was isolated in 1978, followed by GII.P16-GII.17 (2001–2004), GII.P13-GII.17 (2004–2010), GII.Pe-GII.17 (2012–2015) and GII.P3-GII.17 (2011–2015). This indicates that the evolution of GII.17 is sequentially promoted by inter-genotype genetic recombination, which contributes to the exchange between non-GII.17 and GII.17 RdRp genes and promotes the evolution of GII.17 capsid genes. Consequently, these genetic recombinations could have potentially affected the antigenic properties of NoVs and accelerated the emergence of novel epidemic variants or strains [17, 23], for example, GII.P3-GII.17. The highly diverse genome of the rare genotype of GII.17 demonstrated in this study is far beyond what has been reported previously.

Moreover, based on the phylogenetic trees of RdRp and VP1 (**Figure 2B** and **C**), all GII.P3-GII.17 sequences were further subdivided into cluster I and II. Cluster I contains the variants identified from 2013 to 2014, while the variants isolated from 2014 to 2015 comprise Cluster II. Our results suggest that after RdRp recombination with the pandemic genotype of GII.P3, the subsequently emerged GII.17 strains underwent various modifications at the capsid [23, 35], which, consequently, promoted the emergence of new variants (**Figure 3**).

Notably, the co-circulation of the GII.P3-GII.17 variant with the non-epidemic strains of GII.Pe-GII.17 indicates that the evolution of GII.17 strains was driven by multiple mechanisms in different directions within the human population in recent years (**Figure 2A**).

4.2. Shift and variation in capsid structure

Chan et al. [17] and Lu et al. [38] have previously reported that the GII.17 VP1 evolved faster than GII.4 VP1. The rapid evolution of this GII.P3-GII.17 VP1 at a rate of 5.68×10^{-3} nucleotide substitutions per site per year, which is comparable to that of GII.4 VP1 ($5.3–6.3 \times 10^{-3}$ substitutions per site per year), is faster than that in other NoV genotypes such as the epidemic genotypes GII.3 and GII.7 (1.961×10^{-3} and 2.36×10^{-3} nucleotide substitutions per site per year, respectively) [38]. In this study, compared with previous circulating GII.17 variants, the recently emerged GII.17 variant was found to exhibit significant sequence and structure variations.

First of all, the new GII.P3-GII.17 variant showed the most noticeable sequence divergence (76.12% of nucleotide identity) compared to the other GII.17 variants, for example, GII.P16-GII.17, GII.P13-GII.17 and GII.Pe-GII.17 (**Figure 3**), which could be a driving force of antigenic drift of the new GII.17 variant.

Secondly, at the amino acid level, the P domain of the GII.P3-GII.17 variant appeared to have evolved more rapidly than the other variants (**Figure 4**). Specifically, three deletions and two insertions were observed in the new GII.P3-GII.17 variant (**Figure 4**). The homology structure of the P domain also showed that many mutations were located in the human histo-blood group antigen (HBGA) binding loops of the newly emerged GII.P3-GII.17 variant (**Figure 5**). Notably, most amino acid substitutions were found in the P2 subdomain. These results suggest that these mutations most likely affect the HBGA binding property of the new GII.17 variant, which in turn, might expand its host range and prevalence [17, 21, 35]. It is important to note that changes at the antigenic epitopes of the viral capsid might also lead to adaptive advantages that contribute to the rapid spread of the virus.

The observed structural variations led us to hypothesize that, in the beginning of GII.17 evolution, the binding interface of these viruses remained relatively stable from 1978 to 2004, as reflected in the variants GII.P4-GII.17 (1978–1990) and GII.P16-GII.17 (2001–2004). Soon after that, a significant change occurred at the surface loops of the capsid (GII.P13-GII.17, 2004–2010), followed by more dramatic changes in 2011–2015 (GII.P3-GII.17, 2011–2015) (**Figure 2A**). Interestingly, the loop structures of the GII.Pe-GII.17 variant remained relatively constant, even though it was in circulation during the 2012–2015 period, indicating that this variant retained the conventional GII.17 HBGA binding interfaces [35]. Our results also confirmed that the GII.P3-GII.17 variant might have evolved as a unique lineage separated from the other GII.17 variants.

Thirdly, it is also possible that variations in the rate of evolution of the capsid were promoted by recombination, since this recombination resulted in the same capsid lineage that was associated with different RdRp [39]. Recombination breakpoints between RdRp and capsid genes might have allowed GII.17 to acquire RdRp with lower fidelity and/or increased replication efficiency, resulting in a higher rate of evolution for the associated capsid gene [17]. Moreover, within a short time period, the human host might have yet to adapt to the new GII.17 capsid, which might also support the rapid spread of the new GII.17 variant.

4.3. Waterborne GII.17 strains

NoV particles are usually detected in environmental water or in effluents from the wastewater treatment plant throughout the year [40–42]. In fact, the attack rate of waterborne NoV (over 11%) was found to be significantly higher than that associated person to person or environmental transmission [5].

Interestingly, the dataset in this study revealed that over 20% (129 in 646) of GII.17 sequences were collected from water samples or water-related outbreaks. Such waterborne GII.17 outbreaks have been reported in many countries, for example, the USA, in 2005 [43]; South Korea, in 2008–2012 [44]; Guatemala, in 2009 [45]; and China, in 2014–2015 [16, 46]. Moreover, many GII.17 sequences were also detected in environmental water samples from around the world, such as Shandong (China) [47], Singapore [48, 49], Japan [50, 51], South Africa [52, 53], Kenya [54] and New Orleans (USA) [40] .

Since the capsid protects genomic RNA from the environment, it is assumed that capsid degradation could lead to viral inactivation, probably followed by the degradation of the unprotected viral RNA [55]. Our data suggest that the capsid of GII.17 virus is more stable and therefore, can persist longer in water samples. Interestingly, Arthur et al. demonstrated that Tulane virus (a novel human NoV surrogate) also remained stable in surface water (<1 log10 reduction) after 28 days; however, viral load reduced from ≥3.5 to 4 log10 in groundwater by day 21 [56]. Consequently, improving surveillance and systematic monitoring of environmental water samples could provide valuable information on viral circulation and enable further assessment of the emergence of novel NoVs at an earlier stage.

4.4. Nomenclature of the new GII.17 variant

To coordinate the assignment of new genotypes and variants, a dual-typing system based on the complete capsid (VP1) and the partial polymerase (1300 nt) was proposed [3]. In 2014, the RdRp genotype of Kawasaki 323 strain emerged, but it was not assigned to any of the known genotypes in the database. Instead, it was assigned as the GII.P17 RdRp genotype, and the variants of NoVs were named Hu/GII/JP/2014/GII.P17-GII.17 [21]. Meanwhile, the Hu/GII.17/Gaithersburg/2014/US strain was categorized into the same group as the GII.3 strains based on its RdRp sequence. However, based on the bootstrap values (<70%) and genetic distances (<0.143), insufficient confidence was encountered when an attempt was made to classify these variants with any of the known RdRp genotypes [26]. In addition, Fu and coauthors reported that most NoV GII.17 strains detected in the 2000s were affiliated with a GII.3-like RdRp genotype and that the new GII.17 variant might be a recombinant strain that expresses a GII.3-like RdRp gene and a GII.17 capsid gene [14].

Viruses that exhibit a GII.17 VP1 genotype have previously been reported to harbor a GII.P13 ORF1 genotype, although recombinants expressing an ORF1 GII.P16, GII.P3 and GII.P4 genotype have also been identified. Sequence comparison showed that the ORF1 region of the novel GII.17 viruses, which are clustered between the GII.P3 and GII.P13 viruses, has never been detected before. Since this is the first orphan ORF1 sequence associated with GII.17, we decided to designate GII.P17 according to the criteria listed in the proposal for a unified NoV nomenclature and genotyping. The novel GII.17 virus was named Kawasaki 2014 after the first near-complete genome sequence (AB983218) was submitted to GenBank [21]. This typing tool was updated to ensure correct classification of both ORF1 and ORF2 sequences of the new GII.P17-GII.17 viruses [23].

According to the phylogenetic tree of ORF1 (**Figure 2B**), all the recently emerged GII.17 sequences are closely related to GII.P3, with a bootstrap value above 95% (=99%) (**Figure 2B**). Moreover, Simplot analysis also supports that the new GII.17 sequences had undergone recombination between a GII.P3 RdRp and a GII.17 capsid gene, and the recombination breakpoint is located within the ORF1/ORF2 overlapping region. Taken together, our results led us to propose that the new GII.17 sequences should be named as a recombinant genotype, GII.P3-GII.17, or at least be classified as a new GII.17 variant.

5. Conclusion

In conclusion, the genetic diversity and global prevalence of NoV GII.17 from 1978 to 2015 were analyzed in this study. A highly diverse genome was discovered within this rare genotype of GII.17. For example, at least five major recombinant GII.17 clusters were found to have emerged following a particular time order. The circulation of GII.17 changes every 3–5 years due to replacement by a newly emerged variant, and the evolution of GII.17 is sequentially promoted by inter-genotype genetic recombination. Most of the GII.17 sequences were detected in Asian countries including China, South Korea and Japan, and multiple GII.17 variants were found in one country or region. Moreover, the recently emerged GII.P3-GII.17 variant exhibited significant sequence and structure variations and had evolved as a unique lineage. The results presented in this study contribute to the understanding of the evolution and persistence of NoV GII.17 in human population through analyzing recombination in the viral genome. Our findings also provide important insights into the future monitoring of the global circulation of this novel GII.17 variant.

Acknowledgements

We thank the anonymous contributors for collecting and sharing the NoV sequences. This work was supported partially by the National Key Research and Development Program of China (SQ2017YFC160114), the National Natural Science Foundation of China (41376135 and 31570112), Doctoral Fund of Ministry of Education of China (20133104110006) and Innovation Program of Shanghai Municipal Education Commission (14ZZ144), China.

Conflict of interest

The authors have declared that no competing interests exist.

Author details

Yongxin Yu[1,2], Yingjie Pan[1,2], Shuling Yan[1,3] and Yongjie Wang[1,2]*

*Address all correspondence to: yjwang@shou.edu.cn

1 College of Food Science and Technology, Shanghai Ocean University, Shanghai, China

2 Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation (Shanghai), Ministry of Agriculture, China

3 Institute of Biochemistry and Molecular Cell Biology, University of Göttingen, Göttingen, Germany

References

[1] Glass RI, Parashar UD, Estes MK. Norovirus gastroenteritis. New England Journal of Medicine. 2009;**361**(18):1776-1785. DOI: 10.1056/NEJMra0804575

[2] Zheng D-P, Ando T, Fankhauser RL, Beard RS, Glass RI, Monroe SS. Norovirus classification and proposed strain nomenclature. Virology. 2006;**346**(2):312-323. DOI: 10.1016/j.virol.2005.11.015

[3] Kroneman A, Vega E, Vennema H, Vinjé J, White PA, Hansman G, Green K, Martella V, Katayama K, Koopmans M. Proposal for a unified norovirus nomenclature and genotyping. Archives of Virology. 2013;**158**(10):2059-2068. DOI: 10.1007/s00705-013-1708-5

[4] Hall AJ, Vinjé J, Lopman B, Park GW, Yen C, Gregoricus N, Parashar U. Updated Norovirus Outbreak Management and Disease Prevention Guidelines. US Department of Health and Human Services, Centers for Disease Control and Prevention, U.S.

Government Printing Office (GPO), Washington, DC, 20402-9371; 2011. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6003a1.htm

[5] Matthews J, Dickey B, Miller R, Felzer J, Dawson B, Lee A, Rocks J, Kiel J, Montes J, Moe C. The epidemiology of published norovirus outbreaks: A review of risk factors associated with attack rate and genogroup. Epidemiology and Infection. 2012;**140**(07):1161-1172. DOI: 10.1017/S0950268812000234

[6] Siebenga JJ, Vennema H, Zheng D-P, Vinjé J, Lee BE, Pang X-L, Ho EC, Lim W, Choudekar A, Broor S. Norovirus illness is a global problem: Emergence and spread of norovirus GII. 4 variants, 2001-2007. Journal of Infectious Diseases. 2009;**200**(5):802-812. DOI: 10.1086/605127

[7] Van Beek J, Ambert-Balay K, Botteldoorn N, Eden J, Fonager J, Hewitt J, Iritani N, Kroneman A, Vennema H, Vinje J. Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II 4, late 2012. Euro Surveillance. 2013;**18**(1):8-9. DOI: 10.2807/ese.18.01.20345-en

[8] Bull RA, White PA. Mechanisms of GII. 4 norovirus evolution. Trends in Microbiology. 2011;**19**(5):233-240. DOI: 10.1016/j.tim.2011.01.002

[9] Lu J, Sun L, Fang L, Yang F, Mo Y, Lao J, Zheng H, Tan X, Lin H, Rutherford S. Gastroenteritis outbreaks caused by norovirus GII. 17, Guangdong Province, China, 2014-2015. Emerging Infectious Diseases. 2015;**21**(7):1240. DOI: 10.3201/eid2107.150226

[10] Zhang X-F, Huang Q, Long Y, Jiang X, Zhang T, Tan M, Zhang Q-L, Huang Z-Y, Li Y-H, Ding Y-Q. An outbreak caused by GII. 17 norovirus with a wide spectrum of HBGA-associated susceptibility. Scientific Reports. 2015;**5**:17687. DOI: 10.1038/srep17687

[11] Wang H-B, Wang Q, Zhao J-H, Tu C-N, Mo Q-H, Lin J-C, Yang Z. Complete nucleotide sequence analysis of the norovirus GII. 17: A newly emerging and dominant variant in China, 2015. Infection, Genetics and Evolution. 2016;**38**:47-53. DOI: 10.1016/j.meegid.2015.12.007, 10.1111/jam.13052

[12] Xue L, Cai W, Wu Q, Zhang J, Guo W. Direct sequencing and analysis of the genomes of newly emerging GII. 17 norovirus strains in South China. Journal of Applied Microbiology. 2016;**120**(4):1130-1135. DOI: 10.1111/jam.13052

[13] Xue L, Wu Q, Cai W, Zhang J, Guo W. Molecular characterization of new emerging GII. 17 norovirus strains from South China. Infection, Genetics and Evolution. 2016;**40**:1-7. DOI: 10.1016/j.meegid.2016.02.026

[14] Fu J, Ai J, Jin M, Jiang C, Zhang J, Shi C, Lin Q, Yuan Z, Qi X, Bao C. Emergence of a new GII. 17 norovirus variant in patients with acute gastroenteritis in Jiangsu, China, September 2014 to March 2015. Euro Surveillance. 2015;**20**(24):21157. DOI: 10.2807/1560-7917.ES2015.20.24.21157

[15] Han J, Ji L, Shen Y, Wu X, Xu D, Chen L. Emergence and predominance of norovirus GII. 17 in Huzhou, China, 2014-2015. Virology Journal. 2015;**12**(1):1. DOI: 10.1186/s12985-015-0370-9

[16] Qin M, Dong X-G, Jing Y-Y, Wei X-X, Wang Z-E, Feng H-R, Yu H, Li J-S, Li J. A waterborne gastroenteritis outbreak caused by Norovirus GII. 17 in a hotel, Hebei, China, December 2014. Food and Environmental Virology. 2016;**8**(3):180-186. DOI: 10.1007/s12560-016-9237-5

[17] Chan MC, Lee N, Hung T-N, Kwok K, Cheung K, Tin EK, Lai RW, Nelson EAS, Leung TF, Chan PK. Rapid emergence and predominance of a broadly recognizing and fast-evolving norovirus GII. 17 variant in late 2014. Nature Communications. 2015;**6**:10061. DOI: 10.1038/ncomms10061

[18] Lee C-C, Feng Y, Chen S-Y, Tsai C-N, Lai M-W, Chiu C-H. Emerging norovirus GII. 17 in Taiwan. Clinical Infectious Diseases. 2015;**61**(11):1762-1764. DOI: 10.1093/cid/civ647

[19] Gao Z, Liu B, Huo D, Yan H, Jia L, Du Y, Qian H, Yang Y, Wang X, Li J. Increased norovirus activity was associated with a novel norovirus GII. 17 variant in Beijing, China during winter 2014-2015. BMC Infectious Diseases. 2015;**15**(1):1. DOI: 10.1186/s12879-015-1315-z

[20] Chen H, Qian F, Xu J, Chan M, Shen Z, Zai S, Shan M, Cai J, Zhang W, He J. A novel norovirus GII. 17 lineage contributed to adult gastroenteritis in Shanghai, China, during the winter of 2014-2015. Emerging Microbes and Infection. 2015;**4**(11):e67. DOI: 10.1038/emi.2015.67

[21] Matsushima Y, Ishikawa M, Shimizu T, Komane A, Kasuo S, Shinohara M, Nagasawa K, Kimura H, Ryo A, Okabe N. Genetic analyses of GII. 17 norovirus strains in diarrheal disease outbreaks from December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino acid substitutions in the capsid region. Euro Surveillance. 2015;**20**(26):21173. DOI: 10.2807/1560-7917.ES2015.20.26.21173

[22] Thanh HD, Nguyen TH, Lim I, Kim W. Emergence of norovirus GII. 17 variants among children with acute gastroenteritis in South Korea. PLoS One. 2016;**11**(5):e0154284. DOI: 10.1371/journal.pone.0154284. eCollection 2016

[23] De Graaf M, Van Beek J, Vennema H, Podkolzin A, Hewitt J, Bucardo F, Templeton K, Mans J, Nordgren J, Reuter G. Emergence of a novel GII. 17 norovirus-end of the GII. 4 era. Euro Surveillance. 2015;**20**(26):21178. DOI: 10.2807/1560-7917.ES2015.20.26.21178

[24] Medici M, Tummolo F, Calderaro A, Chironna M, Giammanco G, De Grazia S, Arcangeletti M, De Conto F, Chezzi C, Martella V. Identification of the novel Kawasaki 2014 GII. 17 human norovirus strain in Italy, 2015. Euro Surveillance. 2015;**20**(35):30010. DOI: 10.2807/1560-7917.ES.2015.20.35.30010

[25] Dinu S, Nagy M, Negru D, Popovici E, Zota L, Oprișan G. Molecular identification of emergent GII. P17-GII. 17 norovirus genotype, Romania, 2015. Euro Surveillance. 2016;**21**(7):30141. DOI: 10.2807/1560-7917.ES.2016.21.7.30141

[26] Parra GI, Green KY. Genome of emerging norovirus GII. 17, United States, 2014. Emerging Infectious Diseases. 2015;**21**(8):1477. DOI: 10.3201/eid2108.150652

[27] Bull RA, Tanaka MM, White PA. Norovirus recombination. Journal of General Virology. 2007;**88**(12):3347-3359. DOI: 10.1099/vir.0.83321-0

[28] Yu Y, Cai H, Hu L, Lei R, Pan Y, Yan S, Wang Y. Molecular epidemiology of oyster-related human noroviruses and their global genetic diversity and temporal-geographical distribution from 1983 to 2014. Applied and Environmental Microbiology. 2015;**81**(21):7615-7624. DOI: 10.1128/AEM.01729-15

[29] Kearse M, Moir R, Wilson A, Stoneshavas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;**28**(12):1647. DOI: 10.1093/bioinformatics/bts199

[30] Kroneman A, Vennema H, Deforche K, Avoort H, Penaranda S, Oberste M, Vinjé J, Koopmans M. An automated genotyping tool for enteroviruses and noroviruses. Journal of Clinical Virology. 2011;**51**(2):121-125. DOI: 10.1016/j.jcv.2011.03.006

[31] Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution. 2011;**28**(10):2731. DOI: 10.1093/molbev/msr121

[32] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 2007;**33**(Database Issue):501-504. DOI: 10.1093/nar/gki025

[33] Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. Journal of Virology. 1999;**73**(1):152-160. Available from: http://jvi.asm.org/content/73/1/152.long

[34] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L. Swiss-model: Modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research. 2014;**42**(Web Server Issue):W252. DOI: 10.1093/nar/gku340

[35] Singh BK, Koromyslova A, Hefele L, Gürth C, Hansman GS. Structural evolution of the emerging 2014/15 GII. 17 noroviruses. Journal of Virology. 2016;**40**:1-7. DOI: 10.1016/j.meegid.2016.02.026

[36] DeLano WL. PyMOL. San Carlos: DeLano Scientific; 2002. p. 700. Available from: https://pymol.org/2/

[37] Eden J-S, Tanaka MM, Boni MF, Rawlinson WD, White PA. Recombination within the pandemic norovirus GII. 4 lineage. Journal of Virology. 2013;**87**(11):6270-6282. DOI: 10.1128/JVI.03464-12

[38] Lu J, Fang L, Zheng H, Lao J, Yang F, Sun L, Xiao J, Lin J, Song T, Ni T. The evolution and transmission of epidemic GII. 17 noroviruses. Journal of Infectious Diseases. 2016;**214**(4):556-564. DOI: 10.1093/infdis/jiw208

[39] de Graaf M, van Beek J, Koopmans MP. Human norovirus transmission and evolution in a changing world. Nature Reviews Microbiology. 2016;**14**(7):421-433. DOI: 10.1038/nrmicro.2016.48

[40] Montazeri N, Goettert D, Achberger EC, Johnson CN, Prinyawiwatkul W, Janes ME. Pathogenic enteric viruses and microbial indicators during secondary treatment of municipal wastewater. Applied and Environmental Microbiology. 2015;**81**(18):6436-6445. DOI: 10.1128/AEM.01218-15

[41] Myrmel M, Lange H, Rimstad E. A 1-year quantitative survey of noro-, adeno-, human boca-, and hepatitis E viruses in raw and secondarily treated sewage from two plants in Norway. Food and Environmental Virology. 2015;**7**(3):213-223. DOI: 10.1007/s12560-015-9200-x

[42] Prevost B, Lucas FS, Ambert-Balay K, Pothier P, Moulin L, Wurtzer S. Deciphering the diversities of astroviruses and noroviruses in wastewater treatment plant effluents by a high-throughput sequencing method. Applied and Environmental Microbiology. 2015;**81**(20):7215-7222. DOI: 10.1128/AEM.02076-15

[43] Yee EL, Palacio H, Atmar RL, Shah U, Kilborn C, Faul M, Gavagan TE, Feigin RD, Versalovic J, Neill FH. Widespread outbreak of norovirus gastroenteritis among evacuees of Hurricane Katrina residing in a large "megashelter" in Houston, Texas: Lessons learned for prevention. Clinical Infectious Diseases. 2007;**44**(8):1032-1039. DOI: 10.1086/512195

[44] Cho H, Lee S, Kim W, Lee J, Park P, Cheon D, Jheong W, Jho E, Lee J, Paik S. Acute gastroenteritis outbreaks associated with ground-waterborne norovirus in South Korea during 2008-2012. Epidemiology and Infection. 2014;**142**(12):2604-2609. DOI: 10.1017/S0950268814000247

[45] Arvelo W, Sosa SM, Juliao P, López MR, Estevéz A, López B, Morales-Betoulle ME, González M, Gregoricus NA, Hall AJ. Norovirus outbreak of probable waterborne transmission with high attack rate in a Guatemalan resort. Journal of Clinical Virology. 2012;**55**(1):8-11. DOI: 10.1016/j.jcv.2012.02.018

[46] Wang X, Yong W, Shi L, Qiao M, He M, Zhang H, Guo B, Xie G, Zhang M, Jin M. An outbreak of multiple norovirus strains on a cruise ship in China, 2014. Journal of Applied Microbiology. 2016;**120**(1):226-233. DOI: 10.1111/jam.12978

[47] Tao Z, Xu M, Lin X, Wang H, Song L, Wang S, Zhou N, Zhang D, Xu A. Environmental surveillance of genogroup I and II noroviruses in Shandong Province, China in 2013. Scientific Reports. 2015;**5**:17444. DOI: 10.1038/srep17444

[48] Aw T, Gin KH. Environmental surveillance and molecular characterization of human enteric viruses in tropical urban wastewaters. Journal of Applied Microbiology. 2010;**109**(2):716-730. DOI: 10.1111/j.1365-2672.2010.04701.x

[49] Aw T, Gin KH. Prevalence and genetic diversity of waterborne pathogenic viruses in surface waters of tropical urban catchments. Journal of Applied Microbiology. 2011;**110**(4):903-914. DOI: 10.1111/j.1365-2672.2011.04947.x

[50] Kitajima M, Oka T, Haramoto E, Takeda N, Katayama K, Katayama H. Seasonal distribution and genetic diversity of genogroups I, II, and IV noroviruses in the Tamagawa River, Japan. Environmental Science and Technology. 2010;**44**(18):7116-7122. DOI: 10.1021/es100346a

[51] Kitajima M, Haramoto E, Phanuwan C, Katayama H, Furumai H. Molecular detection and genotyping of human noroviruses in influent and effluent water at a wastewater treatment plant in Japan. Journal of Applied Microbiology. 2012;**112**(3):605-613. DOI: 10.1111/j.1365-2672.2012.05231.x

[52] Mans J, Netshikweta R, Magwalivha M, Van Zyl WB, Taylor MB. Diverse norovirus genotypes identified in sewage-polluted river water in South Africa. Epidemiology and Infection. 2013;**141**(02):303-313. DOI: 10.1017/S0950268812000490

[53] Murray TY, Mans J, Taylor MB. Human calicivirus diversity in wastewater in South Africa. Journal of Applied Microbiology. 2013;**114**(6):1843-1853. DOI: 10.1111/jam.12167

[54] Kiulia N, Mans J, Mwenda J, Taylor M. Norovirus GII. 17 predominates in selected surface water sources in Kenya. Food and Environmental Virology. 2014;**6**(4):221-231. DOI: 10.1007/s12560-014-9160-6

[55] de la Noue AC, Estienney M, Aho S, Perrier-Cornet J-M, de Rougemont A, Pothier P, Gervais P, Belliot G. Absolute humidity influences the seasonal persistence and infectivity of human norovirus. Applied and Environmental Microbiology. 2014;**80**(23):7196-7205. DOI: 10.1128/AEM.01871-14

[56] Arthur SE, Gibson KE. Environmental persistence of Tulane virus—A surrogate for human norovirus. Canadian Journal of Microbiology. 2016;**62**(5):449-454. DOI: 10.1139/cjm-2015-0756

Edited by Ibrokhim Abdurakhmonov

Genotyping, a methodological process of detecting the allelic content of loci in a given genome, helps to reveal differences between and among individuals of comparisons. It helps in diagnostics and treatment of diseases, molecular breeding of agricultural crops, barcoding of unique biological materials, solving forensics issues, and/or controlling the spread of pathogens by tracing the origin of outbreaks. This *Genotyping* book aims to provide readers with an overview of the basics of genotyping process, available approaches and protocols, as well as novel, low-cost, high-throughput whole-genome genotyping tools and genotype data handling with the examples of genotyping applications in some organisms.

Published in London, UK

IntechOpen

ISBN 978-1-78923-280-6
9 781789 232806