Khalid Rehman Hakeem
Adeel Malik
Fazilet Vardar-Sukan
Munir Ozturk *Editors*

Plant Bioinformatics

Decoding the Phyta

Springer

Plant Bioinformatics

Khalid Rehman Hakeem • Adeel Malik
Fazilet Vardar-Sukan • Munir Ozturk
Editors

Plant Bioinformatics

Decoding the Phyta

Springer

Editors
Khalid Rehman Hakeem
Department of Biological Sciences
Faculty of Science
King Abdulaziz University
Jeddah, Saudi Arabia

Adeel Malik
Department of Microbiology
and Molecular Biology
Chungnam National University
Daejeon, South Korea

Fazilet Vardar-Sukan
Department of Bioengineering
Faculty of Engineering
Ege University
Bornova, İzmir, Turkey

Munir Ozturk
Centre for Environmental Studies
& Botany Department
Ege University
Bornova, İzmir, Turkey

ISBN 978-3-319-67155-0 ISBN 978-3-319-67156-7 (eBook)
https://doi.org/10.1007/978-3-319-67156-7

Library of Congress Control Number: 2017958881

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

(780–850 CE)
To the founder of classical algebra Muhammad ibn Musa al-Khwarizmi. He was a famous mathematician, astronomer, geographer, and scholar in the "House of Wisdom" (Dār al-Ḥikma) in Baghdad under the caliphate of al-Maʾmūn and is popularly known as the "father of algebra."

Foreword

Bioinformatics by etymological definition is the combination of data and knowledge regarding functional biological processes, particularly recombinant DNA. This new systematic approach permits science to interface between real and abstract knowledge. In the last few decades, bioinformatics has become more important due to the advent of modern information technology.

Bioinformatics encompasses the integration of engineering, mathematics, and statistics along with computer science in order to interpret and understand biological data. However, the type of biological data to be analyzed will determine the level of bioinformatics to be used. For instance, conventional bioinformatics can be used to analyze nucleotides and/or DNA sequences, while more complex structural bioinformatics is mainly used to analyze protein structure and function. Thus, bioinformatics can be a powerful tool for the development of biotechnology and industrial processes. Moreover, bioinformatics can provide a better understanding of and faster solutions to problems in pharmaceutical, medical, agricultural, and environmental fields, among others. Likewise, bioinformatics is a reliable, cost-effective approach to expensive laboratory processes insofar as it is able to predict outcomes through mathematical/statistical modeling of scientific research.

Bioinformatics facilitates the integration of different molecular techniques with high production processes in reduced time, thereby making engineering and industrial processes more feasible. Thus, one advantage in using bioinformatics is that it allows the process to be more reliable and predictable. Despite the reliability of bioinformatics, however, its application will depend on the types of biological tools and/or approaches used in bioinformatics.

The application of bioinformatics also depends on the type of biological system to be analyzed or interpreted. This dependency is analogous to "the chicken and the egg," where bioinformatics relies upon conventional statistics and/or stochastic mathematics that only analyze related variables. Nonetheless, with the advent of new scientific approaches such as epigenetics, synthetic biology, microarrays, and other molecular biology techniques that require multidimensional relationships, there is a need for more complex mathematical-computational modeling. An example of such complex modeling is neuronal modeling, which is able to integrate different

variables within a network system. These new, advanced biological approaches require the use of engineering, computational, and molecular genetics techniques to develop biological systems with specific functions. Bioinformatics can be used to understand the operation of genetic systems by designing and integrating the genetic parts and the physiology from a wide range of different organisms.

Undoubtedly, bioinformatics will evolve with the progress of science, especially as regards biological systems, and the future of bioinformatics will depend on the strength of the scientific knowledge base. New fields of science such as nanotechnology can make a strong impact on both biological sciences and bioinformatics. For instance, if we look at DNA as composed of hundreds of atoms that function as a microprocessor transducing millions of input and output information, we can investigate the effect of energy and/or current (i.e., electrical fields) on the structure and function of large biological systems (e.g., nucleotide sequences, proteins). Therefore, it is expected that bioinformatics in the future will provide a better understanding of biological systems not only at the molecular level but also at the atomic level. This will require the use of a more complex mathematical-computational approach such as neuronal modeling, a better understanding of stereochemistry and biophysics, and a better ability to standardize genetic components. Furthermore, in addition to implementing the criteria mentioned above, following up with experimental laboratory work will guarantee the success of bioinformatics for understanding complex biological systems.

Bogota, Colombia Raul Cuero

Preface

Considering the immense significance of plants for life on Earth, the major foci of research in modern plant biology have been to (a) select plants that best fit the purposes of humans; (b) develop crop plants superior in quality, quantity, and farming practices when compared to natural (wild) plants; and (c) explore strategies to help plants to adapt biotic and abiotic/environmental stress factors. However, the development of methods, technologies, and implementations for a better mechanistic representation of the complex plant system has been increasingly witnessed in current exhaustive plant research. In particular, with the advancement in technology, a huge amount of biological data is emerging from multi-omics approaches aimed at addressing numerous aspects of plant systems under biotic or abiotic stresses. Thus, to decipher plant strategies to combat various stresses, a proper management, analysis, and interpretation of this high-throughput data is required. The field of plant bioinformatics has become a panacea for the highlighted issue where the analysis of the huge data sets available in databases is made possible with specific software. Despite the fact that the field of plant bioinformatics is evolving at a rapid pace, the information on the cross-talks and/or critical digestion of research outcomes in context with plant bioinformatics is scarce.

In view of the above, taking into account authoritative chapters contributed by eminent scientists and researchers in the arena of plant bioinformatics, the current edited volume is aimed to (i) introduce fundamental and applied bioinformatics research in the field of plant life sciences; (ii) enlighten the potential users toward the recent advances in the development and application of novel computational methods available for the analysis and integration of plant omics data; (iii) highlight relevant databases, software, tools, and web resources developed till date to provide ease of access for researchers working to decipher plant responses toward stresses; (iv) present critical cross-talks on the available high-throughput data versus plant bioinformatics, bioinformatical versus experimental analyses of plant small RNAs, bioinformatics significance in the new crop disease emergence and biotic/abiotic stress tolerance, and functional genomics approaches in plant research; (v) provide the role of different areas of bioinformatics such as genomics, proteomics, systems biology, etc. in agriculture; and (vi) summarize challenges and provide

recommendations to overcome the limitations in employing computational methods to solve problems in the current context.

We believe that the present volume could be of great interest among research students and the teaching community and could also be used as a reference material by professional researchers.

We are highly grateful to all our contributors for readily accepting our invitation and for not only sharing their knowledge and research but for venerably integrating their expertise in dispersed information from diverse fields in composing the chapters and enduring editorial suggestions to finally produce this venture. We greatly appreciate their commitment. We are also thankful to Prof. Raul Cuero for writing the foreword. Last but not the least, we are also thankful to the Springer International team for their generous cooperation at every stage of the book's production.

Jeddah, Saudi Arabia — Khalid Rehman Hakeem
Daejeon, South Korea — Adeel Malik
Bornova, İzmir, Turkey — Munir Ozturk
Bornova, İzmir, Turkey — Fazilet Vardar-Sukan

Contents

The original version of this book was revised. An erratum to this book can be found at https://doi.org/10.1007/978-3-319-67156-7_19

About the Editors

Khalid Rehman Hakeem, PhD is working as associate professor at the Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia. He obtained his MSc (environmental botany) as well as PhD (botany) from Jamia Hamdard, New Delhi, India, in 2006 and 2011, respectively. He conducted his postdoctoral research in the fields of plant ecology and plant biotechnological studies from Universiti Putra Malaysia from 2012 to 2013. Dr. Hakeem has more than 9 years of teaching and research experience in plant ecophysiology, biotechnology and molecular biology, plant bioinformatics, plant-microbe-soil interactions, as well as environmental sciences. A recipient of several fellowships at both national and international levels, Dr. Hakeem has so far edited and authored more than 20 books with international publishers. He has also to his credit more than 100 research publications in peer-reviewed international journals, including 35 book chapters with international publishers. He is also the editorial board member and reviewer of several high-impact international journals. Dr. Hakeem is currently engaged in studying the plant processes at ecophysiological as well as proteomic levels.

Adeel Malik, PhD is currently working as a research professor at Department of Microbiology and Molecular Biology, Chungnam National University, Daejeon, South Korea. He obtained his PhD (2009) from the Department of Biosciences, Jamia Millia Islamia (JMI), New Delhi, India. During his PhD, he developed computational methods for the prediction of carbohydrate binding sites in proteins using sequence and evolutionary information. He obtained his postdoctoral fellowship from the School of Computational Sciences, Korea Institute for Advanced Study (KIAS), Seoul, South Korea (2011–2012). As a part of his research, he investigated plant lectin-carbohydrate interactions via community-based network analysis by using glycan array data. He worked as an assistant professor at the School of Biotechnology, Yeungnam University, South Korea, and later at Perdana University Centre for Bioinformatics (PU-CBi), Malaysia. His research interests include developing computational methods for studying protein-carbohydrate interactions and applying bioinformatics approaches to explore the role of glycogen in various biological processes. He has published about 19 research articles in high-impact journals including 3 book chapters.

Munir Ozturk, PhD, DSc was born in Kashmir (1943) and holds PhD + DSc degrees in Ecology & Environmental Sciences from Ege University, Turkey. He is the author of several papers on ecological studies as well as biomonitoring in different habitats and is a member in the editorial board of as well as reviewer in several journals. Dr. Öztürk has received fellowships from the Alexander von Humboldt Foundation and Japanese Society for the Promotion of Science. He has worked at the University of Chapel Hill, North Carolina at Chapel Hill using the grant from the National Science Foundation, USA; and as well as a Consultant Fellow at the Faculty of Forestry, University Putra Malaysia, and as a "Distinguished Visiting Scientist" at the ICCBS, Karachi University, Pakistan. His fields of scientific interest are; Plant Ecophysiology, the; Conservation of Plant Diversity; Biosaline Agriculture and Crops, Pollution, Biomonitoring, and Medicinal/Aromatic Plants. The current number of his publications lies around 450. These include over 40 books, nearly 55 book chapters, more than 170 papers in impact factor journals, and more than 150 presentations in "National and International Conferences, Workshops, and Symposia." He has acted as guest editor for several international journals.

Fazilet Vardar-Sukan, PhD is a chemical engineer who graduated from Ege University, İzmir, Turkey, with a PhD in biochemical engineering from University College London, UK. She has 35 years of teaching and research experience and has over 150 publications and nearly 1,500 citations in the fields of scale-up and mass and momentum transfer in bioreactors, reutilization of agro-industrial waste through bio-industries, and R&D & I management and support. She is the founding head of the Bioengineering Department at Ege University since 2000 and is a pioneer in the field in Turkey. She has been involved in 21 EU-supported projects and 27 national and international projects supported by different national and international funds or industries in the fields of biotechnology, R&D & I support, and science and society, working as either a coordinator, partner, or researcher. She is a referee in numerous scientific journals. She is the holder of the Turkish Scientific and Technological Council Incentive Award in Bioengineering in 1989 and has three patent applications, one of them being a PCT. She is currently the chairperson of the National Biotechnology Strategy Committee of the Ministry of Science, Industry and Technology of Turkey.

Plant Bioinformatics: Next Generation Sequencing Approaches

L.F. De Filippis

Contents

1 Introduction

1.1 Short History

The origin of the discipline of 'molecular biology' or for that matter the area of 'bioinformatics' is difficult to determine. In the short time of 75–100 years, we intend to cover only milestones in key discoveries, more or less in chronological order. We begin with the reported discovery of DNA by Johann Friedrich Miescher in 1869, who discarded the possibility that it might be related to heredity. Jensen and Evans (1935) positioned a single amino acid (a terminal phenylalanine) in the insulin molecule, and the sequence of insulin was further characterised by Sanger's group in 1951 (Sanger et al. 1955). Franklin and Gosling described fundamental research on the molecular and crystalline structure of DNA (Franklin and Gosling 1953a, b), and Watson and Crick interpreted this data to produce a model of the bonding and structure of the DNA molecule (Watson and Crick 1953). Brown et al. (1955) described pig and sheep insulin, and Kendrew determined the first three-dimensional (3D) structure of a protein (Kendrew et al. 1958). Muirhead and Perutz (1963) described the amino acid sequence of haemoglobin, and Dayhoff et al. (1981) produced the first genetic atlas of protein sequence and structure. Protein structure was a complex puzzle, and complete amino acid sequences required the

L.F. De Filippis (✉)
School of Life Sciences, Faculty of Science, University of Technology Sydney (UTS),
P.O. Box 123, Broadway/Sydney, NSW 2007, Australia
e-mail: lou.defilippis@uts.edu.au

K.R. Hakeem et al. (eds.), *Plant Bioinformatics*,
https://doi.org/10.1007/978-3-319-67156-7_1

resolution of many different challenges, as a result the 3D structure of insulin would not be known for another 15 years (Adams et al. 1969). This was the era of manual sequencing projects that could last decades, and the sequence of the first enzyme (a ribonuclease) was determined after 8 years of research (Hirs et al. 1960). In the 1970s, the first sequence of the 24 base pair (bp) *lac* operator (Gilbert and Maxam 1973) and the viral genome of the bacteriophage MS2 (Fiers et al. 1976) were published. Projects of this period paved the way for 3D structures of proteins, but without the sequence information, the electron density maps could not have been interpreted (Wyckoff et al. 1967).

The term bioinformatics was apparently used early in 1977 by Hogeweg when describing her field of research at the University of Utrecht (Hogeweg 1978; Hogeweg and Hesper 1978). The discipline as a field of biology had little impact on molecular biology for another 10 years, although Bedbrook et al. (1977) was instrumental in adopting the phrase 'plant molecular genetics'. Bioinformatics appeared to grow almost by necessity from the needs of researchers to access and analyse, at first, biomedical data which was increasing at an alarming rate. The rapid collection of biomedical and genetic data was a direct consequence of a series of chemical and biological techniques that yielded large quantities of basic molecular 'sequence' information. As well as these advances, the development of algorithms and computational resources necessary to analyse, manipulate and store these growing quantities of data was crucial (Attwood et al. 2011). Together, the integration of these two disciplines (or areas of science) gave birth to the field of bioinformatics. But the history from about 1970 is complex and developed along a number of pathways, including the rise and spread of large volumes of data and its distribution worldwide. During this period, some of the databases developed to store the accumulating data, and some of the organisations and infrastructure created, attempted to put these databases on a more solid financial footing (Karahoca et al. 2012).

Up until the 1970s the sequencing of nucleic acids had remained a problem, due to issues related to molecular size and ease of purification. It was possible to sequence some tRNAs, because they were short (many smaller than 100 nucleotides) and tRNA molecules could be purified with some effort. Chromosomal DNA molecules, however, were in a different category containing many millions of nucleotides. In the mid-1970s, the longest fragment that could have been reliably sequenced in a single experiment was about 150–200 base pairs (bp), and fragments of around half a million base pairs per chromosome were beyond the methods of the time. During the late 1970s, however, Sanger et al. (1977b) had developed a technology (to be known as the 'Sanger Method') that made it possible to work with much longer nucleotide fragments and allowed the complete sequencing of the 5,386 bases long single-stranded bacteriophage X174 (Sanger et al. 1977a, 1978). The 'Sanger Method' and technologies codeveloped by Maxam and Gilbert (1977) permitted the efficient and accurate sequencing of even longer sequences. These were landmark achievements, providing the first evidence of the non-universality of the genetic code and overlapping sequences in genes (Sanger 1988; Dodson 2005). In 1986 the first RFLP map of a plant genome was published by Bernatsky

and Tanksley (1986). But it was automation, storage, improved techniques and distribution of results from the mid-1980s onwards that significantly increased biological and genetic productivity.

In view of the 'high-throughput' sequences and equipment present today, these long time periods now seem almost unbelievable. The challenges and potential of 'sequencing technology' to aid in our understanding of the biochemical functions and evolutionary histories of nucleic acids and proteins were critical to molecular biology. In the following 10 years, time-consuming manual sequencers were replaced with automated sequencers, which increased the rate of information available (Ronni and Hichem 2011). The final link in the technologies was the ability to handle very large amounts of information and the use of computers to help analyse and store sequence and structural data. Initially the idea that molecular information could be collected and distributed in electronic form was not only new but also posed significant challenges. Consider, for a moment, the concepts we take for granted today; e.g. e-mail, the Internet and the World Wide Web had not yet emerged. Therefore there was no easy way to distribute data from a central database, other than by posting computer tapes and/or discs to users. This model of data distribution was difficult and slow, was costly and led some of the first databases to adopt pricing and/or data-sharing policies that threatened to drive away many potential users.

The last 30–35 years have been extremely important, giving rise to many new molecular structures and DNA sequences, to new categories of RNA and protein families and finally to new databases to store them. This period of discovery has been remarkable as two major developments have taken place only recently, i.e. the World Wide Web and high-throughput DNA sequencing. Together, these two technologies would promote an overwhelming explosion of biological data but would also spur their global dissemination. Numerous organism-specific databases to store the emerging genomic data were published and placed on the Web. Yet some scientists questioned the value of this genomic *gold rush*, and its usefulness was not entirely clear as the majority of data was mostly non-coding and impossible to interpret. The assumed hidden *genetic treasure troves* in the data were beginning to look impossible to find and uninspiring and perhaps suggested that molecular biology had entered a somewhat vague era, much like *high-tech coin collecting* (Hunter 2006).

We have come a very long way in a story spanning not much more than 75 years, where now bioinformatics has given us 'complete' catalogues of DNA and protein sequences, including genomes and proteomes of organisms across biology. It has furnished the requisite software to help analyse molecular genetic data on an unprecedented level. It has yielded the possibilities to understand more about evolutionary processes and ultimately a great deal more about plants, their productivity, diseases, metabolic activity, physiology, biochemistry and genetics. Therefore, a definition of 'bioinformatics' I would like to use is that bioinformatics builds mathematical and computer models of biological plant processes to infer relationships between components of a more complex system.

1.2 Next-Generation Sequencing (NGS)

Advancements in high-throughput next-generation sequencing (NGS) technologies and the fast growing volume of biological data meant that a diversity of data sources (databases and Web servers) have been created to facilitate data management, accessibility and analysis. Bioinformatic tasks mean that researchers often need to be skilful in using the data and in extracting information for further analysis and more detailed and more specific information searches. Data integration in bioinformatics aims to establish automated and efficient ways to integrate large, heterogeneous biological datasets from multiple sources. However, these aims are difficult to achieve, as data sources can be heterogeneous in dissemination formats (Zhang et al. 2011). Ultra high-throughput sequencing, also known as 'deep sequencing' and 'high-throughput sequencing' or as I prefer to use 'next-generation sequencing' (NGS), is beginning to impact heavily on the study of biology and genetics and has plant and agricultural implications. This technology has reduced the cost and increased the throughput of genomic sequencing by more than three or four orders of magnitude in just a few years, a trend which is almost certain to accelerate in the next decades (Metzker 2010). For example, using NGS, it is now possible to discover novel disease-causing mutations (Ley et al. 2008) and detect traces of plant pathogenic microorganisms within plant cells and tissues (Isakov et al. 2011). The amount of data produced by a single ultrahigh-throughput sequence run is often very large and can reach millions of reads of various lengths per experiment (Mardis 2008). The storage, processing, querying, parsing, analysing and interpreting such a large amount of data is a significant task containing many problems and challenges (Koboldt et al. 2010).

NGS technologies are evolving with increase in data efficiency and throughput (Mardis 2008). This rate of change and improvement is accompanied by a variety of different sequencing platforms, having both great similarities and many differences (Ekblom and Galindo 2011; Egan et al. 2012). The initial step in the deep, high-throughput sequencing process is random fragmentation of the nucleotides of interest, in order to increase output by simultaneously sequencing millions of fragments. These template fragments can then either undergo clonal amplification, in which they are ligated with adapter molecules and amplified using PCR (polymerase chain reaction) (Roche, Illumina, Life Technology), or the adapter fragments can be used as the sequencing templates themselves (single-molecule templates) (Pacific BioSciences, BioRad) (Salgotra et al. 2014).

(a) *Clonal amplified* template preparations require higher amounts of purified initial DNA. Since the technique relies on PCR amplification, errors might be introduced to the target before the sequencing process is initiated. The amount of introduced errors is related to the fidelity of the DNA polymerase used (Chan 2009). These potential background errors could be considered actual sequence variants in 'downstream' analysis. PCR utilisation might also result in amplification bias and misrepresentation of high GC content DNA, requiring additional assessment hampering uniformity of results (Chiu et al. 2010). Simultaneously

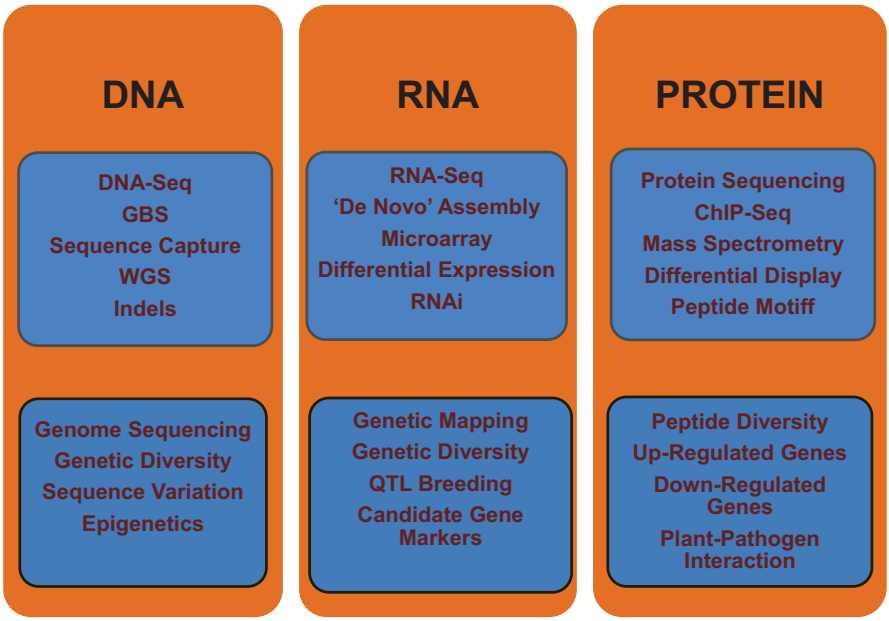

Fig. 1 Overview of NGS applications for plant genetics and breeding. Three different sources of initial starting material are separated (*top*), sequencing and associated technologies (*centre*) and applications (*bottom*) (Data extracted from Ekblom and Galindo 2012; Barabaschi et al. 2016)

sequencing templates are further complicated by potential different extension rates that cause asynchronous sequencing, resulting in a high background.

(b) *Single-molecule* template sequencing does not require PCR amplification, thus making it an appropriate tool for use in quantification experiments and/or in cases where the initial amounts of DNA are low. Because sequencing is performed on single molecules and are inferred from extremely weak signals, the correcting effect of 'simultaneous sequence template' results in high error rates (Schadt et al. 2010a, b). Therefore a high-sequencing fidelity technology must be used (Metzker 2010).

Downstream uses of next-generation sequencing (NGS) include (Fig. 1):

(a) *Whole-genome 'shotgun' sequencing (WGS)*: whole-genome assembly and genome comparisons within and between plant species and cultivars
(b) *Targeted region sequencing (Exome-Seq)*: reference mapping, nucleotide mutations and especially single nucleotide polymorphism (SNP) variant calling
(c) *Whole transcriptome sequencing (RNA-Seq)*: expression quantification and novel splice junction detection (i.e. exons and introns) in plants
(d) *Chromatin immunoprecipitation sequencing (ChIP-Seq)*: regional plant DNA (chromatin) and plant protein interaction-associated sequencing
(e) *Random regions sequenced across samples (RAD-Seq)*: next-generation studies in plant variant detection for population genetics

NGS has already had considerable impact on three primary areas of plant science and agriculture and will continue to produce large amounts of information with various impact and understanding in these disciplines. These three areas of plant biology are briefly discussed below.

1.3 Molecular Markers

Functionally characterised sequences can be identified from next-generation DNA sequences and functional markers (FMs) for important traits have been developed with increasing ease. FMs have been developed from polymorphic sites within genes that causally affect target trait variation, i.e. based on metabolic functional characterisation of the polymorphisms (Sleator 2010) and/or allelic variants of functional genes. Linkage disequilibrium (LD)-based association mapping and homologous recombinants have been developed to identify so-called 'perfect' markers for use in crop improvement. Compared with many other molecular markers, FMs derived from the functionally characterised sequences of genes, and their use provide opportunities to develop high-yielding plant genotypes resistant to various stresses and diseases quickly. Recent progress in the area of plant molecular biology and genomics has the potential to initiate a new 'Green Revolution', which is of vital importance for the development of much improved crop germplasm (Gupta 2008). Exact linkage of markers and genes to traits must lead to more efficient plant selection, and genomic technologies are being applied to the improvement of crop plants with encouraging results (Schnable 2013; He et al. 2014). The genomic revolution, which started in the 1990s, has greatly improved our understanding of the genetic make-up of a wide group of living organisms, including now several crop plant species. Complete genome sequences of *Arabidopsis* (Arabidopsis Genome Initiative 2001), rice (International Rice Genome Sequencing Project 2005) and soybean (Schmutz et al. 2010) have provided the basis for understanding the relationships amongst genes, proteins and phenotypes. Complete genomic sequences of more plant genomes in the near future should improve further information use in crop breeding programmes significantly (Henry 2012; Michael and Jackson 2013) (Table 1).

For about 25–30 years, DNA markers have been the most widely used molecular markers in crop improvement, owing to their abundance and polymorphism. Most of these markers can be selectively neutral because they are usually located in non-coding and non-regulatory regions of DNA (McKay and Latta 2002). The first plant DNA markers were based on restriction fragment length polymorphism (RFLPs) (Bernatsky and Tanksley 1986), and early hybridisation-based, isotope-labelled RFLP techniques were difficult and time-consuming, eventually replaced by safer, less complex and more cost-effective PCR-based markers. Molecular markers now include:

Table 1 Whole plant genomes sequenced or near sequenced (Mbp), showing genome size (Mbp), chromosome number and technology used (only data available publically)

Plant	Species	Genome sequenced	Genome size	Chrom. number	Technology
Cassava	*Manihot esculenta*	533	760	8	454 Sanger
Castor bean	*Ricinus communis*	350	400	10	Sanger
Poplar	*Populus trichocarpa*	410	485	19	Sanger
Medic	*Medicago truncatula*	214	307	8	Sanger
Lotus	*Lotus japonica*	315	472	6	Sanger
Soy	*Glycine max*	950	1100	20	Sanger
Apple	*Malus x domestica*	603	742	8	454 Sanger
Strawberry	*Fragaria vesca*	209	240	7	454 Illumina
Peach tree	*Prunus persica*	227	269	8	Sanger
Cucumber	*Cucumis sativus*	203	880	14	İllumina Sanger
Arabidopsis	*Arabidopsis thaliana*	115	125	5	Sanger
Arabidopsis	*Arabidopsis lyrata*	207	207	8	Sanger
Papaya	*Carica papaya*	135	367	9	Sanger
Chocolate	*Theobroma cacao*	326	430	10	454 Sanger
Sweet orange	*Citrus sinensis*	319	380	9	454 Sanger
Mandarin	*Citrus clementina*	296	370	9	Sanger
Eucalypt tree	*Eucalyptus grandis*	641	650	22	Sanger
Grape	*Vitis vinifera*	715	416	19	Sanger
Potato	*Solanum tuberosum*	727	844	12	454 İllumina Sanger
Sorghum	*Sorghum bicolor*	730	735	10	Sanger
Corn (maize)	*Zea mays*	2300	2650	10	Sanger
Foxtail millet	*Setaria italica*	405	515	–	Sanger
Rice	*Oryza sativa*	389	400	12	Sanger
Grass	*Brachypodium distachyon*	272	355	5	Sanger
Moss	*Selaginella moellendorffii*	215	86	27	Sanger
Slime mould	*Physcomitrella patens*	480	518	27	Sanger

Data modified from Llaca (2012)

(a) RFLPs and other Southern blot-based markers (Botstein et al. 1980)
(b) PCR-based markers, such as random amplification of polymorphic DNA (RAPD) (Williams et al. 1990), random amplification of microsatellite DNA (RAMP) (Wu and Tanksley 1993), amplified fragment length polymorphism (AFLP) (Vos et al. 1995), microsatellite or simple sequence repeats (SSR) (Powell et al. 1996), sequence characterised amplified regions (Paran and Michelmore 1993) and cleaved amplified polymorphic sequences (Konieczny and Ausubel 1993)
(c) Sequence-based markers, such as single nucleotide polymorphism (SNP) (Gupta 2008), which are now the most important and can be applied to a large number of plant species

The majority of the molecular markers have been developed either from genomic DNA libraries (RFLPs and SSRs) or from random PCR amplification of genomic DNA (RAPDs, RAMPs) or both (AFLPs). Direct array technology (DArT) however commonly uses SNP as a base (Sansaloni et al. 2011). When some of these markers are used for marker-assisted selection in plant breeding, they have some limitations owing to some markers being dominant, genetic recombination may give rise to false positives and some produce inconsistent results. High-throughput sequencing techniques and technical developments in NGS of plant species have led to an increase in identification of important variations at the single base pair level (Ray and Satya 2014).

1.4 Plant Breeding

A growing global population and shrinking arable land areas require more efficient plant breeding, in terms of the time taken and the costs. Novel strategies assisted by some molecular markers have proven effective for agricultural plant improvements. Fortunately, cutting-edge sequencing technologies of plant genomes result in detecting, with great efficiency and numbers, genetic variations form the base for plant breeding and increase the potential of marker development for important agricultural traits. Transgenic plants containing artificially inserted genes also have significant economic benefit to farming and agriculture. In both the classical and modern (i.e. transgenic) plant breeding approaches, markers are important to accelerate genetic improvement. Although thousands of articles have been published with the term 'marker-assisted selection' (MAS) or 'quantitative trait loci' (QTLs) or 'molecular markers', a large gap still exists between the expectations and actual applications of molecular markers to practical plant breeding (Egan et al. 2012). The term 'next-generation plant breeding' is increasingly becoming popular in crop breeding programmes and in agriculture in general (Schnable 2013; Davey et al. 2011). Being a frontier area of crop science and business, it can gain considerable interest amongst the scientific community and policymakers, and in so doing funds may flow from entrepreneurs and research funding organisations to this extremely important area of plant breeding.

Plant breeding is a continuous attempt to alter genetic architecture of crop plants for efficient utilisation as food, fodder, fibre, fuel or other end use. Although the scientific concepts in plant breeding originated well over 100 years ago, domestication and selection by humans of desirable traits have contributed a great deal to ensure food security (Gepts 2004). During the past few decades, well-supported crop improvement programmes for major crops have started reaping benefits from cutting-edge technologies in the biological sciences, particularly in the form of molecular markers and transgenic crop development. In combination with conventional phenotype-based selection, the current generation of plant breeding practices have developed. Different types of plant molecular markers have been developed and extensively used during the last three decades for identifying linkage between genes and markers, discovering quantitative trait loci (QTLs), pyramiding desired genes and performing marker-assisted foreground and background selection for introgression of desired

traits (Varshney and Tuberosa 2007). However, these markers have been primarily based on electrophoretic separation of DNA fragments, which limits detection of genetic polymorphism. In large plant breeding populations, traditional genotyping may take up to several months depending on marker systems, adding more cost to breeding programmes. Next-generation plant breeding aims to develop more efficient technologies and programmes for low-cost, high-throughput genotyping and screening of large populations in a shorter time (Varshney et al. 2009).

1.5 Molecular Ecology

All biological disciplines that depend on DNA sequence data have been fundamentally changed in the last few years due to the development and emergence of next-generation sequencing (NGS); and our knowledge of biology, particularly evolutionary genomics, has grown. NGS creates huge amounts of data, presenting many problems to computational biologists, bioinformaticians and end users (especially ecologists and taxonomists) endeavouring to assemble and analyse NGS data. A comprehensive discussion of these challenges is outside the scope of this review, but several papers in these disciplines address some of these issues and possible strategies in dealing with them (e.g. Grover et al. 2012; Ilut et al. 2012; Kvam et al. 2012). NGS data is very cost-effective, and molecular ecologists are now starting to take advantage of sequencing information and embracing the discipline of 'ecological genomics' (Gilad et al. 2009). By shifting genomics from laboratory-based studies of model plant species towards studies of natural populations of ecologically important plants, researchers can now start to address important ecological and evolutionary questions on a scale and precision that was unrealistic only a few years ago. In the last 30 years, a number of DNA fingerprinting methods such as RFLP, RAPD, RAMP, AFLP, SSR and DArT primarily used in marker development for molecular plant breeding have found their role in ecology, genetic diversity and species and population genetics. However, it remains a daunting task to identify highly polymorphic and closely linked molecular markers for targeted traits in molecular marker-assisted population genetics. NGS technology is far more powerful than any existing genetic DNA fingerprinting methods mentioned above in generating DNA markers and continues to present problems and challenges in plant molecular ecology.

In this chapter, we provide an overview of many representative Web-based resources available for use in NGS plant research, with particular emphasis on recent progress related to crop species and crop improvement. We describe sequence-related resources, such as molecular markers, whole-genome platforms and protein-coding and non-coding transcripts, and provide recent sequencing technology updates. We then review resources important for genetic map-based approaches to plant breeding (e.g. QTL analyses, TILLING, near-isogenic lines and allele mining) and population genetic diversity studies (e.g. percent polymorphism, genetic differentiation, heterozygocity) (Travis et al. 2002). We also describe the current status of resources and technologies for transcriptomics, proteomics and metabolomics; however, some of these fields are more comprehensively described in other literature listed

(Akula et al. 2009; Zhao and Grant 2010). NGS applications have been divided into technologies based on starting plant material, like DNA, RNA and protein. This appears to us a logic separation as in many investigations, usually only one of these extracted plant metabolites is readily available for use. Resources for use in NGS research will be discussed, and the integration of computer programmes and datasets (i.e. data banks) across plant species in comparative genomics are outlined.

Bioinformatics and Web addresses for plants have been reviewed by a number of authors (Baginsky 2009; Varshney et al. 2009; Mochida and Shinozaki 2010; Jackson et al. 2011; Memon 2012; De Filippis 2013), and this review will basically cover new areas in NGS application studies, and topics which require more detailed explanation have been updated for crop plants. The excellent review by Mochida and Shinozaki (2010) and De Filippis (2013) has provided the framework for this review, and we intend to concentrate on more recent developments and focus on bioinformation and implications in crop improvement and population genetics, although the technology, instrumentation platforms, statistics and computational programmes and databases used with all plants must be covered.

NGS pre-analysis and post-analysis concepts are introduced, and important advance considerations for alignment, assembly and variation detection are discussed. Currently, the deep sequencing user is faced with an abundance of deep sequencing data and analysis tools, both publicly and commercially available. We intend to point out various aspects to be considered when choosing a tool and emphasise the relevant challenges and possible limitations so as to assist the user in picking the most suitable platform. Therefore our focus will be on fundamental concepts of the analysis process and its challenges amid an increasing number of published software programmes and sites. A brief overview is presented of current NGS methods and associated technologies (e.g. microarray and mass spectroscopy), highlighting strengths and possible drawbacks with regard to different applications and different aspects of post-sequencing analysis (e.g. data alignment, assembly, variant detection, RNA interference and bioactive peptides). Finally, we intend to cover areas of further research and conclusions covered from such a broad area of plant molecular biology and bioinformatics.

2 Next-Generation Sequencing, Computer Programmes and Data Banks

2.1 Computers in Molecular Biology

Fundamental mathematical and algorithmic concepts underlying computational molecular biology are now almost completely reliant on computers. Physical and genetic mapping, sequence analysis (including alignment and probability models), genomic re-arrangements, phylogenetic inference, computational proteomics and systemic modelling of the whole cell could not be possible without computers. Bioinformatics being a computer-reliant technology that supports the life sciences means that tools and systems perform a diverse range of functions including data

collection, data mining, data analysis, data management, data integration, simulation, statistics and visualisation. Biologists that simplistically reduce bioinformatics to the application of computers in biology sometimes fail to recognise the rich intellectual content of bioinformatics.

2.1.1 File Formats

Due to the complex nature of biology, there are a wide variety of biological data types, e.g. sequence data, gene expression data, protein-protein interaction data and pathway data (Karasavvas et al. 2004). Data sources store different data types in different formats (Li 2006): flat file (e.g. tab-delimited file), sequence file (e.g. FASTA), structure file (e.g. PSF-protein structure file) and XML file (e.g. KGML-KEGG markup language). Data sources often adopt preferable data formats, even for the same information which often can be incompatible. The most common initial form of computer output format in bioinformatics is either a sequence FASTA file including a numerical quality QUAL score (Ewing and Green 1998) or the FASTQ format. FASTQ is a text-based format for storing both a biological sequence and its corresponding quality score. Both the sequence and quality score are encoded with a single ASCII character for brevity (Cock et al. 2010).

2.1.2 Quality Control of Data

Searching for rare sequence variants is often the primary aim of researchers; however base overexpression and the more common sequence duplication (Gomez-Alvarez et al. 2009), usually an artefact of PCR amplification and other library preparation processes introduce problems. This creates a skewed coverage distribution that may subsequently bias computer models. If these are sequenced, they can profoundly affect 'downstream' analysis unless removed (e.g. clipping). The clipping process removes any tag remnants from the sequence reads eliminating data from reads composed mainly of or even solely of tags. Trimming may also be required to the sequences by removal from either the 5′ or the 3′ ends of a number of bases in the read, and this is especially true for poly-A or poly-T tails.

2.2 Data Analysis

2.2.1 Sequence Alignment

Bioinformatics and molecular biology analyses also often begin with comparing DNA or amino acid sequences by aligning them. Pairwise alignment, for example, is used to measure the similarities between a query sequence and each of those in a database like Basic Local Alignment Search Tool (BLAST) (Sects. 3.4.2, 5.3.1);

BLAST is the most often used bioinformatic tool (Altschul et al. 1990; Camacho et al. 2009) in biology. Evolutionary history amongst sequences can be better reflected, when more than two sequences are aligned, in multiple sequence alignment (MSA). Most alignment analyses involve an initial step of mapping the deep sequencing reads against a reference genome of either the sequenced species or a related organism with sufficient genetic resemblance. This step presents a computational challenge due to the sheer amount of short reads produced in deep sequencing experiments. When choosing a computer alignment tool, one needs to consider the memory and time requirements and limitations and the appropriateness of the tool to the questions being asked.

2.2.2 Multiple Sequence Alignment (MSA)

MSA assumes that the sequences compared are derived from a common ancestral sequence. The process of MSA building is to infer homologous positions between the input sequences, and gaps are placed in the sequences in order to align these in homologous positions. The gaps represent evolutionary events of their own. Gaps (also called indels – Sect. 2.4.2) are caused by either insertions or deletions of nucleotides or amino acids on a particular lineage during evolution. Building an MSA, therefore, is to reconstruct the evolutionary history of the sequences involved. While it is easy to understand that the quality of MSAs affects the quality of phylogenetic tree reconstruction, the effects of MSA quality go far beyond this. Some examples of bioinformatic tools that utilise information extracted from MSAs include profile building in similarity searches (e.g. PSI-BLAST: Altschul et al. 1997), motif/profile recognition (e.g. PROSITE: Hulo et al. 2008), profile-hidden Markov models for protein families/domains (e.g. Pfam: Finn et al. 2010) and protein secondary structure predictions (Pirovano and Heringa 2010). Due to its significant impact on many bioinformatics and molecular evolutionary studies, MSA is one of the most scrutinised bioinformatic fields (Kemena and Notredame 2009; Thompson et al. 2011). However, detailed assessment via MSAs in plants requires great caution and is usually reserved for experienced computer power users.

2.3 Assembly

Assembly refers to the process of piecing together short DNA/RNA sequences into longer ones (e.g. contigs) which are then grouped to form scaffolds for computational reconstructing a sample's genetic code. When the assembly process is performed with the assistance of a reference genome, it is referred to as mapping assembly; however if no reference genome is available, it is called 'de novo' assembly (Slate et al. 2009). Deep sequencing data presents a more compound assembly problem due to higher amounts of sequences that are significantly shorter. Though it adds complexity to the process, the significant increase in throughput enables the

successful realisation of whole plant genome de novo assembly, as reviewed by Barabaschi et al. (2016). Sequencing errors, uneven genome coverage and reads too short to be informative in repeated regions now require the development of a new breed of computational assembly tools designed specifically for short reads.

2.4 Variant Calling

Variant calling in plants refers to the identification of single nucleotide polymorphisms (SNPs), insertions and deletions (indels), copy number variations (CNVs) and other types of structural variations (e.g. inversions, translocations, etc.) in a sequenced sample (Durbin et al. 2004, 2010). Detection of these variants from deep sequencing data requires, in most cases, both a reference genetic sequence to compare the sequence data against (Goodswen et al. 2010) and/or specialised variant calling software that utilise probabilistic methods for correctly inferring variants. The process is complicated by areas of low coverage, sequencing errors, misalignment caused by either low complexity and repeat regions or adjacent variants and library preparation biases (e.g. PCR duplication) (Chan 2009).

2.4.1 Single Nucleotide Polymorphisms (SNP)

After aligning deep sequencing reads against a reference genome, SNPs can be inferred from the results by simply denoting each base that is inconsistent between reference and sample, i.e. the SNPs. Sequencing has for some time now shifted from fragment-based polymorphism identification to sequence-based single nucleotide polymorphism (SNP) identification to expedite marker identification and to increase the number of informative markers. This straightforward inference of mismatches results in a massive amount of alleged SNPs, many of which suffer from inaccuracies such calling a mismatch in the wrong location, homozygosity and heterozygosity discrepancies and even calling a mismatch in the correct location but with the wrong base (Goodswen et al. 2010).

2.4.2 Insertions and Deletions (Indels)

Indels are the second most common type of polymorphisms and the most common structural variant, comprising of short indels (<1,000 kb) and large (>1,000 kb) structural variants (Sect. 2.4.4). Many indels range between 2 and 16 bases in length (Mullaney et al. 2010) (also referred to as micro-indels). Indel frequency has been shown to vary across the genome, with lower rates in conserved and functional regions and increased rates in ‘hot spots’ for genetic variation. The average indel rate is approximately one indel in 5.1–13.2 kb of plant DNA (Albers et al. 2010; Mills et al. 2006). Indel detection is routine and quite easy with NGS, and indels

have been implicated in plant diseases, gene expression and functionality and viral infection and can be used as genetic markers in natural plant populations (Liu et al. 2014).

2.4.3 Structural Variants

Structural variants (Feuk et al. 2006) identified by NGS are defined as genomic alterations that involve segments of DNA that are larger than 1 kb. They include:

(a) Copy number variations (CNV), which are sections of DNA with a variable copy number when compared to a reference genome. Insertions, deletions and duplications are typical CNVs.
(b) Segmental duplications, where several copies of DNA segments that are almost identical (>90%) can appear in a variable number of copies, are also considered a type of CNV.
(c) Inversions, segments in the DNA that are reversed in orientation.
(d) Translocations and inter- and intra- chromosomal location shift in a DNA segment without changing the total DNA content.

2.4.4 Variant Classification

Calling variants using deep sequencing data often results in a multitude of detected variations, even after strict and effective quality filtration. NGS data may reveal thousands to millions of different variations (Imelfort et al. 2009). These variations can result in biological effects through introduction of different amino acids into protein sequences, early termination of coding sequences and alteration of regulatory elements and splice sites. Essential steps following the variant calling process are annotating detected variants and elucidating their effect and biological significance and separating relevant informative variations from neutral, non-functional ones. In a large potential list spanning so many variants, manual annotation of each variant effect is neither feasible nor accurate, and advanced computational methods must be used. These methods are beginning to become available and are grouped into targeted region sequencing (i.e. Exome-Seq) (Fig. 1).

2.5 Data Banks (Data Bases)

The Bioinformatics Links Directory (Brazas et al. 2010) classified almost 1,500 unique publicly available data bank sources. Based primarily on their function, data banks can be classified into the six diverse categories below (most of the sites are cited in Tables 2, 4, 6 and 7):

Table 2 Integrative (Web-based) database sites for general NGS techniques and analysis in plants

Database name	Plant species and purpose	Uniform resource locator (URL)
NCBI – National Center for Biotechnology Information-HOME	Extensive resources for plant, animal, human and microbial genetics, data banks, tools	http://www.ncbi.nlm.nih.gov/
GenBank	Often used data bank for species sequence search and deposition	http://www.ncbi.nlm.nih.gov/genbank/ftp/
ExPASy	Data bank for protein (nucleic acid) sequence and analysis tools	http://www.expasy.org/
BLAST	Tools – blastx (protein), blast n (nucleotide), Magic-BLAST (NGS)	https://blast.ncbi.nlm.nih.gov/Blast.cgi
PSI-BLAST	Protein similarity searches using BLAST (specialised)	http://www.biology.wustl.edu/gcg/psiblast.html
Mascot	Need log-in privilege but free, for advance MS fingerprint analyses	http://www.matrixscience.com/
PROSITE	Database of protein domains, families and functional analysis	http://prosite.expasy.org/
UniProtKB	Free, accessible protein sequence analysis tool	http://www.uniprot.org/
Swiss-Prot	Combined with UniProt and ExPASy protein sequence	http://www.uniprot.org/ http://www.expasy.org/
TrEMBL	Combined with UniProt and ExPASy protein sequence	http://www.uniprot.org/ http://www.expasy.org/
BLOCKS	Protein sequence database – final release 1990; but still active	http://blocks.fhcrc.org/blocks/
Multiple Sequence Alignment Viewer	Tool used after alignment for viewing and editing sequences	http://www.ncbi.nlm.nih.gov/tools/msaviewer/
Sequence Viewer	Tool used after alignment for viewing and editing sequences	http://www.ncbi.nlm.nih.gov/projects/sviewer/
Variation Viewer	Tool used after alignment for simple editing of sequences	https://www.ncbi.nlm.nih.gov/variation/view/
RefSeq	Tool used after alignment for viewing and editing sequences	https://www.ncbi.nlm.nih.gov/refseq/

Table 2 (continued)

Database name	Plant species and purpose	Uniform resource locator (URL)
PDB	Protein Data Bank – worldwide deposition, 3D structure, peptides	http://www.wwpdb.org/
CATH	Protein 3D structure, function and evolution into superfamilies	http://www.cathdb.info/
DIP	Database of interacting proteins, and catalogue, with searches	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
IntAct	Molecular interactions site, free database and search capacity	http://www.ebi.ac.uk/intact/
MINT	Molecular interactions data base, for protein-protein interactions	http://mint.bio.uniroma2.it/
GO	Gene ontology – tools, finder and mapper of function cellular genes	http://go.princeton.edu/
MAPS	Mutations and polymorphisms surveyor; for TILLING, polyploids	http://comailab.genomecenter.ucdavis.edu/index.php/MAPS
SeqAnt	Sequence annotate site, open source for advance WGS	https://omictools.com/sequence-annotator-tool
SuiteMSA	Mass sequence alignment (MSA) and annotation tools	http://evolution.gs.washington.edu/phylip/software.etc2.html#SuiteMSA
indel-Seq-Gen	Sequence simulation download	http://bioinfolab.unl.edu/~cstrope/iSG/#Introduction
dbSNP	SNP finder and annotation	http://www.ncbi.nlm.nih.gov/snp
AnnTools	SNP, indels, SNV, CNV and mutations from microarray data	http://anntools.sourceforge.net/
FFGED	Filamentous fungal gene expression database	http://bioinfo.townsend.yale.edu/
NCBI Taxonomy	Classification nomenclature of described plant organisms	http://www.ncbi.nlm.nih.gov/taxonomy
PRIMER	Population software and gene diversity statistics and indices	http://www.primer-e.com/

Table 3 Next-generation sequencing comparison between functional markers (FMs), genetic molecular markers (GMMs), random DNA markers (RDMs) and genomic selection (GS); detailing important features of each method

Feature	FMs	GMMs	RDMs	GS
Function of markers	Known	Known majority of the time	Unknown majority of the time	Unknown majority of the time
Requirement of sequence data	Genes and EST data essential	Gene and EST data essential	Required for SSRs, SNPs; not required for RFLPs, RAPDs, AFLPs, RAMP, etc.	Sequence for SNP required
Selection of markers	Limited	Limited	Limited	Entire genomic markers
Function of polymorphic sites	Functional motif	Not known	Not known	Not known
Utility in marker-assisted selection	Great, as FMs from polymorphic sites within genes are involved in phenotypic variations	Great, if marker is derived from gene involved in expression of the trait	High for SSRs, SNPs; moderately low for RFLPs, RAPDs, AFLPs, etc.	Less effective in plant breeding
Labour involved	Less	Less	Moderately more	Moderately more for statistical analysis
Number of markers required	Low	Low	High for SSRs and SNPs, moderately low for RFLPs, RAPDs, AFLPs etc.	High
Costs of generation of the markers	Low	Low	Moderately high	High, more markers are required
Utility of markers to functional diversity of genetic resources	High	Moderately low	Moderately low	High

Data modified from Salgotra et al. (2014)