# GENOTYPING BY SEQUENCING FOR CROP IMPROVEMENT

EDITED BY

HUMIRA SONAH | VINOD GOYAL | S. M. SHIVARAJ | RUPESH K. DESHMUKH

WILEY

**Genotyping by Sequencing for Crop Improvement**

# Genotyping by Sequencing for Crop Improvement

*Edited by*

*Humira Sonah*
*National Agri-Food Biotechnology Institute*
*Punjab, India*

*Vinod Goyal*
*CCS Haryana Agriculture University*
*Hisar, India*

*S.M. Shivaraj*
*Laval University*
*Quebec City, QC, Canada*

*Rupesh K. Deshmukh*
*National Agri-Food Biotechnology Institute*
*Punjab, India*

WILEY

This edition first published 2022
© 2022 John Wiley & Sons Ltd

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at http://www.wiley.com/go/permissions.

The right of Humira Sonah, Vinod Goyal, S.M. Shivaraj, and Rupesh K. Deshmukh to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

*Registered Offices*
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

*Editorial Office*
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

*Limit of Liability/Disclaimer of Warranty*
The contents of this work are intended to further general scientific research, understanding, and discussion only and are not intended and should not be relied upon as recommending or promoting scientific method, diagnosis, or treatment by physicians for any particular patient. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of medicines, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each medicine, equipment, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

*Library of Congress Cataloging-in-Publication Data*

Names: Sonah, Humira, editor. | Goyal, Vinod, editor. | Shivaraj, S. M., editor. | Deshmukh, Rupesh editor.
Title: Genotyping by sequencing for crop improvement / edited by Humira Sonah, Vinod Goyal, S. M. Shivaraj, Rupesh K. Deshmukh.
Description: First edition. | Hoboken, NJ, USA : John Wiley & Sons, Inc., [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021046855 (print) | LCCN 2021046856 (ebook) | ISBN 9781119745655 (hardback) | ISBN 9781119745662 (adobe pdf) | ISBN 9781119745679 (epub)
Subjects: LCSH: Genetics–Technique. | Gene mapping. | Genomics. | Plant genomes.
Classification: LCC QK981.45 .G46 2021 (print) | LCC QK981.45 (ebook) | DDC 572.8/62–dc23/eng/20211001
LC record available at https://lccn.loc.gov/2021046855
LC ebook record available at https://lccn.loc.gov/2021046856

Cover Design: Wiley
Cover Image: © Billion Photos/Shutterstock

Set in 9.5/12.5pt STIXTwoText by Straive, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

*Dedicated to the two most eminent agricultural scientists of Canada whose work in plant genomics and breeding helped in food security and inspired many young scientists worldwide.*

Prof. Richard Bélanger
Département de phytologie
Université Laval, Canada

Prof. François Belzile
Département de phytologie
Université Laval, Canada

Dr. Humira Sonah
Dr. Vinod Goyal
Dr. S.M. Shivaraj
Dr. Rupesh K. Deshmukh

# Contents

**List of Contributors** *xv*
**Preface** *xix*

1 **Molecular Marker Techniques and Recent Advancements** *1*
*Dharminder Bhatia and Gagandeep Singh Bajwa*
1.1 Introduction *1*
1.2 What is a Molecular Marker? *2*
1.3 Classes of Molecular Markers *2*
1.3.1 Hybridization-based Markers *2*
1.3.1.1 Restriction Fragment Length Polymorphism (RFLP) *2*
1.3.1.2 Diversity Array Technology (DArT™) *3*
1.3.2 Polymerase Chain Reaction (PCR)-based Markers *3*
1.3.2.1 Simple-Sequence Repeats (SSRs) *3*
1.3.2.2 Sequence-Tagged Sites (STSs) *4*
1.3.2.3 Randomly Amplified Polymorphic DNAs (RAPDs) *4*
1.3.2.4 Sequence Characterized Amplified Regions (SCARs) *6*
1.3.2.5 Amplified Fragment Length Polymorphism (AFLP) *6*
1.3.2.6 Expressed Sequence Tags (ESTs) *6*
1.4 Sequencing-based Markers *7*
1.4.1 Single-Nucleotide Polymorphisms (SNPs) *7*
1.4.2 Identification of SNP in a Pregenomic Era *7*
1.5 Recent Advances in Molecular Marker Technologies *7*
1.5.1 Genotyping-by-Sequencing (GBS) *8*
1.5.2 Whole-Genome Resequencing (WGR) *9*
1.5.3 SNP Arrays *10*
1.5.4 Kompetitive Allele-Specific PCR (KASP™) *11*
1.6 SNP Databases *12*
1.7 Application of Molecular Markers *13*
1.7.1 Application of Molecular Markers in Crop Improvement *13*
1.7.2 Role of Molecular Markers in Germplasm Characterization *14*
1.7.3 Deployment of Molecular Markers in Plant Variety Protection and Registration *14*
1.8 Summary *15*
References *15*

**2 High-throughput Genotyping Platforms** *22*
*Sandhya Sharma, Kuldeep Kumar, Kishor Tribhuvan, Rita, Sandeep Kumar, Priyanka Jain, Swati Saxena, Joshita Vijayan, Harsha Srivastava, and Kishor Gaikwad*
2.1 Introduction *22*
2.2 SNP Genotyping Platforms *23*
2.2.1 SNP Genotyping Versus SNP Discovery *23*
2.2.2 Types of SNP Genotyping Platforms *24*
2.2.2.1 Allelic Discrimination *24*
2.2.2.2 Allelic Detection *30*
2.2.3 Custom Assay Technologies *34*
2.2.4 Summary *34*
References *35*

**3 Opportunity and Challenges for Whole-Genome Resequencing-based Genotyping in Plants** *38*
*Surbhi Kumawat, Gaurav Raturi, Pallavi Dhiman, Sreeja Sudhakarn, Nitika Rajora, Vandana Thakral, Himanshu Yadav, Gunashri Padalkar, Yogesh Sharma, Vinaykumar Rachappanavar, and Manish Kumar*
3.1 Introduction *38*
3.2 Basic Steps Involved in Whole-Genome Sequencing and Resequencing *40*
3.3 Whole-Genome Resequencing Mega Projects in Different Crops *41*
3.3.1 1K Arabidopsis Genomes Resequencing Project *41*
3.3.2 3K Rice Genomes Resequencing Project *41*
3.3.3 Soybean Whole-Genome Resequencing *42*
3.3.4 Chickpea *42*
3.3.5 Pigeon pea *42*
3.3.6 *Vitis* *43*
3.4 Whole-Genome Pooled Sequencing *43*
3.5 Pinpointing Gene Through Whole-Genome Resequencing-based QTL Mapping *44*
3.6 Online Resources for Whole-Genome Resequencing Data *45*
3.6.1 SNP Seek *45*
3.6.2 Rice Functional and Genomic Breeding *45*
3.6.3 Genome Variation Map *45*
3.7 Applications and Successful Examples of Whole-Genome Resequencing *45*
3.8 Challenges for Whole-Genome Resequencing Studies *46*
3.9 Summary *48*
References *48*

**4 QTL Mapping Using Advanced Mapping Populations and High-throughput Genotyping** *52*
*Sweta Sinha, Brij Kishore Kushwaha, and Rupesh K. Deshmukh*
4.1 Introduction *52*

4.2 The Basic Objectives of QTL Mapping *53*
4.3 QTL Mapping Procedure *53*
4.4 The General Steps for QTL Mapping *53*
4.5 Factors Influencing QTL Analysis *53*
4.6 QTL Mapping Approaches *54*
4.7 Statistical Methods for QTL Mapping *57*
4.8 Software for QTL Mapping *61*
4.9 Bi-parental Mapping Populations *61*
4.10 QTL Mapping Using Bi-parental Populations *63*
4.11 Multiparental Mapping Populations *64*
4.11.1 Nested Association Mapping (NAM) *64*
4.11.2 Multi-advanced Generation Inter-cross Populations (MAGIC) *65*
4.12 QTL Mapping Using Multiparental Populations *65*
4.13 Use of High-throughput Genotyping for QTL Mapping *66*
4.13.1 PCR-based SNP Genotyping *67*
4.14 Next-Generation Sequencing-based Genotyping *69*
4.14.1 Restriction-Site-Associated DNA Sequencing (RAD-seq) *69*
4.14.2 Genotyping-by-Sequencing *69*
4.14.3 Whole-Genome Resequencing *70*
4.15 Challenges with QTL Mapping Using Multiparental Populations and High-throughput Genotyping *72*
References *72*

**5 Genome-Wide Association Study: Approaches, Applicability, and Challenges** *80*
*Akshay S. Sakhare, Suneetha Kota, Santosh Rathod, Brajendra Parmar, and Viswanathan Chinnusamy*
5.1 Introduction *80*
5.2 Methodology to Conduct GWAS in Crops *81*
5.3 Statistical Modeling in GWAS *84*
5.4 Efficiency of GWAS with Different Marker Types *85*
5.5 Computational Tools for GWAS *86*
5.6 GWAS Challenges for Complex Traits *86*
5.7 Factors Challenging the GWAS for Complex Traits *87*
5.8 GWAS Applications in Major Crops *88*
5.8.1 Maize *88*
5.8.2 Rice *92*
5.8.3 Wheat *94*
5.8.4 Barley *95*
5.8.5 Pearl Millet *97*
5.8.6 Sugarcane *98*
5.9 Candidate Gene Identification at GWAS Loci *98*
5.10 Meta-GWAS *101*
5.11 GWAS vs. QTL Mapping *102*
References *102*

**6 Genotyping of Seeds While Preserving Their Viability** *112*
*Vinaykumar Rachappanavar, Arushi Padiyal, and Jitender Kumar Sharma*
6.1 Introduction *112*
6.1.1 Genotyping *114*
6.1.2 Genotyping-by-Sequencing *115*
6.2 Genotyping-by-Sequencing with Minimum DNA *116*
6.3 DNA Extraction from Half Grain *117*
6.3.1 DNA Extraction from Rice Seeds *119*
6.3.2 DNA Extraction from Wheat and Barley Seeds *120*
6.3.3 DNA Extraction from Maize Seeds *120*
6.3.4 DNA Extraction from Soybean Seeds *121*
6.3.5 DNA Extraction from Cotton Seeds *121*
6.3.6 DNA Extraction from Papaya Seeds *121*
6.3.7 DNA Extraction from Watermelon Seeds *122*
6.4 GBS with Half Seed *122*
6.5 Applications of GBS as Diagnostic Tool *125*
6.5.1 Germplasm Conservation and Quality Control *125*
6.5.2 Tracking Crop Varieties *126*
6.5.3 Sex Determination *126*
6.5.4 Transgenic Detection *128*
6.5.5 Detection of Seed-borne Diseases *128*
6.6 Summary *129*
References *129*

**7 Genomic Selection: Advances, Applicability, and Challenges** *141*
*Naina Garewal, Riya Joon, Ravneet Kaur, and Kashmir Singh*
7.1 Introduction *141*
7.2 Natural Selection *142*
7.3 Breeding Selection *143*
7.4 Marker-assisted Selection *144*
7.5 Genomic Selection *145*
7.6 Genotyping for Genomic Selection *147*
7.7 Integration of Genomic Selection in MAS Program *150*
7.8 The Efficiency of Genomic Selection for Complex Traits *150*
7.9 Integration of Genomic Selection in the Varietal Trial Program *152*
7.10 Cost Comparison of GS vs MAS *155*
References *155*

**8 Analytical Pipelines for the GBS Analysis** *161*
*Vinaykumar Rachappanavar, Arushi Padiyal, Jitender Kumar Sharma, and Narender Negi*
8.1 Introduction *161*
8.2 Applications of NGS *164*
8.3 NGS Sequencing Platforms *164*
8.3.1 Sequencing by Synthesis *164*

8.3.1.1 Roche 454 Pyrosequencing *164*
8.3.1.2 Illumina *165*
8.3.1.3 Ion Torrent *165*
8.3.2 Sequencing by Ligation *165*
8.3.2.1 SOLiD *165*
8.3.2.2 Polonator *165*
8.3.3 Single-Molecule Sequencing *166*
8.3.3.1 Helicos *166*
8.3.3.2 Pacific Bioscience *166*
8.3.3.3 ChIP-Sequencing *167*
8.4 Tools for NGS Data Analysis *167*
8.5 Generalized Procedure for NGS Data Analysis *171*
8.5.1 Assessment of Quality *172*
8.5.2 Aligning Sequences *174*
8.5.3 Identification of Variants *174*
8.6 Variant Annotation *175*
8.6.1 Visualization of NGS Data *175*
8.7 Role of NGS Informatics in Identifying Variants *175*
8.8 Genotyping by Sequencing *176*
8.9 Analytical Pipelines for GBS *176*
8.10 Comparison of GBS Pipelines *180*
References *181*

**9 Recent Advances and Applicability of GBS, GWAS, and GS in Maize** *188*
*Anshuman Tiwari, Shalu Choudhary, Jayendra Padiya, Abhijit Ubale, Venugopal Mikkilineni, and Bharat Char*
9.1 Introduction *188*
9.2 Maize Genetics *189*
9.3 Importance of Genomics and Genotyping-based Applications in Maize Breeding Programs *190*
9.4 GBS-based QTL Mapping in Maize *191*
9.5 GBS Protocols and Analytical Pipelines for Maize *199*
9.6 Maize Genome Sequencing and Resequencing *201*
9.6.1 Maize Resequencing *201*
9.7 Genotyping-by-Sequencing-based GWAS and GS Efforts in Maize *202*
9.8 Summary *209*
References *210*

**10 Recent Advances and Applicability of GBS, GWAS, and GS in Soybean** *218*
*Prashant Raghunath Shingote, Dhananjay Narayanrao Gotarkar, Ravindra Ramrao Kale, Omkar Maharudra Limbalkar, and Dhiraj Lalji Wasule*
10.1 Introduction *218*
10.1.1 Importance of Soybean for Global Food Security *218*
10.1.2 Challenges in Soybean Production *220*
10.1.3 Soybean Genetic Improvement *221*

10.2 GBS Efforts in Soybean *221*
10.3 High-Density Linkage Maps in Soybean *222*
10.4 GBS Protocols and Analytical Pipelines for Soybean *225*
10.5 GBS-based QTL Mapping Efforts in Soybean *227*
10.6 Soybean Genome Sequencing and Resequencing *230*
10.7 GBS-based GWAS Efforts in Soybean *233*
10.7.1 The General Procedure for Association Mapping *233*
10.7.2 Approaches Used for Association Studies *234*
10.8 GBS-based Genomic Selection Efforts in Soybean *240*
References *243*

**11 Advances and Applicability of Genotyping Technologies in Cotton Improvement** *250*
*Shubham Bhardwaj, Vikas Devkar, Amit Kumar, Alish Alisha, Shivani Sharma, Rupesh K. Deshmukh, and Gunvant B. Patil*
11.1 Introduction *250*
11.2 Challenges due to Polyploidy in Cotton *251*
11.3 Applications of Genomics and Genotyping for Cotton Breeding Programs *251*
11.4 Genotyping Efforts in Cotton *253*
11.5 High-Density Linkage Maps in Cotton *255*
11.6 Whole-Genome Sequencing of Cotton Germplasm *257*
11.7 Application of GBS Technology in Cotton Research *259*
11.8 GBS-based Bi-Parental QTL Mapping and Association Mapping in Cotton *260*
11.9 Summary and Outlook *262*
References *262*

**12 Recent Advances and Applicability of GBS, GWAS, and GS in Millet Crops** *270*
*Pankaj S. Mundada, Swapnil B. Kadam, Anupama A. Pable, and Vitthal T. Barvkar*
12.1 Introduction *270*
12.2 GBS Efforts in Millet Crops *272*
12.3 High-density Linkage Maps in Millet Crops *275*
12.4 GBS-based QTL Mapping Efforts in Millet Crops *277*
12.5 Genome Sequencing and Resequencing of Millet Crops *280*
12.5.1 Pearl Millet *281*
12.5.2 Broomcorn Millet *281*
12.5.3 Finger Millet *282*
12.5.4 Foxtail Millet *282*
12.5.5 Sorghum *283*
12.6 GBS-based GWAS Efforts in Millet Crops *285*
12.7 GBS-based Genomic Selection (GS) Efforts in Millet Crops *287*
12.8 Summary *290*
References *290*

**13 Recent Advances and Applicability of GBS, GWAS, and GS in Pigeon Pea** *295*
*Anuradha Singh and Nisha Singh*
13.1 Introduction *295*
13.2 Pigeon Pea Sequencing and Resequencing *296*

13.3 Development of Pigeon Pea High-density Genotyping Platforms *297*
13.4 Development of High-density Linkage Maps in Pigeon Pea *297*
13.5 QTL Analysis Using High-density Genotyping Platforms and GBS *299*
13.6 GWAS Efforts in Pigeon Pea *299*
13.7 Genomic Selection (GS) Efforts in Pigeon Pea *302*
13.8 Summary *302*
References *303*

**14 Opportunity and Challenges for High-throughput Genotyping in Sugarcane** *306*
*Prathima P. Thirugnanasambandam, Avinash Singode, Lakshmipathy Thalambedu, and Senthilkumar Shanmugavel*
14.1 Introduction *306*
14.2 Sugarcane Genome and Genetics *307*
14.3 Genetic Studies and Marker Systems *308*
14.4 Genotyping-by-Sequencing (GBS) *309*
14.5 SNP Calling Using GBS Pipelines *312*
14.6 Sugarcane Genome Sequencing *313*
14.7 Linkage and QTL Mapping in Sugarcane *315*
14.8 GWAS in Sugarcane *318*
14.9 Genomic Selection in Sugarcane *320*
14.10 Summary *322*
References *322*

**15 Recent Advances and Applicability of GBS, GWAS, and GS in Polyploid Crops** *328*
*Vandana Thakral, Himanshu Yadav, Gunashri Padalkar, Surbhi Kumawat, Gaurav Raturi, Virender Kumar, Rushil Mandlik, Nitika Rajora, and Manipal Singh*
15.1 Introduction *328*
15.2 Challenges for Genotyping in Polyploidy Crops *329*
15.3 Genotyping Platforms for Barley *330*
15.4 Long-Read Sequencing-based Genotyping in Polyploid Canola *331*
15.5 Peanut Genotyping with Targeted Amplicon Sequencing *333*
15.6 SNP Genotyping Methods and Platforms Available for Sugarcane *334*
15.7 Recent Advances and Applicability of GBS, GWAS, and GS in Polyploidy Crop Species *336*
15.7.1 Linkage Maps and Mapping Software Tools for Polyploids *336*
15.8 Haplotype-based Genotyping *338*
15.9 GBS Analytical Pipelines for Polyploids *339*
15.9.1 Fast-GBS *339*
15.9.2 PolyRAD *339*
15.9.3 UGbS-Flex *339*
15.9.4 HaploTag *340*
15.10 GBS-based QTL Mapping Efforts in Polyploids *340*
15.10.1 Wheat *341*
15.10.2 Potato *341*
15.10.3 Sugarcane *342*
15.10.4 Canola *342*

15.10.5 Peanut *343*
15.10.6 Cotton *343*
15.11 GWAS and GS Using High-throughput Genotyping in Polyploidy Crops *343*
15.11.1 Wheat *344*
15.11.2 Potato *344*
15.11.3 Sugarcane *344*
15.11.4 Canola *344*
15.11.5 Peanut *345*
15.11.6 Cotton *345*
References *345*

**16 Recent Advances and Applicability of GBS, GWAS, and GS in Oilseed Crops** *355*
*Sanskriti Vats, Yogesh Sharma, Virender Kumar, Rushil Mandlik, Surbhi Kumawat, Himanshu Yadav, Pallavi Dhiman, Vandana Thakral, Md Aminul Islam, and Sreeja Sudhakaran*
16.1 Introduction *355*
16.2 GBS Efforts in Oilseed Crops *356*
16.3 High-density Linkage Maps for Oilseed Crops *357*
16.3.1 High-density Linkage Maps for Peanut *357*
16.3.2 High-density Linkage Maps for Sunflower *358*
16.3.3 High-density Linkage Maps of Sesamum *358*
16.3.4 High-density Linkage Maps of Safflower *359*
16.3.5 High-density Linkage Maps of Linseed *359*
16.3.6 High-density Linkage Maps of Soybean *359*
16.3.7 High-density Linkage Maps of Rapeseed *360*
16.4 GBS Protocols and Analytical Pipelines *360*
16.5 GBS-based QTL Mapping Efforts in Oilseed Crops *361*
16.5.1 GBS-based QTL in Peanut *361*
16.5.2 GBS-based QTL in Soybean *364*
16.5.3 GBS-based QTL in Rapeseed *364*
16.6 GBS-based GWAS Efforts in Oilseed Crops *365*
References *367*

**Index** *371*

# List of Contributors

**Alish Alisha**, Department of Gene Expression, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

**Gagandeep Singh Bajwa**, Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, Punjab, India

**Vitthal T. Barvkar**, Department of Botany, Savitribai Phule Pune University, Pune, Maharashtra, India

**Shubham Bhardwaj**, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

National Institute of Plant Genome Research (NIPGR), New Delhi, India

**Dharminder Bhatia**, Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, Punjab, India

**Bharat Char**, Mahyco Research Centre, Mahyco Private Limited, Jalna, Maharashtra, India

**Viswanathan Chinnusamy**, Division of Plant Physiology, ICAR-IARI, New Delhi, India

**Shalu Choudhary**, Mahyco Research Centre, Mahyco Private Limited, Jalna, Maharashtra, India

**Rupesh K. Deshmukh**, Agricultural Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Vikas Devkar**, Department of Plant and Soil Science, Institute of Genomics for Crop Abiotic Stress Tolerance (IGCAST), Texas Tech University, Lubbock, TX, USA

**Pallavi Dhiman**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Kishor Gaikwad**, ICAR – National Institute for Biotechnology, New Delhi, India

**Naina Garewal**, Department of Biotechnology, Panjab University, Chandigarh, India

**Dhananjay Narayanrao Gotarkar**, International Rice Research Institute, Los Baños, Philippines

**Md Aminul Islam**, Department of Botany, Majuli College, Majuli, Assam, India

**Priyanka Jain**, ICAR – National Institute for Biotechnology, New Delhi, India

**Riya Joon**, Department of Biotechnology, Panjab University, Chandigarh, India

**Swapnil B. Kadam**, Department of Botany, Savitribai Phule Pune University, Pune, Maharashtra, India

**Ravindra Ramrao Kale**, ICAR-Indian Institute of Rice Research, Hyderabad, Telangana, India

**Ravneet Kaur**, Department of Biotechnology, Panjab University, Chandigarh, India

**Suneetha Kota**, ICAR-IIRR, Hyderabad, Telangana, India

**Amit Kumar**, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Kuldeep Kumar**, ICAR – Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

**Manish Kumar**, Department of Seed Science and Technology, Dr. Yashwant Singh Parmar University of Horticulture and Forestry, Solan, Himachal Pradesh, India

**Sandeep Kumar**, Xcelris Lab Pvt Ltd., Ahmedabad, Gujarat, India

**Virender Kumar**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Surbhi Kumawat**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Brij Kishore Kushwaha**, Department of Molecular Biology and Genetic Engineering, Bihar Agricultural University, Sabour Bhagalpur, Bihar, India

**Omkar Maharudra Limbalkar**, ICAR-Division of Genetics, Indian Agriculture Research Institute, New Delhi, India

**Rushil Mandlik**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Venugopal Mikkilineni**, Mahyco Research Centre, Mahyco Private Limited, Jalna, Maharashtra, India

**Pankaj S. Mundada**, Department of Botany, Savitribai Phule Pune University, Pune, Maharashtra, India

Department of Biotechnology, Yashavantrao Chavan Institute of Science, Satara Maharashtra, India

**Narender Negi**, Department of Fruit Science, ICAR-NBPGR Regional Station, Shimla, Himachal Pradesh, India

**Anupama A. Pable**, Department of Microbiology, Savitribai Phule Pune University, Maharashtra, India

**Gunashri Padalkar**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Jayendra Padiya**, Mahyco Research Centre, Mahyco Private Limited, Jalna, Maharashtra, India

**Arushi Padiyal**, Department of Seed Science and Technology, Dr. Yashwant Singh Parmar University of Horticulture and Forestry, Solan, Himachal Pradesh, India

**Brajendra Parmar**, ICAR-IIRR, Hyderabad, Telangana, India

**Gunvant B. Patil**, Department of Plant and Soil Science, Institute of Genomics for Crop Abiotic Stress Tolerance (IGCAST), Texas Tech University, Lubbock, TX, USA

**Vinaykumar Rachappanavar**, Department of Seed Science and Technology, Dr. Yashwant Singh Parmar University of Horticulture and Forestry, Solan, Himachal Pradesh, India

Department of Agriculture, MS Swaminathan School of Agriculture, Shoolini University, Solan, Himachal Pradesh, India

**Nitika Rajora**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Santosh Rathod**, ICAR-IIRR, Hyderabad, Telangana, India

**Gaurav Raturi**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Rita**, ICAR – National Institute for Biotechnology, New Delhi, India

**Akshay S. Sakhare**, Division of Plant Physiology, ICAR-IARI, New Delhi, India

**Swati Saxena**, ICAR – National Institute for Biotechnology, New Delhi, India

**Senthilkumar Shanmugavel**, Crop Improvement Division, ICAR – Sugarcane Breeding Institute, Coimbatore, Tamil Nadu, India

**Jitender Kumar Sharma**, Department of Agriculture, School of Agriculture, Baddi University of Emerging Sciences & Technology, Baddi, Himachal Pradesh, India

**Sandhya Sharma**, ICAR – National Institute for Biotechnology, New Delhi, India

**Shivani Sharma**, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Yogesh Sharma**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Prashant Raghunath Shingote**, Vasantrao Naik College of Agricultural Biotechnology, Dr. Panjabrao Deshmukh Krishi Vidyapeeth, Akola, Maharashtra, India

**Harsha Srivastava**, ICAR – National Institute for Biotechnology, New Delhi, India

**Anuradha Singh**, Department of Genomics, ICAR – National Institute on Plant Biotechnology, New Delhi, India

**Kashmir Singh**, Department of Biotechnology, Panjab University, Chandigarh, India

**Manipal Singh**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Nisha Singh**, Department of Genomics, ICAR – National Institute on Plant Biotechnology, New Delhi, India

**Avinash Singode**, ICAR – Indian Institute of Millets Research, Hyderabad, Telangana, India

**Sweta Sinha**, Department of Molecular Biology and Genetic Engineering, Bihar Agricultural University, Sabour Bhagalpur, Bihar, India

**Sreeja Sudhakaran**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Lakshmipathy Thalambedu**, Crop Improvement Division, ICAR – Sugarcane Breeding Institute, Coimbatore, Tamil Nadu, India

**Vandana Thakral**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

**Prathima P. Thirugnanasambandam**, Crop Improvement Division, ICAR – Sugarcane Breeding Institute, Coimbatore, Tamil Nadu, India

**Anshuman Tiwari**, Mahyco Research Centre, Mahyco Private Limited, Jalna, Maharashtra, India

**Kishor Tribhuvan**, ICAR – Indian Institute of Agricultural Biotechnology, Ranchi, Jharkhand, India

**Abhijit Ubale**, Mahyco Research Centre, Mahyco Private Limited, Jalna, Maharashtra, India

**Sanskriti Vats**, Agricultural Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

Regional Centre for Biotechnology, Faridabad, Haryana (NCR Delhi), India

**Joshita Vijayan**, ICAR – National Institute for Biotechnology, New Delhi, India

**Dhiraj Lalji Wasule**, Vasantrao Naik College of Agricultural Biotechnology, Dr. Panjabrao Deshmukh Krishi Vidyapeeth, Akola, Maharashtra, India

**Himanshu Yadav**, Department of Agriculture Biotechnology, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

# Preface

Recent advances in sequencing technology and computational resources have accelerated genomics and translational research in crop science. The technological advances have provided many opportunities in genomics-assisted plant breeding to address issues related to food security. Among the several applications, genotyping-by-sequencing (GBS) technology has evolved as one of the frontier areas facilitating high-throughput plant genotyping. The GBS approaches have proved effective for the utilization in genotyping-based applications like quantitative trait loci (QTL) mapping, genome-wide association study (GWAS), genomic selection (GS), and marker-assisted breeding (MAB). Considering the current affairs in plant breeding, we decided to compile the advances in GBS methods, statistical approaches to analyze the GBS data, and its applications including QTL mapping, GWAS, and GS in crop improvement.

Presently, the food produced around the world is adequate for the existing population. However, the constantly increasing population mounting pressure on a food production system. Hence efficient utilization of technological advances and existing knowledge is essential to enhance food production to match the growing food demand. In this direction, most of the countries around the globe have adopted advanced genomic methodologies to breed superior plant genotypes. Among such technological advances, the high-throughput genotyping using GBS has shown promising results in different crop plants. The GBS has predominantly been used for germplasm evaluation, evolutionary studies, development of dense linkage map, QTL mapping, GWAS, GS, and MAB. The cost-effectiveness and whole-genome coverage make GBS more reliable than other next-generation sequencing (NGS) techniques.

This book describes advanced molecular markers, high-throughput genotyping platforms, whole-genome resequencing (WGR), QTL mapping using advanced mapping populations, analytical pipelines for the GBS analysis, advances in GWAS, advances in GS, application of GBS, GWAS, and GS in different crop plants. The different marker types including traditional and advanced markers used in plant genotyping have been presented in great detail. DNA extraction directly from seeds without germination can save time and effort. Several modified and crop-specific nondestructive seed DNA extraction protocols have been compiled and presented. Many advanced genotyping platforms are now available which cater to specific research purposes because of the differences in terms of reaction chemistry involved, cost, method of signal detection, and flexibility in the protocols. Such advanced platforms along with their principles have been discussed. The WGR

methodology and available resources have been covered in detail. The WGR has emerged as a powerful method to identify genetic variation among individuals. The recent advancement in WGR includes pool-Seq which provides an alternative to individual sequencing and a cost-effective method for GWAS. Compared to biparental populations the multiparental population provides an opportunity to interrogate multiple alleles and to provide an increased level of recombination and mapping resolution of QTLs. The use of such improved populations in the era of high-throughput genotyping has been presented in one of the chapters. The dedicated section focused on the basic principle of GWAS, the efficiency of different markers, candidate gene identification, meta-GWAS, and statistical methods involved in GWAS analysis has been included. For genetic mapping, and marker-assisted selection, rapid and quality DNA isolation is mandatory to accelerate the whole process. A focused section about GS has been included which gives an account of the basic concept, advances, applicability, and challenges of GS. Similarly, a separate chapter is included which discusses the analytical pipelines used for GBS data. Application of technologies such as GBS, GWAS, and GS in different crop categories like cereals, pulses, oilseeds, and commercial crops has been discussed in different chapters.

Here, we have tried to compile basic aspects and recent advances in GBS, GWAS, and GS in plant breeding. We believe that the book will be helpful to researchers and scientists to understand and plan future experiments. This book will enable plant scientists to explore GBS application more efficiently for basic research as well as applied aspects in various crops improvement projects.

Editors
Dr. Humira Sonah
Dr. Vinod Goyal
Dr. S. M. Shivaraj
Dr. R. K. Deshmukh

# 1

# Molecular Marker Techniques and Recent Advancements

*Dharminder Bhatia and Gagandeep Singh Bajwa*

*Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, Punjab, India*

## 1.1 Introduction

Plant selection and systematic breeding efforts led to the development of present-day improved cultivars of crop plants. From a historical perspective, increased crop yield is the result of genetic improvement (Fehr 1984). Markers play an important role in the selection of traits of interest. Markers can be morphological, biochemical, or molecular in nature. Morphological markers are visual phenotypic characters such as growth habit of the plant, seed shape, seed color, flower color etc. Biochemical markers are the isozyme-based markers characterized by variation in molecular form of enzyme showing a difference in mobility on an electrophoresis gel. Very few morphological and biochemical markers are available in plants, and they are influenced by developmental stage and environmental factors. Since a large number of economically important traits are quantitative in nature, which are affected by both genetic and environmental factors, the morphological and biochemical markers-based selection of traits may not be much reliable. The subsequent discovery of abundantly available DNA-based markers made possible the selection of almost any trait of interest. DNA-based markers are not affected by the environment. Besides, these markers are highly reproducible across labs and show high polymorphism to distinguish between two genetically different individuals or species.

In the last four decades, DNA-based molecular marker technology has witnessed several advances from low throughput hybridization-based markers to high-throughput sequencing-based markers. These advances have been possible due to critical discoveries such as polymerase chain reaction (PCR) (Mullis et al. 1986), Sanger sequencing method (Sanger et al. 1977), automation of Sanger sequencing (Shendure et al. 2011), next-generation sequencing (NGS) technologies (Mardis 2008), and development of bioinformatics tools. This chapter will briefly discuss different types of molecular markers while particularly focusing on recent developments in molecular marker technologies. These

*Genotyping by Sequencing for Crop Improvement*, First Edition.
Edited by Humira Sonah, Vinod Goyal, S.M. Shivaraj, and Rupesh K. Deshmukh.
© 2022 John Wiley & Sons Ltd. Published 2022 by John Wiley & Sons Ltd.

developments have expedited the mapping and cloning of several loci governing important traits, precise trait selection, and transfer into elite germplasm.

## 1.2 What is a Molecular Marker?

DNA or molecular marker is a fragment of the DNA that is associated with a particular trait in an individual. These molecular markers aid in determining the location of genes that control key traits.

Generally, molecular markers do not represent the gene of interest but act as “flags” or “signs.” Similar to genes, all the molecular markers occupy a specific position within the chromosomes. Molecular markers located close to genes (i.e. tightly linked) are referred to as “gene tags.”

DNA-based molecular markers are the most widely used markers predominantly due to their abundance. They arise from different classes of DNA mutations such as substitution mutations (point mutations), rearrangements (insertions or deletions), or errors in replication of tandemly repeated DNA. These markers are selectively neutral because they are usually located in noncoding regions of DNA. Unlike morphological and biochemical markers, DNA markers are practically unlimited in number and are not affected by environmental factors and/or the developmental stage of the plant.

DNA markers show genetic differences that can be visualized by using a gel electrophoresis technique and staining ethidium bromide or hybridization with radioactive or colorimetric probes. Markers that can identify the difference between two individuals are referred to as polymorphic markers, whereas those that do not distinguish the individuals are called monomorphic markers. Based on how polymorphic markers can discriminate between individuals, they are described as codominant or dominant. Codominant markers indicate differences in size whereas dominant markers reveal differences based on their presence or absence. The different forms of a DNA marker in the form of band size on gels are known as marker “alleles.” Dominant marker has only two alleles whereas codominant markers may have many alleles.

## 1.3 Classes of Molecular Markers

Based on the method of their detection, DNA markers are broadly classified into three groups: (i) hybridization-based, (ii) PCR-based, and (iii) DNA sequence-based molecular markers. Molecular markers have been discussed earlier in several reviews (Collard et al. 2005; Semagn et al. 2006; Gupta and Rustgi 2004) and book chapters (Mir et al. 2013; Singh and Singh 2015), which readers can also consult for more details. However, a brief description of each of these markers has been presented below.

### 1.3.1 Hybridization-based Markers

#### 1.3.1.1 Restriction Fragment Length Polymorphism (RFLP)

These are the first molecular markers used by Grodzicker et al. (1975) in adenovirus and Botstein et al. (1980) in human genome mapping. These were first used in plants by Helentjaris et al. (1986). In this type of marker, polymorphism is detected by cutting DNA

into fragments by the use of restriction enzymes followed by hybridization of radioactively labeled DNA probes which are single or low copy DNA fragments and visualized by autoradiography. DNA probes could be genomic clones, cDNA clones, or even cloned genes. The RFLP markers show co-dominance and are highly reliable in linkage analysis and breeding (Semagn et al. 2006). However, this technique requires a large quantity of DNA, labor-intensive, relatively expensive, and hazardous. RFLP shows polymorphism in two different species if they differ due to point mutations, insertion/deletion, inversion, translocation, and duplication.

#### 1.3.1.2 Diversity Array Technology (DArT™)

This is a high-throughput DNA polymorphism analysis method which combines microarray and restriction-based PCR methods. It is similar to AFLP where hybridization is used for the detection of polymorphism. It can able to provide a comprehensive genome coverage even in those organisms not having genome sequence information (Jaccoud et al. 2001). Diversity array technology (DArT) is a solid-state open platform method for analyzing DNA polymorphism. DArT procedure includes (i) Generating a diversity panel and (ii) Genotyping using a diversity panel. The diversity panel is generated using a set of lines representing the breadth of variability in germplasm (~10 lines). An equal quantity of DNA from each representative line is pooled followed by restriction with two to three restriction endonucleases (REs) and ligation of RE-specific adaptors. Later DNA fragments are amplified using adaptor complementary primers. The representation fragments are ligated to vector and transformed into *Escherichia coli* cells. The transformed cells with recombinant DNA are selected and amplified using M13 forward and reverse primer. The amplified DNA is isolated and purified. The purified DNA is coated onto polylysine-coated glass slides to generate a diversity array.

For genotyping, the representation fragments of the target genotypes are prepared in the same as in the diversity panel. The DNA fragments are column purified and fluorescently labeled with two different dyes (Cy3 or Cy5). The labeled DNA fragments are used for hybridization onto the diversity array. Two representative panels – one labeled with Cy3 and another with Cy5 – can be hybridized simultaneously and hybridization signal intensities are measured for each spot. DArT, thus detects DNA polymorphism at several hundred genomic loci in a single array without relying on sequence information.

### 1.3.2 Polymerase Chain Reaction (PCR)-based Markers

#### 1.3.2.1 Simple-Sequence Repeats (SSRs)

Simple-sequence repeats (SSRs) (Litt and Luty 1989) are also known as microsatellites or short tandem repeats (STRs) or simple sequence length polymorphism (SSLP). These are widely used markers and are also referred to as the mother of all the markers. These are STRs, generally of one to eight nucleotide length. These are found dispersed throughout the genome and are hypervariable. These repeat regions are flanked with unique sequences that are highly conserved. The flanking unique sequences are used to design complementary primers which can be assayed with PCR. SSRs are highly polymorphic and codominant markers. These show polymorphism as a result of the variable number of repeat units. Before the era of genome sequencing, it was difficult to develop SSRs due to the extensive cost and labor involved in the identification of repeat regions and flanking unique

sequences. However, with the availability of genome sequences of several organisms, the development of SSR has become very easy which involves in silico identification of STRs, designing of SSR from flanking unique sequences, and validation through experimentation. SSR markers have shown immense application in population genetic analysis, gene mapping, and cloning due to their abundance in the genome and high polymorphism, and very high reproducibility across labs. SSR-based linkage maps have been developed in several important crop plants such as rice (Temnykh et al. 2000; McCouch et al. 2002; Orjuela et al. 2010), wheat (Roder et al. 1998), maize (Sharopova et al. 2002), potato (Milbourne et al. 1998), etc.

#### 1.3.2.2 Sequence-Tagged Sites (STSs)

Sequence-tagged sites (STSs) were first developed for physical mapping of the human genome by Olsen et al. (1989). STS is the short unique sequences developed from polymorphic RFLP probe or AFLP fragment which is linked to desirable traits. The RFLP probes or AFLP fragments showing polymorphism are end-sequenced and primers are designed to specifically amplify these fragments. STS markers are co-dominant and highly reproducible. For example, STS markers have been developed for RFLP markers linked with bacterial blight resistance genes *xa5, xa13,* and *Xa21* (Huang et al. 1997). One major limitation of these types of markers is the reduced polymorphism than the corresponding RFLP probe.

#### 1.3.2.3 Randomly Amplified Polymorphic DNAs (RAPDs)

Williams et al. (1990) first developed these markers to amplify DNA without prior sequence information. In this type of marker, the arbitrary decamer sequences are used as primers at low annealing temperatures for DNA amplification. These markers are referred to as dominant markers because the polymorphism is determined based on the presence or absence of a particular amplified fragment. Polymorphism may also be due to varying brightness of bands at a particular locus due to copy number differences. These markers have been used for constructing linkage maps in several species (Hunt 1997; Laucou et al. 1998) and also for tagging genes of economic importance. However, due to the dominant nature, these may not be appropriate for genetic mapping and marker-assisted selection (MAS). One major limitation of these markers is the lack of repeatability in certain cases. Variations of RAPD include AP-PCR (arbitrarily primed PCR) and DAF (DNA amplification fingerprinting (Table 1.1).

**Table 1.1** Details of the other important molecular markers.

| Marker | Description |
|---|---|
| Variable number tandem repeat (VNTR) or minisatellites | A short DNA sequence (10–100 bp) is present as tandem repeats and is a highly variable copy number |
| DNA amplification fingerprinting (DAF) | A variation of RAPD, where 4–5 bp single and arbitrary primer is used to detect polymorphism |
| Arbitrary-primed PCR (AP-PCR) | A variation of RAPD, where 18–32 bp long single and arbitrary primer is used to detect polymorphism |

**Table 1.1** (Continued)

| Marker | Description |
|---|---|
| Inter-simple sequence repeat (ISSR) | Primers are designed based on the repeat region of microsatellites. These primers are used to amplify the region between two microsatellites. The stretches of unique DNA in between or flanking the SSRs are amplified. A single SSR-based primer is used to prime PCR |
| Selective amplification of microsatellite polymorphic loci (SAMPL) | A modification of ISSR, where SSR-based primer is used along with AFLP primer. The template is identical to the AFLP template and the rare cutter primer is replaced by SSR-based primer |
| Cleaved amplified polymorphic sequences (CAPS) | These markers are also called PCR-RFLP, where amplified PCR product is digested with endonucleases to reveal polymorphism. These are used when PCR product does not show polymorphism and restriction enzyme site present in amplified PCR product may detect polymorphism |
| Derived cleaved amplified polymorphic sequences (dCAPS) | A variation of CAPS, where a primer containing one or more mismatches to template DNA is used to create a restriction enzyme recognition site in one allele but not in another due to the presence of SNP. Thus, obtained PCR product is subjected to restriction enzyme digestion to find the presence or absence of the SNP |
| Single-strand conformational polymorphism (SSCP) | DNA fragments of size ranging from 200 to 800 bp were amplified by PCR using specific primers (20–25 bp), followed by gel-electrophoresis of single-strand DNA to detect nucleotide sequence variation. The method is based on a principle that the secondary structure of single-strand DNA molecule changes significantly if it harbors mutation. This method detects nucleotide variation without sequencing a DNA sample |
| Denaturing/temperature gradient gel electrophoresis (DGGE, TGGE) | These methods reveal polymorphism due to differential movement of the same genomic double-stranded region with different base-pair composition. As an example, the AT-rich region would have a lower melting temperature than the GC-rich region |
| Target region amplification polymorphism (TRAP) | This method employs primers designed from the EST database for detecting polymorphism around a selected candidate gene. This includes two primers of 18 bp, one of which is designed from targeted EST and the other is an arbitrary primer |

#### 1.3.2.4 Sequence Characterized Amplified Regions (SCARs)

These markers overcome the limitation of RAPDs. In this case, the RAPD fragments that are linked to a gene of interest are cloned and sequenced. Based on the terminal sequences, longer primers (20 mer) are designed. These SCAR primers more specifically amplify a particular locus. These are similar to STS markers in design and application. The presence or absence of the band indicates variation in sequences. The SCAR markers thus are dominant in nature. These, however, can be converted to codominant markers in certain cases by digesting the amplified fragment with tetranucleotide recognizing restriction enzymes. There are several examples where the RAPD markers linked to the gene of importance have been converted to SCAR markers (Joshi et al. 1999; Liu et al. 1999; Kasai et al. 2000; Akkurt et al. 2007; Chao et al. 2018).

#### 1.3.2.5 Amplified Fragment Length Polymorphism (AFLP)

This marker technique was developed by Vos et al. (1995) and is patented by Keygene (www.keygene.com). In this technique, DNA is cut into fragments by a combination of restriction enzymes which are frequent (four bases) and rare (six bases) cutters that generate restriction overhangs on both sides of fragments. This is followed by the annealing of double-stranded oligonucleotide adapters of a few oligonucleotide bases with respective restriction overhangs. The oligonucleotide adapters are designed in such a way that the original restriction sites are not reinstated and also provide the PCR amplification sites. The fragments are PCR amplified and visualized on agarose gel. This method produces many restriction fragments enabling the polymorphism detection. The number of amplified DNA fragments can be controlled by selecting different number or composition of bases in the adapters. The stringent reaction conditions used for primer annealing make this technique more reliable. This method is a combination of both RFLP and PCR techniques and is extremely useful in the detection of polymorphism between closely related genotypes. Like RAPD, AFLP is a dominant marker and is not preferred for genetic mapping studies and MAS. AFLP maps have been constructed in several species and integrated into already existing RFLP maps e.g. tomato (Haanstra et al. 1999), rice (Cho et al. 1997), and wheat (Lotti et al. 2000).

#### 1.3.2.6 Expressed Sequence Tags (ESTs)

These markers are developed by end sequencing (generally 200–300 bp) of random cDNA clones. The sequence thus obtained is referred to as expressed sequence tags (ESTs). A large number of ESTs have been synthesized in several crop plants and are available in the EST database at NCBI (https://www.ncbi.nlm. nih.gov/dbEST/). These markers were originally developed to identify gene transcripts and have played important role in the identification of several genes and the development of markers such as RFLP, SSR, SNPs, CAPS, etc. (Semagn et al. 2006). However, EST-based SSRs show less polymorphism as compared to genomic DNA-based SSRs. Since EST markers are from expressed sequence regions, these are highly conserved among the species and can be used for synteny mapping. Most of these could also be functional genes. A large number of EST markers have been used in rice for developing a high-density linkage map (Harushima et al. 1998) and for chromosome bin mapping in wheat using deletion stocks (Qi et al. 2003). In addition to these, several other molecular marker variants have been developed. The description of those markers is presented in Table 1.1.

## 1.4 Sequencing-based Markers

### 1.4.1 Single-Nucleotide Polymorphisms (SNPs)

Single-nucleotide polymorphisms (SNPs) are more abundant resulted from single-base pair variations. These are evenly distributed in a whole genome that can tag almost any gene or locus of a genome (Brookes 1999). However, the distribution of SNPs varies among species with 1 SNP per 60–120 bp in maize (Ching et al. 2002) and 1 SNP per 1000 bp in humans (Sachidanandam et al. 2001). SNPs are more prevalent in the noncoding region. In the coding region, SNPs could be synonymous or nonsynonymous. In synonymous SNPs, there is no change in the amino acid resulting in no phenotypic differences. However, phenotypic differences could be produced due to modified mRNA splicing (Richard and Beckman 1995). In nonsynonymous SNPs, change in amino acid results in phenotypic differences. SNPs are mostly bi-allelic and cause polymorphism due to nucleotide base substitution. The two types of nucleotide base substitutions result in SNPs. A transition substitution occurs between purines (A, G) or between pyrimidines (C, T). This type of substitution constitutes two-thirds of all SNPs. A transversion substitution occurs between a purine and pyrimidine. SNPs can be detected by the alignment of the similar genomic region of two different species. The SNPs have only two alleles compared to typical multiallele SSLP; however, this disadvantage can be compensated by using the high density of SNPs.

### 1.4.2 Identification of SNP in a Pregenomic Era

Initially, identification of SNP markers was laborious and expensive and involved allele-specific sequencing (Ganal et al. 2009). This includes sequencing of unigene-derived amplicons using Sanger's method from two or more than two lines. In an experiment, about 350 bp of the RFLP clone, A-519 was end sequenced in soybean and the flanking amplification primers were designed (Coryell et al. 1999). Primers were used to screen for allele diversity using PCR from ten genotypes and the amplicons were sequenced followed by sequence comparison to identify SNP. SNPs were also identified through mining a large number of EST sequences in EST databases, which are generated through improved sequencing technologies (Soleimani et al. 2003). These SNPs are further validated using PCR (Batley et al. 2003). These approaches allowed the identification of mainly gene-based SNPs, but their frequency is generally low. Additionally, SNPs located in low-copy noncoding regions and intergenic spaces could not be identified.

Several assays have been developed for genotyping based on identified SNPs which include, allele-specific hybridization, primer extension, oligonucleotide ligation, and invasive cleavage (Sobrino et al. 2005). Besides, DNA chips, allele-specific PCR, and primer extension were also attractive options since these are suitable for automation and can be used for the development of dense genetic maps. Allele-specific hybridization was used for the identification of polymorphism in 570 genotypes of soybean (Coryell et al. 1999).

## 1.5 Recent Advances in Molecular Marker Technologies

The improvement of Sanger sequencing technology in the 1990s combined with the beginning of EST and genome sequencing projects in model plants led to the spurt in the identification of variation at the single-base resolution (Wang et al. 1998). From 2005 onward, the

emergence of NGS platforms such as Roche 454, Illumina HiSeq2500, ABI 5500xl SOLiD, Ion Torrent, PacBio RS, Oxford Nanopore, and advances in bioinformatics tools simplified the process of identification of genome-wide SNPs and changed the face of molecular marker technology. NGS-based genotyping platforms such as genotyping-by-sequencing (GBS), whole-genome resequencing (WGR), and high-density SNP arrays helped to type thousands of SNPs in a single reaction in hundreds of individuals.

### 1.5.1 Genotyping-by-Sequencing (GBS)

GBS is an NGS-based reduced representation sequencing technique for the identification of genome-wide SNPs and genotyping large populations (Bhatia et al. 2013). GBS is a one-step approach for the identification and utilization of markers in a single reaction. It is a complexity reduction procedure where a combination of restriction enzymes is used to separate low copy sequences from high copy repetitive regions. In general, GBS involves the sequencing of fragments generated through restriction digestion of the genome on the NGS platform. In this process, the DNA of the population is digested with RE followed by ligation of RE-specific adaptors containing genotype-specific barcode sequences and sites for binding PCR and sequencing primers (Figure 1.1). The fragments thus generated can be PCR amplified and an equal volume of PCR product from different individuals are pooled in a tube. The fragments in the pool can be selected based on their size and sequenced on

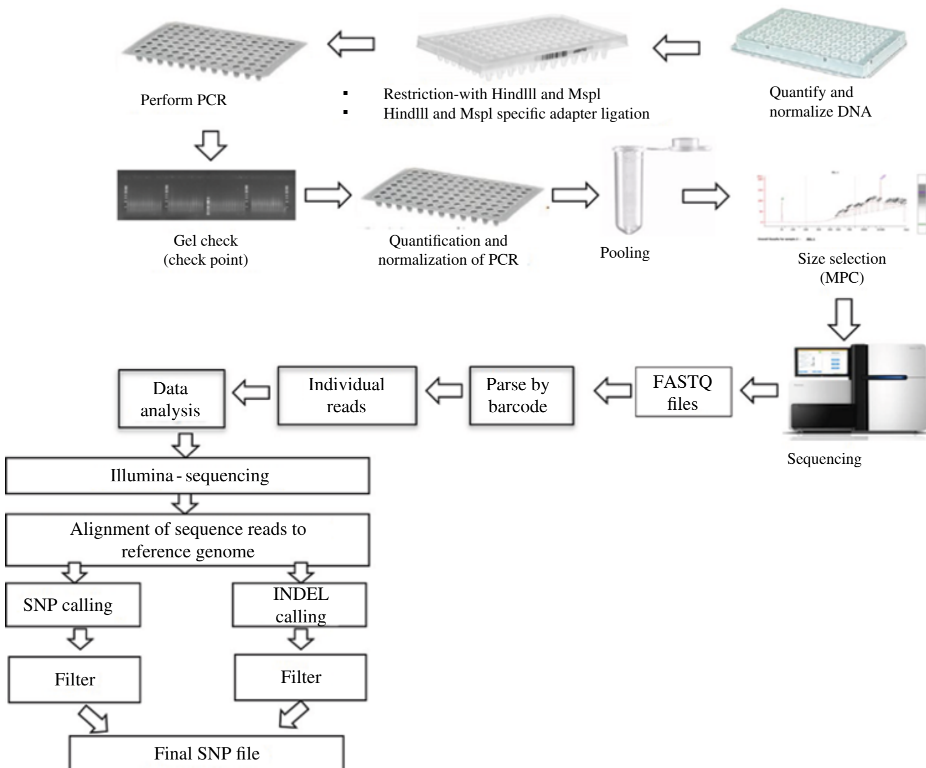


**Figure 1.1** An example of GBS and GBS data analysis workflow for identification of SNP markers.