



<b>SECURITY (DISSEMINATION LEVEL)</b>	Confidential
<b>CONTRACTUAL DATE OF DELIVERY</b>	30.04.2019
<b>ACTUAL DATE OF DELIVERY</b>	12.10.2021
<b>DELIVERABLE NUMBER</b>	D 7.1
<b>TYPE</b>	Deliverable
<b>STATUS AND VERSION</b>	Final
<b>NUMBER OF PAGES</b>	193
<b>WP CONTRIBUTING TO THE DELIVERABLE</b>	WP 7
<b>LEAD BENEFICIARY</b>	UNITUS
<b>OTHER CONTRIBUTORS</b>	BOKU, IFVC, IHAR, NPPC
<b>AUTHOR(S)</b>	Pagnotta Mario A., Forte Paola, Kuzmanović Ljiljana, Čeran Marina, Meglič Vladimir, Pipan Barbara, Plich Jarosław, Sedlar Aleš, Timperio Annamaria, Turco Silvia, Bürstmayr Hermann, Ehn Magdalena, Titan Primož
<b>KEYWORDS</b>	Genotyping, molecular methods, selection, transcriptomics, statistics, bioinformatics
<b>ABSTRACT (FOR DISSEMINATION)</b>	<p>The document contains material for improved genotyping training. The booklet is divided in several chapters, with suggested further reading material and links to tutorial. The four sections are: 1) Genetic background, 2) Germplasm conservation, evaluation, and utilisation (breeding), 3) Molecular procedures and tools and 4) Statistics and bioinformatics.</p> <p><b>Audience:</b> The material is aimed at providing background knowledge for young researchers and scientists with a new blend of applied and fundamental R&amp;D skills required to, (a) further improve the selection and breeding of organic crops (b) apply breeding and agronomic methodologies/approaches developed in the project to other crops and/or (c) contribute to transferring technologies developed into commercial practice. This training material will provide background material for the ECOBREED workshops on Advanced Genotyping.</p>
<b>DOCUMENT ID</b>	D 7.1 Production of materials for Advanced Genotyping training

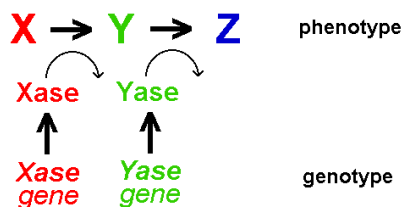


**ecobreed**  
IMPROVING CROPS



Funded by European Union  
Horizon 2020  
Grant agreement No 771367

# Miscellanea of didactic material for the training courses on Advanced Genotyping<sup>1</sup>



<sup>1</sup> This booklet has been developed within the ECOBREED project (Increasing the efficiency and competitiveness of organic crop breeding), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 771367.

## D 7.1 Production of materials for improved genotyping training

### AUTHORS/EDITORS:

Bürstmayr Hermann: University of Natural Resources and Life Sciences, Austria  
[hermann.buerstmayr@boku.ac.at](mailto:hermann.buerstmayr@boku.ac.at)

Ćeran Marina: Institute of Field and Vegetable Crops, Serbia [marina.ceran@ifvcns.ns.ac.rs](mailto:marina.ceran@ifvcns.ns.ac.rs)

Ehn Magdalena: University of Natural Resources and Life Sciences, Austria  
[magdalena.ehn@boku.ac.at](mailto:magdalena.ehn@boku.ac.at)

Forte Paola: The University of Tuscia, Italy  
[pforte64@gmail.com](mailto:pforte64@gmail.com)

Kuzmanović Ljiljana: The University of Tuscia, Italy  
[ljiljanakuzmanovic@gmail.com](mailto:ljiljanakuzmanovic@gmail.com)

Meglič Vladimir: Agricultural Institute of Slovenia, Slovenia  
[vladimir.meglic@kis.si](mailto:vladimir.meglic@kis.si)

Pagnotta Mario A.: The University of Tuscia, Italy  
[pagnotta@unitus.it](mailto:pagnotta@unitus.it)

Pipan Barbara: Agricultural Institute of Slovenia, Slovenia  
[barbara.pipan@kis.si](mailto:barbara.pipan@kis.si)

Plich Jarosław: The Plant Breeding and Acclimatization Institute – National Research Institute, Poland  
[j.plich@ihar.edu.pl](mailto:j.plich@ihar.edu.pl)

Primož Titan: RGA research genetics and agrochemistry, Ltd., Slovenia  
[primoz.titan@rga.si](mailto:primoz.titan@rga.si)

Sedlar Aleš: Agricultural Institute of Slovenia, Slovenia  
[ales.sedlar@kis.si](mailto:ales.sedlar@kis.si)

Timperio Annamaria: The University of Tuscia, Italy  
[timperio@unitus.it](mailto:timperio@unitus.it)

Turco Silvia: The University of Tuscia, Italy  
[silvia.turco@unitus.it](mailto:silvia.turco@unitus.it)

Edited by Mario A. Pagnotta  
The University of Tuscia, Viterbo – Italy  
Via San Camillo de Lellis snc

## D 7.1 Production of materials for improved genotyping training

### TOPICS BY AUTHORS

Topics	Author/Editor
What is Genotyping?	Mario A. Pagnotta Barbara Pipan
Overview of Quantitative Genetics.	Mario A. Pagnotta
Population genetics and HW equilibrium	Mario A. Pagnotta
Genetic resources and their conservation	Mario A. Pagnotta
Overview of breeding methods and trait selection in agriculture	Barbara Pipan Vladimir Meglič Primož Titan
DNA extraction methods	Barbara Pipan
PCR (Polymerase Chain Reaction) – main principles	Ljiljana Kuzmanović
Gel electrophoresis	Ljiljana Kuzmanović
Markers and Molecular Tools	Mario A. Pagnotta
HRM High-Resolution Melting	Paola Forte
Real Time PCR	Mario A. Pagnotta
Methods/strategies for QTL identification	Ljiljana Kuzmanović
MAS	Paola Forte
MAS in bean	Barbara Pipan
MAS for bunt resistance	Hermann Bürstmayr Magdalena Ehn
MAS in organic potato breeding	Jarosław Plich
Association genetics	Ljiljana Kuzmanović
Sequences	Aleš Sedlar Silvia Turco
Chromosome engineering	Ljiljana Kuzmanović
Fluorescent <i>In-situ</i> hybridization - FISH or GISH	Paola Forte
Transcriptomics and its applications	Ljiljana Kuzmanović
Transcriptomics – biological interpretation of gene expression data	Aleš Sedlar
Proteomics	Annamaria Timperio
Statistical tests (Mean, St dev, Variance, correlation, ANOVA, etc.).	Mario A. Pagnotta
Comparison among methods and statistical software packages to analyse genetic diversity by means of codominant markers	Mario A. Pagnotta
What is Bioinformatics? Manage sequences	Silvia Turco
Genotyping by Sequence	Marina Čeran
Bioinformatics resources	Aleš Sedlar
Data science in R	Aleš Sedlar
Genomic selection in R	Marina Čeran
TASSEL 3.0 Genotyping by Sequencing (GBS) pipeline documentation	Marina Čeran
Tassel Association	Marina Čeran



## D 7.1 Production of materials for improved genotyping training

### Contents

<b>Summary</b>	<b>9</b>
<b>Introduction</b>	<b>10</b>
<b>Section 1 - Genetic background</b>	<b>12</b>
<b>1. What is genotyping?</b>	<b>12</b>
1.1 <i>Why we perform SNP genotyping?</i>	12
1.1.1 Disease association	12
1.1.2. Population genomics	12
1.1.3. Trait selection in agriculture	13
1.1.4. Microorganisms	13
<i>Online Tutorials</i>	13
<i>Further reading</i>	13
<b>2. Overview of quantitative genetics</b>	<b>14</b>
2.1 <i>Heritability</i>	14
<i>Further Reading</i>	15
<b>3. Population Genetics and Hardy-Weinberg equilibrium</b>	<b>16</b>
3.1. <i>Hardy-Weinberg equilibrium frequency consideration</i>	16
3.2. <i>Hardy- Weinberg equilibrium</i>	18
3.3. <i>Hardy- Weinberg equilibrium, alteration of the conditions</i>	19
3.3.1. Population size – Genetic drift	19
3.3.2. Type of mating	20
3.3.3. Absence of mutations	20
3.3.4. Absence of migration	20
3.3.5. Absence of selection	21
<i>Online Tutorials</i>	22
<i>Further reading</i>	22
<b>Section 2 - Germplasm conservation, evaluation, and utilisation (breeding)</b>	<b>23</b>
<b>4. Genetic resources and their conservation</b>	<b>23</b>
4.1 <i>Species definition and gene pool</i>	23
4.2 <i>The importance of genetic resources</i>	24
4.3 <i>Plant breeding and genetic resources</i>	25
4.4 <i>In-situ and Ex-situ Strategies of Plant Conservation</i>	26
4.4.1 In-situ Conservation	26
4.4.2 Ex-situ Conservation	27
<i>Online Tutorials</i>	28
<i>Further reading</i>	29
<b>5. Overview of breeding methods</b>	<b>30</b>
5.1. <i>Mass selection</i>	30
5.2. <i>Pure line selection</i>	30
5.3. <i>Pedigree selection</i>	30
5.4. <i>Recurrent selection</i>	31
<i>Online Tutorials</i>	31
<i>Further reading</i>	31
<b>6. Trait selection in agriculture</b>	<b>32</b>
<i>Further reading</i>	34
<b>Section 3 - Molecular procedures and tools</b>	<b>36</b>
<b>7. DNA extraction methods</b>	<b>36</b>
7.1. <i>Extraction methods from different plants</i>	36
7.1.1. Problematic plants for DNA extraction	36
7.1.2. Non-problematic plants for DNA extraction	40
<i>Further reading</i>	41
<b>8. Polymerase Chain Reaction (PCR) – main principles</b>	<b>43</b>
8.1. <i>PCR mixture components</i>	43
8.1.1. Primers – characteristics and principles for design	43
8.1.2. DNA polymerase I	44
8.1.3. DNA template	45

## D 7.1 Production of materials for improved genotyping training

8.1.4. MgCl <sub>2</sub>	45
8.2. Main PCR types	45
Online tutorials	46
Further reading	46
<b>9. Gel electrophoresis</b>	<b>47</b>
Online Tutorials	47
<b>10. Markers and Molecular Tools</b>	<b>48</b>
10.1. Molecular Assessment of Genetic Diversity	49
10.1.1. Biochemical Markers	50
10.1.2. Molecular Markers	50
10.2. Non-PCR-Based Techniques	50
10.2.1. Restriction-Hybridisation Techniques	50
10.3. Markers Based on Amplification Techniques (PCR-Derived)	51
10.3.1. Random Amplified Polymorphic DNA (RAPD)	51
10.3.2. Amplified Fragment Length Polymorphism (AFLP)	52
10.3.3. Sequence Specific PCR Based Markers	52
10.3.4. Microsatellite-Based Marker Technique	53
10.4. Single Nucleotide Polymorphisms (SNPs)	53
10.5. Markers Based on Other DNA than Genomic DNA	54
10.6. Transposable Elements-Based Molecular Markers	54
10.7. RNA-Based Molecular Markers	55
10.8. Real-Time PCR	55
10.9. Diversity Arrays Technology (DArT)	55
10.10. New Generation of Sequencing Technology	56
10.11. Genotyping by sequencing	56
Online Tutorials	57
Further reading	57
<b>11. High-Resolution Melting (HRM)</b>	<b>59</b>
11.1. Genetic Fingerprinting	59
11.2. Mapping Genes and Development of Trait-Linked Markers	59
11.3. Testing Food Products and Seeds	59
11.4. HRM in Autopolyploid Species	59
11.5. Schematic phases of identification of SNPs in genes	61
<b>12. Real Time PCR</b>	<b>64</b>
Online Tutorials	65
Further reading	65
<b>13. Methods/strategies for QTL identification</b>	<b>66</b>
13.1 Types of mapping populations	67
13.2. Statistical analysis in genetic mapping and QTL detection	67
Online Tutorials	67
Further reading	67
<b>14. Molecular-assisted selection - MAS</b>	<b>69</b>
14.1. Marker-assisted backcrossing (MABC)	71
14.2. Marker-assisted gene pyramiding	71
14.3. Marker-assisted recurrent selection (MARS)	72
14.4. Genomic selection (GS)	72
Online Tutorials	72
Further reading	72
<b>15. Marker-assisted selection (MAS) in bean</b>	<b>74</b>
15.1. Introduction	74
15.2. MAS in bean	74
Further reading	76
<b>16. Marker assisted selection for bunt resistance in wheat</b>	<b>77</b>
Further reading	80
<b>17. The use of marker assisted selection (MAS) in organic potato breeding</b>	<b>81</b>
Further reading	82
<b>18. Association genetics</b>	<b>84</b>

## D 7.1 Production of materials for improved genotyping training

18.1. LD principles	84
18.2. Steps in AM	85
18.2.1. Selection of association mapping panel/population	85
18.2.2. Genotyping of the mapping population	85
18.2.3. Analysis of population structure	85
18.2.4. Phenotyping of mapping population for the trait of interest	85
18.2.5. Identifying associations between genotypes and phenotypes	86
Software commonly used in associate genetics	86
Online Tutorials	87
Further reading	87
<b>19. Sequences</b>	<b>88</b>
19.1. What is sequencing	88
19.1.1. DNA sequencing	88
19.2. Comparison of high-throughput sequencing methods	90
19.2.1. RNA sequencing	90
19.2.2. Protein sequencing	90
19.3. Working with sequences	90
19.3.1. Basic Local Alignment Search Tool (BLAST)	91
19.3.2. Online resources for comparative, evolutionary and functional genomics	91
19.4. Applications of sequencing technology in genotyping	91
Software commonly used for sequence analysis	92
Online Tutorials	92
Further reading	92
<b>20. Chromosome engineering</b>	<b>93</b>
20.1. Chromosome engineering in durum wheat	95
20.2. Pyramiding multiple transfers by chromosome engineering in durum wheat	96
<b>21. Fluorescent <i>In-situ</i> hybridization - FISH or GISH</b>	<b>98</b>
21.1. Probe preparation, labelling and control	98
21.2. Preparation and pre-treatment of cytological preparations (on slide)	99
21.3. <i>In-situ</i> hybridisation, of the probe placed on the cytological preparation	99
21.4. Post hybridisation washes and detection	100
21.5. Closure of the slides and counterstaining	101
Online Tutorials	102
Further reading	102
<b>22. Transcriptomics</b>	<b>104</b>
22.1. Alternative splicing (AS) of mRNA	104
22.2. ncRNAs (miRNAs and siRNAs) and their regulatory functions	104
22.3. Technologies for transcriptome analysis	105
22.3.1. Microarrays	105
22.3.2. SAGE (Serial Analysis of Gene Expression)	105
22.3.3. MPSS (Massively Parallel Signature Sequencing)	105
22.3.4. RNA-Seq (whole transcriptome shotgun sequencing)	106
22.4. Applications of transcriptomic analysis in plant research and crop breeding	106
22.5. Transcriptome databases	106
Examples	107
Online Tutorials	107
Further reading	108
<b>23. Biological interpretation of gene expression (transcriptomics) data</b>	<b>109</b>
23.1. GO and MapMan plant ontologies	109
23.2. MapMan	109
23.3. GoMapMan	110
23.4. GSEA	111
Further reading	111
<b>24. Proteomics</b>	<b>113</b>
Online Tutorials	116
<b>Section 4 - Statistics and bioinformatics</b>	<b>117</b>
<b>25. Statistical tests</b>	<b>117</b>

## D 7.1 Production of materials for improved genotyping training

25.1. Averages and dispersion indexes	117
25.1.1. Mean, median and mode	117
25.1.2. Dispersion indexes	118
25.2. Correlation	122
25.3. Regression	124
<b>26. Comparison among methods and statistical software packages to analyse germplasm genetic diversity by means of codominant markers</b>	<b>126</b>
26.1 Gene diversity	126
26.2. Genetic distance	128
26.3. Data Input	129
26.4. Data Analysis	131
26.4.1. GenAEx	131
26.4.2. GDA	135
27.4.3. Popgene	138
26.4.4. Power Marker	139
26.4.5. Cervus	140
26.4.6. Arlequin	141
26.4.7. Structure	142
26.5. Conclusions	144
Online tutorials	145
Further reading	145
<b>27. What is Bioinformatics and sequence management</b>	<b>147</b>
27.1. Bioinformatics applications	147
27.2. Bioinformatic data formats	148
27.3. Downstream analysis	149
Further reading	149
<b>28. Genotyping by Sequence (GBS) Bioinformatics Pipeline</b>	<b>151</b>
28.1. Vocabulary	151
28.2. GBS Discovery Pipeline	151
28.3. Sequence	151
28.4. Tag Counts	152
28.5 Tag alignment (TOPM)	152
28.6. Tags by Taxa (TBT)	153
28.7. SNP calling	153
28.8. GBS Production Pipeline	154
Online tutorial	155
<b>29. Bioinformatics resources</b>	<b>156</b>
29.1. Bioinformatics resources	156
29.1.1. Databases	156
29.1.2. Software, tools and open-source bioinformatics software	157
29.1.3. Web services in bioinformatics	157
29.1.4. Education platforms	157
29.2. Bioinformatics pipelines and workflow management	158
29.2.1. Examples of genotyping bioinformatics pipelines and applications	158
Online resources	159
Further reading	159
<b>30. Data science in R</b>	<b>160</b>
30.1. Packages	160
30.2. Tidyverse	160
30.2.1. Tidyverse and how to use it	160
30.2.2. Tidyverse packages	161
Online resources for R	162
30.3 Case study; Genomic selection in R	162
30.3.1. Install the rrBLUP Package	162
30.3.2. Sample Files	165
30.3.3. Load the Sample Files	165
30.3.4 Impute Missing Markers	167

## D 7.1 Production of materials for improved genotyping training

30.3.4. Training and Validation Populations	168
30.3.5. Run mixed.solve	169
30.3.6. Common Errors	172
<i>Software and resources</i>	173
<b>31. Tutorial I: TASSEL 3.0 Genotyping by Sequencing (GBS) pipeline documentation</b>	<b>174</b>
31.1. <i>Introduction</i>	174
31.2. <i>Recommended directory (folder) structure for a GBS analysis</i>	174
<i>Online resources</i>	185
<b>32. Tutorial II: Association Mapping in Tassel</b>	<b>186</b>
32.1. <i>Reading in data</i>	186
32.2. <i>Fit a generalized linear model (GLM)</i>	186
32.3. <i>Make a Manhattan and QQ-Plot</i>	187
32.4. <i>Calculate a kinship matrix</i>	188
32.5. <i>Run PCA</i>	188
32.6. <i>Run MLM</i>	190
32.7. <i>QQ and Manhattan Plot from MLM</i>	191
32.8. <i>Investigate LD on Chromosome 1</i>	191
32.9. <i>Make an LD Plot</i>	192
32.10. <i>Make a Cladogram</i>	193

## D 7.1 Production of materials for improved genotyping training

### Summary

ECOBREED project activities are aimed at increasing the competitiveness of organic farming. The main and innovative elements are the development of improved genotyping, evaluation of advanced phenotyping, farmer-participatory evaluation and breeding, identification and selection of target traits suitable for an organic production environment to increase the availability of organic seed and varieties in Europe.

Recently, genotyping has recorded a large increase in knowledge and methodologies but not always registering a parallel development of the knowledge and skills for utilising the techniques available. As a result, breeders have often reported difficulties in finding people with advanced genotyping knowledge and expertise. This is also due to the different aspects covered by genotyping requiring competence, among others, in genetics, molecular biology, proteomic, and bioinformatics.

To fulfil this gap ECOBREED project plans to deliver training workshops on advanced genotyping for **researchers, young scientists, breeders, and technicians**. Training workshops will focus on some/all of the following topics: (i) methods/strategies for QTL identification, (ii) association genetics, (iii) marker-assisted breeding, (iv) applications of transcriptomic and (v) proteomic methods/approaches in crop breeding, (vi) genomics and (vii) bioinformatics integration into the field of molecular breeding.

The objective of present training material is to provide a general knowledge on Advanced Genotyping. The material is aimed at providing background knowledge for young researchers and scientists with a new blend of applied and fundamental R&D skills required to, (a) further improve the selection and breeding of organic crops (b) apply breeding and agronomic methodologies/approaches developed in the project to other crops and/or (c) contribute to transferring technologies developed into commercial practice.

The document is divided in several chapters, with suggested further reading material and links to tutorials. The chapters are grouped in four sections which are: 1) Genetic background, 2) Germplasm conservation, evaluation, and utilisation (breeding), 3) Molecular procedures and tools and 4) Statistics and bioinformatics.



## D 7.1 Production of materials for improved genotyping training

### Introduction

Traditional plant breeding, involving quantitative genetic/phenotypic selection approaches which compare genotypes for a wide range of traits over a range of contrasting environments, is extremely time consuming and expensive and is therefore difficult to fully exploit especially for the smaller organic and low-input market sectors. There is therefore an urgent need to develop improved genotyping, phenotyping and farmer-participatory breeding approaches which are more efficient, reduce costs and/or can be used to target traits suited to an organic production environment.

The increase of knowledge and the development of new techniques in genotyping are proceeding very fast. In addition, the decreasing costs in biotechnological approaches and the increase capacity in managing meta-data make it possible to adopt sophisticated procedures also in not very advanced laboratories. Conversely, there is often limited opportunity for detailed training in the several aspects of genotyping such as (a) plant molecular biology, (b) molecular assisted breeding/selection and (c) associated biometrics. As a result, breeding companies and crop research institutions increasingly report difficulties in recruiting Early Stage Researchers, who combine knowledge and skills in the areas of (a) traditional plant breeding, (b) plant genetics and physiology and (c) applied plant molecular biology/biometrics.

The ECOBREED project aims to provide advanced training in genotyping with particular focus of the traits suited to an organic production environment and to facilitate rapid technology transfer from theory into commercial practice. The training workshops will be addressed to **researchers, young scientists, breeders, and technicians**; focusing on some of the following topics: germplasm management and conservation, methods/strategies for QTL identification, association genetics and marker-assisted breeding, applications of transcriptomic and proteomic methods/approaches in crop breeding, use of statistics and bioinformatics to integrate genomics into the field of molecular breeding.

The objective of this training material is to provide a general background covering the various aspects necessary for advanced genotyping. These range from germplasm structure and conservation, plant breeding, molecular markers and tools, proteomics and finally biostatistics and bioinformatics aspects. Therefore, the booklet is divided into several chapters, with suggestions for further reading material and links to tutorials. The chapters are grouped in four sections which are: 1) Genetic Background, 2) Germplasm Conservation, Evaluation and Utilisation (Breeding), 3) Molecular Techniques and Tools, and 4) Statistics and Bioinformatics.

This material, like the training courses offered under the ECOBREED project, is aimed at young researchers and scientists with a new mix of applied and basic R&D skills required to (a) further improve the selection and breeding of organic crops, (b) apply the breeding and agronomic methods/approaches to other crops, and/or (c) contribute to the transfer of the developed technologies into commercial practise. This training material will serve as background material for the ECOBREED workshops on advanced genotyping. The workshops may focus on only part of the



## **D 7.1 Production of materials for improved genotyping training**

topics presented here. Nevertheless, this training material will be distributed to course participants so that they have references on the topics covered in the specific training course, but also on the other topics related to advanced genotyping.

# Section 1 - Genetic background

Objectives:

- To evaluate the range of methods available for genotyping with a particular focus on SNPs and their applications
- To provide an understanding of quantitative genetics
- To provide an understanding of population dynamics together with the different aspects this can be monitored and predicted by the Hardy-Weinberg law.

## 1. What is genotyping?

Genotyping is the process that detects genetic differences that can lead to major changes in phenotype, including both physical differences that provide unique and pathological changes underlying disease. Genotyping examines DNA sequences using biological assays and compares this to reference sequences. It reveals the alleles that have been inherited from the parents. It has a vast range of uses across basic scientific research, medicine, and agriculture.

Traditionally genotyping is the use of DNA sequences to define biological populations by use of molecular tools. It does not usually involve defining the genes of an individual. Current methods of genotyping include the use of molecular markers using restriction enzyme, polymerase chain reaction (PCR), or DNA sequencing. Genotyping is important in research of genes and gene variants associated with traits. Due to current technological limitations, almost all genotyping is partial. That is, only a small fraction of an individual's genotype is determined, such as with (epi) GBS (Genotyping by sequencing) or RADseq. New mass-sequencing technologies promise to provide whole-genome genotyping (or whole genome sequencing) in the future.

When genotyping transgenic plants, a single genomic region may be all that needs to be examined to determine the genotype.

### 1.1 Why we perform SNP genotyping?

SNP genotyping has many applications including:

#### 1.1.1 Disease association

Genome-wide association studies (GWAS) can identify connections between SNPs and common disease risk by comparing the polymorphisms across two different populations (one healthy and one diseased).

In addition to risk stratification, GWAS can begin to unravel the biological processes underlying disease states by identifying potential causal factors.

#### 1.1.2. Population genomics

SNPs also have implications for evolutionary biology, so GWAS can be useful in identifying forms of genetic variation that underlie phenotypic differences between healthy individuals. Understanding this normal genetic variation across different populations helps us to understand how different groups have evolved and diverged which may have implications for protecting species against future environmental challenges.

## D 7.1 Production of materials for improved genotyping training

### 1.1.3. Trait selection in agriculture

Understanding genetic variation has a particular benefit in the agricultural world, where trait selection in plants and livestock has been used for centuries to increase performance, yield and quality.

While traditional selective breeding involved purely observational methods (selecting only plants or animals with superior phenotypic traits, such as size or strength, for breeding), modern selective breeding relies heavily on molecular techniques, including SNP genotyping.

Selective breeding pressures have generated animal breeds and plant varieties with more desirable phenotypes and changes to specific genomic regions associated with these phenotypes. Detecting these functionally relevant genetic changes helps us to understand which particular genes and sequences are associated with specific traits. This is useful for designing new and more intelligent breeding programs.

### 1.1.4. Microorganisms

Single-celled organisms, such as bacteria, also have SNPs. SNP genotyping can discriminate between bacterial isolates and can also be used to characterize strains of antibiotic resistance. SNP-based strain detection is relevant in both clinical and agricultural research and has been used to study a range of infectious diseases in both humans and plants.

#### Online Tutorials

- Brief description of what is genotyping? What does genotyping mean? Genotyping meaning, definition and explanation; <https://www.youtube.com/watch?v=NxqqRNt3ub0>
- High-throughput genotyping solutions for challenges in commercial plant breeding presented by T. Osborn director of Molecular Breeding technology (LCG group); <https://www.youtube.com/watch?v=zvwQjez6AIQ>
- SNP Genotyping Technologies by CD Genomics; <https://www.youtube.com/watch?v=pIWYBLY9OaM>

#### Further reading

- Gaj P, Maryan N, Hennig EE, Ledwon K, Paziewska A. 2012. Pooled sample-based GWAS: A cost-effective alternative for identifying colorectal and prostate cancer risk variants in the Polish population. *PLOS One* 7(4):e35307.
- Kwok P-Y, Chen X. 2003. Detection of single nucleotide polymorphisms. *Curr. Issues. Mol. Biol* 5:43–60.
- Rathnayake I, Hargreaves M, Huygens F. 2011. SNP diversity of *Enterococcus faecalis* and *Enterococcus faecium* in a Southeast Queensland waterway, Australia, and associated antibiotic resistance gene profiles, *BMC Microbiol.* 11(1):201.
- Reddy MPL, Wang H, Liu S, Bode B, Reed JC, Steed RD, She JX. 2011. Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population. *Genes. Immun.* 12(3):208–212.
- Sengstake S, Bablishvili N, Schuitema A, Bzekalava N, Abadia E, de Beer J, Bergval I. 2014. Optimizing multiplex SNP-based data analysis for genotyping of *Mycobacterium tuberculosis* isolates, *BMC Genomics*, 15(1):572.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101(1):5-22.

## D 7.1 Production of materials for improved genotyping training

### 2. Overview of quantitative genetics

The characters studied by Mendel have only two distinct possibilities. The pea seeds were either round or wrinkled, with a colour either yellow or green, etc. The so call qualitative traits have a variability which is not continuous (discrete variation).

Other traits do not fall into discrete classes. Rather, when a segregating population is analysed, a continuous distribution of phenotypes is found. An example is seed size. When a large seed line is crossed with a small seed line, the seed size of the F1 is intermediate to the two parents. Furthermore, when the F1 plants are inter-mated, the distribution of seed size in the F2 ranges from the small to large size. The distribution resembles the bell-shaped curve for a normal distribution.

These types of traits are called continuous traits and cannot be analysed in the same manner as discontinuous traits. Because continuous traits are often measured and given a quantitative value, they are often referred to as quantitative traits, and the area of genetics that studies their mode of inheritance is called quantitative genetics.

Many important agricultural traits such as crop yield, weight gain, plant height, colour intensity, etc. are quantitative traits, and much of the pioneering research into the modes of inheritance of these traits was performed by agricultural geneticists.

Due to the continuous distribution of phenotypic values, quantitative genetics must employ many other statistical methods (such as the effect size, the mean and the variance) to link phenotypes (attributes) to genotypes. Some phenotypes may be analysed either as discrete categories or as continuous phenotypes, depending on the definition of cut-off points, or on the metric used to quantify them. Mendel himself had to discuss this matter in his famous paper, (Mendel, 1866). “Versuche über Pflanzen Hybriden”. *Verhandlungen Naturforschender Verein in Brünn*] especially with respect to his peas attribute tall/dwarf, which actually was “length of stem”. Analysis of quantitative trait loci, or QTL, is a more recent addition to quantitative genetics, linking it more directly to molecular genetics.

The reasons for the normal distribution of quantitative traits have been justified by the Johannsen and Nilsson-Ehle experiments. The first demonstrated that variation in some traits is due to environmental variation which affects the phenotype expression while the second attributed variation to the effect of multigenes with additive actions. These several pairs of genes are called **quantitative trait loci (QTL)** and this is called **polygenic inheritance** or the **multiple-factor hypothesis**.

#### 2.1 Heritability

Heritability measures the fraction of phenotype that can be attributed to genetic variation.

The phenotype can be modelled as the sum of genetic and environmental effects:  
Phenotype (P) = Genotype (G) + Environment (E).

Likewise, the phenotypic variance in the trait  $\sigma^2(P)$  is:

$$\sigma^2(P) = \sigma^2(G) + \sigma^2(E) + 2 \text{Cov}(G,E).$$

## D 7.1 Production of materials for improved genotyping training

In a planned experiment  $\text{Cov}(G,E)$  can be controlled and held at 0. In this case, heritability is:

$$h_b^2 = \sigma^2(G) / \sigma^2(P).$$

$h_b^2$  is the broad-sense heritability. This reflects all the genetic contributions to a population's phenotypic variance including additive, dominant, and epistatic (multi-gene interactions).

A particularly important component of genetic variance is the additive variance,  $\sigma^2(A)$ , which is the variance due to the average effects (additive effects) of the alleles. Since each parent passes a single allele per locus to each offspring, parent-offspring resemblance depends upon the average effect of single alleles. Additive variance represents, therefore, the genetic component of variance responsible for parent-offspring resemblance. The additive genetic portion of the phenotypic variance is known as Narrow-sense heritability and is defined as:  $h_n^2 = \sigma^2(A) / \sigma^2(P)$ .

### Further Reading

- Davies SW, Scarpino SV, Pongwarin T, Scott J, Matz MV. 2015. Estimating Trait Heritability in Highly Fecund Species G3: Genes, Genomes, Genetics 5: 2639-2645. <https://doi.org/10.1534/g3.115.020701>.
- Wray N, Visscher P. 2008. Estimating trait heritability. Nature education, 1(1), 29.

## D 7.1 Production of materials for improved genotyping training

### 3. Population Genetics and Hardy-Weinberg equilibrium

Population genetics evaluates the inheritance in groups of individuals (populations) and includes the genetic constitution of a population and the ways in which it varies from one generation to another.

The **population** is a community of individuals united by kinship and mating relationships. Individuals who exchange a common gene pool (sum of the genotypes of all individuals). The population is the fundamental unit in evolution, because it is the smallest unit in which there is an exchange of genes and therefore alterations on which natural selection and evolution could act.

Since population genetics is about multiple individuals and the mating systems between them, it is not the individual cross that is important, but the pool of crosses that occur between individuals. So, it is not important to know the genotype of a single parent, but the frequency of the genotype in the population and the frequency of occurrence of different alleles. Frequency is the ratio between the event and the total possibility. The genotype frequency is the ratio of the number of individuals with that genotype and the total number of individuals as shown in the following example.

If we have a population with three possible flower colours: red, pink, and white; given respectively by three genotypes,  $C^R C^R$ ,  $C^R C^r$ , and  $C^r C^r$ , their frequency is:

Colour	Genotype	Number	Frequency
red	$C^R C^R$	122	$122/760=0.161$
pink	$C^R C^r$	416	$416/760=0.547$
white	$C^r C^r$	222	$222/760=0.292$
Total		760	1

Since there are 2 alleles per individual the alleles and the homozygote individuals have two alleles of the same type frequency which is:

$$(f)C^R = [(2*122)+416]/(2*760) = (f)R + \frac{1}{2} (f)Rr = 0.4345$$

$$(f)C^r = [(2*222)+416]/(2*760) = (f)r + \frac{1}{2} (f)Rr = 0.5655$$

When the number of alleles present at a locus is greater than two, the calculation of frequencies is based on the same rules that apply for two alleles.

$$k(k + 1) / 2 = \text{number of possible genotypes}$$

with

$$k = \text{number of alleles per locus}$$

#### 3.1. Hardy-Weinberg equilibrium frequency consideration

Population genetics and thus Hardy-Weinberg's law uses Mendelian genetics, but does not take into account the individual crosses of two individuals: a male parent

## D 7.1 Production of materials for improved genotyping training

with a female parent, but the crosses of the entire population, i.e. not focusing on the individual gametes of the parents, but the ones of the entire population. So, in population genetics we need to consider probability effects, more than deterministic ones, and think in terms of frequency (see above).

If in the Mendelian genetics considering the crossing of the father (with genotype Aa) with the mother (with genotype Aa) the gametes of the genotypes are analysed (A and a for both) and therefore their possible assortment (through the Punnet square).

	A	a
A	AA	Aa
a	Aa	aa

Building the Punnet square it is taken for granted that the possible gametes produced by both parents are produced with the same frequencies, 50% A and 50% a.

If, instead, we consider a population the frequencies of the gametes produced by the population are equivalent to their presence in the population and can have any frequency value between 0 and 1 (100%). Obviously in a locus with only two alleles (A and a) the sum of their frequency must be one. Then indicating with p the frequency of allele A and with q that of allele a:

$$p + q = 1$$

This is whether the population is in Hardy-Weinberg equilibrium or not.

Therefore, considering the intersections of the population (and not of a single pair) our Punnet square must be "enriched" with the frequencies of the two alleles in the population:

	A (p)	a (q)
A (p)	AA	Aa
a (q)	Aa	aa

The probability that the three genotypes (AA, Aa, aa) will be formed will not be 25% for each cell but will depend on the frequency of the alleles. It will be:

	A (p)	a (q)
A (p)	AA (p <sup>2</sup> )	Aa (pq)
a (q)	Aa (pq)	aa (q <sup>2</sup> )

$$p^2 (AA) + 2pq (Aa) + q^2 (aa)$$



## D 7.1 Production of materials for improved genotyping training

Let us now consider the frequencies of genotypes (genotypic frequencies) and alleles (gene or allelic frequencies) of a population of N individuals with D dominant, H heterozygotes and R recessive.

$$\begin{array}{l} \text{D AA} \quad \quad \quad \text{D/N} = x \\ \text{H Aa} \quad \text{genotypic frequencies} \quad \quad \text{H/N} = y \quad \quad x + y + z = 1 \\ \text{R aa} \quad \quad \quad \text{R/N} = z \end{array}$$

AA individuals will produce 2D gametes A

Aa individuals will produce H gametes A and H gametes a

Individuals aa will produce 2R gametes a

Therefore, the following gametes will be obtained:

$$A = 2D + H$$

and

$$a = 2R + H$$

with frequencies

$$f(A) = (2D + H) / 2N = p \text{ and } f(a) = (2R + H) / 2N = q$$

$$p + q = 1$$

### 3.2. Hardy- Weinberg equilibrium

The Hardy–Weinberg principle considers the genetic and genotype frequency for a single locus in a population and states: “*allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences*”. A population would be in Hardy-Weinberg equilibrium if 5 conditions (evolutionary forces) would take place:

1. population size – should be large
2. type of mating –should be random
3. absence of mutations
4. absence of migration
5. absence of selection

If the above 5 conditions are met, the population is in Hardy-Weinberg equilibrium, and the consequences are:

1. It keeps constant over time (as long as the 5 conditions exist) the genetic frequencies, that is the frequencies of the alleles. Using the agreed values for allele frequencies, in the case of two alleles, p and q remain constant (with the same percentage) in the different generations.
2. It maintains constant over time (as long as the 5 conditions exist) the genotypic frequencies, that is the frequencies of AA, Aa, and aa.

## D 7.1 Production of materials for improved genotyping training

3. It is possible to compute the gene (allele) frequencies from the genotypic frequencies, and vice versa, by the formula of the developing of the binomial square:  $(p + q)^2 = p^2 + 2pq + q^2$

Development of the binomial that, of course, is always true from the algebraic point of view, but from the genetic point of view it is true only in the case of a population in equilibrium, where  $p$  and  $q$  are the allele frequencies of  $A$  and  $a$ , while  $p^2$ ,  $2pq$ ,  $q^2$  are the genotypic frequencies, respectively for  $AA$ ,  $Aa$  and  $aa$ .

The population reaches the equilibrium in a single generation in which the 5 conditions are verified.

$$\begin{array}{l} p^2 \text{ (AA)} \\ 2pq \text{ (Aa)} \quad (p + q)^2 \\ q^2 \text{ (aa)} \end{array}$$

The above is used if only two alleles  $A$  and  $a$ , are possible for that locus. If instead three alleles occur at a locus, the formula would be a trinomial square development ( $(p+q+r)^2=p^2+q^2+r^2+2pq+2pr+2qr=1$ ) and so on for higher numbers of alleles. It should be noted that the square terms (i.e.  $p^2+q^2+r^2$ , etc.) are homozygote frequencies while the others (i.e.  $2pq+2pr+2qr$ , etc.) are heterozygotes.

### 3.3. Hardy- Weinberg equilibrium, alteration of the conditions

The Hardy-Weinberg law is useful also in the cases where the conditions are not fulfilled. Hereafter are reported the situation for each of the conditions.

#### 3.3.1. Population size – Genetic drift

The variance  $\sigma^2 = pq / 2N$  and standard deviation  $\sigma = \sqrt{pq/2N}$

The smaller the sample, then the greater is the standard deviation and the interval  $\pm 2d\sigma$  in which the gene frequency can vary due only to random effects. Genetic drift is a sampling error due to the limited size of the population (sample). Just as throwing a coin into the area it has a 50% chance of being heads and 50% tails, if the coin is thrown only few times the deviation from the 50/50 value can be very large. For example, if the coin is thrown only 4 times it could easily happen that all heads come out. Similarly, in the case of the assortments of the population, alleles could be fixed, not because they are favourable, but only due to the effect of chance, because the crossings (launches) are few.

In small populations there is genetic drift and a reduction in variability, as in the case of selection. But in the case of selection the alleles that are fixed are those favourable to overcome certain conditions, while in the case of genetic drift the fixed alleles are random, they could be favourable, but also unfavourable. This is why small populations (and genetic drift) are the entry point for extinction: Alleles are randomly fixed; if they are not advantageous, they are eliminated by selection followed by extinction.

Indicating with  $x$  the deterministic effects (selection, mutation and migration) it can be calculated if the variations of the genetic frequency are determined by deterministic effects or by genetic drift if:

$$4 N x \ll 1$$

## D 7.1 Production of materials for improved genotyping training

### 3.3.2. Type of mating

This requirement should be not considered similar to the lack of randomness due to the small number of the population with consequent genetic drift.

There are four possible reasons for lack of casual mating.

- i. unions between genetically similar individuals
- ii. unions between phenotypically similar individuals
- iii. unions between genetically dissimilar individuals
- iv. unions between phenotypically dissimilar individuals

### 3.3.3. Absence of mutations

The mutations could occur with a frequency of about  $p_0u \cong 10^{-5}$

If a mutation occurs with a  $u$  frequency and changes the allele A into the allele a, after a generation the A frequency would be:

$$p_1 = p_0 - p_0u = p_0(1-u)$$

hence after  $t$  generations it would be:  $p_t = p_0(1-u)^t$

But if there is also a retro mutation which changes the allele a into A with a frequency of  $v$ , the final frequency after a generation would be:

$$p_1 = p_0 - p_0u + q_0v$$

so the increment (decrement) of A would be:

$$\Delta p = p_1 - p_0 = p_0 - p_0u + q_0v - p_0 = q_0v - p_0u \rightarrow \Delta p = 0 \rightarrow q_0v = p_0u$$

Reminding that:  $p = (1-q)$

Hence:

$$p = v/(u + v) \quad \text{and} \quad q = u/(u + v)$$

### 3.3.4. Absence of migration

Migration is a term to describe the movement of individuals, often used in the case of animals. Consequently, a more general term is **gene flow**, which also indicates the displacement due to pollen or seeds.

If we indicate with

$m$  = immigrants;  $1-m$  = natives; the allele A would have a frequency of  $q_0$  in native  $q_m$  in the immigrant group.

After a generation of immigration the new frequency  $q_1$  would be:

$$q_1 = mq_m + (1 - m)q_0$$

number of immigrants by their frequency plus number of natives by their frequency.

Developing the formula and grouping equal terms

$$q_1 = m(q_m - q_0) + q_0$$

The change of frequencies in a generation would be:

$$\Delta q = \text{natives} = m(q_m - q_0)$$

## D 7.1 Production of materials for improved genotyping training

It is a function of the migration rate ( $m$ ) for differences in terms of frequency between the two populations ( $q_m - q_0$ ).

A new equilibrium would be reached when  $\Delta q = 0$ , constancy of the frequencies. This would occur when  $m(q_m - q_0) = 0$ , that is or when  $m = 0$ ; zero migration, it is returned in the requisite wanted by the law of HW, absence of migration, or when  $(q_m - q_0) = 0$  and that is when  $q_m = q_0$  in particular, it will be  $q_0$  that will arrive to have the value of  $q_m$  and therefore the arrival of new gene flow with value  $q_m$  will not alter the frequencies. This observation is particularly important in the case of the management of reserves since the knowledge of the direction of the genetic flow and the study of the genetic structure of the population from where this flow arrives (immigrants) gives us information on how the population of the reserve will evolve (natives).

### 3.3.5. Absence of selection

It is not selection that is important, but fitness, adaptive value measured as reproductive capacity i.e. the contribution to the next generation. It describes an individual's reproductive success and is equal to the average contribution to the gene pool of the next generation that is made by individuals of the specified genotype or phenotype. The fitness of a genotype is manifested through its phenotype, which is also affected by the developmental environment.

Selection coefficient ( $s$ ) = 1- selective value ( $w$ )

The selection could act against several different genotypes or genotype combinations. We can compute the function of the allele frequency variation for each of the different selection situations following the schemes with  $\Delta q = q_1 - q_0$ .

#### Selection against recessive homozygotes

	Genotype			Total	a frequency
	AA	Aa	aa		
Initial frequency	$p^2$	$2pq$	$q^2$	1	$q$
Fitness $w$	1	1	$1-s$		
Contribution to the next generation	$p^2$	$2pq$	$q^2(1-s)$	$1-sq^2$	
Normalized frequency	$p^2/(1-sq^2)$	$2pq/(1-sq^2)$	$q^2(1-s)/(1-sq^2)$	1	$q_1 = (q-sq^2)/(1-sq^2)$
Changes in frequency					$\Delta q = (spq^2)/(1-sq^2)$

#### Selection against dominant phenotypes

	Genotype			Total	A frequency
	AA	Aa	aa		
Initial frequency	$p^2$	$2pq$	$q^2$	1	$p$
Fitness $w$	$1-s$	$1-s$	1		
Contribution to the next generation	$p^2(1-s)$	$2pq(1-s)$	$q^2$	$1-s+sq^2$	
Normalised Frequency	$p^2(1-s)/(1-s+sq^2)$	$2pq(1-s)/(1-s+sq^2)$	$q^2(1-s+sq^2)$	1	$p_1 = (p(1-s)/(1-s+sq^2))$
Changes in Frequency					$\Delta p = -(spq^2)/(1-s+sq^2)$

## D 7.1 Production of materials for improved genotyping training

### Selection against heterozygotes

	Genotype				
	AA	Aa	aa	Total	A frequency
Initial frequency	$p^2$	$2pq$	$q^2$	1	q
Fitness w	1	1-s	1		
Contribution to the next generation	$p^2$	$2pq(1-s)$	$q^2$	$1-s+sq^2$	
Normalised Frequency	$p^2/(1-2spq)$	$2pq(1-s)/(1-2spq)$	$q^2/(1-2spq)$	1	$q_1=(q-spq)/(1-2spq)$
Changes in Frequency					$\Delta p=(spq)/(1-2spq)$

### Online Tutorials

- Hardy-Weinberg equation by Khan Academy; <https://www.khanacademy.org/science/biology/her/heredity-and-genetics/v/hardy-weinberg> (Be careful the required conditions are 5 not 3 as mentioned in the video)
- Applying the Hardy-Weinberg equation by Khan Academy; <https://www.khanacademy.org/science/biology/her/heredity-and-genetics/v/applying-hardy-weinberg>
- The Hardy-Weinberg Principle: Watch your Ps and Qs by ThePenguinProfChannel; <https://www.youtube.com/watch?v=oG7ob-MtO8c>
- Hardy Weinberg equilibrium problems with 3 alleles by Shomu's Biology; <https://www.youtube.com/watch?v=Zi7DnMLUudQ>

### Further reading

- Bosco F, Castro D, Briones M. 2012. Neutral and Stable Equilibria of Genetic Systems and the Hardy-Weinberg Principle: Limitations of the Chi-Square Test and Advantages of Auto-Correlation Functions of Allele Frequencies. *Frontiers in Genetics* 3: 276. DOI=10.3389/fgene.2012.00276.

# Section 2 - Germplasm conservation, evaluation, and utilisation (breeding)

Section objectives:

- To provide an understanding of techniques available for germplasm conservation, evaluation, and utilisation in plant breeding.
- To provide knowledge about germplasm classification and relationships among genetic resources.
- To evaluate the tools available for genetic resource conservation.
- To summarize the international agreements in place for germplasm conservation.
- To describe the range of methods available for molecular plant breeding.
- To evaluate traits to be selected for in organic breeding programs

## 4. Genetic resources and their conservation

To run any breeding program and/or to undertake genotypic characterisation, it is important to use a wide range of genetic variation. Hence, it is imperative to properly conserve germplasm and understand the taxonomic relationships and cross-ability potential in the selected plant genetic resources.

The importance of conserving genetic diversity is to have sources of variability for using in breeding programs. This is ever more and more important due to the increase and range of available technologies as well as the urgency of increasing rates of genetic erosion and overall loss of biodiversity.

### 4.1 Species definition and gene pool

The initial taxonomic relationships among species can be obtained by their botanical nomenclature: species belonging to the same genera are genetically closer to each other than are species belonging to different genera. This is also the case for species from same or different tribes, families, etc. The basic concept of species, which is the lower level of the taxonomic definitions, originates from the ability to cross and produce fertile progeny from those individuals belonging to the same species. Species are considered to be reproductive communities, and isolated from other such communities.

The definition of species, in the past, has focused only on the individual's crossing capability: 'species are merely those strongly marked races or local forms which, when in contact, do not intermix, and when inhabiting distinct areas are incapable of producing a fertile hybrid offspring'. In the animal kingdom, a horse and donkey are different species since despite being able to cross; they do not produce fertile offspring. The cross produces either a mule or a hinny, depending upon whether the cross is donkey ( $\sigma$ ) by horse ( $\text{♀}$ ), or horse ( $\sigma$ ) by donkey ( $\text{♀}$ ). Darwin's theories shifted attention away from uniformity to variation and from the general to the specific. Consequently, the definition of species has become more difficult, and now involves different levels of understanding of inter-species hybridisation and the barriers to the

## D 7.1 Production of materials for improved genotyping training

process: morphological, biological, recognition, cohesion, evolutionary, cladistic, ecological and phylogenetic.

The earliest challenge for successful crossing is during the first metaphase of meiosis, where homologous chromosomes face each other along the equator of the cell. The greater the degree of homology between the chromosomes, the more equilibrated the cellular meiosis. When the homology is weak, the first metaphase is unbalanced which results in sterile cells. The probability of balance/unbalance during cell division depends upon the homology between homologous chromosomes and hence with the potential for mutations from gene to genome levels.

For conserving and using plant genetic resources effectively, it is important that species are classified and divided into gene pools, as proposed by Harlan and de Wet (1971). Two individuals belong to: the same primary gene pool (GP-1) if one is a cultivated form (GP-1A) and the other is its wild form (GP-1B); the secondary gene pool (GP-2) if gene transfer between them is possible, but difficult; to the tertiary gene pool (GP-3) if gene transfer is very difficult or impossible unless sophisticated techniques are used, such as embryo culture, etc. We can now also include a fourth gene pool (GP-4), where gene transfer can take place only through the application of genetic engineering. This classification has practical implications for the selection of germplasm for plant breeding programs.

### 4.2 The importance of genetic resources

When developing a comprehensive plant breeding programme, this should address not only the selection of the species, but also the choice of the correct parental material and a programme of crossing activity. The Italian scientist, Nazareno Strampelli, was one of the first to use the potential of genetic resources. In 1900, he performed the first documented cross between wheat varieties, which resulted in a hybrid variety that incorporated leaf rust resistance (from the Rieti variety) with lodging resistance (from the Noè variety). Moreover, he also utilised wheat wild relatives, such as *Dasypyrum* spp. *Secale* spp. in his breeding programs.

During the 1920s and 1930s, the Russian botanist, Nikolai Ivanovich Vavilov, undertook a series of excursions around the world. He noted that species diversity was not randomly distributed but was higher for each crop in a particular region of the world. He identified eight centres of origin of cultivated species. These are still in use and are as follows: (1) Mexico-Guatemala, (2) Peru-Ecuador-Bolivia, (2A) Southern Chile, (2B) Southern Brazil, (3) Mediterranean, (4) Middle East, (5) Ethiopia, (6) Central Asia, (7) Indo-Burma, (7A) Siam-Malaya-Java, and (8) China.

Within each centre of diversity, the variation found is higher for specific cultivated crops. Later studies distinguished between centres of origin i.e., where levels of diversity are high, with the presence of wild progenitors (from where the species probably originated) and centres of diversity, i.e., where the species variation is high, but without the presence of those crop wild relatives. For example, the Middle East is the centre of origin for wheat while Ethiopia is a centre of wheat diversity, with only tetraploid species present. Harlan (1971) then identified larger areas where domestication took place and, considering their magnitude, defined those areas as “not-centre”.

At the Earth Summit, held in Rio de Janeiro, Brazil, in 1992, the Convention on Biodiversity (CBD) was signed. Biodiversity, and the high rates of its loss, was



## D 7.1 Production of materials for improved genotyping training

brought to the attention of the wider public and the focus of the genetic erosion shifted to the issue of broader conservation of biodiversity. The CBD is a framework convention, with a broad mandate on the conservation and sustainable use of biodiversity and the equitable sharing of benefits arising from the use of genetic resources. As defined in Article 2 of the CBD: 'Biological diversity' means the variability among living organisms from all sources including, inter alia, terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems. The CBD, while addressing the wider issues of biodiversity, also shift the focus from conservation approached to the conservation of biological diversity *in-situ* conservation (Art. 8 of CBD).

The Multilateral System of the Treaty (MLS) aims to conserve and use sustainably plant genetic resources for food and agriculture through the fair and equitable sharing of benefits arising out of their use (Articles 10-13). The MLS includes all plant genetic resources for food and agriculture that are listed in Annex I (<http://www.fao.org/plant-treaty/areas-of-work/the-multilateral-system/annex1/en/>), under the management and control of the Contracting Parties and in the public domain. It is in harmony with the Nagoya Protocol of the CBD. «The Contracting Parties agree that benefits arising from the use of plant genetic resources for food and agriculture that are shared under the Multilateral System should flow primarily, directly and indirectly, to farmers in all countries, especially in developing countries, and countries with economies in transition, who conserve and sustainably utilise plant genetic resources for food and agriculture».

### 4.3 Plant breeding and genetic resources

Public awareness of genetic resources conservation has been steadily increasing, especially as a key component of sustainable agriculture as well as mitigating the impact of large-scale monocultures and the consequential effects produced by these human activities. Despite this recognition of the importance of crop genetic resources, which has been well known since plant breeders started their activities (XIX century), but the awareness of its importance and the danger of erosion and disappearance is limited almost exclusively to scientists directly involved. Public concerns are generally focused on the extinction of some endangered minor wild species, but nobody at that level worries about the dramatic impact that shrinking of genetic variation in crop plants and their wild relatives may have on future food production. Since a limited component of modern societies, except farmers, are concerned with agriculture itself. Even if crop genetic resources represent the basis of agricultural development, they provide an enormous reservoir of useful genes and gene complexes that endow plants to cope with evolving resources and habitats. Without the availability of a reserve of variation and different alleles able to let the crops react differently to the different needs, which could be resistance to both biotic and abiotic factors, but also the needs of new products different in colour, shape, test, etc., the breeder would not have the starting material to do his work.

Genes for adaptability are the main resource for affording challenges of a changing environment. Unfortunately, the destruction of natural ecosystems has severely reduced the genetic variability in wild species, whereas the replacement of local varieties with improved ones has virtually eliminated landraces. Most of the

## D 7.1 Production of materials for improved genotyping training

abundant genetic resources available a few decades ago have been lost forever. The concern for genetic resources started during the 1950s when scientists started travelling world-wide and especially in developing countries when it was pointed out what was happening with the cultivation of newly acquired crop land and the spreading of modern uniform varieties. Problems shifted from technical, to financial and finally to political grounds. A wide movement took place; the onset of recombinant DNA technology, at the beginning of the 80s shifted the attention from Genetic Resources to advanced biotechnology as possible source of variation, and to some extent alleviated the concerns. The limits of manipulations based on molecular and tissue culture techniques have been recognised and it is becoming clear that only an integrated approach between traditional and advanced techniques will produce the best results. It is obvious, in fact, that useful genes have to be identified, their role in metabolic pathways clarified, and their membership to gene families and/or complexes ascertained, and this requires the availability of the widest genetic variation, for as many species as possible. For this reason, seed companies started to be interested both in genetic resources and biotechnology.

### 4.4 *In-situ* and *Ex-situ* Strategies of Plant Conservation

*In-situ* conservation focuses on conserving organisms in their natural environment while *ex-situ* conservation entails the preservation of genetic resources in collections external to the natural environment in which the organisms are found e.g. genebanks.

The conservation of plant genetic resources, and more specifically, crop genetic resources, historically focused mainly on *ex-situ* collections as seed collection and seed storage was the main method for plant breeders to maintain their collections. The conservation of PGR *in-situ* and *ex-situ* are often seen as competing conservation strategies. The CBD's Article 9 stresses, however, that the two approaches must be complementary and applied in combination (CBD 1992). It is vital to maintain linkages between these two strategies. One of the key linkages is the use of *ex-situ* material to improve *in-situ* populations or to reintroduce extinct species or varieties into cultivation. Consequently, *ex-situ* materials also perform a role of a safety net as species and varieties may be lost *in-situ* due to extreme events or habitat destruction. Often, however, it is the breeding system of the species which dictates the strategy for preservation. In the case of apomictic species (which produce seed asexually) or in vegetatively propagated species, it may be possible to choose between on-farm conservation, field genebanks as well as techniques such as cryopreservation. Once again, the use of both methodologies would be the most successful path to maximising the chances of survival.

#### 4.4.1 *In-situ* Conservation

*In-situ* conservation maintains the organisms in their original habitat, or agroecological environment, the plants are still undergoing evolution in the field; hence this method of conservation is dynamic. It tends to focus on protected areas such as biospheres and nature reserves but also includes on-farm conservation which has now increased greatly in profile and importance. This approach can simply target one species or the ecosystem as a whole. The latter is now preferred as the ecosystem functions can remain intact while the former is used in cases where an extinct target species requires reintroduction or if there is a key/focal species whose

## D 7.1 Production of materials for improved genotyping training

presence is important to the functioning of the ecosystem. An overview of different forms of *in-situ* conservation is given below:

1. Genetic Reserves: These involve the establishment of protected areas, managed by conservationists with the aim of monitoring and maintaining the diversity present (Maxted et al. 2008). It is most likely that one area will not be sufficient to cover the diversity present and therefore a network of sites is required. A buffer zone and a transition zone for each reserve are also generally recommended (Dulloo et al. 2008).
2. On-farm Conservation: The conservation of diversity within a farming system generally entails the preservation and use of agricultural varieties e.g., landraces which have been maintained by farmers over time to allow for adaptation to local environments. In this form of conservation, not only does the agroecosystem maintain its integrity but the traditional knowledge required for the propagation and survival of those crops and varieties is also conserved. On-farm management, however, includes the introduction of modern cultivars in the agroecosystem, thus affecting landrace diversity originally present and is not generally included in the definition of on-farm conservation (Maxted et al. 2008).
3. Home Gardens: The resources conserved in home gardens are considered a form of *in-situ* conservation as significant amounts of diversity are often found. The produce here is for local consumption, and more specifically, for consumption by the household within which they are grown (Eyzaguirre and Linares 2004). Generally, home gardens are valuable reserves of local fruit, vegetable, medicinal and herb diversity.

### 4.4.2 Ex-situ Conservation

*Ex-situ* conservation maintains the organisms in a place different from where they have been originated and aims to maintain the original gene and allele frequency, hence the conservation is static. Breeders and researchers most often obtain their resources from genebanks and botanical gardens. There are approximately 1500 genebanks around the world, containing 6.5 million samples; 83% of these are held in national government genebanks (Global Crop Diversity Trust 2007). The major crops (wheat, rice, potato, banana/plantain) are well represented in *ex-situ* collections. Crop wild relatives and minor crops (yams, coconut, amaranth) are less well represented but these are now gaining in importance with genebanks accepting regional responsibility for local minor crops (Scarascia-Mugnozza and Perrino 2002). *Ex-situ* collections are also maintained in facilities other than genebanks, such as botanic gardens and field genebanks (Swaminathan 2002). Engels et al. (2008) provide a comprehensive overview with a list of useful references for each of the *ex-situ* conservation methods. A summary of the different types of *ex-situ* conservation is given below:

1. Botanic Gardens and Arboreta: Plants of major socio-economic importance such as medicinal, ornamental, aromatic are the species most often found in these facilities. The plants are most often displayed in gardens with the focus of high species diversity but not on varietal diversity. These gardens often also maintain seed banks to maximise the amount of diversity conserved (Laliberté 1997). The first botanical gardens were set up in Italy during the second half of the XIV century, in Pisa (1544), Padua and Florence (1545) followed by Holland (1593), Paris (1635) and Edinburgh (1690).

## D 7.1 Production of materials for improved genotyping training

2. **Field Genebanks:** Field genebanks are the principal method for conserving species that do not “breed true”, or retain the production characteristics of the parent (i.e. fruit trees, forest plants), or those of which the seeds cannot tolerate the desiccation and cooling used for storage (i.e. recalcitrant seeds) (Fowler and Hodgkin 2004). These plants are kept either in fields or in greenhouses.
3. **Seed and Pollen storage:** This is the most used form of *ex-situ* conservation with 90% of the world’s 6 million accessions stored as seed (de Vicente 2006). The seeds are dried to 3-7% moisture content (5% for pollen) and stored at below zero °C. The seeds need to be assessed at regular intervals for viability (as this tends to decrease with time) and regenerated when necessary. The regeneration process should be undertaken considering that the same diversity and allele frequency should be maintained, hence avoiding actions that change the original gene frequencies.
4. **In-vitro Storage:** In-vitro storage involves the storing of plants in a synthetic nutrient medium in a sterile environment. This approach is most commonly used for species with recalcitrant seeds, apomictic species and vegetatively propagated species. Endangered species are also being conserved using this method. A part of the plant (most often buds and embryos but also stems, leaves and flower buds) is maintained, often in a glass tube, with temperatures set according to the tolerance levels of the individual species. Pence et al. (2002) provide a detailed guide on *in-vitro* collection techniques.
5. **Cryopreservation:** Cryopreservation, very low temperature storage of seeds or embryos, is being developed as an alternative for a number of species. The germplasm is stored in liquid nitrogen (-196°C). A major disadvantage of this technique is that it is expensive as well as being specific to a particular species, sometime even variety specific. The main advantage, however, is that the material may be stored, theoretically, for an unlimited period of time.
6. **DNA Storage:** In this method, DNA is extracted and stored at -20°C in an ethanol solution. This is not a conservation strategy per se but the information obtained is useful for monitoring genetic changes over time, assisting in the development of molecular markers and provides a greater understanding of taxonomy at the genetic level. Thus, alternative conservation measures are required to conserve the germplasm in its entirety. A detailed discussion of this subject is provided in de Vicente (2006).

### Online Tutorials

- Ensuring long-term preservation of rice genetic diversity by IRRI's International Rice Genebank; [https://www.youtube.com/watch?v=RdwclF\\_cm5Y](https://www.youtube.com/watch?v=RdwclF_cm5Y)
- What are Genetic Resources? by Bio Scholar; <https://www.youtube.com/watch?v=EUOMy1Z9Yk8>
- The International Treaty on Plant Genetic Resources (IT PGRFA) by Bio Scholar; <https://www.youtube.com/watch?v=eC3vQBDvNAs>
- Why does plant diversity matter? By Kew Royal Botanical Garden; <https://www.youtube.com/watch?v=ZLC1kapyBjI>

## D 7.1 Production of materials for improved genotyping training

### Further reading

- De Vicente (Ed). 2006. DNA banks – providing novel options for genebanks? Topical Views in Agricultural Biodiversity. Bioversity International, Rome, Italy.
- Dulloo ME, Labokas J, Iriondo JM, Maxted N, Lane A, Laguna E, Jarvis A, Kell SP. 2008. Genetic reserve location and design. In *Conserving Plant Genetic Diversity in Protected Areas* (Eds J.M. Iriondo, N. Maxted and M.E. Dulloo). Cabi International.
- Engels JMM, Maggioni L, Maxted N, Dulloo ME. 2008. Complementing in situ conservation with ex situ measures. In *Conserving Plant Genetic Diversity in Protected Areas* (Eds J.M. Iriondo, N. Maxted and M.E. Dulloo). Cabi International.
- Eyzaguirre PB, Linares OF. 2004. *Home Gardens and Agrobiodiversity*. Smithsonian Books, Washington, DC, USA.
- Fowler C, Hodgkin T. 2004. Plant genetic resources for food and agriculture: assessing global availability. *Annu. Rev. Environ. Resour.* 29, 143-179.
- Harlan JR, De Wet JMJ. 1971. Toward a rational classification of cultivated plants. *Taxon*, 20 (4), 509-517.
- Laliberté B. 1997. Botanic garden seed banks/genebanks worldwide, their facilities, collections and networks. *Bot. Gardens Cons. News*, 2(9), 18-23.
- Maxted N, Iriondo JM, Dulloo ME, Lane A. 2008. The integration of PGR conservation with protected area management. In *Conserving Plant Genetic Diversity in Protected Areas* (Eds Iriondo J.M., Maxted N., Dulloo M.E.). CAB International.
- Pence VC, Sandoval JA, Villalobos VM, Engelmann F. 2002. In vitro collecting techniques for germplasm conservation. *IPGRI Technical Bulletin No.7*. Bioversity International, Rome, Italy.
- Scarascia-Mugnozza GT, Perrino P. 2002. History of ex situ conservation and the use of PGR. In *Managing Plant Genetic Diversity* (J.M.M. Engels, V.R. Rao, A.H.D. Brown and M.T: Jackson Eds). IPGRI, Italy.
- Swaminathan MS. 2002. Conservation and development of genetic diversity. In *Managing Plant Genetic Diversity* (J.M.M. Engels, V.R. Rao, A.H.D. Brown and M.T: Jackson Eds). IPGRI, Italy.



### 5. Overview of breeding methods

#### 5.1. Mass selection

Mass selection is the simplest selection method for autogamous and allogamous crops. For mass selection, the first step is to compose a base population which is a mixture of genotypes. Followed by selecting and harvesting (in bulk) to sow as a mixture in the next season. This way, several cycles would be repeated in consecutive seasons to increase the frequency of desirable traits. Mass selection is divided into a positive and negative selection. Positive selection is to select plants containing desirable traits individually, based on phenotype in a population and harvest in bulk to produce the next generation's population. On the contrary, the method to discard undesirable traits and harvest the remaining plants in bulk is called negative selection. Besides those artificial selections, natural selection is also involved in mass selection. Active artificial selections are conducted by breeders but growing the populations in certain environments also leads to natural selection pressures to increase the frequency of genotypes with desirable traits. So, natural selection will complement the artificial selection in mass selection. This method is simple and easy to control in bulk. In addition, the final homogeneous population could be well-adapted to the local conditions. However, selection is for traits with low heritability, mass selection would not be a suitable method to select desirable traits as the phenotype is not highly represented by its genotype and is more affected by environmental conditions.

#### 5.2. Pure line selection

Pure line selection is a similar method as the positive mass selection, the difference is that the selected plants are harvested separately, and the seeds of each plant are also kept separate in early generations. In the next season, the seeds from each plant are sown and evaluated by breeders based on a single row. In this case, each row would be the criteria to be evaluated, not each single plant as all plants in a row are derived from a homozygous parent. It means that individuals are genetically similar but that phenotypic differences are due to non-genetic variation (within pure lines = heritability is 0). Rows showing low performance and high variability are discarded and the remaining rows are harvested in bulk. Finally, each harvested row in bulk is tested several years and the best pure line will be released in the breeding program.

#### 5.3. Pedigree selection

Pedigree selection is the most common selection method in autogamous crops. Two promising homozygous parents (P1 and P2) are selected by breeders and new traits from P2 are introduced to P1 by removing anthers from P1, to generate the F1 population which is a new combination but not segregating. From the F2 onwards wheat plants containing desirable traits are selected and harvested individually as the F2 population which is segregating. Those individual wheat plants are sown in individual rows to comprise the F3 population and within each row, plants containing desirable traits are selected and harvested individually. At this stage planting rows which show undesirable traits or segregation are completely discarded. The same procedure is repeated up to the F7 population. As generations go by, wheat plants become more homogeneous with an increase in homozygosity. However, if wheat plants containing desirable traits have a low stability, individual wheat plants are not

## D 7.1 Production of materials for improved genotyping training

able to be effectively evaluated exactly based on genotypic variation. The main difference between the pure line and pedigree selection can be explained as individual selection and bulk selection. For pure line selection, each favourable trait is selected and sown again but it will be harvested in bulk in the next season. On the contrary, pedigree selection is based on each single favourable trait. Like in pure line selection, at first, each individual plant is selected and sown again. In the next generation, each single favourable plant is harvested and sown individually within several cycles in order to increase homozygosity among plants with favourable traits.

### 5.4. Recurrent selection

Recurrent selection aims to increase the frequency of desirable traits of several parents and upgrade its germplasm. First of all, several homozygous parents are randomly intercrossed to generate an F1 population which will be intercrossed again. In this case, single cross, three-way cross or double-cross hybrids would be exerted. Those new segregating combinations are sown as a mixture and grown as the F2 population. Each plant in the population containing favourable traits will be selected and harvested individually and sown in a row for another generation. Also, planting rows which show undesirable traits are completely discarded. The selected rows are harvested in bulk and sown again as a bulk population. This new population reverts to the recurrent selection for its second cycle and the procedure will be repeated in several cycles to increase the frequency of favourable traits. The method is used for autogamous and allogamous crops.

#### Online Tutorials

- An Introduction to Plant Breeding by Scholar's wing; <https://www.youtube.com/watch?v=8ATRfaiaOLg>
- Plant Breeding by Science & Agriculture Botany; <https://www.youtube.com/watch?v=zvwRh06kyts>
- Breeding Methods by vikas mangal; <https://www.youtube.com/watch?v=BZJX9mgxuoY>

#### Further reading

- Ali N, Heslop-Harrison JS, Ahmad H, Graybosch RA, Hein GL, Schwarzacher T. 2016. Introgression of chromosome segments from multiple alien species in wheat breeding lines with wheat streak mosaic virus resistance. *Heredity* 117: 114.
- Biagetti M, Vitellozzi F, Ceoloni C. 1999. Physical mapping of wheat-*Aegilops longissima* breakpoints in mildew-resistant recombinant lines using FISH with highly repeated and low-copy DNA probes. *Genome* 42: 1013-1019.
- Ceoloni C, Forte P, Gennaro A, Micali S, Carozza R, Bitti A. 2005. Recent developments in durum wheat chromosome engineering. *Cytogenet. Genome Res.* 109: 328-334.
- Ceoloni C, Forte P, Kuzmanović L, Tundo S, Moschetti I, De Vita P, Virili ME, D'Ovidio R. 2017. Cytogenetic mapping of a major locus for resistance to Fusarium head blight and crown rot of wheat on *Thinopyrum elongatum* 7EL and its pyramiding with valuable genes from a *Th. ponticum* homoeologous arm onto bread wheat 7DL. *Theor. Appl. Gen.* 130: 2005-2024.
- Ceoloni C, Kuzmanovic L, Forte P, Gennaro A, Bitti A. 2014. Targeted exploitation of gene pools of alien Triticeae species for a sustainable and multifaceted improvement of the durum wheat crop. *Cr. Sci.* 65: 96-111.



### 6. Trait selection in agriculture

The development of improved varieties has made a major contribution to the increased productivity and quality of plants used for their food, feed, fibre or esthetic value. Selection of an appropriate variety is one of the key decisions that a farmer has to make since the variety will define the limits of performance that can be achieved in any environment. Plant breeding has been a part of agriculture since humans first selected one type of plant or seed in preference to another instead of randomly taking what nature provided. This led to elimination of undesirable characters such as seed dormancy and shattering. Preferential selection to meet human needs resulted in a broad range of cultivated types within species. The overall objective of plant breeding is to improve those characteristics of a species that contribute to its economic value. The part of plant having economic value may be leaf, stem, root, flower, seed or fruit. Selection can be made for direct improvement of the plant part or for the characters that are related to reliability of production, harvestability and marketability. There is a vast list of characteristics considered by plant breeders. Traits of primary importance for plant breeding common for many species are: yield – the amount of production per unit area; resistance to pests and diseases – genetic resistance is the most effective mean of biological control; seed composition – the value of seed may be influenced by its chemical composition, content, quality and nutritive value; forage quality – animal productivity is related to the quality of the forage consumed; tolerance to mineral stress – where crops are grown on soils with undesirable characteristics; tolerance to environmental stress – temperature and water extremes can cause major reductions in crop productivity; adaptability to mechanisation – modification of certain characteristics enables highly mechanised crop production and harvesting.

Recent world issues concerning environmental protection, agriculture sustainability, food safety and quality have largely impacted as well on potato breeding objectives. Traditionally breeders consider more than 50 traits during field and laboratory evaluation and selection which can be grouped into three categories: yield; quality characteristics of tubers; and resistance to biotic and abiotic stress.

Variety development involves the application of knowledge provided by a number of scientific disciplines and their integration in an effective program. Central disciplines of plant breeding are considered agronomy, horticulture, genetics and in the last decade different 'omics' technologies (e.g. genomics, proteomics, metabolomics). The method by which a crop is produced and utilised determines the characters that are important for selection and the conditions under which the characters should be evaluated. Knowledge of the inheritance of a character is basic, which as well as qualitative and quantitative genetics contribute to the understanding of plant behaviour in breeding process. Today's plant breeding utilises as a foundation the genetic principles initiated by the classic investigations of Gregor Mendel rediscovered in the early 1900s. Mendel employed the sound scientific principle of reducing a complex question to its component parts for study and then bringing the parts together for the final conclusions. He was able to accurately describe inheritance mechanisms based on assumptions of paired units and random transmission of the units from parent to progeny. His laws of segregation and independent assortment are as valid today as at the time when they were

## D 7.1 Production of materials for improved genotyping training

discovered. Since then, a vast number of plant inheritance studies has occurred in some of which the trait of interest was simply inherited. For a trait to be defined as simply inherited a single gene or tightly grouped cluster of linked genes, inherited as a unit, must be responsible. For any given population structure simply inherited traits segregate among progeny at expected Mendelian ratios and include those traits in which a completely dominant phenotype can be qualitatively scored. Quantitative traits are determined by the action and interaction of two or more genes or gene  $\times$  environment interactions and can be defined as one whose genetic component does not follow strict Mendelian inheritance. Many traits of primary interest to breeders are genetically quite complex (e.g. yield, dormancy, nutritional traits) and provide significant analytical challenges requiring dense linkage maps and well replicated sets of phenotypic data. The application of molecular genetics is an important contribution and addition to plant breeding programs. Because agronomic traits are quantitatively inherited, quantitative trait loci (QTL) discovery represents a valuable tool for enhancing yield and yield stability of crop production while maximising its sustainability. Genomics approaches will allow more efficient discovery and manipulation of QTL and will become increasingly important for coping with the challenges faced by crop production. QTL studies allow us to investigate cause-effect relationships between traits. A better understanding of the QTL that underlines these traits would provide new momentum for more targeted selection programs based on marker assisted selection which is already an important component of different breeding programs, particularly in the private sector.

The success of a breeding program is dependent on two main factors: having the necessary variation and being able to manipulate it to produce a stable new variety. In the past and as well as at present the variation exploited in most breeding programs is derived from naturally occurring variants and wild relatives of crop species. The distribution of variability was found by Vavilov, who demonstrated the existence of centres of origin – centres of diversity of cultivated plants in which can be found the highest level of genetic variability of a species. It allows the breeder to identify sources of variation for specific characteristics. Traditionally, diversity is assessed by measuring variation in phenotypic traits such as flower colour, growth habit or quantitative agronomic traits like yield potential, stress tolerance, etc., which are of direct interest to users. This approach has certain limitations: genetic information provided by morphological characters is often limited and expression of quantitative traits is subject to strong environmental influence.

Plant breeders are continually searching for germplasm which might be useful to meet the objectives of their breeding programs. A great diversity is available for most crop species. Conservation of crop germplasm diversity involves the establishment of *in-situ* and *ex-situ* genebanks. The major activities for *ex-situ* genebanks include assembling, conserving, characterising, and providing easy access to germplasm for scientists, breeders, and other users. More than six million accessions are currently assembled in over 1300 genebanks worldwide. The assessment and characterisation of diversity in germplasm collections is important to plant breeders for crop improvement and to genebank curators for the efficient and effective management of their collection. Adequate characterisation with respect to agronomic and morphological traits is necessary to facilitate the utilisation of germplasm. The value of the germplasm collection depends upon the availability of information relative to

## D 7.1 Production of materials for improved genotyping training

the accessions. Morphological and agronomic traits as well as reaction to biotic and abiotic stresses that are known to be in the individual accessions increase the importance of the germplasm. Moreover, systematic description leads to a more efficient use of germplasm in the collection. Morphological and agronomic characters of plants are best scored at different growth stages of the crop; thus, characterisation is done at three different stages, vegetative, reproductive, and at post-harvest stages. Post-harvest characteristics are scored in the laboratory, from the panicle samples that are taken at harvest time. *Ex-situ* conservation of biodiversity can result in large collections that are difficult to characterise, evaluate, utilise, and maintain. An important task for genebank curators is to find a way to preserve the widest range of genetic diversity within crop species as well as to improve the knowledge and utilisation of the genetic resources. To alleviate management difficulties, the identification and use of core collections has been suggested.

The study of phenotypic and genetic diversity to identify groups with similar genotypes is important for conserving, evaluating, and utilising genetic resources, for studying the diversity of pre-breeding and breeding germplasm, and for determining the uniqueness and distinctness of the phenotypic and genetic constitution of genotypes with the purpose of protecting the breeder's intellectual property rights. To pursue these objectives, various types of attributes are commonly measured in each genotype: continuous phenotypic variables such as morpho-agronomic traits (maturity, height, phenology, etc.); discrete phenotypic variables such as grain colour and texture, resistance to diseases and insects, etc. (these are usually multi-state variables); and discrete genetic marker characteristics using molecular markers of choice.

Conventional breeding has contributed to the huge increase in crop yield during the past century. Enhancing further gains in crop production will require a multidisciplinary effort for the identification of agronomical superior alleles and their introgression into desired germplasm. This challenge is very important in view of climate change, the shrinking availability of irrigation water and phosphate fertiliser, the increasing cost of nitrogen fertiliser and fuels, and the necessity to improve the long-term sustainability of crop production.

### Further reading

- Ceoloni C, Forte P, Kuzmanovic L, Tundo S, Moschetti I, De Vita P, Virili ME, D'Ovidio R. 2017. Cytogenetic mapping of a major locus for resistance to *Fusarium* head blight and crown rot of wheat on *Thinopyrum elongatum* 7EL and its pyramiding with valuable genes from a *Th. ponticum* homoeologous arm onto bread wheat 7DL. *Theor. Appl. Genet.* 130: 2005-2024.
- Ceoloni C, Kuzmanovic L, Forte P, Virili ME, Bitti A. 2015. Wheat-perennial Triticeae introgressions: major achievements and prospects. In: Molnár-Láng M, Ceoloni C, Doležel J (Eds.), *Alien Introgression in Wheat - Cytogenetics, Molecular Biology, and Genomics*. Springer, pp. 273–313.
- Ceoloni C, Kuzmanović L, Gennaro A, Forte P, Giorgi D, Grossi MR, Bitti A. 2014. Genomes, chromosomes, and genes of perennial triticeae of the genus *Thinopyrum*: the value of their transfer into wheat for gains in cytogenomic knowledge and 'precision' breeding. In: Tuberosa R, Graner A, Frison E. (Eds.), *Advances in Genomics of Plant Genetic Resources*. Springer, Dordrecht, The Netherlands, pp. 333–358.

## D 7.1 Production of materials for improved genotyping training

- Forte P, Virili ME, Kuzmanović L, Moscetti I, Gennaro A, D'Ovidio R, Ceoloni C. 2014. A novel assembly of *Thinopyrum ponticum* genes into the durum wheat genome: pyramiding Fusarium head blight resistance onto recombinant lines previously engineered for other beneficial traits from the same alien species. *Mol. Breed* 34: 1701–1716.
- Gennaro A, Forte P, Carozza R, Savo Sardaro ML, Ferri D, Bitti A, Borrelli GM, D'Egidio MG, Ceoloni C. 2007. Pyramiding different alien chromosome segments in durum wheat: feasibility and breeding potential. *Isr. J. Plant Sci.* 55: 267-276.
- Gennaro A, Forte P, Panichi D, Lafiandra D, Pagnotta MA, D'Egidio MG, Ceoloni C. 2012. Stacking small segments of the 1D chromosome of bread wheat containing major gluten quality genes into durum wheat: transfer strategy and breeding prospects. *Mol. Breed* 30: 149-167.
- Kuzmanović L, Gennaro A, Benedettelli S, Dodd IC, Quarrie SA, Ceoloni C. 2014. Structural-functional dissection and characterization of yield-contributing traits originating from a group 7 chromosome of the wheatgrass species *Thinopyrum ponticum* after transfer into durum wheat. *J. Exp. Bot.* 65, 509–525.
- Kuzmanović L, Ruggeri R, Able JA, Bassi FM, Maccaferri M, Tuberosa R, De Vita P, Rossini F, Ceoloni C. 2018. Yield performance of chromosomally engineered durum wheat-*Thinopyrum ponticum* recombinant lines in a range of contrasting rain-fed environments. *Field Crops Res.* 228: 147-157.
- Meglič V. 2013. Morpho-agronomic traits. Maloy S.R., Hughes K.T. (ed.) *Brenner's encyclopedia of genetics*. Second ed. vol. 4, 475-477.
- Moore G. 2009. Early stages of meiosis in wheat and the role of *Ph1*. In: Muehlbauer G, Feuillet C (eds) *Genetics and Genomics of the Triticeae*. *Plant Genetics and Genomics: Crops and Models*. Pp237-252.
- Singh RP, Huerta-Espino J, Rajaram S, Crossa J. 1998. Agronomic effects from chromosome translocations 7DL.7Ag and 1BL.1RS in spring wheat. *Crop Sci.* 38: 27–33.
- Vitellozzi F, Ciaffi M, Dominici L, Ceoloni C. 1997. Isolation of a chromosomally engineered durum wheat line carrying the common wheat *Glu-D1d* allele. *Agronomie* 17: 413-419.

### Section 3 - Molecular procedures and tools

Objective:

- To evaluate the available molecular tools by illustrating the different procedures and protocols

#### 7. DNA extraction methods

As molecular marker technology is evolving into a more and more valuable tool for creating new plant cultivars (Kikuchi et al. 2017; Dayteg et al. 2017), it is important to provide good quality, high yield genetic material and a consistent method for its extraction. This can serve as a basis for further molecular genetic analysis (Abdel Latif and Osman 2017), for instance PCR and real time PCR analysis, Southern blotting, restriction enzyme digestion, NGS-based applications, etc. A number of different commercial kits for DNA extraction are available on the market nowadays, differing in isolation technology, sample type and amount; time needed per run, elution volume, DNA yield and potential downstream applications. Most commonly, these kits are based on solid-phase nucleic acid purification (Tan and Yiap 2009) and performed by using a spin column, operated under centrifugal force (Gjerse et al. 2009). That results in fast and efficient DNA purification in comparison to conventional methods, such as cetyltrimethylammonium bromide (CTAB) or sodium dodecyl sulfate (SDS) methods (Tan and Yiap 2009). When preparing plant tissue for the DNA extraction it is very important to consider:

1. the type of plant tissue;
2. that we are selecting the young and healthy parts of the plant tissue;
3. the amount of starting plant material (too much is not always better);
4. how is the starting material stored (fresh, -20°C, -80°C);
5. the appropriate homogenisation method (time of grinding is also important);
6. the use of extra additions to the homogenisation buffer if needed.

#### 7.1. Extraction methods from different plants

##### 7.1.1. Problematic plants for DNA extraction

Plant samples usually contain high amounts of secondary metabolites whose content varies among species. Different commercial kits or DNA extraction methods will thus give different results when used with different plant species and even plant tissue for further molecular applications (Pipan et al. 2013; Derlink et al. 2014; Maras et al. 2015; Rusjan et al. 2015; Sinkovič et al. 2017); therefore, the extraction methods need to be optimised for each material to ensure the best possible outcome (Sahu et al. 2012). Apple leaves contain various polyphenolic compounds, the most frequently mentioned being flavonoids, while phenolic and hydroxycinnamic acids have also been identified (Mikulic Petkovsek et al. 2010). Phenolic compounds bind irreversibly to nucleic acids, making it resistant to different modifying enzymes (Manoj et al. 2007). This can lead to DNA degradation, contamination and low yield (Azmat et al. 2012) and therefore interfere with its use in various types of analyses (Souza et al. 2012). There was a study from Pipan et al. (2018) about the *Comparison of six*



## D 7.1 Production of materials for improved genotyping training

genomic DNA extraction methods for molecular downstream applications of apple tree (*Malus X domestica*). Six different kits were tested, and their performance compared on five to ten samples of apple (*Malus X domestica*) leaves (Table 7.1), genomic DNA was extracted following manufacturers' protocols.

Table 7.1. Commercial kits and samples used for DNA extraction from *Malus X domestica*. Kit 3 includes two different lysis buffers and both of them were tested, each with 5 samples.

Kit no.	Commercial name	Manufacturer	Amount of starting material [mg]	Number of samples
1	E.Z.N.A. SP Plant DNA Kit	Omega Bio-tek	80–90	5
2	E.Z.N.A. Plant DNA DS Mini Kit	Omega Bio-tek	40–50	5
3a	NucleoSpin Plant II – Lysis Buffer PL1	Macherey-Nagel	80	5
3b	NucleoSpin Plant II – Lysis Buffer PL2	Macherey-Nagel	80	5
4	Invisorb Spin Plant Mini Kit	Stratec Biomedical AG	80	5
5	DNeasy Plant Mini Kit	Qiagen	80	5
6	DNeasy Plant Pro Kit	Qiagen	50	8

An amplified by touchdown PCR using 12 simple sequence repeat markers was performed to test DNA quality. The quality of DNA and PCR products was proven on agarose gel; additionally, DNA concentrations were determined quantitatively using fluorimeter Qubit 3.0. (Table 7.2).

Table 7.2. DNA concentrations and PCR amplification ratios from samples of *Malus X domestica*, extracted with various commercial kits.

DNA kit	isolation	DNA concentration [ng/μl]			Amplification ratio
		Min	Max	Average	
1		10.7	> 1000	> 512.9	47%
2		92.8	> 1000	> 503.2	88%
3a		19.2	81.2	61.3	98%
3b		95.6	232.0	138.7	95%
4		48.0	74.4	63.4	95%
5		0.9	75.6	19.5	78%
6		10.7	> 1000	13.0	92%

1 – E.Z.N.A. SP Plant DNA Kit (Omega Bio-tek), 2 – E.Z.N.A. Plant DNA DS Mini Kit (Omega Bio-tek), 3 – NucleoSpin Plant II (Macherey-Nagel); 3a – Lysis Buffer PL1, 3b – Lysis Buffer PL2, 4 – Invisorb Spin Plant Mini Kit (Stratec Biomedical AG), 5 – DNeasy Plant Mini Kit (Qiagen), 6 – DNeasy Plant Pro Kit (Qiagen). Concentrations were measured on Qubit 3.0 fluorimeter with dsDNA Broad Range Assay Kit (Thermo Scientific) with a range from 2 to 1000 ng/μl, therefore concentrations exceeding this were not precisely determined. Amplification ratio was calculated by dividing the number of samples, successfully amplified during reaction, with number of all samples, taking into account 12 SSR markers used in the study. Average DNA concentration and amplification ratio is based on five (kits 1–5) or eight samples (kit 6).

## D 7.1 Production of materials for improved genotyping training

Results show high level of variation for concentrations and DNA purities; the highest yield (more than 512 ng/ $\mu$ l) was obtained with E.Z.N.A. SP Plant DNA Kit (Omega bio-tek), but DNA was not pure. The purest DNA was obtained with DNeasy Plant Pro Kit (Qiagen), which on the other hand resulted in the lowest concentration (13 ng/ $\mu$ l).

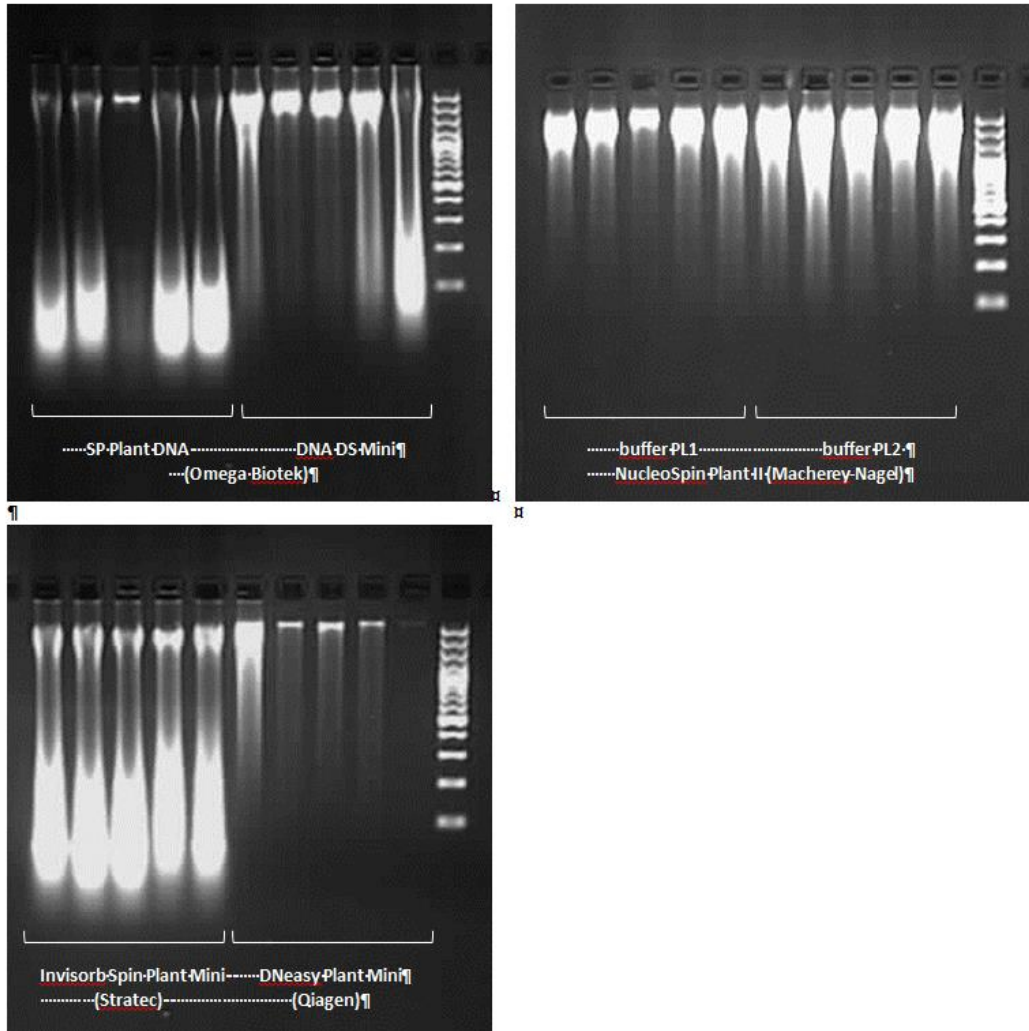


Fig. 7.1: Results from gel electrophoresis, run with 6  $\mu$ l of each DNA sample on 2% agarose gel.

Despite big differences in DNA yields, all kits were sufficient for PCR amplification. It was concluded that choosing suitable methods for different plant species plays a big role in the quality and yield of DNA, and its downstream applications. From the evaluated commercial kits, extraction with DNeasy Plant Pro Kit (Qiagen) was the most efficient, as it resulted in the purest DNA. Despite its relatively low DNA yield, concentration was still high enough for further PCR amplification. Results indicated the most optimal method for DNA extraction for other (problematic) plant species used in molecular studies.

The same as for Apple tree leaves, there are also other agronomically important plant species which are as problematic as apple leaves including *Vitis* plants, blueberries, and buckwheat. For *Vitis*, two already proven related commercial kits (one with two different add-ing buffers in homogenization step) were tested and



## D 7.1 Production of materials for improved genotyping training

compared the results of the DNA extraction success on agarose gel, on spectrophotometer (Nanodrop 2.0) and fluorimeter (Qubit 3.0).

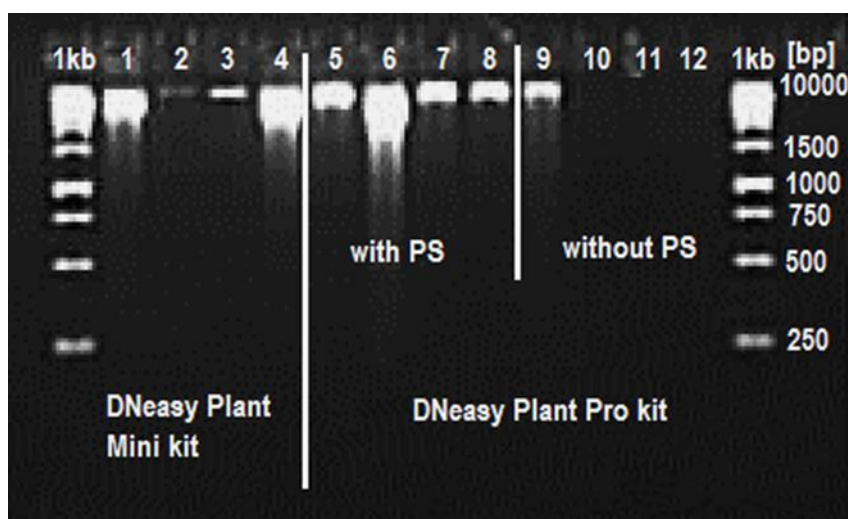


Fig. 7.2. Gel electrophoresis comparison of two extraction kits on *Vitis* young leaves.

Table 7.3. List of plant material included in optimisation procedure and cumulative results of *Vitis* DNA extraction.

Lab label	Actual label	Application of extraction method	Nanodrop concentration [ng/ul]	Nanodrop ratio 260/280*	Nanodrop ratio 260/230**	Qubit concentration [ng/ul]
1	12	DNeasy	49.7	1.79	1.93	198
2	18	Plant Mini	2.9	1.30	0.68	3.40
3	35	Kit	4.5	1.28	1.14	12.9
4	25	(Qiagen)	87.3	1.82	2.03	400
5	36	DNeasy	28.4/30.3	1.89/1.65	0.91/0.74	110
6	33	Plant Pro	94.1	1.80	1.96	400
7	25	Kit	22.6	1.78	1.12	90.0
8	29	(Qiagen) <u>with PS</u>	23.4	1.70	0.63	59.6
9	31	DNeasy	23.3	1.59	0.85	65.4
10	30	Plant Pro Kit	4.2	0.71	0.44	Too low to measure
11	25	(Qiagen) <u>without PS</u>	2.2	3.52	0.34	Too low to measure
12	29		5.2	1.40	0.51	Too low to measure

\*purity of DNA: ratio ~ 1.8 is generally accepted.

\*\*secondary measure of nucleic acid purity which indicates the presence of contaminants (~ 2.0.-2.2 is generally accepted).

In the *Vitis* case, the best option was the use of DNeasy Plant Pro Kit (Qiagen) with PS added in the homogenisation step.

The DNA extraction method used also depends on the genotyping procedure. For the NGS-based application it is important that pure, not degraded and an adequate quantity of DNA is obtained. For some other project including in ECOBREED,

## D 7.1 Production of materials for improved genotyping training

commercial widely used kits have been tested. For GBS of runner bean (*Phaseolus coccineus* L.) DNeasy Plant Mini kit (Qiagen) was used and for GWAS in buckwheat optimisation procedure and quality checks for DNeasy Plant Pro Kit (Qiagen) are being evaluated, instead of the classic CTAB method.

### 7.1.2. Non-problematic plants for DNA extraction

For plant tissues which are classified as not problematic are leaves of beans, brassicas, cereals (e.g., wheat, barley). For those, the automated extraction methods can be used. In the Genetic laboratory of the Agricultural institute of Slovenia, magnetic DNA extraction is being used on two robots with different capacity of the samples operating with different reaction volumes. Both applications are used in combination with commercial kits for magnetic DNA extraction (mostly Qiagen kits as

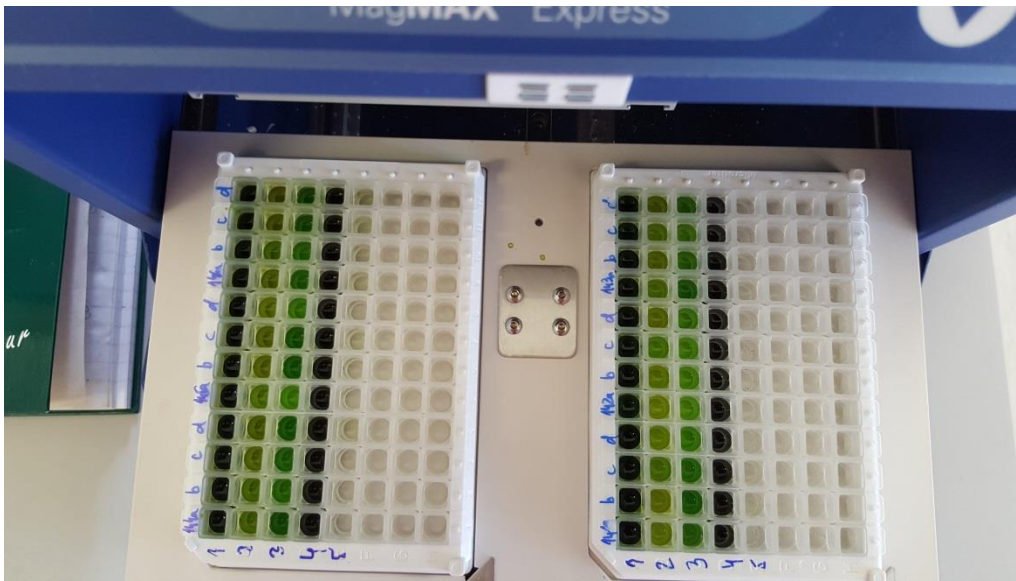


Fig. 7.3. Example of MagMAX™ DNA extraction.

BioSprint 96 DNA Plant kit). First is KingFisher™ system. Those instruments utilise revolutionary particle separation technology that uses permanent magnetic rods and disposable tip combs to move particles through the purification process, providing quality yields with excellent reproducibility and reducing hands-on time compared to manual methods. The second is the MagMAX™ Express magnetic particle processors, which are ideal for researchers who want to automate their MagMAX™ sample preparation kits. The MagMAX™ Sample Preparation System (nucleic acid isolation kits plus magnetic particle processor) offers accelerated and automated nucleic acid isolation utilising MagMAX™ magnetic bead-based technology on MagMAX™ Express Magnetic Particle Processors. MagMAX™ technology was developed to extract both genomic and viral nucleic acid from diverse sample matrices for several downstream applications. An example is for common bean leaf extraction steps on a plate performing on MagMAX™ instrument.

In conclusion it is important to point out that optimised DNA extraction methodology is crucial step when obtaining suitable template for further genotyping applications. Moreover, required time and cost of a particular method should not be ignored, especially when dealing with a high number of samples.

## D 7.1 Production of materials for improved genotyping training

### Online Tutorials

- DNA extraction by Lab Center at DNALC;  
[http://labcenter.dnalc.org/labs/dnaextraction/dnaextraction\\_d.html](http://labcenter.dnalc.org/labs/dnaextraction/dnaextraction_d.html)
- Plant genomic DNA extraction by Rahul Patharkar;  
<https://www.youtube.com/watch?v=p2RARQj0X9Y>
- Plant genomic DNA isolation by Disha Lifesciences Ltd;  
<https://www.youtube.com/watch?v=9-olhcH4NI4>

### Further reading

- Abdel Latif A, Osman G. 2017. Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize. *Plant Methods*. 13(1).
- Azmat MA, Khan IA, Cheema HMN, Rajwana IA, Khan AS, Khan AA. 2012. Extraction of DNA suitable for PCR applications from mature leaves of *Mangifera indica* L. *J Zhejiang Univ-Sci B (Biomed Biotechnol)*. 13(4):239–243.
- Dayteg C, Tuvešson S, Merker A, Jahoor A, Kolodinska-Brantestam A. 2017. Automation of DNA marker analysis for molecular breeding in crops: practical experience of a plant breeding company. *Plant Breed*. 126:410–415.
- Derlink M, Pipan B, Pavlovčič P, Jones, LE, Meglič V, Symondson WO, Virant-Doberlet M. 2014C. Characterization of eleven polymorphic microsatellite markers for leafhoppers of the genus *Aphrodes* (Hemiptera: Cicadellidae). *Conserv Genet Resour*. 6(4):933–935.
- Gjerse DT, Hoang L, Hornby D. 2009. RNA purification and analysis: sample preparation, extraction, chromatography. 1st ed. Weinheim, Germany: Wiley-VCH.
- Kikuchi T, Kasajima I, Morita M, Yoshikawa N. 2017. Practical DNA markers to estimate apple (*Malus X domestica* Borkh.) skin color, ethylene production and pathogen resistance. *J Hortic*. 4(4):211.
- Manoj K, Tushar B, Sushama C. 2007. Isolation and purification of genomic DNA from black plum (*Eugenia jambolana* Lam.) for analytical applications. *Int. J. Biotechnol Biochem*. 3(1):49–55.
- Maras M, Pipan B, Šuštar Vozlič J, Todorović V, Đurić G, Vasić M, Meglič V. 2015. Examination of genetic diversity of Common bean from the Western Balkans. *J. Am. Soc. Hortic. Sci*. 140(4):308–316.
- Mikulic Petkovsek M, Slatnar A, Stampar F, Veberic R. 2010. The influence of organic/integrated production on the content of phenolic compounds in apple leaves and fruits in four different varieties over a 2-year period. *J Sci Food Agric*. 90:2366–2378.
- Pipan B, Šuštar Vozlič J, Meglič V. 2013. Genetic differentiation among sexually compatible relatives of *Brassica napus* L. *Genetika*. 45(2):309–327.
- Pipan B., Zupančič M., Blatnik E., Dolničar P., Meglič V. 2018. Comparison of six genomic DNA extraction methods for molecular downstream applications of apple tree (*Malus X domestica*). *Cogent food & agriculture*, p, no. 4, p. 1-10.
- Rusjan D, Pelengić R, Pipan B, Or E, Javornik B, Štajner N. 2015. Israeli germplasm: phenotyping and genotyping of native grapevines (*Vitis vinifera* L.). *Vitis*. 54:87–89.
- Sahu SK, Thangaraj M, Kathiresan K. 2012. DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *ISRN Mol. biol*. [6 p.]. DOI: 10.5402/2012/205049.
- Sinkovič L, Pipan B, Meglič V, Kunstelj N, Nečemer M, Zlatič E, Žnidarčič D. 2017. Genetic differentiation of Slovenian sweet potato varieties (*Ipomoea batatas*) and effect of different growing media on their agronomic and nutritional traits. *Ital. J. Agron*. 12(4):350–356.

## D 7.1 Production of materials for improved genotyping training

- Souza HAV, Muller LAC, Brandão RL, Lovato MB. 2012. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis*. Genet. Mol. Res. 11(1):756–764.
- Tan SC, Yiap BC. 2009. DNA, RNA, and Protein Extraction: The Past and The Present. J. Biomed. Biotechnol. [10 p.]. DOI: 10.1155/2009/574398.

## D 7.1 Production of materials for improved genotyping training

### 8. Polymerase Chain Reaction (PCR) – main principles

PCR is a widely used method in molecular biology for making copies of a specific DNA segment. Most PCR methods amplify DNA fragment lengths of between 0.1 and 10-kilo base pairs (kbp). By PCR, a single copy (or more) of a DNA sequence is exponentially amplified to generate thousands to millions copies of that particular DNA segment. The main steps of PCR are represented in Fig. 8.1. During the denaturation step (94-96°C), the DNA double helix is separated by breaking the hydrogen bonds between the nucleotides to allow copying of the given fragment. At the annealing step, the temperature is lowered to 50-65°C (depending on the primers used) to allow binding of primers to regions flanking the DNA stretch to be amplified. Finally, at the extension step (68-78°C), DNA polymerase synthesises the new DNA strand based on the DNA template.

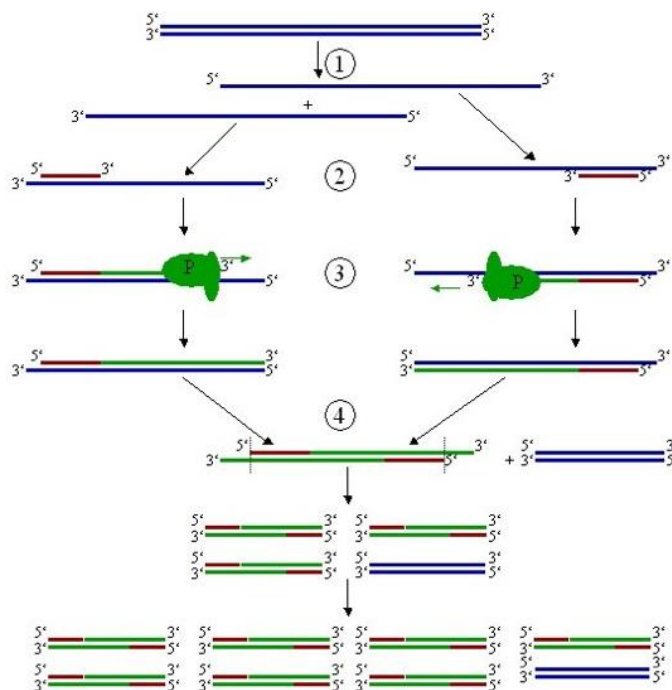


Fig. 8.1. Scheme of a polymerase chain reaction and main stages: initialisation and denaturation (1), annealing (2), extension (3) and final elongation (4) (in blue, the DNA to amplify; in red, primers; in green, the newly synthesised DNA strand; P, DNA polymerase).

#### 8.1. PCR mixture components

##### 8.1.1. Primers – characteristics and principles for design

At the beginning of a PCR, the fragment to be amplified must be flanked by short single-stranded DNA sequences named primers ("forward" -  $P_f$  and "reverse" -  $P_r$ ) which serves as a probe for initiation of the reaction (Fig. 8.2a). Primers are indispensable for PCR because many DNA polymerases i.e. enzymes that catalyse DNA replication, cannot begin an *ex-novo* synthesis of a new strand. Each primer of a primer pair is oriented towards the other one with its 3'-OH terminals and must respect the principle of complementarity between the nucleotides in the site that flanks the stretch of DNA to be amplified (Fig. 8.2b).

## D 7.1 Production of materials for improved genotyping training



Fig. 8.2. Orientation of primers with respect to the DNA sequence to amplify (a) and principle of complementarity between the primer and the DNA sequence (b).

The extension of the two primers during PCR always occurs in the 5' - 3' direction, as indeed happens during the replication *in-vivo*. The primers serve to form an initial duplex with the DNA template filament and to provide the free 3'-OH site that the DNA polymerase I recognizes as suitable for starting replication. Design of primer sequence is one of the most critical factors for PCR and in general, should produce primer pairs where each primer is 17-24 base pair long, has around 50% GC content, and the two sequences must not allow the formation of internal H bonds e.g. primer-dimers or hairpin loops (Fig. 8.3).



Fig. 8.3. Examples of secondary structure that can form within a primer pair due to the excess of complementarity between the sequences of the two member primers.

Annealing temperatures ( $T_a$ ) of the two primers in a pair must be similar and lower than the "melting" temperature of the DNA template sequence. To determine the final  $T_a$  for a given primer pair, Wallace's rule should be initially followed: for each A/T, and each G/C base in the primer sequence, 2 and 4°C are added, respectively, and summed all together [ $T_a = 2 * (A + T) + 4 * (G + C)$ ]. This rule serves to provide an indication of the pairing temperature of a certain primer. The higher the  $T_a$ , the more specific is the pairing. The lower the  $T_a$ , the more possibility there is for a non-specific pairing, which would lead to amplification of products other than the one of interest.

Different types of molecular markers depend on the specific primer sequences (RAPD, AFLP, SSR, EST, etc., see chapter below).

### 8.1.2. DNA polymerase I

The enzyme *Taq* polymerase, normally used in PCRs, was identified and isolated from the DNA of the bacterium *Thermus aquaticus*) and is stable at high temperatures necessary for the denaturation of double-stranded DNA. Its optimal functioning is at 68-72°C, depending on the type of *Taq* and of the producer. The action of this enzyme is based on formation of the phosphodiester bond between a 3'-OH terminal of a primer nucleotide and the alpha phosphate on the free 5'-triphosphate (dATP, dGTP, dCTP, dTTP), with pyrophosphate release (PPi). The enzyme has a 5'- 3' exonuclease activity, with no 3'-5' exonuclease activity.



## D 7.1 Production of materials for improved genotyping training

### 8.1.3. DNA template

The amount of DNA template often determines the specificity of the reaction, the more DNA the more specific is the reaction. Moreover, the quality of the DNA should be sufficient for normal DNA polymerase activity, hence protocols for DNA extraction should eliminate all excessive amounts of proteins, sugars and other compounds that could interfere with the DNA polymerase.

### 8.1.4. $MgCl_2$

Optimal concentration of divalent cations ( $Mg^{++}$ ) is necessary and critical for the *Taq* action. It is important to pay attention that the reaction solution does not contain an excess of chelating agents, such as EDTA, which could capture magnesium, and make it unavailable for the DNA polymerase.

It is essential to apply good laboratory practice to avoid contamination of the PCR mixture by unwanted exogenous DNA, which could represent the biggest problem a successful outcome of the experiment. Therefore, the area where a PCR mixture is prepared must be different from that where the DNA samples are managed (ideally under a hood used exclusively for PCR preparation). Pipettes to be used for PCR mixture preparation should not be used for any other solvent or DNA/RNA, and reagents divided in small aliquots. Frequent change of gloves, thorough cleaning of work surfaces and instrumentation, immediate closure of the tubes immediately after use should also make part of the routine practice.

## 8.2. Main PCR types

**Standard PCR** - use of a common DNA polymerase and conditions of amplification.

Long-range PCR - refers to the amplification of DNA lengths up to 40kbp by using specific methods and reagents. Usually, DNA polymerases with proofreading activity are used and longer extension times. It is used for the analysis/cloning of long DNA fragments, amplification of particularly long gene sequences, or the analysis of chromatin rearrangements.

Asymmetric PCR – refers to preferential amplification of one of the two DNA strands of the template. An excess of one of the two primers is used, leading to the exponential amplification of the targeted strand. This type of PCR is used in sequencing and hybridisation.

High fidelity PCR – uses *Taq* polymerases with high ability of accurate replication of the desired template i.e., with combined low non-incorporation rates and proofreading activity. Its main applications are sequencing of *in-vitro* amplified material, cloning, protein expression or gene studies, SNP analysis, or RNA analysis by RT-PCR (see ahead).

Hot-start PCR - use of DNA polymerase type which can be gradually activated only after exposure to 95°C for 2-15 minutes (depending on the manufacturer). This PCR reduces non-specific amplification during the initial set up stages of the PCR and allows for PCR mixture preparation at room temperature.

Touch down PCR – programmed to perform cycles in which the annealing temperature is progressively lowered during the various phases of the PCR, from an initial higher value with respect to the expected  $T_m$ , to a value lower than the expected  $T_m$ . It aims to reduce non-specific DNA amplification.



## D 7.1 Production of materials for improved genotyping training

**Multiplex PCR** – use of more than 1 primer pair in a single PCR mixture to produce amplicons of various sizes. In this way, multiple polymorphic sequences or genes can be targeted in a single reaction and reduce the time of the overall analyses. Annealing temperatures for each of the primer sets must be optimised to work correctly within a single reaction. In addition, the amplicon profiles i.e., band sizes must be different enough to enable easy visualisation by gel electrophoresis.

**Nested PCR** – PCR that involves two pairs of primers used in two successive PCR reactions, where the second pair of primers amplifies the PCR product of the first one. It increases the specificity of the reaction and reduces the non-specific DNA amplification.

**Real Time PCR** - also known as quantitative PCR (qPCR), represents a method of simultaneous amplification and quantification of DNA. It is commonly used to determine with high precision the number of target DNA copies in the sample. The qPCR uses fluorescent dyes, such as Sybr Green to measure the amount of amplified product in real time.

**Reverse transcriptase Polymerase chain reaction (RT-PCR) for RNA** – technique that combines the reverse transcription of RNA into DNA (i.e. cDNA) and amplification of specific DNA targets using standard PCR. It is primarily used to measure the amount of a specific RNA. RT-PCR can be used without qPCR, to enable molecular cloning, sequencing or simple detection of RNA.

### Online tutorials

- Basis of a PCR by Applied Biological Materials – abm; <https://www.youtube.com/watch?v=matsiHSuoOw>
- Primer3 web site is a widely used program for designing PCR primers, but also for designing hybridization probes and sequencing primers; <http://primer3.ut.ee/>

### Further reading

- BatchPrimer3 v1.0, A high throughput web application for PCR and sequencing primer. <https://probes.pw.usda.gov/batchprimer3/index.html>
- GrainGenes online tool for designing genome-specific primers in polyploidy species. <https://wheat.pw.usda.gov/GG3/node/248>

## D 7.1 Production of materials for improved genotyping training

### 9. Gel electrophoresis

Electrophoresis is a technique used to separate macromolecules such as proteins and nucleic acids, which differ in size, charge or conformation. It is one of the most widely used techniques in biochemistry and molecular biology.

When charged molecules are placed in an electric field, they migrate toward either the positive or negative pole according to their charge. In contrast to proteins, which can have a net positive or negative charge, nucleic acids have a consistent negative charge imparted by their phosphate backbone and migrate toward the anode. Proteins and nucleic acids are electrophoresed within a matrix or “gel”. Most commonly, the gel is casted in the shape of a thin block, with wells for loading the sample. The gel is immersed within an electrophoresis buffer that provides ions to carry a current and some type of buffer to maintain the pH at a relatively constant value.

The gel itself is composed of either agarose or polyacrylamide, each of which has attributes suitable to the task.

**Agarose** is a polysaccharide extracted from seaweed. It is typically used at concentrations from 0.5 to 2%. The higher the agarose concentration the “stiffer” the gel. Agarose gels are extremely easy to prepare; agarose powder should be simply mixed with buffer solution, melted by heating, and then poured. It is also non-toxic. Agarose gels have a large range of separation, but relatively low resolving power. By varying the concentration of agarose, fragments of DNA from about 200 to 50,000 bp can be separated using standard electrophoretic techniques. When preparing agarose gel a fluorescent dye, such as ethidium bromide, is used for staining nucleic acids and for enabling the bands’ visualisation under the UV light, once the electrophoresis is finished. *Ethidium bromide is a known mutagen and should be handled as a hazardous chemical - wear gloves while handling it!*

**Polyacrylamide** is a cross-linked polymer of acrylamide. The length of the polymer chains is dictated by the concentration of acrylamide used, which is typically between 3.5 and 20%. Polyacrylamide gels are significantly more time-consuming and delicate to prepare than agarose gels. As the oxygen inhibits the polymerisation process, they must be poured between glass plates. *Acrylamide is a potent neurotoxin and should be handled with care!* Disposable gloves should be worn when handling solutions of acrylamide, and a mask when weighing out powder. Polyacrylamide is considered to be non-toxic, but polyacrylamide gels should also be handled with gloves due to the possible presence of free acrylamide. Polyacrylamide gels have a rather small range of separation, but very high resolving power. In the case of DNA, polyacrylamide is used for separating fragments of less than about 500 bp. However, under appropriate conditions, fragments of DNA differing in length by a single base pair are easily resolved. In contrast to agarose, polyacrylamide gels are used extensively for separating and characterising mixtures of proteins.

#### Online Tutorials

- Web site illustrating Gel Electrophoresis procedures with videos therein. By Exploratorium Teacher Institute; <https://www.exploratorium.edu/snacks/gel-electrophoresis>
- Agarose Gel Electrophoresis of DNA fragments amplified using PCR by Schools Project. University of Bath; <https://www.youtube.com/watch?v=kjJ56z1HeAc>

### 10. Markers and Molecular Tools

Understanding the molecular basis of the essential biological phenomena in plants is crucial for the effective conservation, management, and efficient utilisation of plant genetic resources (PGR). The assessment of genetic diversity within and between populations is evaluated using morphological, biochemical, and molecular characterisation and evaluation:

1. *Morphological* characterisation does not require expensive technology, but large amounts of land are often required for these experiments, making it possibly more expensive than molecular assessment. These traits are often susceptible to phenotypic plasticity; conversely, this allows assessment of diversity in the presence of environmental variation.
2. *Biochemical* analysis is based on the separation of proteins into specific banding patterns. It is a fast method which requires only small amounts of biological material. However, only a limited number of enzymes are available and thus, the resolution of diversity is limited.
3. *Molecular* analyses comprise a large variety of DNA molecular markers, which can be employed for analysis of variation. Different markers have different genetic qualities (they can be dominant or co-dominant, can amplify anonymous or characterised loci, can contain expressed or non-expressed sequences, etc.).

The concept of genetic markers is not a new one; Gregor Mendel employed phenotype-based genetic markers in his experiments. Later, phenotype-based genetic markers for *Drosophila melanogaster* led to the founding of the theory of genetic linkage, occurring when particular genetic loci or alleles for genes are inherited jointly. The limitations of phenotype-based genetic markers led to the development of DNA-based markers, i.e. molecular markers. A molecular marker can be defined as "*a genomic locus, detected through probe or specific starters (primer) which, by virtue of its presence, distinguishes unequivocally the chromosomal trait which it represents as well as the flanking regions at the 3' and 5' extremities*".

Molecular markers may or may not correlate with phenotypic expression of a genomic trait. They offer numerous advantages over conventional, phenotype-based alternatives as they are stable and detectable in all tissues regardless of growth, differentiation, development, or defence status of the cell. Additionally, they are not confounded by environmental, pleiotropic and epistatic effects.

An ideal molecular marker should possess the following features: (i) be polymorphic and evenly distributed throughout the genome; (ii) provide adequate resolution of genetic differences; (iii) generate multiple, independent and reliable markers; (iv) be simple, quick and inexpensive; (v) need small amounts of tissue and DNA samples; (vi) link to distinct phenotypes; and (vii) require no prior information about the genome of an organism. Nevertheless, no molecular marker presents all the listed advantages.

The different methods of molecular assessment differ from each other with respect to important features such as genomic abundance, level of polymorphism detected, locus specificity, reproducibility, technical requirements and cost. Depending on the need, modifications in the techniques have been made, leading to a second generation of advanced molecular markers.

## D 7.1 Production of materials for improved genotyping training

Genetic or DNA based marker techniques such as Restriction Fragment Length Polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), Simple Sequence Repeats (SSR) and Amplified Fragment Length Polymorphism (AFLP) are now in common use for ecological, evolutionary, taxonomical, phylogenetic and genetic studies of plant sciences. These techniques are well established and their advantages and limitations have been well documented (Ayad et al. 1995; Agarwal et al. 2008).

A new class of advanced techniques has emerged, primarily derived from combination of the earlier, more basic techniques. These advanced marker techniques combine advantageous aspects of several basic techniques. In particular, the newer methods incorporate modifications in the basic techniques, thereby increasing the sensitivity and resolution in detecting genetic discontinuity and distinctiveness. The advanced marker techniques also utilise newer classes of DNA elements such as retrotransposons, mitochondrial and chloroplast-based microsatellites, allowing increased genome coverage. Techniques such as RAPD and AFLP are also being applied to cDNA-based templates (i.e. sequences of complementary DNA obtained by mRNA retrotranscription) to study patterns of gene expression and uncover the genetic basis of biological responses. The recent development of high-throughput sequencing technology provides the possibility of analysing high numbers of samples over smaller periods of time. The present review details the molecular techniques of genetic variability and their application to plant sciences.

### 10.1. Molecular Assessment of Genetic Diversity

Analyses of genetic diversity are usually based on assessing the diversity of an individual using either allozymes (i.e., variant forms of an enzyme that are coded for by different alleles at the same locus) or molecular markers, which tend to be selectively neutral. It has been argued that the rate of loss of diversity of these neutral markers will be higher than those which are associated with fitness. In order to verify this, Reed and Frankham (2003) conducted a meta-analysis of fitness components in three or more populations and in which heterozygosity, and/or heritability, and/or population size were measured. Their findings, based on 34 datasets, concluded that heterozygosity, population size, and quantitative genetic variation, which are all used as indicators of fitness, were all significantly positively correlated with population fitness. Genetic variability within a population can be assessed through the:

1. number (and percentage) of polymorphic genes in the population
2. number of alleles for each polymorphic gene
3. proportion of heterozygous loci per individual

Protein methods, such as allozyme electrophoresis, and molecular methods, such as DNA analysis, directly measure genetic variation, giving a clear indication of the levels of genetic variation present in a species or population (Karp et al. 1996) without direct interference from environmental factors. However, they have the disadvantage of being relatively expensive, time consuming and require high levels of expertise and materials for analysis. Given below is an overview of the different types of markers used for assessing genetic diversity (adapted from Spooner et al. 2005).

## D 7.1 Production of materials for improved genotyping training

### 10.1.1. Biochemical Markers

The use of biochemical markers involves the analysis of seed storage proteins and isozymes. This technique utilises enzymatic functions and is a comparatively inexpensive yet powerful method of measuring allele frequencies for specific genes.

Allozymes, being allelic variants of enzymes, provide an estimate of gene and genotypic frequencies within and between populations. This information can be used to measure population subdivision, genetic diversity, gene flow, genetic structure of species, and comparisons among species out-crossing rates, population structure and population divergence, such as in the case of crop wild relatives. Major advantages of these types of markers consist in assessing co-dominance, absence of epistatic and pleiotropic effects, ease of use, and low costs. Disadvantages of isozymes include: (i) there are only a few isozyme systems per species (no more than 30) with correspondingly few markers; (ii) the number of polymorphic enzymatic systems available is limited and the enzymatic loci represent only a small and not random part of the genome (the expressed part) - therefore, the observed variability may be not representative of the entire genome; (iii) although these markers allow large numbers of samples to be analysed, comparisons of samples from different species, loci, and laboratories are problematic, since they are affected by extraction methodology, plant tissue, and plant stage.

### 10.1.2. Molecular Markers

Molecular markers work by highlighting differences (polymorphisms) within a nucleic sequence between different individuals. These differences include insertions, deletions, translocations, duplications and point mutations. They do not, however, encompass the activity of specific genes.

In addition to being relatively impervious to environmental factors, molecular markers have the advantage of: (i) being applicable to any part of the genome (introns, exons, and regulation regions); (ii) not possessing pleiotropic or epistatic effects; (iii) being able to distinguish polymorphisms which do not produce phenotypic variation and finally, (iv) being some of them co-dominant. The different techniques employed are based either on restriction-hybridisation of nucleic acids or techniques based on Polymerase Chain Reaction (PCR), or both. In addition, the different techniques can assess either multi-locus or single-locus markers. Multi-locus markers allow simultaneous analyses of several genomic loci, which are based on the amplification of casual chromosomal traits through oligonucleic primers with arbitrary sequences. These types of markers are also defined as *dominant* since it is possible to observe the presence or the absence of a band for any locus, but it is not possible to distinguish between heterozygote ( $a/-$ ) conditions and homozygote for the same allele ( $a/a$ ). By contrast, single-locus markers employ probes or primers specific to genomic loci and can hybridise or amplify chromosome traits with well-known sequences. They are defined as co-dominant since they allow discrimination between homozygote and heterozygote loci.

## 10.2. Non-PCR-Based Techniques

### 10.2.1. Restriction-Hybridisation Techniques

Molecular markers based on restriction-hybridisation techniques were employed relatively early in the field of plant studies and combined the use of *restriction endonucleases* and the hybridisation method (Southern 1975). Restriction



## D 7.1 Production of materials for improved genotyping training

endonucleases are bacterial enzymes able to cut DNA, identifying specific palindrome sequences and producing polynucleotidic fragments with variable dimensions. Any changes within sequences (i.e. point mutations), mutations between two sites (i.e. deletions and translocations), or mutations within the enzyme site, can generate variations in the length of restriction fragment obtained after enzymatic digestion.

RFLP and Variable Numbers of Tandem Repeats (VNTRs) markers are examples of molecular markers based on restriction-hybridisation techniques. In RFLP, DNA polymorphism is detected by hybridising a chemically-labelled DNA probe to a Southern blot of DNA digested by restriction endonucleases, resulting in a differential DNA fragment profile. The RFLP markers are relatively highly polymorphic, co-dominantly inherited, highly replicable, and allow the simultaneous screening of numerous samples. DNA blots can be analysed repeatedly by stripping and re-probing (usually eight to ten times) with different RFLP probes. Nevertheless, this technique is not very widely used as it is time-consuming, involves expensive and radioactive/toxic reagents and requires large quantities of good quality genomic DNA. Moreover, the pre-requisite of prior sequence information for probe construction contributes to the complexity of the methodology. These limitations led to the development of a new set of less technically complex methods known as PCR-based techniques.

### 10.3. Markers Based on Amplification Techniques (PCR-Derived)

The use of this kind of marker has been exponential, following the development by Mullis et al. (1986) of the Polymerase Chain Reaction (PCR). This technique consists of the amplification of several discrete DNA products, derived from regions of DNA which are flanked by regions of high homology with the primers. These regions must be close enough to one another to permit the elongation phase.

The use of random primers overcame the limitation of prior sequence knowledge for PCR analysis and being applicable to all organisms, facilitated the development of genetic markers for a variety of purposes. PCR-based techniques can further be subdivided into two subcategories: (1) arbitrarily primed PCR-based techniques or sequence non-specific techniques; and (2) sequence targeted PCR-based techniques. Based on this, two different types of molecular markers have been developed: RAPD and AFLP.

#### 10.3.1. *Random Amplified Polymorphic DNA (RAPD)*

RAPDs were the first PCR-based molecular markers to be employed in genetic variation analyses. RAPD markers are generated through the random amplification of genomic DNA using short primers (decamers), separation of the obtained fragments on agarose gel in the presence of ethidium bromide and finally, visualisation under ultraviolet light. The use of short primers is necessary to increase the probability that, although the sequences are random, they are able to find homologous sequences suitable for annealing. DNA polymorphisms are then produced by “rearrangements or deletions at or between oligonucleotide primer binding sites in the genome” (Williams et al. 1991). As this approach requires no prior knowledge of the genome analysed, it can be employed across species using universal primers. The major drawback of this method is that the profiling is dependent on reaction conditions which can vary between laboratories; even a difference of a degree in temperature is sufficient to

## D 7.1 Production of materials for improved genotyping training

produce different patterns. Additionally, as several discrete loci are amplified by each primer, profiles are not able to distinguish heterozygous from homozygous individuals (Bardakci 2001). Arbitrarily Primed Polymerase Chain Reaction (AP-PCR) and DNA Amplification Fingerprinting (DAF) are independently developed methodologies, which are variants of RAPD. For AP-PCR, a single primer, 10–15 nucleotides long, is used and involves amplification for initially two PCR cycles at low stringency. Thereafter, the remaining cycles are carried out at higher stringency by increasing the annealing temperatures.

### **10.3.2. Amplified Fragment Length Polymorphism (AFLP)**

To overcome the limitation of reproducibility associated with RAPD, AFLP technology was developed by the Dutch company, Keygene (Vos et al. 1995). This method is based on the combination of the main analysis techniques: digestion of DNA through restriction endonuclease enzymes and PCR technology. It can be considered an intermediate between RFLPs and RAPDs methodologies as it combines the power of RFLP with the flexibility of PCR-based technology.

The primer pairs used for AFLP usually produce 50–100 bands per assay. The number of amplicons per AFLP assay is a function of the number selective nucleotides in the AFLP primer combination, the selective nucleotide motif, GC content, and physical genome size and complexity. AFLP generates fingerprints of any DNA regardless of its source, and without any prior knowledge of the DNA sequence. Most AFLP fragments correspond to unique positions on the genome and hence can be exploited as landmarks in genetic and physical mapping. The technique can be used to distinguish closely related individuals at the sub-species level and can also map genes.

The origins of AFLP polymorphisms are multiple and can be due to: (i) mutations of the restriction site which create or delete a restriction site; (ii) mutations of sequences flanking the restriction site, and complementary to the extension of the selective primers, enabling possible primer annealing; (iii) insertions, duplications or deletions inside amplification fragments. These mutations can cause the appearance/disappearance of a fragment or the modification (increase or decrease) of an amplified-restricted fragment.

### **10.3.3. Sequence Specific PCR Based Markers**

A different approach to arbitrary PCR amplification consists of the amplification of target regions of a genome through specific primers. With the advent of high-throughput sequencing technology, abundant information on DNA sequences for the genomes of many plant species has been generated. Expressed Sequence Tags (EST) of many crop species have been generated and thousands of sequences have been annotated as putative functional genes using powerful bioinformatics tools. ESTs are single-read sequences produced from partial sequencing of a bulk mRNA pool that has been reverse transcribed into cDNA. EST libraries provide a snapshot of the genes expressed in the tissue at the time of, and under the conditions in which, they were sampled (Bouck and Vision 2007). Despite these advantages, however, EST-SSRs are not without their drawbacks. One of the concerns with SSRs in general is the possibility of null alleles, which fail to amplify due to primer site variation, do not produce a visible amplicon. Because the cDNA from which ESTs are derived lack introns, another concern is that unrecognised intron splice sites could



## D 7.1 Production of materials for improved genotyping training

disrupt priming sites, resulting in failed amplification. Lastly, as EST-SSRs are located within genes, thus more conserved across species, they may be less polymorphic than anonymous SSRs. Although the use of EST possesses these limitations, several features of EST sequence libraries make them a valuable resource for conservation and evolutionary genetics. ESTs are an inexpensive source for identifying gene-linked markers with higher levels of polymorphism, which can also be applied to closely related species. EST libraries are also a good starting point for developing tools to study gene expression such as microarrays or quantitative PCR assays.

### 10.3.4. *Microsatellite-Based Marker Technique*

Microsatellites or Simple Sequence Repeats (SSR) are sets of repeated sequences found within eukaryotic genomes (Morgante and Olivieri 1993). These consist of sequences of repetitions, comprising basic short motifs generally between 2 and 6 base-pairs long. Polymorphisms associated with a specific locus are due to the variation in length of the microsatellite, which in turn depends on the number of repetitions of the basic motif. Variations in the number of tandemly repeated units are mainly due to strand slippage during DNA replication where the repeats allow matching via excision or addition of repeats (Schlotterer and Tautz 1992). As slippage in replication is more likely than point mutations, microsatellite loci tend to be hypervariable. Microsatellite assays show extensive inter-individual length polymorphisms during PCR analysis of unique loci using discriminatory primers sets.

Microsatellites are highly popular genetic markers as they possess: co-dominant inheritance, high abundance, enormous extent of allelic diversity, ease of assessing SSR size variation through PCR with pairs of flanking primers and high reproducibility. However, the development of microsatellites requires extensive knowledge of DNA sequences, and sometimes they underestimate genetic structure measurements, hence they have been developed primarily for agricultural species, rather than wild species. Initial approaches were principally based on hybridisation techniques, whilst more recent techniques are based on PCR (Gupta and Varshney 2000). Major molecular markers based on assessment of variability generated by microsatellites sequences are: STMSs (*Sequence Tagged Microsatellite Site*), SSLPs (*Simple Sequence Length Polymorphism*), SNPs (*Single Nucleotide Polymorphisms*), SCARs (*Sequence Characterised Amplified Region*) and CAPS (*Cleaved Amplified Polymorphic Sequences*).

### 10.4. **Single Nucleotide Polymorphisms (SNPs)**

Single nucleotide variations in genome sequence of individuals of a population are known as SNPs. SNPs are the most abundant molecular markers in the genome. They are widely dispersed throughout genomes with a variable distribution among species. The SNPs are usually more prevalent in the non-coding regions of the genome. Within the coding regions, when an SNP is present, it can generate either non-synonymous mutations that result in an amino acid sequence change (Sunyaev et al. 1999), or synonymous mutations that do not alter the amino acid sequence. Synonymous changes can, however, modify mRNA splicing, resulting in phenotypic differences. Improvements in sequencing technology and an increase in the availability of the increasing number of EST sequences have made analysis of genetic variation possible directly at the DNA level.

## D 7.1 Production of materials for improved genotyping training

Many SNP genotyping analyses are based on allele-specific hybridisation, oligonucleotide ligation, primer extension or invasive cleavage (Sobrino et al. 2005). Genotyping methods, including DNA chips, allele-specific PCR and primer extension approaches based on SNPs, are particularly attractive for their high data throughput and for their suitability for automation. They are used for a wide range of purposes, including rapid identification of crop cultivars and construction of ultra-high-density genetic maps.

### 10.5. Markers Based on Other DNA than Genomic DNA

There are also other highly informative approaches used to study genetic variation based on organelle microsatellite sequence detection; in fact, due to their uniparental mode of transmission, chloroplast (cpDNA) and mitochondrial genomes (mtDNA) exhibit different patterns of genetic differentiation compared to nuclear alleles (Provan et al. 1999; Breidenbach et al. 2019). Consequently, in addition to nuclear microsatellites, marker techniques based on chloroplast and mitochondrial microsatellites have also been developed. The cpDNA, maternally inherited in most plants, has proved to be a powerful tool for phylogenetic studies. Due to increasing numbers of recent examples of intra-specific variation observed in cpDNA, there is additional potential for within-species genetic variation analysis. CpDNA has been preserved well within the genome, and consequently has been employed widely for studying plant populations through the use of PCR-RFLP and PCR sequencing approaches. They are also employed in the detection of hybridisation/introgression (Bucci et al. 1998), in the analysis of genetic diversity and in obtaining the phylogeography of plant populations.

Mitochondrial DNA in plants, in contrast, has been demonstrated to be an unsuitable tool for studying phylogenesis and genetic diversity, being quantitatively scarce. At the nuclear level, another type of sequence employed largely for studying genetic diversity is ribosomal RNA (rRNA). Ribosomal RNA genes are placed on the specific chromosomal loci *Nor* and organised in tandem repeats which can be repeated up to thousands of times. Since some regions of rRNA are well preserved in eukaryotes, it represents a very useful phylogenetic tool. Conversely, other regions such as the “Internal Transcriber Spacers” (ITS) are so variable that they can be used to analyse polymorphism at the intra-specific level.

### 10.6. Transposable Elements-Based Molecular Markers

Although transposon insertions can have deleterious effects on host genomes, transposons are considered important for adaptative evolution, and can be instrumental in acquiring novel traits (Miller et al. 1997; May and Dellaporta 1998; Girard and Freeling 1999). Retrotransposons have so far received little attention in the assessment of genetic diversity, despite their contribution to genome structure, size, and variation. Additionally, their dispersion, ubiquity and prevalence in plant genomes provide an excellent basis for the development of a set of marker systems, to be used alone or in combination with other markers, such as AFLPs and SSRs. Retrotransposon-based molecular analysis relies on amplification using a primer corresponding to the retrotransposon and a primer matching a section of the neighbouring genome. To this type of class of molecular markers belong: *Sequence-Specific Amplified Polymorphism* (S-SAP), *Inter-Retrotransposon Amplified Polymorphism* (IRAP), *Retrotransposon-Microsatellite Amplified Polymorphism*

## D 7.1 Production of materials for improved genotyping training

(REMAP), *Retrotransposon-Based Amplified Polymorphism* (RBIP) and finally, *Transposable Display* (TD).

### 10.7. RNA-Based Molecular Markers

Studies of mechanisms which control genetic expression are essential to better understand biological responses and developmental programming in organisms. PCR-based marker techniques such as cDNA-SSCP, cDNA-AFLP and RAP-PCR are used for differential RNA studies, using selective amplification of cDNA.

### 10.8. Real-Time PCR

Real-time polymerase chain reaction is a laboratory technique based on the polymerase chain reaction, amplifying and simultaneously quantifying a targeted DNA molecule (Heid et al. 1996). It enables both detection and quantification (as absolute number of copies or relative amount when normalised to DNA input or additional normalising genes) of a specific sequence in a DNA sample. The procedure follows the general principle of polymerase chain reaction; its key feature is that the amplified DNA is quantified as it accumulates in the reaction in *real time* after each amplification cycle. Two common methods of quantification are: (i) the use of fluorescent dyes that intercalate with double-stranded DNA and (ii) modified DNA oligonucleotide probes that fluoresce when hybridised with a complementary DNA. The major advantage of this technique consists in its sensitivity and speed due to the system of detection (spectrophotometric respect to ethidium bromide) and the quick changes of temperature. Real-time PCR is, therefore, particularly suitable for molecular markers based on PCR amplifications. In fact, the number of conservation and phylogenetic studies are now increasingly using real-time PCR for assessment of genetic variation (Pagnotta et al. 2009).

### 10.9. Diversity Arrays Technology (DArT)

DArT is a generic and cost-effective genotyping technology. It was developed to overcome some of the limitations of other molecular marker technologies such as RFLP, AFLP and SSR (Akbari et al. 2006). DArT is an alternative method to time-consuming hybridisation-based techniques, typing simultaneously several thousand loci in a single assay. DArT is particularly suitable for genotyping polyploid species with large genomes, such as wheat. This technology generates whole-genome fingerprints by scoring the presence/absence of DNA fragments in genomic representations generated from samples of genomic DNA. DArT technology consists of several steps: (i) complexity reduction of DNA; (ii) library creation; (iii) the microarray of libraries onto glass slides; (iv) hybridisation of fluoro-labelled DNA onto slides; (v) scanning of slides for hybridisation signal and (vi) data extraction and analysis. DArT acts by reducing the complexity of a DNA sample to obtain a “representation” of that sample. The main method of complexity reduction used relies on a combination of restriction enzyme digestion and adapter ligation, followed by amplification. However, an infinite range of alternative methods can be used to prepare genomic representations for DArT analysis. DArT markers for a new species are discovered by screening a library of several thousand fragments from a genomic representation prepared from a pool of DNA samples that encompass the diversity of the species. The microarray platform makes the discovery process efficient because all markers on a particular DArT array are scored simultaneously. For each complexity reduction method, an independent collection of DArT markers can be

## D 7.1 Production of materials for improved genotyping training

assembled on a separate DArT array. The number of markers for a given species, therefore, is only dependent on: (i) the level of genetic variation within the species (or gene pool); and (ii) the number of complexity reduction methods screened.

### 10.10. New Generation of Sequencing Technology

The recent development of “high throughput sequencing” technologies makes DNA sequencing particularly important for conservation biology. These technologies have the potential to remove one of the major impediments to implementing genomic approaches in non-model organisms, including many of conservation relevance, i.e., the lack of extensive genomic sequence information. These technologies, in fact avoid the expense, complication, and biases associated with traditional clone-based sequencing by using direct amplification of DNA templates (Bentley 2006; Mardis 2008). The three pre-eminent technologies to be commercialized are 454 (Roche), Solexa (Illumina), and SOLiD (Applied Biosystems). The 454 sequencing is a pyrosequencing-based method that utilises emulsion PCR to achieve high throughput, parallel sequencing. Solexa’s sequencing-by-synthesis (SBS) approach is based on a simplified library construction method and reversible fluorescence termination chemistry in the sequencing reaction, which produces 35-bp reads. Supported oligonucleotide ligation and detection (SOLiD) sequencing has some features in common with the other two technologies but, unlike the other two technologies, uses ligation-based sequencing technology. These new approaches to DNA sequencing enable the generation of 0.1–4 gigabases of DNA sequence in one to seven days with reagent costs being between US\$ 3,400 and 8,500. Due to the differences in fragment read lengths of sequencing, the target of each of these technologies is different: the shorter length and lower price per base of Solexa and SOLiD. This makes these approaches well suited to whole genome resequencing, where a novel genome sequence can be assembled and then compared to a reference sequence, that is, when the genome sequence of the species already exists. The 454 sequencing, on the other hand, with longer read lengths (soon to be upward of 400 bp per sequence) can also be used for obtaining the first glimpse of a species’ genome or transcriptome.

### 10.11. Genotyping by sequencing

With the reduction of sequence cost and the speed up of the procedures next generation sequencing (NGS) technologies allow large-scale genome wide variation in populations to be obtained. Hence, genotyping by sequencing (GBS) has become popular to identify large scale variation in species both with and without a reference genome. Using GBS it is possible to identify thousands of single nucleotide polymorphism (SNP) markers, which can be used to analyse genetic variation within and between populations, and facilitate the analysis and dissection of complex traits, especially those involved in adaptive selection (Elshire 2011). GBS has two advantages (a) lower cost compared to the other techniques to identify SNPs in different species and crops and (b) it provides satisfactory results in the characterisation of germplasm, population studies and breeding.

## D 7.1 Production of materials for improved genotyping training

### Online Tutorials

- Markers Molecular/Genetic/DNA, Biochemical and Phenotypic by XploreBio; <https://www.youtube.com/watch?v=Quk-Dh65iHY>
- SNP (single nucleotide polymorphism) by XploreBio; <https://www.youtube.com/watch?v=aaJkGFrWzFQ>
- Introduction to Sequencing by Synthesis. Illumina Sequencing by Synthesis. By Illumina; <https://www.youtube.com/watch?v=fCd6B5HRaZ8>
- SSR Marker? Causes of SSR variation? Advantages, how to design SSR marker? by XploreBio; <https://www.youtube.com/watch?v=iGN2tFCLPZ0>
- genotyping mapping population using SSR marker by Genomics Lab; <https://www.youtube.com/watch?v=dI0ITCBxNgE>
- Genetic Markers | RAPD, RFLP, AFLP by Shomu's Biology; <https://www.youtube.com/watch?v=JVM4LpCuT7g>

### Further reading

- Agarwal M, Shrivastava N, Padh H. 2008. Advances in molecular marker techniques and their application in plant sciences. *Plant Cell Rep.* 27: 617-631.
- Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, Uszinski G, Mohler V, Lehmsiek A, Kuchel H, Hayden M.J, Howes N, Sharp P, Vaughan P, Rathmell B, Huttner E, Kilian A. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor. Appl. Genet.* 113: 1409-1420.
- Ayad WG, Hodking A, Jaradat A, Rao VR. 1995. Molecular genetic techniques for plant genetic resources. IPGRI Workshop: Rome, Italy.
- Bardakci F. 2001. Random amplified polymorphic DNA (RAPD) markers. *Turk. J. Biol.* 25: 185-196.
- Bentley DR. 2006. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 16: 545-552.
- Bouck A, Vision T. 2007. The molecular ecologist's guide to expressed sequence tags. *Mol. Ecol.* 16: 907-924.
- Breidenbach N, Gailing O, Krutovsky KV. 2019. Development of novel polymorphic nuclear and chloroplast microsatellite markers in coast redwood (*Sequoia sempervirens*). *Plant Genetic Res.* 17(3), 293-297.
- Bucci G, Anzidei M, Madaghiele A, Vendramin GG. 1998. Detection of haplotypic variation and natural hybridization in halepensis complex pine species using chloroplast simple sequence repeat (SSR) markers. *Mol. Ecol.* 7: 1633-1643.
- Elshire RJ. 2011. A Robust, Si 571 mple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *Plos One* 6, doi:10.1371/journal.pone.0019379.
- Girard L, Freeling M. 1999. Regulatory changes as a consequence of transposon insertion. *Dev. Genet.* 25: 291-296.
- Gupta PK, Varshney RK. 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113: 163-185.
- Heid CA, Stevens J, Livak KJ, Williams PM. 1996. Real time quantitative PCR. *Genome Res.* 6: 986-994.
- Karp A, Seberg O, Buiatt, M. 1996. Molecular techniques in the assessment of botanical diversity. *Ann. Bot.*, 78: 143-149.
- Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24: 133-141.
- May BP, Dellaporta SL. 1998. Transposon sequences drive tissue-specific expression of the maize regulatory gene R-s. *Plant J.* 13: 241-248.



## D 7.1 Production of materials for improved genotyping training

- Miller W, McDonald J, Pinsker W. 1997. Molecular domestication of mobile elements. *Genetica* 100: 261-270.
- Morgante M, Olivieri AM. 1993. PCR-amplified microsatellites as markers in plant genetics. *Plant J.* 3: 175-182.
- Mullis KB, Faloona FA, Scharf SJ, Saiki SK, Horn GT, Erlich HA. 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor. Symp. Quant. Biol.*, 51: 263-273.
- Nadeem MA., Nawaz MA., Shahid M.Q., Doğan Y., Comertpay G, Yıldız M, Baloch FS. 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing, *Biotechnology & Biotechnological Equipment*, 32:2, 261-285, DOI: 10.1080/13102818.2017.1400401.
- Pagnotta MA, Mondini L, Porceddu E. 2009. Quantification and organization of WIS2-1A and BARE-1 retrotransposons in different genomes of *Triticum* and *Aegilops* species. *Molecular Genetics and Genomics*. 282: 245-255.
- Provan J, Russell JR, Booth A, Powell W. 1999. Polymorphic chloroplast simple-sequence repeat primers for systematic and population studies in the genus *Hordeum*. *Mol. Ecol.* 8: 505-511.
- Reed D.H, Frankham R. 2003. Correlation between fitness and genetic diversity. *Cons. Biol.*, 17: 230-237.
- Schlotterer C, Tautz D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* 20: 2211-2215.
- Southern E. 1975. Detection of specific sequences among DNA fragments separated by gel-electrophoresis. *J. Mol. Biol.* 98: 503.
- Spooner D, van Treuren R, de Vicente MC. 2005. *Molecular Markers for Genebank Management*. Bioversity International: Rome, Italy.
- Sunyaev S, Hanke J, Aydin A, Wirkner U, Zastrow I, Reich J, Bork P. 1999. Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *J. Mol. Med.* 77: 754-760.
- Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23: 4407-4414.
- Williams JGK, Kubelik AR, Livak K.J, Rafalski JA, Tingey SV. 1991. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531-6535.

### 11. High-Resolution Melting (HRM)

High resolution melting (HRM) is a new generation of mutation scanning and genotyping technology. The detection and study of spontaneous and induced mutations is essential for; the understanding of changes in phenotypic variation, evolution of species and exploration of natural or induced genetic diversity as a viable strategy for the improvement of all major food crops. Whereby HRM analysis is a method to facilitate rapid and accurate mapping, sequencing, and analysis of genes, enabling rapid, high-throughput identification of variants in the regions of interest without sequencing. Recent plant applications of HRM analysis are reported by Simko (2016).

#### 11.1. Genetic Fingerprinting

Genetic fingerprinting is used for a survey of single nucleotide polymorphisms (SNPs) and simple sequence repeat (SSR) markers and detecting insertions and deletions in plant species. Because HRM can simplify screening of candidate loci, the method has been applied for rapid identification of high-quality loci that can be used for designing molecular markers (Arthofer et al. 2011; Mondini et al. 2012).

#### 11.2. Mapping Genes and Development of Trait-Linked Markers

The HRM method has been applied to the development of molecular linkage maps, mapping, tagging, cloning of genes, and for selecting material with a desirable combination of genes (Wang-Pruski 2012).

#### 11.3. Testing Food Products and Seeds

It is used for DNA-based identification and differentiation of cultivars and closely related species, genotyping pathogenic microorganism in food, screening for genetically modified organisms, uncovering fraudulent substitutions and detection of food allergens (Druml and Cichna-Markl 2014).

#### 11.4. HRM in Autopolyploid Species

In addition to SNP discovery and validation, HRM has the ability to detect allelic dosage in polyploid species. HRM is a post-PCR method that allows detecting polymorphism in double-stranded DNA by comparing profiles of melting curves (Fig. 11.1) and has been developed to detect SNPs (single nucleotide polymorphisms) in small PCR amplicons. This technique measures temperature-induced strand separation and is therefore able to detect variations as small as one base difference between samples.

PCR is performed in the presence of a dye that binds to double-stranded DNA (dsDNA). This dye shows low levels of fluorescence when unbound but is highly fluorescent in the bound state (Fig. 11.1-2). After PCR is completed the amplicon (typically 50 to 500 bp long) is gradually denatured by increasing temperature in steps of ca. 0.01-8°C to 0.2-8°C. During this stage, termed melting analysis, the gradually denaturing amplicon releases the fluorescent dye.



## D 7.1 Production of materials for improved genotyping training

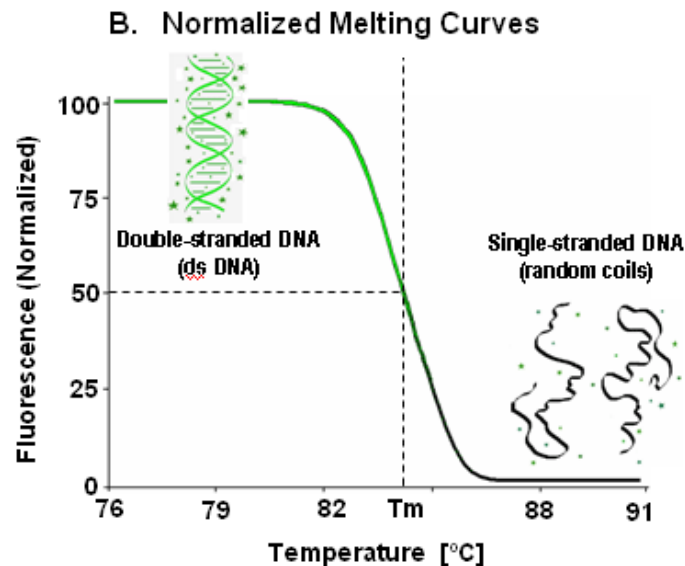


Fig. 11.1. Example of a melting curve plot. Highlighted the point where the signal light emission falls as a function of the temperature increase at the instant in which the denaturing of the double strand of the labelled DNA occurs, losing fluorescence (from <http://hrm.gene-quantification.info/>).

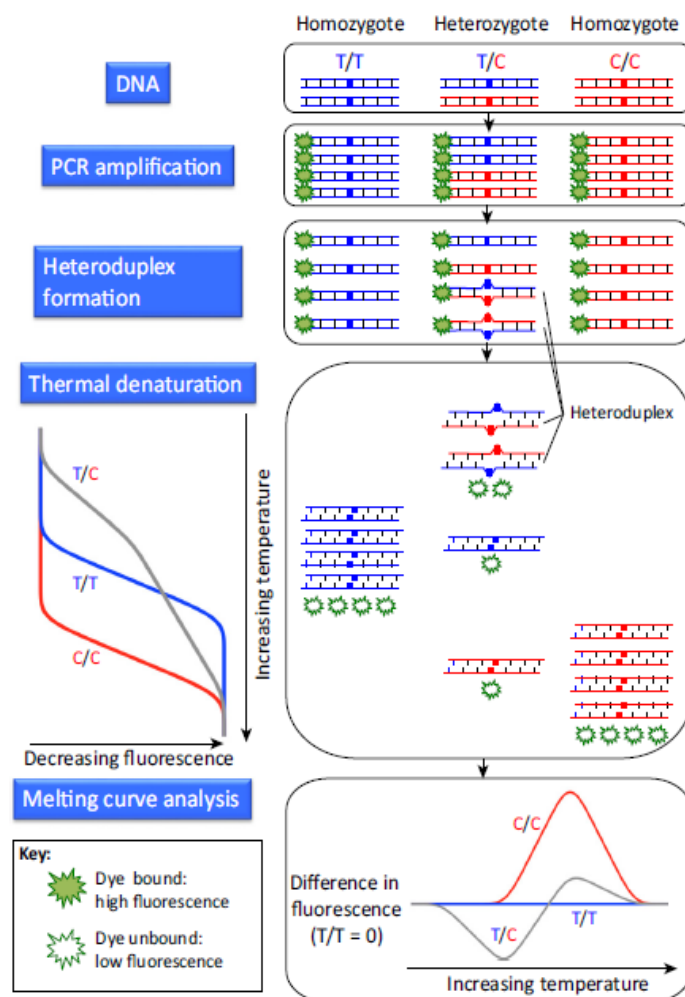
A melting curve can be charted from the diminishing fluorescence emission values plotted against increasing temperature (Fig. 11.2). The unique shape of the melting curve depends on the amplicon's length, sequence, GC content and complementarity of DNA strands.

Since DNA melting is such a simple process that requires no more than PCR and a generic DNA dye, efforts to increase its information content eventually led to the use high resolution melting curve analysis. High resolution melting was made possible by progress along three fronts, dye chemistry, instrument resolution and data analysis (Erali and Twitter 2010).

As examples, multi-well instruments with greater practical utility were introduced to the market. The firsts multi-well HRM instruments were the Rotor-Gene 6000 (Corbett Life Science) and the LightScanner (Idaho Technology). These two instruments were introduced at about the same time but employed fundamentally different technical innovations to achieve HRM. The more recently introduced LightCycler 480 (Roche Molecular Systems) is capable of HRM and thermal cycling.

## D 7.1 Production of materials for improved genotyping training

### Schematic Overview of High-Resolution DNA Melting (HRM) Analysis



Trends in Plant Science

Fig. 11.2. Double-stranded DNA (dsDNA) extracted from tissue is used as a template in PCR amplification. Fluorescent dye that binds to dsDNA is added to the reaction. The dye shows high fluorescence in the bound state but shows low levels of fluorescence when unbound (for simplicity, dsDNA dye binds are indicated only at the end of PCR amplicons). After the final PCR extension, heteroduplex formation is facilitated by subjecting amplicons to 95 °C for 30 s, followed by cooling to 25 °C for 30 s (this step can be omitted if only different homoduplexes are analysed). If a heterozygote sample (T/C in this example) contains amplicons of two different alleles in equal amount, approximately 50% of all amplicons will form heteroduplexes, while homoduplexes of each allele will represent approximately 25% of all amplicons. When homoduplexes and heteroduplexes are submitted to slowly increasing temperature, they gradually denature (melt), releasing fluorescent dye. In this example, denaturation of heteroduplexes (T–G, C–A) occurs at the lowest temperature, followed by denaturation of T–A homoduplexes, and then by denaturation of C–G homoduplexes. The graph on the left side shows decreasing fluorescence recorded for each of the three genotypes (T/T, T/C, and C/C) as plotted against increasing temperature. The final plot range of melting temperatures (from Simko 2016).

### 11.5. Schematic phases of identification of SNPs in genes

**Step 1:** As a prerequisite, HRM accuracy depends on high quality PCR, therefore from the most accurate knowledge of the gene sequence/s to be analysed, obtained

## D 7.1 Production of materials for improved genotyping training

from the dedicate databases, on which primers are designed, with specific software and optimisation of PCR reactions.

**Step 2:** PCR amplifications: HRM are best performed on real time PCR thermocycler, therefore the amplification mixture will contain the necessary reagents for amplification (see chapter 12) and a specific fluorescent dye.

**Step 3:** Melt sample in real time: consisted of post-PCR phases with a decreasing ramp temperature (about 75-95°C) that automatically followed the PCR reaction.

**Step 4:** Data analysis: before HRM curves are plotted, the raw data is first normalised. Melt curves are normally plotted with fluorescence on the Y axis and temperature on the X axis. This is similar to real-time PCR amplification plots but with the substitution of temperature for cycle number. As with real-time PCR plots, the fluorescence axis of HRM plots is normalised onto a 0 to 100% scale (Fig. 11.3).

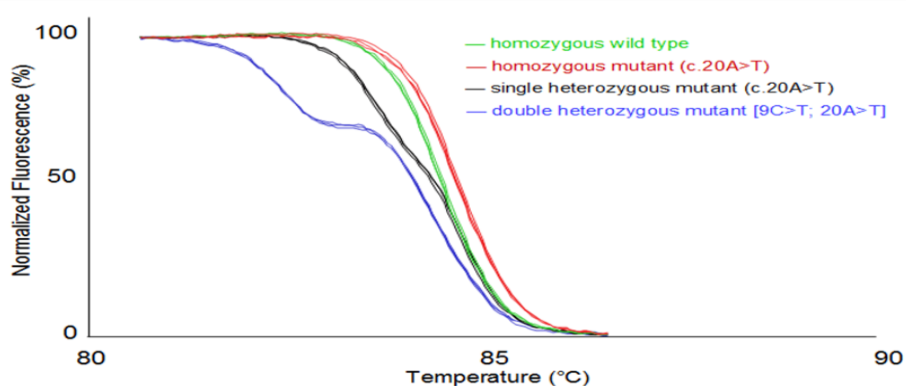


Fig. 11.3. HRM analysis by comparison of different normalized melting curves for different DNA samples analysed for the selection of possible mutations in specific DNA sequences distinguishable by the different denaturation point linked to the increase in temperature and the consequent drop of the emitted fluorescence, making it possible to identify the different genotypes (from: <http://hrm.gene-quantification.info/>).

Because different genetic sequences melt at slightly different rates, they can be viewed, compared, and detected using these curves. Even a single base change will cause differences in the melting curve.

To sum up, a melting curve is then performed using high data acquisition rates, and data are finally analysed using an instrument-specific gene scanning software, by three basic steps (by: <http://hrm.gene-quantification.info/>) (Fig. 11.4):

- i. **Normalisation:** the pre-melt (initial fluorescence) and post-melt (final fluorescence) signals of all samples are set to uniform, relative values from 100% to 0% (Figure 11.4)
- ii. **Temperature shifting:** the temperature axis of the normalised melting curves is shifted to the point where the entire double-stranded DNA is completely denatured. Samples with heterozygous SNPs can then easily be distinguished from the wild type by the different shapes of their melting curves (Fig. 11.4)
- iii. **Difference Plot:** the differences in melting curve shape are further analysed by subtracting the curves from a reference curve. This helps cluster samples automatically into groups that have similar melting curves (e.g. those that are heterozygote as opposed to homozygotes) (Fig. 11.4).

# D 7.1 Production of materials for improved genotyping training

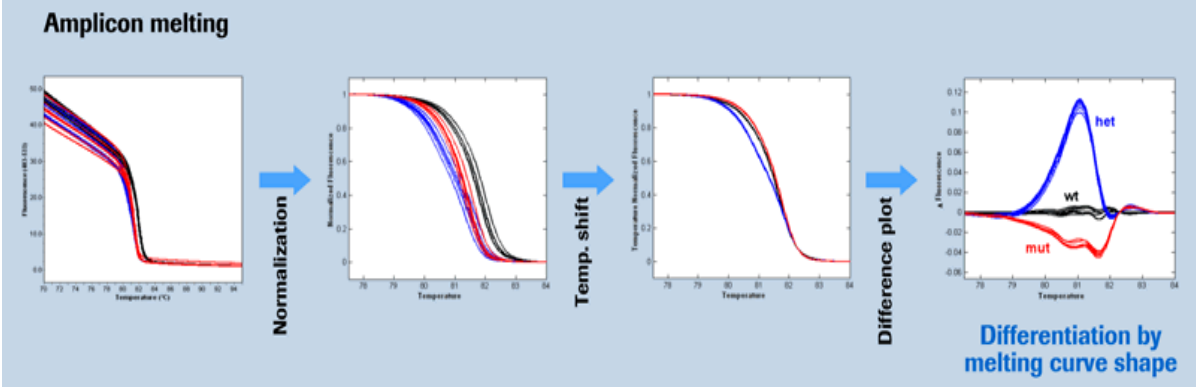


Fig. 11.4: Genotype differentiation by melting curve analyses (source: gene-quantification.info/).

### 12. Real Time PCR

In this chapter is shown how the use of the Real-time PCR technique can be used to assess the genetic variability present in five accessions of einkorn (*T. monococcum*,  $2n = 2x = 14$ ), emmer (*T. turgidum*,  $2n = 4x = 28$ ) and spelt (*T. spelta*,  $2n = 6x = 42$ ). Variability as copy number, present in the genome, of one of the most important retrotransposons present in *Triticum* genomes, WIS2-1A. WIS2-1A is the first retrotransposon found in wheat (Harberd et al. 1987) and was primarily observed as an insertion into a High-Molecular-Weight (HMW) storage protein gene in *T. aestivum*. It represents an ancient DNA element that probably was already present in the common diploid ancestor of the *Triticae* tribe. Retrotransposons, in general, are particularly important in the *Triticum* species, since about 80% of the wheat genome consists of repetitive elements (Smith and Flavell 1975). The development of real time PCR has made possible direct monitoring of the amplification reaction during its progress.

The model real-time PCR amplification curve of fluorescence intensity against cycle number was sigmoidal with the amplification product from the early cycles remaining below the limit of detection. However, assay progress was monitored from the onset of the log-linear phase. At the “threshold cycle” ( $C_T$ ), fluorescence became detectable, and since the cycle at which this occurred was dependent upon the quantity of template,  $C_T$  was used to estimate template quantity during the quantitative real-time PCR progress. Copy number was then determined by comparing the amplification curve, derived from each test sample, with the standard curve. One-way analysis (SPSS version 14 software) was performed to test differences in copy number between ploidy levels.

Real-time PCR assay has advantages over conventional methods based on agarose gel electrophoresis: it does not include several laborious post-PCR handling steps; its specificity is directly confirmed by melting curve profiling and it has a great sensibility. The results of present example demonstrate the difference between species, but also within species, in WIS2-1A presence. The copy number of WIS 2-1A retrotransposons in einkorn varied, excluding the calibrator, from 2 to 10 per ng of template in einkorn, from 11 to 27 in emmer, from 19 to 27 in spelt, from 8 to 42 in *Ae. speltoides*, from 26 to 30 in *Ae. tauschii*, from 5 to 10 in *Ae. sharonensis* and from 20 to 28 in *T. urartu*. Furthermore, as expected, the lowest copy number was observed for *T. monococcum* which represent the diploid level present among hulled wheat. On the other hand, a similar number of copies have been observed in *T. dicoccum* (tetraploid) and in *T. spelta* (exaploid). In line with the results of Moore *et al.* (1991) that in species such as *T. aestivum* (A, B and D genomes), *Ae. longissima* (S genome), *Ae. squarrosa* (the proposed donor of D genome), and *T. monococcum*, revealed that the genome B has a higher number of WIS 2-1A retroelements than the A and D genomes. In a previous study in barley a strong correlation between the retrotransposons copy number and genome size was observed, more recently in wheat it has been demonstrated that the wheat genome A has a higher transposable element content than the B and D genomes. Therefore, this work confirms previous results where it has been observed that the A ancestral genome may have undergone differential genome expansion caused by Class I elements prior to speciation of

## D 7.1 Production of materials for improved genotyping training

the tetraploid wheat ancestor; hence the number of retrotransposons is not linearly linked to the ploidy level of the wheat species.

*Ae. speltoides* presents the highest number of copies and in *T. urartu* and *Ae. tauschii* have been observed several copies like *T. dicoccum* and *T. spelta*. These results are surprising, as confirmed by recent work, the relative retrotransposon copy number it is not in strict relation with the ploidy level. The highest number of copies observed in species considered to be the donors than in species considered to be recipients, can be justified considering the genome size reduction due to *Illegitimate Recombination* events. This shows the great evolutionary importance of transposable genome elements. In fact, in addition to polyploidisation, retrotransposon amplification has been a major cause of genome expansion; in contrast to this process, a counterbalancing force has been demonstrated to be the “unequal homologous recombination” between LTRs of different elements belonging to the same family, resulting in a net loss of DNA and in particular a deletion of the internal portion of retrotransposons (Devos et al. 2002, Jianxin et al. 2004, Gu et al. 2006).

Since species have different amounts of ‘C’ due to their polyploidy level it is interesting to compute the number of retrotransposons per unit of DNA content. WIS 2-1A copy number per unit of DNA content was similar in einkorn, emmer and spelt.

### Online Tutorials

- Real-time PCR page web with video by .r-biopharm; <https://food.r-biopharm.com/it/tecnologie/real-time-pcr/>
- Real Time PCR - Basic simple animation by MrSimpleScience; <https://www.youtube.com/watch?v=EaGH1eKfvC0>
- The principle of Real Time PCR, Reverse Transcription, quantitative rt-PCR by Biomedical and Biological Sciences; [https://www.youtube.com/watch?v=DH7o9Df5\\_50](https://www.youtube.com/watch?v=DH7o9Df5_50)
- Web page on What is Real-Time PCR (qPCR)? By Biorad; <https://www.bio-rad.com/it-it/applications-technologies/what-real-time-pcr-qpcr?ID=LUSO4W8UU>

### Further reading

- Devos K M, Brown JKM, Bennetzen J. 2002. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Res.* 12: 1075-1079.
- Gu YQ, Salse J, Coleman D, Dupen A, Crossman C, Lazo G, Huo N, Belcram H, Ravel C, charmet G, Charles M, Anderson O, Chalhoub B. 2006. *Genetics.* 174: 1493-150.
- Harberd N, Flavell R, Thompson R. 1987. Identification of a transposon- like insertion in a *Glu-1* allele of wheat. *Mol. Genet. Genomics.* 209:326-332.
- Jianxin M, Devos KM, Bennetzen JL. 2004. Analyses of LTR-Retrotransposon Structure Reveal Recent and Rapid genomic DNA Loss in Rice. *Genome Res.* 14:860-869.
- Moore G, Lucas H, Batty N, Flavell R. 1991. A family of retrotransposons and associated genomic variation in wheat. *Genomics* 10:461-468.
- Smith D, Flavell A. 1975. Characterisation of the wheat genome by renaturation kinetics. *Chromosoma* 50:223-242.



### 13. Methods/strategies for QTL identification

Quantitative Trait Locus (QTL) is a locus (DNA region) containing genetic factor(s) with additive effects on a specific polygenic quantitative trait i.e. a trait controlled by multiple genes and their interaction with the environment. Unlike monogenic traits, polygenic traits are not inherited according to classical Mendelian rules (discrete values), but their phenotypes vary along a continuous gradient of a bell curve. In crop plants, typical quantitative traits are yield, quality, flowering time or tolerance to abiotic stresses. By QTLs affecting only a portion of the variability of a given trait, the breeding process can be rather complicated.

Therefore, accurate identification of QTLs for agronomically useful traits is of paramount importance. To achieve progress in crop improvement for polygenic traits, mapping QTLs in the genome of crop species using molecular markers is indispensable. QTL mapping simply refers to finding an association between a genetic marker and a measurable phenotype. There are two main approaches to map QTLs:

- 1. Linkage mapping method** - conventional mapping method, depends upon genetic recombination during the construction of mapping populations and has relatively low mapping resolution. The main steps include 1) creation of suitable mapping population ( $F_2$ , DH, BC, RIL, NIL), 2) selection of molecular markers (e.g., SSR, SNP) and construction of a linkage map, 3) genotyping of the mapping population and 4) linkage analysis with appropriate software package.
- 2. Association or linkage disequilibrium (LD) mapping** - complementary to linkage mapping and takes advantage of historic recombination events accumulated over hundreds of generations, thus providing higher resolution and greater allele numbers. Owing to the dramatic reduction in costs of sequence technologies, association mapping has been conducted in many plants, including the major crops such as wheat, soybean, potato, maize, rice, sorghum, tomato, etc.

Steps in association mapping are 1) selection of association mapping panels from natural populations or germplasm collections with wide genetic diversity; 2) genotyping of the mapping population; 3) LD quantification by using molecular marker data and assessment of the population structure; 4) phenotyping of the panel and 5) correlation of phenotypic and genotypic data with an appropriate statistical approach to identify 'marker tags' positioned closely to the targeted trait.

There are two main types of association mapping: 1) candidate gene-based association mapping, which analyses polymorphisms of selected candidate genes and 2) genome-wide association mapping, which surveys genetic variation in the whole genome to associate allelic variation across the genome with various complex traits. For more details see Chapters on association mapping.

The key distinction between association and linkage mapping lies in whether recombination events occur in populations or families. Both of these methods share a consistent strategy for identifying molecular markers that are linked to QTL. It is also important to make a distinction between the terms linkage and LD, which are often confused. Linkage refers to the correlated inheritance of loci located on the same chromosome, whereas LD refers to the correlation between alleles in a population, but not necessarily on the same chromosome.

## D 7.1 Production of materials for improved genotyping training

### 13.1 Types of mapping populations

**Bi-parental** – progeny obtained from the cross between two selected parents that have contrasting phenotypes for the trait of interest ( $F_2$ , BC, DH, RIL, NIL). They are typically used for classical linkage mapping, are easy to obtain, yet by combining the genomes of only two parents, a relatively narrow genetic base (low mapping resolution and large genetic intervals) is included that cannot adequately represent wider allelic diversity.

**Multi-parent** – emerged as next-generation mapping resources and combine diverse genetic founder contributions with high levels of recombination obtained. These populations derive from structured inter-mating between more than two well-characterised parents and maximize allelic diversity. In this way, the flaws of the bi-parental populations can be overcome by allowing the derivation of individuals which feature diverse levels and patterns of recombination and new genotype and haplotype combinations. The two most used multi-parent populations are the: (i) Nested Association Mapping (NAM) population - derived by crossing a single inbred parent to a successive collection of diverse inbred lines; and the (ii) Multi-parent Advanced Generation InterCrosses (MAGIC) population - developed by inter-crossing of multiple (typically four, eight or sixteen) parental lines in a balanced funnel crossing scheme.

**Natural populations** - collection of a sample population including elite cultivars, landraces, wild relatives, and exotic accessions are typically used for association mapping. Analysis of such populations include phenotyping and estimating broad-sense heritability of traits of interest, determining the genotypes of the population entries, quantification of the LD extent of the selected population and testing the associations between genotypes and phenotypes using appropriate statistical approaches.

### 13.2. Statistical analysis in genetic mapping and QTL detection

Linkage mapping analysis comprehend single-marker analysis (SIM), simple interval mapping (1 QTL at the time), composite interval mapping (CIM; identifies more QTL at the time, it is more precise).

#### Online Tutorials

- QTL Mapping Part 1 by Kristin Bishop-von Wettberg; <https://www.youtube.com/watch?v=1JSw1gl3-RI>
- QTL Mapping Part 2 by Kristin Bishop-von Wettberg; <https://www.youtube.com/watch?v=lu0SjECydK8>
- Genome sequence and QTL identification for major agronomic traits of mung bean (*Vigna radiata*) by Suk-Ha Lee Seoul; <https://www.youtube.com/watch?v=d17D5V0tgMo>

#### Further reading

- Bernardo R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48: 1649-1664.
- Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM. 2009. Defining gene and QTL networks. *Curr. Opin. Plant Biol.* 12: 241-246.

## D 7.1 Production of materials for improved genotyping training

- Lincoln SE, Daly MJ, Lander ES. 1993. Mapping genes controlling quantitative traits using MAPMAKER/QTL version 1.1: a tutorial and reference manual. Whitehead Institute for Biometrical Research, Cambridge, Mass.
- McIntyre CL, Mathews KL, Rattey A, Chapman SC, Drenth J, Ghaderi M, Reynolds M, Shorter R. 2010. Molecular detection of genomic regions associated with grain yield and yield-related components in an elite bread wheat cross evaluated under irrigated and rainfed conditions. *Theor. Appl. Genet.* 120: 527-541.
- Yang J, Hu C, Hu H, Yu R, Xia Z, Ye X, Zhu J. 2008. QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics* 24: 721-723.
- Parent B, Shahinnia F, Maphosa L, Berger B, Rabie H, Chalmers K, Kovalchuk A, Langridge P, Fleury, D. 2015. Combining field performance with controlled environment plant imaging to identify the genetic control of growth and transpiration underlying yield response to water-deficit stress in wheat. *J. Exp. Bot.* 66: 5481-5492.
- Quraishi UM, Pont C, Ain Q, Flores R, Burlot L, Alaux M, Quesneville H, Salse J. 2017. Combined genomic and genetic data integration of major agronomical traits in bread wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 8:1843.
- Salvi S, Tuberosa R. 2015. The crop QTLome comes of age. *Current Opinion Biotech.* 32: 179-185.
- Tinker NA, Mather DE. 1995. MQTL: software for simplified composite interval mapping of QTL in multiple environments. *J. Agric. Genomics* 1: 1-5.
- Tondelli A, Francia E, Visioni A, Comadran J, Mastrangelo AM, Akar T, Al- Yassin A, Ceccarelli S, Grando S, Benbelkacem A, van Eeuwijk FA, Thomas WTB, Stanca AM, Romagosa I, Pecchioni N. 2014. QTLs for barley yield adaptation to Mediterranean environments in the 'Nure'×'Tremois' biparental population. *Euphytica* 197: 73-86.
- Utz HF, Melchinger AE. 1996. PLABQTL: a program for composite interval mapping of QTL. *J. Quant. Trait Loci* 2: 1-5.
- Utz HF. 2000. Introduction to PLABQTL. Institute of Plant Breeding, Seed Science and Population Genetics. University of Hohenheim, Germany. URL <https://www.uni-hohenheim.de/fileadmin/einrichtungen/plant-breeding/software/pgintro.pdf>.
- WGIN project website, provide genetic and molecular resources for wheat genetic stocks, mapping populations, molecular markers and marker technologies, trait identification and evaluation. <http://www.wgin.org.uk/about.php>.

### 14. Molecular-assisted selection - MAS

Principal goals of global plant breeding have typically aimed at improved yields, nutritional qualities, and other traits of commercial value. During the last thirty years, many studies have led to a rapid increase in knowledge of plant genome sequences and the physiological and molecular role of various genes, which have revolutionised molecular genetics and its own efficiency in genetic improvement programs (Nadeem et al. 2018).

Genetic mapping of major genes and quantitative traits loci (QTLs) for many important agricultural traits has increased the integration of biotechnology with the conventional breeding process. Therefore, DNA marker technology (see specific chapter 11) derived from research in molecular genetics and genomics, offers promising prospective for plant breeding. Owing to genetic linkage, molecular markers can be used to detect the presence of allelic variation in genes underling these traits. The use of molecular markers in plant breeding is called marker-assisted selection (MAS) and is a component of the new discipline of 'molecular breeding' (Collard and Mackill 2008). The volume of publications on the development and to a lesser extent application of markers for assisting plant breeding has increased dramatically during recent decades. The annual number of articles containing the term "marker assisted selection" surpassed 1000 in 2003 (Fig. 14.1.) (Xu and Crouch 2008) and probably an updated estimate would show further growth.

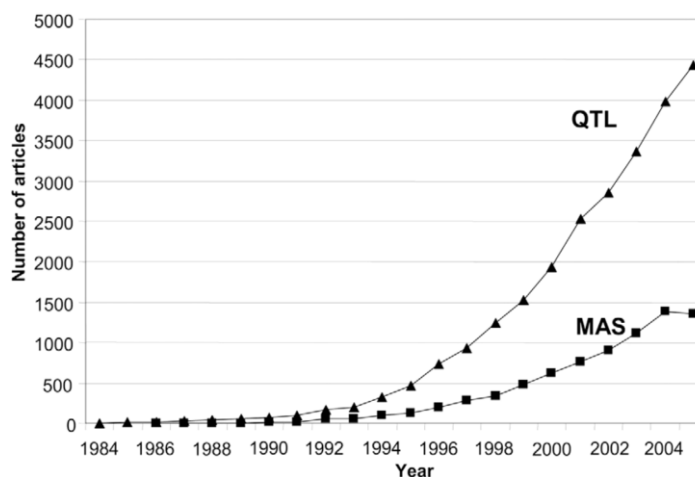


Fig. 14.1. The numbers of articles with the terms quantitative trait locus or quantitative trait loci (QTL) and marker-assisted selection (MAS) by years (1984–2005) from Google Scholar (4 Aug. 2007) (from Xu and Crouch, 2008).

Compared to classical breeding, MAS offers advantages in shortening the times of selection to obtain the desired phenotype based on the genotype identified with the markers (Collard et al. 2005), whose molecular profile is not influenced by environmental factors, but needs more specific and complex equipment and facilities. The first requirements for marker-assisted breeding (MAS) in plants should have: a) an appropriate marker system and reliable markers; b) quick DNA extraction and high throughput marker detection systems; c) knowledge of genetic linkage map and marker-trait association; d) quick and efficient data processing and management (Jiang 2013).

## D 7.1 Production of materials for improved genotyping training

Plant breeders mostly use MAS for the identification of suitable dominant or recessive alleles across a generation and for identification of the most favourable individuals across the segregating progeny (Francia et al. 2005) (Fig. 14.2.).

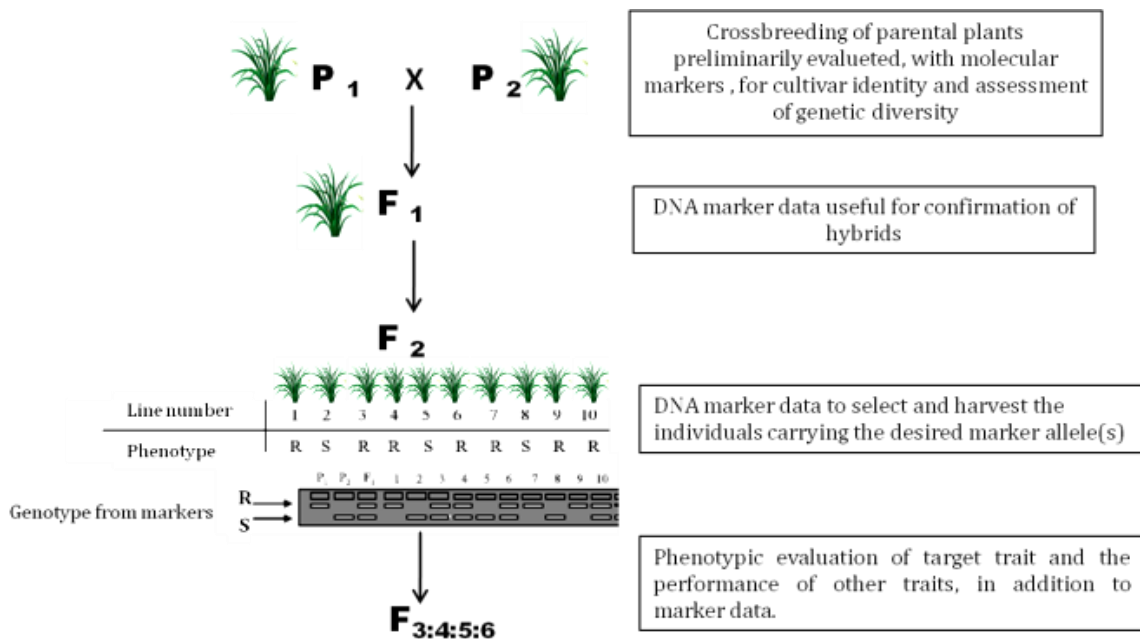


Fig. 14.2. An example of MAS approach. R= resistant, S= susceptible genotype.

An ideal DNA marker for MAS should be co-dominant (Fig. 14.3), evenly distributed throughout the genome, highly reproducible, at low cost, and having ability to detect higher level of polymorphism (Nadeem et al. 2018).

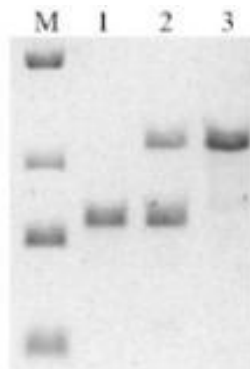


Fig. 14.3. Example of co-dominant marker. Lanes: M) DNA ladder; 1) dominant homozygote genotype; 2), heterozygote genotype; 3) recessive homozygote genotype.

For MAS it was suggested to limit the number of genes undergoing selection to three to four, if there are QTLs selected based on linked markers, and to five to six if there are known loci selected directly (Hospital 2003). The number of individuals in the screened population increases exponentially with the increase of target loci involved. The markers should be in the region of gene sequences or be close enough to the gene/QTL of interest (<5cM) to ensure that only a minor proportion of the selected individuals will be recombinants. From the point of both effectiveness and efficiency, for a single QTL it is usually suggested to use two markers (i.e. flanking

## D 7.1 Production of materials for improved genotyping training

markers) that are tightly linked to the gene/QTL of interest (Jiang 2013). Main schemes used for MAS are:

1. Selection based on mapped loci (QTL / genes): Marker-assisted backcrossing (MABC); Marker-assisted gene pyramiding (MAGP); Marker-assisted recurrent selection (MARS)
2. Selection with markers without map information of QTL / genes: Genomic selection (GS).

### 14.1. Marker-assisted backcrossing (MABC)

Special case of MAS in which breeding favourable alleles, to one or more loci, are transferred from a donor parent to an elite line through various cycles of backcross assisted selection with markers. Three general levels of marker-assisted backcrossing (MAB) have been described (Holland 2004) as:

1. 'Foreground selection': Selection for the allele most associated with the target gene (allele) provided by the parent donor (Fig. 14.4a)
2. 'Recombinant Selection': selection for the recurrent parent's alleles to the markers flanking the target gene, for reduced 'linkage drag' alongside the target gene (Fig. 15.4b)
3. 'Background selection': selection for alleles of the recurrent parent in the rest of the genome (optional) (Fig. 14.4c).

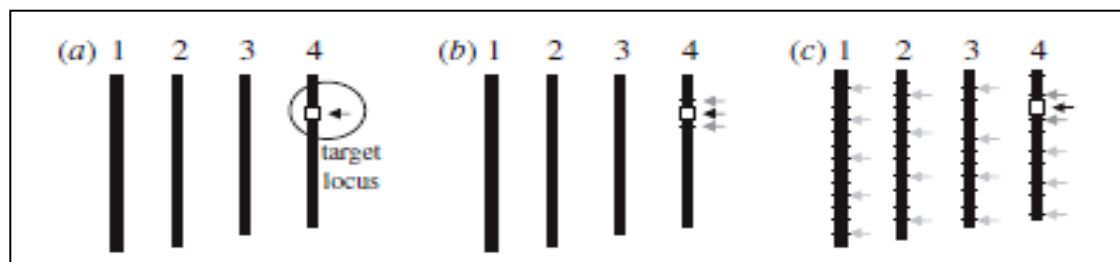


Fig. 14.4. Levels of selection during marker-assisted backcrossing. A hypothetical target locus is indicated on chromosome 4. (a) Foreground selection, (b) recombinant selection and (c) background selection (Collard and Mackill 2008).

Advantages of backcross assisted versus conventional MAS are: a) faster recovery of the recurrent parent's genome (often elite cultivar) and reduction of the "linkage drag" problem; b) like any type of MAS, it is not influenced by environmental factor; c) efficient selection of recessive alleles and individuals with event recombination near the target gene; d) best use of breeding program resources as the number of lines to keep per backcross cycle and the number of cycles to be performed are lower than in the traditional program.

### 14.2. Marker-assisted gene pyramiding

Marker-assisted gene pyramiding (MAGP) is one of the most important applications of DNA markers to plant breeding. This is a technique to enable transfer into a cultivar of QTLs/genes for single or multiple traits. This technique is mainly applied to increase the level of resistance to particular diseases and insects through the selection of two or more genes simultaneously (Nadeem et al. 2018).



## D 7.1 Production of materials for improved genotyping training

MAS has been successfully applied to pyramid many desired genes in various crops (Ye et al. 2008; Gupta et al. 2010; Li et al. 2010; Wang et al. 2012).

### 14.3. Marker-assisted recurrent selection (MARS)

As defined by Ribaut et al. (2010), MARS is a recurrent selection scheme using molecular markers for the identification and selection of multiple genomic regions involved in the expression of complex traits to assemble the best-performing genotype within a single or across related populations.

MARS is a scheme which allows performing genotypic selection and intercrossing in the same crop season for one cycle of selection (Jiang, 2013); it's specially involved with the improvement of the F<sub>2</sub> population that is achieved through one cycle of MAS (having phenotypic data with marker scores) followed by performing 2–3 cycles of marker-based selections (having marker scores only) (Nadeem et al. 2018). This approach has also been effectively used by Monsanto for improvement of several traits in corn, soybean, and sunflower (Eathington et al. 2007).

### 14.4. Genomic selection (GS)

Genomic selection is based on markers without significant testing and without identifying *a priori* a subset of markers associated with the trait (Bernardo and Yu 2007). GS is a form of MAS, where marker effects across the entire genome (explaining entire phenotypic variation) are simultaneously estimated and used to calculate genomic estimated breeding values (GEBV) (Meuwissen et al. 2001; Heffner et al. 2009; Nakaya and Isobe 2012). Selection is then based on this breeding value rather than on a subset of significant markers, that are generally used in MAS (Gupta et al. 2010). While the MAS is commonly used for only major QTL/genes, so that breeding benefits are limited by the proportion of the genotypic/phenotypic variance explained by markers associated with major QTLs, genomic selection (GS) could and should identify the whole of quantitative traits, that generally are controlled by a few major genes and many minor QTL/genes (Gupta et al. 2010).

#### Online Tutorials

- Description of Marker Assistant Selection by Dan Quiin for Shomu's Biology; <https://www.youtube.com/watch?v=OfDyT8E8dI>
- Genomic Selection by Mark Sorrells Cornell University for Borlaug Global Rust Initiative; [https://www.youtube.com/watch?v=s\\_FD7o5svZE](https://www.youtube.com/watch?v=s_FD7o5svZE)
- Genomic Selection - Theory and Tools by Aaron Lorenz University of Nebraska –Lincoln for iPlant Collaborative; <https://www.youtube.com/watch?v=PKc-IWKBD0c>
- Eurofin web pages with links to Molecular Breeding and Genomic Technology video; <https://www.eurofinsus.com/biodiagnostics/our-services/molecular-breeding/>
- DNA Extraction and Marker Assisted Selection by TomatoLab; <https://www.youtube.com/watch?v=yI8M9z4N4Y8>

#### Further reading

- Bernardo R, Yu J. 2007. Prospects for genome wide selection for quantitative traits in maize. *Crop Sci.* 47:1082-1090.

## D 7.1 Production of materials for improved genotyping training

- Collard BC, Jahufer MZ, Brouwer JB, Pang ECK. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* 142(1–2):169–196.
- Collard BCY, Mackill DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Phil. Trans. R. Soc. B* 363, 557–572. doi:10.1098/rstb.2007.2170.
- Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK. 2007. Molecular markers in a commercial breeding program. *Crop Sci.* 47:S154–S163.
- Francia E, Tacconi G, Crosatti C, Barabaschi D, Bulgarelli D, Dall’Aglio E, Valè G. 2005. Marker assisted selection in crop plants. *Plant Cell Tiss Org.* 82 (3):317–342.
- Gupta PK, Langridge P, Mir RR. 2010. Marker-assisted wheat breeding: present status and future possibilities. *Mol Breed.* 26(2):145–161.
- Heffner EL, Sorrells ME, Jannink JL. 2009. Genomic selection for crop improvement. *Crop Sci* 49:1–12.
- Holland JB. 2004. Implementation of molecular markers for quantitative traits in breeding programs-challenges and opportunities. In *Proc. 4th Int. Crop Sci. Congress., Brisbane, Australia, 26 September-1 October.*
- Hospital F. 2003. Marker-assisted breeding. In: H.J. Newbury (ed.), *Plant molecular breeding.* Blackwell Publishing and CRC Press, Oxford and Boca Raton, pp. 30-59.
- Jiang GL. 2013. Molecular Markers and Marker-Assisted Breeding in Plants. *Plant Breeding from Laboratories to Fields, Chapter 3.* <http://dx.doi.org/10.5772/5258>.
- Li X, Han Y, Teng W, Zhang S, Yu K, Poysa V, Anderson T, Ding J, Li W. 2010. Pyramided QTL underlying tolerance to *Phytophthora* root rot in mega-environment from soybean cultivar ‘Conrad’ and ‘Hefeng 25’. *Theor. Appl. Genet.* 121, 651-658.
- Meuwissen T, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.
- Nadeem MA, Nawaz MA, Shahid MQ, Dogan Y, Comertpay G, Yıldız M, Hatipoglu R, Ahmad F, Alsaleh A, Labhane N, Ozkan H, Chung G, Baloch FS. 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology & biotechnological equipment*, vol. 32, no. 2, 261–285. <https://doi.org/10.1080/13102818.2017.1400401>.
- Nakaya A, Isobe SN. 2012. Will genomic selection be a practical method for plant breeding? *Annals of Botany*, DOI:10.1093/aob/mcs109.
- Ribaut JM, de Vicente MC, Delannay X. 2010. Molecular breeding in developing countries: challenges and perspectives. *Current Opinion in Plant Biology*, 13, 213–218. DOI 10.1016/j.pbi.2009.12.011.
- Wang X, Jiang G-L, Green M, Scott RA, Hyten DL, Cregan PB. 2012. Quantitative trait locus analysis of saturated fatty acids in a population of recombinant inbred lines of soybean. *Mol. Breeding* 30(2), 1163-1179. DOI 10.1007/s11032-012-9704-0.
- Xu Y, Crouch JH. 2008. Marker-Assisted Selection in Plant Breeding: From Publications to Practice. *Crop Sci.* 48, 391–407. doi: 10.2135/cropsci2007.04.0191.
- Ye G, Smith KF. 2008. Marker-assisted gene pyramiding for inbred line development: basic principles and practical guidelines. *Int. J. Plant Breed.* 2(1), 1–10.

# 15. Marker-assisted selection (MAS) in bean

## 15.1. Introduction

Molecular breeding (MB) may be defined in a broad-sense as the use of genetic manipulation performed at DNA molecular levels to improve characters of interest in plants and animals, including genetic engineering or gene manipulation, molecular marker-assisted selection, genomic selection, etc. More often, however, molecular breeding implies molecular marker-assisted breeding (MAB) and is defined as the application of molecular biotechnologies, specifically molecular markers, in combination with linkage maps and genomics, to alter and improve plant or animal traits on the basis of genotypic assays. This term is used to describe several modern breeding strategies, including marker-assisted selection (MAS), marker-assisted backcrossing (MABC), marker-assisted recurrent selection (MARS), and genome-wide selection (GWS) or genomic selection (GS) (Guo-Liang Jiang 2013; Ribaut et al. 2010).

## 15.2. MAS in bean

The approach is more effective in selection for simple and single gene traits and has been applied for selection for resistance genes for various problematic common bean diseases of viral, bacterial and fungal origin (Blair et al. 2007; Tryphone et al. 2013). Selection for quantitative traits such as quantitative resistance or drought tolerance presents a greater challenge as it can involve multiple major and minor QTLs controlling the trait (Assefa et al. 2013). For decades DNA markers have been the most widely used molecular markers in crop improvement, due to their abundance and polymorphisms. Numerous molecular linkage maps have been established in different mapping populations of common beans on which genes and QTL for a wide variety of traits have been mapped. Economically important complex traits are usually controlled by multiple genes. Molecular markers are powerful tools to analyse the genetic control of complex traits such as drought tolerance that affects yield components, phenology, rooting pattern traits (Blair et al. 2010) as well as photosynthate acquisition, accumulation, and remobilisation to grain (Asfaw et al. 2012). Segregation mapping has been used to evaluate quantitative trait loci (QTL) that control multigenic traits such as biomass production and yield partitioning (Collins et al. 2008). Regarding stress response in addition to QTLs for abiotic stress (drought tolerance) QTL have been found for resistance to various pathogens (e.g. *Colletotrichum* sp.). Furthermore, several QTL have been identified for yield, morphological traits, traits associated with plant maturity and plant nutritional value. Micronutrient accumulation in plants is controlled by many genes as was demonstrated by mutational studies. Discovered were markers associated with Fe, Zn, P and phytic acid concentration in seed (Blair et al. 2009). Guzman Maldonado et al. (2003) identified QTL associated with seed mass, Ca, Fe, Zn and tannin content in the seed. QTL associated with tannin content were found and will be important for the genetic improvement of bio-fortified beans that have either higher tannins for their health benefits or lower tannins to increase iron bioavailability. Molecular markers for plant ability to uptake nutrients and form symbiosis were identified in several studies and associated with morphological traits associated with root architecture and formation (Ochoa et al. 2006, Beebe et al. 2006, Yan et al. 2004). These markers are potentially very useful in trait selection and breeding applications. Despite an

## D 7.1 Production of materials for improved genotyping training

increase in information about QTL, practical utilisation is just beginning. Recently, Pipan et al. (2018a) reported about marker-assisted evaluation and trait-specific selection of *Phaseolus vulgaris* L. Their study proposed the panel of 75 different types of trait-related DNA markers including KASP\_SNP markers (24), nSSRs (35), SCAR (5), EST-SSR (2), CAPS (1) and locus-specific PCR based markers (8), respectively covering a broad range of economically important traits (bean common mosaic virus, bean rust, anthracnose and angular leaf spot resistance; drought/heat tolerance; low phosphorus uptake and root morphology; high zinc and iron content; earliness and high yield).

Common bean is a crop featured in breeding programs worldwide, including Agricultural Institute of Slovenia (AIS). It is financially supported by Slovenian Ministry of agriculture, forestry and food. This is a small and well-orientated program with emphasis to achieve biotic and abiotic tolerance in combination with better agronomically important traits (Pipan et al. 2017a) and favourable nutritional value (Sinkovič et al. 2018a; 2018b) of newly developed varieties. The first step in bringing all traits together into a final set of the most promising hybrids, it is essential to identify parental lines with superior and desirable traits derived from European germplasm (mainly from the Andean gene pool). To discover potential parental lines, it is fundamental to perform the evaluations of common bean genetic resources. The most recognised characterizes in common bean are morphological seed characteristics, especially for dry beans. A recent study published by Sinkovič et al. (2019) presents the overall morphological seed evaluation on both, common bean, and runner bean (*Phaseolus coccineus* L.) collections, stored in Slovene genebank at AIS (2019). Additional evaluations of the parents are performed during the growing season according to different descriptors: International Union for the Protection of New Varieties of Plants (UPOV), Community Plant Variety Office (CPVO), International Board for Plant Genetic Resources (IBPGR), and Improvement of sustainable *Phaseolus* production in Europe for human consumption (Phaselieu). To develop new varieties with desired traits at AIS hand cross pollination of superior parental genotypes were performed as described by Ivančič (2002) and Pipan et al. (2017b). Positive selection occurred in the F<sub>2</sub> hybrids is combining morphologic/phenotypic and molecular characteristics. Morpho-agronomic traits are following the objectives of the national common bean breeding program for snap and dry beans; with marker assisted selection (MAS) using different types of DNA markers (Pipan et al. 2018b) associated with nine agronomical important traits and breeding attributes was introduced into the breeding program in 2018 (Pipan et al. 2019). A routine performance of MAS in combination with phenotypic selection enables and accelerates efficient identification of the elite hybrids to develop new varieties in the common bean breeding process. To obtain distinctive, uniform, and stable varieties six generations of self-pollination; positive selection from F<sub>3</sub> to F<sub>6</sub> is on the morphologic/phenotypic level considering characters described by CPVO protocol for *P. vulgaris* were performed (CPVO-TP/012/4).

## D 7.1 Production of materials for improved genotyping training

### Further reading

- Asfaw A, Blair MW, Struik PC. 2012. Multienvironment quantitative trait loci analysis for photosynthate acquisition, accumulation, and remobilization traits in common bean under drought stress. *G3: Genes| Genomes| Genetics*, 2(5), 579-595.
- Beebe SE, Rojas-Pierce M, Yan X, Blair MW, Pedraza F, Munoz F, Tohme J, Lynch JP. 2006. Quantitative trait loci for root architecture traits correlated with phosphorus acquisition in common bean. *Crop Science*, 46(1), 413-423.
- Blair MW, Knewton SJ, Astudillo C, Li CM, Fernandez AC, Grusak MA. 2010. Variation and inheritance of iron reductase activity in the roots of common bean (*Phaseolus vulgaris* L.) and association with seed iron accumulation QTL. *BMC Plant Biology*, 10(1), 1-12.
- Blair MW, Galeano CH, Tovar E, Torres MCM, Castrillón AV, Beebe SE, Rao IM. 2012. Development of a Mesoamerican intra-genepool genetic map for quantitative trait loci detection in a drought tolerant x susceptible common bean (*Phaseolus vulgaris* L.) cross. *Molecular Breeding*, 29(1), 71-88.
- Collins NC, Tardieu F, Tuberosa R. 2008. Quantitative trait loci and crop performance under abiotic stress: where do we stand? *Plant physiology*, 147(2), 469-486.
- Guo-Liang J. (2013). *Molecular Markers and Marker-Assisted Breeding in Plants, Plant Breeding from Laboratories to Fields*, Sven Bode Andersen, IntechOpen, DOI: 10.5772/52583.
- Ivančič A. 2002. Hibridizacija pomembnejših rastlinskih vrst. Maribor: Fakulteta za kmetijstvo, 776 p. Ochoa in sod., *Crop Sci.* (2006) 46(4): 16091621.
- Pipan B, Dolničar P, Sedlar A, Šuštar Vozlič J, Sinkovič L, Meglič V. 2018a. Application of molecular data to construct common bean core collection. <http://www.globalengage.co.uk/pgc/pr18.html>.
- Pipan B, Meglič V. 2019. Introduction of DNA markers in the process of common bean (*Phaseolus vulgaris* L.) breeding. Čeh B. (Ed.), et al. *New challenges in agronomy: proceedings of symposium*. Ljubljana: Slovensko Agronomsko Društvo. p. 96-102.
- Pipan B, Sedlar A, Šuštar Vozlič J, Meglič V. 2017a. Development of the common bean core collection referring to the Central and South Eastern European germplasm. *Book of abstracts*. <http://iclgg2017.hu/wp-content/uploads/2017/03/ICLGG2017-Book-of-abstracts.pdf>.
- Pipan B, Sinkovič L, Meglič V. 2018b. Marker-assisted evaluation and trait-specific selection of accessions from Central and Eastern European Common bean germplasm. Skočaj M. (Ed.). *Genetika 2018: book of abstracts*. <http://genetika2018.alfa-faktor.si>.
- Pipan B, Šuštar Vozlič J, Meglič V, Germšek B, Mavrič Pleško I, Dolničar P. 2017b. Cross pollination of common bean (*Phaseolus vulgaris* L.) in the breeding at the Agricultural Institute of Slovenia. Čeh, B. et al. (Eds) *New challenges in agronomy: proceedings of symposium*. Ljubljana: Slovensko Agronomsko Društvo. p. 33-39.
- Ribaut JM, de Vicente MC, Delannay X. 2010. Molecular breeding in developing countries: challenges and perspectives. *Current Opinion in Plant Biology* 13: 1-6.
- Sinkovič L, Pipan B, Kolmanič A, Nečemer M, Šibul F, Nemeš I, Tepić AN, Meglič V. 2018a. Identification and quantification of nutritinally important compounds in different legume species. Rozman V., Antunović Z. (Eds). *Book of abstracts*, ISSN 2459-5551). Osijek p. 157.
- Sinkovič L, Pipan B, Sinkovič E, Meglič V. 2019. Morphological seed characterization of common (*Phaseolus vulgaris* L.) and runner (*Phaseolus coccineus* L.) bean germplasm: a Slovenian gene bank example. *Biomed research international*.
- Sinkovič L, Pipan B, Tepić-Horecki A, Šibul F, Meglič V. 2018b. Nutritional composition of common bean (*Phaseolus vulgaris* L.) as green beans grains. Pojić M., Kokić B. (Eds.). *Foodtech Congress*, 23-25. 10. 2018b, Novi Sad, Serbia p. 150.
- Yan ER, Wang XH, Huang JJ. 2006. Shifts in plant nutrient use strategies under secondary forest succession. *Plant and Soil*, 289(1), 187-197.



### 16. Marker assisted selection for bunt resistance in wheat

Resistance breeding for bunt has recently witnessed a renaissance due to the increasing popularity of organic farming and the use of untreated seed. Assessment of bunt infection in field experiments is time and resource demanding, requires inoculated field experiments and allows only one selection cycle per year. Bunt resistance in wheat may range from quantitative to qualitative, depending on the resistance sources. On the one hand, for resistance to common bunt in wheat a gene-for-gene interaction between the pathogen and its host (Reed 1928; Bressman 1931), and 16 resistance race-specific genes denominated as *Bt1-Bt15* and *Btp* have been postulated so far (Goates 2012). On the other hand, there are also forms of race-non-specific resistance against bunt and for many cultivars, it is not unambiguously defined which resistance genes they harbour (Gaudet and Puchalski 1989).

The use of molecular markers for resistance breeding is an attractive option to directly target bunt resistance genes and select resistant genotypes during variety development following the crossing of parents. Marker assisted selection (MAS) is particularly attractive for a small number of loci which harbour resistance genes with strong effects. The main advantages of MAS for bunt resistance are:

1. Targeting known resistance loci becomes possible.
2. Combining resistance loci which are phenotypically not distinguishable is feasible.
3. Rapid back-crossing into adapted germplasm and therefore higher selection gain per unit time.
4. Marker analysis is cost-effective.
5. Some loci confer resistance to common bunt and dwarf bunt, allowing broad spectrum resistance selection.

For implementation of MAS useful markers that target genomic regions which harbour bunt resistance loci are required. Markers for MAS should ideally:

1. Be tightly linked or directly inside to the actual resistance gene(s) (= diagnostic).
2. Possess a high prediction ability to distinguish resistant from susceptible alleles.
3. Be easy, fast and reliably assessable through genotyping.

Mapping of bunt resistance loci and development of markers is therefore a precondition for MAS. Mapping for either QTL (quantitative trait locus/loci) or large effect *Bt*-genes has been on the research agenda since 1996. Mapping projects clearly mirror the advancement of genotyping methods. While RAPD markers were used in the early projects (1996), SCAR (sequence characterized amplified region) markers and SSR (simple sequence repeat) markers were implemented later on and more recently high throughput SNP (single nucleotide polymorphism) markers or GBS (genotyping by sequencing) markers were applied.

A range of QTLs with low to moderate effects have been discovered, but also several large effect QTLs which most likely tag major bunt resistance genes (Table 16.1). If QTLs mapping approaches are applied, QTLs which explain more than 30% of the phenotypic variation for a trait can be considered large effect QTLs and appear as such very attractive for MAS, such as the resistance genes *Bt9*, *Bt10*, *Q.DB.ui-7DS*, and a QTL on chromosome 1B for which linked markers are available (Table 16.1.).



## D 7.1 Production of materials for improved genotyping training

So far, no bunt resistance gene has been cloned and no diagnostic markers are available. Therefore, resistance selection can rely on linked markers only. MAS, using linked markers, is feasible but requires some additional precautions:

1. For starting MAS, the correct resistance donor must be used.
2. Polymorphic markers need to be identified and tested that allow clear separation between the allele of the resistance donor and the cultivars, which are used as recurrent parents in the breeding program.
3. Markers of choice are typically SSR markers that demand gel-separation post PCR or SNP markers. Modern SNP analysis can be performed gel-free, e.g. using the KASP® method (Semagn et al. 2014).
4. Ideally, markers flanking the resistance locus should be employed in MAS to minimise the probability of selecting false positive progeny.

## D 7.1 Production of materials for improved genotyping training

Table 16.1. Mapped resistance loci for bunt resistance in wheat.

Population	Type <sup>1)</sup>	Population Size	Gene/QTL	Source <sup>3)</sup>	Chr	R <sup>2</sup> <sup>4)</sup>	Markers	Marker Type <sup>5)</sup>	CB/DB <sup>6)</sup>	Comment	Reference
Attila x CDC Go	RIL	167	QCbt.dms-1B.2	CDC Go	1B	18.7	BS00086854_51; wspn_Ex_c5679_9976893	SNP	CB		Zou et al. (2017)
Attila x CDC Go	RIL	167	QCbt.dms-3A	CDC Go	3A	7.9	RAC875_c17451_896; RAC875_c57584_240	SNP	CB		Zou et al. (2017)
PI554099 x Cortez	DH	91	B19	PI554099	6DL	52.7	Xgpm4005-1; Xgpm7433; Xwmc773; Xgpm7303; Xgpm362; and more	SSR and SNP	CB		Steffan et al. (2017)
Carberry x AC Cadillac	DH	261	QCbt.spa-1B	Carberry	1B	5.0-18.0	wPe-743523	GBS	CB		Singh et al. (2016)
Carberry x AC Cadillac	DH	261	QCbt.spa-4B	Carberry	4B	7.6	wPe-744434-Xwmc617	GBS, SSR	CB		Singh et al. (2016)
Carberry x AC Cadillac	DH	261	QCbt.spa-4D	Carberry	4D	4.6	wPe-9747		CB	possibly	Singh et al. (2016)
Carberry x AC Cadillac	DH	261	QCbt.spa-6D	AC Cadillac	6D	20.2-46.2	wPe-1695	GBS	CB		Singh et al. (2016)
Carberry x AC Cadillac	DH	261	QCbt.spa-7D	Carberry	7D	3.5-6.4	Xwmc273	SSR	CB		Singh et al. (2016)
Rio Blanco x IDO444	RIL	159	Q.DB.ui-7DS	IDO444	7D	53.4	wPe-2565	SNP, STS	DB		Chen et al. (2016)
Rio Blanco x IDO444	RIL	159	Q.DB.ui-1A	IDO444	1A	9.9	Xgfa2129	SSR	DB		Chen et al. (2016)
Rio Blanco x IDO444	RIL	159	Q.DB.ui-2B	IDO444	2B	3.7	Xwmc317	SSR	DB		Chen et al. (2016)
McKenzie x BW711	DH	338	QCbt.spa-7B.1	BW711	7B	n.a.	Xgwm573	SSR	CB		Knox et al. (2013)
Trinitella x Plko	RIL	88	n.a. <sup>2)</sup>	Trinitella	1B	30.0	Xgwm273	SSR	CB		Dumalassova et al. (2012)
F94976GM28 x Uman	RIL	115	n.a.	F94976GM28	3B	n.a.	Xborc180, Xwmc623, Xwmc808, Xgwm285	SSR	CB	possibly Bt1.1	Ciura (2011)
8405-JC3C/Bizzard//2*8405-JC3C	DH	78	n.a.	Bizzard	1B	81.0	Xgwm374; Xgwm264; Xborc128	SSR	CB		Wang et al. (2009)
RL4452 x AC Domain	DH	185	Q.Cbt.crc-1B.1	AC Domain	1B	21.0	Xgwm374.1-Xwmc818	SSR	CB		Fofana et al. (2008)
RL4452 x AC Domain	DH	185	Q.Cbt.crc-1B.2	AC Domain	1B	8.0	GluB1-Xgwm274	SSR	CB		Fofana et al. (2008)
RL4452 x AC Domain	DH	185	Q.Cbt.crc-7	AC Domain	7A	3.0	Xgwm63-Xwmc633	SSR	CB		Fofana et al. (2008)
Glenlea x AC Taber	DH	62	Bt10	AC Taber	6D	n.a.	Xgwm469	SSR	CB	loose linkage	Menzies et al. (2006)
Laura x RL5407	RIL (BSA)	59	n.a.	RL5407	n.a.	n.a.	UBC548(590); UBC274(988)	RAPD	CB		He & Hughes (2003)
BW553 x Neeppawa	F2	199	Bt10	BE553	n.a.	n.a.	FSD, RSA	SCAR	CB		Laroche et al. (2000)
BW553 x Neeppawa	NILs; F2		Bt10	BW553	n.a.	n.a.	Primer196	RAPD	CB		Demeke et al. (1996)

1) RIL = recombinant inbred line; DH = doubled haploid; BSA = bulked segregant analysis; NIL = near isogenic line; F2 = filial generation 2

2) not available

3) donor line of the resistance improving allele

4) percentage of the phenotypic variance explained by marker or QTL

5) SNP = single nucleotide polymorphism; SSR = simple sequence repeat; GBS = genotyping by sequencing; STS = sequence tagged site; RAPD = random amplified polymorphic DNA; SCAR = sequence characterised amplified region

6) CB = common bunt; DB = dwarf bunt.

## D 7.1 Production of materials for improved genotyping training

### Further reading

- Bressman EN. 1931. Varietal resistance, physiologic specialisation, and inheritance studies in bunt of wheat. Oregon Agricultural Experiment Station Bulletin 281, 44 281:1-44.
- Chen JL, Guttieri MJ, Zhang JL, Hole D, Souza E, Goates B. 2016. A novel QTL associated with dwarf bunt resistance in Idaho 444 winter wheat. *Theor. Appl. Genet.* 129:2313-2322.
- Ciuca M. 2011. A Preliminary Report on the Identification of SSR Markers for Bunt (*Tilletia* sp.) Resistance in Wheat. *Czech J. Genet. Plant Breed.* 47:S142-S145.
- Demeke T, Laroche A, Gaudet DA. 1996. A DNA marker for the Bt-10 common bunt resistance gene in wheat. *Genome* 39:51-55. doi:10.1139/g96-007.
- Dumalasova V, Simmonds J, Bartos P, Snape J. 2012. Location of genes for common bunt resistance in the European winter wheat cv. Trintella. *Euphytica* 186:257-264.
- Fofana B, Humphreys DG, Cloutier S, McCartney CA, Somers DJ. 2008. Mapping quantitative trait loci controlling common bunt resistance in a doubled haploid population derived from the spring wheat cross RL4452 x AC Domain. *Mol Breed* 21:317-325 9.
- Gaudet DA, Puchalski BL. 1989. Races of common bunt (*Tilletia caries* and *Tilletia foetida*) of wheat in western Canada. *Can J Plant Pathol* 11:415-418.
- Goates BJ. 2012. Identification of New Pathogenic Races of Common Bunt and Dwarf Bunt Fungi, and Evaluation of Known Races Using an Expanded Set of Differential Wheat Lines. *Plant Dis* 96:361-369. doi:10.1094/pdis-04-11-0339.
- He C, Hughes GR. 2003. Development of RAPD markers associated with common bunt resistance to race T1 (*Tilletia tritici*) in spelt wheat. *Plant Breed* 122:375-377.
- Knox RE, Campbell HL, Depauw RM, Gaudet D, Puchalski B, Clarke FC. 2013. DNA markers for resistance to common bunt in 'McKenzie' wheat. *Can J Plant Pathol* 35:328-337. doi:10.1080/07060661.2013.763292.
- Laroche A, Demeke T, Gaudet DA, Puchalski B, Frick M, McKenzie R. 2000. Development of a PCR marker for rapid identification of the Bt-10 gene for common bunt resistance in wheat. *Genome* 43:217-223. doi:10.1139/gen-43-2-217.
- Menzies JG, Knox RE, Popovic Z, Procnier JD. 2006. Common bunt resistance gene Bt10 located on wheat chromosome 6D. *Can. J. Plant Sci.* 86:1409-1412. doi:10.4141/p06-106.
- Reed GM. 1928. Physiologic races of bunt of wheat. *Am. J. Bot.* 15:157-170. doi:10.2307/2435660.
- Semagn K, Babu R, Hearne S, Olsen M. 2014. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol. Breed.* 33:1-14. doi:10.1007/s11032-013-9917-x.
- Singh A, Knox RE, DePauw RM, Singh AK, Cuthbert RD, Kumar S, Campbell HL. 2016. Genetic mapping of common bunt resistance and plant height QTL in wheat. *Theor. Appl. Genet.* 129:243-256. doi:10.1007/s00122-015-2624-8.
- Steffan PM, Torp AM, Borgen A, Backes G, Rasmussen SK. 2017. Mapping of common bunt resistance gene Bt9 in wheat. *Theor. Appl. Genet.* 130:1031.
- Wang S, Knox RE, DePauw RM, Clarke FR, Clarke JM, Thomas JB. 2009. Markers to a common bunt resistance gene derived from 'Blizzard' wheat (*Triticum aestivum* L.) and mapped to chromosome arm 1BS. *Theor. Appl. Genet.* 119:541-553.
- Zou J, Semagn K, Chen H, Iqbal M, Asif M, N'Diaye A, Navabi A, Perez-Lara E, Pozniak C, Yang RC, Graf RJ, Randhawa H, Spaner D. 2017. Mapping of QTLs associated with resistance to common bunt, tan spot, leaf rust, and stripe rust in a spring wheat population. *Mol. Breed.* 37:14. doi:10.1007/s11032-017-0746-1.

### 17. The use of marker assisted selection (MAS) in organic potato breeding

Developing a new potato cultivar is a long-lasting and laborious process which takes about 12 years from initial crossing to the registration of a new cultivar. To develop one potato cultivar selection starts with about 80'000-100'000 seedlings. During the whole selection process, the potato breeding lines are assessed in terms of about 70 various traits (Zimnoch-Guzowska and Flis 2017). Due to specific requirements, the selection of potato cultivars dedicated for organic production system is more difficult than those for conventional. One of the main challenges for organic potato producers are pests and diseases. In conventional production system this problem could be (at least partly) resolved with the use of synthetic pesticides, which are prohibited in organic agriculture. Therefore, in the case of organic farming the best solution is breeding of potato cultivars resistant to the most important diseases.

The disease commonly thought to be the most important factor limiting yield of organic potatoes is late blight, caused by the oomycete *Phytophthora infestans*. In organic agriculture there is no highly effective control measure against late blight management and development of resistant varieties can play a key role in sustainable control of the disease for organic farmers (Pacilly et al. 2016). The use of genetic resistance to develop resistant cultivars has long been one of the primary goals of potato breeding. The two main types of potato resistance to *P. infestans* are R-gene-based (qualitative) and non-R-gene-based (quantitative) resistance. Both types of resistance were extensively used in resistance breeding programs, and both have pros and cons. Qualitative resistance is governed by major resistance genes (R-genes) that encode immunity through a hypersensitive reaction (HR), leading to cell death and rapid localisation of the pathogen thus preventing further colonisation of the host tissue. This resistance is considered non-durable because it exerts a strong selection pressure on *P. infestans* populations, which can adapt rapidly due to selection and multiplication of virulent and more aggressive races (Plich et al. 2016). However, in the last decades several potentially more durable R-genes providing resistance against broad spectrum of *P. infestans* races, have been identified. Quantitative resistance is caused by an array of different genes with minor effects, is race non-specific (RNS) and is considered durable. RNS resistance can lead to the reduction in late blight severity and deceleration of disease progression but does not provide complete protection against *P. infestans* infection. What's more, its polygenic nature makes it difficult to use in breeding programs (Plich et al. 2015; 2016).

Currently, pyramiding of various broad-spectrum R-genes in one cultivar is considered the best solution to breed potato cultivars with a high level of durable resistance against late blight (Zhu et al. 2012, Plich et al. 2017). Examples of such promising genes are *RB/Rpi-blb1* (Song et al. 2003), *Rpi-blb3* (Park et al. 2005), *Rpi-phu1* (Śliwka et al. 2006) which is identical to *Rpi-vnt1.1*, *Rpi-rzc1*, (Śliwka et al. 2012), *Rpi-Smira2/R8* (Rietman 2011) etc. The use of these R-genes in breeding of cultivars dedicated for organic production is highly sought after. However, phenotypic evaluation of thousands of potato individuals during the selection process is highly time- and work-consuming. In the case of pyramiding of R-genes the use of a set of various isolates with various (sometimes non-existing yet) combination of

## D 7.1 Production of materials for improved genotyping training

virulence/avirulence genes is required for proper phenotyping. Marker assisted selection (MAS) together with marker-assisted gene pyramiding (MAGP) allow breeders to select potato individuals with desired R-gene compositions without laborious phenotypical tests. In the literature there are many examples of the successful use of DNA markers for selection of potato clones possessing a particular R-gene or specific combination of R-genes against late blight (Śliwka et al. 2010, Plich et al. 2017). As evidence of usefulness of MAGP is the potato cultivar Gardena, which is included in the ECOBREED project Working Collection of potato cultivars. This cultivar was bred by HZZ Zamarte in co-operation with IHAR-PIB (Poland) and possesses two R-genes conferring resistance against late blight: *Rpi-phu1* and *Rpi-smira1*. Presence of both these genes was confirmed with the use of DNA markers and later by phenotypical tests.

Another example, of using DNA markers in organic potato breeding, concerns resistance to *potato virus Y* (PVY). PVY (genus *Potyvirus*, family *Potyviridae*) is the potato virus with the highest economic impact on seed potato production worldwide. It is transmitted in a non-persistent manner by various species of aphids. Use of insecticides, while effectively controls the vectors, have a low impact on the spread of PVY because of the short time needed to transmit the non-persistent virus. Aphids often transmit PVY prior to being killed by the insecticide (Dupuis et al. 2017). Therefore, breeding of resistant cultivars is considered the best strategy to manage PVY infections in both conventional and organic potatoes.

Potato resistance to PVY is provided by two types of genes denoted by *Ry* and *Ny*. The former controls extremely rapid virus localization (extreme resistance - ER) and the later hypersensitive response (HR). These genes differ in specificity to various strains of PVY. The most suitable for organic (and conventional) potato breeding is the *Ry-fsto* gene (also designated as *Ry-sto*) derived from *S. stoloniferum* located on potato chromosome XII (Flis et al. 2005; Song et al. 2005). The *Ry-fsto* gene provides durable resistance against all known strains of PVY and as it is not linked with male-sterility, is more convenient to use in breeding programs. There are also DNA markers suitable for selection of potato clones/cultivars possessing this gene (Witek et al. 2006, Milczarek et al. 2017).

Since monogenic resistance is relatively easy to use in breeding programs, R-gene based resistance against many diseases is widely used in potato breeding. What's more, in contrary to most polygenic traits, the selection of individuals containing R-genes could be successfully supported using molecular markers. Although the use of molecular markers in organic breeding arouses many controversies, DNA markers provide a great opportunity to enhance the effectiveness of selection of new potato cultivars suitable for organic agriculture and should be widely used in organic potato breeding.

### Further reading

- Dupuis B, Cadby J, Goy G, Tallant M, Derron J, Schwaerzel R, Steinger T. 2017. Control of potato virus Y (PVY) in seed potatoes by oil spraying, straw mulching and intercropping. *Plant Pathology* Vol 66, Issue 6. <https://doi.org/10.1111/ppa.12698>.



## D 7.1 Production of materials for improved genotyping training

- Flis B, Hennig J, Strzelczyk-Żyta D, Gebhardt C, Marczewski W. 2005. The *Ry-fsto* gene from *Solanum stoloniferum* for extreme resistance to Potato virus Y maps to potato chromosome XII and is diagnosed by PCR marker GP122718 in PVY resistant potato cultivars. *Molecular Breeding* 15: 95–101.
- Milczarek D, Plich J, Tatarowska B, Flis B. 2017. Early selection of potato clones with resistance genes: the relationship between combined resistance and agronomical characteristics. *Breeding Science* 67: 416-420. doi:10.1270/jsbbs.17035.
- Pacilly FCA, Groot JCJ, Hofstede GJ, Schaap BF, Lammerts van Bueren ET. 2016. Analysing potato late blight control as a social-ecological system using fuzzy cognitive mapping *Agron. Sustain. Dev.* 36: 35. <https://doi.org/10.1007/s13593-016-0370-1>.
- Park TH, Gros J, Sikkema A, Vleeshouwers VG, Muskens M, Allefs S, Jacobsen E, Visser RG, van der Vossen EA. 2005. The late blight resistance locus *Rpi-blb3* from *Solanum bulbocastanum* belongs to a major late blight R gene cluster on chromosome 4 of potato. *Mol Plant Microbe Interact.* 18(7): 722-729.
- Plich J, Tatarowska B, Lebecka R, Śliwka J, Zimnoch-Guzowska E, Flis B. 2015. R2-like gene contributes to resistance to *Phytophthora infestans* in Polish potato cultivar Bzura. *Am. J. Potato Res.* 92:350–358. <https://doi.org/10.1007/s12230-015-9437-9>.
- Plich J, Tatarowska B, Milczarek D, Flis B. 2017. Pyramiding of resistance genes against *Phytophthora infestans* in potato. *Biuletyn Inst. Hodowli i Aklimatyzacji Roślin* 281: 69-76.
- Plich J, Tatarowska B, Milczarek D, Zimnoch-Guzowska E, Flis B. 2016. Relationships between race-specific and race-non-specific resistance to potato late blight and length of potato vegetation period in various sources of resistance. *Field Crops Research* Vol. 196 pp. 311-324. <https://doi.org/10.1016/j.fcr.2016.04.033>.
- Rietman H. 2011. Putting the *Phytophthora infestans* Genome Sequence at Work; Identification of Many New R and Avr Genes in *Solanum* PhD Thesis. Wageningen University.
- Song J, Bradeen JM, Naess SK, Raasch JA, Wielgus SM, Haberlach JT, Liu J, Kuang H, Austin-Phillips S, Buell CR, Helgeson JP, Jiang J. 2003. Gene *RB* cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. *Proceedings of the National Academy of Science of the United States of America* 100: 9128–9133.
- Song YS, Hepting L, Schweizer G, Hartl L, Wenzel G, Schwarzfischer A. 2005. Mapping of extreme resistance to PVY (*Rysto*) on chromosome XII using anther-culture-derived primary dihaploid potato lines. *Theor. Appl. Genet.* 111:879–887.
- Śliwka J, Jakuczun H, Chmielarz M, Hara-Skrzypiec A, Tomczyńska I, Kilian A, Zimnoch-Guzowska E. 2012. Late blight resistance gene from *Solanum ruiz-ceballosii* is located on potato chromosome X and is linked to violet flower color. *BCM Genetics* 13: 11.
- Śliwka J, Jakuczun H, Kamiński P, Zimnoch-Guzowska E. 2010. Marker-assisted selection of diploid and tetraploid potatoes carrying *Rpi-phu1*, a major gene for resistance to *Phytophthora infestans*. *Journal of Applied Genetics.* 51(2): 133-140.
- Śliwka J, Jakuczun H, Lebecka R, Marczewski W, Gebhardt C, Zimnoch-Guzowska E. 2006. The novel, major locus *Rpi-phu1* for late blight resistance maps to potato chromosome IX and is not correlated with long vegetation period. *Theoretical and Applied Genetics* 113: 685–695.
- Witek K, Strzelczyk-Żyta D, Hennig J, Marczewski W. 2006. A multiplex PCR approach to simultaneously genotype potato towards the resistance alleles *Ry-fsto* and *Ns*. *Mol Breed.*18:273–275. doi: 10.1007/s11032-006-9021-6.
- Zhu S. 2014. R gene stacking by trans- and cisgenesis to achieve durable late blight resistance in potato. PhD thesis. Wageningen University.
- Zimnoch-Guzowska E, Flis B. 2017. Rola hodowli w integrowanej ochronie i produkcji ziemniaka. Rozdział w monografii “Metodyka integrowanej ochrony dla doradców” - wydawca IOR PIB; ISBN: 978-83-64655-32-6: 68-77.



### 18. Association genetics

Association genetics/mapping is one of the recently developed approaches in plant genetics and breeding, based on analysis of the linkage disequilibrium (LD) structure across the genome, as well as on the association of a phenotype of interest (i.e., a single or multiple traits) with specific loci. Association genetics ultimately aims at identifying “marker-trait” associations to be used in marker-assisted selection (MAS) strategies. This approach offers a possibility of high-resolution mapping of multiple genomic regions involved in genic control of the desired trait, without the need to construct mapping populations. Therefore, it is of particular interest for studies of multigenic traits.

Initially, association studies were introduced in human genetics while in the last 10-15 years the approach finds its application also in crop genetics; in particular in QTL studies regarding complex traits such as yield. This was due to an increase in the number of available molecular markers, lowering of sequencing costs as well as increasing availability of newly annotated sequences of several species. Most of the association studies in crop species regarded cereals (barley, maize, rice, pearl millet, sorghum, and wheat) and polygenic traits such as yield, yield components, disease resistance, processing quality, etc. Dicotyledonous crops have been so far less included in association studies. Among these crops, members of *Fabaceae* and *Solanaceae* families were the most relevant ones. However, soybean had few studies concerning its economic importance.

Association mapping (AM) uses germplasm collections ideally composed of well-characterised population structure, with minimal genetic structure, and that is made up of genotypes varying enough for single or multiple traits of interest. As the association panels of genotypes should guarantee a good coverage of alleles for a specific trait(s), AM then can be used to analyse the high amount of recombination possibly occurring in the various individuals from the population under study and can deliver a high-resolution power of LD mapping. By using large and diverse natural populations some downsides of classic QTL mapping such as limited variability, recombination frequency and sample size can be overcome.

#### 18.1. LD principles

LD refers to a statistically significant possibility of simultaneous inheritance of alleles at different loci and can involve loci situated on the same, but also on different chromosomes. A non-random association of alleles at these different loci is said to be the LD. If two genes are in LD, it means that particular alleles of each gene are inherited together more often than would be expected by chance. Apart from the physical distance between the involved loci, their LD can be a result of some kind of functional interaction of specific allele combinations.

LD is measured by the frequency of allele co-occurrence and is expressed using gamete frequencies. As LD usually describes smaller genomic regions concerning classical QTL mapping, a considerably greater number of molecular markers is needed for sufficient coverage and detection of “marker-trait” associations. Fortunately, the fast development of high throughput sequencing techniques in the last decade allows for the development of an ample array of molecular markers which gives the possibility to directly study statistical associations, i.e., LD between markers and complex traits.

## D 7.1 Production of materials for improved genotyping training

### 18.2. Steps in AM

#### 18.2.1. Selection of association mapping panel/population

Genotypes included should have representative genetic variability that covers different phenotypic characteristics, geographic origin, evolutionary paths, and in some cases to belong to different, yet related species. The population structure must not be high, i.e. it should not contain diverse individuals which do not markedly cluster.

#### 18.2.2. Genotyping of the mapping population

Two approaches can be employed:

- i. Candidate Gene (CG) analysis is based on previous knowledge and studies, a candidate gene and markers are chosen for polymorphism analysis within the association panel. This analysis is used when the function of a gene is known, and the gene is involved in biological or biochemical pathways causing trait variability. A haplotype sequencing approach is usually applied around the candidate gene to compensate for possible low variability of the gene. Moreover, the availability of ever-increasing annotated databases helps in discovering useful genetic markers.
- ii. Genome-Wide Association (GWA) analysis refers to association testing with markers covering most of the genome and the trait of interest. A range of different molecular marker types can be employed for GWA (e.g., SSR, DArT, CAPS, etc.), yet analysis with SNPs seem to be needed at some point of the genotyping as they are abundant, accurate and are becoming easily affordable since the costs of high-throughput sequencing have dramatically reduced nowadays. The number of molecular markers required for GWA is significantly higher with respect to the CG analysis, to be able to characterise all chromosomes. In addition, the shorter the genomic sequence extended by a certain LD, the higher the number of molecular markers needed. Typically, GWA studies include hundreds of thousands of genetic markers for adequate coverage. For AM in plant species, GWA studies are favoured by an increasing number of genomes sequenced which, together with the next-generation sequencing technologies, allow for the identification of a large number of genetic markers.

#### 18.2.3. Analysis of population structure

Analysis of population structure is carried out based on genotyping data to characterize LD and allele unequal distribution across the genome, i.e., to identify the association of different allele co-inheritances not dependent on physical linkage. A drawback of this analysis is the occurrence of false-positive “marker-trait” associations, especially in highly structured populations. This can be adjusted by using a large number of independent genetic markers across the genome.

#### 18.2.4. Phenotyping of mapping population for the trait of interest

The phenotyping of mapping population is an equally important step as is the genotyping. High-precision genotyping is not informative enough if the phenotyping does not produce good quality data. In recent years, phenotyping platforms have been constantly improved to produce significant quantities of data yet their quality remains a critical concern. Therefore, sufficient replicas and/or environments should be included in the experimental design to characterise possible genotype ×

## D 7.1 Production of materials for improved genotyping training

environment interactions, an aspect that has been overlooked in many association studies so far. It is likely, that more investment will be directed to the phenotyping process in the future, having in mind a continuous decrease in genotyping costs.

### 18.2.5. Identifying associations between genotypes and phenotypes

Using an appropriate statistical method, association of a marker with segregation of desired phenotype is calculated. Several approaches have been in use: haplotype analysis, Hardy–Weinberg equilibrium, population structure determination, missing genotype data etc. For population structure analyses, different information is included, yet the most important is the number and location of null loci. Most software uses a frequency-based approach for population structure. For “marker-trait” associations to be significant, the *LOD* score (i.e., statistical estimate of how closely two loci are linked) needs to be higher than a threshold (normally higher than 3), and the effects of one of the alleles on the trait of interest must be large. Although methods for GWAS are standardised, they still do not give all the answers for multi-environment QTL and QTL x environment analyses.

#### Software commonly used in associate genetics:

- TASSEL – mostly used, evaluate traits associations, evolutionary patterns, and LD; applies general linear model and mixed linear model and can manage a wide range of indels (insertion and deletions), which other softwares do not; <https://www.maizegenetics.net/tassel>
- GAPIT - R package that implements advanced statistical methods including the compressed mixed linear model (CMLM) and CMLM-based genomic prediction and selection; it can handle large datasets over 10000 individuals and 1 million single-nucleotide polymorphisms with minimal computational time; <http://www.zzlab.net/GAPIT/>
- STRAT - enables valid case-control studies even in the presence of population structure; <https://web.stanford.edu/group/pritchardlab/software/STRAT.html>
- Bimbam – for Bayesian IMputation-Based Association Mapping, it can handle both large association studies (e.g., genome scans) and smaller studies of candidate genes/regions; <http://stephenslab.uchicago.edu/software.html#bimbam>
- GOLD - provides a graphical summary of LD in human genetic data, suitable for the analysis of dense genetic maps; <http://csg.sph.umich.edu/abecasis/GOLD/>
- Powermarker - designed especially for SSR/SNP data analysis but handles a variety of marker data; <https://brcwebportal.cos.ncsu.edu/powermarker/>
- EMMA (Efficient Mixed-Model Association) - performs association mapping and simultaneously corrects for relatedness and population structure which speeds up the computation process; <http://mouse.cs.ucla.edu/emma/>
- ASREML - specially designed for general linear mixed models using Residual Maximum Likelihood (REML) to estimate the variance components; <https://www.vsni.co.uk/software/asrem/>
- JMP Genomics – applies techniques from simple case-control association to complex mixed models, and easy controls for population structure, and identifies and understand rare genomic variants; [https://www.jmp.com/en\\_us/software/genomics-data-analysis-software.html](https://www.jmp.com/en_us/software/genomics-data-analysis-software.html)

## D 7.1 Production of materials for improved genotyping training

### Online Tutorials

- Genetic Association Studies - Tales from the Genome by Udacity; <https://www.youtube.com/watch?v=HIF73Hu1Vmw>
- Linkage Disequilibrium by Physeo; <https://www.youtube.com/watch?v=DvrAuMyu4wU>
- Genome-wide association studies revisited by Rosie Redfield University of British Columbia; <https://www.youtube.com/watch?v=5sgPkRXR6pE>
- Genome-wide association studies revisited by Rosie Redfield University of British Columbia; <https://www.youtube.com/watch?v=-WrmAvL711Y>

### Further reading

- Álvarez MF, Mosquera T, Blair MW. 2015. The use of association genetics approaches in plant breeding. *Plant Breed. Rev.* 38: 17-68.
- Biscarini F, Cozzi P, Casella L, Riccardi P, Vattari A, Orasen G, Greco R. 2016. Genome-wide association study for traits related to plant and grain morphology, and root architecture in temperate rice accessions. *PLoS ONE* 11: e0155425.
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635.
- Challa S, Neelapu NRR. 2018. Genome-wide association studies (GWAS) for abiotic stress tolerance in plants. In: *Biochemical, physiological and molecular avenues for combating abiotic stress tolerance in plants* (Eds) Wani SH, Academic Press, pp 135-150.
- Chang HX, Brown PJ, Lipka AE, Domier LL, Hartman GL. 2016. Genome-wide association and genomic prediction identifies associated loci and predicts the sensitivity of Tobacco ringspot virus in soybean plant introductions. *BMC genom.*,17: 153.
- Fang C, Luo J. 2019. Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J.* 97: 91-100.
- Ingvarsson PK, Street, NR. 2011. Association genetics of complex traits in plants. *New Phytol.* 189: 909–922.
- Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinf.*, 28(18), 2397-2399.
- Ma F, Xu Y, Ma Z, Li L, An D. 2018. Genome-wide association and validation of key loci for yield-related traits in wheat founder parent Xiaoyan 6. *Mol. Breed.* 38: 91.
- Mérida-García R, Liu G, He S, Gonzalez-Dugo V, Dorado G, Gálvez S, Hernandez P. 2019. Genetic dissection of agronomic and quality traits based on association mapping and genomic selection approaches in durum wheat grown in Southern Spain. *PloS one* 14, e0211718.
- Molero G, Joynson R, Pinera-Chavez FJ, Gardiner LJ, Rivera-Amado C, Hall A, Reynolds MP. 2019. Elucidating the genetic basis of biomass accumulation and radiation use efficiency in spring wheat and its role in yield potential. *Plant Biotechnol. J.* 17(7), 1276-1288. <https://doi.org/10.1111/pbi.13052>.
- Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *The American J. of Human Genetics* 65, 220-228.
- Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate genes and quantitative traits. *PLoS Genetics* 3: e114.
- Sharma SK, MacKenzie K, McLean K, Dale F, Daniels S, Bryan GJ. 2018. Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3* 8: 3185-3202.

# 19. Sequences

## 19.1. What is sequencing

In molecular biology and biochemistry, sequencing is the process of primary structure determination of complex polymeric biological molecules such as DNA, RNA and proteins that are composed of simpler monomeric units, such as nucleotides, ribonucleotides and amino acids, respectively. The resulting sequence is represented by letters and summarises the structure of the molecule. The DNA molecule consists of paired and linked nucleotides that contain a phosphate, deoxyribose sugar and one of the four nitrogenous nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T).

### 19.1.1. DNA sequencing

The sequence of DNA encodes the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the key importance DNA has to living things, knowledge of DNA sequences is useful in practically any area of biological research. For example, in medicine it can be used to identify, diagnose, and potentially develop treatments for genetic diseases. Similarly, research into pathogens may lead to treatments for contagious diseases. Biotechnology is a burgeoning discipline, with the potential for many useful products and services.

Chain termination sequencing (Sanger sequencing) developed by Frederick Sanger was the first widely used DNA sequencing method. The technique uses sequence-specific termination of a DNA synthesis reaction using modified nucleotide substrates.

Extension is initiated at a specific site on the template DNA by using a short oligonucleotide 'primer' complementary to the template at that region. The oligonucleotide primer is extended using a DNA polymerase, an enzyme that replicates DNA. Included with the primer and DNA polymerase are the four deoxynucleotide bases (DNA building blocks), along with a low concentration of a chain terminating nucleotide (most commonly a di-deoxynucleotide). Limited incorporation of the chain terminating nucleotide by the DNA polymerase results in a series of related DNA fragments that are terminated only at positions where that particular nucleotide is used. The fragments are then size separated by electrophoresis on a slab polyacrylamide gel, or more commonly now, in a narrow glass tube (capillary) filled with a viscous polymer.

An alternative to the labelling of the primer is to label the terminators instead, commonly called 'dye terminator sequencing'. A major advantage of this approach is the complete sequencing set can be performed in a single reaction, rather than the four needed with the labelled-primer approach. This is accomplished by labelling each of the dideoxynucleotide chain-terminators with a separate fluorescent dye, which fluoresces at a different wavelength. This method is easier and quicker than the dye primer approach but, may produce more uneven data peaks (different heights), due to a template dependent difference in the incorporation of the large dye chain-terminators. This problem has been significantly reduced with the introduction of new enzymes and dyes that minimise incorporation variability. This method is now used for the vast majority of sequencing reactions as it is both simpler and cheaper.



## D 7.1 Production of materials for improved genotyping training

The major reason for this is that the primers do not have to be separately labelled (which can be a significant expense for a single-use custom primer), although this is less of a concern with frequently used 'universal' primers. This is changing rapidly due to the increasing cost-effectiveness of second- and third-generation systems from Illumina, 454, ABI, Helicos, and Dover.

Next-generation sequencing (NGS), or high-throughput sequencing, is the term used to describe several different modern sequencing technologies including Illumina (Solexa) sequencing, Roche 454 sequencing, Ion torrent: Proton/PGM sequencing and SOLiD sequencing. Life Technologies SOLiD system, called sequencing by ligation, where a probe bound to a fluorophore hybridizes to a DNA fragment and is ligated to an adjacent oligonucleotide for detection. According to the emission spectrum of the labelled probe and its anti-complementarity with the base, the sequence can be revealed (Goodwin et al. 2016).

These recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology. These technologies apply to genome sequencing, genome resequencing, transcriptome profiling (RNA-Seq), DNA-protein interactions (ChIP-sequencing), and epigenome characterization. Resequencing is necessary, because the genome of a single individual of a species will not indicate all of the genome variations among other individuals of the same species.

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that complement the sequencing process, producing thousands or millions of sequences concurrently. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods. In ultra-high-throughput sequencing as many as 500'000 sequencing-by-synthesis operations may be run in parallel. Such technologies led to the ability to sequence an entire human genome in as little as one day. As of 2019, corporate leaders in the development of high-throughput sequencing products included Illumina, Qiagen and ThermoFisher Scientific.

However, also these NGS have some drawbacks. For example, even if the output amount of data is high, the fragments are still quite short (Illumina sequence fragments of 250-300 bp) and this becomes a problem when repetitive regions have to be sequenced, haplotype alleles or mRNAs isoforms to be identified. Plus, they keep the PCR bias from previous generation techniques.

Third generation sequencing techniques (TGS) started to be developed around 2011. They are mostly single molecule real time sequencing using different technologies. Pacific Biosciences produces a platform for single molecule real time sequencing that uses the so-called Single molecules real time (SMRT) bells to be sequenced in a chip-platform with wells called zero-mode waveguide (ZMW), a structure small enough to observe a single nucleotide being incorporated and the released fluorescence, thanks to the polymerase attached to the bottom of the well. The Oxford Nanopore technology uses Nanopore which the DNA molecule passes through while being sequenced. One of the latest technologies is MinION, a portable device with 512 Nanopores where the DNA passes through and changes the



## D 7.1 Production of materials for improved genotyping training

electronic current according to the nucleotide (Lu et al. 2016). These techniques are specified in long reads, even up to 80Kb, which would overcome all above mentioned problems for NGS, easily solving the repeated regions, different alleles and all the PCR bias.

### 19.2. Comparison of high-throughput sequencing methods

#### 19.2.1. RNA sequencing

The structure of RNA nucleotides is like DNA nucleotides, with the difference that the sugar is ribose, and one of the bases, thymine is replaced by a demethylated form uracil (U). Sequencing of messenger RNA (mRNA) provides a snapshot of actively expressed genes in a tissue at the time of sampling. Because RNA has low stability in the cell and is prone to nuclease degradation, it is usually transcribed to cDNA and sequenced as described for DNA sequencing. The bulk of RNA expressed in cells is ribosomal RNAs or small RNAs, detrimental for cellular translation, but often not the focus of a study.

Derived from the exons these mRNAs are to be later translated to proteins that support particular cellular functions. The expression profile therefore indicates cellular activity, particularly desired in the studies of diseases, cellular behaviour, responses to reagents or stimuli. Eukaryotic RNA molecules are not necessarily co-linear with their DNA template, as introns are excised. This gives a certain complexity to map the read sequences back to the genome and thereby identify their origin. For more information on the capabilities of next-generation sequencing applied to whole transcriptomes see: RNA-Seq and MicroRNA Sequencing.

#### 19.2.2. Protein sequencing

Protein sequencing is the process of determination of partial or complete amino acid sequence of a protein or peptide that enables protein identification or characterisation of its post-translational modifications. Partial sequence information is typically enough to identify the protein in reference databases. The most widely used method for direct protein sequencing and sequence identification is mass spectrometry. Another major method is Edman degradation using a protein sequencer that is today used mostly characterisation of the N-terminus of a protein. Among other methods are peptide mass fingerprinting and protease digests. A protein sequence can also be inferred from the sequence of the gene encoding the protein.

### 19.3. Working with sequences

FASTA format is a commonly used text-based format for nucleotide or amino acid sequences. A sequence in FASTA format begins with a single-line description (begins with a greater-than symbol ">") symbol and is followed by lines of sequence data. Sequences from the NCBI database are labelled with unique identifiers (SeqID) in the header line.

FASTQ is an extended FASTA format for storing both a biological sequence and its corresponding quality scores, which are both represented by a single ASCII character. Different filename extensions can be used for a text file containing FASTA formatted sequences.

Sequences in FASTA and FASTQ formats are simple and easy to manipulate (concentrate into a multisequence file or parse to single sequences) using text-

## D 7.1 Production of materials for improved genotyping training

processing tools and scripting languages like the R programming language, Python, Ruby, and Perl. Format converters for conversion between Sanger, Solexa and Illumina 1.3+ formats include Biopython, EMBOSS, BioPerl, BioRuby, BioJava, MAQ and fastx\_toolkit.

### 19.3.1. Basic Local Alignment Search Tool (BLAST)

BLAST is a widely used and useful tool for performing a sequence similarity search. It can compare nucleotide or protein sequence queries for regions of similarity with thousands of sequences uploaded in databases and calculate the statistical significance for the similarity. In this way information on new and unknown DNA and protein sequences can be gained by identifying the most similar sequences for which functional information is already known. BLAST uses heuristics and produces results quickly. Confidence of the alignment can also be evaluated by the calculated “expect value”, an estimation of number of matches occurring at a given score by chance.

BLAST algorithm performs local alignments which can identify sequences with similar functional domains and motifs or only a shorter part of the sequence being similar. With local alignment, mRNA sequences can be aligned to the coding region of the DNA sequence, which is important for genome assembly. BLAST enables sequence identification, evaluation of functional and evolutionary relationships between sequences, and identification of members of gene families. Some of the BLAST applications are:

- BLASTN for aligning distant nucleotide sequences
- BLASTP for comparing protein sequences
- BLASTX for translating a nucleotide query and comparing it against a protein sequence database
- TBLASTN for translating a protein query and comparing it against a nucleotide sequence database
- megaBLAST for finding very similar nucleotide sequences in the same or closely related species

### 19.3.2. Online resources for comparative, evolutionary and functional genomics

PLAZA (<https://bioinformatics.psb.ugent.be/plaza>) is a plant-oriented online resource for comparative, evolutionary and functional genomics. The PLAZA platform consists of multiple independent instances focusing on different plant clades, while also providing access to a consistent set of reference species. Each PLAZA instance contains structural and functional gene annotations, gene family data and phylogenetic trees and detailed gene collinearity information.

### 19.4. Applications of sequencing technology in genotyping include:

- genotyping by sequencing,
- utilising whole genome sequence for identifying new molecular markers,
- confirmation of the sequence of amplified amplicons (verification of genes / alleles),
- enriched sequencing of resistance genes,
- utilisation of sequences for gene annotation,
- comparison and orthologue analysis.

## D 7.1 Production of materials for improved genotyping training

### Software commonly used for sequence analysis:

- <https://blast.ncbi.nlm.nih.gov/>
- The BLAST Sequence Analysis Tool; <https://www.ncbi.nlm.nih.gov/books/NBK153387/>
- PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics; <https://www.ncbi.nlm.nih.gov/pubmed/29069403>
- [https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=BlastHelp](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp)

### Online Tutorials

- Web page describing the Genome Insights platform with video therein. By Keygene; <https://www.keygene.com/technology/1-genome-insights/>.
- Illumina Sequencing technology by Illumina; <https://www.youtube.com/watch?v=womKfikWlxM>

### Further reading

- Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771.
- Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351.
- Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics and Bioinformatics* 14, 265–279.
- Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *PNAS U. S. A.* 74, 560–4.
- Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *PNAS U. S. A.* 74, 5463–7.
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X-z, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature.* 452: 872–876. doi:10.1038/nature06884. ISSN 0028-0836. PMID 18421352.

### 20. Chromosome engineering

Chromosome engineering (CE) represents a complex cytogenetic approach, which allows the transfer of chromosomal segments harbouring genes of interest from related species to cultivated wheats. It is a powerful technology for obtaining targeted genetic gains in wheat by attaining to unexploited variability present in wheat wild relatives (WWRs). By using the *Ph* (Pairing Homoeologous) system, CE promotes recombination between closely related chromosomes of wheat and WWRs to induce homoeologous pairing and an exchange of genetic material between them i.e. crossing-over (Fig. 20.1). To obtain wheat-alien recombinant lines, CE employs natural sexual means in plants, and in this way, it represents a valid GMO alternative for the enlargement of the genetic basis of cultivated wheats. CE approach integrated with continuously developing techniques of genome and chromosome analysis (e.g., marker-assisted selection, association mapping, next-generation sequencing, *in-situ* hybridisation), represents a unique platform for the creation of novel and breeder-friendly genetic stocks.

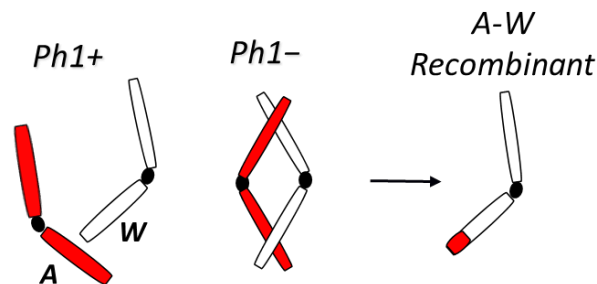


Fig. 20.1. Action of *Ph* (Pairing homoeologous) gene on homoeologous pairing and recombination (*Ph1+*, actively expressed gene; *Ph1-*, mutated and inactive gene; A, alien; W, wheat) Permissions for the figure: Prof. Carla Ceoloni, University of Tuscia, Viterbo, Italy).

The first steps in the application CE methodologies for the transfer of exotic genes from Triticeae species into cultivated wheat were made in the 1960s and 1970s, by the great cytogeneticist Ernest Sears. This was possible after the complex control of chromosome pairing was understood in polyploid wheats and their hybrids with closely related species. Multiple genes are involved in the control of inter- and intra-specific pairing, to limit the pairing and recombination only to homologous chromosomes of each genome. In particular, the *Ph1* gene, located on the 5B chromosome of both bread and durum wheat, is the main regulator of chromosome pairing between the homeologs i.e. chromosomes that share only partial homology. Initially, the *Ph1* function seemed to be the blocking of homoeologous pairing, yet from recent studies, it emerged that the *Ph1* function is dual: 1) it does not prevent the homoeologous from pairing but rather strongly promotes homologous pairing and 2) it impedes the crossing-over between the homeologs. The isolation of *ph1* mutants in bread and durum wheat offered a very important tool to overcome the interspecific mating incompatibility among the Triticeae species.

In inter-specific hybrids created by using the *ph1* mutants, or in wheat *ph1* mutant genotypes in which single alien chromosomes were added to or used to substitute wheat chromosomes (i.e., in addition and substitution lines, respectively),

## D 7.1 Production of materials for improved genotyping training

homoeologous chromosomes can “recognise” themselves and then to pair and recombine. To be fully effective, the *ph1* mutation has to be in a homozygous state. By crossing a *ph1* genotype with a *Ph1* one, no further recombination is possible.

The introgression products obtained by CE and *ph1* mutants are overall balanced as they result from the exchanges between homeologs. Accordingly, CE offers a possibility to subject primary recombinants to further events of recombination to progressively diminish the amount of alien chromatin. This is possible when *Ph1* remains in the condition of recessive homozygosity (*ph1/ph1*) throughout multiple generations, before re-establishing the normal condition of dominance. It is possible to obtain a similar result also by crossing two wheat-alien recombinants that have shared portions of alien chromatin (i.e. homologous), and both recombinants containing the gene(s) of interest. The result may be a new recombinant with a significant reduction in exotic DNA (Fig. 20.2). Normally, the selection favours the recombinants with a lower amount of alien DNA, which is highly important when having in mind that the alien segment contains also non-target genes and some of them may have adverse effects on plant characteristics. This is the phenomenon known as “linkage drag”. To avoid such events, it is necessary to have efficient selection systems for the traits of interest and to introgress the least possible alien chromatin into wheat, particularly in the case of tetraploid durum wheat, which concerning hexaploid bread wheat has lower buffering capacity for chromosome manipulations.

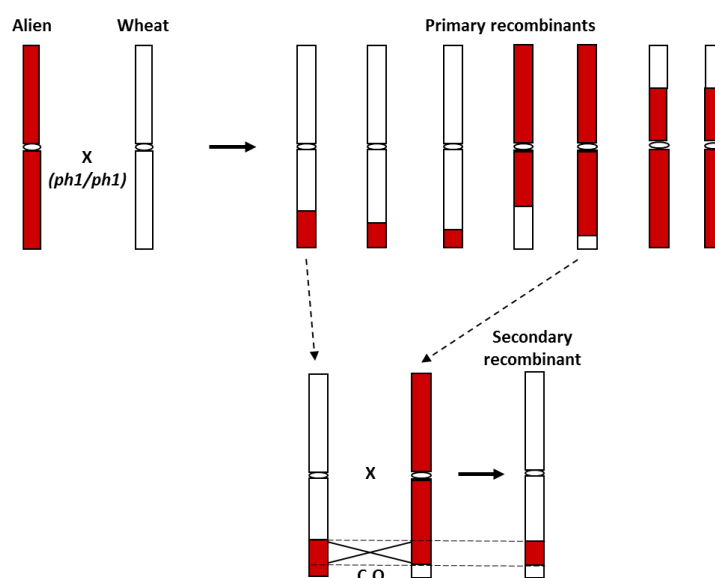


Fig. 20.2. Creation of primary and secondary recombinants through chromosome engineering. Each chromosome scheme represents a separate genotype (C.O., crossing-over) (Permissions for the figure Ljiljana Kuzmanovic, University of Tuscia, Viterbo, Italy).

In bread wheat, many genotypes containing entire chromosome arms transferred from wild species were obtained without any negative effect and were largely used in cultivation. The most known is the 1RS arm centromeric translocation from rye (*Secale cereale*) to bread and durum wheat chromosome 1B. This translocation is known to carry multiple resistance genes (*Lr26*, leaf rust; *Sr31*, stem rust; *Yr9*, yellow rust; *Pm8*, powdery mildew) and to increase the yield. Another example is the

## D 7.1 Production of materials for improved genotyping training

transfer onto wheat chromosome 6A of the 6AgL arm of *Agropyron elongatum*, harbouring an efficient resistance gene to stem rust *Sr26*. However, both transfers described were associated with negative effects on grain and gluten quality, which limited their extensive exploitation in breeding.

### 20.1. Chromosome engineering in durum wheat

Fine chromosome manipulations through CE are particularly important in durum wheat, which, due to its lower ploidy level is less tolerant to extensive chromosome alterations with respect to bread wheat. This is also one of the reasons for the lower number of cases of chromosomally engineered durum wheat genotypes so far. However, given the big economic importance of durum wheat in the Mediterranean Basin and other areas worldwide (e.g. North America's Great Plains, Mexico, Kazakhstan, South Australia etc.), and that the cultivation areas of the crop will likely enlarge and be dislocated in new zones, to have a sustainable approach for its genetic improvement such as CE is a desirable asset.

In recent decades, some important results in durum wheat breeding through CE were obtained at the University of Tuscia (Viterbo, Italy). Stable and breeding exploitable durum wheat-alien recombinant lines were produced, by introgression of some very useful genes, including the following:

1. *Lr19+Sr25+Yp* genes, conferring efficient resistance to leaf rust, stem rust, and increased carotenoid content of the endosperm, respectively, transferred from chromosome arm 7eL<sub>1</sub>L of the decaploid tall grass *Thinopyrum ponticum* onto 7AL arm of durum wheat (Ceoloni et al. 2005);
2. *Pm13* gene conferring resistance to powdery mildew, transferred from 3S<sup>1</sup>S arm of the diploid *Aegilops longissima*, onto 3BS arm of durum wheat (Biagetti et al. 1999);
3. *Glu-D1* and *Glu-D3* genes, one of the major factors determining gluten quality, transferred from the 1D chromosome of bread wheat onto 1A chromosome of durum wheat (Vitelozzi et al. 1999; Gennaro et al. 2012).

Different recombinant lines were obtained for each of these transfers, but those that were selected for breeding purposes had the shortest alien segment possible still containing genes of interest. In all cases, those segments occupied about 20-25% of the recipient chromosome arm, as revealed by genomic *in-situ* hybridisation (GISH). GISH represents an extremely valid tool for characterisation and isolation of primary recombinants, yet in breeding programs, it is faster and cheaper to use appropriate molecular markers. In this way, genetic maps of chromosomal regions of interest can be enriched with various co-dominant and/or dominant markers, easily visualised by PCR, which then allows an easy determination of homozygous and heterozygous genotypes in segregating populations for each of the recombinants described above.

Additional assessment of the agronomic performance of three of the *Lr19+Sr25+Yp* recombinants having 23, 28 or 40% of 7AL arm replaced by the 7eL<sub>1</sub>L segment, revealed that the same 7eL<sub>1</sub>L segments harbour also several QTLs for important yield-related traits such as grain number/spike, flag leaf dimensions, tiller number, biomass and yield (Fig. 20.3). This allowed for the first time for a structural-functional dissection of overall 40% of 7eL<sub>1</sub>L DNA present on durum wheat 7AL arm in the three recombinants analysed. In drought-prone environments, the recombinants R5 and R23 were shown also to be significantly more productive with



## D 7.1 Production of materials for improved genotyping training

respect to their controls lacking the segment (higher grain number and yield, higher biomass, and harvest index), which is encouraging for their future use in breeding pipelines, and variety constitution. The recombinant R5 was registered in 2010 under the name Cincinnato as the first chromosomally engineered European variety of durum wheat. In addition to good productive features of the Cincinnato genotype across a variety of environments challenged with drought or leaf rust, the genotype gave good results as for quality parameters, in association with the presence of the *Yp* (=Yellow pigment) gene on its  $7e_1L$  segment. Namely, an increase in the yellow index of semolina and carotenoid pigments i.e. antioxidants were observed, which are of particular importance in contrasting the onset of pathologies and aging process in humans as well as of deterioration of quality of food products.

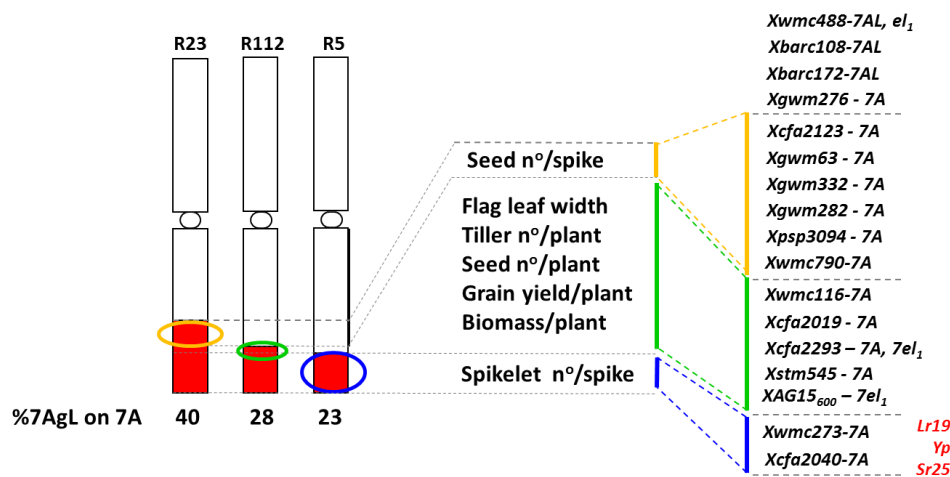


Fig. 20.3. Cytogenetic map of three durum wheat-*Th. ponticum* recombinants assessed for agronomic traits with association of some of the traits with specific alien DNA segment (white area, wheat 7A chromosome; red area,  $7e_1L$  chromosome) (adapted from Kuzmanović et al. 2014).

## 20.2. Pyramiding multiple transfers by chromosome engineering in durum wheat

The examples of alien segment introgressions in durum wheat described above validate CE as an effective approach to improve wheat when used together with proper analytical tools for selection. Moreover, reaching some even more ambitious goals of breeding through this unconventional methodology does not seem unrealistic. One of them is to combine i.e. pyramid multiple alien chromosomal segments within a single wheat genotype, and so increase the content of target genes in a single individual. This strategy can be pursued in two different ways:

1. **Stacking multiple alien segments** through crossing between single recombinants - applied a few times in bread wheat and only in one case of durum wheat. In bread wheat for example, the widely used 1RS centromeric translocation on 1B chromosome was combined in one case with the  $7e_1L$  translocation on 7D arm (Singh et al. 1998) and another with *Th. intermedium* 4Ai#2S translocation on 4D chromosome (Ali et al. 2016), to combine multiple resistance genes from rye (see above) and *Th. ponticum/Th. intermedium*. As for durum wheat, the only case reported so far is the combination of segments described above, originating from

## D 7.1 Production of materials for improved genotyping training

*Th. ponticum* (7eL<sub>1</sub>L), *Ae. longissima* (3S<sup>1</sup>S) and *T. aestivum* (1D), harbouring *Lr19+Sr25+Yp*, *Pm13* and *Glu-D3* genes, respectively (Gennaro et al. 2007). The triple recombinant in durum wheat showed normal segregation of the alien segments i.e. good tolerance of the recipient genotype, and improved yield and gluten quality, which makes it a good candidate for the development of dual-purpose durum wheat, suitable for both pasta and bread making.

2. **Nesting of multiple alien segments** through homeologues recombination – more exploited in bread than in durum wheat. In both species, in recent years important results were obtained in combining portions of group 7 chromosomes from *Thinopyrum* species onto 7D and 7A chromosomes of bread and durum wheat, respectively. That is, in the alien segment of *Th. ponticum* containing *Lr19+Sr25+Yp* genes (accession eL<sub>1</sub>), two other and different alien segments were nested in separate CE events, each one of them containing one highly efficient QTL against Fusarium diseases. In one case, the nested segment originated from another accession (eL<sub>2</sub>) of *Th. ponticum* (*Fhb-7eL<sub>2</sub>* QTL). In the other case, the segment was from the diploid *Th. elongatum* (*Fhb-7E* QTL). By these CE manipulations, in bread and durum wheat became available effective genetic sources for Fusarium head blight and Fusarium crown rot, the two devastating wheat diseases, which in recent decades have become a major threat for yields and quality (mycotoxin accumulation in grains) due to climatic changes and spreading of these diseases in new areas.

Based on chromosome engineering, it is possible to foresee that by combining the “gene pyramiding” strategy with classical and marker-assisted breeding, it will be possible to create an array of well-diversified new genotypes, able to significantly contribute to alleviating problems associated with current socio-economic and environmental challenges (population increase, climate change etc.).

### 21. Fluorescent *In-situ* hybridization - FISH or GISH

Hybridization *in-situ* (ISH) is a direct method of localising DNA sequences on chromosomes. It is based on the ability of the denatured DNA molecules (probes) to form duplexes with homologous DNA sequences of chromosomes on a slide. Better performances, of this molecular cytological method, could be obtained with the development of simpler and more efficient DNA tagging systems and better visualisation of the hybridisation signal like fluorescence probes: they are labelled by the insertion in the sequence of the nucleic acid of a modified nucleotide linked to a molecule of biotin or digoxigenin, or a fluorochrome (Leitch et al. 1994).

On the base of probes utilisation we distinguish: 1) Fluorescent *in-situ* hybridisation (FISH) that is used to study the distribution of individual DNA sequences on chromosomes, more useful in karyotyping, phylogenetic analysis or helping the rapid characterisation of the repetitive fractions of a genome in natural populations and agricultural plants (Chester et al. 2010; Cuadrado et al. 1997, 2010); 2) Genomic *in-situ* hybridisation (GISH), in which whole genomic DNA is used like a probe (Anamthawat - Jonsson et al. 1990), that is commonly used in characterization of genomes and chromosomes in hybrid polyploidy, hybrid plants, partial allopolyploids, polyhaploids and recombinant breeding lines (Raina et al. 2001; Chester et al. 2010; Salina et al. 2018).

Both GISH and FISH analysis are used for physical mapping and in their relationship with genetic maps (Ceoloni et al. 2017; Gennaro et al. 2012; Jiang and Gill 2006).

The articulation and the specifications for the implementation of FISH protocols are different especially according to the type of probe (short DNA sequences towards whole genomes) or chromosomal target (mitotic or meiotic chromosomes, tissues, or fibers); but for a schematic vision FISH/GISH (Fig. 24.1) work protocol's is divided into steps which can be summarised as follows:

1. probe preparation, labelling and control
2. preparation and pre-treatment of cytological preparations (on slide)
3. *in-situ* hybridisation, of the probe placed on the cytological preparation
4. post hybridisation washes and detection
5. closure of the slides and counter staining
6. observation by fluorescence microscopy.

#### 21.1. Probe preparation, labelling and control

Different methods are used for labelling DNA depending on the probes: Nick Translation (Table 21.1.), for genomic DNA or BAC clones (Anamthawat - Jonsson et al. 1990; Schwarzacher et al. 1992; Zhang et al. 2000); PCR for plasmid clones that were amplified and labelled; Random priming for SSR-oligonucleotides (Cuadrado and Jouve 2010) (it should be noted that for SSR-oligonucleotides the hybridisation protocol and detection is different from the one described below).

Probes sizes are checked for fragment length on gel electrophoresis; while dye incorporation is tested with chemical-colorimetric Dot blot reaction if biotin or digoxigenin-dUTP are used in the labelled reaction.

## D 7.1 Production of materials for improved genotyping training

Table 21.1. Nick Translation (Fermentas) labelling mixture with digoxigenin and/or biotin.

Components	initial concentration	Final conc.
Reaction buffer	10X	1X
d(GAC)TP mix	0.4 mM	0.02mM
dTTP	1 mM	0.004mM
Dig-11-dUTP / Bio-11-dUTP	1mM	0.02mM
DNasi I, RNasi-free	0.002U/ $\mu$ l	0.002U
DNA Polimerasi I, E. Coli	10U/ $\mu$ l	20U
DNA (if genomic sheared up to 10-12 Kbp)		20 ng/ $\mu$ l
H <sub>2</sub> O distilled		Up to the volume

The mixture should be at 15 for 2 h

### 21.2 Preparation and pre-treatment of cytological preparations (on slide)

The hybridisation process involves the preparation of cytological preparations (slides), on which the chromosomes (often in mitotic metaphase) of the "receiving" plant (chromosomal DNA is a target) have been fixed for the pairing between probes DNA. Recovery of mitotic cells depends on the species, but frequently excised root-tips are used. The preparation of slides for *in-situ* hybridisation provides a removal or digestion treatment of the cell wall to favour the access of DNA probe towards the chromosomes fixed on a slide.

Before hybridisation, the preparation of chromosomal slides consists of pre-treatments with RNase, fixing with 4% paraformaldehyde and dehydration with alcohol washes.

### 21.3. *In-situ* hybridisation, of the probe placed on the cytological preparation

DNA, previously prepared as probes, is added to the test tube to specific components of the hybridisation mixture suitable to assist the hybridisation process; the final composition is shown in Table 21.2.

The hybridisation mixture is then denatured, at a temperature of around 70°C, when also formamide is present. Then placed on ice, and finally aliquoted on each slide and covered with a coverslip.

The next phase is the key one: the hybridisation, carried out in a wet chamber in a specific thermal cycler or in a thermostat, at specific temperatures (denaturation and annealing) determined according to the cases by the degree of similarity between the probe DNA and the chromosomal DNA target, by the type of probe used and the aim of the hybridisation. The duration of the hybridisation must always be modulated according to the results sought and therefore again by the type of interaction between the probe and the cytological preparation.

## D 7.1 Production of materials for improved genotyping training

Table 21.2. Final composition of hybridisation mixture.

Compounds	concentration
SSC	2X
SDS	0.25%
Formamide	50%
Dextran Sulfate	10%
Salmon sperm DNA	0,125mg/ml
Probe DNA Dig/Bio	About 100 ng/slide
Block DNA (autoclaved unlabelled genomic DNA, only in GISH)	In greater quantity than the probe
H <sub>2</sub> O	Add if necessary to complete volume reaction

### 21.4. Post hybridisation washes and detection

The post-hybridisation washes allow the elimination of the excess of the unhybridised probes and the breakage of non-specific bonds between probes and chromosomal DNA. After removal of the coverslip, the preparations are washed in 2xSSC at the same hybridisation temperature and then incubated at 37°C for 20'. Finally, they are washed at 37°C in 4 x SSCTween (0.2% Tween in 4xSSC).

In the case of indirect marking of the probes, with labelled nucleotides not directly linked to fluorochromes - for example with digoxigenin and / or biotin-, to visualise them the slides are subjected to a treatment with specific antibodies conjugated to fluorochromes (detection phase).

First, a saturating solution is placed on the slides, for example Blocking Reagent at 1% in 4xSSC / Tween, for few minutes at room temperature. Once the excess solution has been removed, a Block Reagent 1% solution is placed in 4xSSC/Tween containing the calibrated amount of antibody bound with the fluorochrome, for example "anti-digoxigenin-FITC" and/or "Cy3-streptavidin". The glass cover is affixed and the whole is kept at 37°C, in a humid chamber, for 1h. Fluorescein isothiocyanate, FITC, characterized by a green fluorescence ( $\lambda = 525$  nm) will detect the probe marked with digoxigenin while avidin - Cy3 that emits in red ( $\lambda = 570$  nm) will identify the probe labelled with biotin.

The slides at the end of the incubation are subjected to several washes, in 4xSSC/Tween at 37°C, again to remove all that is not strictly connected to the chromosomal preparation, and which would give a big disturbance in the detection of the light signal under the microscope.

## D 7.1 Production of materials for improved genotyping training

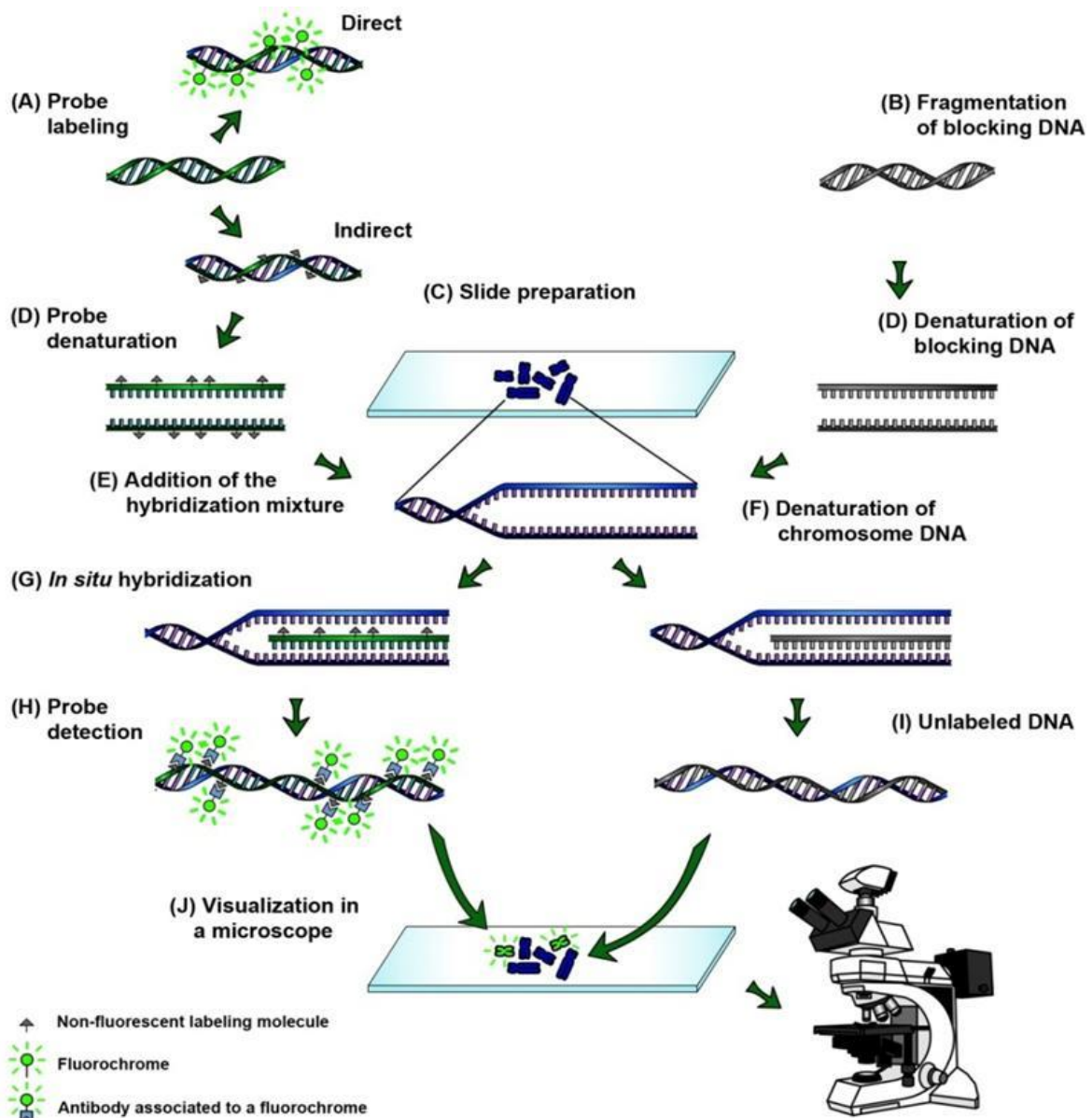


Fig. 21.1. Main steps of the genomic in situ hybridization (GISH).

(A) Direct and indirect probe labelling. (B) Fragmentation of the blocking DNA. (C) Slide preparation. (D) Probe and blocking DNA denaturation in a hybridisation mixture. (E) Addition of the hybridisation mixture with the probe and the blocking DNA. (F) Denaturation of the chromosome DNA. (G) *In-situ* hybridisation of probe and blocking DNA in the target sequence of the chromosome. (H) Detection of the probe in the chromosome DNA of one parent, in an indirect labelling. (I) Chromosome DNA molecule of the second parent associated to the unlabelled blocking DNA. (J) Visualisation of hybridisation signals associated to a probe (green) in a fluorescence microscope. Unmarked chromosomes are visualised with a counter-staining (blue). When the probe labelling is direct, the detection step of the GISH can be excluded. The fluorochromes are the signalling molecules and can be directly visualised in a fluorescence microscope with the appropriate filter. Santelmo Vasconcelos and Ana C. Brasileiro-Vidal (from Patussi Brammer et al. 2013.).

### 21.5. Closure of the slides and counterstaining

The fluorescent stain DAPI (4', 6-diamidino-2 phenylindole), characterised by blue fluorescence ( $\lambda = 450 \text{ nm}$ ), is the most widely used for the staining of the "background" chromatin (counter colour). For this purpose, DAPI is previously



## D 7.1 Production of materials for improved genotyping training

dissolved in antifade mounting medium, a compound that allows fluorescence to be preserved for longer, and then positioned on the slides.

The slides are closed again with glass cover-slips and placed in the dark at -20°C, before observation by an epifluorescence microscope with appropriate filters.

### Online Tutorials

- In situ Hybridization (ISH) and Fluorescence in Situ Hybridization (FISH) by Creative Diagnostics; <https://www.youtube.com/watch?v=KuhnEd8lCyo> *in situ hybridization*

### Further reading

- Anamthawat-Jónsson K, Schwarzacher T, Leitch AR, Bennett MD, Heslop-Harrison JS, 1990. Discrimination between closely related Triticeae species using genomic DNA as a probe. *Theor. Appl. Genet.* 79:721-728.
- Brammer SP, Vasconcelos S, Poersch LB, Oliveira AR, Brasileiro-Vidal AC, Anderson SB. 2013. Genomic in situ Hybridization in Triticeae: A Methodological Approach. *Plant Breeding from Laboratories to*, 1-22. <http://dx.doi.org/10.5772/52928>.
- Ceoloni C, Forte P, Kuzmanovic L, Tundo S, Moscetti I, De Vita P, Virili ME, D'Ovidio R. 2017. Cytogenetic mapping of a major locus for resistance to Fusarium head blight and crown rot of wheat on *Thinopyrum elongatum* 7EL and its pyramiding with valuable genes from a *Th. Ponticum* homoeologous arm onto bread wheat 7DL. *Theor. Appl. Genet.* DOI 10.1007/s00122-017-2939-8.
- Chester M, Leitch AR, Soltis PS, Soltis DE. 2010. Review of the Application of Modern Cytogenetic Method (FISH/GISH) to the Study of Reticulation (Polyploidy/Hybridisation) Genes, 1, 166-192. doi:10.3390/genes1020166.
- Cuadrado A, Schwarzacher T, Jouve N. 2000. Identification of different chromatin classes in wheat using *in-situ* hybridization with simple sequence repeat oligonucleotides. *Theor. Appl. Genet.* 101:711–717.
- Cuadrado A, Vitellozzi F, Jouve N, Ceoloni C. 1997. Fluorescence in situ hybridization with multiple repeated DNA probes applied to the analysis of wheat/rye chromosome pairing. *Theor. Appl. Gen.* 94: 347-355.
- Cuadrado Á, Jouve N. 2010. Chromosomal detection of simple sequence repeats (SSRs) using non-denaturing FISH (ND-FISH). *Chromosoma* 119:495–503. DOI 10.1007/s00412-010-0273-x.
- Gennaro A, Forte P, Panichi D, Lafiandra D, Pagnotta MA, D'Egidio MG, Ceoloni C. 2012. Stacking small segments of the 1D chromosome of bread wheat containing major gluten quality genes into durum wheat: transfer strategy and breeding prospects. *Mol. Breed.* 30: 149–167.
- Jiang J, Gill BS. 2006. Current status and status of fluorescence in situ hybridization (FISH) in plant genome research. *Genome* 49: 1057-1068. doi:10.1139/G06-076.
- Leitch AR, Schwarzacher T, Jackson D, Leitch IJ. 1994. *In situ Hybridization: a practical guide*. Royal Microscop. Soc., Microscopy Handbooks 27, BIOS Scient Publ Ltd Oxford UK.
- Raina SN, Rani V. 2001. GISH technology in plant genome research. *Methods in Cell Sci* 23: 83–104.
- Salina EA, Adonina IG. 2018. Cytogenetics in the Study of Chromosoma Rearrangement during Wheat Evolution and Breeding. *Cytogenetics-Past, Present and Further Perspectives*, 10.

## D 7.1 Production of materials for improved genotyping training

- Schwarzacher T, Anamthawat-Jonsson K, Harrison GE, Islam AKMR, Jia JZ, King IP, Leitch AR, Miller TE, Reader SM, Rogers WJ, Shi M, Heslop-Harrison JS. 1992. Genomic in situ hybridization to identify alien chromosomes and chromosome segments in wheat. *Theor. Appl. Genet.* 84: 778-786.
- Zang P, Kynast R, Friebe B. 2000. Protocol for gish genomic in situ hybridization - Kansas State University; <https://www.k-state.edu/wgrc/images/pdfs/GISH.pdf>.

### 22. Transcriptomics

Transcriptomics is the study of the transcriptome i.e. complete set of RNA molecules present in a cell. Transcriptome profiling is a powerful approach for identification of genes differentially expressed under different conditions and/or in different tissue/cell types. Compared to genomic studies, transcriptomics is *per se* more challenging as it must manipulate RNA, which is a molecule more unstable than DNA by being single-stranded and more prone to breaking down. RNA is more sensitive to heat and nuclease activity, the latter easily found everywhere and difficult to remove completely. Moreover, the amount of RNA in a cell is constantly changing compared to the stable amount of DNA. Therefore, the main goal of transcriptome analysis is to elucidate these RNA fluctuations in the frame of a studied biological phenomenon. Measuring gene expression in different tissues and/or under different conditions can shed more light on an organism's biology or functions of previously unknown genes.

Transcriptome consists of coding and non-coding (nc)RNAs i.e. translated or non-translated into protein, respectively. Coding or messenger (m)RNA has an intermediary role of passing the genetic information from the DNA, while the ncRNA (microRNAs, small interfering RNAs and long non-coding RNAs) has diverse supplementary regulation functions. In the past, most of the studies were focused on mRNA i.e., the expressed genes, which account for less than 5% of the genome in most species. On the other hand, during the last two decades the knowledge on variety in forms and functions of ncRNAs was constantly increasing, mainly due to advances in transcriptome sequencing and analysis technologies, so that transcriptomic studies have started to consider abundantly also the micro (mi)RNAs or Small Interfering (si)RNAs. With respect to traditional techniques for expression analysis such as northern blotting and *in-situ* hybridisation or real-time RT-PCR, which allows the analysis on a single or very small transcript group, new technologies offer the possibility of the simultaneous study of the entire transcriptome.

Transcriptomics in plants has been used to study transcriptional regulatory elements and to unravel mechanisms involved in the control of transcriptional regulation.

#### 22.1. Alternative splicing (AS) of mRNA

The pre-mRNA contains both introns and exons which can be processed by spliceosome (i.e. introns removed), to give multiple transcripts from a single gene (= isoforms) and so increase the complexity of the transcriptome. AS was studied in various plants (*Arabidopsis*, *Brachypodium*, rice, wheat, soybean, etc.) for various aspects (recognition of splice site, *cis*-regulatory sequences in the pre-mRNA, alternative polyadenylation, etc.) thanks to the availability of genome and transcriptome sequences and to the fact that about 20% of plant genes has AS. The main phenomena strongly influenced by AS in plants are plant development, photosynthesis, flowering and transition from vegetative to generative phases.

#### 22.2. ncRNAs (miRNAs and siRNAs) and their regulatory functions

ncRNAs are involved in plant response to pathogens, immunity, a wide range of abiotic stress tolerances (heat, drought, salinity, cold, heavy metals, nutrients,

## D 7.1 Production of materials for improved genotyping training

oxidation, hypoxia, UV-B), plant development, hormone signalling, reproductive function and are studied in tomato, oilseed rape, Arabidopsis, potato, rice, wheat etc.

**miRNAs** are 19-25 nt long stretches, derived from double-stranded or hairpin-like RNA (60-70 nt) and encoded by the host. They target and regulate the expression of numerous mRNA with which miRNA has partial complementarity. **siRNAs** are 21-22 nt long, generated from long double stranded RNAs, and in contrast to miRNA, can be exogenous or endogenous in origin, and specifically regulate the same genes that they originate from. Different aspects of ncRNA processing were studied in plants, including the formation of miRNA precursors and siRNA RISC complex, the exportation of mRNA out of the nucleus, etc.

### 22.3. Technologies for transcriptome analysis

Studies of the whole transcriptome started in the 1990s with techniques available at the time such as RNA-DNA hybridisation, subtractive hybridisation, subtraction cDNA libraries, and differential display. It is not until the progress in advanced sequencing technologies, from the year 2000 onward, that transcriptomics gained much more power and became the approach of choice for gene expression studies. There are two key methodologies for large-scale transcriptome profiling: hybridisation-based method (Microarray technology) and sequencing-based method (RNA-seq, MPSS, SAGE).

#### 22.3.1. Microarrays

Microarrays allow expression analysis of multiple (tens of thousands) chosen genes (= *probes*) through their hybridisation with fluorescently labelled cDNA or cRNA (anti-sense RNA) of the tested samples (= *target*). The chosen *probes* (*reporters* or *oligos*) are portions of the chosen genes or other DNA elements, which can be spotted on, or directly synthesised on solid support (glass, plastic, silicon biochip), and are denominated “cDNA” or “oligonucleotide” arrays, respectively. The first report on this technique was published in 1995 and until late 2000 it was the method of choice for transcription profiling. To apply microarray techniques, previous knowledge on gene sequence and/or function is needed for designing a specific probe set to be included on a biochip.

#### 22.3.2. SAGE (Serial Analysis of Gene Expression)

SAGE (Serial Analysis of Gene Expression) allows identification and quantitative comparison of a large number (thousands) of expressed genes without prior knowledge of their sequence, by using short gene tags. Namely, cDNA derived from RNA is digested into 11bp-long tag fragments by restriction enzymes that cut at specific sites, and then fragments are PCR amplified, transformed into bacteria for multiplication, and finally subjected to sequencing. SAGE is highly efficient in detecting rare or novel transcripts, yet very few studies in plants have used SAGE for transcriptome analysis.

#### 22.3.3. MPSS (Massively Parallel Signature Sequencing)

MPSS is another tag-based method that does not require knowledge on gene sequences or identity, similar to SAGE but with some considerably different technical steps. With respect to SAGE, MPSS uses longer cDNA tags (about 16-20 nt) and therefore is more specific. MPSS also has larger tag libraries and each tag is cloned onto a single microbead, for each microbead to contain only one tagged sequence. It allows the identification of millions of transcripts and is very sensitive to those with

## D 7.1 Production of materials for improved genotyping training

low abundance. The disadvantage of the technique is laborious cloning of the transcripts on the microbeads, and no free access for all researchers, limiting MPSS's application. MPSS technique was a precursor of RNA-seq development.

### 22.3.4. RNA-Seq (*whole transcriptome shotgun sequencing*)

By combining high-throughput sequencing technologies and computational procedures, it allows identification and quantification of all RNA content in a cell at a given time. It analyses both mRNAs and ncRNAs and helps in the discovery of new RNA classes. RNA-Seq generates reads of 30 to 10000 bp that are then aligned to reference genome or to each other (*de novo* transcriptome assembly). There are various applications and protocols but most include common RNA isolation and creation of cDNA, adding of adapters to generate a library to be used for sequencing. The crucial step represents the library preparation, which should be produced from intact mRNAs, captured with their poly A tail so that contaminating rRNAs and tRNAs are eliminated. For small RNAs sequencing, RNAs are captured by size. RNA-Seq protocols commonly used are called *single-* or *paired-end* experiment, depending on if single, or both end of the molecule is/are sequenced (50-100 bp at each end of a 200-400 bp-long segment), respectively.

The advantage of RNA-Seq, due to its power of deep sequencing is that it can reveal novel information, possibly missed by array-based platforms, as there is no need for knowledge of the sequence. Since the technique was described for the first time in 2006, it developed rapidly and overtook array-based approaches.

### 22.4. Applications of transcriptomic analysis in plant research and crop breeding

- Transcriptome assembly, including both coding mRNA and non-coding RNA.
- Simple mRNA profiling to quantify gene expression in a cell under different conditions and to identify up- and down- regulated genes, not dependent on the reference genome and so suitable for less studied species.
- Co-expression networks - analysis of similar expression patterns across tissues and conditions and construction of the expression atlases.
- Discovery of gene pathways involved in plant development and response to biotic and abiotic stresses to help in breeding for more resilient crops with higher yield under stress.
- Discovery of genes involved in secondary metabolite pathways.
- Gene function discovery.
- Marker discovery (e.g. SSRs, SNPs) and marker-assisted breeding - identifying functionally relevant molecular tags (novel genes, QTLs, alleles, haplotypes) that regulate qualitative as well as complex quantitative traits of agricultural importance.
- Discovery of rare genes, splice junctions, gene fusions, especially with poorly studied species.
- RNA editing (post-transcriptional alterations).
- Allele or genome-specific expression, particularly in polyploidy species.
- Domestication and evolutionary studies.

### 22.5. Transcriptome databases

The amounts of data being generated by the Microarray and RNAseq experiments are massive and ever-increasing, but they are also publicly available in various

## D 7.1 Production of materials for improved genotyping training

dedicated databases (DBs), serving as repositories. The data can be deposited as raw, processed, alone, or integrated on platforms also containing data on the corresponding genotype, phenotypes, gene networks, literature etc. On one hand, DBs can include general data obtained from different high-throughput technologies that require some bioinformatic expertise to be used (plant derived data of this kind represent less than 20% of the total). On the other hand, data contained in DBs may already be annotated, or mapped on standardised vocabularies and ontologies, may comprise co-expression DBs or the individual gene expression DBs, and hence more easily exploitable by users without specific bioinformatic skills.

### Examples

- ArrayExpress; [www.ebi.ac.uk/arrayexpress/](http://www.ebi.ac.uk/arrayexpress/)
- Gene Expression Omnibus (GEO); [www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)
- Rice Expression Database (RED) - repository of rice gene expression profiles derived entirely from RNA-Seq on tissues spanning an entire range of rice growth stages under a variety of biotic and abiotic treatments; <http://expression.ic4r.org/>
- RiceFRIEND - gene co-expression database in rice based on a large collection of microarray data derived from various tissues/organs at different stages of growth and development under natural field conditions, and phytohormone treatments; <https://ricefrend.dna.affrc.go.jp/>
- Expression Atlas - open science resource for gene expression across species and conditions (microarray and RNA-Seq), to help finding information on the abundance and localization of RNA (and proteins). Plant species included in Expression Atlas among all are rice, wheat, maize, tomato, and potato; <https://www.ebi.ac.uk/gxa/about.html>
- GENEINVESTIGATOR - high performance search engine to investigate in a single analysis gene transcriptional regulation across thousands of experimental conditions; <https://geneinvestigator.com/>
- NONCODE - is an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs); <http://www.noncode.org/>
- Wheat Expression Browser - RNA-seq data analysis and visualization platform for wheat; <http://www.wheat-expression.com/>
- KnetMiner – biological knowledge network integrating different data types, including transcriptome data, to allow identification of candidate genes; <http://knetminer.rothamsted.ac.uk/>

### Online Tutorials

- The Beginner's Guide to RNA-Seq by Applied Biological Materials; <https://www.youtube.com/watch?v=8IAVfKbRK3I>
- Intro to Transcriptomics by David Tabb; <https://www.youtube.com/watch?v=YofduWH3GVo>
- Introduction to RNA-Seq by BIO Platform; <https://www.youtube.com/watch?v=g0mmAgETTM8>
- OmicsLogic, Transcriptomics 2020, Next Generation Sequencing, RNA Seq Analysis by Pine Biotech; <https://www.youtube.com/watch?v=86v3iiiXkL8>
- Plant Transcriptomics Virtual Lab Simulation by Labster; [https://www.youtube.com/watch?v=DZfGS6\\_mUGw](https://www.youtube.com/watch?v=DZfGS6_mUGw)



## D 7.1 Production of materials for improved genotyping training

### Further reading

- Agarwal P, Parida SK, Mahto A, Das S, Mathew IE, Malik N, Tyagi AK. 2014. Expanding frontiers in plant transcriptomics in aid of functional genomics and molecular breeding. *Biotechnol. J.* 9: 1480-1492.
- Evans TG. 2015. Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation. *J. Exp. Biol.* 218: 1925-1935.
- Gong L, Zhang H, Gan X, Zhang L, Chen Y, Nie F, Shi L, Li M, Guo Z, Song Y. 2015. Transcriptome profiling of the potato (*Solanum tuberosum* L.) plant under drought stress and water-stimulus conditions. *PLoS One* 10: e0128041.
- Lawas LMF, Zuther E, Jagadish SK, Hinch DK. 2018. Molecular mechanisms of combined heat and drought stress resilience in cereals. *Curr. Opin. Plant Biol.* 45: 212–217.
- Liu H, Searle IR, Watson-Haigh NS, Baumann U, Mather DE, Able AJ, Able JA. 2015. Genome-wide identification of microRNAs in leaves and the developing head of four durum genotypes during water deficit stress. *PLoS One* 10: e0142799.
- Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. 2017. Transcriptomics technologies. *PLoS Comput Biol* 13: e1005457.
- Lu QH, Wang YQ, Song JN, Yang HB. 2018. Transcriptomic identification of salt-related genes and de novo assembly in common buckwheat (*F. esculentum*). *Plant Physiol. Biochem.* 127: 299-309.
- Nejat N, Ramalingam A, Mantri N. 2018. Advances in Transcriptomics of Plants. In: *Plant Genetics and Molecular Biology*, Springer, Cham. pp. 161-185.
- Ramesh SV, Govindasamy V, Rajesh MK, Sabana AA, Praveen S. 2019. Stress-responsive miRNAome of Glycine max (L.) Merrill: molecular insights and way forward. *Planta*: 1-18. <https://doi.org/10.1007/s00425-019-03114-5>.
- Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, Van Ex F, Pasha A, Khedikar Y, Robinson SJ, Cory AT, Florio T, Concia L, Juery C, Schoonbeek H, Steuernagel B, Xiang D, Ridout CJ, Chalhoub B, Mayer KFX, Benhamed M, Latrasse D, Bendahmane A, International Wheat Genome Sequencing Consortium, Wulff BBH, Appels R, Tiwari V, Datta R, Choulet F, Pozniak CJ, Provar NJ, Sharpe AG, Paux E, Spannagl M, Bräutigam A, Uauy C. 2018. The transcriptional landscape of polyploid wheat. *Science* 361, eaar6089.
- Rodrigues CM, Mafra VS, Machado MA. 2014. Transcriptomics. In: (Eds.) Borem A and Fritsche-Neto R, *Omics in Plant Breeding*. JohnWiley & Sons, Inc. pp 33-57.
- Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics* 10: 57–63.
- Zhang R, Marshall D, Bryan GJ, Hornyik C. 2013. Identification and characterization of miRNA transcriptome in potato by high-throughput sequencing. *PLoS One* 8: e57233.

### 23. Biological interpretation of gene expression (transcriptomics) data

#### 23.1. GO and MapMan plant ontologies

Important for proper biological interpretation of gene expression data are ontologies that classify genes according to their function and enable assessment of their representation in a large-scale transcriptomic study.

Two of the most widely used plant specific ontologies are Gene Ontology, organized as a direct acyclic graph into collections of concepts and their relations (The Gene Ontology Consortium, 2015), and MapMan ontology, organized as a hierarchical tree of functional categories (Thimm et al. 2004) that each have their own advantages (Klie and Nikoloski 2012). MapMan ontology was created specifically to study the functional pathways in plants and its organisation allows for compartmentalised manual curation of ontology down to the level of specific metabolic pathways and genes (Rotter et al. 2007; 2009; Ramšak et al. 2014). While up-to-date ontologies facilitate the functional interpretation of large-scale data, in turn, their curation also depends on the knowledge obtained from such studies. MapMan ontology and software have gained reputation as an important tool in mapping gene associations and showing pathways in plants, being a flexible tool relying on improvements from the scientific community.

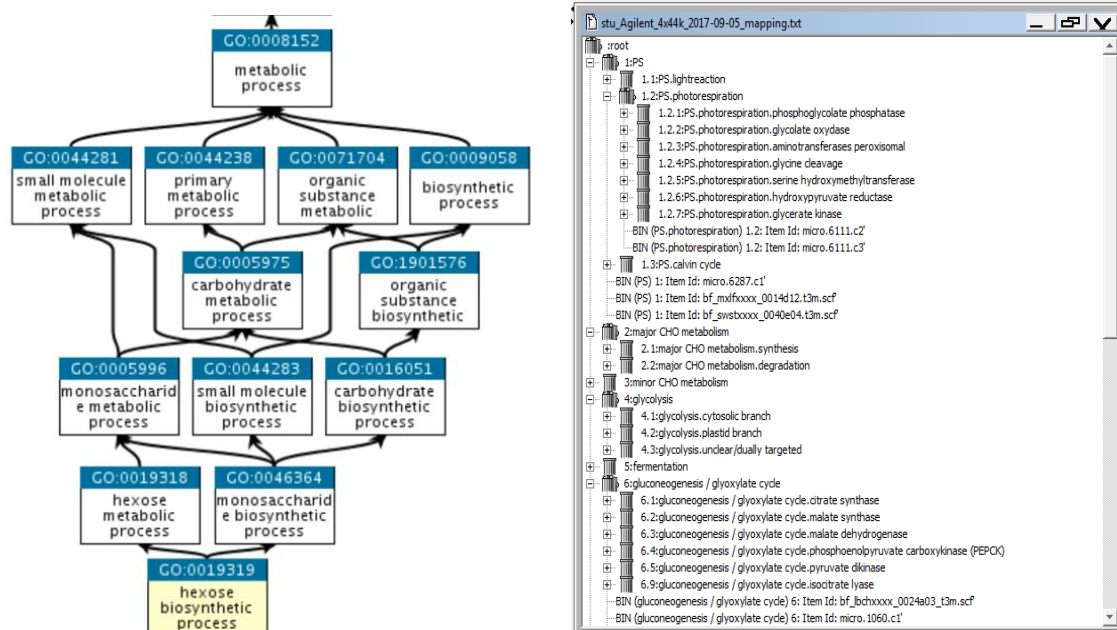


Fig. 23.1. Ontologies organisation. (a) Gene Ontology is organised as a direct acyclic graph (<http://geneontology.org/docs/ontology-documentation/>). (b) MapMan ontology is organised as a hierarchical tree of functional categories.

#### 23.2. MapMan

MapMan is a tool for displaying large data sets onto visual diagrams that symbolically depict biological functional pathways (Thimm et al. 2004). Individual genes are represented by discrete signals and are initially organised into functional blocks rather than as pathways. This allows even for genes whose function is only

## D 7.1 Production of materials for improved genotyping training

approximately known to be assigned to tentative categories. Grouping of genes allows for detection of trends that are less apparent when a list of individual genes is manually inspected. Defined functional categories can be a metabolic process, a particular cellular function (such as protein synthesis), a biological response (such as genes involved in metabolism or stress response) or a particular type of enzyme (e.g. cytochrome P450) for example in the case of the large gene families encoding classes of enzymes for which the function of most of the members is not well known. The use of hierarchical categories and diagrams with increasing detail enables analysis of different functional areas at different levels of resolution, depending on the question of interest and the amount of prior information available.

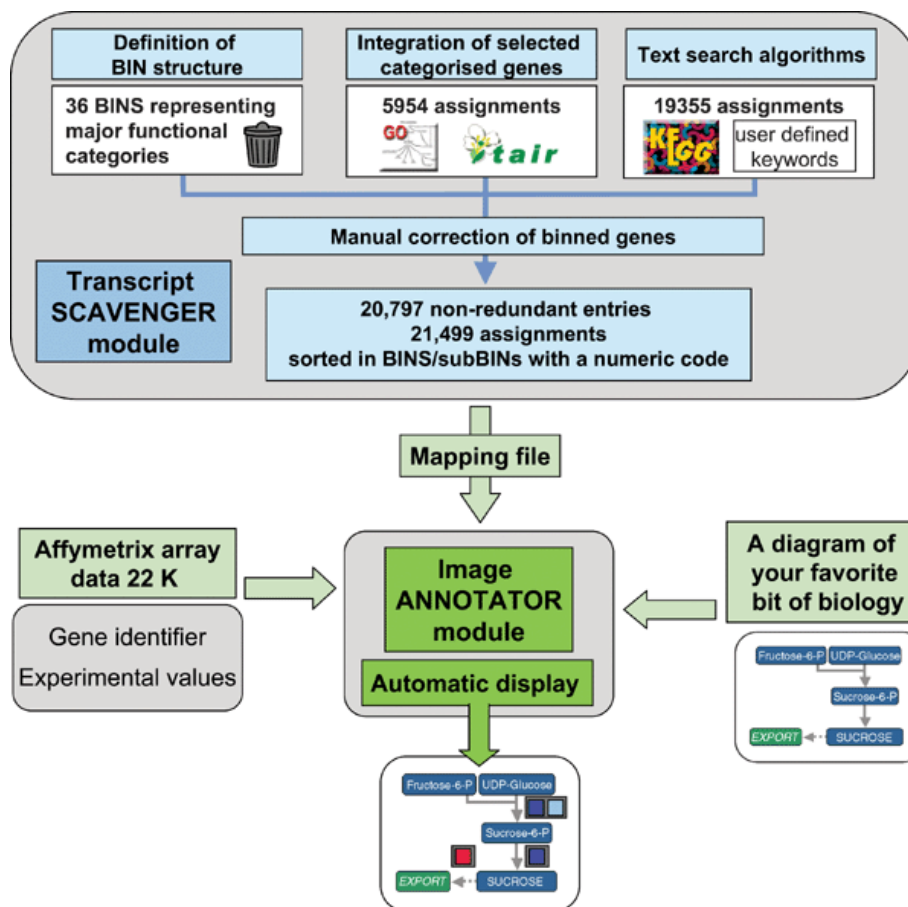


Fig. 23.2. Overview of the structure of MapMan (Thimm et al. 2004).

### 23.3. GoMapMan

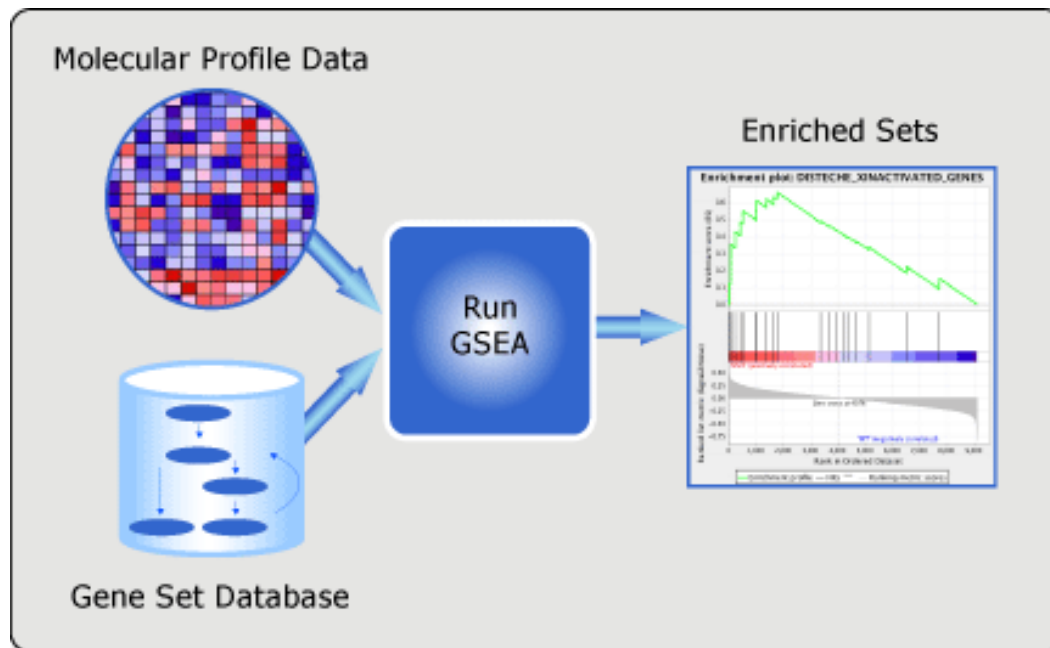
GoMapMan (<http://www.gomapman.org/>) is an open web-accessible resource for gene functional annotations in plants (Ramsak et al. 2013). The platform was developed to facilitate improvement, consolidation and visualisation of gene annotations across several plant species including model species and crops. The basis for GoMapMan is plant specific MapMan ontology, organised in the form of a hierarchical tree of biological concepts describing gene functions. GoMapMan is divided into three subsets. Protein GoMapMan is a resource for protein-coding genes of various plants described with MapMan ontology terms and grouped according to implemented orthologue groups. Metabolite GoMapMan is a resource for plant metabolites functionally annotated with MapMan ontology terms. SmallRNA

## D 7.1 Production of materials for improved genotyping training

GoMapMan is a resource for Plant small RNA's functionally annotated with MapMan ontology terms.

### 23.4. GSEA

Computational method called Gene Set Enrichment Analysis (GSEA) is used to determine if an *a priori* defined set of genes shows statistically significant differences between different biological states or phenotypes (Subramanian et al. 2005). Different ontologies, such as Gene Ontology and MapMan ontology can be used to define sets of genes according to their common function, structure or origin.



#### Further reading

- Klie S, Nikoloski Z. 2012. The Choice between MapMan and Gene Ontology for Automated Gene Function Prediction in Plant Science. *Front. Genet.* 3. <https://doi.org/10.3389/fgene.2012.00115>.
- Ramšak Ž, Baebler Š, Rotter A, Korbar M, Mozetič I, Usadel B, Gruden K. 2013. GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic Acids Res.* 42. <https://doi.org/10.1093/nar/gkt1056>.
- Rotter A, Camps Céline, Lohse Marc, Kappel Christian, Pilati Stefania, Hren Matjaž, Stitt M, Coutos-Thévenot P, Moser C, Usadel B, Delrot S, Gruden K. 2009. Gene expression profiling in susceptible interaction of grapevine with its fungal pathogen *Eutypa lata*: extending MapMan ontology for grapevine. *BMC Plant Biol.* 9, 104.
- Rotter A, Baebler Š, Stitt M, Gruden K. 2007. Adaptation of the MapMan ontology to biotic stress responses: application in solanaceous species. *Plant Methods* 3, 10.
- Sedlar A, Gerič Stare B, Mavrič Pleško I, Dolničar P, Maras M, Šuštar-Vozlič J, Baebler Š, Gruden K, Meglič V. 2017. Expression and regulation of programmed cell death associated genes in systemic necrosis of PVYNTN susceptible potato tubers. *Plant Pathol.* 5, 1238-1252.

## D 7.1 Production of materials for improved genotyping training

- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550.
- Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J. Cell Mol. Biol.* 37, 914–939.
- Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M. 2009. A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ.* 32, 1211–1229.

### 24. Proteomics

In modern biology the term "Omics" refers to the analytical techniques used to characterise and quantify pools of biological molecules and to define their roles, relationships and actions in cells, tissues, and organisms.

Therefore, Proteomics is defined as a large-scale study that aims to investigate the entire protein inventory of a cell, "the Proteome", at a specific time point.

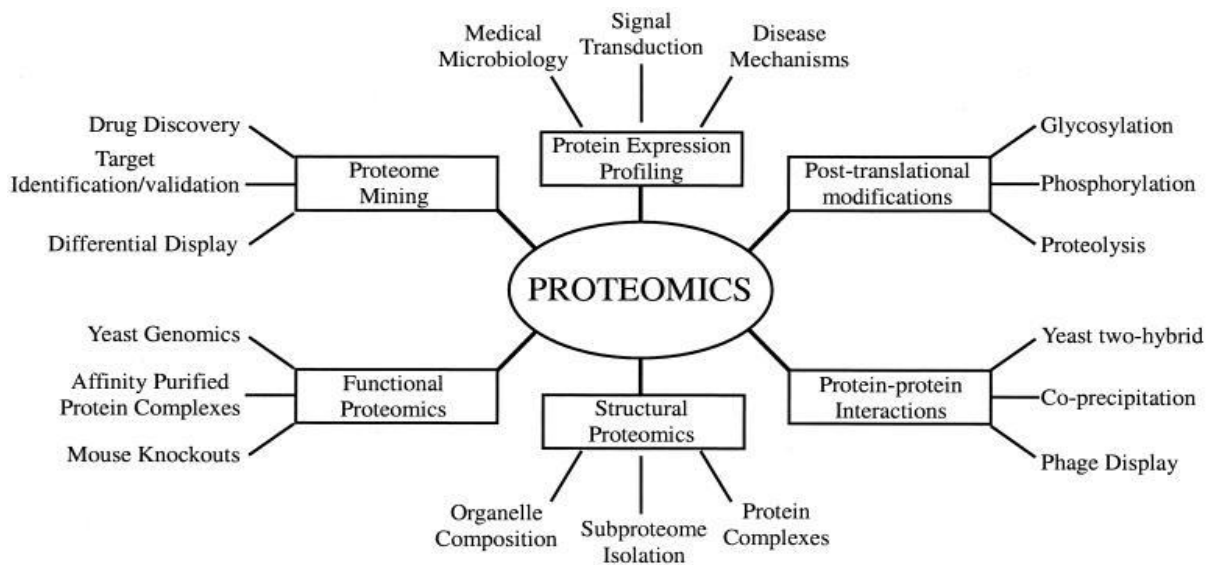


Fig. 24.1. Pathways of Proteomics.

Many different areas of study are now grouped under the rubric of proteomics (Fig. 24.1.). These include protein-protein interaction studies, protein modifications, protein function, and protein localisation studies to name a few. The aim of proteomics is not only to identify all the proteins in a cell but also to create a complete three-dimensional (3-D) map of the cell indicating where proteins are located. Proteomics to achieve these ambitious goals require the involvement of different disciplines such as molecular biology, biochemistry, and bioinformatics.

A generalised workflow of a proteomics experiment can be summarised, as shown below:

- 1 Sample preparation
- 2 Protein extraction
- 3 Protein separation (2-D Gel Electrophoresis or HPLC)
- 4 Protein Identification (Mass Spectrometry)
- 5 Data analysis

Proper sample preparation for MS-based analysis is a critical step in the proteomics workflow, because it can be both variable and time consuming. The Protein extraction protocols should ensure that most, if not all, proteins in a cell or its organelles are extracted to determine their function, structure, and interactions. Quality and reproducibility of sample extraction and *preparation* significantly impacts MS results. After a good extraction the proteins or peptides are separated and analysed by LC-based workflow or Gel-based workflow (Fig. 24.2.).



## D 7.1 Production of materials for improved genotyping training

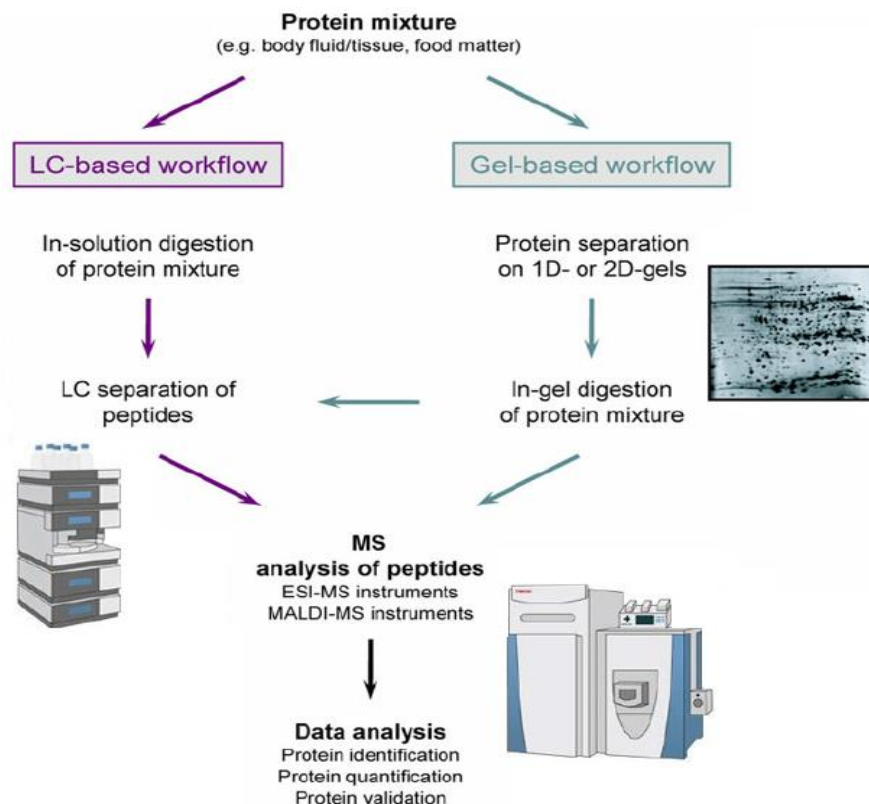


Fig. 24.2. LC-based workflow or Gel-based workflow.

A powerful tool for molecular analysis is represented by Mass Spectrometry (MS). Basically, MS measures the mass-charge ratio ( $m/z$ ) of the gas phase ions. Mass Spectrometers consist of a source of ions that converge the molecules of the analytes into ions in the gaseous phase, a mass analyser that separates the ionized analytes on the basis of an  $m/z$  ratio and a detector that records the number of ions for each value  $m/z$ . This application allows identifying and quantifying a large number of proteins and peptides from complex mixtures. Fig. 24.3. shows a general outline of the main components and functions of a mass spectrometer.

Quantification, in mass spectrometry-based proteomic applications, can be performed using either labelled or unlabelled approaches. Labelled approaches rely on mass-tagging of peptide reference standards either by direct synthesis or through chemical or metabolic means.

Unlabelled approaches rely on run-to-run comparisons, between test and reference samples. A labelled method, known as Isotope-Coded Affinity Tags (ICAT), relies on the labelling of protein samples, from two different sources with two chemically identical reagents that differ only in mass as a result of isotope composition. Differential labelling of samples by mass allows the relative amount of protein between two samples to be quantified in the mass spectrometer.

As mentioned earlier, quantification can also be achieved without labelling. There are a few different methods of quantification without labelling; these include spectral counting, peptide ion intensity counting and Spectral TIC counting.

## D 7.1 Production of materials for improved genotyping training

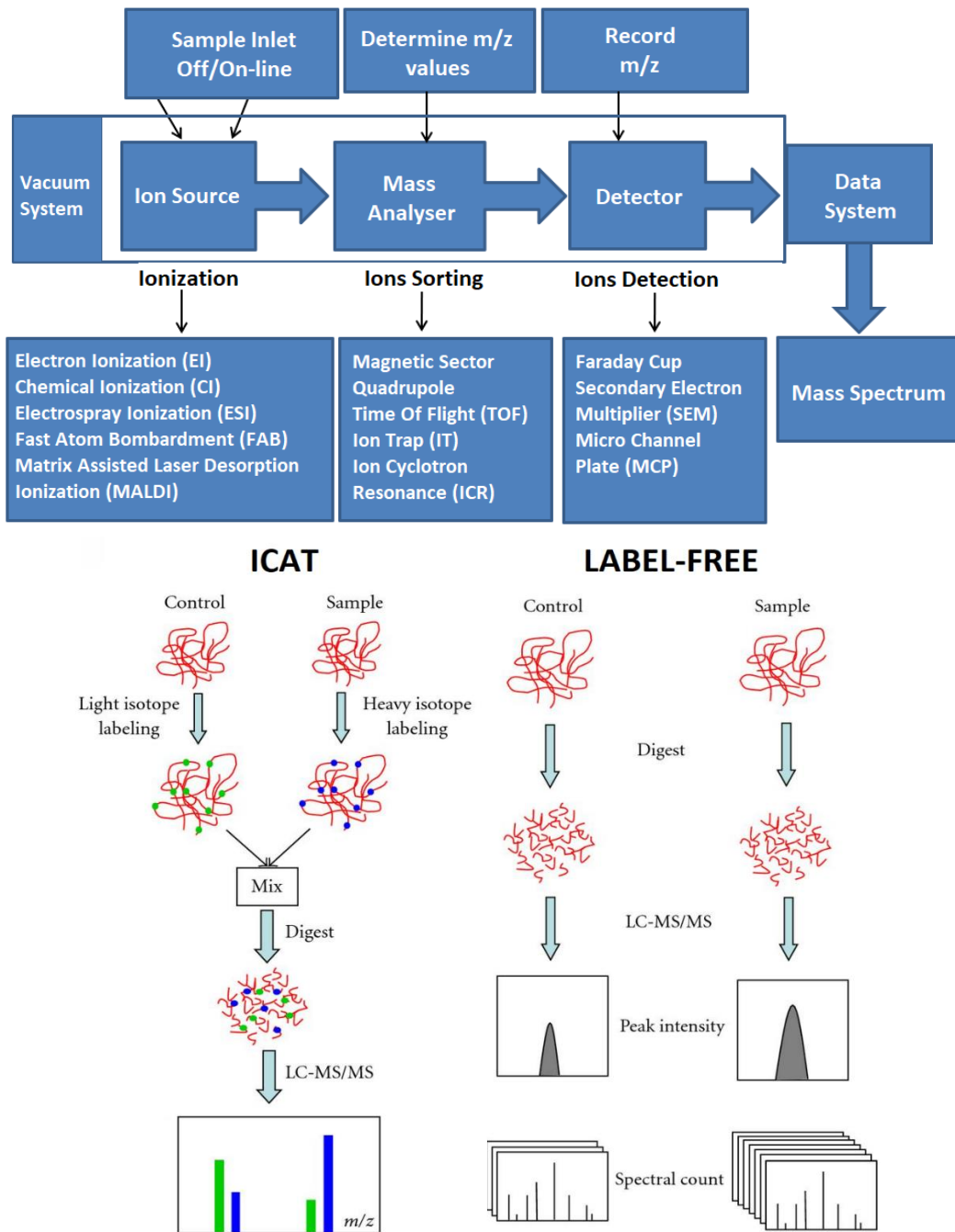


Fig. 24.3. Main components and functions of a mass spectrometer.

## D 7.1 Production of materials for improved genotyping training

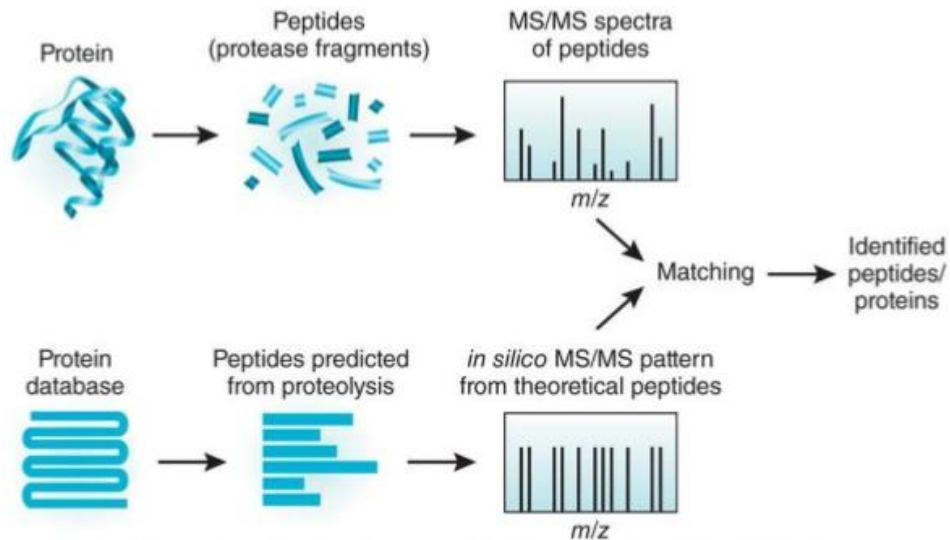


Fig. 24.4. Protein identification strategy by database search.

Software and Databases allow protein structural information harvested from Edman sequencing or MS to be used for protein identification. For example, Mascot is a software search engine that uses mass spectrometry data to identify proteins from peptide sequence databases. Mass spectrometers measure the molecular weights of peptides in a sample; Mascot then compares these molecular weights against a database of known peptides. Mascot then computes a score based on the probability that the peptides from a sample match those in the selected protein database. The more peptides Mascot identifies from a particular protein, the higher the Mascot score for that protein. Fig 24.4. shows a typical protein identification strategy by database search (<http://www.matrixscience.com/>).

### Online Tutorials

- Introduction to Proteomics by Matthew Padula University of Technology, Sydney; <https://www.youtube.com/watch?v=5paHhTq87Ak>
- Proteomics sample preparation by Matthew Padula University of Technology, Sydney; <https://www.youtube.com/watch?v=8c903dLrq7g>
- Mass spectrometry for proteomics - part one by Matthew Padula University of Technology, Sydney; <https://www.youtube.com/watch?v=v8EsEWwrJWs>
- Mass spectrometry for proteomics - part two by Matthew Padula University of Technology, Sydney; <https://www.youtube.com/watch?v=L6MHSb7I820>

# Section 4 - Statistics and bioinformatics

Section objectives:

The huge amount of data produced by the new technologies gives the necessity to manage them with adequate knowledge of statistics and bioinformatics. This is the objective of this section.

## 25. Statistical tests

### 25.1. Averages and dispersion indexes

#### 25.1.1. Mean, median and mode

Mean, median, and mode are three kinds of "averages" (Fig. 25.1). The "**mean**" is the "average" you're used to, where you add up all the numbers and then divide by the number of samples.

$$X = \sum x/n$$

The "**median**" is the "middle" value in the list of numbers (n). To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median. If n is odd, the median is the central value of the series itself; the number i provides the position of the data in the series with the following formula:

$$i = \frac{n+1}{2}$$

For example, if we have the following numbers (already ordered in a crescent way):

13, 13, 13, 13, 14, 14, 16, 18, 21

Since the numbers are nine, the middle one will be the  $(9 + 1) \div 2 = 10 \div 2 = 5^{\text{th}}$  number:

13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.

if n is even, none of the values is the central value of the series itself; the median is between the two central values and its position i will be:

$$\frac{n}{2} < i < \frac{n}{2} + 1$$

For example, if we have 1, 5, 8, 12, 23, 35. n = 6

$$\frac{6}{2} < i < \frac{6}{2} + 1 \Rightarrow 3 < i < 4$$

therefore, the median is between the third and fourth data of the series so between 8 and 12, it could be done also as the mean between 8 and 12 so 10.

The "**mode**" is the value that occurs most often. If no number in the list is repeated, then there is no mode for the list.

## D 7.1 Production of materials for improved genotyping training

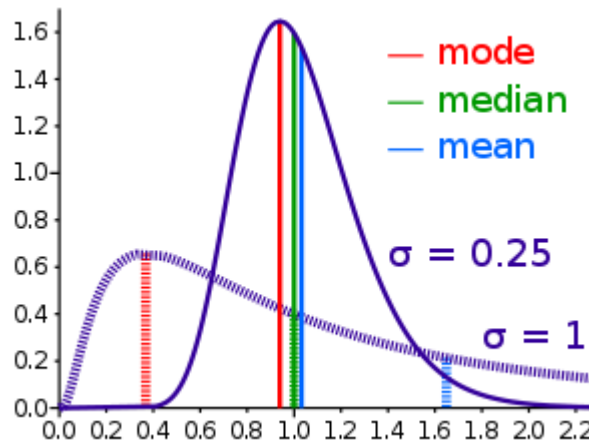


Fig. 25.1. Average curves.

A data distribution contains a complex set of information that is difficult to handle. The use of a central tendency index involves a strong simplification, and by itself does not provide exhaustive information on the distribution. It is also necessary to understand how the data is dispersed around the central trend index. Therefore, other than the central points to describe the curve we should have its magnitude, so we can distinguish among the curves in the following figure which have the same mean, median and mode (Fig. 25.2).

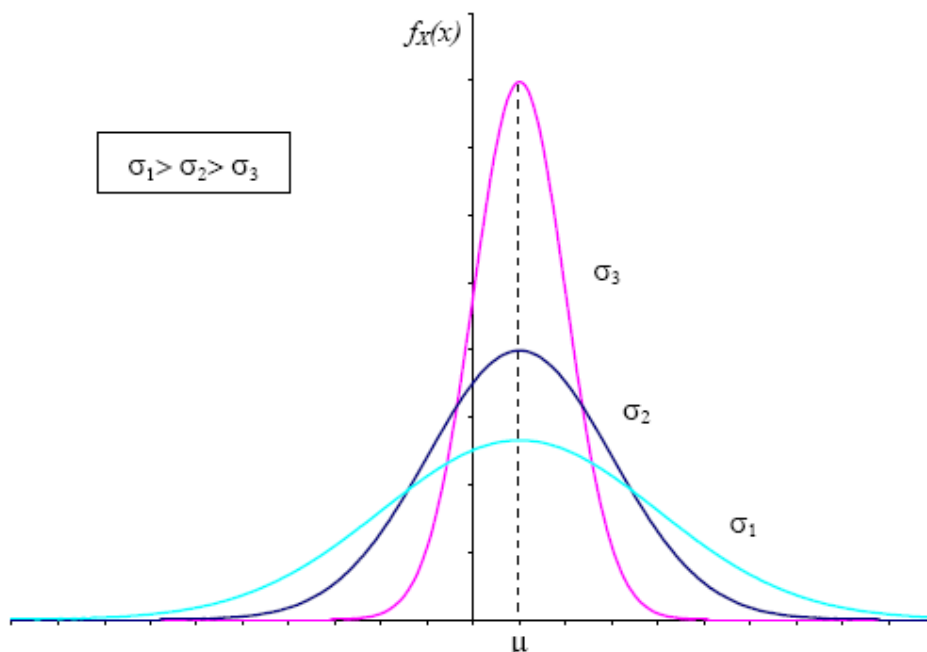


Fig. 25.2. Populations with different variation values and same means.

### 25.1.2. Dispersion indexes

These indexes serve to measure the dispersion (or variability) of a given distribution of data. For this reason, they are defined as dispersion indices or variability indices. The dispersion indexes can assume only positive values (it does not make sense to speak of negative dispersion) or zero values (in the cases in which all the observed data are equal to each other).

## D 7.1 Production of materials for improved genotyping training

The **range** of a set of data is the difference between the largest and smallest values.

The **variance** ( $\sigma^2$ ) of a data set is defined as the average of the squared differences between the data and the average of the data itself. It assumes the minimum value of 0 when the data are all equal to each other and increases as the variability of the data increases. The formulas for calculating variance are different depending on whether the data is grouped in classes or not. For not grouped data it could be used the formula:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

or

$$\sigma^2 = \frac{\sum_i x_i^2}{n} - \left( \frac{\sum_i x_i}{n} \right)^2$$

This is the average of squares plus the square of the average.

The **standard deviation** (SD or  $\sigma$ ), is the square root of the variance ( $\sigma = \sqrt{\sigma^2}$ ) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data. In addition to expressing the variability of a population, the standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. This derivation of a standard deviation is often called the "standard error" of the estimate or "**standard error of the mean**" when referring to a mean. It is computed as the standard deviation of all the means that would be computed from that population if an infinite number of samples were drawn and a mean for each sample were computed.

It is very important to note that the standard deviation of a population and the standard error of a statistic derived from that population (such as the mean) are quite different but related (related by the inverse of the square root of the number of observations). The reported margin of error of a poll is computed from the standard error of the mean (or alternatively from the product of the standard deviation of the population and the inverse of the square root of the sample size, which is the same thing) and is typically about twice the standard deviation - the half-width of a 95% confidence interval.

In science, many researchers report the standard deviation of experimental data, and only effects that fall much farther than two standard deviations away from what would have been expected are considered statistically significant - normal random error or variation in the measurements is in this way distinguished from likely genuine effects or associations.



## D 7.1 Production of materials for improved genotyping training

For example the weight variable was found in a sample of 20 subjects. In this sample the average is 70 kg and the standard deviation is 10.7. It can be stated that the subjects differ on average by 10.7 kg from the average weight of 70 kg.

In the normal distribution there is a fix proportion of the standard deviation and the relative population identified. In the Fig. 25.3. the dark blue is one standard deviation on either side of the mean. For the normal distribution, this accounts for 68.27 percent of the set; while two standard deviations from the mean (medium and dark blue) account for 95.45 percent; three standard deviations (light, medium, and dark blue) account for 99.73 percent; and four standard deviations account for 99.994 percent. The two points of the curve that are one standard deviation from the mean are also the inflection points.

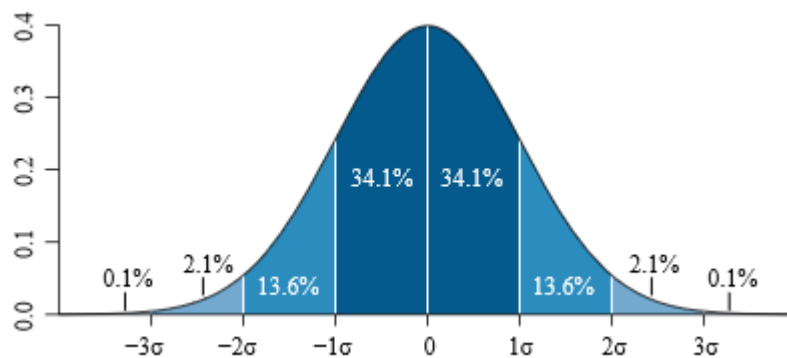


Fig. 25.3. Fix proportion of the standard deviation in a population with normal distribution.

The **coefficient of variation** is given by the ratio between the standard deviation and the absolute value of the average of the data:

$$C V = \frac{\sigma}{|\bar{x}|}$$

It is an index of relative variability, which takes into account not only the standard deviation of the data but also its average. For this reason it is very useful for making comparisons in terms of variability between "different" phenomena between them.

For example, in the gynecology and obstetrics department of a hospital, the weight of a sample of 80 male newborns and the weight of their respective dads were detected. The data obtained are reported in Table 25:

Table 25.1. The weight of a sample of 80 male newborns and the weight of their respective dads.

group	mean	Standard deviation
male new-borns	3.4 Kg	0.8
dad	82 Kg	15

One wonders if, with respect to the weight variable, there is more variability in the group of infants or in the group of fathers.

## D 7.1 Production of materials for improved genotyping training

Of course, the standard deviations comparison is not very helpful. They strongly depend on the average of the data on which they were calculated. To be able to make a comparison on the variability of the two groups it is appropriate to calculate the respective coefficients of variation:

$$\text{New-born CV} = 0.8/3.4 = 0.24$$

$$\text{Dads CV} = 15/82 = 0.18$$

Looking at the results it can be concluded that the group of children has a greater variability than that of the group of dads.

Some particular values of the CV that can be useful in the study of a data distribution:

- $CV = 0$ , in this case the standard deviation is equal to 0. All the data are equal to each other and the average can be considered as a perfect index to represent them.

- $CV \geq 0.5$ , in this case the standard deviation is more than half of the average. The average, in this case, cannot be considered a good index to represent the data.

- $CV \leq 0.5$ , in this case the standard deviation is less than half the average. The average, in this case, can be considered a good index to represent the data.

### 26.1.2.5. Quartile

In statistics, the quartile coefficient of dispersion is a descriptive statistic which measures dispersion and which is used to make comparisons within and between data sets.

Given a series of data, it is defined as Quantile of index  $p$  and is indicated by  $Q_p$ , the data below which a percentage  $p$  of data is located. For example, the median can be considered as the quantile  $Q_{50}$ , i.e. the data below which 50% of the data is located.

The Quartiles, the Deciles and Percentiles are nothing more than special Quantiles characterized by different values that the parameter  $p$  can assume.

The Quartiles divide the distribution of the data into 4 equal parts.

They are:

$Q_{25}$  indicates the data below which 25% (1/4) of the values are located.

$Q_{50}$  indicates the figure below which is 50% (2/4) of the values. It coincides with the Median.

$Q_{75}$  indicates the figure below which 75% (3/4) of the values are located.

The statistic is easily computed using the first ( $Q_1$  or  $Q_{25}$ ) and third ( $Q_3$  or  $Q_{75}$ ) quartiles for each data set. The quartile coefficient of dispersion is:

$$\frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The quantile calculation has different procedures depending on whether the data are grouped in classes or not. We report in detail the formulas for the calculation of percentiles (from them by analogy we can immediately derive the formulas for the calculation of deciles and quartiles).

## D 7.1 Production of materials for improved genotyping training

The position  $i$  of a given percentile, called  $p$  the percentile index and  $n$  the sample size, is given by the following formula:

$$i = \frac{n + 1}{100} \cdot p$$

For example, given a series of 19 ordered values, the position of the 20<sup>th</sup> percentile will be:

$$i = \frac{19 + 1}{100} \cdot 20 = 4$$

Therefore  $Q_{20}$  will assume the value of the fourth datum observed.

If the value  $i$  calculated in the previous formula is integer then the percentile sought coincides exactly with the data occupying the position  $i$ . If, on the other hand, the  $i$  value is not integer, linear interpolation must be used to determine the value of the percentile sought: if  $i = a, b$  then

$$Q_p = X_a + b \cdot (X_{a+1} - X_a)$$

For example, given the following numerical series: 1, 3, 4, 5, 8, 10, 12, 13, 15 ( $n = 9$ )

To calculate the 80<sup>th</sup> percentile the computation is:

$$Q_{80} = X_{\frac{9+1}{100} \cdot 80} = X_{\frac{10}{100} \cdot 80} = X_8 = 13$$

While if the data would be even such as 1, 3, 4, 5, 8, 10, 12, 13, 15, 18 ( $n = 10$ ), the computation would be:

$$Q_{80} = X_{\frac{10+1}{100} \cdot 80} = X_{\frac{11}{100} \cdot 80} = X_{8.8}$$

Position of the 8.8 data is valid only from a theoretical point of view. In order to calculate the value of  $Q_{80}$  we use linear interpolation from the 8<sup>th</sup> position (13) we computed the proportion 0.8 from 13 to 15:

$$Q_{80} = 13 + 0.8 \cdot (15 - 13) = 14.6$$

### 25.2. Correlation

The correlation aims to quantify the (linear) relationship between two or more random variables. The relationship between each pair of variables is associated with a number  $r$ , which could range from -1 to 1. It indicates the degree of correlation.

In most cases there is no manipulation of variables by the investigator.

Given  $N$  observations of 2 variables  $X_1$  and  $X_2$ , we can calculate their covariance as

$$\text{Cov}(X_1, X_2) = \Sigma_{12} = \frac{\sum (X_{1i} - \bar{X}_1) \cdot (X_{2i} - \bar{X}_2)}{N-1}$$

and their correlation as:

## D 7.1 Production of materials for improved genotyping training

$$\text{Corr}(X_1, X_2) = R_{12} = \frac{1}{N-1} \sum \frac{(x_{1i} - \bar{x}_1)}{\sigma_1} \cdot \frac{(x_{2i} - \bar{x}_2)}{\sigma_2}$$

The correlation is related to the angle between  $X_1$  and  $X_2$ .

If the variables are related,  $R \sim 1$

If they are anticorrelated,  $R \sim -1$

If they are weakly correlated,  $R \sim 0$

To determine the statistical significance of  $r$ , the experimental  $r$  should be compared with the  $r$  on the Pearson's correlation table (Table 25.2.). Where  $\alpha$  is the level of significance and  $df$  are the degree of freedom, which in the correlation cases is  $n-2$  since a degree of freedom is lost for each variable.

NOTE: we always talk about linear correlation

If  $R \sim 0$ , it cannot be excluded that between the two variables there is a nonlinear correlation (e.g., quadratic).

Table 25.2. Pearson's correlation table.

df \ $\alpha$	0.2	0.1	0.05	0.02	0.01	0.001	df \ $\alpha$	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999	35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000	40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139	45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068	50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883	60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904	70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260	80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115	90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047	100	0.127947	0.163782	0.194604	0.230079	0.253979	0.321095
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305	125	0.114477	0.146617	0.174308	0.206245	0.227807	0.288602
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962	150	0.104525	0.133919	0.159273	0.188552	0.208349	0.264316
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998	175	0.096787	0.124036	0.147558	0.174749	0.193153	0.245280
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351	200	0.090546	0.116060	0.138098	0.163592	0.180860	0.229840
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934	250	0.081000	0.103852	0.123607	0.146483	0.161994	0.206079
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657	300	0.073951	0.094831	0.112891	0.133819	0.148019	0.188431
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429	350	0.068470	0.087814	0.104552	0.123957	0.137131	0.174657
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163	400	0.064052	0.082155	0.097824	0.115997	0.128339	0.163520
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781	450	0.060391	0.077466	0.092248	0.109397	0.121046	0.154273
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208	500	0.057294	0.073497	0.087528	0.103808	0.114870	0.146436
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378	600	0.052305	0.067103	0.079920	0.094798	0.104911	0.133787
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230	700	0.048427	0.062132	0.074004	0.087789	0.097161	0.123935
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710	800	0.045301	0.058123	0.069234	0.082135	0.090909	0.115981
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768	900	0.042711	0.054802	0.065281	0.077450	0.085727	0.109385
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360	1000	0.040520	0.051993	0.061935	0.073484	0.081340	0.103800
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446	1500	0.033086	0.042458	0.050582	0.060022	0.066445	0.084822
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988	2000	0.028654	0.036772	0.043811	0.051990	0.057557	0.073488
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956	3000	0.023397	0.030027	0.035775	0.042457	0.047006	0.060027
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317	4000	0.020262	0.026005	0.030984	0.036773	0.040713	0.051996
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047	5000	0.018123	0.023260	0.027714	0.032892	0.036417	0.046512
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119							

The relationships could be represented graphically where the observations are represented as points on the Cartesian plane with the two variables on the two axes. The degree of correlation is high when the points are close to the "regression line", while it is low when the points far from the "regression line" (Fig. 25.4.).

The correlation could be utilised for:

1. **Forecast:** prediction of the value of a target variable based on the value of a predictor variable
2. **Validation:** comparison between the results of a new test and existing tests
3. **Reliability:** replicability of tests/experiments

Verification of theoretical predictions: verification of an expected relationship between two variables.

## D 7.1 Production of materials for improved genotyping training

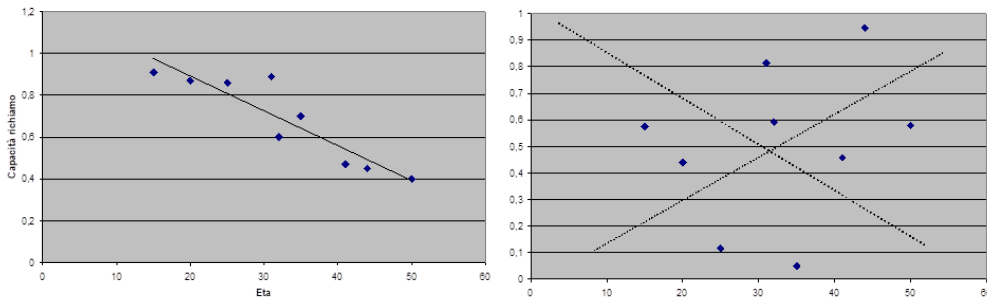


Fig. 25.4. Graphical representation of high correlation (left) versus low correlation (right).

There are some risks of interpretation:

1. A correlation relationship can be "spurious": it does not necessarily imply a relationship of cause and effect. For example: correlation between the number of priests and the number of murders.



Fig. 25.5. Example of direct (left) and spurious (right) correlations.

2. A correlation is valid only for a limited range of values, it should avoid the extrapolation of experimental results, but use only results with a wide range of X and Y values (Fig. 25.6).

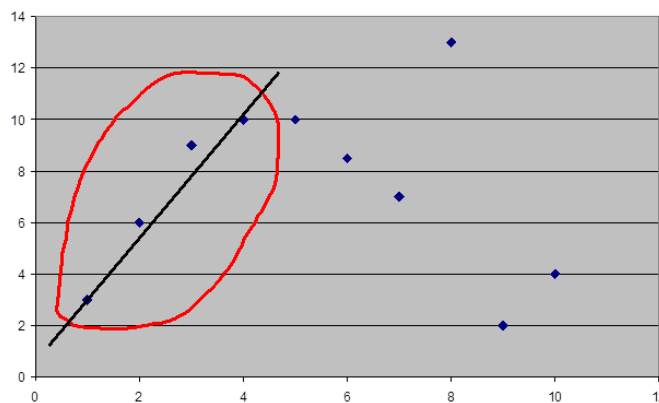


Fig 25.6. Example of correlation valid only for a limited range of values.

The coefficient of determination measures the percentage of the variability of Y explained by the variability of X.

### 25.3. Regression

The regression is a statistical technique for identifying the optimal regression line ("best fit") for a given set of observations. The regression identifies an equation that describes a linear relationship between two variables. This equation can be

## D 7.1 Production of materials for improved genotyping training

represented as a line (regression line). Regression analysis allows a display (facilitate understanding) of the relationship between two variables. It identifies the central tendency of the relationship (as well as the average) identifies the central tendency for a set of observations. With the regression it is possible to predict the value of Y for an unknown X (interpolation/extrapolation).

The optimal regression line is defined as that line which minimises the vertical distances between the observations and the line itself.

In the linear equation  $Y=aX+b$

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sigma_{XY}}{\sigma_X^2} \qquad a = \bar{Y} - b\bar{X}$$

The regression line should not be used to predict values of Y for values of X lower or higher than those included in the sample and the relationship between the two variables must be linear.



### 26. Comparison among methods and statistical software packages to analyse germplasm genetic diversity by means of codominant markers

Genetic diversity of germplasm is assessed by collecting key information i.e. (i) allele number per locus; (ii) genotype number per locus; (iii) gene diversity; (iv) PIC (polymorphism information content) values; (v) observed and (vi) expected heterozygosity; (vii) diversity within and between populations and (viii) the genetic distance among the analysed populations. The analyses are usually performed using a variety of molecular markers grouped into two categories: dominant and co-dominant markers (see section below on Markers and Molecular Tools). The dominant markers produce a series of bands with unknown relationships (i.e. could be allelic variants of the same genes or mark different genome regions). Hence, without knowing the allelic situation, each band is recorded as a locus with two possible alleles band presence (generally scored as 1) or band absence (generally scored as 0). The obtained 0/1 matrix is then used in statistical analyses.

The present section offers a guide to the principles that are the base of the most common analyses. It focuses on some of the most widely used computer programs in population genetics, run under Windows OS. The advantages and disadvantages of the various software packages are mentioned to facilitate their appropriate selection and use.

Most of the statistical computations use parameters based on the Hardy–Weinberg principle (see above section on Population Genetic and Hardy-Weinberg equilibrium). Here, it should be remembered that the square alleles' frequencies (i.e.,  $p^2+q^2+r^2$ , etc.) are homozygote frequencies while the others (i.e.,  $2pq+2pr+2qr$ , etc.) are heterozygote frequencies. Considering several alleles  $i$  with a frequency  $p_i$ , the frequency of all the homozygote is  $\sum p_i^2$  and heterozygote frequency can be calculated as the complementary difference from the homozygote frequency (i.e.  $2pq = 1 - (p^2+q^2)$  or  $1 - \sum p_i^2$ ).

#### 26.1 Gene diversity

The gene diversity index is calculated for each locus and population according to Nei (1973), utilizing the Hardy-Weinberg formula  $He = 1 - \sum_{i=1}^n p_i^2$  hereafter simplified as  $He = 1 - \sum p_i^2$ , which is the heterozygosity expected if a population is in Hardy-Weinberg equilibrium. In analogy, the genetic identity (J) is  $\sum p_i^2$  (homozygotes). However, since He could be computed for all populations, including non-random mating systems (e.g. autogamous which by definition are not in Hardy-Weinberg equilibrium being pure lines with homozygosity for all loci); the terminology for He is *gene diversity* rather than expected heterozygosity.

In a small population the alleles per locus can be skewed, especially when compared to large populations (Petit et al. 1998). Unbiased heterozygosity is as for the above-mentioned heterozygosity multiplied by the factor  $2n/(2n-1)$  (Nei 1978). As a result, the larger the population, the lower are the differences between the biased and unbiased expected heterozygosity. This detail is often not sufficiently elaborated upon in the literature, as many papers do not mention whether unbiased or biased He is used.

## D 7.1 Production of materials for improved genotyping training

The variability between and within populations can be calculated according to Nei (1973) taking into account different allele frequencies in whole populations or in sub-populations. The nomenclature used is:  $H_T$  for total observed diversity;  $H_S$  for within-population diversity and  $D_{ST}$  for the between-population diversity, with  $H_T = H_S + D_{ST}$ .

Similarly, Wright's fixation indices  $F_{IS}$ ,  $F_{ST}$  and  $F_{IT}$  (Wright 1965) is often used, also the F-statistics which are based on the expected level of heterozygosity. The measures describe the different levels of population structures such as variance of allele frequencies within populations ( $F_{IS}$ ), variance of allele frequencies between populations ( $F_{ST}$ ), and inbreeding coefficient of an individual relative to the total population ( $F_{IT}$ ), all of which are related to heterozygosity at various levels of population structure. The terms mentioned above are represented by the formula  $1 - F_{IT} = 1 - F_{IS} + 1 - F_{ST}$ , where I is the individual, S the sub-population and T the total population.  $F_{IT}$  thus refers to the individual in comparison to the total,  $F_{IS}$  is the individual in comparison with the sub-population, and  $F_{ST}$  is the sub-population in comparison with the total. As shown in Fig. 26.1, total F, indicated by  $F_{IT}$ , can be partitioned into  $F_{IS}$  (or  $f$ ) and  $F_{ST}$  (or  $\theta$ ).

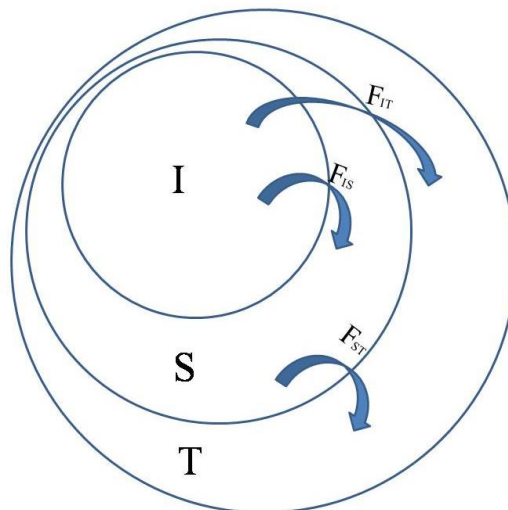


Fig. 26.1. Diagram of the relationships between the gene diversity components. I=individual, S=sub-population, T=total population.

$F_{ST}$  can be calculated using the formula:  $F_{ST} = (H_T - H_S) / H_T$  where  $H_T$  is the proportion of heterozygotes in the total population and  $H_S$  the average proportion of heterozygotes in sub-populations.

Considering a series of loci  $l$  in  $n$  populations by using the complementary sum of allele frequency  $(1 - \sum p_i^2)$  different figures can be obtained. In particular:

For each locus and each population,  $H_e = (1 - \sum p_{i(lg)}^2)$  where  $p_{i(lg)}$  is the  $i^{\text{th}}$  allele frequency of the  $l^{\text{th}}$  locus in the  $g^{\text{th}}$  population.

The average of the above  $H_e$  over populations gives the genetic diversity within a population for each locus, while the average of all the loci within population diversity gives  $H_S$ . The formula can thus be written as  $H_S = (\sum_l (\sum_g (1 - \sum p_{i(lg)}^2) / g) / l)$  where  $(1 - \sum p_{i(lg)}^2)$  indicates the expected heterozygosity for each locus in each population,  $g$  indicates the number of populations and the loci number.

## D 7.1 Production of materials for improved genotyping training

The total genetic diversity  $H_T$  is calculated using the allele frequency  $p_{i(l)}$  for each locus over all populations and calculating the mean over loci:  $H_T = \sum (1 - \sum p_{i(l)}^2) / l$ .

The between population component of diversity is calculated using the formula  $D_{ST} = H_T - H_S$ .

The between population component may also be expressed in relation to the total genetic diversity (for each locus and overall loci) as  $G_{ST} = H_T / D_{ST}$  (Nei 1973).

Table 26.1 shows an example extracted from Turpeinen et al. (2001) where different parameters are computed for three populations analysed with two markers. The  $H_T$  for each locus corresponds to the Polymorphic Information Content (PIC) of that locus, which in other words, consists of the capacity of that locus (or better a marker) to assess polymorphism and diversity. Botstein et al. (1980) proposed an adjustment of this value as:

$$PIC = 1 - \sum_{i=1}^n p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2$$

where  $P_i$  and  $P_j$  are the population frequency of the  $i^{\text{th}}$  and  $j^{\text{th}}$  alleles. The PIC proposed by Botstein and colleagues (Botstein et al. 1980) subtracts from the  $H_e$  value an additional probability ( $\sum \sum 2p_i^2 p_j^2$ ) due to the fact that linked individuals do not add information to the overall variation.

Table 26.1. Allelic situation and computation of the genetic parameters in three populations analysed using two markers where each one has three possible alleles (source: (Turpeinen et al. 2001)).

Locus\pop	Pop1	Pop2	Pop3				Mean
Locus 1		10	10				
167	0.00	0	0				0.00
168	0.50	0	0.9				0.47
172	0.50	1	0.1				0.53
He	0.50	0.00	0.18	$H_S$	0.23	$H_T$	<b>0.50</b>
Locus 2							
218	0.50	0.00	0.10				0.20
221	0.10	1.00	0.10				0.40
224	0.40	0.00	0.80				0.40
He	0.58	0.00	0.34	$H_S$	0.31	$H_T$	<b>0.64</b>
	$H_T$	$H_S$	$D_{ST}$	$G_{ST}$			
Locus 1	0.50	0.23	0.27	0.54			
Locus 2	0.64	0.31	0.33	0.52			
Mean	<b>0.57</b>	<b>0.27</b>	<b>0.30</b>	<b>0.53</b>			

### 26.2. Genetic distance

Genetic diversity ( $H_e$ ) and genetic identity ( $J$  or  $H_o$ ) are also used to estimate the genetic distance within and between populations, since two populations with high identity in their genes are closer than two with high diversity. If  $J_x = \sum p_{xi}^2$  is the probability of identity in population  $x$  with  $p_{xi}$  the frequency of the  $i^{\text{th}}$  allele and  $J_y =$

## D 7.1 Production of materials for improved genotyping training

$\Sigma p_{yi}^2$  is the probability of identity in population y, the probability of identity in both populations is  $J_{xy} = \Sigma p_{xi}p_{yi}$  as described by Nei (Nei 1972; Nei and Roychoudhury, 1974). The probability of identity in population x for all normalized loci is  $I = J_{xy} / \sqrt{(J_x J_y)}$  and in turn the genetic distance is  $D = -\ln I = -\ln (J_{xy} / \sqrt{(J_x J_y)})$ . In a small sample set with many loci, any bias can be corrected using  $\check{D} = -\ln G_{xy} / \sqrt{(G_x G_y)}$ , where  $G_x$  and  $G_y$  are  $(2n_x J_x - 1) / (2n_x - 1)$  and  $(2n_y J_y - 1) / (2n_y - 1)$  over the  $I$  loci studied, respectively, and  $G_{xy} = J_{xy}$  (Nei 1987). In this case  $\check{D}$  could be negative, due to sampling errors, and hence considered as zero.

Various software packages can be used to calculate the above-mentioned parameters; they often use different parameters and have their own advantages and disadvantages. In general, for analyses of genetic diversity, characteristics required in statistical software are: i) precision (no bugs), accuracy and reproducibility, ii) user friendliness (e.g., no need for command line scripts), iii) clear output in terms of graphical options and iv) that it is open access. Here are a list of compared some software packages that run under Microsoft Windows, which are generally used to calculate population genetic analyses. The software packages are:

- GenAIEx (Peakall and Smouse 2012), <http://biology-assets.anu.edu.au/GenAIEx/Welcome.html>
- GDA (Lewis and Zaykin 2012), <http://en.bio-soft.net/dna/gda.html>
- Popogene (Yeh et al. 1999), <https://sites.ualberta.ca/~fyeh/popogene.html>
- Power Marker (Liu and Muse 2005), <http://statgen.ncsu.edu/powermarker/index.html>
- Cervus (Kalinowski et al. 2007), [www.fieldgenetics.com](http://www.fieldgenetics.com)
- Arlequin (Excoffier et al. 2005), <http://cmpg.unibe.ch/software/arlequin3/>
- Structure v 2.3 (Pritchard et al. 2000), <http://web.stanford.edu/group/pritchardlab/structure.html>

Software description and comparison is carried out using examples of data obtained with SSR markers (hence co-dominant) on nine durum wheat populations from three Ethiopian regions as described by Mondini et al. (2010). For this example, the analyses of 10 genotypes per population are reported.

### 26.3. Data Input

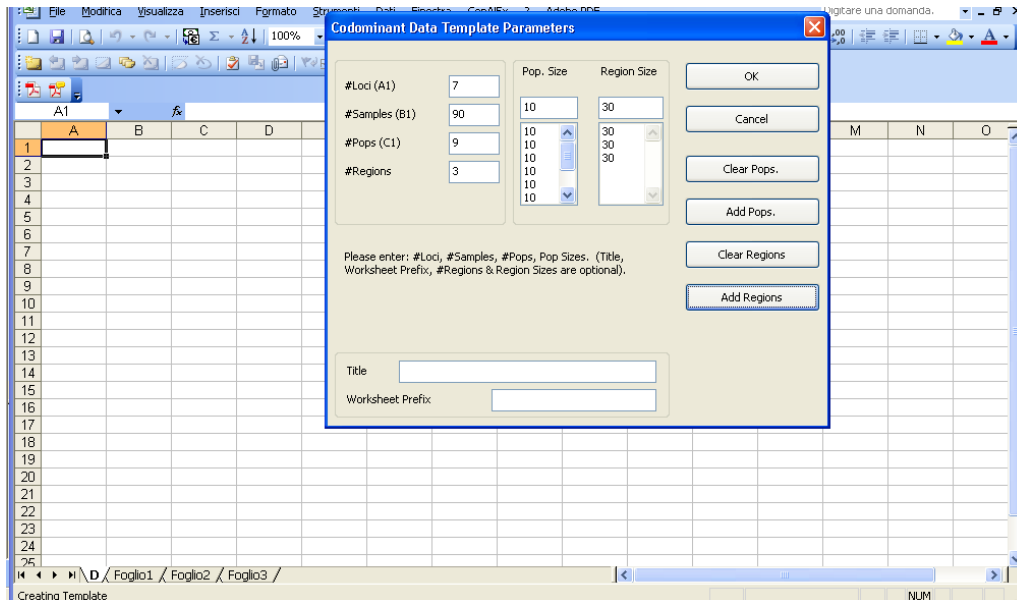
One of the first issues is the data format required as the various software packages use different data-file formats. Small discrepancies such as a single comma or space can make the data unreadable or misclassified. As a result, it often takes more time to organise the data into the correct format than to run the analysis itself. Some programs may offer the possibility of importing/exporting data from/to other formats, thereby avoiding reformatting data manually and making it easier and faster to analyse a given data set with different programs. This is of particular importance where the data set may require the use of more than one application and/or analyses offered by different software packages.

The amplicons generated from markers are distinguished after electrophoresis runs in submarine gel or in capillary by sequencer; in the later cases the results, as alleles call, can be exported from sequencer into a Microsoft Excel file. Excel seems to be the easier and universal way to insert data. As such, GenAIEx (Peakall and Smouse 2012) which is an Excel macro rather than a full software package, is first considered. GenAIEx software version 6.5 has a template function for co-dominant,

## D 7.1 Production of materials for improved genotyping training

binary and haploid data, creating a framework on which data insertion can be easily carried out starting from cell C4. After the data are inserted, they can be analysed directly by GenAIEx or alternatively be exported to other formats specific to other commonly used statistical software. The present example entails seven loci, 90 samples, nine populations, and three regions as indicated in the template (Fig. 26.2.a).

(a)



(b)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	7	90	9	10	10	10	10	10	10	10	10	10	3	30	30	30				
2	Codominantdatatem	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9	Region			Region2	Region3					
3	Sample Pop	WMC24	BARC213	BARC8	wms124	WMC177	WMC170	CFA2278												
4	0	7441	0	0	214	214	286	286	213	213	202	202	202	202	150	150				
5	2	7441	161	176	0	0	278	278	213	213	202	202	0	0	150	150				
6	3	7441	176	176	210	210	286	286	215	215	240	240	240	240	150	150				
7	4	7441	161	161	214	214	286	286	213	213	202	202	202	202	150	150				
8	5	7441	161	161	216	216	286	286	215	215	242	242	202	202	150	150				
9	6	7441	161	161	216	216	286	286	215	215	242	242	202	202	150	150				
10	7	7441	161	161	214	214	286	286	215	215	242	242	202	202	150	150				
11	8	7441	0	0	210	210	278	278	211	211	206	206	206	206	147	147				
12	9	7441	161	161	0	0	274	292	211	211	202	202	206	240	150	150				
13	10	7441	0	0	218	220	286	286	0	0	240	240	206	206	150	150				
14	11	7755	161	161	214	214	278	278	213	213	216	242	208	208	150	150				
15	12	7755	176	176	216	216	278	278	213	213	242	242	204	208	150	150				
16	13	7755	161	161	222	222	278	278	213	213	242	242	208	208	150	150				
17	14	7755	0	0	208	208	278	278	213	213	242	242	202	202	147	147				
18	15	7755	161	161	208	208	278	278	215	215	242	242	206	206	147	147				
19	16	7755	176	176	212	222	278	278	211	211	242	242	202	202	147	147				
20	17	7755	176	176	212	212	0	0	215	215	242	242	202	202	147	147				
21	18	7755	161	161	212	220	290	290	213	213	242	242	204	204	150	150				
22	19	7755	161	161	212	212	272	286	213	213	242	242	206	206	150	150				
23	20	7755	161	161	212	212	278	278	213	213	242	242	202	202	150	150				
24	21	7761	173	173	218	218	278	278	213	213	242	242	204	204	150	150				
25	22	7761	176	176	216	216	278	278	213	213	242	242	202	206	150	150				
26	23	7761	176	176	218	218	278	278	213	213	242	242	206	206	150	150				
27	24	7761	161	176	218	218	278	278	213	213	242	242	206	206	150	150				
28	25	7761	165	176	220	220	278	278	213	213	242	242	206	206	150	150				
29	26	7761	176	176	0	0	278	278	213	213	206	206	240	240	150	150				

Fig. 26.2. Structure of the data inserted by GenAIEx, the Excel macro for genetic analyses. a) template, b) data in D sheet.

The results are stored in an Excel sheet where the loci and the populations are indicated with consecutive numbering; it is possible, however, to change these to the

## D 7.1 Production of materials for improved genotyping training

correct locus and population names. Being co-dominant data, each locus will have two columns for the two alleles (Fig. 26.2.b). GenAIEx can also be used to import or export data from or to other software packages, although it is very important to pay attention to the codes used by the different software to indicate missing data. For example, the alleles can be easily named with their molecular weight in bp (base pair), however the null allele (which **is not** missing data) could be named as zero, but zero is considered missing for some software such as GenAIEx when co-dominance is the option selected. In such cases, it is important to rename the null allele, for example, by substituting zero with 1.

### 26.4. Data Analysis

The same data was then analysed using various software packages and the various outputs compared and reported here.

#### 26.4.1. GenAIEx

GenAIEx is available at <http://biology-assets.anu.edu.au/GenAIEx/Welcome.html>. It is an Excel macro used for statistical analysis in genetics, so you should be registered for Office package which is not open source. By using the “Frequency...” option it is possible to compute allele frequency, heterozygosity, F-stat and polymorphism by population and by locus, some genetic distances (i.e., Nei distance, Nei unbiased distance, pairwise  $F_{ST}$ ) together with some graphic options (Fig. 26.3.).

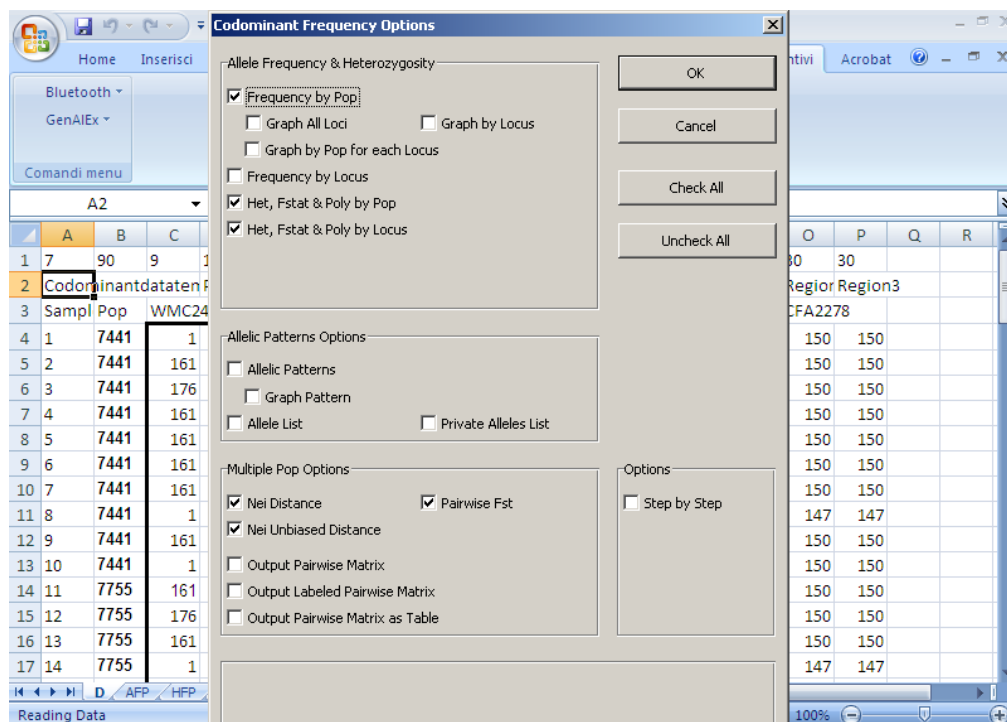


Fig. 26.3. Co-dominant frequency options of GenAIEx Excel macro.

One of the positive aspects of GenAIEx is that the different output-sheets display the base of the statistic used. There are also options for graphic (i.e. Allele Frequencies by Population with Graph over Loci or Graphs by Population and Locus) that provide a quick overview of allele distribution among populations. The most important outputs are in the sheets “HFP” and “HFL” where the different statistical



## D 7.1 Production of materials for improved genotyping training

parameters by locus (Table 26.2.) and/or by populations (Table 26.3.) are provided. The parameters are:

- N: (number of genotypes)
- Na: (No. of Different Alleles)
- Ne: (No. of Effective Alleles =  $1 / (\sum p_i^2)$ )
- I: (Shannon's Information Index =  $-1 * \sum (p_i * \ln (p_i))$ )
- Ho: (Observed Heterozygosity = No. of Hets / N)
- He: (Expected Heterozygosity =  $1 - \sum p_i^2$ )
- uHe: (Unbiased Expected Heterozygosity =  $(2N / (2N-1)) * He$ )
- F: (Fixation Index =  $(He - Ho) / He = 1 - (Ho / He)$ )
- Fis: (Mean He - Mean Ho) / Mean He)
- Fit:  $(H_T - \text{Mean Ho}) / H_T$ ,  $F_{ST} (H_T - \text{Mean He}) / H_T$ )
- Nm:  $([(1 / F_{ST}) - 1] / 4)$
- $H_T$ : Total Expected Heterozygosity =  $1 - \sum tp_i^2$ .

Where  $tp_i$  is the frequency of the  $i^{\text{th}}$  allele for the total and  $\sum tp_i^2$  is the sum of the squared total allele frequencies.

Table 26.2. GenAlEx output of the data in Fig. 26.2 per locus.

		WMC24	BARC213	BARC8	wms124	WMC177	WMC170	CFA2278	Mean	SE
<b>N</b>	<b>Mean</b>	9.333	8.667	9.889	9.556	10.000	9.667	9.889		
	<b>SE</b>	0.333	0.236	0.111	0.242	0.000	0.167	0.111		
<b>Na</b>	<b>Mean</b>	3.222	4.444	3.222	1.667	3.444	4.000	1.444		
	<b>SE</b>	0.547	0.475	0.401	0.289	0.475	0.408	0.176		
<b>Ne</b>	<b>Mean</b>	2.167	3.374	1.949	1.424	2.106	2.742	1.176		
	<b>SE</b>	0.287	0.411	0.322	0.212	0.308	0.292	0.100		
<b>I</b>	<b>Mean</b>	0.825	1.266	0.747	0.321	0.829	1.099	0.183		
	<b>SE</b>	0.154	0.131	0.145	0.140	0.158	0.132	0.081		
<b>Ho</b>	<b>Mean</b>	0.289	0.143	0.035	0.000	0.122	0.117	0.000		
	<b>SE</b>	0.084	0.032	0.017	0.000	0.057	0.029	0.000		
<b>He</b>	<b>Mean</b>	0.466	0.657	0.395	0.198	0.436	0.585	0.113		
	<b>SE</b>	0.074	0.051	0.075	0.087	0.079	0.065	0.054		
<b>uHe</b>	<b>Mean</b>	0.493	0.698	0.416	0.209	0.459	0.617	0.119		
	<b>SE</b>	0.078	0.054	0.079	0.092	0.084	0.068	0.057		
<b>F</b>	<b>Mean</b>	0.426	0.803	0.887	1.000	0.693	0.726	1.000		
	<b>SE</b>	0.119	0.045	0.054	0.000	0.130	0.106	0.000		
<b>Pops</b>	<b>F<sub>IS</sub></b>	0.381	0.783	0.913	1.000	0.720	0.799	1.000	0.799	0.081
	<b>F<sub>IT</sub></b>	0.566	0.838	0.954	1.000	0.767	0.853	1.000	0.854	0.058
	<b>F<sub>ST</sub></b>	0.300	0.253	0.471	0.308	0.167	0.269	0.210	0.282	0.037
	<b>Nm</b>	0.584	0.739	0.281	0.562	1.246	0.680	0.941	0.719	0.116

## D 7.1 Production of materials for improved genotyping training

Table 26.3. GenAlEx output of the data in Fig. 26.2 per population.

<b>Mean and SE over Loci for each Pop</b>		<b>N</b>	<b>Na</b>	<b>Ne</b>	<b>I</b>	<b>Ho</b>	<b>He</b>	<b>uHe</b>	<b>F</b>
Pop1	Mean	9.000	3.286	2.400	0.921	0.068	0.520	0.551	0.857
	SE	0.436	0.421	0.348	0.147	0.025	0.077	0.081	0.059
Pop2	Mean	9.714	3.286	2.268	0.853	0.073	0.475	0.501	0.750
	SE	0.184	0.565	0.412	0.174	0.029	0.083	0.088	0.142
Pop3	Mean	9.857	2.286	1.535	0.450	0.057	0.249	0.262	0.832
	SE	0.143	0.522	0.254	0.186	0.043	0.103	0.108	0.090
Pop4	Mean	9.571	2.714	1.697	0.622	0.089	0.347	0.367	0.783
	SE	0.297	0.360	0.246	0.138	0.041	0.074	0.079	0.106
Pop5	Mean	9.714	4.286	2.934	1.088	0.221	0.541	0.571	0.635
	SE	0.184	0.778	0.509	0.245	0.097	0.119	0.125	0.143
Pop6	Mean	9.714	3.571	2.461	0.900	0.164	0.477	0.504	0.694
	SE	0.286	0.719	0.550	0.217	0.096	0.101	0.107	0.161
Pop7	Mean	9.286	2.429	1.733	0.539	0.122	0.303	0.320	0.547
	SE	0.286	0.528	0.358	0.195	0.068	0.105	0.111	0.171
Pop8	Mean	9.429	2.857	1.932	0.687	0.066	0.370	0.392	0.840
	SE	0.297	0.595	0.359	0.209	0.036	0.108	0.114	0.064
Pop9	Mean	9.857	2.857	2.245	0.716	0.046	0.383	0.404	0.886
	SE	0.143	0.705	0.481	0.262	0.033	0.138	0.145	0.056
<b>Grand Mean and SE over Loci and Pops</b>									
		<b>N</b>	<b>Na</b>	<b>Ne</b>	<b>I</b>	<b>Ho</b>	<b>He</b>	<b>uHe</b>	<b>F</b>
Total	Mean	9.571	3.063	2.134	0.753	0.101	0.407	0.430	0.755
	SE	0.090	0.198	0.136	0.067	0.019	0.034	0.036	0.039
Population	<b>%P</b>								
Pop1	100.00%								
Pop2	100.00%								
Pop3	57.14%								
Pop4	100.00%								
Pop5	85.71%								
Pop6	85.71%								
Pop7	71.43%								
Pop8	71.43%								
Pop9	57.14%								
Mean	80.95%								
SE	5.83%								

## D 7.1 Production of materials for improved genotyping training

The three levels of fixation indexes ( $F_{IS}$ ,  $F_{IT}$ ,  $F_{ST}$ ) are computed per locus and not per population as in other programs such as Arlequin (see below). The output of different genetic distance such as Nei's distance, Nei's unbiased distance, pairwise  $F_{ST}$  are reported in Table 26.4.

Table 26.4. Computation of different parameters of distance between populations. Sheets NeiP, uNeiP and  $F_{ST}P$ . A) Nei's genetic distance (Nei 1972); B) Pairwise Population Matrix of Nei's Unbiased Genetic Distance; C) Pairwise Population  $F_{ST}$  Values.

A)

	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
Pop2	0.406	0.000							
Pop3	0.569	0.234	0.000						
Pop4	0.602	0.224	0.032	0.000					
Pop5	0.615	0.401	0.236	0.222	0.000				
Pop6	0.513	0.250	0.120	0.127	0.249	0.000			
Pop7	0.947	0.619	0.598	0.577	0.445	0.495	0.000		
Pop8	0.624	0.540	0.398	0.376	0.163	0.416	0.579	0.000	
Pop9	0.392	0.386	0.336	0.290	0.237	0.374	0.619	0.251	0.000

B)

	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
Pop2	0.347	0.000							
Pop3	0.527	0.200	0.000						
Pop4	0.553	0.183	0.008	0.000					
Pop5	0.548	0.342	0.194	0.173	0.000				
Pop6	0.453	0.199	0.085	0.085	0.189	0.000			
Pop7	0.901	0.582	0.577	0.549	0.399	0.456	0.000		
Pop8	0.573	0.497	0.372	0.343	0.112	0.373	0.550	0.000	
Pop9	0.341	0.344	0.310	0.258	0.186	0.330	0.589	0.217	0.000

C)

	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
Pop2	0.142	0.000							
Pop3	0.248	0.140	0.000						
Pop4	0.214	0.105	0.044	0.000					
Pop5	0.162	0.139	0.141	0.096	0.000				
Pop6	0.157	0.108	0.105	0.076	0.104	0.000			
Pop7	0.279	0.246	0.364	0.241	0.181	0.234	0.000		
Pop8	0.215	0.211	0.257	0.163	0.080	0.187	0.290	0.000	
Pop9	0.179	0.197	0.234	0.143	0.110	0.190	0.308	0.152	0.000

## D 7.1 Production of materials for improved genotyping training

GenAlEx can calculate the Molecular Analysis of Variance AMOVA which partitions genetic variability into different components (Table 26.5) including, or not at an individual level.

Table 26.5. AMOVA (Analyses of Molecular Variance) output of GenAlEx

Source	df	SS	MS	Est. Var.	%
Among Regions	2	27.828	13.914	0.033	2%
Among Pops	6	71.567	11.928	0.515	24%
Within Pops	171	277.550	1.623	1.623	75%
Total	179	376.944		2.171	100%

### 26.4.2. GDA

GDA can be downloaded at <http://en.bio-soft.net/dna/gda.html>. Data can be exported from GenAlEx to GDA, but it is necessary manually to change the file extension. A useful tool of GDA is the possibility of easily re-running the analysis excluding/including loci and/or populations.

The descriptive statistics offered by GDA are: (i) number of alleles per population (A), (Na in GenAlEx); (ii) polymorphic alleles per locus, (not available in GenAlEx); (iii) expected (He) and (iv) observed (Ho) heterozygosity. Observed heterozygosity is in line with GenAlEx output, while the He is here the unbiased expected heterozygosity (uHe in GenAlEx). GDA outputs per population and per locus are reported in Table 26.6. Table 26.7 shows the private alleles, another useful option present in GDA. In Table 26.8, genetic distances computed in agreement with Nei (1972; 1978) are shown; the first is the unbiased genetic distance of GenAlEx, while the second is equal to the genetic distance reported in GenAlEx.

## D 7.1 Production of materials for improved genotyping training

Table 26.6. Descriptive statistics output of GDA per population (A) and per locus (B). Where n is the number of observations, P the polymorphism, A the allele number, Ap the polymorphic alleles number, He the expected heterozygosity and Ho the observed heterozygosity.

A)

<b>Population</b>	<b>N</b>	<b>P</b>	<b>A</b>	<b>Ap</b>	<b>He</b>	<b>Ho</b>
Pop1	9.00	1.00	3.29	3.29	0.55	0.07
Pop2	9.71	1.00	3.29	3.29	0.50	0.07
Pop3	9.86	0.57	2.29	3.25	0.26	0.06
Pop4	9.57	1.00	2.71	2.71	0.37	0.09
Pop5	9.71	0.86	4.29	4.83	0.57	0.22
Pop6	9.71	0.86	3.57	4.00	0.50	0.16
Pop7	9.29	0.71	2.43	3.00	0.32	0.12
Pop8	9.43	0.71	2.86	3.60	0.39	0.07
Pop9	9.86	0.57	2.86	4.25	0.40	0.05
Mean	9.57	0.81	3.06	3.58	0.43	0.10

B)

<b>Locus</b>	<b>N</b>	<b>P</b>	<b>A</b>	<b>Ap</b>	<b>He</b>	<b>Ho</b>
WMC24	84.00	1.00	8.00	8.00	0.66	0.30
BARC213	78.00	1.00	12.00	12.00	0.89	0.14
BARC8	89.00	1.00	12.00	12.00	0.75	0.03
wms124	86.00	1.00	3.00	3.00	0.29	0.00
WMC177	90.00	1.00	10.00	10.00	0.53	0.12
WMC170	87.00	1.00	11.00	11.00	0.80	0.11
CFA2278	89.00	1.00	2.00	2.00	0.15	0.00
All	86.14	1.00	8.29	8.29	0.58	0.10

## D 7.1 Production of materials for improved genotyping training

Table 26.7. Private alleles (i.e., allele present in a single population).

Locus	Allele	Frequency	Found in
WMC24	171	0.050	Pop5
WMC24	153	0.050	Pop5
WMC24	169	0.150	Pop5
BARC213	204	0.200	Pop6
BARC213	224	0.050	Pop4
BARC8	248	0.100	Pop8
BARC8	242	0.100	Pop7
BARC8	272	0.050	Pop2
BARC8	274	0.050	Pop1
WMC177	246	0.300	Pop9
WMC177	212	0.100	Pop8
WMC177	204	0.100	Pop7
WMC177	220	0.150	Pop5
WMC177	222	0.050	Pop5
WMC170	214	0.100	Pop8
WMC170	220	0.100	Pop8
WMC170	248	0.050	Pop6
WMC170	230	0.050	Pop4

Table 26.8. Genetic distances computed by GDA. Above the diagonal Nei (1978) distance; below the diagonal Nei (1972) distance.

Pop1		0.35	0.53	0.55	0.55	0.45	0.90	0.57	0.34
Pop2	0.41		0.20	0.18	0.34	0.20	0.58	0.50	0.34
Pop3	0.57	0.23		0.01	0.19	0.09	0.58	0.37	0.31
Pop4	0.60	0.22	0.03		0.17	0.09	0.55	0.34	0.26
Pop5	0.62	0.40	0.24	0.22		0.19	0.40	0.11	0.19
Pop6	0.51	0.25	0.12	0.13	0.25		0.46	0.37	0.33
Pop7	0.95	0.62	0.60	0.58	0.45	0.49		0.55	0.59
Pop8	0.62	0.54	0.40	0.38	0.16	0.42	0.58		0.22
Pop9	0.39	0.39	0.34	0.29	0.24	0.37	0.62	0.25	

Based on Nei's genetic distance computed in Table 26.6, GDA builds up a dendrogram with UPGMA methodology (Fig. 26.4.).



## D 7.1 Production of materials for improved genotyping training

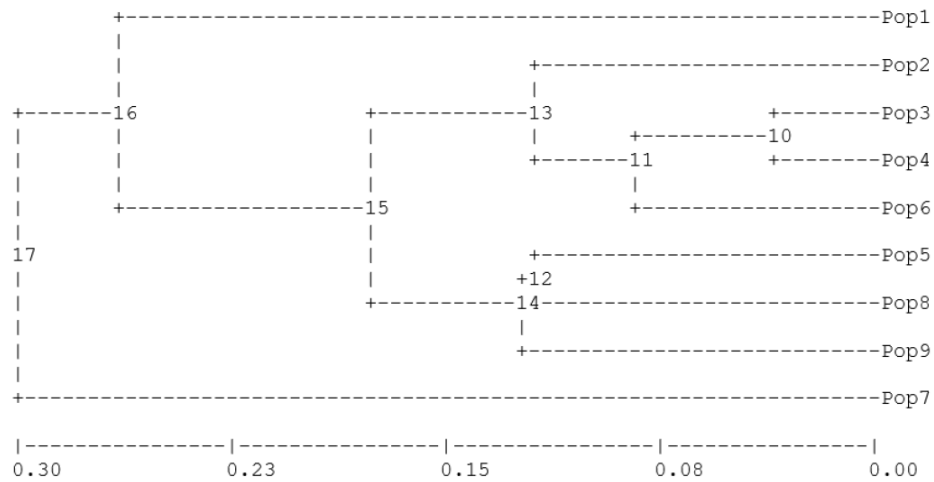


Fig. 26.4. UPGMA dendrogram based on Nei genetic distance.

The graphic output is as a text file. To improve options for the quality of graphs, it is necessary to use other software such as TreeView (Page 2001). The graphic quality and options are not considered here as it is the ability of the statistical software to export the dendrogram codes to be then used in the graphical software that is of prime importance.

### 27.4.3. Popgene

Popgene offers two versions for either 32 or 16 bit Windows operating environments, and can be downloaded at <https://sites.ualberta.ca/~fyeh/popgene.html>. It immediately divided the analysis depending on whether it deals with dominant or co-dominant markers. For diploid data it performs a genotypic frequency, HW test (not commonly found in other packages), Fixation index, Allele frequency, Allele number, Effective allele number, Polymorphic loci, Observed and Expected Homozygosity and Heterozygosity, Shannon index, Homogeneity test, F-statistics ( $F_{IT}$ ,  $F_{ST}$ ,  $F_{IS}$ ), Gene flow, Genetic distance (following Nei (1972; 1978)). It also produces a dendrogram using UPGMA of the Nei's distance, Neutrality test, Linkage disequilibrium (LD) between two loci. In the cases of several alleles per locus, the required input is not straightforward, based on the Mendelian convention (Fig. 26.5), i.e., providing a letter for each allele, but it is possible to export the Popgene format from GenAlEx. However, a significant disadvantage is that it assigns the same letter to alleles from different loci, as if they were the same allele. This creates confusion and errors especially when reading the tables of "Allele Frequency".

## D 7.1 Production of materials for improved genotyping training

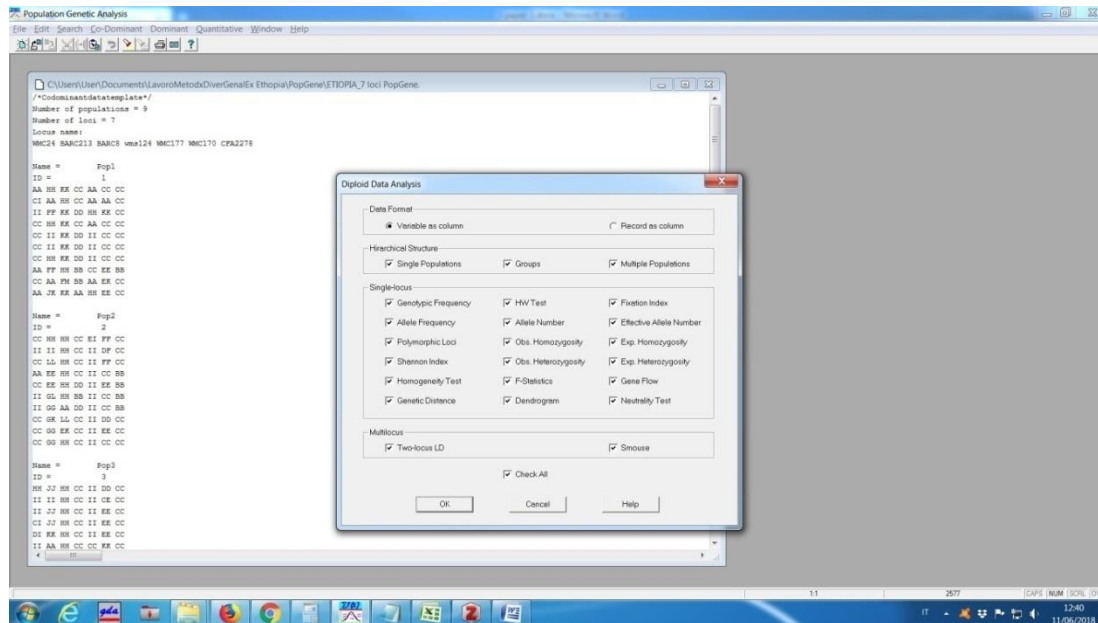


Fig. 26.5. Popgene input file and analyses options.

### 26.4.4. Power Marker

Power Marker, like GDA, was developed at the North Carolina State University and uses as reference Genetic Data Analysis by Weir (1990). The original download source for Power Markers, <http://www.powermarker.net/> (Liu and Muse 2005), seems to have expired, however the program and the manual can be found at <http://statgen.ncsu.edu/powermarker/index.html>.

Data input is very easy, entering the allelic phase separated by space, tab, and/or commas. It is possible to indicate up to three category levels. In this example, we used: genotype, population, and region. The program is suitable for microsatellite data; however, it also works with haplotypes. The data can be reduced by sub-selection of genotypes or markers on the basis of particular parameters such as the level of missing data, heterozygosity, or diversity. Outputs have their own format which can be easily converted into Excel files. A very useful tool is the internal link with TreeView (Page 1996) graphic program (<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>) used to display genotype relationships (Trees) with good graphical resolution. However, in order to use this function, the user must also install the TreeView program.

The summary table (Fig. 26.6.) illustrates information such as: (i) allele frequency, (ii) genotype number, (iii) number of observations, (iv) allele, (v) gene diversity, (vi) heterozygosity and (vii) PIC. Number of observations, allele, gene diversity, heterozygosity, and PIC are equivalent to the values reported in the GDA respectively as  $n$ ,  $A$ ,  $H_e$ , and  $H_o$ . In Power Marker the expected heterozygosity (which is not unbiased expected heterozygosity as in GDA) is named "gene diversity". It should be noted that the PIC values are here computed according to Botstein et al. (1980). The main disadvantage of Power Marker is that outputs always refer to the markers rather than to populations as per GDA. To show values per population, it is necessary to create a subset of data where only one population is considered each time. Another disadvantage is that the output does not report the options chosen, so naming the folders with self-explaining labels is imperative.

## D 7.1 Production of materials for improved genotyping training

Marker	Major Allele Freq	GenotypeNo	SampleSize	No. of obs.	AlleleNo	Availability	GeneDiversity	Heterozygosity	PIC
WMC24	0.5119	14.0000	90.0000	84.0000	8.0000	0.9333	0.6532	0.2976	0.6151
BARC213	0.1987	18.0000	90.0000	78.0000	12.0000	0.8667	0.8696	0.1410	0.8683
BARC8	0.4438	13.0000	90.0000	89.0000	12.0000	0.9889	0.7415	0.0337	0.7267
wms124	0.8372	3.0000	90.0000	86.0000	3.0000	0.9556	0.2801	0.0000	0.2614
WMC177	0.6778	11.0000	90.0000	90.0000	10.0000	1.0000	0.5185	0.1222	0.5078
WMC170	0.3161	16.0000	90.0000	87.0000	11.0000	0.9667	0.7906	0.1149	0.7721
CFA2278	0.9213	2.0000	90.0000	89.0000	2.0000	0.9889	0.1433	0.0000	0.1344
Mean	0.5581	11.0000	90.0000	86.1429	8.2857	0.9571	0.5710	0.1014	0.5551

Fig. 27.6. Power Marker output for the genetic data.

After the user has computed allele frequency by using the “phylogeny” option, it is possible to calculate frequency based on utilising several methods. The only equivalent method to the other software in this paper is Nei’s genetic distance (Nei 1972). In addition, Power Maker can compute Pairwise Linkage Disequilibrium, where the output is displayed for each marker in the order, they were inserted in the data file (Fig. 26.7.). Therefore, it is crucial that the marker results be entered in the “right” order, which is important only if a genetic map with marker positions along the chromosomes is available.

Row	Marker1	Marker2	Mutual Infor	Multi-Allelic	Multi-Allelic	ChiSquare df	ChiSquare val	ChiSquare p-
1	GPW2140	GPW2283	0.2109	0.0169	0.3918	168	444.9015	0.0000
2	GPW2283	GPW1010	0.2286	0.0107	0.4720	546	815.8862	0.0000
3	GPW1010	BARC106	0.2347	0.0148	0.4803	390	976.5542	0.0000
4	BARC106	GPW2138	0.2353	0.0555	0.5058	210	917.2628	0.0000
5	GPW2138	GPW2279-1	0.1908	0.0164	0.3553	98	387.1889	0.0000
6	GPW2279-1	GPW2279-2	0.3364	0.0161	0.4759	48	68.1617	0.0293
7	GPW2279-2	WMC513-1	0.4549	0.0107	0.6010	170	398.2509	0.0000
8	WMC513-1	BARC170-1	0.2659	0.0123	0.4470	90	149.9942	0.0001
9	BARC170-1	BARC170-2	0.5442	0.0430	0.9756	45	367.8286	0.0000
10	BARC170-2	WMC468	0.2123	0.0375	0.5057	45	151.4449	0.0000
11	WMC468	CFD257	0.0782	0.0080	0.1414	30	80.4793	0.0000
12	CFD257	GPW2244	0.0841	0.0120	0.2661	45	133.4443	0.0000
13	GPW2244	GPW2228	0.2052	0.0234	0.3693	144	857.2807	0.0000
14	GPW2228	WMS637-2	0.0634	0.0264	0.4397	28	52.8339	0.0031
15	WMS637-2	WMC161	0.0328	0.0128	0.2895	14	26.7485	0.0208
16	WMC161	BARC343	0.2151	0.0279	0.4982	165	886.8320	0.0000
17	BARC343	WMC262	0.2412	0.0195	0.5451	154	522.4451	0.0000
18	WMC262	CFD88	0.1261	0.0104	0.3000	126	680.6362	0.0000
19	CFD88	WMS160	0.2323	0.0318	0.5468	243	647.6450	0.0000
20	WMS160	BARC78-1	0.1306	0.0088	0.3082	207	211.5212	0.4000
21	BARC78-1	GPW356-1	0.1391	0.0065	0.2267	117	149.4079	0.0232
22	GPW356-1	BARC52-2	0.1381	0.0073	0.2955	98	168.6862	0.0000
23	BARC52-2	BARC184	0.3161	0.0327	0.3849	14	272.1096	0.0000
24	BARC184	WMC219	0.0985	0.0179	0.3351	45	69.5696	0.0108
25	WMC219	WMS269-1	0.3318	0.0082	0.5035	216	616.3428	0.0000
26	WMS269-1	WMS269-2	0.3081	0.0255	1.0000	5	22.0000	0.0005
27	WMS269-2	WMC513-2	0.3069	0.0504	0.7852	16	76.2367	0.0000
28	WMC513-2	BARC327	0.0976	0.0438	0.4311	14	24.3119	0.0420
29	BARC327	BARC153	0.0935	0.0060	0.3748	22	64.1180	0.0000
30	BARC153	WMS637-1	0.1665	0.0265	0.7788	7	22.7060	0.0019
31	WMS637-1	BARC78-2	0.4629	0.0390	0.6390	28	98.2248	0.0000
32	BARC78-2	BARC52-1	0.3009	0.0334	0.4842	15	17.7955	0.2736
33	BARC52-1	GPW356-2	Non un num	Non un num	Non un num	0	0.0000	Non un num

Fig. 26.7. Power Marker output for the Pairwise Linkage Disequilibrium.

### 26.4.5. Cervus

Cervus is primarily designed for assignment of parents to their offspring using genetic markers. Nevertheless, it is sometimes used for genetic analysis. It is available for download at [www.fieldgenetics.com](http://www.fieldgenetics.com). The input data sheet is not as user-friendly as some of the other programs, but this can be converted from GenePop which in turn can be converted from GenAIEx.

It calculates PIC value as per Botstein et al. (1980) and He is unbiased. In crossed populations, Cervus computes the average non-exclusion probability for a series of related genotypes such as first and second parent, parent pair, identity, and sib

## D 7.1 Production of materials for improved genotyping training

identity (Table 26.9). Moreover, it also tests Hardy-Weinberg equilibrium. The program is particularly useful for animal population genetics.

Table 26.9. Cervus output reporting the number of alleles per locus (k), number of individuals (N), observed (Hobs) and expected (Hexp) heterozygosity, PIC, Combined non-exclusion probability for first parent (NE1P), second parent (NE2P), parent pair (NE-PP), identity (NE-I) and sib identity (NE-SI), the Hardy-Weinberg equilibrium significance (HW), and the F test (F).

Locus	k	N	HObs	HExp	PIC	NE-1P	NE-2P	NE-PP	NE-I	NE-SI	HW	F(Null)
WMC24	8	84	0.298	0.663	0.615	0.748	0.575	0.387	0.160	0.460	***	+0.3786
BARC213	12	78	0.141	0.886	0.868	0.391	0.242	0.090	0.026	0.317	ND	+0.7244
BARC8	12	89	0.034	0.754	0.727	0.620	0.435	0.230	0.086	0.397	***	+0.9145
wms124	3	86	0.000	0.285	0.261	0.960	0.858	0.754	0.536	0.742	ND	+0.9766
WMC177	10	90	0.122	0.527	0.508	0.835	0.655	0.448	0.243	0.549	***	+0.6174
WMC170	11	87	0.115	0.804	0.772	0.565	0.388	0.204	0.067	0.367	***	+0.7509
CFA2278	2	89	0.000	0.146	0.134	0.989	0.933	0.879	0.742	0.863	ND	+0.8551
Mean	8.29			0.580	0.555	0.081	0.012	0.000	0.000	0.007		

### 26.4.6. Arlequin

Arlequin, available at <http://cmpg.unibe.ch/software/arlequin3/>, produces output displayed in a browser page, and thus is not ideal for conversion into a word document. On the other hand, the particular computation run by Arlequin is AMOVA (Analyses of Molecular Variance) as described by Excoffier et al. (1992). It considers haplotype, and with 90 genotypes, the total degree of freedom is 179 [(90\*2)-1] = 2N-1 (Table 26.10). The AMOVA output is very similar to the GenAIEx one (Table 26.5.).

Table 26.10. AMOVA (Analyses of Molecular Variance) output of Arlequin.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Expected mean square
Among region	2 (R-1)	27.828	0.03310 $V_a$	1.52	$N\sigma_a^2 + 2\sigma_b^2 + \sigma_c^2$
Among populations within region	6 (P-R)	71.567	0.51523 $V_b$	23.73	$2\sigma_b^2 + \sigma_c^2$
Within populations	171 (2N-P)	277.550	1.62310 $V_c$	74.75	$\sigma_c^2$
Total	179 (2N-1)	376.944	2.17144		$\sigma_T^2$

Where:

$$\sigma_a^2 = F_{ct} \sigma_T^2, \sigma_b^2 = (F_{ST} - F_{CT}) \sigma_T^2, \sigma_c^2 = (1 - F_{st}) \sigma_T^2,$$

$$F_{ST} = (\sigma_a^2 + \sigma_b^2) / \sigma_T^2, F_{SC} = \sigma_b^2 / (\sigma_b^2 + \sigma_c^2), F_{CT} = \sigma_a^2 / \sigma_T^2,$$

$$F_{ST} = 0.252 = F_{IT}$$

$$F_{SC} = 0.240 = F_{IS}$$

$$F_{CT} = 0.015 = F_{ST}$$

## D 7.1 Production of materials for improved genotyping training

He and Ho are reported for each locus within each population and produce the same average outcome as the GDA software. Linkage Disequilibrium, where the deviation from random association between alleles at different loci (Lewontin and Kojima 1960), expressed as  $D = p_{ij} - p_i p_j$ , is a potentially useful additional feature of Arlequin. However, although the instruction manual asserts that computation of the linkage disequilibrium coefficient (D) is possible, this seems not to be true. On the contrary, significance is reported as the P values of  $\chi^2$  with 1000 permutations. Moreover, the number of loci linked to each locus for each population analysed is provided. Unfortunately, even when the locus name is inserted, it is not reflected in the output, where the loci are simply numbered starting at zero. Similarly, the populations are numbered as pop1#, pop2#, pop3#, etc. rather than using the given name. This could easily lead to mistakes and confusion. In addition, there are sometimes discrepancies between the data saved in the browser output file and that saved as an xls file.

### 26.4.7. Structure

Structure software (Pritchard et al. 2000) is available to download at <http://web.stanford.edu/group/pritchardlab/structure.html>. Preparation of the data file in order to run Structure presents some problems. Conversion from GenAIEx is not straightforward since (i) an extra space is required at the end of the second row to allow the program to read the last number and (ii) population name is not converted automatically. Moreover, particular care must be taken when dealing with missing data and their code, for doing so differs from other software packages (in Structure “-9” is used as default, but it is possible to set it differently). However, with suitable modification, it is easy to convert files directly from Excel by saving it as text file.

In Structure, the analysis should be set in agreement with the population information and the procedures used in population sampling. Useful information to assist clustering includes three possible options: (i) considering individuals with or without common ancestry, (ii) with or without using sampling locations and (iii) to set the allele frequencies as either independent or dependent in each population.

Fundamentally, Structure performs a K-mean cluster analyses. As with all K-mean cluster methods, in Structure the analysis should be performed trying different values for the number of clusters (K). Clearly, some logical pre-cluster division can be argued in agreement with the data typologies, number of populations, regions, groups, etc. Nevertheless, several runs with different K values should be performed and compared. Moreover, it is sometimes useful to run single populations alone to test if they include different subpopulations. Evanno et al. (2005) has suggested the use of  $\Delta K$  to aid determination of the correct number of clusters. This should help in most situations but should not be used as an exclusive criterion. The Structure Harvester software, available on-line (Earl and vonHoldt 2012), has been developed to determine Evanno computation.

The Structure output can be displayed as a “triangle plot” in which two clusters are plotted at two vertices and all the others at the third (Fig. 26.8).

## D 7.1 Production of materials for improved genotyping training

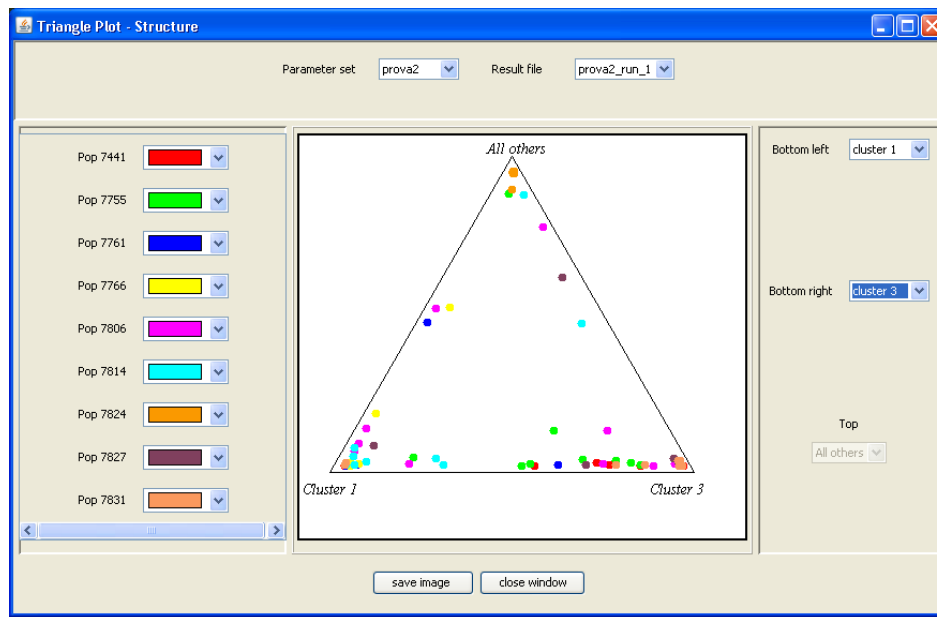


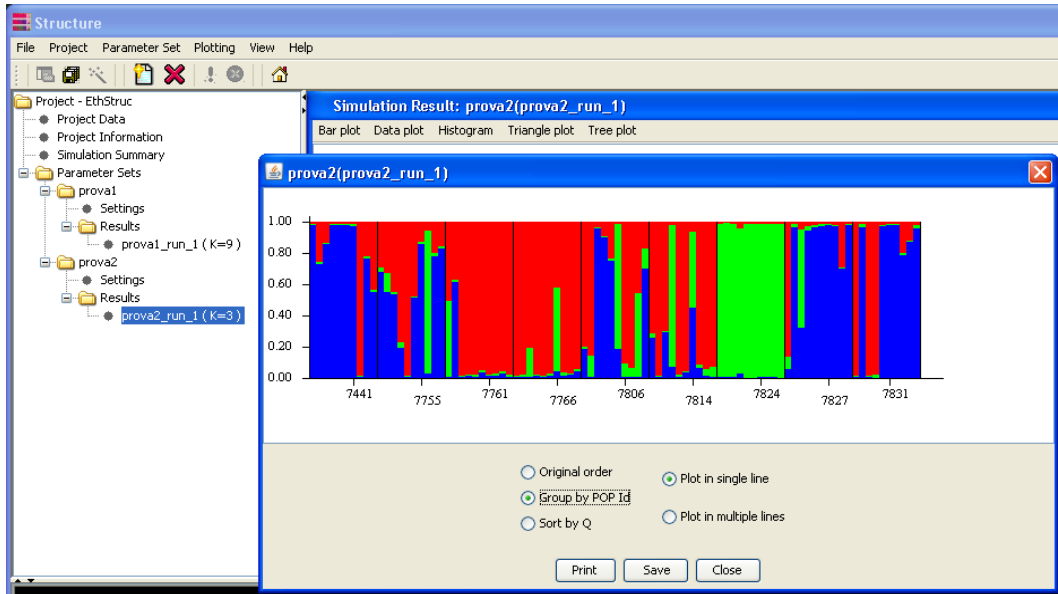
Fig. 26.8. Structure output of triangle plot with the relationships among clusters.

When more than three clusters are obtained, it can require further classification. However, a more useful, frequently used, output is the bar plot on where the clusters are shown using different colours that can then be divided to highlight populations or can be sorted by the Q value (Figs. 26.9.a and 26.9.b). The picture gives a clear idea of how the individuals are divided among clusters/populations and hence the population similarity and the collection structure. Structure can provide the histograms of  $F_{st}$ ,  $\alpha$ , and likelihood for each cluster, as well as a tree plot of distance among clusters. It is also possible to plot average proportion of the Q values directly on a geographic map (Pagnotta et al. 2017).



## D 7.1 Production of materials for improved genotyping training

A)



B)

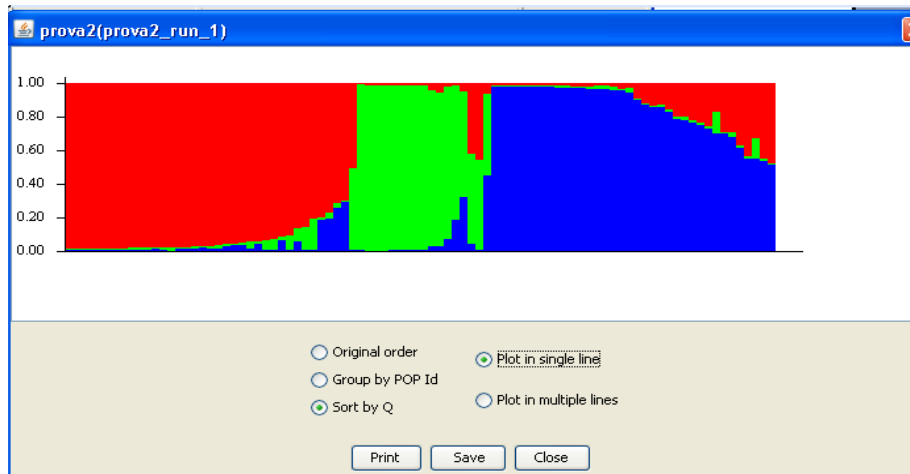


Fig. 26.9. Structure output of bar plot clusters either by population (A) or sorted by Q value (B) are reported with different colours.

### 26.5. Conclusions

The different software packages available often use different methods and tools to describe populations. Table 26.11. provides an overview of the main functions available for each program assessed in this manual. Overall, the author recommends using GenAlEx and/or Power Marker to insert data, subsequently exporting/importing and converting as required. In addition, either GDA and/or Power Marker can be used to perform most of the statistical analyses required for measuring genetic diversity, such as % of polymorphism, allele number, polymorphic allele number, expected and observed heterozygosity. In GDA these parameters refer both to the loci and to the populations, while in Power Marker several sub-sets of data should be run per population. Power Marker also computes PIC values, while GDA also computes private alleles. Both programs have different methodologies for computing population distance. Finally, GenAlEx and Arlequin are useful for determining Analyses of Molecular Variance and Structure provides a clear illustration of population clustering.

## D 7.1 Production of materials for improved genotyping training

Table 26.11. Comparison of different characteristics of most frequently used software.

Software	GenAEx	GDA	Popgene	Power Marker	Cervus	Arlequin	Structure
Insert data	Excel	Text	Text	Excel	Text	Text	Text
Descriptive statistic:							
<i>Genetic diversity</i>		X	X	X		X	
<i>Degree of polymorphism</i>	X		X	X			
<i>Heterozygosity Expected</i>	X	X	X	X	X	X	
<i>heterozygosity</i>	X	X	X	X	X	X	
<i>Number of alleles</i>	X	X	X	X	X	X	
<i>Private alleles</i>		X					
<i>Effective allele number</i>	X		X				
<i>PIC</i>				X	X		
Gene flow			X				
Homogeneity test			X				
Genetic distance	X	X	X	X		X	X
Graphic options		X	X	X			X
Fisher parameters (Fis, Fit, Fst)	X	X	X			X	
MANOVA	X					X	
LD			X	X		X	

### Online tutorials

- Index with several tutorial for GenAEx; <https://biology-assets.anu.edu.au/GenAEx/Tutorials.html>
- PopGen32 – Dendrogram demonstration video by Matheus Francisco; <https://www.youtube.com/watch?v=y4Omq1jhZBU>
- How to use the Structure software by Genomics Lab; <https://www.youtube.com/watch?v=Gl1nUSqFAfA>
- Instruction for Arlequin (in Spanish) by Genética II; <https://www.youtube.com/watch?v=3iE5LEqqIwx>

### Further reading

- Kalinowski ST, Taper ML, Marshall TC. 2007. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Mol. Eco.* 16, 1099–1106.
- Lewis PO, Zaykin D. 2012. *Genetic Data Analysis: computer program for the analysis of allelic data.* Version 1.1. Albuquerque: University of New Mexico.
- Lewontin RC., Kojima K. 1960. The evolutionary dynamics of complex polymorphisms. *Evo.* 14, 458-472.
- Liu K, Muse SV. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinf.* 21, 2128–2129.
- Mondini L, Farina A, Porceddu E, Pagnotta MA. 2010. Analysis of durum wheat germplasm adapted to different climatic conditions. *Ann. App. Bio.* 156, 211–219.

## D 7.1 Production of materials for improved genotyping training

- Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York, 512.
- Nei M. 1973. Analysis of Gene Diversity in Subdivided Populations. *PNAS* 70, 3321–3323.
- Nei M. 1972. Genetic Distance between Populations. *The Am. Nat.* 106, 283–292.
- Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89, 583–590.
- Nei M, Roychoudhury AK. 1974. Sampling variances of heterozygosity and genetic distance. *Genet.* 76, 379-390.
- Page RDM. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in Biosciences* 12, 357-358.
- Pagnotta MA. 2018. Comparison among methods and statistical software packages to analyze germplasm genetic diversity by means of codominant markers. *Multidisciplinary Sci. Journal*, 1, 197-215.
- Pagnotta MA, Fernández JA, Sonnante G, Egea-Gilabert C. 2017. Genetic diversity and accession structure in European *Cynara cardunculus* collections. *PLOS ONE* 12, e0178770.
- Peakall R, Smouse PE. 2012. GenAIEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinf.* 28, 2537–2539.
- Petit RJ, Mousadik AE, Pons O. 1998. Identifying populations for conservation on the basis of genetic markers. *Cons. Bio.* 12, 844–855.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.
- Turpeinen T, Tenhola T, Manninen O, Nevo E, Nissilä E. 2001. Microsatellite diversity associated with ecological factors in *Hordeum spontaneum* populations in Israel. *Mol. Ecol.* 10, 1577–1591.
- Weinberg W. 1908. On the demonstration of heredity in man. In: Boyer SH (1963) *Papers on human genetics*. Prentice Hall, Englewood Cliffs, NJ.
- Weir BS. 1990. *Genetic data analysis. Methods for discrete population genetic data*. Sinauer Associates, Inc. Publishers.
- Wright S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19, 395–420.
- Yeh FC., Yang RC, Boyle T, Ye ZH, Mao JX. 1999. POPGENE, version 1.32: the user friendly software for population genetic analysis. Mol Bio and Biotech Centre, Univ of Alberta, Edmonton, AB, Canada.

### 27. What is Bioinformatics and sequence management

The term **Bioinformatics** was invented by Paulien Hogeweg, a Dutch theoretical biologist and Ben Hesper in 1970 as "the study of informatics processes in biotic systems".

Since then and with the advent of new technologies, the field on bioinformatics has become more and more prominent in daily life and science. There are many definitions trying to better describe this field, but as the term itself suggests, bioinformatics is the interpretation of biological process through new computational methods. In fact, the term is often used and confused with computational biology, which it is just a part of:

***Computational Biology**, which includes many aspects of **bioinformatics**, is the science of using biological data to develop algorithms or models to understand biological systems and relationships. Until recently, biologists did not have access to very large amounts of data. This data has now become commonplace, particularly in molecular biology and genomics. Researchers were able to develop analytical methods for interpreting biological information but were unable to share them quickly among colleagues.*

People are studying bioinformatics in different ways. Some scientists are programmers devoted to developing new computational tools, both from software and hardware viewpoints, for the better handling and processing of biological data. Other scientists are more the users of the efforts of the programmers and want to better understand biological process using computational tools.

#### 27.1. Bioinformatics applications

The applications are broad, from medical to environmental science. An easy query search using a common search engine "bioinformatics applications" helps to realise the scope of this topic.

For example, it has been used for **bioremediation**: *Deinococcus radiodurans* is a bacterium with a high resistance to radiation, desiccation, and extreme environmental conditions. Thus, its genome has been completely sequenced and this information used to genetically modify other bacteria with the aim to confer the same resistance to specific conditions (Makarova et al. 2001).

In **drug discovery**, the huge amount of high-throughput data is being used to screen new drug candidates, their possible pathway and interactions, receptors and downstream consequences (Xia 2017). In the same way, **personalised medicine** is taking over, with a target therapy according to the genetic makeup of the individual, as well as **preventive medicine** (Suwinski et al. 2019).

By knowing in advance, the genome sequence of a bacterium, an effective **antibiotic can be designed**, avoiding those structures that could induce resistance (Mason et al. 2018).

Bioinformatics can also be used for **evolutionary** and phylogenetic studies with the aim to identify the origins of a particular species (Prost et al. 2019).

In the field of plant pathology, bioinformatics has been used for **viral diagnostic** and virus diversity studies (Turco et al. 2018).

## D 7.1 Production of materials for improved genotyping training

As **climate change** is becoming a major issue, we need to breed new crop varieties. Genomics can be used to adapt crops to climate change, by identifying SNPs related to a particular genetic marker that can be used for Marker Assisted Breeding (MAB), as it has been done for the *Sub1A* gene in rice, which confers resistance to submerging (the ability of a rice plant to survive and continue growing after being completely submerged in water for several days, Septiningsih et al. 2009). Genomic selection (GS) is a computational simulation to plan breeding experiments based on the crop genotypes and is already carried out in wheat and maize (Wang et al. 2018).

CRISPR/Cas can be applied to disrupt crop genes, particularly those associated with susceptibility to pests. For instance, the system was recently used to enhance blast resistance in rice by targeting the *OsERF922* gene (Wang et al. 2016). With the same approach, crop improvement has been carried out in tomato (Ruggieri et al. 2019).

If we think carefully about all these applications, it should immediately be clear that **sequencing and genotyping** are the common core: by knowing the sequences, the genome variants between isolates, individual and species the above-mentioned applications (and many more) have been possible.

### 27.2. Bioinformatic data formats

Since the classical automated Sanger sequence machines are from applied biosystem, usually the output files are in an *.abi* format which can be easily visualised by software like FinchTV or MEGA. The sequence can be exported in a **FASTA** format, which has a first mandatory text string defined by an arrow (>), followed by the ID of the sequence and starting from the second line, the sequence itself.

Instead, the output file of the NGS machines is a **FASTQ** file, which combines the information of a FASTA file (read sequence) and a QUAL file carrying the quality PHRED score, and the information is summarised in four lines. The first one is a title line, defined by “@”, for read identification and optional description. The second one is the sequence line, usually in upper case and without any tabs or spaces. The third line starts with a “+” sign, to make clear the end of the sequence line and the beginning of the quality string. The last line contains the PHRED quality information in ASCII printable character. Each character corresponds to a value derived from the formula  $Q \text{ PHRED} = -10 \log_{10} (P_e)$  which estimates the probability that the corresponding base call is incorrect (Cock et al. 2010).

The sequence fragments contained in the FASTQ files can be aligned to a Reference genome using tools like BWA (Burrow-Wheeler Alignment <https://sourceforge.net/projects/bio-bwa/>, Burrow and Wheeler 1994), giving an output file which would be in SAM or in its binary BAM format.

The Sequence/Alignment/Map (SAM) is a TAB-delimited text format containing a header section and an alignment section. The header contains the reference name and its length, as defined in the FASTA file used for mapping. The alignment sections is defined by 11 mandatory fields among which the first column is the name of the read (QNAME), followed by a FLAG that can be 4, 0 or 16 if the read unmapped or mapped in the forward or reverse strand, respectively. The third column is the reference sequence name (RNAME) and POS indicates the starting position of the alignment between the read and the reference. MAPQ and CIGAR indicate the

## D 7.1 Production of materials for improved genotyping training

quality of the mapping and the possible presence of insert/deletion using the word M (match/mismatch), I (insertion), D (deletion).

For downstream analysis, the **SAM** file can be exported by the manipulating SAM-format tool Samtools (<http://samtools.sourceforge.net/>) in a Binary Alignment/Map (**BAM**) file, compressed in BGZF format. The file contains the same information but in a binary format readable only by machines (Li et al. 2009). Samtools is also able to extract information like mapped or unmapped sequences, the consensus sequence, and the variant call format **VCF** file containing all the possible variants and SNPs which would be important for genotyping.

### 27.3. Downstream analysis

Once the samples have been sequenced, FASTQ files created and quality control passed, the reads are ready for further analysis. This of course depends on the experimental design which could be a **genome assembly** followed by **variant calling for genotyping**, their involvement in genome expression (**RNAseq**) and thus, **annotations** or the identification of different species in an environmental matrices (**Metagenomics**), the evolutionary placement in a Phylogenetic tree (**Phylogenetics**).

It is important to mention that these analyses (and many more) are made possible and easier also thanks to the huge science community willing to share their results in the genomic **databases**. In fact, since Sanger sequenced the first protein in 1955, the necessity to create databases to store biological information became urgent. Thus, the first database was created by Margaret O. Dayhoff within a short period after the Insulin protein sequence was made available in 1956: *The Atlas of Protein Sequence and Structure*. Around mid-nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structures of proteins were studied and the well-known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972.

Nowadays there are huge consortiums like NCBI, genome databases like Ensembl or databases specific for a particular species, like the TAIR project for Arabidopsis. There are annotation databases, SNPs and variants, genome expression profiles, protein sequences or structures and so on. All of them are make that beautiful world that is scientific research available to anyone interested.

#### Further reading

- Hood LE, Hunkapiller MW, Smith LM. 1987. Automated DNA sequencing and analysis of the human genome. *Genomics* 1, 201–212.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.
- Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJI. 2001. Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Mic. Mol. Biol. Rev.* 65, 44–79.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–80.



## D 7.1 Production of materials for improved genotyping training

- Mason A, Foster D, Bradley P, Golubchik T, Doumith M, Gordon NC, Peto T. 2018. Accuracy of Different Bioinformatics Methods in Detecting Antibiotic Resistance and Virulence Factors from *Staphylococcus aureus* Whole-Genome Sequences. *J. Clin. Microbiol.* 56.
- Ruggieri V, Calafiore R, Schettini C, Rigano MM, Olivieri F, Frusciante L, Barone A. 2019. Exploiting Genetic and Genomic Resources to Enhance Heat-Tolerance in Tomatoes. *Agronomy* 9, 22.
- Septiningsih EM, Pamplona AM, Sanchez DL, Neeraja CN, Vergara GV, Heuer S, Mackill DJ. 2009. Development of submergence-tolerant rice cultivars: the Sub1 locus and beyond. *Ann. Bot.* 103, 151–60.
- Turco S. 2017. siRomics for universal diagnostics of plant viral disease and virus diversity studies. Phd Thesis doi: 10.5451/unibas-006776645.
- Turco S, Golyaev V, Seguin J, Gilli C, Farinelli L, Boller T, Pooggin MM. 2018. Small RNA-Omics for Virome Reconstruction and Antiviral Defense Characterization in Mixed Infections of Cultivated Solanum Plants. 31, 707–723.
- van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681.
- Wang F, Wang C, Liu P, Lei C, Hao W, Gao Y, Zhao K. 2016. Enhanced Rice Blast Resistance by CRISPR/Cas9-Targeted Mutagenesis of the ERF Transcription Factor Gene OsERF922. *PLoS One* 11, e0154027.
- Wang X, Xu Y, Hu Z, Xu C. 2018. Genomic selection methods for crop improvement: Current status and prospects. *Crop J.* 6, 330–340.
- Xia X. 2017. Bioinformatics and Drug Discovery. *Curr. Top. Med. Chem.* 17, 1709–1726.

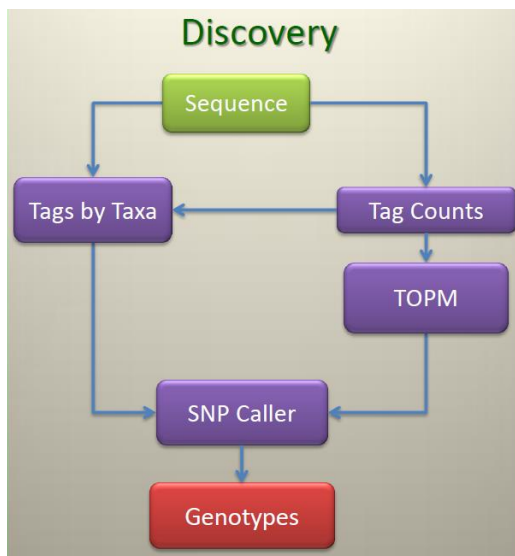
## D 7.1 Production of materials for improved genotyping training

### 28. Genotyping by Sequence (GBS) Bioinformatics Pipeline

#### 28.1. Vocabulary

- **Sequence File:** Text file containing DNA sequence and supplemental information from the Illumina Platform.
- **Key File** Text file used to assign a GBS Bar Code to Taxa
- **GBS Tag** DNA sequence consisting of a cut site remnant and additional sequence.
- **GBS Bar Code** A short known sequence of DNA used to assign a GBS Tag to its original Taxa
- **Taxa** An individual sample

#### 28.2. GBS Discovery Pipeline



#### 28.3. Sequence

Example of data:

##### 1. Raw Sequence (Qseq)

```
HWI-ST397 0 3 68 15896 200039 0 1 GTCGATTCTGCTGACTTCATGGCTTCTGTTGAI
HWI-ST397 0 3 68 15960 200043 0 1 GAGAATCAGCTTTTCCAACACCTTGAGTTTGA
HWI-ST397 0 3 68 15831 200053 0 1 ATGTA CTGCA CCGTTGCAAGCGAGCACCACCA
HWI-ST397 0 3 68 15867 200049 0 1 CCAGCTCAGCTGCATTCTTTCAAAAAC TTCC
HWI-ST397 0 3 68 15943 200048 0 1 GATTTTACTGCACATCGGTCTTGTCACACCAG
HWI-ST397 0 3 68 15812 200062 0 1 TCACCCAGCATCACGCCCTTCACATCCAGTA
HWI-ST397 0 3 68 15888 200067 0 1 CTTGACTGCCACCATGAATATGTGTTCCAAGTC
HWI-ST397 0 3 68 15969 200067 0 1 CCACA ACTGCTCCATCTTTCCATGAGACATTG
HWI-ST397 0 3 68 15786 200078 0 1 GTATTCTGCACACGAATCAGCTGAGACACCAA
HWI-ST397 0 3 68 15830 200072 0 1 AATATGCCAGCAGTTAAGAGAGTTCAAGATCC
HWI-ST397 0 3 68 15863 200073 0 1 CTCCCTGCGGGTGC GCGGACCCATCTTCAG
HWI-ST397 0 3 68 15762 200088 0 1 TGGTACGTCTGCGGAATGGCGTTTTTATGCC
HWI-ST397 0 3 68 15903 200085 0 1 GGACCTACTGCCAAGAACGGCTCACCCATC
HWI-ST397 0 3 68 15921 200082 0 1 GAGAATCAGCGTGTACGGGGCACGGGGTGAC
HWI-ST397 0 3 68 15984 200085 0 1 TTCTCCAGCCGCATGGGCCGGAGACCAGAGA
HWI-ST397 0 3 68 15788 200096 0 1 GCGT CAGCA AATGCCCAACAGCCAAGTCAG
HWI-ST397 0 3 68 15842 200099 0 1 TAGGCCATCAGCTGACTTCCCGGGTGTGGAG
HWI-ST397 0 3 68 15876 200105 0 1 GGACCTACTGCCGGCGGACGAAAGCGGTTG
HWI-ST397 0 3 68 15937 200097 0 1 CTCCTGTTGAAGCATGTGCAAAAGAGCTTGT
HWI-ST397 0 3 68 15958 200102 0 1 CGCCTATCTGCCCTCGCCGGTCATGGGGAG
```

## D 7.1 Production of materials for improved genotyping training

### 2. Key File

Flowcell	Lane	Barcode	DNASample	LibraryPlate	Row	Column	LibraryPrepID	PlateName
81PVTABXX	2	CTCC	Sample_1	1	A	1	1	Plate_A
81PVTABXX	2	TGCA	Sample_2	1	A	2	2	Plate_A
81PVTABXX	2	ACTA	Sample_3	1	A	3	3	Plate_A
81PVTABXX	2	CAGA	Sample_4	1	A	4	4	Plate_A
81PVTABXX	2	AACT	Sample_5	1	A	5	5	Plate_A
81PVTABXX	2	GCGT	Sample_6	1	A	6	6	Plate_A
81PVTABXX	2	TGCGA	Sample_7	1	A	7	7	Plate_A
81PVTABXX	2	CGAT	Sample_8	1	A	8	8	Plate_A
81PVTABXX	2	CGCTT	Sample_9	1	A	9	9	Plate_A
81PVTABXX	2	TCACC	Sample_10	1	A	10	10	Plate_A
81PVTABXX	2	CTAGC	Sample_11	1	A	11	11	Plate_A
81PVTABXX	2	ACAAA	Sample_12	1	A	12	12	Plate_A
81PVTABXX	2	TTCTC	Sample_13	1	B	1	13	Plate_A
81PVTABXX	2	AGCCC	Sample_14	1	B	2	14	Plate_A
81PVTABXX	2	GTATT	Sample_15	1	B	3	15	Plate_A
81PVTABXX	2	CTGTA	Sample_16	1	B	4	16	Plate_A
81PVTABXX	2	ACCGT	Sample_17	1	B	5	17	Plate_A
81PVTABXX	2	GTAA	Sample_18	1	B	6	18	Plate_A
81PVTABXX	2	GGTTGT	Sample_19	1	B	7	19	Plate_A

### 28.4. Tag Counts

With information from the key file, each sequence file is processed, tags are identified and counted. If a tag is shorter than 64 bases it is padded. The tags and counts are put into a tag count file for each sequence file.

*QseqToTagCountsPlugin / FastqToTagCountsPlugin*

Master Tag Counts: The individual tag count files are merged into a master tag count file. A minimum count is specified at the merge stage to exclude tags with low counts (likely sequencing errors).

*MergeMultipleTagCountsPlugin*

Conversion of Tags to Fastq: Sequence aligners do not work with the tag count file format. In preparation for the alignment step, the tag count file is converted to fastq format.

*TagCountsToFastqPlugin*

### 28.5 Tag alignment (TOPM):

The GBS pipeline uses an external aligner to do the initial alignment.

- The current version uses bowtie2 which produces the alignment in the SAM format.

## D 7.1 Production of materials for improved genotyping training

We convert the SAM file into our tags on physical map format (TOPM)

### *SAMConverterPlugin*

tag	chr	str	start	end	variant1
CAGCTCAGCGAGCACACAGCCAGCAGACCAACATCAATGGCTGTTGTGTATTTCAGTAGCACCA	10	1	85,497,730	85,497,793	* *
CAGCTCAGCGAGCTAGCCCCACTGCCAGCTTAGTTGTCGTGACAGCCATTGGGCTGAAAAAA	*	*	*	*	* *
CAGCTCAGCGAGCTGACTGACACACATGGGCTGGCTGGCTGGCCGCTGAAAAA	4	1	11,254,510	11,254,558	40 C
CAGCTCAGCGAGCTGACTGACACACATGGGCTGGCTGGCTGGCCGCTGAAAAA	4	1	11,254,510	11,254,562	40 T
CAGCTCAGCGAGCTGTTGGGCGCAGGCTCGCACCTCCGCGGCCAACCGCCCTCCCTCGACG	1	-1	182,626,947	182,626,884	* *
CAGCTCAGCGAGCTGTTGGGCGCAGGCTCGCACCTCCGCGGCCAACCGCCCTCCCTCGACG	*	*	*	*	* *
CAGCTCAGCGAGCTTCCTCCGTCCAGTCGGAGAAGGCTCCGGTGATTTGGAACTGCGCGTGTCC	3	1	160,516,139	160,516,202	* *
CAGCTCAGCGAGCTTCCTCCGTCCAGTCGGAGAAGGCTCCGGTGATTTGGAACTGCGCGTGTCT	3	1	160,516,139	160,516,202	* *
CAGCTCAGCGAGCTTCTTGTGTCGCTTATTGCGATTTCTCCCTGGCCTGCGTGTGGTAACAT	5	1	191,129,125	191,129,188	* *
CAGCTCAGCGAGCTTGGCGAGAAGGGAGCTCATCACCCCTGCTGCTGAAAAA	*	*	*	*	* *
CAGCTCAGCGAGGCAAGATCCGGACGGCGAGCCGAGGAATCCACGCAGAAAAA	*	*	*	*	* *
CAGCTCAGCGAGGCCAGATCCGGACGGCGAGCCGAGGAATCCACGCAGAAAAA	5	1	215,021,313	215,021,360	* *
CAGCTCAGCGAGTGCCGCGAACGTAAGCAAGGAGGCACCCTCAGGCCTGAAAAA	7	-1	162,758,542	162,758,493	28 C
CAGCTCAGCGAGTGCCGCGAACGTAAGCAAGGAGGCACCCTCAGGCCTGAAAAA	7	-1	162,758,542	162,758,493	28 T
CAGCTCAGCGATCGCCAGCGCCATCCTGCATCTCGCGCTCGGCGCGTCCAGTCCGGTGACCATC	6	1	88,066,897	88,066,960	* *

So Far We Have

- ✓ Identified and counted GBS tags
- ✓ Converted tag counts file to fastq
- ✓ Aligned the tags to a reference
- ✓ Converted the alignment to TOPM

### 28.6. Tags by Taxa (TBT)

In this step we identify which tags are present in which taxa using:

- Original Sequence Files
- Key File
- Master Tag Count File

Recently migrated to HDF5 file format.

- Efficient storage
- Large data sets

### *SeqToTBTHDF5Plugin*

Tags By Taxa Additional Operations: If many TBTs have been created they are merged into one TBT. Taxa that were sequenced multiple times are also merged. The TBT table is pivoted in preparation for SNP calling.

### *ModifyTBTHDF5Plugin*

### 28.7. SNP calling

Files used in SNP calling are:

–TOPM

## D 7.1 Production of materials for improved genotyping training

- TBT
- Pedigree File (optional)

### Some Key Settings

- mnF MinimumF (inbreeding coefficient)
- mnMAF Minimum Minor Allele Frequency
- mnMAC Minimum Minor Allele Count
- mnLCov Minimum Locus Coverage

### TagsToSNPByAlignmentPlugin

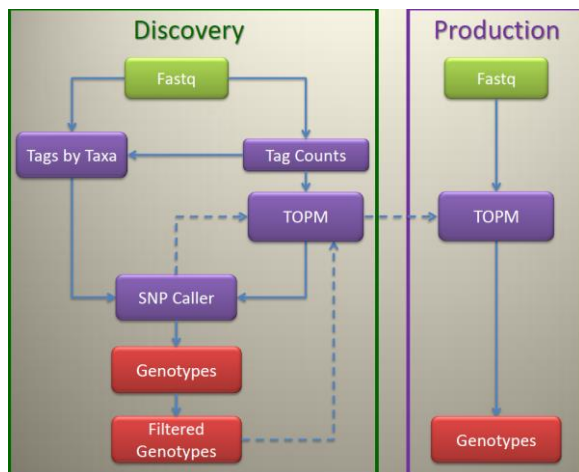
### Hap Map

rs#	alleles	chrom	pos	strand	SgSBRILO67:633Y5AAXX:2:C9	SgSBRILO19:633Y5AAXX:2:C3												
S1_2100	A/G	1	2100	+	N	N	N	N	N	N	N	R	N	A	N			
S1_2163	T/C	1	2163	+	N	N	N	N	N	N	T	C	T	T	N			
S1_13837	T/G	1	13837	+	N	N	N	N	N	N	N	G	N	N	N			T
S1_14606	C/T	1	14606	+	N	N	C	N	N	N	T	T	T	T	C			
S1_2061	T/A	1	20601	+	T	N	N	N	N	N	N	A	N	N	N			
S1_68332	C/T	1	68332	+	N	N	N	N	N	N	N	N	N	N	N			
S1_68596	A/T	1	68596	+	A	N	N	N	N	N	N	N	N	A	N			
S1_69309	G/A	1	69309	+	N	G	N	N	N	N	N	A	N	N	N			
S1_79955	T/G	1	79955	+	N	T	G	T	T	N	T	T	N	N	N			
S1_79961	T/G	1	79961	+	N	T	T	T	T	N	T	T	N	N	N			
S1_80584	G	1	80584	+	N	N	N	N	N	N	N	N	N	N	N			G
S1_80647	C/T	1	80647	+	N	N	N	N	N	N	N	C	N	N	N			C
S1_81274	T/G	1	81274	+	N	N	N	N	N	N	T	G	N	N	N			
S1_108834	G/A	1	108834	+	N	N	N	N	N	N	N	N	N	N	N			
S1_112345	T/G	1	112345	+	N	N	N	N	N	N	K	T	N	N	N			
S1_115359	C/T	1	115359	+	N	N	N	N	N	N	N	T	C	N	T			
S1_115362	T/C	1	115362	+	N	N	N	N	N	N	N	C	N	N	N			N
S1_115405	G/A	1	115405	+	G	G	A	N	N	G	G	G	G	G	N			
S1_115516	T/G	1	115516	+	N	N	T	N	N	N	T	T	N	N	N			T
S1_116694	A/G	1	116694	+	N	A	G	N	N	N	G	A	N	N	N			
S1_119016	C/T	1	119016	+	N	N	N	N	C	N	N	C	N	N	N			
S1_155366	T/C	1	155366	+	N	T	N	N	N	N	N							

## 28.8. GBS Production Pipeline

Why another pipeline?

- The last maize build (30'000 taxa) with the discovery pipeline took over 3 months.
- Most common alleles have been identified after the first few discovery builds.
- Use the information from the discovery pipeline to call SNPs in new runs quickly.
- Improve efficiency and automate.



Running the Production Pipeline:

- Required Files:

## D 7.1 Production of materials for improved genotyping training

- Sequence file (fastq or qseq)
- Key file
- Production TOPM
- •TASSEL 3 Standalone and RawReadsToHapMapPlugin
- Running the Pipeline:
  - One lane processed at a time
  - HapMap files by chromosome
  - ~7 minutes

### Testing Production Pipeline

- Compared HapMap files produced by Discovery Pipeline and Production Pipeline
- Site Comparison:
  - Discovery 48,139
  - Production 47,676
  - Difference due to maximum 8 alleles
- 99.98% correlation of genetic distance matrices

### Next Steps in Pipeline Development

- •Hierarchical Data Format – supports very large data sets and complex data structures
- Working to fuse TOPM, TBT, Keyfile, and Pedigree File into one HDF5 repository
- Continued improvements to SNP caller
- Ability to use tags not present in the reference

#### Online tutorial

- Genotyping by Sequence (GBS) Method Overview by Rob Elshire AgResearch Limited; <https://www.youtube.com/watch?v=NGqKJ0TnL9o>
- Genotyping by Sequence (GBS) Method by Genomics Lab; <https://www.youtube.com/watch?v=jbJMIjHFdJI>



### 29. Bioinformatics resources

Bioinformatics is a broad term that refers to the interdisciplinary field combining biological sciences with computer technology, with focus on utilisation of software, algorithms and statistics for data analysis and interpretation. It includes numerous databases in which large quantities of sequencing, genotyping and ontology data are stored, software tools implementing analysis algorithms that can be complex licensed point and click software combining several applications, as well as freeware solutions developed by the scientific community to tackle specific analysis.

Bioinformatics is employed in the field of gene and protein expression and regulation analysis, and includes annotation, data integration and systems biology approaches. In high-throughput transcriptomic techniques, such as microarrays, expressed cDNA sequence tag (EST) sequencing and RNA-Seq require, bioinformatics is required already in pre-processing and statistical analysis of the raw data to separate the gene-expression signal from the background noise. Similar approaches are necessary for assessment of differentially expressed proteins in the sample using protein microarrays. The use of high throughput (HT) mass spectrometry (MS) for protein expression analysis generates large amounts of mass data that need to be compared against predicted masses from protein sequence databases. Gene regulation can be further explored by sequence analysis of regulatory elements such as promoters and regulators. Gene expression and regulation analysis includes comparison of expression datasets in a wide variety of experimental conditions (changing environmental conditions, presence and absence of abiotic and biotic stress). Clustering algorithms (such as k-means clustering, hierarchical clustering) enable determination of co-expressed genes while ontology assignments enable identification of over-represented functional categories. Network analysis enables integration of the expression data into biological pathways, enabling functional interpretation of the studied processes. Integration of different functionally connected data types such as gene and protein expression datasets, metabolite information, regulatory elements information, can be performed. Systems biology enables modelling of interactions within metabolic pathways and prediction of scientific hypothesis.

In narrower term bioinformatics is used to refer to specific analysis pipelines integrating complex analysis steps that in genomics research are usually performed by different algorithms and software tools. Pipeline can be a well-documented description of steps, the software used and analysis settings, or it can employ programming to run a sequence of steps, feeding results from one step as an input to the next step in the pipeline.

#### 29.1. Bioinformatics resources

##### 29.1.1. Databases

Databases contain various types of information, such as DNA and protein sequences, molecular structures, phenotypes, metabolic pathways and ontologies. The contained data may be derived from experiments or subsequent data analysis. They may be specific to a particular organism, pathway or molecule of interest. Alternatively, they can incorporate data compiled from multiple other databases. These databases vary in their format, access mechanism, and whether they are public or not. Well known and commonly used databases include:

## D 7.1 Production of materials for improved genotyping training

- biological sequence analysis: Genbank, UniProt
- structure analysis: Protein Data Bank (PDB)
- finding Protein Families and Motif Finding: InterPro, Pfam
- Next Generation Sequencing: Sequence Read Archive
- Network Analysis: Metabolic Pathway Databases (KEGG, BioCyc)
- Interaction Analysis Databases, Functional Networks
- design of synthetic genetic circuits: GenoCAD.

### 29.1.2. Software, tools and open-source bioinformatics software

Software tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services available from various bioinformatics companies or public institutions.

Many free and open-source software tools have existed and continued to grow since the 1980s. The combination of a continued need for new algorithms for the analysis of emerging types of biological readouts, the potential for innovative *in-silico* experiments, and freely available open code bases have helped to create opportunities for all research groups to contribute to both bioinformatics and the range of open-source software available, regardless of their funding arrangements. The open source tools often act as incubators of ideas, or community-supported plug-ins in commercial applications. They may also provide *de facto* standards and shared object models for assisting with the challenge of bioinformation integration.

The range of open-source software packages includes titles such as Bioconductor, BioPerl, Biopython, BioJava, BioJS, BioRuby, Bioclipse, EMBOSS, NET Bio, Orange with its bioinformatics add-on, Apache Taverna, UGENE and GenoCAD.

### 29.1.3. Web services in bioinformatics

SOAP- and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads.

Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment), and BSA (Biological Sequence Analysis). The availability of these service-oriented bioinformatics resources demonstrate the applicability of web-based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

### 29.1.4. Education platforms

Resources for learning concepts of bioinformatics include online software platforms, massive online open courses (MOOC) such as:

- Rosalind
- Swiss Institute of Bioinformatics Training Portal,
- The Canadian Bioinformatics Workshops,
- Coursera's Bioinformatics Specialization (UC San Diego),
- Genomic Data Science Specialisation (Johns Hopkins),

## D 7.1 Production of materials for improved genotyping training

- EdX's Data Analysis for Life Sciences XSeries (Harvard),
- Masters In Translational Bioinformatics (University of Southern California).

### 29.2. Bioinformatics pipelines and workflow management

The advantage of the stepwise data analysis pipelines is that can be automated and routinely and repeatedly utilised on datasets, can be optimised for large datasets, thus enabling faster analysis times, repeatable analysis, and reduced chances of errors. Due to a wide variety of program languages, algorithms, and available resources and customised code writing, different pathways may be employed for achieving the desired outcome. This document is meant to illustrate the wide scope and the flexibility of the bioinformatics approach in genotyping by providing examples and references.

A bioinformatics workflow management system is a specialised form of a workflow management system designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, in a Bioinformatics application. Such systems are designed to:

1. provide an easy-to-use environment for individual application scientists themselves to create their own workflows,
2. provide interactive tools for the scientists enabling them to execute their workflows and view their results in real-time,
3. simplify the process of sharing and reusing workflows between scientists, and
4. enable scientists to track the provenance of the workflow execution results and the workflow creation steps.

Some of the platforms providing this service are Galaxy, Kepler, Taverna, UGENE, Anduril and HIVE.

#### 29.2.1. Examples of genotyping bioinformatics pipelines and applications

Bioinformatics analysis of genotyping techniques such as microsatellites, SNPs and GBS can include data filtering, allele and genotype calling, as well as data management and visualisation applications. Advanced genotyping techniques utilise NGS, various genotyping and sequencing platforms (ABI-LifeTechnologies, Illumina, Affymetrix, Agilent...) and generate high quantities of data.

Bioinformatics tools enable:

Sequence assembly of the whole genome can be a valuable source for genetic marker identification. Bioinformatics pipelines include steps such as sequence assembly and annotation.

SNP genotyping

RNA sequencing is a tool to discover and type SNPs in genes (Rogier et al. 2018). Analysis pipeline includes read quality control, trimming, mapping, mapping post-treatment (such as duplicated read removal), variant detection, variant filtering

GBS employs high throughput sequencing of parallel genotypes. Tools enabling accurate, efficient, and user-friendly analysis include:

Mining sequence variants

For instance, a plant-based SNP annotation pipeline consisting of 3 phases (Reads alignment and SNP prediction, Model evaluation, Annotation, and visualisation) was implemented in Linux and consisted of a Pearl wrapper script that

## D 7.1 Production of materials for improved genotyping training

for a seamless analysis called a series of individual Perl scripts addressing pre-processing and processing steps by utilising various tools and Bioconductor R packages (Bhardway, 2018).

### Online resources:

- Foundations of Computational and Systems Biology MIT Course; <https://web.archive.org/web/20071222091912/http://ocw.mit.edu/OcwWeb/Biology/7-91JSpring2004/LectureNotes/index.htm>
- Computational Biology: Genomes, Networks, Evolution Free MIT Course; <http://stellar.mit.edu/S/course/6/fa19/6.047/>
- <https://academic.oup.com/nar/issue/47/D1>
- [https://en.wikipedia.org/wiki/List\\_of\\_biological\\_databases](https://en.wikipedia.org/wiki/List_of_biological_databases)

### Further reading

- Baxevanis AD, Ouellette BFF. 2005. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, third edition. Wiley.
- Baxevanis AD, Petsko GA, Stein LD, Stormo GD. 2007. Current Protocols in Bioinformatics. Wiley.
- Durbin R, Eddy S, Krogh A, Mitchison G. 1998. Biological sequence analysis. Cambridge University Press.
- Kohane IS, Butte AJ, Kho A. 2002. Microarrays for an Integrative Genomics. The MIT Press.
- Tisdall J. 2001. Beginning Perl for Bioinformatics. O'Reilly.

### 30. Data science in R

R is a widely used free programming language and software environment for statistical computing and analysis, development of statistical software, data mining and data analysis. It was developed and supported by the R Foundation for Statistical Computing and has an increasingly large base of users contributing to its development. In recent years the popularity of R has increased among statisticians, data miners and researchers, which is evident by increased citations in the scientific literature as well as introduction into statistics curriculums. The basic as well as more advanced R applications are a subject of several excellent on-line tutorials.

A wide variety of statistical and graphical techniques are implemented in R and its libraries, including modelling (linear and non-linear), classical statistical tests, time-series analysis, classification, clustering. New functions in R can be easily integrated into packages and an active community has provided numerous extensions and analysis packages available to all users. In addition to the many functions written in R, other programming languages such as C, C++ and Fortran can be applied for computationally intensive tasks. R object can also be directly manipulated by C, C++, Java, .NET or Python code. R users have provided many valuable extensions and packages performing specific functions and devoted to specific areas of study.

R also has static graphics and can produce quality graphs ready for publication, including mathematical symbols. Additional packages offer dynamic and interactive graphics. Comprehensive documentation in numerous formats can be created by an integrated Rd documentation format.

#### 30.1. Packages

R installation includes a core set of packages and more than 15'000 additional packages are available at repositories such as Comprehensive R Archive Network (CRAN), Bioconductor, Omegahat and GitHub. Packages created by the user community allow specialized statistical techniques, graphical devices, import/export capabilities and reporting tools (knitr, Sweave). The R packaging system is also used by researchers to create compendia to organise research data, code, and report files in a systematic way for sharing and public archiving.

The CRAN website lists a wide range of tasks in fields such as Finance, Genetics, High Performance Computing, Machine Learning, Medical Imaging, Social Sciences and Spatial Statistics to which R has been applied and for which packages are available. FDA has identified R as suitable for interpreting data from clinical research. Other R package resources include Crantastic, a community site for rating and reviewing all CRAN packages, and R-Forge, a central platform for the collaborative development of R packages, R-related software, and projects. R-Forge also hosts many unpublished beta packages, and development versions of CRAN packages. The Bioconductor project provides R packages for the analysis of genomic data and includes object-oriented data-handling and analysis tools for data from Affymetrix, cDNA microarray, and next-generation high-throughput sequencing methods.

#### 30.2. Tidyverse

##### 30.2.1. *Tidyverse and how to use it*

**Tidyverse** is a collection of packages for data science that enable intuitive data management, analysis, plotting and reproducible programming. All the associated

## D 7.1 Production of materials for improved genotyping training

packages can be installed and loaded by simple console commands `install.packages("tidyverse")` and `library("tidyverse")`. Among them are `tidyr` – a set of functions for tidying data, `dplyr` – a set of functions for data manipulation and `ggplot2` – a set of functions for graphical presentation of data. Tidyverse introduces an intuitive syntax and functions with all packages sharing data representations and API design. For instance, the `%>%` operator passes the object on the left side as first argument in the function on the right side. The best place to learn about all the packages in the tidyverse and how they fit together is *R for Data Science*.

### 30.2.2. Tidyverse packages

The commonly known and widely used packages:

- `ggplot2`: advanced data visualisation [SO\\_doc](#)
- `dplyr`: fast (Rcpp) and coherent approach to data manipulation [SO\\_doc](#)
- `tidyr`: tools for data tidying [SO\\_doc](#)
- `readr`: for data import
- `purrr`: makes your pure functions purr by completing R's functional programming tools with important features from other languages, in the style of the JS packages `underscore.js`, `lodash` and `lazy.js`.
- `tibble`: a modern re-imagining of data frames
- `magrittr`: piping to make code more readable [SO\\_doc](#)

Packages for manipulating specific data formats:

- `hms`: easily read times
- `stringr`: provide a cohesive set of functions designed to make working with strings as easy as possible
- `lubridate`: advanced date/times manipulations [SO\\_doc](#)
- `forcats`: advanced work with factors.

Data import:

- `DBI`: defines a common interface between the R and database management systems (DBMS)
- `haven`: easily import SPSS, SAS and Stata files [SO\\_doc](#)
- `httr`: the aim of `httr` is to provide a wrapper for the `curl` package, customized to the demands of modern web APIs
- `jsonlite`: a fast JSON parser and generator optimized for statistical data and the web
- `readxl`: read .xls and .xlsx files without need for dependency packages [SO\\_doc](#)
- `rvest`: `rvest` helps you scrape information from web pages [SO\\_doc](#)
- `xml2`: for XML

and modelling:

- `modelr`: provides functions that help you create elegant pipelines when modelling
- `broom`: easily extract the models into tidy data

Finally, tidyverse suggest the use of:

- `knitr`: the amazing general-purpose literate programming engine, with lightweight API's designed to give users full control of the output without heavy coding work. [SO\\_docs](#): one, two



## D 7.1 Production of materials for improved genotyping training

- rmarkdown: Rstudio's package for reproducible programming. SO\_docs: one, two, three, four.

### Online resources for R

- Tidyverse website; [www.tidyverse.org](http://www.tidyverse.org)
- “R for data science” online book; <https://r4ds.had.co.nz/>
- RStudio website; <https://www.rstudio.com/>
- Graphical cheatsheets for “Tidyverse” packages summarising all the data manipulation, plotting etc.; <https://www.bioinfoacademy.com/>; [Practical data science with R](#)
- [https://en.wikipedia.org/wiki/R\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

## 30.3 Case study; Genomic selection in R

### Learning Objectives:

- Download the package and load the sample files
- Impute missing markers using A.mat()
- Define the training and validation populations
- Run mixed.solve() and determine the accuracy of predictions

### Overview of rrBLUP package

- Download from CRAN-version 4; Must use R version 2.14.1 or greater
- Uses ridge regression BLUP for genomic predictions
- Predicts marker effects through mixed.solve()
- A.mat() command can be used to impute missing markers; Mixed.solve does not allow NA marker values
- Define the training and validation populations

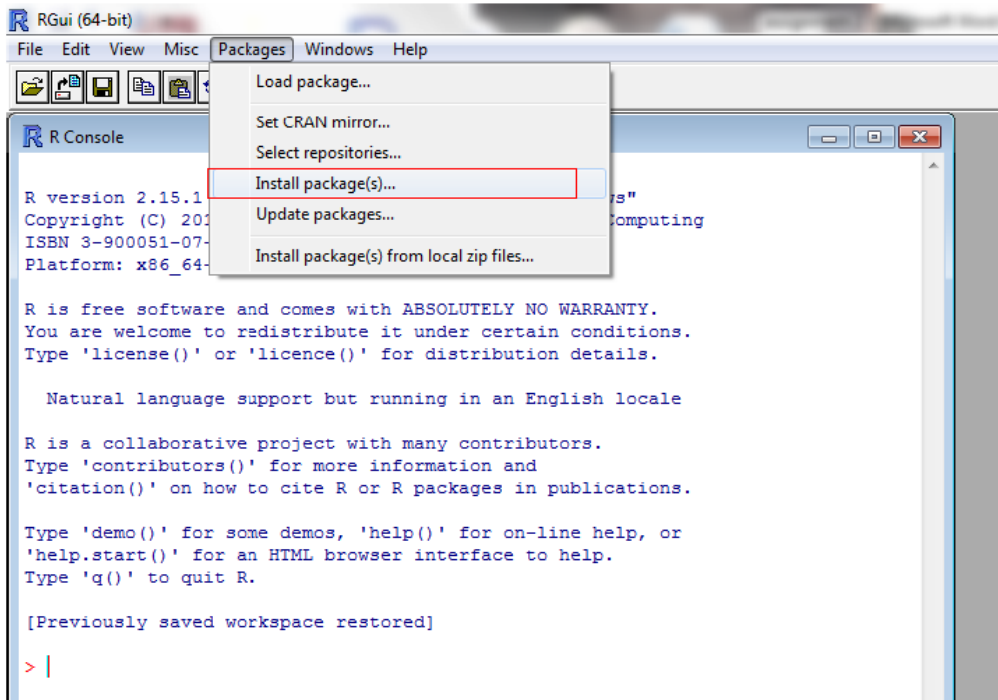
### One Step vs. Two Steps

- One step  
Uses a mixed model analysis for the plot data
- Two step  
Adjusted means are calculated across locations  
Means are then used in ridge regression blup  
Computationally more efficient and faster

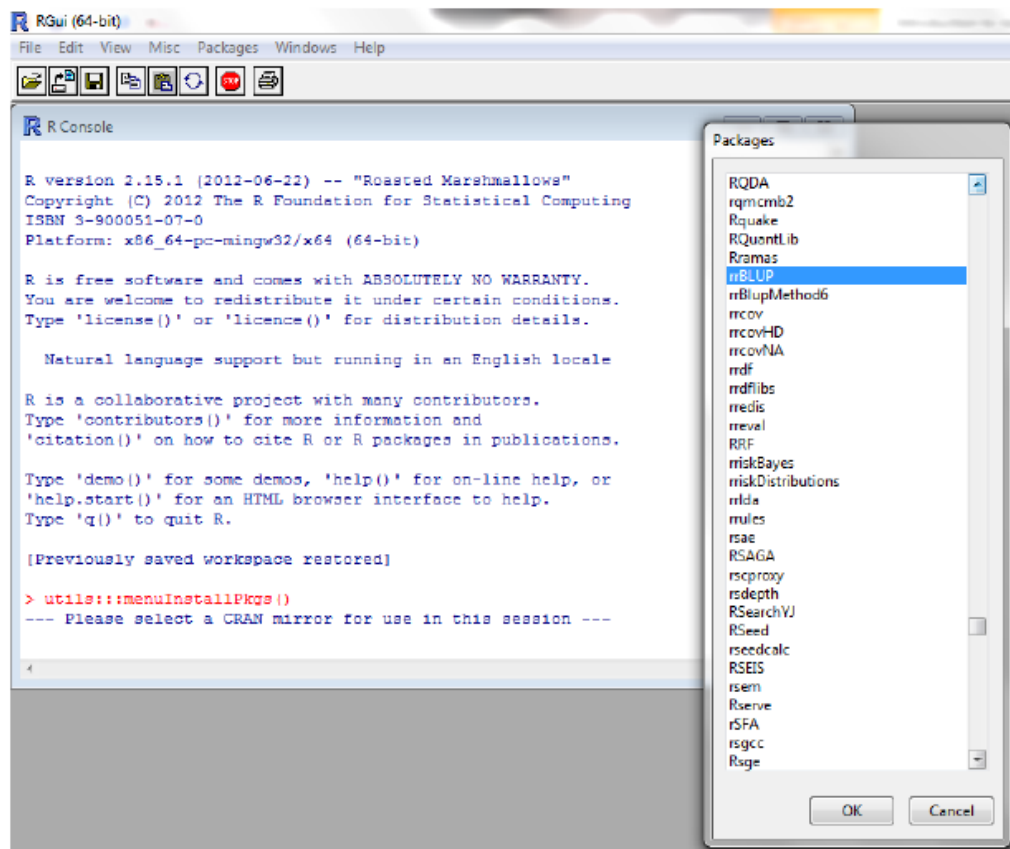
#### 30.3.1. Install the rrBLUP Package

1. Launch R->Packages->Install Package
2. Select CRAN Mirror nearest you

## D 7.1 Production of materials for improved genotyping training



### 3. Select the rrBLUP package

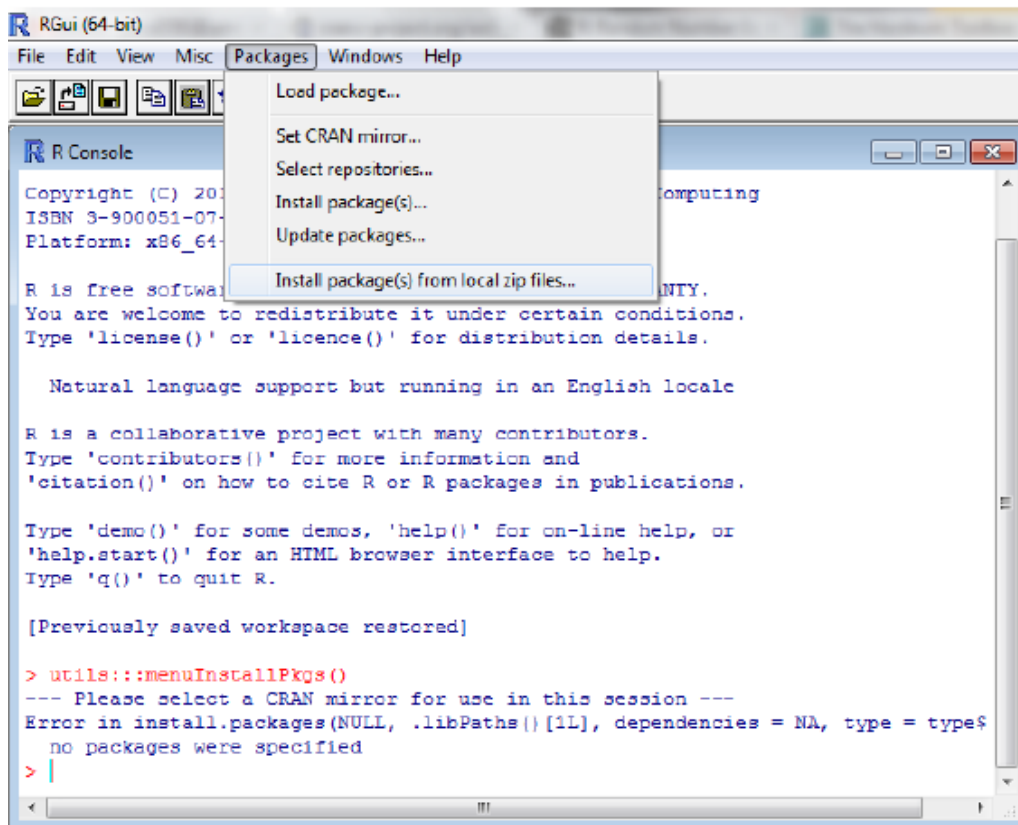


### 4. Install the package by a zip file

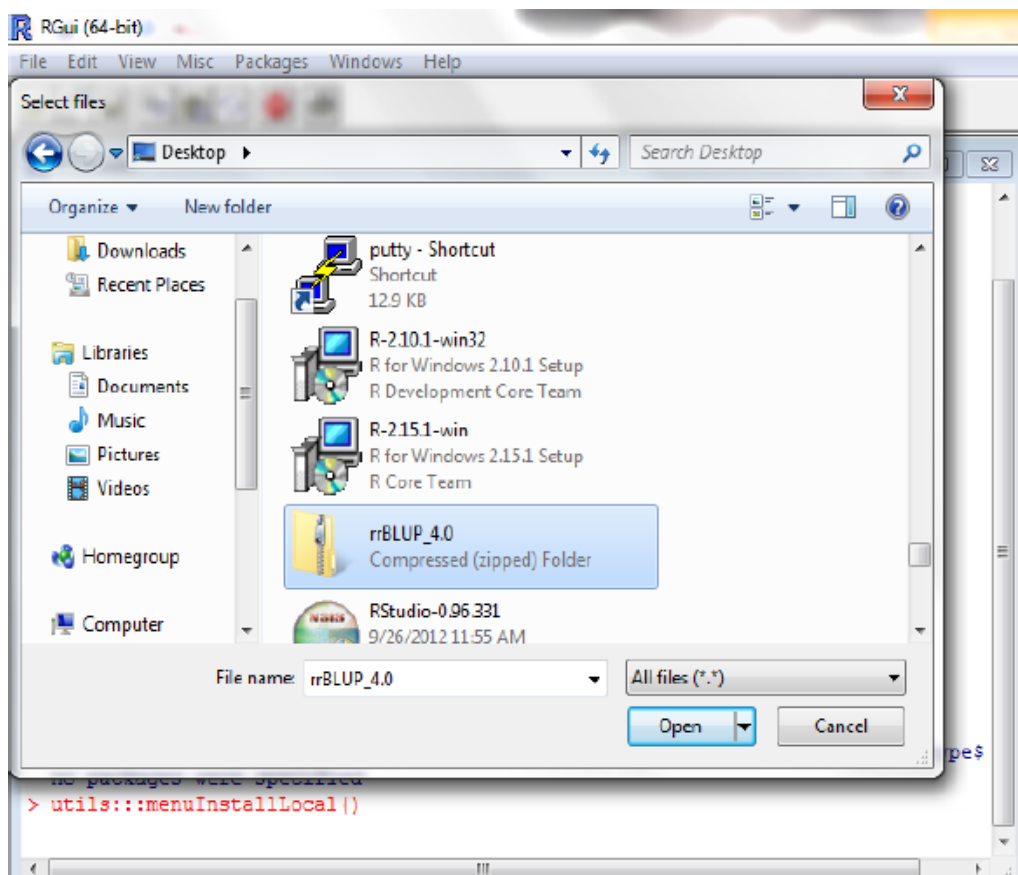
<http://cran.r-project.org/web/packages/rrBLUP/index.html>

Packages->install package from local zip files

## D 7.1 Production of materials for improved genotyping training

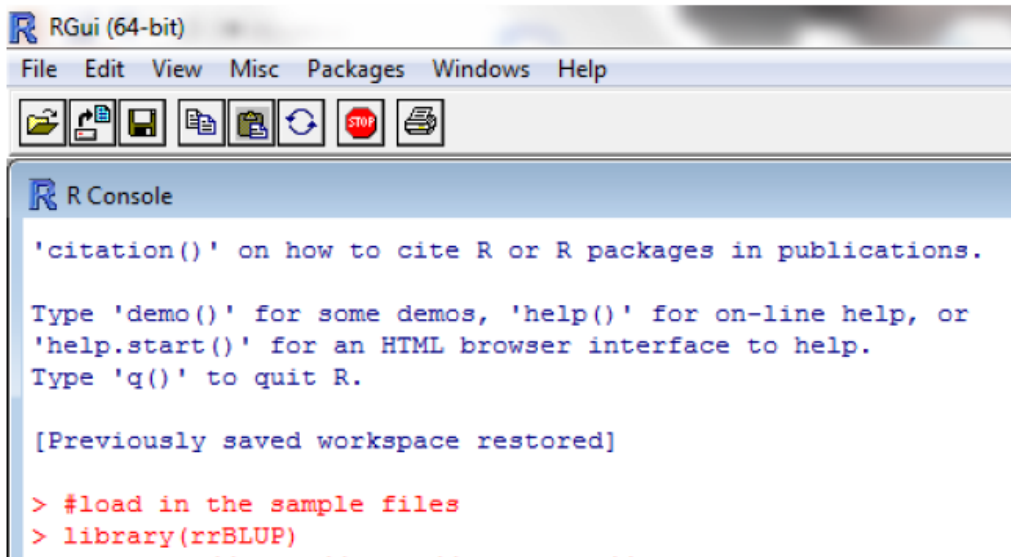


5. Select the package from saved location



## D 7.1 Production of materials for improved genotyping training

Now that the package is installed, the library must be loaded every time R is opened.



```
R R Console
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

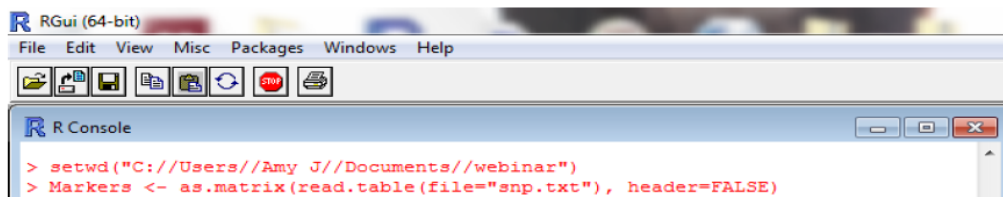
> #load in the sample files
> library(rrBLUP)
```

### 30.3.2. Sample Files

- Files downloaded from the Hordeum Toolbox <http://hordeumtoolbox.org/>
- University of Minnesota barley breeding program preliminary yield trial-St. Paul location in 2009
- Phenotypic traits-yield, plant height and heading date
- 1178 markers, 164 NA markers
- 1 = homozygous for parent 1, 0 = heterozygous, and -1 homozygous for parent 2
- Markers must be in the {-1,0,1} format for rrBLUP

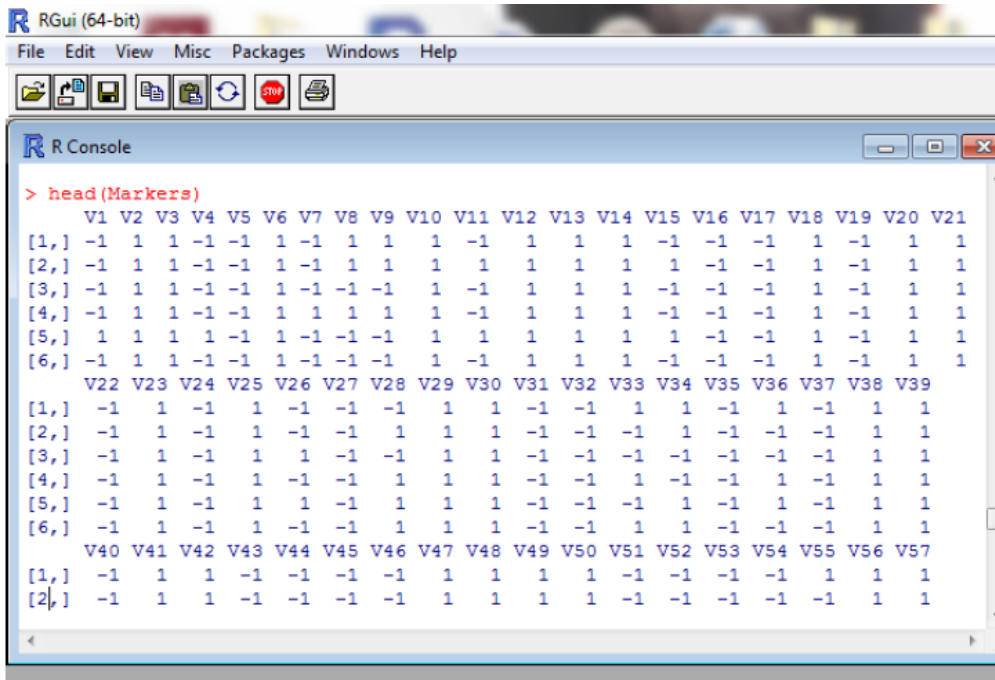
### 30.3.3. Load the Sample Files

1. Setwd()-Set the working directory to the location of sample files
2. Read.table command used for .txt files
3. Read.csv command used for .csv files
4. Header=F since sample marker file does not have a header with marker names
5. head() command used to see the first 5 lines of a file
6. Useful to see if data was loaded correctly



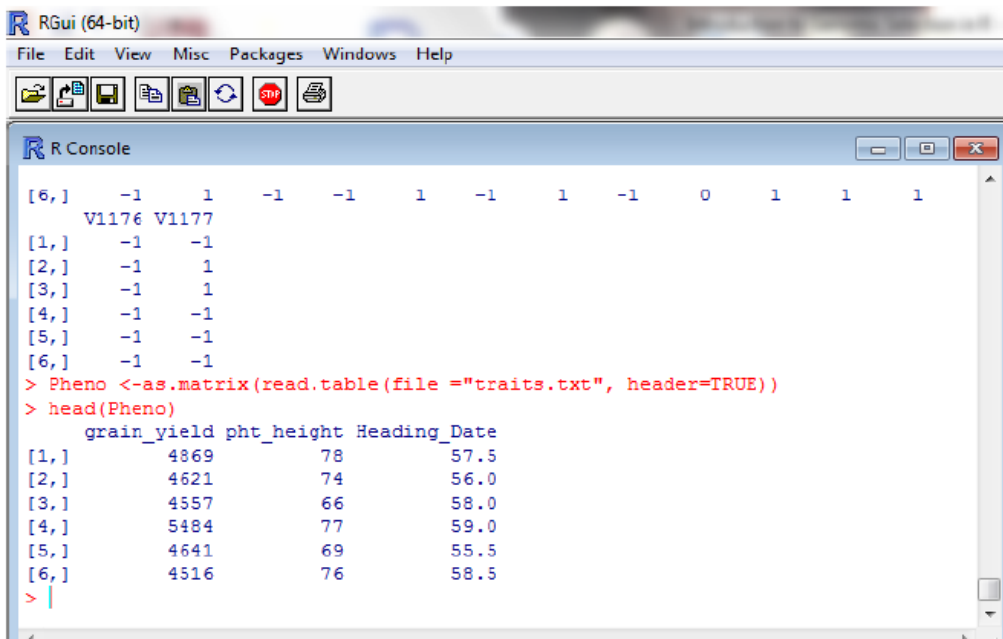
```
R R Console
> setwd("C://Users//Amy J//Documents//webinar")
> Markers <- as.matrix(read.table(file="snp.txt"), header=FALSE)
```

## D 7.1 Production of materials for improved genotyping training



```
RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
> head(Markers)
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21
[1,] -1 1 1 -1 -1 1 -1 1 1 1 -1 1 1 1 1 -1 -1 1 -1 1 1
[2,] -1 1 1 -1 -1 1 -1 1 1 1 1 1 1 1 1 -1 -1 1 -1 1 1
[3,] -1 1 1 -1 -1 1 -1 -1 -1 1 -1 1 1 1 1 -1 -1 -1 1 -1 1
[4,] -1 1 1 -1 -1 1 1 1 1 1 -1 1 1 1 1 -1 -1 -1 1 -1 1
[5,] 1 1 1 1 -1 1 -1 -1 -1 1 1 1 1 1 1 1 -1 -1 1 -1 1
[6,] -1 1 1 -1 -1 1 -1 -1 -1 1 -1 1 1 1 1 -1 -1 -1 1 -1 1
  V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39
[1,] -1 1 -1 1 -1 -1 -1 1 1 -1 -1 1 1 -1 1 -1 1 1
[2,] -1 1 -1 1 -1 -1 -1 1 1 -1 -1 -1 1 -1 -1 -1 1 1
[3,] -1 1 -1 1 1 -1 -1 1 1 -1 -1 -1 -1 -1 -1 -1 1 1
[4,] -1 1 -1 1 -1 -1 1 1 1 -1 -1 1 -1 -1 1 -1 1 1
[5,] -1 1 -1 1 1 -1 1 1 1 -1 -1 -1 1 -1 1 -1 1 1
[6,] -1 1 -1 1 -1 -1 -1 1 1 1 -1 -1 1 1 -1 -1 -1 1 1
  V40 V41 V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57
[1,] -1 1 1 -1 -1 -1 -1 1 1 1 1 -1 -1 -1 -1 1 1 1
[2,] -1 1 1 -1 -1 -1 -1 1 1 1 1 -1 -1 -1 -1 -1 1 1
```

7. Load the phenotype file and use the head command to see the first five lines
8. Header=T since phenotype files have column names
9. Markers and phenotypes must be in matrix format



```
RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
[6,] -1 1 -1 -1 1 -1 1 -1 0 1 1 1
  V1176 V1177
[1,] -1 -1
[2,] -1 1
[3,] -1 1
[4,] -1 -1
[5,] -1 -1
[6,] -1 -1
> Pheno <-as.matrix(read.table(file ="traits.txt", header=TRUE))
> head(Pheno)
  grain_yield pht_height Heading_Date
[1,] 4869 78 57.5
[2,] 4621 74 56.0
[3,] 4557 66 58.0
[4,] 5484 77 59.0
[5,] 4641 69 55.5
[6,] 4516 76 58.5
> |
```

10. Determine the size of the matrices
11. dim() command gives the number of rows and columns
12. 96 observations and 1178 markers, 3 traits

## D 7.1 Production of materials for improved genotyping training

```
-  
> Pheno <-as.matrix(read.table(file ="traits.txt", header=TRUE))  
> head(Pheno)  
      grain_yield pht_height Heading_Date  
[1,]      4869         78         57.5  
[2,]      4621         74         56.0  
[3,]      4557         66         58.0  
[4,]      5484         77         59.0  
[5,]      4641         69         55.5  
[6,]      4516         76         58.5  
>  
> dim(Markers)  
[1]   96 1178  
> dim(Pheno)  
[1]  96   3  
> |
```

### 30.3.4 Impute Missing Markers

rrBLUP mixed.solve does not allow for missing markers

Imputed value is the population mean for that marker

Useful for SNP data since level of missing data is low

In the sample files 164 markers are missing out of 1178 (0.14%)

A.mat also calculates the additive relationship matrix

max.missing-maximum proportion of missing data

If 50% of markers are missing data then markers are not imputed

impute method- imputes the mean of the markers

return.imputed-prints out the imputed results if set to TRUE

```
> #what if markers are NA?  
> #impute with A.mat  
> impute=A.mat(Markers,max.missing=0.5,impute.method="mean",return.imputed=T)  
> |
```

>impute=A.mat(Markers,max.missing=0.5,impute.method="mean",return.imputed=  
T)

```
> Markers_impute=impute$imputed
```

Rename imputed marker matrix as Markers\_impute

impute\$imputed-returns the imputed marker matrix

impute\$A-returns the additive relationship matrix

```
>impute$imputed
```



## D 7.1 Production of materials for improved genotyping training

- `>impute$imputed`

Imputed marker value

Marker value left NA if more than 50% missing data

```

R Console
V137 V138 V139 V140 V141 V142 V143 V144 V145 V146 V147 V148 V149 V150
[1,] -1 1 1 1 1 1 -1 -1 1 -1 0.8043478 1 1 1 1
[2,] -1 1 1 1 1 1 -1 -1 -1 -1 1.0000000 1 1 1 1
[3,] -1 1 1 1 1 1 -1 -1 1 -1 1.0000000 1 1 1 1
[4,] 1 -1 1 1 1 1 -1 -1 1 -1 1.0000000 1 1 1 1
[5,] -1 1 1 1 1 1 -1 -1 -1 -1 1.0000000 1 -1 1 1
[6,] -1 1 1 1 1 1 -1 -1 -1 -1 1.0000000 1 1 1 1
V151 V152 V153 V154 V155 V156 V157 V158 V159 V160 V161 V162 V163 V164 V165
[1,] 1 1 -1 1 -1 1 1 -1 1 1 1 -1 -1 1 1 -1
[2,] 1 1 -1 1 -1 -1 1 -1 1 1 1 -1 -1 1 1 -1
[3,] 1 1 -1 1 -1 1 1 -1 1 1 1 -1 -1 1 1 -1
[4,] 1 1 -1 1 -1 1 1 -1 1 1 1 -1 -1 1 1 -1
[5,] 1 1 -1 1 -1 1 1 -1 1 1 1 -1 -1 1 1 1
[6,] 1 1 -1 1 -1 -1 1 -1 1 1 1 -1 -1 1 1 1
V166 V167 V168 V169 V170 V171 V172 V173 V174 V175 V176 V177 V178 V179 V180
[1,] 1 1 1 1 NA 1 1 1 -1 -1 1 1 1 -1 -1 1
[2,] 1 1 1 1 NA 1 1 1 -1 -1 1 1 1 -1 -1 1
[3,] 1 1 1 1 NA 1 1 1 -1 -1 1 1 1 -1 -1 1
[4,] 1 1 1 1 NA 1 1 1 -1 -1 1 1 1 -1 -1 1

```

Remove markers that had more than 50% missing data

NA values are not allowed in mixed.solve

Two markers in the SNP file must be removed

Column 169 and 562

New dimensions show 2 less columns

Use `Markers_impute2` as marker matrix for estimating marker effects

```

> Markers_impute2=Markers_impute[,-c(169,562)]
> dim(Markers_impute)
[1] 96 1178
> dim(Markers_impute2)
[1] 96 1176
> |

```

### 30.3.4. Training and Validation Populations

- Training population-genotyped and phenotyped
- Validation population-phenotype values estimated based on marker effects calculated from training population
- Code is set that 60% of the total population is the training population
- 40% validation population
- 58 (60% of total population of 96) random numbers sampled to determine which individuals are in the training population
- Individuals are the row numbers for the phenotypes and marker matrices
- Sampled numbers will be different every time the code is run and will affect the correlation accuracy

## D 7.1 Production of materials for improved genotyping training

```
> train= as.matrix(sample(1:96, 58))
> head(train)
      [,1]
[1,]  52
[2,]  82
[3,]  50
[4,]  14
[5,]   7
[6,]  80
> |
```

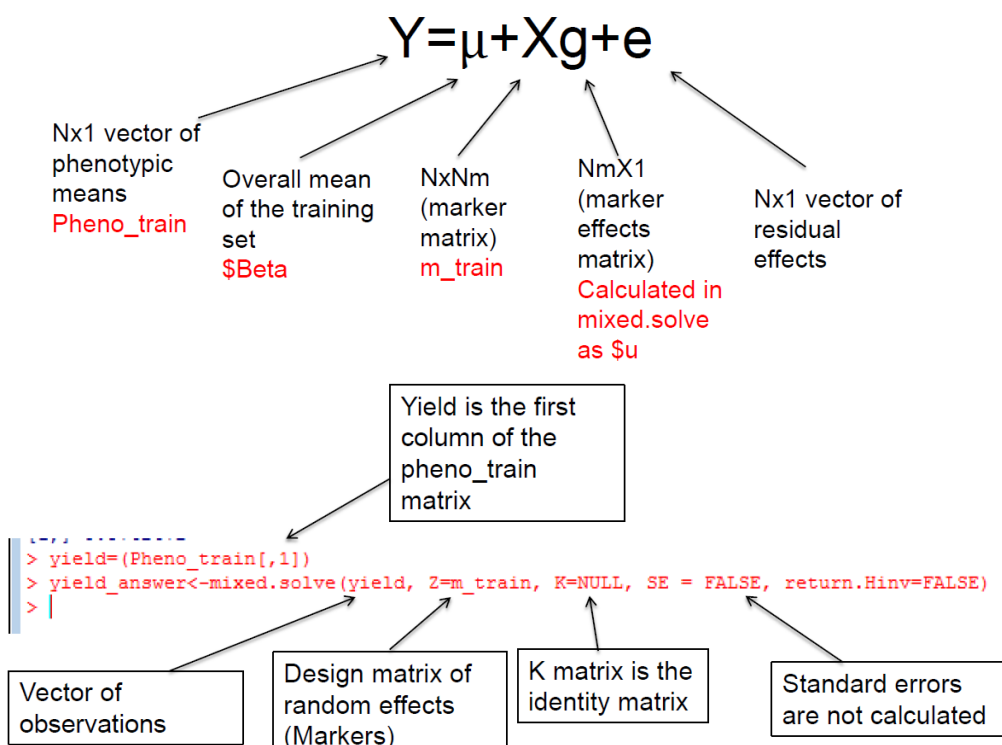
- Validation population is 40% of the total population
- `setdiff()` command determines the numbers that are not in the training population and will be part of the validation population

```
> test<-setdiff(1:96,train)
> test
 [1]  6 12 18 19 22 23 26 27 28 29 33 34 36 40 41 43 47 48 53 54 55 56 57 58 59 62 66
[28] 68 71 75 77 79 83 84 86 90 91 96
> |
```

- `Pheno_train` and `m_train` are the phenotype and marker matrices for the values in the training population
- `Pheno_valid` and `m_valid` will be the validation populations

```
> Pheno_train=Pheno[train,]
> m_train=Markers_impute2[train,]
> Pheno_valid=Pheno[test,]
> m_valid=Markers_impute2[test,]
> |
```

### 30.3.5. Run `mixed.solve`



## D 7.1 Production of materials for improved genotyping training

Yield\_answer\$u is the output of the marker effects

head(e) shows the marker effects for the first five markers

```
> yield_answer<-mixed.solve(yield, Z=m_train, K=NULL, SE = FALSE, return.Hinv=FALSE)
> YLD = yield_answer$u
> e = as.matrix(YLD)
> head(e)
      [,1]
V1  1.1380597
V2 -0.1141220
V3  0.4970927
V4  2.2986051
V5  0.3579770
V6 -0.1141220
> |
```

m\_valid\*e = marker validation matrix times the marker effects

Pred\_yield=predicted yield based on the marker effects of the training population with the grand mean added in

```
> pred_yield_valid = m_valid %*% e
> pred_yield=(pred_yield_valid[,1])+yield_answer$beta
> pred_yield
 [1] 4745.698 4621.133 4742.935 4601.210 4671.582 4636.899 4552.350 4486.954
 [9] 4589.440 4601.534 4508.288 4656.675 4462.313 4493.898 4668.741 4498.701
[17] 4708.654 4593.296 4441.527 4705.500 4597.538 4089.056 4177.749 4261.560
[25] 4107.757 4207.431 4454.215 4713.850 4740.123 4537.690 4585.838 4526.935
[33] 4570.133 4512.558 4613.167 4412.658 4747.170 4872.127 4774.157 4697.992
[41] 4640.538 4576.519 4707.957 4658.228 4772.145 4596.747 4371.145 4779.256
[49] 4427.464 4525.557 4305.716 4564.654 4450.188 4634.591 3989.726 4068.685
[57] 4043.495 3886.869
```

### Determine Correlation Accuracy:

- Correlation between the predicted yield values and the observed yield values
- Accuracy will change slightly each time due to different individuals sampled for the training and validation populations

```
> yield_valid = Pheno_valid[,1]
> YLD_accuracy <-cor(pred_yield_valid, yield_valid, use="complete" )
> YLD_accuracy
      [,1]
[1,] 0.2521498
> |
```

- Plant Height

```
> PHT_HT=(Pheno_train[,2])
> PHT_HT_answer<-mixed.solve(PHT_HT, Z=m_train, K=NULL, SE = FALSE, return.Hinv=FALSE)
> PHT_HT = PHT_HT_answer$u
> e = as.matrix(PHT_HT)
> pred_PHT_HT_valid = m_valid %*% e
> PHT_HT_valid = Pheno_valid[,2]
> PHT_HT_accuracy <-cor(pred_PHT_HT_valid, PHT_HT_valid, use="complete" )
> PHT_HT_accuracy
      [,1]
[1,] 0.4055428
> |
```

## D 7.1 Production of materials for improved genotyping training

- Heading Date

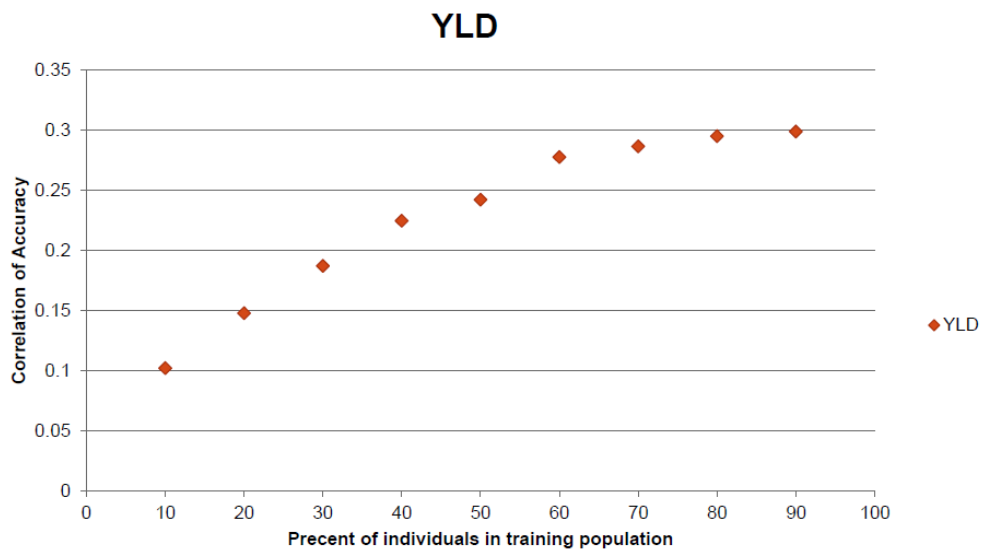
```
> HD_DATE=(Pheno_train[,3])
> HD_DATE_answer<-mixed.solve(HD_DATE, Z=m_train, SE = FALSE, return.Hinv=FALSE)
> HD_DATE = HD_DATE_answer$u
> e = as.matrix(HD_DATE)
> pred_HD_DATE_valid = m_valid %*% e
> HD_DATE_valid = Pheno_valid[,3]
> HD_DATE_accuracy <-cor(pred_HD_DATE_valid, HD_DATE_valid, use="complete" )
> HD_DATE_accuracy
      [,1]
[1,] 0.5205029
> |
```

- Correlation accuracy with 500 iterations

```
>
> ##### cross validation for many cycles for yield only
> traits=1
> cycles=500
> accuracy = matrix(nrow=cycles, ncol=traits)
> for(r in 1:cycles)
+ {
+ train= as.matrix(sample(1:96, 38))
+ test<-setdiff(1:96,train)
+ Pheno_train=Pheno[train,]
+ m_train=Markers_impute2[train,]
+ Pheno_valid=Pheno[test,]
+ m_valid=Markers_impute2[test,]
+
+ yield=(Pheno_train[,1])
+ yield_answer<-mixed.solve(yield, Z=m_train, K=NULL, SE = FALSE, return.Hinv=F$
+ YLD = yield_answer$u
+ e = as.matrix(YLD)
+ pred_yield_valid = m_valid %*% e
+ pred_yield=(pred_yield_valid[,1])+yield_answer$beta
+ pred_yield
+ yield_valid = Pheno_valid[,1]
+ accuracy[r,1] <-cor(pred_yield_valid, yield_valid, use="complete" )
+ }
> mean(accuracy)
[1] 0.2305713
```

- Correlation accuracy is different for each trait
- Values will be different every time it is run since different lines will be included in the training or validation sets
- Accuracy is affected by training size, validation size, number of markers and heritability
- Effects of training population size on accuracy

## D 7.1 Production of materials for improved genotyping training



### 30.3.6. Common Errors

- Headers incorrectly input

```
> Pheno <-as.matrix(read.table(file ="traits.txt", header=F))
> head(Pheno)
      V1      V2      V3
[1,] "grain_yield" "pht_height" "Heading_Date"
[2,] "4869"        "78"        "57.5"
[3,] "4621"        "74"        "56"
[4,] "4557"        "66"        "58"
[5,] "5484"        "77"        "59"
[6,] "4641"        "69"        "55.5"
> train= as.matrix(sample(1:96, 38))
> test<-setdiff(1:96,train)
> Pheno_train=Pheno[train,]
> m_train=Markers_impute2[train,]
> Pheno_valid=Pheno[test,]
> m_valid=Markers_impute2[test,]
> yield=(Pheno_train[,1])
> yield_answer<-mixed.solve(yield, Z=m_train, K=NULL, SE = FALSE, return.Hinv=F$
Error in crossprod(x, y) :
  requires numeric/complex matrix/vector arguments
```

- NA Markers

```
> train= as.matrix(sample(1:96, 38))
> test<-setdiff(1:96,train)
> Pheno_train=Pheno[train,]
> m_train=Markers[train,]
> Pheno_valid=Pheno[test,]
> m_valid=Markers[test,]
> yield=(Pheno_train[,1])
> yield_answer<-mixed.solve(yield, Z=m_train, K=NULL, SE = FALSE, return.Hinv=FALSE)
Error in eigen(Hb, symmetric = TRUE) : infinite or missing values in 'x'
```

- Incorrect matrix dimensions
- Removed one individual from phenotype matrix

## D 7.1 Production of materials for improved genotyping training

```
> #####
> #define the training and test populations
> #training-60% validation-40%
> train= as.matrix(sample(1:96, 38))
> test<-setdiff(1:96,train)
> Pheno_train=Pheno[train,]
> m_train=Markers_impute2[train,]
> Pheno_valid=Pheno[test,]
Error: subscript out of bounds
> m_valid=Markers_impute2[test,]
>
> #####
> yield=(Pheno_train[,1])
> yield_answer<-mixed.solve(yield, Z=m_train, K=NULL, SE = FALSE, return.Hinv=FS
> YLD = yield_answer$u
> e = as.matrix(YLD)
> pred_yield_valid = m_valid %*% e
> pred_yield=(pred_yield_valid[,1])+yield_answer$beta
> yield_valid = Pheno_valid[,1]
Error: object 'Pheno_valid' not found
> YLD_accuracy <-cor(pred_yield_valid, yield_valid, use="complete" )
Error in is.data.frame(y) : object 'yield_valid' not found
> YLD_accuracy
Error: object 'YLD_accuracy' not found
> |
```

- Read in values as characters instead of numeric
- Quotes around values

```
> Pheno <-as.matrix(read.table(file ="traits.txt", header=F))
> head(Pheno)
      V1      V2      V3
[1,] "grain_yield" "pht_height" "Heading_Date"
[2,] "4869"        "78"        "57.5"
[3,] "4621"        "74"        "56"
[4,] "4557"        "66"        "58"
[5,] "5484"        "77"        "59"
```

### Software and resources:

- rrBLUP reference manual; <http://cran.r-project.org/web/packages/rrBLUP/rrBLUP.pdf>
- Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255. doi: 10.3835/plantgenome2011.08.0024
- <https://pbgworks.org/sites/pbgworks.org/files/Introduction%20to%20Genomic%20Selection%20in%20R.pdf> (all figures are taken from this source)



## D 7.1 Production of materials for improved genotyping training

### 31. Tutorial I: TASSEL 3.0 Genotyping by Sequencing (GBS) pipeline documentation

#### 31.1. Introduction

The GBS analysis pipeline is an extension to the program TASSEL, and, as such, GBS commands are run as TASSEL plugins in the following format (Linux or Mac operating system):

```
run_pipeline.pl -fork1 -PluginName --plugin-option -endPlugin -runfork1
```

Each step of the pipeline is specified with a "fork" command and a number, since TASSEL can run several processes at once, then split and recombine their results. The fork option is followed by the name of the plugin, and any plugin-specific options. If no plugin options are provided, the program will print a list of available options. -endPlugin signals the end of plugin-specific options, and -runfork1 then runs the specified plugin.

See <http://www.maizegenetics.net/tassel/docs/TasselPipelineCLI.pdf> for general instructions on how to install the TASSEL 3.0 Standalone Build on your computer. These GBS-specific instructions assume that you have unzipped the standalone into the directory (folder)

```
/programs
```

and then renamed the directory

```
/programs/tassel3.0_standalone
```

to

```
/programs/tassel
```

If not, you will have to edit the example commands appropriately (e.g., replace "tassel" with "tassel3.0\_standalone").

Note that many of the GBS commands ("plugins") produce a large amount of console output ("stdout") that is not discussed below (at least not in this incomplete draft). Much of this output is very informative, so you will likely find it helpful to either copy and paste it to a text log file or redirect stdout to both the console and a log file (e.g., using '| tee GBSlogfile20110915.txt' in linux).

#### 31.2. Recommended directory (folder) structure for a GBS analysis

A dot (.) will represent the working directory (folder) for your analysis, which will be your current working directory:

(e.g., /home/myUserName/myGBSstudyName).

The example commands below don't create the directories (and will fail if they don't already exist), so at the start of the analysis, create the following directories inside your working directory.

```
./qseq
```

```
./tagCounts
```

## D 7.1 Production of materials for improved genotyping training

*./topm*  
*./mergedTagCounts*  
*./tbt*  
*./mergedTBT*  
*./hapmap*  
*./hapmap/unfilt* (for output from TagsToSNPByAlignmentMTPlugin)  
*./hapmap/mergedSNPs* (for output from MergeDuplicateSNPsPlugin)  
*./hapmap/filt* (for output from GBSHapMapFiltersPlugin)

### QseqToTagCountPlugin

**Summary:** Derives a tagCount list for each qseq file in the input directory (and all sub-directories thereof). Keeps only good reads having a barcode and a cut site and no N's in the useful part of the sequence. Trims off the barcodes and truncates sequences that (1) have a second cut site, or (2) read into the common adapter.

**Input:** (i) Barcode key file (see example); (ii) Directory (folder) containing qseq files

**Output:** Directory (folder) containing a corresponding tagCount file for every qseq file in the input directory

**Arguments:** QseqToTagCountPlugin

-i	Input directory containing .qseq text or gzipped text files. NOTE: Directory will be searched recursively, and should be written without a slash after its name.
-k	Key file listing barcodes for each sample.
-e	Enzyme used to create the GBS library (ApeKI or PstI).
-s	Maximum number of good reads per lane. Default is 200,000,000.
-c	Minimum number of times a tag must be present to be output. Default is 1.
-o	Output directory to contain .cnt files, one per .qseq file. Defaults to input directory (the default is not recommended – it is best to use a separate directory).

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -QseqToTagCountPlugin -i ./qseq  
-k myGBSkey.txt -e ApeKI -o ./tagCounts -endPlugin -runfork1
```

**Gory Details:** This is the initial step of a GBS analysis. This step reads a user-supplied key file (mandatory argument **-k**) in tab-delimited text format which indicates, for each lane of interest from a flow-cell, which barcodes are assigned to which sample (a short example key file is provided in Appendix 1). It then recursively searches the specified input directory (mandatory argument **-i**) and all subfolders for qseq files matching one of the flowcell/lane combinations in the key file and with the following acceptable file naming conventions:

FLOWCELL\_LANE\_qseq.txt (example: **42A87AAXX\_2\_qseq.txt**)  
FLOWCELL\_LANE\_qseq.txt.gz (example: **42A87AAXX\_2\_qseq.txt.gz**)

## D 7.1 Production of materials for improved genotyping training

code\_FLOWCELL\_s\_LANE\_qseq.txt

(example: **10225395\_42A87AAXX\_s\_2\_qseq.txt**)

code\_FLOWCELL\_s\_LANE\_qseq.txt.gz

(example: **10225395\_42A87AAXX\_s\_2\_qseq.txt.gz**)

Note that both compressed (\*.gz) and uncompressed (\*.txt) files can be read – we recommend using compressed files to save disk storage space. The “code” part of the latter two file name examples is a numerical tracking code generated by the sequencing centre. The GBS pipeline doesn’t actually use this code, so you can substitute any text or numbers (or use one of the first two conventions). The underscores are essential for correct parsing of the parts of each qseq file name (only FLOWCELL and LANE are actually used by our pipeline).

For each qseq file that has a match in the key file, QseqToTagCountPlugin finds all reads that begin with one of the expected barcodes immediately followed by the expected cut site remnant (CAGC or CTGC for *ApeKI*) and trims them to 64 bases (including the cut site remnant but removing the barcode). Reads containing N within the first 64 bases after the barcode are rejected. If a read contains either a full cut site (from incomplete digestion or chimera formation) or the beginning of the common adapter (from restriction fragments less than 64bp) within the first 64 bases it is truncated appropriately and padded to 64 bases with polyA. The actual length of truncated (or full 64 base) reads is recorded in the output tagCount file.

The output of QseqToTagCountPlugin is a single tagCount file in the specified output directory (mandatory argument **-o**) for every matching qseq file in the input directory. The tagCount files are named after their corresponding qseq file, with \*.txt.gz or \*.txt replaced by \*.cnt. The tagCount files are binary, and can only be read by our pipeline. They contain the 64 base sequence of each good, barcoded tag (padded with polyA if truncated), the actual length of the tag (before padding with polyA), and the number of times that tag was observed in the corresponding flowcell lane. The tags are sorted by their sequence.

The enzyme used to create the GBS library is indicated via mandatory option **-e**. currently, our pipeline only accepts *ApeKI* or *PstI*. The **-s** option (*maximum number of good reads per lane*) is used to set an upper limit on memory usage. So far we have not encountered a qseq (or fastq) file with more than 200 million good, barcoded reads (the default).

We usually keep the **-c** option (*minimum number of times a tag must be present to be output*) at its default value of 1. In a typical analysis, we generally combine the results of multiple lanes, or even multiple flow cells (via the next step, MergeMultipleTagCountPlugin). Tags that occur only once in a given flowcell lane might occur multiple times in other lanes, so they might be real (i.e., not from sequencing error).

### **MergeMultipleTagCountPlugin:**

**Summary:** Merges each tagCount file in the input directory into a single “master” tagCount list. Only keeps tags with a total count (after merger) greater than or equal

## D 7.1 Production of materials for improved genotyping training

to that specified in option **-c** (*minimum number of times a tag must be present to be output*).

**Input:** Input directory (folder) containing tagCount (\*.cnt) files

**Output:** Merged tagCount file (it is best to send this to a separate directory from the input directory)

**Arguments:** *MergeMultipleTagCountPlugin*

-i	Input directory containing tagCount (*.cnt) files.
-o	Output file name (should be in a separate directory from the input).
-c	Minimum number of times a tag must be present to be output. Default is 1.
-t	Specifies that reads should be output in FASTQ text format.

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -MergeMultipleTagCountPlugin  
-i ./tagCounts -o ./mergedTagCounts/myMasterTags.cnt -c 5 -endPlugin -runfork1
```

**Gory Details:** The MergeMultipleTagCountPlugin step merges multiple tagCount files produced by the QseqToTagCountPlugin step (from multiple lanes and/or flow-cells) into a single “master” tagCount file. (For a description of the tagCount file format, see QseqToTagCountPlugin.) All tagCount (\*.cnt) files in the specified input directory (argument **-i**) are merged.

To remove rare or singleton tags that possibly result from sequencing errors, we use the **-c** option (*minimum number of times a tag must be present to be output*). A **-c** option setting of 10 is typical, but when deciding on an appropriate cutoff, you should consider the number of individuals in your analysis, the expected coverage (about 0.4-0.5x for maize with *ApeKI*), the expected segregation ratio, minimum minor allele frequency of interest, etc. The merged tagCount output file is used as a master tag list for two subsequent steps: the QseqToTBTPPlugin step and alignment to the reference genome. The output is in (binary) tagCount format by default, which serves as the input format for the QseqToTBTPPlugin step.

We typically perform the alignment to the reference genome with external software, such as BWA. To obtain a master tagCount file in FASTQ format for use as input to BWA, invoke the **-t** option.

### **SAMConverterPlugin:**

**Summary:** Converts a SAM format alignment (\*.sam) file produced by the Linux program BWA into a tagsOnPhysicalMap (\*.topm.bin) file that can be used by the TagsToSNPByAlignmentMTPlugin for calling SNPs.

**Input:** SAM format alignment (\*.sam) file produced by the Linux program BWA

**Output:** tagsOnPhysicalMap (\*.topm.bin) file that can be used by the TagsToSNPByAlignmentMTPlugin for calling SNPs

**Arguments:** SAMConverterPlugin

## D 7.1 Production of materials for improved genotyping training

-i	Input SAM format alignment (*.sam) file produced by the Linux program BWA.
-o	Output tagsOnPhysicalMap (*.topm.bin) file that can be used by the TagsToSNPByAlignmentMTPlugin for calling SNPs. We recommend using the extension *.topm.bin.

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -SAMConverterPlugin
-i ./mergedTagCounts/myAlignedMasterTags.sam
-o ./topm/myMasterTags.topm.bin -endPlugin -runfork1
```

**Gory Details:** The next draft of this documentation will include more information on running BWA using the output of MergeMultipleTagCountPlugin (using -t option) as input to BWA and then using BWA to produce a \*.sam file to be used as input to this SAMConverterPlugin.

### QseqToTBTPugin:

**Summary:** Generates a TagsByTaxa file for each qseq file in the input directory (or in subfolders thereof). One TagsByTaxa file is produced per qseq file. Require a master list of tags of interest, which may come either from a tagCount or tagsOnPhysicalMap file.

**Input:** (i) Barcode key file; (ii) Directory (folder) containing qseq files

Output: Directory (folder) containing a corresponding tagsByTaxa file for every qseq file in the input directory

### Arguments: QseqToTBTPugin

-i	Input directory containing .qseq files.
-k	Barcode key file.
-e	Enzyme used to create the GBS library (ApeKI or PstI).
-o	Output directory.
-c	Minimum taxa count within a qseq file for a tag to be output. Default is 1.
-t	Tag count file listing unique reads (mutually exclusive with option -m).
-m	Physical map file listing unique reads (mutually exclusive with option -t).

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -QseqToTBTPugin -i ./qseq
-k myGBSkey.txt -e ApeKI -o ./tbt -t ./mergedTagCounts/myMasterTags.cnt
-endPlugin -runfork1
```

## D 7.1 Production of materials for improved genotyping training

**Gory Details:** Similar to QseqToTagCountPlugin, QseqToTBTPugin parses qseq files for good reads that contains a barcode and cut site remnant and that have no N's in the first 64 bases after the barcode, and trims them to 64 bases (not including the barcode). As in QseqToTagCountPlugin, QseqToTBTPugin appropriately truncates reads that contain either a full cut site or the beginning of the common adapter within the first 64 bases, and pads them to 64 bases with polyA. In a given GBS analysis, the same key file (-k option), containing the names of the taxa corresponding to each barcode in each lane, is used for both plugins.

The difference between QseqToTBTPugin and QseqToTagCountPlugin is that QseqToTBTPugin uses the barcode information to keep track of which taxa each tag of interest is observed in. Each good read in each qseq file is checked for a match against a set of tags of interest. A tagsByTaxa output file is produced for every qseq file in the input directory that has a matching flow-cell and lane in the key file. Each output file is named after its corresponding input qseq file but with ".txt.gz" or ".txt" replaced by ".tbt.bin". Each output tagsByTaxa file is in binary format (only readable by our pipeline), but can be thought of as a grid where the rows are the tags of interest, the columns are taxa names (including flow-cell, lane and well information) and the cells indicate whether or not a particular tag was observed in a particular taxon. The actual length in bases of each tag (not including the polyA padding) is also recorded.

The set of tags of interest are those that are present in the input master tagCount file (using the -t option) or tagsOnPhysicalMap file (using the mutually exclusive -m option). The -t option is generally used, using the output of MergeMultipleTagCountPlugin as the -t option input file.

Our pipeline currently supports only the enzymes ApeKI and PstI (-e option).

We generally leave the -c option (*minimum taxa count within a qseq file for a tag to be output*) at its default value of 1. Filtering of tags based upon the number of taxa they appear in would be better performed at the MergeTagsByTaxaFilesPlugin step, but is not currently implemented. With the default -c option of 1, tags that are in the master tagCount file but are not present in a given qseq file will not be output into the corresponding tagsByTaxa file – this is a good thing, as it saves disk space (no need to store rows containing nothing but zeros).

The multiple tagsByTaxa files produced by this QseqToTBTPugin can be merged into a single master tagsByTaxa file in the next step, MergeTagsByTaxaFilesPlugin.

### MergeTagsByTaxaFilesPlugin

**Summary:** Merges all \*.tbt.bin files present in the input directory and all of its subdirectories.

**Input:** Directory (folder) containing multiple tagsByTaxa (\*.tbt.bin) files (produced by QseqToTBTPugin)

**Output:** Merged tagsByTaxa file (it is best to send this to a separate directory from the input directory)



## D 7.1 Production of materials for improved genotyping training

### Arguments: MergeTagsByTaxaFilesPlugin:

-i	Input directory containing multiple tagsByTaxa (*.tbt.bin) files.
-o	Output file name (should be in a separate directory from the input). We recommend using the extension *.tbt.bin.
-x	Merges tag counts of taxa with identical names. Not performed by default.

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -MergeTagsByTaxaFilesPlugin  
-i ./tbt -o ./mergedTBT/myStudy.tbt.bin -endPlugin -runfork1
```

**Gory Details:** This step merges the separate tagsByTaxa files produced by the QseqToTBTPlugin (or FastqToTBTPlugin) into a single, experiment-wide tagsByTaxa file for all of the flow cell lanes in your experiment. Currently, only the presence or absence of each tag in each taxon is recorded.

The **-x** option (off by default) can be invoked to merge the tag counts of taxa with identical names in the key file but from different flow cells, lanes or barcodes within a lane. However, we recommend leaving in any duplicated taxa for now as they can be used in a later step (GBSHapMapFiltersPlugin or MergIdenticalTaxaPlugin) to check error rates.

### TagsToSNPByAlignmentMTPlugin

**Summary:** Aligns tags from the same physical location against one another, calls SNPs from each alignment, and then outputs the SNP genotypes to a HapMap format file (one file per chromosome).

**Input:** (i) TagsByTaxa file (\*.tbt.bin) indicating the presence or absence of each tag of interest in each taxon; (ii) TagsOnPhysicalMap file (\*.topm.bin) containing genomic position of each tag of interest

**Output:** Directory (folder) containing a HapMap format genotype file (one file per chromosome).

### Arguments: TagsToSNPByAlignmentMTPlugin:

-i	Input TagsByTaxa (*.tbt.bin) file.
-o	Output directory. Defaults to current directory (the default is not recommended – it is best to use a separate directory such as './hapmap/unfilt').
-m	TagsOnPhysicalMap (*.topm.bin) file containing genomic position of tags.
-mnF	Minimum value of F (inbreeding coefficient). Not tested by default.
-mnMAF	Minimum minor allele frequency. Defaults to 0.01. SNPs that pass <i>either</i> the specified minimum minor allele frequency (mnMAF) or count (mnMAC) will be output.
-mnMAC	Minimum minor allele count. Defaults to 10. SNPs that pass <i>either</i> the specified minimum minor allele count

## D 7.1 Production of materials for improved genotyping training

	(mnMAC) or frequency (mnMAF) will be output.
-mnLCov	Minimum locus coverage, <i>i.e.</i> , the proportion of taxa with a genotype. Defaults to 0.1.
-inclRare	Include the rare alleles (3 <sup>rd</sup> or 4 <sup>th</sup> states) at site. These are ignored by default (genotype set to missing).
-inclGaps	Include sites where the major or minor allele is a gap. These sites are ignored by default.
-s	Start chromosome.
-e	End chromosome.

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -TagsToSNPByAlignmentMTPlugin
-i ./mergedTBT/myStudy.tbt.bin -m ./topm/myMasterTags.topm.bin
-o ./hapmap/unfilt -s 1 -e 10 -mnF 0.9 -endPlugin -runfork1
```

**Gory Details:** In this step, a sequence alignment is created for each set of tags that align to the exact same genomic position and strand (where the starting point is defined by the barcode end of the tag) and SNPs are called from each alignment. Tags with multiple or unknown physical genomic positions are not used. The SNP calls from each set of alignments are written to a genotype file in HapMap format, with one HapMap file produced per chromosome. These output HapMap files are named after the input tagsByTaxa file, with “.tbt.bin” replaced with “.c#.hmp.txt”, where # is a chromosome number (*e.g.*, \*.c1.hmp.txt)

If you are working with highly homozygous inbred lines or a selfing species, then be sure to use the **-mnF** (minimum F) option (we suggest setting mnF to 0.9), where ‘F’ means ‘inbreeding coefficient’. In species like maize which contain abundant paralogs (from ancient chromosomal duplications), this can filter out numerous bad SNPs.

The options **-mnMAF** (minimum minor allele frequency) and **-mnMAC** (minimum minor allele count) can be used to filter out SNPs with rare minor alleles that possibly result from sequencing errors. Keep in mind that SNPs that pass *either* of these criteria will be output, so there is no point in having one of them set stringent but the other too lax. If you are working with a biparental family with 1:1 segregation you might try a mnMAF of 0.2 and a highly stringent mnMAC close to your total number of taxa, so that it is irrelevant (in that case, only the mnMAF will matter). With unrelated individuals and no way to test segregation or LD, you might want to try a mnMAF of 0.02 (and a highly stringent mnMAC close to your total number of taxa).

The **-mnLCov** (minimum locus coverage) option can be used to filter out SNPs with very high amounts of missing data from the output. These most likely result from large restriction fragments (>400 bp) that are not amplified as efficiently in the PCR steps of the GBS protocol. The default value mnLCov of 0.1 should suffice for most purposes.

We recommend that you do not invoke the **-inclRare** option, so that 3<sup>rd</sup> and 4<sup>th</sup> allelic states (*i.e.*, triallelic and quadra-allelic SNPs) are ignored (genotypes set to missing). Any 3<sup>rd</sup> and 4<sup>th</sup> allelic states are far more likely to result from sequencing error than biological reality.

## D 7.1 Production of materials for improved genotyping training

Similarly, we recommend that you do not invoke the **-inclGaps** option, so that small indels are not scored. Any 3<sup>rd</sup> and 4<sup>th</sup> allelic states are far more likely to result from sequencing error than biological reality. Because of alignment issues for small indels (multiple equally likely alignments), they can end up being positioned slightly differently in replicate runs of the plugin. Also, because our tags are all 64 bases (or smaller) in length, small indels in the middle of a tag alignment always result in artefactual, compensatory small indels of equal size at or near the end of the tag alignment. However, if you are interested in maximizing marker saturation (for example, for GWAS or for fine-mapping of a QTL), then you might want to invoke `inclGaps`: there will almost certainly be numerous sets of tag alignments that contain no SNPs but do contain a small indel. Note that with `inclGaps` invoked, a three base indel (for example) will be output as three consecutive single base gaps in the HapMap file (plus an additional three artefactual, single base gaps). If the insertion is not present in the reference genome, the three real gaps will all have the same position (the base in the reference genome immediately preceding the insertion). Essentially they are redundant scorings of the same indel.

The HapMap genotype files that we generate save disk space and memory by using single letters to represent phase unknown, diploid genotypes. Heterozygotes are represented by IUPAC nucleotide codes:

A = A/A  
C = C/C  
G = G/G  
T = T/T  
M = A/C  
R = A/G  
W = A/T  
S = C/G  
Y = C/T  
K = G/T  
N = missing data

The “MT” in the name of this plugin indicates that it was initially written to run faster by using multiple threads on multiple CPUs. However, we found that this caused some difficult to trace bugs, so the multiple threading is currently disabled.

Genotypes from tags matching the minus strand of the reference genome are complemented so that they are recorded relative to the plus strand. Hence, all SNPs in the output are relative to the plus strand. For restriction fragment smaller than 128bp, the (plus and minus strand) reads from opposite ends can overlap and assay the same SNPs. Hence, the output of `TagsToSNPByAlignmentMTPlugin` will contain some duplicate SNPs, each with different patterns of missing data. These duplicate SNPs can be merged in the next step of the analysis, with the `MergeDuplicateSNPsPlugin`.

### **MergeDuplicateSNPsPlugin**

**Summary:** Finds duplicate SNPs in the input HapMap file, and merges them if they have the same pair of alleles (not necessarily in the same major/minor order) and if

## D 7.1 Production of materials for improved genotyping training

their mismatch rate is no greater than the threshold specified by **-maxMisMat**. If **-callHets** is on, then genotypic disagreements will be called heterozygotes; otherwise they will be set to missing (callHets is off by default).

**Input:** Input HapMap file(s). Use a plus sign (+) as a wild card character to specify multiple chromosome numbers (each chromosome in a separate file).

**Output:** HapMap files (one per chromosome) in which duplicate SNPs have been merged

**Arguments:** MergeDuplicateSNPsPlugin

-hmp	Input HapMap file(s). Use a plus sign (+) as a wild card character to specify multiple chromosome numbers (each chromosome in a separate file).
-o	Output HapMap file(s). Use a plus sign (+) as a wild card character to specify multiple chromosome numbers (each chromosome in a separate file).
-misMat	Threshold genotypic mismatch rate above which the duplicate SNPs won't be merged. Defaults to 0.05.
-callHets	When two genotypes at a replicate SNP disagree for a taxon, call it a heterozygote. Defaults to false (=set to missing).
-s	Start chromosome. Default 1.
-e	End chromosome. Default 10.

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -MergeDuplicateSNPsPlugin
-hmp ./hapmap/unfilt/myStudy.c+.hmp.txt
-o ./hapmap/mergedSNPs/myStudy.mergedSNPs.c+.hmp.txt -misMat 0.1 -callHets
-s 1 -e 12 -endPlugin -runfork1
```

**Gory Details:** This step should be run directly after TagsToSNPByAlignmentMTPlugin, using the HapMap file(s) from that step as input.

If the germplasm is not fully inbred, and still contains residual heterozygosity (like the maize NAM or IBM populations do) then **-callHets** should be on and **-maxMisMat** should be set fairly high (0.1 to 0.2, or even higher, depending on the amount of heterozygosity). Because the sequencing coverage is usually less than 1x, most of the time only one allele at a heterozygous SNP will be detected (particularly for ApeKI). Hence, duplicate SNPs genotypes from a true heterozygote may disagree simply because different alleles were sampled by the duplicate assays. Hence, these disagreements are not necessarily errors, and should not necessarily be used to prevent duplicate SNPs from being merged (unless your germplasm is highly inbred, with very little residual heterozygosity).

Indels (gaps) are ignored by this plugin: it makes no attempt to merge apparent duplicate gaps with the same chromosomal position.

### GBSHapMapFiltersPlugin

**Summary:** Reads HapMap format genotype files (one per chromosome), and filters out SNPs with low taxon coverage (missing data at most taxa), high heterozygosity, low (and/or high) minor allele frequency, and (optionally) maximum LD of less than  $R^2$  of 0.5. Taxa with low SNP coverage (missing data at most SNPs) are also removed. All cutoffs are adjustable except for the LD cutoff.

## D 7.1 Production of materials for improved genotyping training

**Input:** (i) Key file; (ii) Directory (folder) containing qseq files

Output: Directory (folder) containing a corresponding tagCount file for every qseq file in the input directory.

**Arguments:** GBSHapMapFiltersPlugin

-hmp	Input HapMap file. Use a plus sign (+) as a wild card character to specify multiple chromosome numbers (each chromosome in a separate file).
-o	Output HapMap file. Use a plus sign (+) as a wild card character to specify multiple chromosome numbers (each chromosome in a separate file).
-mnTCov	Minimum taxon coverage, i.e. the minimum number of taxa that must contain each tag. Defaults to .1.
-mnScov	Minimum presence, i.e. the fraction of sites that must be present in a given taxon. Defaults to .1.
-mnF	Minimum value of F (inbreeding coefficient). Not tested by default.
-mnMAF	Minimum minor allele frequency Default is 0.0, no filtering.
-mxMAF	Maximum minor allele frequency. Default is 1.0, no filtering.
-hLD	Specifies that samples should be filtered for high LD.
-sC	Start chromosome. Default 1.
-eC	End chromosome. Default 10.

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -GBSHapMapFiltersPlugin
-hmp ./hapmap/mergedSNPs/myStudy.mergedSNPs.c+.hmp.txt
-o ./hapmap/filt/myStudy.mergedSNPs.filt.c+.hmp.txt -hLD -mnTCov 0.05
-mnSCov 0.05 -sC 1 -eC 12 -endPlugin -runfork1
```

**Gory Details:** If your study germplasm are from a single biparental cross, then the **-hLD** (high LD) filter (off by default) can be very useful to filter out bad SNPs with high genotyping error or incorrect physical genomic positions.

BiParentalErrorCorrectionPlugin

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -BiParentalErrorCorrectionPlugin
-endPlugin -runfork1
```

### BinaryToTextPlugin

**Summary:** Reads a Binary GBS File and outputs the equivalent text file.

**Input:** Binary File

**Output:** Text File

**Arguments:** BinaryToTextPlugin

-i <filename>	Input Binary File Name.
-o <filename>	Output Text File Name.

## D 7.1 Production of materials for improved genotyping training

-t <type>	Type of input file (TOPM, TBTBit, TagCounts).
-----------	---

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -BinaryToTextPlugin -i  
/Users/terry/users/james/rice.topm.bin -o rice_topm_bin.txt -t TOPM  
-endPlugin -runfork1
```

```
/programs/tassel/run_pipeline.pl -fork1 -BinaryToTextPlugin -i  
/Users/terry/users/james/rice.tbt.bin -o rice_tbt_bin.txt -t TBTBit  
-endPlugin -runfork1
```

```
/programs/tassel/run_pipeline.pl -fork1 -BinaryToTextPlugin -i  
/Users/terry/users/james/rice.cnt -o rice_cnt.txt -t TagCounts  
-endPlugin -runfork1
```

### TextToBinaryPlugin

**Summary:** Reads a Text GBS File and outputs the equivalent binary file.

**Input:** Text File

**Output:** Binary File

**Arguments:** BinaryToTextPlugin

-i <filename>	Input Text File Name.
-o <filename>	Output Binary File Name.
-t <type>	Type of input file (TOPM, TBTBit, TagCounts).

Example command:

```
/programs/tassel/run_pipeline.pl -fork1 -TextToBinaryPlugin -i  
/Users/terry/users/james/rice_topm_bin.txt -o rice.topm.bin -t TOPM  
-endPlugin -runfork1
```

```
/programs/tassel/run_pipeline.pl -fork1 -TextToBinaryPlugin -i  
/Users/terry/users/james/rice_tbt_bin.txt -o rice.tbt.bin -t TBTBit  
-endPlugin -runfork1
```

```
/programs/tassel/run_pipeline.pl -fork1 -TextToBinaryPlugin -i  
/Users/terry/users/james/rice_cnt.txt -o rice.cnt -t TagCounts  
-endPlugin -runfork1
```

### Online resources:

- Glaubitz J, Harriman J, Casstevens T: TASSEL 3.0 Genotyping by Sequencing (GBS) pipeline documentation;  
<https://biohpc.cornell.edu/lab/doc/TasselPipelineGBS20120215.pdf>



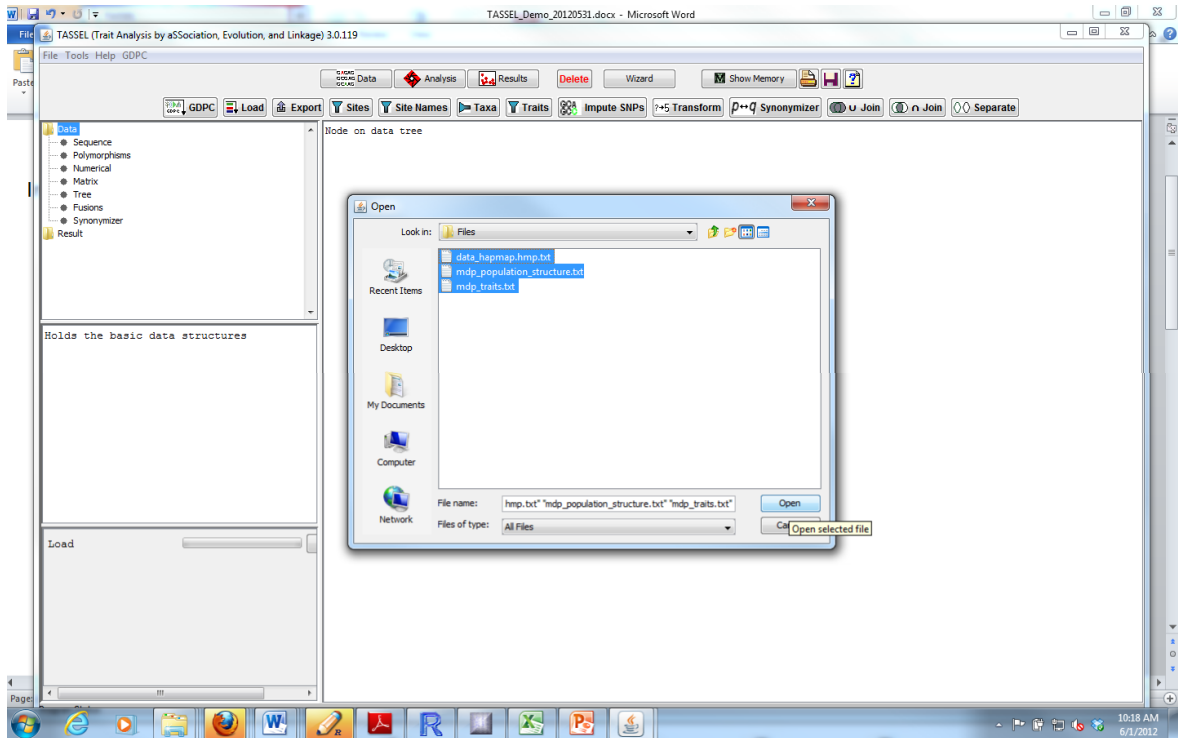
## D 7.1 Production of materials for improved genotyping training

### 32. Tutorial II: Association Mapping in Tassel

[www.maizegenetics.net/tassel](http://www.maizegenetics.net/tassel)

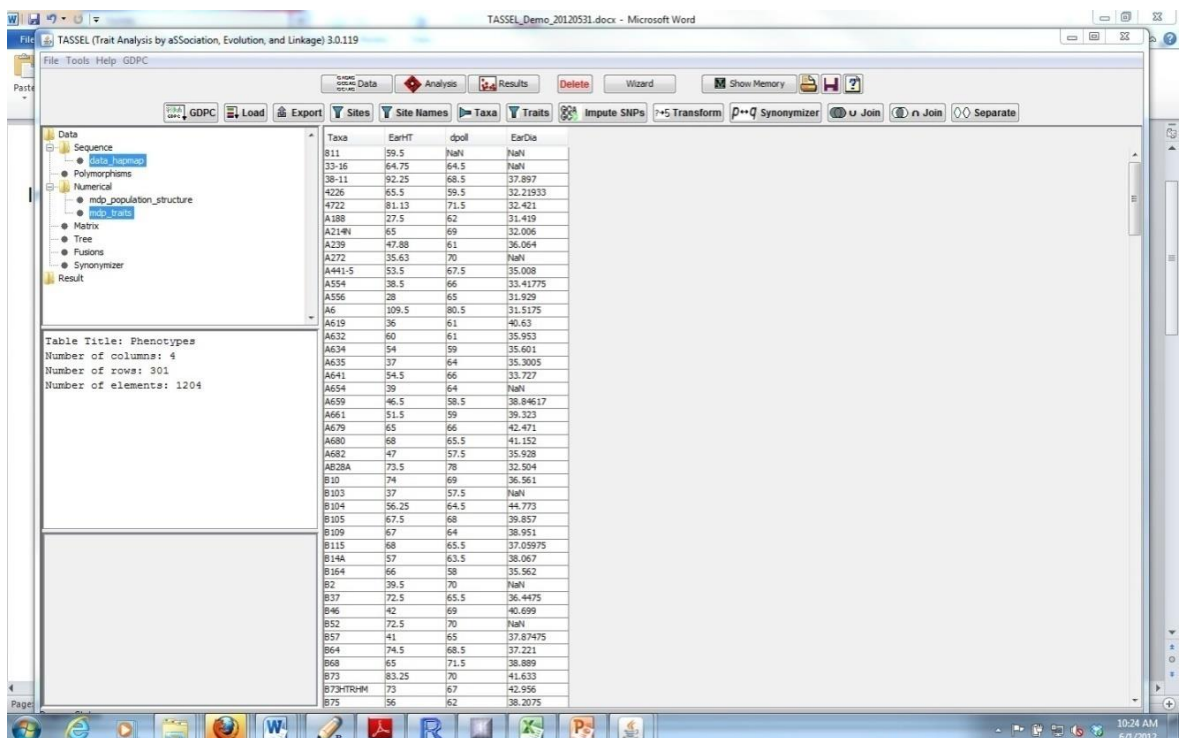
#### 32.1. Reading in data

Data > Load >



#### 32.2. Fit a generalized linear model (GLM)

Select "mdp\_traits" and "data\_hapmap", and join them together



## D 7.1 Production of materials for improved genotyping training

Select “mdp\_traits+data\_hapmap”, then Analysis > GLM > OK

The screenshot shows the TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.119 software interface. The main window displays a list of taxa and their associated data. A dialog box titled "GLM Options" is open, showing the "Use permutation test for markers" checkbox and the "Number of Permutations" set to 1000. The background window shows a table of taxa with columns for EarHT, dpoll, EarDia, and Haplotype.

Taxa	EarHT	dpoll	EarDia	Haplotype
33-16	64.75	64.5	NaN	CCGTGTCTC...
38-11	92.25	68.5	37.897	CCGTGTCTC...
4226	65.5	59.5	32.21933	CCGTGTCTC...
4722	81.13	71.5	32.421	CCGTGTCTC...
A188	27.5	62	31.419	ACGTGTCTC...
A214N	65	69	32.006	CCTAGACT...
A239	47.88	61	36.064	ACTTAAGTC...
A272	35.63	70	NaN	ACTTAAGTC...
A441-5	53.5	67.5	35.008	CCGTGTCTC...
A554	38.5	66	33.41775	CGTTATCTC...
A556	28	65	31.929	CCGTGTCTC...
A6	109.5	80.5	31.5175	ACTTAAGTC...
A619	36	61	40.63	CCGTGTCTC...
A632	60	61	35.953	CCTAGACT...
A634	54	59	35.601	CCTAGACT...
A635	37	64	35.3005	CCTAGACT...
A641	54.5	66	33.727	ACTTAAGTC...
A654	39	64	NaN	ACTTAAGTC...
A659	46.5	58.5	38.84617	AGTT
A661	51.5	59	39.323	CGTT
A679	65	66	42.471	CCTA
A680	68	65.5	41.152	CCTA
A682	47	57.5	35.928	CCTA
A828A	73.5	78	32.504	AGTA
B10	74	69	36.561	ACGT
B103	37	57.5	NaN	ACGT
B104	56.25	64.5	44.773	CCTT
B105	67.5	68	39.857	CCTTATCTC...
B109	67	64	38.951	CCGTGTCTC...
B115	68	65.5	37.05975	CCGTGTCTC...
B14A	57	63.5	38.067	CCTAGACT...
B164	66	58	35.562	CGTTATCTC...
B2	39.5	70	NaN	ACGTGTCTC...
B37	72.5	65.5	36.4475	ACGTGTCTC...
B46	42	69	40.699	CCGTGTCTC...
B52	72.5	70	NaN	CCTTGTCTC...
B57	41	65	37.87475	CGTTATCTC...
B64	74.5	68.5	37.221	CCGTGTCTC...
B68	65	71.5	38.889	CCTAGACT...
B73	83.25	70	41.633	CCTAAGTC...
B73HTRH	73	67	42.956	CCTAAGTC...
B75	56	62	38.2075	CCTAAGTC...
B76	53	61.5	35.33575	ACGTGTCTC...

## 32.3. Make a Manhattan and QQ-Plot

Select GLM model results, then Results > Manhattan Plot/QQ Plot

The screenshot shows the TASSEL software interface with the results of a GLM analysis. The main window displays a Manhattan Plot titled "P-Values by Chromosome for dpoll". The plot shows -Log(P-Value) on the y-axis (ranging from 0.0 to 12.5) and Position on the x-axis (ranging from 0 to 2,000,000,000). The plot is color-coded by chromosome (1-10). A legend at the bottom indicates the color coding for chromosomes 1 through 10. To the right, a QQ Plot titled "Expected -Log(P-Value) vs. -Log(P-Value)" is displayed, showing a diagonal line representing the expected distribution of p-values. Below the plots, a table of results is visible, showing the association between markers and traits.

Marker	EarHT	dpoll	EarDia	Haplotype
9225	245	16.47584	1	2.59225
242	16.57678	0	NaN	
698	235	16.67336	1	3.08698
4412	245	16.33441	1	37.24412
57403	237	15.61917	2	68.57403
97155	241	16.20564	2	27.97155
924	229	15.95274	2	17.0924
447	245	16.36506	2	7.25447
28	223	17.06721	1	4.4228
262	246	16.36832	1	3.10262
9559	230	16.24645	2	63.89559
8191	221	15.80922	1	132.8191
4703	243	16.3595	1	49.24703
9529	240	16.58771	2	2.99529
3385	244	15.90072	1	52.2385
11146	241	16.28879	2	14.11146
809	240	16.66936	2	2.3309
809	246	16.44139	2	2.39809
562	245	16.30265	1	50.7562
232	16.52332	1	9.47482	
3385	243	16.50804	1	6.37877
5424844	242	16.33894	1	54.24844
142.40177	242	15.97467	1	142.40177

## D 7.1 Production of materials for improved genotyping training

### 32.4. Calculate a kinship matrix

Select "data\_hapmap", then Analysis > Kinship > Run

The screenshot shows the TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.119 software interface. The 'Analysis' menu is open, and 'Kinship' is selected. A 'Kinship Options' dialog box is displayed, with 'Model heterozygotes as' set to 'Related to homozygotes' and 'Rescale results between 2 and 0' checked. The main window displays a grid of DNA sequence data for various sites (e.g., 441, 882, 1323, 1764, 2205, 2646, 3087) across different individuals (e.g., 33-16, 4226, 4722, A188, A2149, A239, A272, A441-5, A554, A556, A6, A619, A632, A634, A635, A650, A652, A654, A659, A661, A679, A680, A682, A628A, B10, B103, B104, B105, B109, B115, B14A).

### 32.5. Run PCA

Select "data\_hapmap", then Data > Transform > Create Dataset

The screenshot shows the TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.119 software interface. The 'Data' menu is open, and 'Transform' is selected. A 'Create Dataset' dialog box is displayed, with 'Collapse Non Major Alleles' selected. The main window displays the same DNA sequence data grid as in the previous screenshot.

Select "data\_hapmap\_Collapsed", then Data > Transform > Impute > Create Dataset



## D 7.1 Production of materials for improved genotyping training

The screenshot shows the TASSEL software interface with the 'PCA' dialog box open. The dialog box has three tabs: 'Trans', 'Impute', and 'PCA'. The 'PCA' tab is active, showing the following options:

- Method:  Manhattan Distance,  Euclid Distance,  Unweighted Average,  Weighted Average
- Number of Neighbors (K): 3
- Min. Freq. of Row Data: 0.80

The background shows a data table with columns labeled S0 through S11 and rows labeled with taxon IDs like 33-16, 38-11, 50, 51, etc. A 'Transform' window is also visible in the foreground, showing a list of columns and their corresponding 'Percent Missing Data' values.

Select "data\_hapmap\_Collapsed\_3093\_imputed", then Data > Transform > PCA > Create Dataset

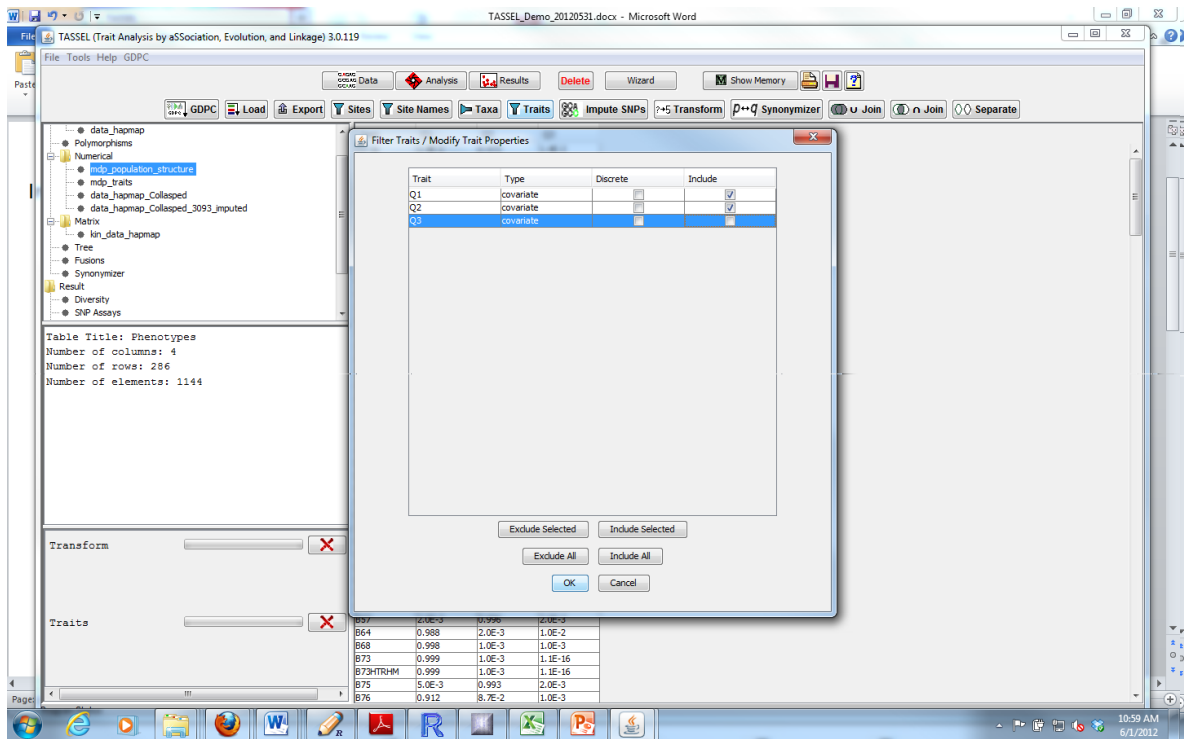
The screenshot shows the TASSEL software interface with the 'PCA' dialog box open. The 'PCA' tab is active, showing the following options:

- Method:  Correlation,  Covariance
- Output:  Eigenvalue  $\geq 0$ ,  Var Prep %  $\geq 0.00032$ ,  Components = 3,093

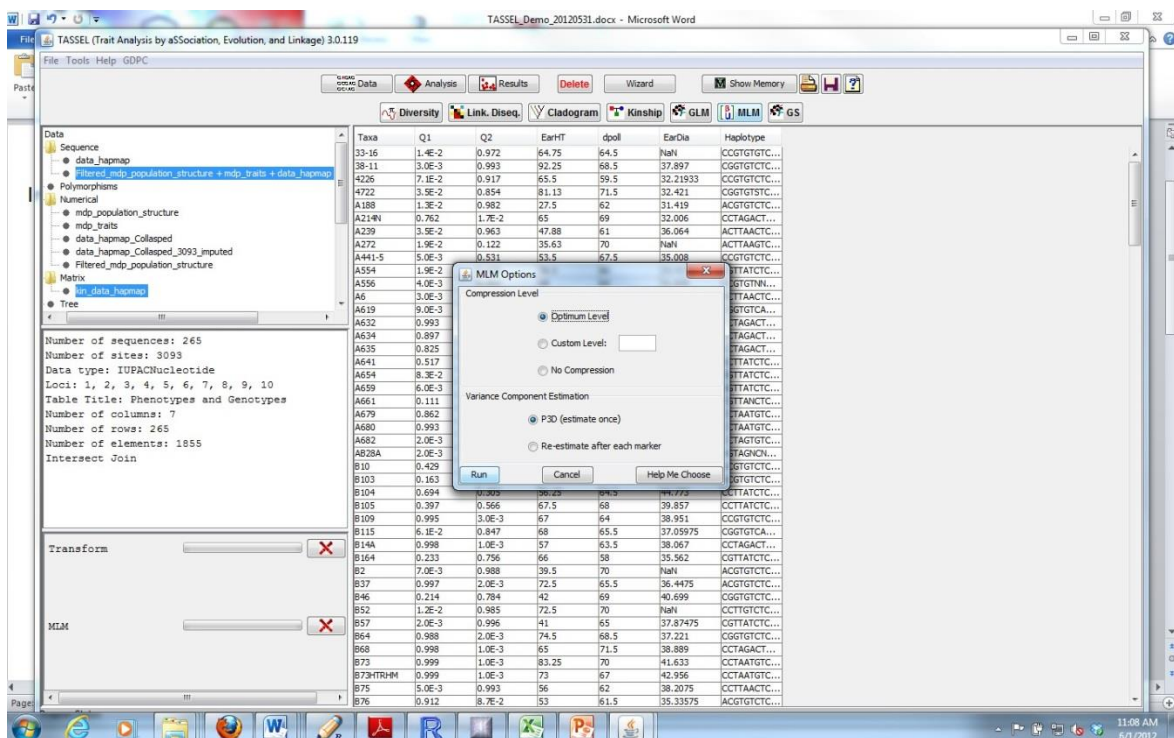
The background shows the same data table as the previous screenshot. The 'Transform' window in the foreground now shows that the 'Percent Missing Data' for all columns is 0.00. The 'Table Title' is 'Phenotypes', with 3094 columns and 276 rows.

## D 7.1 Production of materials for improved genotyping training

### 32.6. Run MLM

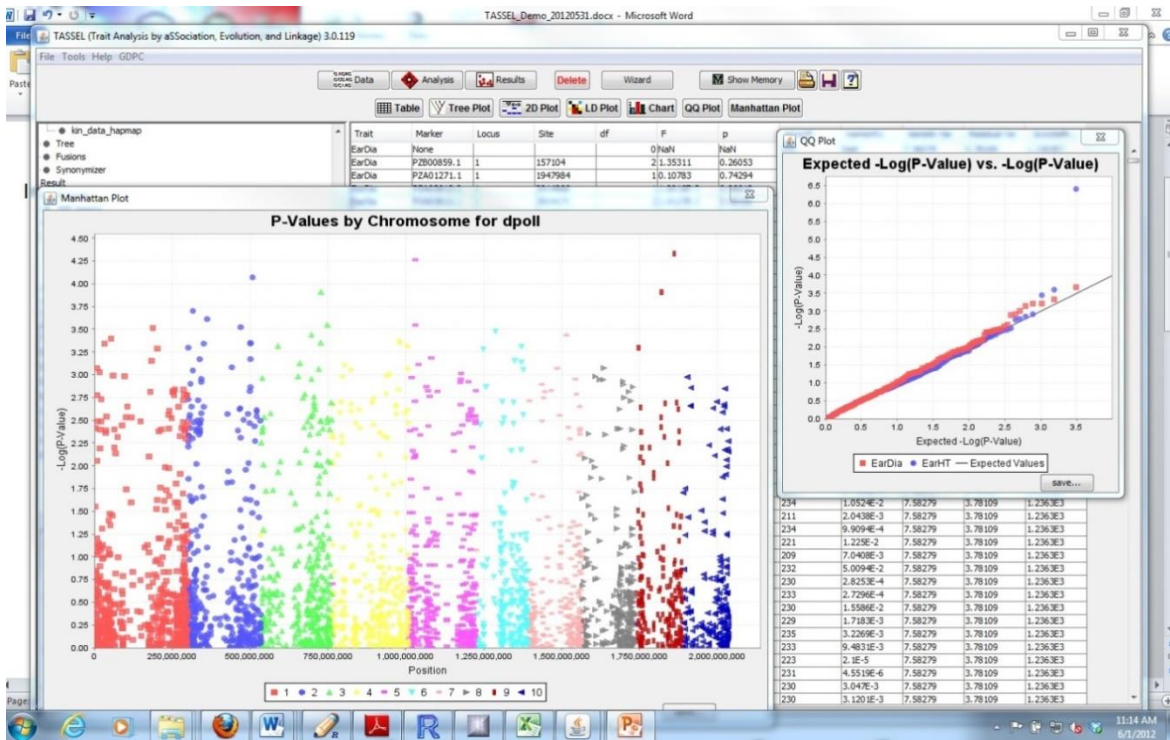


Select "mdp\_population\_structure", then Data > Traits > Deselect Q3 > OK  
 Join "mdp\_traits", "data\_hapmap", and "Filtered\_population\_structure" Select merged file and "kin\_data\_hapmap", then Analysis > MLM > Run



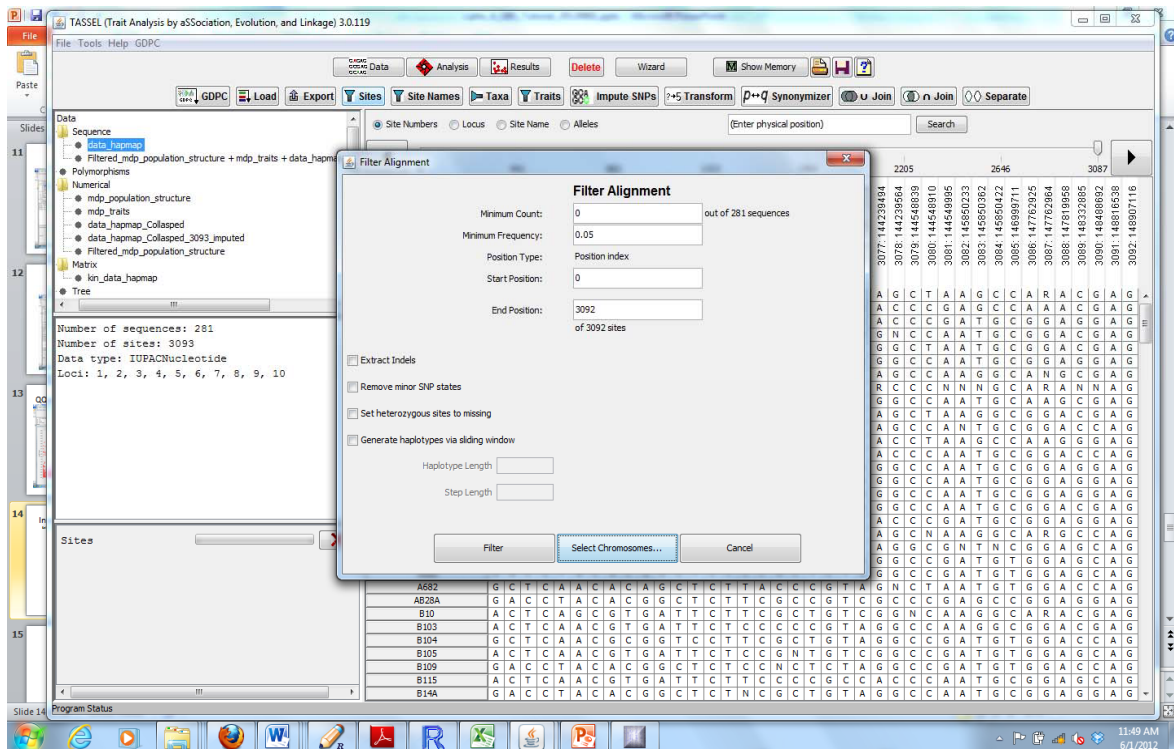
# D 7.1 Production of materials for improved genotyping training

## 32.7. QQ and Manhattan Plot from MLM



## 32.8. Investigate LD on Chromosome 1

Select "data\_hapmap", then Data > Sites > Select Chromosomes

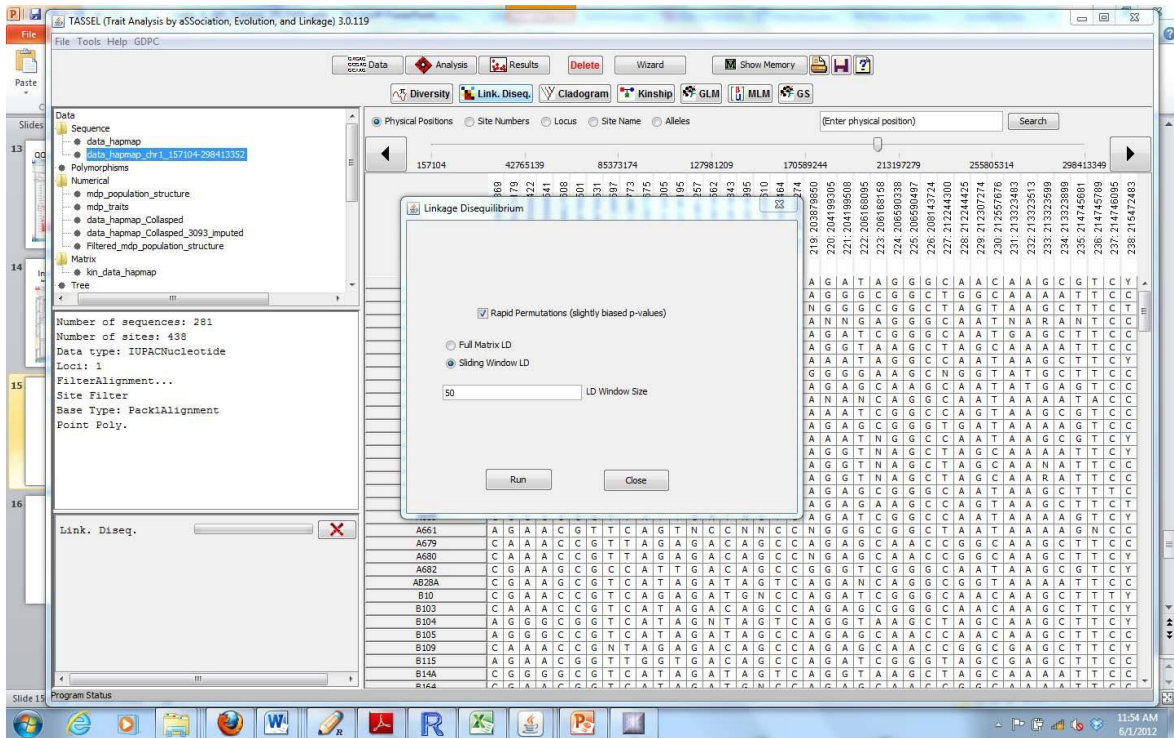




# D 7.1 Production of materials for improved genotyping training

## 32.9. Make an LD Plot

Select the Chr1 filtered data, then Analysis > Link. Diseq. > Run



Select the LD data, then Results > LD Plot

