

1 Next-generation sequencing accelerates crop gene discovery

2

3 Khanh Le Nguyen^{1,2}, Alexandre Grondin¹, Brigitte Courtois^{3,4} and Pascal
4 Gantet^{1,*}

5

6 ¹Université de Montpellier, Institut de Recherche pour le Développement, UMR
7 DIADE, 911 Avenue Agropolis, 34394 Montpellier cedex 5, France

8 ²LMI RICE 2, AGI, Km2 Pham Van Dong, Tu Liem, Hanoi, Vietnam

9 ³CIRAD, UMR AGAP, F-34398 Montpellier, France

10 ⁴Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

11

12 *Correspondence: pascal.gantet@umontpellier.fr (P. Gantet).

13

14

15

16

17 Key words: bulk-segregant analysis; candidate gene; crops; genetics; next-
18 generation sequencing; quantitative trait loci.

19

20

21

22 **Abstract:**

23 The identification and isolation of genes underlying quantitative trait loci (QTLs)
24 associated with agronomic traits in crops has been recently accelerated thanks to
25 next-generation sequencing (NGS)-based technologies combined with plant
26 genetics. With NGS, different revisited genetic approaches, which benefited from
27 higher marker density, have been elaborated. These approaches improved
28 resolution in QTL position and assisted in determining functional causative
29 variations in genes. Examples of QTLs/genes associated with agronomic traits in
30 crops and identified using different strategies based on whole- genome
31 sequencing/resequencing or RNAseq are presented and discussed in this review.
32 More specifically, we will summarize and illustrate how NGS boosted bulk
33 segregant analysis, expression profiling and the construction of polymorphism
34 databases to facilitate the detection of QTLs and causative genes.

35

36 **How NGS boosts QTL and gene determination.**

37 In molecular genetics, quantitative traits are first decomposed in their
38 Mendelian components by **quantitative trait loci (QTL)** analysis. Then, each
39 QTL is fine-mapped or cloned individually. Thousands of QTLs (see Glossary)
40 associated with agronomic traits were found in crops and represent a reservoir of
41 alleles for breeders to create improved varieties [1–4]. *SUBMERGENCE 1*
42 (*SUB1*), a major QTL which confers tolerance to submergence in rice (*Oryza*
43 *sativa*), is probably one of the most successful examples of QTL utilization
44 worldwide [5]. This QTL with large effect was identified in a traditional rice variety,
45 and the underlying gene that is absent from the genome of the reference rice
46 variety, was cloned. The favorable allele of the *SUB1* gene was introgressed into
47 elite cultivars by marker-assisted backcrossing and the improved products were
48 released in several Asian countries. However, very few QTLs were as
49 successfully used in marker-assisted selection (MAS) because they were
50 positioned with insufficient precision, because they explained a low proportion of
51 the trait variation, or because QTL x environment interactions made them

52 useless outside their detection context [6,7]. Before undertaking MAS, one
53 challenge is to reduce the confidence interval (CI) of the QTL position to make
54 the introgressed segment carrying the QTL as small as possible and to avoid
55 possible undesirable side effects due to the other genes carried by the
56 introgressed segment. To reduce the CI of a QTL position, a possible approach
57 is to undertake a meta-analysis of different studies targeting the same trait in the
58 same species [8]. A QTL meta-analysis was effectively applied to different crops
59 to refine the CI regions and seek candidate genes [9–12]. QTLs positioned on a
60 single consensus map and narrowed down by meta-analysis enable the target
61 regions of interest for MAS to be more precisely identified. However, depending
62 on the size of the meta-QTLs, further steps of either fine mapping and positional
63 cloning or association mapping with sufficient marker density are often necessary
64 to identify the shortest target DNA fragment responsible for the phenotypic
65 variation [12,13].

66 **Next-Generation Sequencing** (NGS) designates new sequencing
67 methods (see *Box*) that produce high coverage with lower cost and higher speed
68 than traditional SANGER sequencing [14,15]. Among NGS platforms,
69 genotyping-by-sequencing (GBS), which is a high-throughput sequencing
70 approach, has remarkably increased the number of molecular markers usable in
71 crop genetics [16,17]. The basic features of GBS rely on using restriction
72 enzymes to reduce genome complexity and barcode adapters that allow
73 sequencing of pooled samples. The choice among GBS methods is generally
74 based on the genome size of the studied crop, the extent of linkage
75 disequilibrium and level of heterozygosity of the studied panel, and cost-
76 efficiency considerations. Unlike the earlier low-throughput approaches based on
77 restriction fragment length polymorphism (RFLP) or simple sequence repeats
78 (SSR), GBS enables the identification and genotyping of a massive quantity of
79 single nucleotide polymorphisms (SNPs). These SNPs can be associated with
80 agronomical traits of interest and then used in marker-assisted breeding or to
81 validate trait-linked haplotypes in crops [17][18][19]. This strategy has been
82 successfully used in many important crops [20]. For instance, GBS methods

83 were employed to genotype recombinant inbred line (RIL) populations in rice
84 [20,21], maize and barley [22] and doubled-haploid (DH) populations in wheat
85 [23] in view of QTL mapping. GBS was also applied to provide adequate marker
86 density for **genome-wide association study** (GWAS) of rice traditional
87 populations [24], rice and chickpea (*Cicer arietinum*) multiparent advanced
88 generation intercrosses (MAGIC) [25] and maize (*Zea mays*) nested association
89 mapping populations (NAM) [26].

90 Recently, a shift occurred towards **whole-genome resequencing** (WGR),
91 an approach in which the entire genome of different genotypes is sequenced and,
92 then, compared to a known reference sequence. WGR allows the detection not
93 only of SNPs, but also of insertions-deletions (InDels) and structural variants [27].
94 In addition, alternative approaches targeted to specific parts of the genome such
95 as **RNA-sequencing** (RNAseq) and exome-sequencing have also been
96 developed, allowing scientists to go further in the discovery of the SNPs altering
97 coding sequences [28].

98 In this review, we will summarize genetic approaches combined with
99 NGS-based methods that have been recently developed to speed up the
100 detection of QTLs and their causative genes and their utilization in molecular
101 breeding.

102

103 **Approaches to improve QTL and candidate gene detection**

104 **Bulk segregant analysis (BSA)** represents a simple, effective and cost-saving
105 QTL mapping strategy compared with conventional QTL mapping that requires
106 genotyping and phenotyping of an entire mapping population [29]. In BSA, two
107 bulks of segregant individuals derived from biparental populations (F_2 ,
108 **recombinant inbred lines** (RILs), or **doubled haploids** (DH)), multiparental
109 populations (NAM or MAGIC), natural populations or mutant libraries, are created
110 by pooling DNA from individuals with extreme phenotypic values for the traits of
111 interest [30]. Markers from a genomic region linked to the trait are expected to
112 show a distinct allele frequency between the two bulks, while markers from a
113 region unlinked to the trait will show a similar allele frequency in the two bulks

114 [31,32]. The minimum size of the bulks is determined by the frequency with which
115 unlinked loci might be detected as polymorphic between the bulk samples [29].
116 The smaller the bulk size, the higher the risk of false positives. For example, for
117 a SNP segregating in an F_2 population, the probability of a bulk of n individuals
118 having all the same allele and a second bulk of equal size having all the other
119 allele is $2(1/4)^n (1/4)^n$ when the locus is unlinked to the target gene. With 5
120 individuals in each bulk, this probability is $1.90e^{-06}$ while with 10 individuals, this
121 probability decreases to $1.89e^{-12}$. However, because the phenotype of the
122 individuals composing the bulks should be indisputable and because the
123 confirmation step requires, on a second time, to test individually these plants, it is
124 advisable not to use too large bulks. Bulks of 10-15 plants are commonly used. In
125 BSA, the whole population has to be phenotyped to identify individuals in the tails
126 of the distribution and the method is therefore better suited for traits easy and
127 inexpensive to phenotype. To date, SNPs are the markers of choice for linkage
128 analysis in many crops because of their high density in the genomes and their
129 codominant nature [19][33]. Recent NGS-based methods such as WGR can be
130 efficiently used to determine SNPs between parents of a mapping population [34].
131 Therefore, WGR coupled with a BSA approach provides a coverage of dense
132 informative SNP markers to detect QTLs in mapping populations.

133

134 *QTL-seq approaches*

135 A first example of such an approach is QTL-seq (Figure 1A), which is a modern
136 version of the classical BSA combined with WGR [35]. In this approach, a
137 mapping population derived from a cross between two contrasted parents is used.
138 The progenies are phenotyped, and the tails of the distribution are divided into
139 two extreme bulks of 10-20 individuals, which are sequenced at above 6x
140 coverage. For each genomic position, the proportion of short reads harboring
141 SNPs with the sequence of one of the parents chosen as reference (so-called
142 SNP-index) is estimated and the difference between the SNP-index of the low
143 trait-bulk and that of the high trait-bulk, called $\Delta(\text{SNP-index})$, is calculated. A
144 large $\Delta(\text{SNP-index})$ characterizes the genomic fraction that has an association

145 with the phenotypic value [36]. In chickpea, QTL-seq was applied to two 100-
146 seed weight (SW)-contrasted bulks, each bulk containing 10 F₄ homozygous
147 individuals, which had been produced by single-seed descent (SSD) from a cross
148 between high SW and low SW landraces [37]. A major QTL was detected on
149 chromosome 1. One SNP tightly linked with the SW-QTL was further identified in
150 the coding region of the constitutive photomorphogenic9 (*COP9*) signalosome
151 complex subunit 8 (*CSN8*) gene. This gene was specifically expressed in seeds
152 and was up/downregulated during seed development in high/low SW parent and
153 homozygous mapping individuals, respectively. Moreover, a functional molecular
154 diversity analysis showed that the coding SNP was completely absent from wild
155 accessions while it discriminated the cultivated genotypes, the high and low SW
156 parents and the two bulk mapping individuals. Therefore, QTL-seq combined with
157 differential expression profiling and diversity analysis proved to be efficient not
158 only in scaling-down QTL size, but also in rapidly enabling potential candidate
159 gene identification. The same approach has been successfully used in other
160 crops, such as foxtail millet (*Setaria italica*) [45] and rice [49].

161 Another approach derived from QTL-seq is multiple QTL-seq (mQTL-seq),
162 which can be defined as QTL-seq applied to several mapping populations
163 derived from crosses with at least one common parent (Figure 1 B) [38]. The
164 utilization of multiple mapping populations representing a broader genetic
165 diversity was beneficial for the validation of QTLs, along with narrowing down the
166 detected QTLs to shorter segments for several agronomic traits in chickpea, such
167 as pod number per plant (PN) [38] or plant height [39]. For example, mQTL-seq
168 applied independently to two F₅ mapping populations of chickpea allowed the
169 identification of common significant genomic regions. For each population, two
170 bulks of 10 lines with low/high PN were built. Two major QTLs associated with
171 PN that were previously detected using the entire population were scaled down:
172 *CaqaPN4.1* from 868 kb to 638 kb and *CaqbPN4.2* from 1.8 Mb to 1.3 Mb.
173 Furthermore, mQTL-seq identified a regulatory SNP governing PN in the
174 pentatricopeptide repeat (*PPR*) gene. A gene expression study demonstrated
175 that the *PPR* gene was strongly upregulated in the high-PN bulks and the high-

176 PN parent of the two mapping populations during pollen and pod development
177 [38].

178

179 *MutMap approaches*

180 MutMap approaches combine NGS with BSA in the analysis of a mutated
181 population. Mutagenesis is a classic way to produce material useful in
182 determining the function of a candidate gene. Where QTL-seq uses two
183 contrasted bulks of individuals from any mapping population, MutMap (Figure
184 1C) is a method based on WGR using bulked segregants which are derived from
185 cross between a homozygous recessive mutant and its wild-type parental line
186 [30,40]. The F₂ population is phenotyped and only plants showing the recessive
187 mutant phenotype are bulked. The parental genome sequence is used as the
188 template to detect causal SNPs underlying the mutant phenotype. As with QTL-
189 seq, a SNP-index is computed for each SNP position. MutMap is actually a
190 simplified version of QTL-seq with only the mutant-phenotype bulk sequenced
191 and no possibility to distinguish **segregation distortions** from a true QTL effect.
192 It works only if the mutant allele is recessive and if the mutant phenotype can be
193 easily distinguished from the wild phenotype in F₂ plants. It is applicable in cases
194 of crosses between a mutant and its wild-type progenitor rather than crosses
195 between genetically distant lines. MutMap should probably also be avoided when
196 targeting mutations with small or subtle effects. This method was recently used to
197 isolate mutations causing pale green leaves and semidwarfism in rice [40], and
198 the many-noded dwarf (*mnd*) in barley [40,41]. MutMap was also successfully
199 used to identify the causative gene, *OsRR22*, from a salt-tolerant rice mutant
200 called *hitomebore salt tolerant 1 (hst1)*. Subsequently, the introgression of the
201 *hst1* allele into the elite cultivar Hitomebore by successive backcrosses enabled
202 the release of the improved variety Kaijin, which differed from Hitomebore wild
203 type by only 201 SNPs but had the same salt tolerance as the *hst1* mutant. With
204 the application of MutMap, the new salt-tolerant elite variety Kaijin was
205 developed in only two years and contributed to the restoration of rice production
206 in tsunami-affected areas of Japan [42].

207 An extended version of MutMap, MutMap+ (Figure 1 E), allows the
208 identification of causal mutations without having to cross a mutant and its wild-
209 type parental line. This approach especially suits mutants with early stage
210 lethality or sterility and species for which efficient techniques for crossing are not
211 available. In MutMap+, only plants of the second mutant generation (M_2) that are
212 heterozygous for the mutation are used. To identify those plants, each individual
213 M_2 plants is selfed to obtain M_3 seeds and the segregation of each M_3 progeny
214 (expected to be 3:1 if the M_2 plant was heterozygous) is assessed. The selected
215 M_3 progenies are further analyzed to confirm that the mutation is caused by a
216 single recessive mutation, then two bulks are constituted, the mutant bulk (MB)
217 and the wild-type bulk (WTB). The two bulks are sequenced and a SNP-index is
218 calculated as in the MutMap approach. Although a SNP-index equal to 1 can be
219 caused by irrelevant homozygous SNPs fixed in M_2 , it is possible to detect the
220 true region harboring the causal mutation by comparing SNP-index plots of the
221 wild-type and mutant bulks. Causative SNPs are specific to the mutant bulk.
222 Using MutMap+, causal mutations leading to an early stage lethality in rice
223 seedling were rapidly identified by WGR of a segregating M_3 generation [43].

224 The wild type parental line is often different from the reference sequenced
225 variety. However, MutMap or MutMap+ are inadequate to detect valuable SNPs
226 that are located in the unmapped regions between a wild-type genome and a
227 reference genome. For such situations, MutMap-Gap (Figure 1D) is better suited.
228 MutMap-Gap is a MutMap approach that includes a *de novo* genome sequence
229 assembly to determine SNPs in a specific parental genome region missing in the
230 reference genome. Using mutant lines that were susceptible to a strain attacking
231 the blast resistance gene *Pii*, MutMap-Gap revealed the existence of the *Pii* gene
232 in the rice variety Hitomebore. This gene was absent from the Nipponbare
233 reference sequence [44].

234

235 *Figure 1*

236

237 **NGS-assisted expression profiling**

238 NGS-assisted expression profiling identifies candidate genes having transcripts
239 linked with the phenotype of interest. The availability of NGS-based
240 transcriptome-wide tools provides precise information about the abundances of
241 gene transcripts [45]. Gene expression analysis is a method that has been
242 frequently used to screen among candidate genes underlying a QTL in different
243 crops, e.g., chickpea, potato, and rice [46–48]. A gene becomes a causative
244 candidate when evidence coming from QTL mapping coincides with transcription
245 activities in the conditions where the phenotype of interest was observed. In this
246 context, the expression level of the causative candidate gene correlates with the
247 phenotypic value. Recently, RNA-seq, the direct sequencing of complementary
248 DNA (cDNA) derived from RNA extracts, has been used to cater comprehensive
249 expression profiling of QTL genes in different tissues and organs of contrasted
250 genotypes [49]. RNA-seq provides a global view of the protein-coding regions
251 that only occupy 1-2% of the genome but include many functional variations [50].
252 RNA-seq is an exceptional method to overcome the limitations of previous
253 expression microarrays in which the dissection of different transcripts was
254 dependent on probes designs [51][52]. For example, RNA-seq was performed on
255 the sorghum root tissues of two sorghum (*Sorghum bicolor*) varieties used as
256 parents of a mapping population and revealed that 108 gene transcripts involved
257 in nitrogen metabolism, plant hormone metabolism and glycolysis were
258 differentially expressed. These genes were located in the vicinity of QTLs
259 detected in the mapping population that regulated multiple agronomic traits under
260 normal and low nitrogen conditions [53]. In maize, published RNA-seq data
261 combined with meta-QTL analysis facilitated the identification of candidate genes
262 involved in kernel row number [54]. In soybean (*Glycine max*), RNA-seq
263 contributed to the identification of a novel salt-tolerance gene from a highly salt
264 tolerant wild accession. A combination of two approaches (*de novo* sequencing
265 of the wild accession and QTL mapping in a population derived from a cross
266 between the wild accession and a cultivated one) was used. The results were
267 validated using resequencing data from 23 soybean accessions with contrasted
268 levels of salinity tolerance. *GmCHX1* was identified as the causal gene and

269 shown to encode an ion transporter that reduces the Na⁺/K⁺ ratio under salt
270 stress [55].

271 While QTL mapping enables significant regions related with a trait to be
272 identified, functional genomic analysis, with the support of NGS, provides
273 complete RNA profiles to determine the expression of QTL genes in specific
274 biological conditions. The integration of these two strategies results in the
275 detection of **expression quantitative trait loci** (eQTL) that enable the
276 expression of complex traits governed by multiple QTLs/genes to be explained
277 [56,57] (Figure 2). In the eQTL approach, segregating populations are both
278 genotyped and phenotyped by expression profiling methods such as microarray
279 or RNA-seq to collect the information of transcript abundance. Rather than
280 microarray, RNA-seq is becoming the technique of choice in eQTL analyses
281 because it can determine allele-specific expression and isoform-RNA expression
282 [58]. Thousands of RNA expression levels are analyzed for linkage or association
283 with genetic markers, leading to the detection of variations acting in *cis* or *trans*
284 manners. *Cis*-acting factors are DNA variations located within or near a
285 differentially expressed gene and regulating its transcription. *Trans*-acting factors
286 are distantly mapped elsewhere in the genome and influence the activity of
287 transcription factors that regulate the differentially expressed gene [58,59]. Using
288 this approach in maize, a strong trans-acting eQTL has been successfully fine
289 mapped to an interval of only 186 bp within a class I glutamine amidotransferase
290 domain containing gene [60]. Under the effect of this eQTL, the transcription level
291 of another gene encoding an ABA 8'-hydroxylase was upregulated to 6-fold
292 greater in one parental genotype compared to the other. Although the regulatory
293 mechanisms involving the glutamine amidotransferase protein on ABA 8'-
294 hydroxylase gene expression remained unclear, the cloning of this trans-acting
295 eQTL showed the efficiency of the eQTL approach to identify causative genes.
296 Furthermore, coexpression network databases compiling a large number of
297 microarray studies were developed to further help in identifying functionally
298 related genes. For instance, RiceFRIEND (<http://ricefriend.dna.affrc.go.jp>) was

299 helpful in detecting shared expression networks between candidate genes for
300 panicle development in rice [61,62].

301 Although eQTL is powerful, the application of this method still remains a
302 challenge because of the heavy costs to do experiments with large samples,
303 difficulties in finding an appropriate statistical method to analyze the downstream
304 eQTLs linked with physiological or morphological phenotypes and the
305 computational resources needed to handle the large datasets [63,64]. In this
306 context, the prediction of regulatory cascades and their major hubs during the
307 realization of a trait using systems biology approaches could be a solution [65].

308 *Figure 2*

309

310 **Polymorphism databases expedite the identification of candidate genes**

311 Fast technical progress accompanying the cost decrease of NGS-based
312 methods induced many WGS studies of numerous varieties, particularly in rice
313 [66–69]. Although the sequencing qualities differed in depth and coverage, the
314 results of these studies provided large-scale polymorphism resources that enable
315 the validation of target SNPs and structural variation associated with important
316 agronomic traits. For example, the sequence variability of the granule bound
317 starch synthase gene related to amylose content in rice grain was analyzed using
318 WGS data from 47 elite varieties [68]. New genetic markers were successfully
319 designed to track alleles affecting this trait. In addition, the high density of
320 variations obtained from WGS allowed the development of markers to track
321 alleles/genes involved in other agronomic traits. Moreover, WGS enabled the
322 recombination points closest to the causative gene to be marked, to avoid
323 undesirable effects during MAS.

324 SNP-Seek, the 3K project database (<http://snp-seek.irri.org/>), enabled
325 immediate *in silico* access to sequence variations including SNPs and InDels for
326 the target segment in rice. This resource allowed the validation of a QTL
327 haplotype by identifying varieties that carried either contrasted haplotypes or
328 recombinant haplotypes, phenotyping these varieties, and detecting which allelic
329 variation was responsible for the QTL effect [70,71]. For instance, SNP-Seek

330 facilitated the prediction of novel genes/alleles of resistance to rice blast disease
331 based on sequence and structure variations between the resistant haplotypes
332 and the susceptible ones [72]. In another example, SNP-seek was used to detect
333 mutations in the Effector Binding Elements (EBE) of promoters of rice genes
334 favorable to the proliferation of bacterial blight, making impossible the recognition
335 of EBE by the bacteria Transcription Activator-like Effectors (TALE). Such
336 mutations could improve plant resistance against the bacteria. The mining of
337 such mutations in the 3K database combined with a rapid phenotyping for
338 bacterial blight resistance is used to detect new sources of resistance [72].

339 During rice domestication, important agronomic alleles were fixed in elite
340 varieties but not in wild ones, thus these alleles appear to be very rare among
341 non-elite accessions. The comparison of the sequences of elite varieties with the
342 sequences of non-elite varieties selected from public genomic data revealed
343 SNPs which were fixed in elite varieties but had a low frequency (<5%) in non-
344 elite varieties. For example, this method allowed the detection of an important
345 nonsynonymous mutation in the 9-cis-epoxycarotenoid dioxygenase gene (*Nced*)
346 that was associated with adaption to upland conditions, possibly through
347 significantly higher abscisic acid levels and denser lateral roots [73]. The
348 promising results in rice which facilitated the identification of candidate
349 genes/alleles and generated novel markers for marker-assisted crop breeding,
350 promoted the investigation and the development of SNP databases in other
351 crops [74–76].

352

353 **Concluding remarks**

354 With its broad applications, NGS is becoming an essential tool for crop
355 geneticists to identify and characterize genomic variations associated with
356 agronomical traits. WGR and transcription profiling that contribute to provide
357 comprehensive information on genetic variability and their regulatory
358 mechanisms are the most popular applications of NGS. QTL-seq, MutMap and
359 their extended versions showed efficiency in narrowing down the position of
360 QTLs and precisely detecting their causative variations. RNA-seq provided

361 functional context to candidate genes. As such, a large number of QTLs/eQTLs
362 were found in attempts to break down the genetic mechanisms regulating
363 important agronomic traits.

364 To be successful in the interpretation of NGS data, bioinformatic
365 computational methods are critical elements to delivering accurate assembly,
366 alignment and variant detection [77]. Second-generation sequencing platforms
367 such as SOLiD, Illumina (MiSeq and HiSeq), Roche (454) and Ion torrent
368 produce short reads that range from 35 bp to 700 bp. Short-read sequencing
369 approaches have created a revolution for the *de novo* assembly of new reference
370 genomes, the analysis of population structure or the identification of SNPs and
371 InDels. However, plant genomes are complex with an abundance of repetitive
372 regions, transposons, and genomic structural variations, making short-read
373 approaches insufficient, particularly in the case of large genomes such as wheat
374 or maize [78,79]. Long-read sequencing (up to several kb) produced by third-
375 generation sequencing systems such as PacBio or Oxford Nanopore [80] is, then,
376 a promising way to overcome the limitations of short-read sequencing
377 approaches. The increase in read length allows researchers to span repeats or
378 scaffolding gaps, to solve genomic rearrangements, thus, generating a higher
379 quality assembly [80–82]. It also enables the determination of epigenetic marks
380 in highly variable genomic regions by DNA methylation and their effect on gene
381 expression [83,84]. In polyploid plant species, longer reads are beneficial to
382 detect specific-SNPs enabling the differentiation of a segregating SNP from
383 homeologous sequences [16]. One important advantage of longer read
384 sequencing is to facilitate haplotype phasing, which is a necessary step in the
385 map construction and QTL mapping in heterozygous crops [85,86]. Moreover,
386 the development of longer read sequencing allows a more precise analysis of
387 mRNA structure variation such as exon-intron limits, alternative splicing and RNA
388 isoform [87].

389 Emerging long-read sequencing approaches with their advantages will
390 accelerate the construction of high-quality reference genomes and, combined
391 with genetic approaches, speed up gene discovery in plants. However, the

392 genetic approaches described in this review are all based on a combination of
393 genotyping/sequencing and phenotyping. By comparison, phenotyping has not
394 registered the same progress as genotyping and is often the element limiting the
395 population size for traits complex to phenotype. Progress has also to be made in
396 decreasing phenotyping costs and arduousness. Automatized high throughput
397 phenotyping platforms designed for greenhouse or field conditions can help
398 develop high precision phenotyping, give access to dynamic traits by repeating
399 easily measurements along time, decrease costs and contribute to speed up
400 gene discovery even further [88] (see *Outstanding questions*). To target QTLs
401 with small effects, phenotyping precision will need to be improved. In addition the
402 resolution of genetic determinants of small effect multi-loci dependant traits will
403 beneficiate of the capacity to conduct transcriptome-wide association studies
404 (TWAS) that aims to associate gene expression, SNP in *cis*-regulatory
405 sequences and traits in large population. This approach is starting to be use in
406 medicine to identify genes associated with complex traits (eg. obesity, [92]) and
407 is promizing for application in plant science. Similarly, the systems biology
408 approach that allows to consider globally the regulatory links between all genes
409 involved in the realisation of a trait will help to properly manipulate multi-loci
410 dependant traits (sytems biology approaches for plant breeding have been
411 recently reviewed in [65]). Like medicine, modern plant breeding will require a
412 shift toward the development of multidisciplinary teams able to deal with plant
413 biology, genetics, large scale phenotyping approaches, sequencing,
414 bioinformatics, data analysis, statistic, and mathematics, that is an exciting
415 perspective.

416

417 **ACKNOWLEDGEMENTS**

418

419 LKN was supported by an IRD ARTS PhD thesis fellowship. In addition to
420 institutional funding, this work was supported by the Global Rice Science
421 Partnership (2011-2016) and by the CGIAR Research Program (CRP) on rice
422 agri-food systems (RICE, 2017-2022).

423

425 **References**

- 426 1 Kole, C. *et al.* (2015) Application of genomics-assisted breeding for
 427 generation of climate resilient crops: progress and prospects. *Front. Plant*
 428 *Sci.* 6, 1–16
- 429 2 Zuo, J. and Li, J. (2014) Molecular genetic dissection of quantitative trait
 430 loci regulating rice grain size. *Annu. Rev. Genet.* 48, 99–118
- 431 3 Gupta, P.K. *et al.* (2017) QTL analysis for drought tolerance in wheat:
 432 present status and future possibilities. *Agronomy* 7, 5
- 433 4 Moury, B. *et al.* (2017) Quantitative resistance to plant pathogens in
 434 pyramiding strategies for durable crop protection. *Front. Plant Sci.* 8, 1–9
- 435 5 Bailey-serres, J. *et al.* (2010) Submergence tolerant rice: SUB1' s journey
 436 from landrace to modern cultivar. *Rice* 3, 138–147
- 437 6 Salvi, S. and Tuberosa, R. (2015) The crop QTLome comes of age. *Curr.*
 438 *Opin. Biotechnol.* 32, 179–185
- 439 7 Das, G. *et al.* (2017) Insight into MAS: A molecular tool for development
 440 of stress resistant and quality of rice through gene stacking. *Front. Plant*
 441 *Sci.* 8, 1–9
- 442 8 Goffinet, B. and Gerber, S. (2000) Quantitative Trait Loci: A Meta-
 443 analysis. *Genetics* 155, 463–473
- 444 9 Chardon, F. *et al.* (2004) Genetic architecture of flowering time in maize as
 445 inferred from quantitative trait loci meta-analysis and synteny conservation
 446 with the rice genome. *Genetics* 2185, 2169–2185
- 447 10 Zhang, X. *et al.* (2016) Meta-analysis of major QTL for abiotic stress
 448 tolerance in barley and implications for barley breeding. *Planta* 245, 283–
 449 295
- 450 11 Zhang, L. *et al.* (2010) Genomic distribution of quantitative trait loci for yield
 451 and yield-related traits in common wheat. *J. Integr. Plant Biol.* 52, 996–
 452 1007
- 453 12 Courtois, B. *et al.* (2009) Rice root genetic architecture: Meta-analysis from
 454 a drought QTL database. *Rice* 2, 115–128
- 455 13 Salvi, S. and Tuberosa, R. (2005) To clone or not to clone plant QTLs:
 456 present and future challenges. *Trends Plant Sci.* 10, 297–304
- 457 14 Sanger, F. and Nicklen, S. (1977) DNA sequencing with chain-terminating.
 458 *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467
- 459 15 Egan, A.N. *et al.* (2012) Applications of next - generation sequencing in
 460 plant biology. *Am. J. Bot.* 99, 175–185
- 461 16 Poland, J.A. and Rife, T.W. (2012) Genotyping-by-sequencing for plant
 462 breeding and genetics. *Plant Genome* 5, 92–102
- 463 17 He, J. *et al.* (2014) Genotyping-by-sequencing (GBS), an ultimate marker-
 464 assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.*
 465 5, 484
- 466 18 Varshney, R.K. *et al.* (2014) Harvesting the promising fruits of genomics:
 467 applying genome sequencing technologies to crop breeding. *Plos Biol.* 12,
 468 6

- 469 19 Thomson, M.J. (2014) High-throughput SNP genotyping to accelerate crop
470 improvement. *Plant Breed. Biotechnol.* 2, 195–212
- 471 20 Spindel, J. *et al.* (2013) Bridging the genotyping gap: Using genotyping by
472 sequencing (GBS) to add high-density SNP markers and new value to
473 traditional bi-parental mapping and breeding populations. *Theor. Appl.*
474 *Genet.* 126, 2699–2716
- 475 21 Huang, X. *et al.* (2009) High-throughput genotyping by whole-genome
476 resequencing. *Genome Res.* 19, 1068–1076
- 477 22 Elshire, R.J. *et al.* (2011) A robust , simple genotyping-by-sequencing
478 (GBS) approach for high diversity species. *PLoS One* 6, 1–10
- 479 23 Chapman, J.A. *et al.* (2015) A whole-genome shotgun approach for
480 assembling and anchoring the hexaploid bread wheat genome. *Genome*
481 *Biol.* 16, 1–26
- 482 24 Phung, N.T.P. *et al.* (2016) Genome-wide association mapping for root
483 traits in a panel of rice accessions from Vietnam. *BMC Plant Biol.* 16, 64
- 484 25 Huang, B.E. *et al.* (2015) MAGIC populations in crops□: current status and
485 future prospects. *Theor. Appl. Genet.* 128, 999–1017
- 486 26 Lu, F. *et al.* (2015) High-resolution genetic mapping of maize pan-genome
487 sequence anchors. *Nat. Commun.* 6, 6914
- 488 27 Varshney, R.K. *et al.* (2009) Next-generation sequencing technologies and
489 their implications for crop genetics and breeding. *Trends Biotechnol.* 27,
490 522–530
- 491 28 Scheben, A. *et al.* (2017) Genotyping-by-sequencing approaches to
492 characterize crop genomes□: choosing the right tool for the right
493 application. *Plant Biotechnol. J.* 15, 149–161
- 494 29 Michelmore, R.W. *et al.* (1991) Identification of markers linked to disease-
495 resistance genes by bulked segregant analysis - a rapid method to detect
496 markers in specific genome regions by using segregating populations. *Proc.*
497 *Natl. Acad. Sci.* 88, 9828–9832
- 498 30 Zou, C. *et al.* (2016) Bulk sample analysis in genetics , genomics and
499 crop improvement. *Plant Biotechnol. J.* 14, 1941–1955
- 500 31 Thanda, K. *et al.* (2016) QTL mapping for downy mildew resistance in
501 cucumber via bulked segregant analysis using next□generation
502 sequencing and conventional methods. *Theor. Appl. Genet.* 130, 199–211
- 503 32 Wambugu, P. *et al.* (2017) Sequencing of bulks of segregants allows
504 dissection of genetic control of amylose content in rice. *Plant Biotechnol. J.*
505 1, 1–11
- 506 33 Hayward, A. *et al.* (2015) Molecular marker applications in plants. *Methods*
507 *Mol. Biol.* 1245, 101–18
- 508 34 Magwene, P.M. *et al.* (2011) The statistics of bulk segregant analysis using
509 next generation sequencing. *PloS Comput. Biol.* 7, 1–9
- 510 35 Terauchi, R. *et al.* (2015) Whole genome sequencing to identify genes and
511 QTL in rice. In *Advances in the understanding of biological sciences using*
512 *next generation sequencing (NGS) approaches* 1pp. 33–42
- 513 36 Takagi, H. *et al.* (2013) QTL-seq: Rapid mapping of quantitative trait loci in
514 rice by whole genome resequencing of DNA from two bulked populations.

- 515 *Plant J.* 74, 174–183
- 516 37 Das, S. *et al.* (2014) Deploying QTL-seq for rapid delineation of a potential
517 candidate gene underlying major trait-associated QTL in chickpea. *DNA*
518 *Res.* 22, 193–203
- 519 38 Das, S. *et al.* (2015) MQTL-seq delineates functionally relevant candidate
520 gene harbouring a major QTL regulating pod number in chickpea. *DNA*
521 *Res.* 23, 53–65
- 522 39 Parida, S.K. *et al.* (2017) A genome-wide mQTL-seq scan identifies
523 potential molecular signatures regulating plant height in chickpea. *Plant*
524 *Mol. Biol. Report.* 35, 273–286
- 525 40 Abe, A. *et al.* (2012) Genome sequencing reveals agronomically important
526 loci in rice using MutMap. *Nat. Biotechnol.* 30, 174–178
- 527 41 Mascher, M. *et al.* (2014) Mapping-by-sequencing accelerates forward
528 genetics in barley. *Genome Biol.* 15, R78
- 529 42 Takagi, H. *et al.* (2011) MutMap accelerates breeding of a salt-tolerant rice
530 cultivar. *Nat. Biotechnol.* 1, 1–5
- 531 43 Fekih, R. *et al.* (2013) MutMap+: Genetic Mapping and Mutant
532 Identification without Crossing in Rice. *PLoS One* 8, 1–10
- 533 44 Takagi, H. *et al.* (2013) MutMap-Gap: Whole-genome resequencing of
534 mutant F2 progeny bulk combined with de novo assembly of gap regions
535 identifies the rice blast resistance gene Pii. *New Phytol.* 200, 276–283
- 536 45 Gedil, M. *et al.* (2016) Perspectives on the application of next-generation
537 sequencing to the improvement of Africa' s staple food crops. In *Next*
538 *generation sequencing - Advances, applications and challenges* pp. 287–
539 321
- 540 46 Kujur, A. *et al.* (2015) A genome-wide SNP scan accelerates trait-
541 regulatory genomic loci identification in chickpea. *Sci. reports* 5, 11166
- 542 47 Kloosterman, B. *et al.* (2010) From QTL to candidate gene: genetical
543 genomics of simple and complex traits in potato using a pooling strategy.
544 *BMC Genomics* 11, 158
- 545 48 Daware, A. *et al.* (2016) An efficient strategy combining SSR markers- and
546 advanced QTL-seq-driven QTL mapping unravels candidate genes
547 regulating grain weight in rice. *Front. Plant Sci.* 7, 1–17
- 548 49 Wang, Z. *et al.* (2010) RNA-Seq: a revolutionary tool for transcriptomics.
549 *Nat. Rev. Genet.* 10, 57–63
- 550 50 Li, X. *et al.* (2012) Genic and nongenic contributions to natural variation of
551 quantitative traits in maize. *Genome Res.* 22, 2436–2444
- 552 51 Martin, J.A. and Wang, Z. (2011) Next-generation transcriptome assembly.
553 *Nat. Rev. Genet.* 12, 671–682
- 554 52 Kudo, T. *et al.* (2016) Identification of reference genes for quantitative
555 expression analysis using large-scale RNA-seq data of *Arabidopsis*
556 *thaliana* and model crop plants. *Genes Genet. Syst.* 91, 111–125
- 557 53 Gelli, M. *et al.* (2016) Mapping QTLs and association of differentially
558 expressed gene transcripts for multiple agronomic traits under different
559 nitrogen levels in sorghum. *BMC Plant Biol.* 16, 16
- 560 54 Jiang, Q. *et al.* (2016) Combining meta-QTL with RNA-seq data to identify

561 candidate genes of kernel row number trait in maize. *Crea J.* 61, 4
562 55 Qi, X. *et al.* (2014) Identification of a novel salt tolerance gene in wild
563 soybean by whole-genome sequencing. *Nat. Commun.* 5, 4340
564 56 Majewski, J. and Pastinen, T. (2011) The study of eQTL variations by
565 RNA-seq: From SNPs to phenotypes. *Trends Genet.* 27, 72–79
566 57 Westra, H. and Franke, L. (2014) From genome to function by studying
567 eQTLs. *Biochim. Biophys. Acta - Mol. Basis Dis.* 1842, 1896–1902
568 58 Sun, W. and Hu, Y. (2012) eQTL mapping using RNA-seq data. *Stat.*
569 *Biosci.* 5, 189–219
570 59 Cubillos, F.A. *et al.* (2012) Lessons from eQTL mapping studies: Non-
571 coding regions and their role behind natural phenotypic variation in plants.
572 *Curr. Opin. Plant Biol.* 15, 192–198
573 60 Holloway, B. *et al.* (2011) Genome-wide expression quantitative trait loci
574 (eQTL) analysis in maize. *BMC Genomics* 12, 336
575 61 Sato, Y. *et al.* (2013) RiceFRIEND: A platform for retrieving coexpressed
576 gene networks in rice. *Nucleic Acids Res.* 41, 1214–1221
577 62 Crowell, S. *et al.* (2016) Genome-wide association and high-resolution
578 phenotyping link *Oryza sativa* panicle traits to numerous trait-specific QTL
579 clusters. *Nat. Commun.* 7, 10527
580 63 Lipka, A.E. *et al.* (2015) From association to prediction: Statistical methods
581 for the dissection and selection of complex traits in plants. *Curr. Opin. Plant*
582 *Biol.* 24, 110–118
583 64 Feltus, F.A. (2014) Systems genetics: A paradigm to improve discovery of
584 candidate genes and mechanisms underlying complex traits. *Plant Sci.* 223,
585 45–48
586 65 Lavarenne, J. *et al.* (2018) The Spring of Systems Biology-Driven Breeding.
587 *Trends Plant Sci.* 23, 1–15
588 66 Xu, X. *et al.* (2012) Resequencing 50 accessions of cultivated and wild rice
589 yields markers for identifying agronomically important genes. *Nat.*
590 *Biotechnol.* 30, 105–11
591 67 Huang, X. *et al.* (2010) Genome-wide association studies of 14 agronomic
592 traits in rice landraces. *Nat. Genet.* 42, 961–967
593 68 Duitama, J. *et al.* (2015) Whole genome sequencing of elite rice cultivars
594 as a comprehensive information resource for marker assisted selection.
595 *PLoS One* 10, 1–20
596 69 Alexandrov, N. *et al.* (2015) SNP-Seek database of SNPs derived from
597 3000 rice genomes. *Nucleic Acids Res.* 43, D1023–D1027
598 70 Mori, A. *et al.* (2016) The role of root size versus root efficiency in
599 phosphorus acquisition in rice. *J. Exp. Bot.* 67, 1179–1189
600 71 Wissuwa, M. *et al.* (2016) From promise to application: root traits for
601 enhanced nutrient capture in rice breeding. *J. Exp. Bot.* 67, 3605–3615
602 72 Leung, H. *et al.* (2015) Allele mining and enhanced genetic recombination
603 for rice breeding. *Rice* 8, 34
604 73 Lyu, J. *et al.* (2013) Analysis of elite variety tag SNPs reveals an important
605 allele in upland rice. *Nat. Commun.* 4, 2138
606 74 Doddamani, D. *et al.* (2015) CicArVarDB: SNP and InDel database for

607 advancing genetics research and breeding applications in chickpea.
608 *Database 1*, 1–7

609 75 Xu, C. *et al.* (2017) Development of a maize 55 K SNP array with improved
610 genome coverage for molecular breeding. *Mol. Breed.* 37, 20

611 76 Joshi, T. *et al.* (2014) Soybean knowledge base (SoyKB): a web resource
612 for integration of soybean translational genomics and molecular breeding.
613 *Nucleic Acids Res.* 42, 1245–1252

614 77 Voelkerding, K. V *et al.* (2010) Next generation sequencing for clinical
615 diagnostics-principles and application to targeted resequencing for
616 hypertrophic cardiomyopathy. *J. Mol. Diagnostics* 12, 539–551

617 78 Alkan, C. *et al.* (2011) Limitations of next-generation genome sequence
618 assembly. *Nat. Methods* 8, 61–65

619 79 Torkamaneh, D. *et al.* (2018) Efficient genome □ wide genotyping
620 strategies and data integration in crop plants. *Theor. Appl. Genet.* 131,
621 499–511

622 80 Yuan, Y. *et al.* (2017) Improvements in genomic technologies□: application
623 to crop genomics. *Trends Biotechnol.* 35, 547–558

624 81 Jiao, W. and Schneeberger, K. (2017) The impact of third generation
625 genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.*
626 36, 64–70

627 82 Schatz, M.C. *et al.* (2014) Whole genome de novo assemblies of three
628 divergent strains of rice, *Oryza sativa*, document novel gene space of aus
629 and indica. *Genome Biol.* 15, 506

630 83 Meaburn, E. and Schulz, R. (2012) Seminars in cell & developmental
631 biology next generation sequencing in epigenetics□: insights and
632 challenges. *Semin. Cell Dev. Biol.* 23, 192–199

633 84 Gabrieli, T. *et al.* (2018) Genome-wide epigenetic profiling of 5-
634 hydroxymethylcytosine by long-read optical mapping. *bioRxiv* DOI:
635 <http://dx.doi.org/10.1101/260166>

636 85 Browning, S.R. and Browning, B.L. (2011) Haplotype phasing□: existing
637 methods and new developments. *Nat. Publ. Gr.* 12, 703–714

638 86 Schlötterer, C. *et al.* (2014) Sequencing pools of individuals — mining
639 genome-wide polymorphism data without big funding. *Nat. Publ. Gr.* 15,
640 749–763

641 87 Pastinen, T. (2010) Genome-wide allele-specific analysis: insights into
642 regulatory variation. *Nat. Publ. Gr.* 11, 533–538

643 88 Montes, J.M. *et al.* Novel throughput phenotyping platforms in plant genetic
644 studies. *Trends Plant Sci.* 12, 10–13

645 89 Baird, N.A. *et al.* (2008) Rapid SNP discovery and genetic mapping using
646 sequenced RAD markers. *PLoS One* 3, 1–7

647 90 Peterson, B.K. *et al.* (2012) Double digest RADseq□: an inexpensive
648 method for de novo SNP discovery and genotyping in model and non-
649 model species. *PLoS One* 7, e37135

650 91 Bayer, P.E. *et al.* (2015) High resolution skim genotyping by sequencing
651 reveals the distribution of crossovers and gene conversions in *Cicer*
652 arietinum and *Brassica napus*. *Theor. Appl. Genet.* 128, 1039–1047

653 92 Gusev, Alexander, *et al.* (2016) Integrative approaches for large-scale
654 transcriptome-wide association studies." *Nature genetics* 48.3: 245.
655
656
657

658

659 **Box1. NGS-based genotyping approaches used in crop genetics**

660 **Reduced-representation sequencing (RRS)**

661 In general, the preparation of a sequencing library starts by the digestion of
662 genomic DNA with restriction enzymes, followed by the attachment of barcode
663 adapters and pooling for multiplex sequencing of the samples. In the restriction
664 enzyme-associated DNA sequencing method (RADseq) [89], DNA fragments are
665 further sheared while in a variation of RADseq called double digest restriction-
666 site associated DNA marker generation (ddRADseq) [90], this step is replaced by
667 a digestion with a second enzyme which helps to improve fragment selection by
668 size. The fragments are purified and ligated to common adapters. Finally, they
669 are amplified to produce sequencing libraries. In the genotyping-by-sequencing
670 (GBS)[22], the preparation of sequencing libraries is simplified by eliminating the
671 step of DNA size fractionation. In addition, both barcode adapters and common
672 adapters have overhangs at restriction site and are simultaneously ligated to
673 DNA fragments through sticky-ends. The sequencing is performed by systems
674 such as Illumina or Ion Torrent, producing short-reads of 50 to 150 bp. RRS
675 methods simultaneously detect polymorphisms in the region flanking the
676 restriction site and call genotypes. Among RRS methods, GBS is presently a
677 popular technique for crop genetics since it provides an appropriate SNP density
678 but a compromise has to be found between cost efficiency and sequencing depth,
679 which needs to be high for accurate allele calling, particularly in heterozygous
680 crops. Another advantage of GBS is that, in the absence of a reference genome,
681 the consensus of the read clusters nearby the restriction sites can become a
682 reference. The high rate of missing data due to low sequencing depth and the
683 intrinsic error rate of the sequencing technique are the two main concerns for this
684 approach.

685 **Whole genome resequencing (WGR)**

686 This method supposes that a reference genome is available. Genomic DNA is
687 sheared, ligated to adaptors and amplified. The amplified PCR products are then
688 separated by size and purified to provide the sequencing libraries. Short-reads

689 generated from sequencing are aligned on the reference genome. Skim-based
690 genotyping by sequencing (SkimGBS) was developed for high-resolution whole
691 genome resequencing of mapping populations [91]. After SNPs between the
692 parents are called, the progeny reads are mapped on the same reference and
693 compared to parental SNP data to determine the genotypes and recombination
694 frequencies. In addition, a sliding window approach, which examines collectively
695 consecutive SNPs instead of assessing SNPs individually, was proposed as a
696 method to avoid erroneous SNP calling [21]. Compared to RRS approaches,
697 WGR eliminates several steps in the preparation of sequencing libraries and
698 provides a high-throughput genotyping with low cost per marker point. The
699 polymorphisms detected by WGR are more comprehensive, including not only
700 SNPs but also structural variations, gene conversions, recombination break
701 points, etc. However, the cost per sample remains high depending on the chosen
702 coverage and crop genome size.

703

704 *Figure 1:* Different approaches combining bulk segregant analysis and whole
705 genome resequencing developed to identify genetic variations controlling
706 valuable traits. Mapping populations are generated from biparental crosses
707 (QTL-seq; mQTL-seq) or from crosses between a wild-type and its mutant
708 (MutMap, MutMap-gap); In MutMap+, no cross is generated; The M₂ and M₃
709 generations are obtained from M₁ and M₂, respectively, by selfing; The portions
710 of the population that are pooled as DNA bulks and sequenced are hatched. A
711 SNP index is calculated to identify SNPs linked with the trait of interest. P:
712 parent; WT: wild-type; LB: low-trait bulk; HB: high-trait bulk; MB: mutant bulk;
713 WTB: wild-type bulk.

714

715

716

717 *Figure 2:* Integration of QTL and eQTL detection identify the causative genes
718 involved in the realization and the modulation of a trait.

719 In this example, a segregating population was genotyped and phenotyped
720 leading to the detection of a linkage between the studied trait and a SNP (A or G)
721 located in the promoter of gene1. This defines a QTL. In parallel, a genome wide
722 expression study of the individuals of the population detected a correlation
723 between the expression of gene1 (violet graph) that carries the SNP in its
724 promoter, and the expression of the trait (green graph). This defines a cis-eQTL.
725 The expression of gene 2 (orange graph), for which no significant genetic linkage
726 with the SNP was detected, is also correlated with the expression of the trait
727 (green graph). This defines a trans-eQTL. The functional analysis revealed that
728 the expression level of gene 1 is modulated by the SNP detected in its promoter
729 and that the product of gene 1 is a transcription factor (TF1) that binds to the
730 promoter of gene 2 and modulates its expression. In this example, gene 2
731 controls the trait and gene 1 modulates the intensity of the trait. QTL: quantitative
732 trait loci, eQTL: expression quantitative trait loci, TF1 : transcription factor 1, red
733 triangle: position of the SNP associated with the trait.

734
735

736 **Glossary**

737

738 **Bulk segregant analysis:** Extreme phenotypic individuals from a biparental
739 mapping population are identified and a low-trait and a high-trait bulk are
740 constituted by pooling the DNA of approximately 10 plants of each tail. The two
741 bulks and the two parents are genotyped at a high density to identify molecular
742 markers that have different allelic frequency between the two bulks and establish
743 a link between those markers and the trait of interest.

744

745 **Doubled haploids (DH):** plants produced from the chromosome doubling of F_1
746 haploid plantlets obtained using anther culture. DH lines are perfectly
747 homozygous (fixed).

748

749 **Expression-Quantitative Trait Locus (eQTL):** a genomic locus that regulates
750 gene transcripts. eQTLs analysis tests the association between genetic markers
751 and gene expression level in a segregation population, leading to the
752 identification of regulatory variants located nearby or far away from the target
753 gene.

754

755 **Genome-wide association study (GWAS):** Method used to identify genomic
756 regions/variants statistically associated with the phenotypic values of a diverse
757 panel.

758

759 **Meta-QTL:** QTL resulting from the statistical integration of independent QTL
760 studies leading to QTLs with a smaller confidence interval of the position than the
761 initial QTLs.

762

763 **Multiparent Advanced Generation Intercross:** Mapping population obtained
764 from a complex pyramidal intercrossing scheme involving multiple parents (4-8
765 lines). Intercrossing is carried out for several generations before selfing the
766 plants up to full fixation.

767

768 **Near-isogenic lines (NILs):** Lines developed through several backcrosses on a
769 recurrent parent to obtain a new line with a genome identical to that of the
770 recurrent parent except at a particular locus of interest introgressed from a donor.
771 NILs are among the best materials to validate a QTL.

772

773 **Nested association mapping (NAM):** population generated by the creation of
774 multiple recombinant inbred lines having one common parent. NAM population
775 takes advantages of both linkage and association mapping to increase mapping
776 resolution with a reasonable marker density.

777

778 **Next-generation sequencing (NGS):** a term that encompasses all high-
779 throughput short-read sequencing platforms. NGS can be used to rapidly
780 sequence DNA.

781

782 **Quantitative Trait Locus (QTL):** one of the DNA segments linked with the
783 variation of a quantitative trait.

784

785 **Recombinant inbred lines (RILs):** homozygous lines derived from a biparental
786 cross obtained by selfing plants during several generations up to fixation.

787

788 **RNA-sequencing:** a method to detect the presence and quantity of RNA in a
789 given sample. The total RNA extracted from each sample is converted to cDNA,
790 then sequenced by an NGS platform.

791

792 **Segregation distortion:** a phenomenon in which the segregation ratio of the
793 observed genotypes of a mapping population at a given marker significantly
794 differs from the expected Mendelian ratio for this type of population.

795

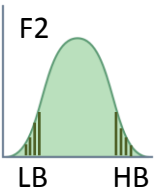
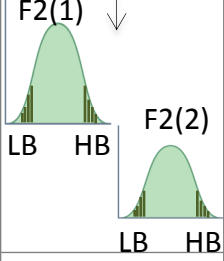

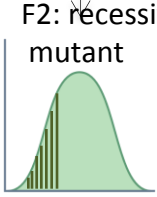
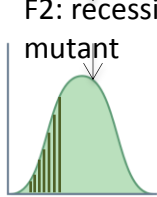
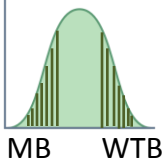
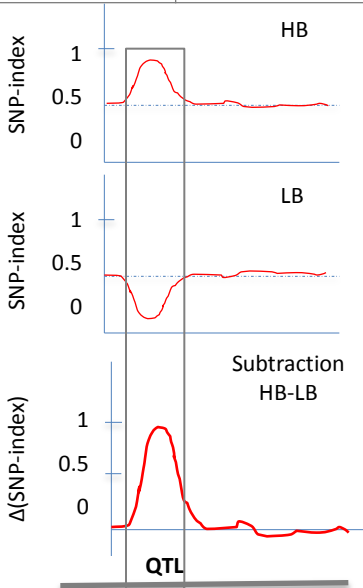
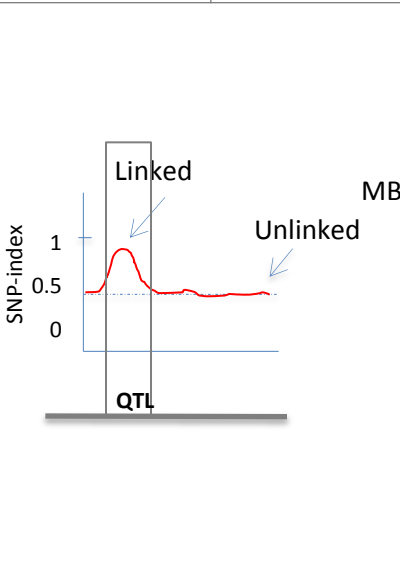
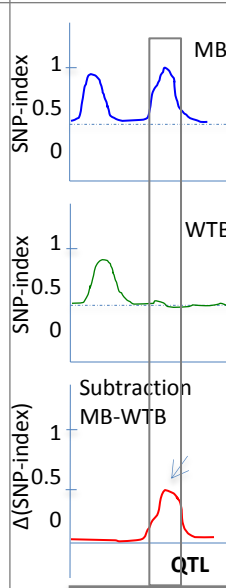
796 **SNP index:** In a biparental mapping population that was sequenced, the
797 proportion of short reads harboring a given SNP with the sequence of one of the
798 two parents chosen as reference.

799

800 **Whole genome resequencing (WGR):** once a reference genome is available for
801 a given species, sequencing of new individuals is performed to identify
802 polymorphisms and structural variations compared to the reference genome.

803

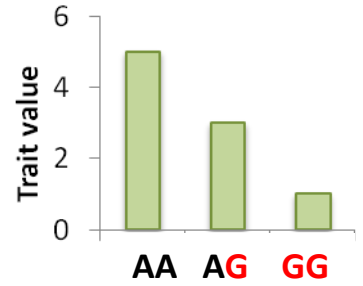
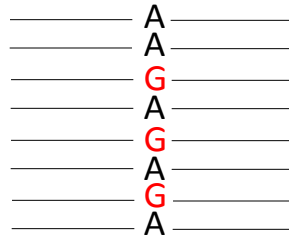
804

	(A) QTL-seq [36]	(B) mQTL-seq [38]	(C) MutMap [40]	(D) MutMap-Gap [44]	(E) MutMap+ [43]	
Genetic resource	Biparental population P1 x P2 ↓ F1 ↓ F2	Multiparental population Cross 1, cross 2...cross n (n>=2) F1(1,2,...,n) F2(1,2,...,n)	WT -> Mutant WT x Mutant F1 F2	WT -> <i>De novo</i> sequencing -> WT-specific regions WT x Mutant F1 F2	WT-> Mutant M2 M3	
Bulked samples	F2 	F2(1) ↓  F2(2) 	F2: recessive mutant 	F2: recessive mutant 	M3: recessive mutant 	
Sample size	10-20 /bulk+parent	10-20/ bulk+parent	20 mutant ind.+WT	20 mutant ind.+WT	20-40/bulk+WT	
SNP-index						

Segregating population

Genotypes

Phenotypes



Gene expression

