Next-Generation Sequencing Accelerates Crop Gene Discovery

- Khanh Le Nguyen, Alexandre Grondin, Brigitte Courtois, Pascal Gantet

Next-generation sequencing accelerates crop gene discovery

Khanh Le Nguyen[1,2], Alexandre Grondin[1], Brigitte Courtois[3,4] and Pascal Gantet[1,*]

[1]Université de Montpellier, Institut de Recherche pour le Développement, UMR DIADE, 911 Avenue Agropolis, 34394 Montpellier cedex 5, France
[2]LMI RICE 2, AGI, Km2 Pham Van Dong, Tu Liem, Hanoi, Vietnam
[3]CIRAD, UMR AGAP, F-34398 Montpellier, France
[4]Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

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

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

Abstract:

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

How NGS boosts QTL and gene determination.

In molecular genetics, quantitative traits are first decomposed in their Mendelian components by **quantitative trait loci (**QTL) analysis. Then, each QTL is fine-mapped or cloned individually. Thousands of QTLs (see Glossary) associated with agronomic traits were found in crops and represent a reservoir of alleles for breeders to create improved varieties [1–4]. *SUBMERGENCE 1* (*SUB1*), a major QTL which confers tolerance to submergence in rice (*Oryza sativa*), is probably one of the most successful examples of QTL utilization worldwide [5]. This QTL with large effect was identified in a traditional rice variety, and the underlying gene that is absent from the genome of the reference rice variety, was cloned. The favorable allele of the *SUB1* gene was introgressed into elite cultivars by marker-assisted backcrossing and the improved products were released in several Asian countries. However, very few QTLs were as successfully used in marker-assisted selection (MAS) because they were positioned with insufficient precision, because they explained a low proportion of the trait variation, or because QTL x environment interactions made them useless outside their detection context [6,7]. Before undertaking MAS, one challenge is to reduce the confidence interval (CI) of the QTL position to make the introgressed segment carrying the QTL as small as possible and to avoid possible undesirable side effects due to the other genes carried by the introgressed segment. To reduce the CI of a QTL position, a possible approach is to undertake a meta-analysis of different studies targeting the same trait in the same species [8]. A QTL meta-analysis was effectively applied to different crops to refine the CI regions and seek candidate genes [9–12]. QTLs positioned on a single consensus map and narrowed down by meta-analysis enable the target regions of interest for MAS to be more precisely identified. However, depending on the size of the meta-QTLs, further steps of either fine mapping and positional cloning or association mapping with sufficient marker density are often necessary to identify the shortest target DNA fragment responsible for the phenotypic variation [12,13].

Next-Generation Sequencing (NGS) designates new sequencing methods (see *Box*) that produce high coverage with lower cost and higher speed than traditional SANGER sequencing [14,15]. Among NGS platforms, genotyping-by-sequencing (GBS), which is a high-throughput sequencing approach, has remarkably increased the number of molecular markers usable in crop genetics [16,17]. The basic features of GBS rely on using restriction enzymes to reduce genome complexity and barcode adapters that allow sequencing of pooled samples. The choice among GBS methods is generally based on the genome size of the studied crop, the extent of linkage disequilibrium and level of heterozygosity of the studied panel, and cost-efficiency considerations. Unlike the earlier low-throughput approaches based on restriction fragment length polymorphism (RFLP) or simple sequence repeats (SSR), GBS enables the identification and genotyping of a massive quantity of single nucleotide polymorphisms (SNPs). These SNPs can be associated with agronomical traits of interest and then used in marker-assisted breeding or to validate trait-linked haplotypes in crops [17][18][19]. This strategy has been successfully used in many important crops [20]. For instance, GBS methods were employed to genotype recombinant inbred line (RIL) populations in rice [20,21], maize and barley [22] and doubled-haploid (DH) populations in wheat [23] in view of QTL mapping. GBS was also applied to provide adequate marker density for **genome-wide association study** (GWAS) of rice traditional populations [24], rice and chickpea (*Cicer arietinum*) multiparent advanced generation intercrosses (MAGIC) [25] and maize (*Zea mays*) nested association mapping populations (NAM) [26].

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

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

Approaches to improve QTL and candidate gene detection

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

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

QTL-seq approaches

A first example of such an approach is QTL-seq (Figure 1A), which is a modern version of the classical BSA combined with WGR [35]. In this approach, a mapping population derived from a cross between two contrasted parents is used. The progenies are phenotyped, and the tails of the distribution are divided into two extreme bulks of 10-20 individuals, which are sequenced at above 6x coverage. For each genomic position, the proportion of short reads harboring SNPs with the sequence of one of the parents chosen as reference (so-called SNP-index) is estimated and the difference between the SNP-index of the low trait-bulk and that of the high trait-bulk, called Δ(SNP-index), is calculated. A large Δ(SNP-index) characterizes the genomic fraction that has an association with the phenotypic value [36]. In chickpea, QTL-seq was applied to two 100-seed weight (SW)-contrasted bulks, each bulk containing 10 F_4 homozygous individuals, which had been produced by single-seed descent (SSD) from a cross between high SW and low SW landraces [37]. A major QTL was detected on chromosome 1. One SNP tightly linked with the SW-QTL was further identified in the coding region of the constitutive photomorphogenic9 (*COP9*) signalosome complex subunit 8 (*CSN8*) gene. This gene was specifically expressed in seeds and was up/downregulated during seed development in high/low SW parent and homozygous mapping individuals, respectively. Moreover, a functional molecular diversity analysis showed that the coding SNP was completely absent from wild accessions while it discriminated the cultivated genotypes, the high and low SW parents and the two bulk mapping individuals. Therefore, QTL-seq combined with differential expression profiling and diversity analysis proved to be efficient not only in scaling-down QTL size, but also in rapidly enabling potential candidate gene identification. The same approach has been successfully used in other crops, such as foxtail millet (*Setaria italica*) [45] and rice [49].

Another approach derived from QTL-seq is multiple QTL-seq (mQTL-seq), which can be defined as QTL-seq applied to several mapping populations derived from crosses with at least one common parent (Figure 1 B) [38]. The utilization of multiple mapping populations representing a broader genetic diversity was beneficial for the validation of QTLs, along with narrowing down the detected QTLs to shorter segments for several agronomic traits in chickpea, such as pod number per plant (PN) [38] or plant height [39]. For example, mQTL-seq applied independently to two F_5 mapping populations of chickpea allowed the identification of common significant genomic regions. For each population, two bulks of 10 lines with low/high PN were built. Two major QTLs associated with PN that were previously detected using the entire population were scaled down: *CaqaPN4.1* from 868 kb to 638 kb and *CaqbPN4.2* from 1.8 Mb to 1.3 Mb. Furthermore, mQTL-seq identified a regulatory SNP governing PN in the pentatricopeptide repeat (*PPR*) gene. A gene expression study demonstrated that the *PPR* gene was strongly upregulated in the high-PN bulks and the high-

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

MutMap approaches

MutMap approaches combine NGS with BSA in the analysis of a mutated population. Mutagenesis is a classic way to produce material useful in determining the function of a candidate gene. Where QTL-seq uses two contrasted bulks of individuals from any mapping population, MutMap (Figure 1C) is a method based on WGR using bulked segregants which are derived from cross between a homozygous recessive mutant and its wild-type parental line [30,40]. The F_2 population is phenotyped and only plants showing the recessive mutant phenotype are bulked. The parental genome sequence is used as the template to detect causal SNPs underlying the mutant phenotype. As with QTL-seq, a SNP-index is computed for each SNP position. MutMap is actually a simplified version of QTL-seq with only the mutant-phenotype bulk sequenced and no possibility to distinguish **segregation distortions** from a true QTL effect. It works only if the mutant allele is recessive and if the mutant phenotype can be easily distinguished from the wild phenotype in F_2 plants. It is applicable in cases of crosses between a mutant and its wild-type progenitor rather than crosses between genetically distant lines. MutMap should probably also be avoided when targeting mutations with small or subtle effects. This method was recently used to isolate mutations causing pale green leaves and semidwarfism in rice [40], and the many-noded dwarf (*mnd*) in barley [40,41]. MutMap was also successfully used to identify the causative gene, *OsRR22*, from a salt-tolerant rice mutant called *hitomebore salt tolerant 1* (*hst1).* Subsequently, the introgression of the *hst1* allele into the elite cultivar Hitomebore by successive backcrosses enabled the release of the improved variety Kaijin, which differed from Hitomebore wild type by only 201 SNPs but had the same salt tolerance as the *hst1* mutant. With the application of MutMap, the new salt-tolerant elite variety Kaijin was developed in only two years and contributed to the restoration of rice production in tsunami-affected areas of Japan [42].

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

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

Figure 1

NGS-assisted expression profiling

NGS-assisted expression profiling identifies candidate genes having transcripts linked with the phenotype of interest. The availability of NGS-based transcriptome-wide tools provides precise information about the abundances of gene transcripts [45]. Gene expression analysis is a method that has been frequently used to screen among candidate genes underlying a QTL in different crops, e.g., chickpea, potato, and rice [46–48]. A gene becomes a causative candidate when evidence coming from QTL mapping coincides with transcription activities in the conditions where the phenotype of interest was observed. In this context, the expression level of the causative candidate gene correlates with the phenotypic value. Recently, RNA-seq, the direct sequencing of complementary DNA (cDNA) derived from RNA extracts, has been used to cater comprehensive expression profiling of QTL genes in different tissues and organs of contrasted genotypes [49]. RNA-seq provides a global view of the protein-coding regions that only occupy 1-2% of the genome but include many functional variations [50]. RNA-seq is an exceptional method to overcome the limitations of previous expression microarrays in which the dissection of different transcripts was dependent on probes designs [51][52]. For example, RNA-seq was performed on the sorghum root tissues of two sorghum (*Sorghum bicolor*) varieties used as parents of a mapping population and revealed that 108 gene transcripts involved in nitrogen metabolism, plant hormone metabolism and glycolysis were differentially expressed. These genes were located in the vicinity of QTLs detected in the mapping population that regulated multiple agronomic traits under normal and low nitrogen conditions [53]. In maize, published RNA-seq data combined with meta-QTL analysis facilitated the identification of candidate genes involved in kernel row number [54]. In soybean (*Glycine max*), RNA-seq contributed to the identification of a novel salt-tolerance gene from a highly salt tolerant wild accession. A combination of two approaches (*de novo* sequencing of the wild accession and QTL mapping in a population derived from a cross between the wild accession and a cultivated one) was used. The results were validated using resequencing data from 23 soybean accessions with contrasted levels of salinity tolerance. *GmCHX1* was identified as the causal gene and shown to encode an ion transporter that reduces the Na+/K+ ratio under salt stress [55].

While QTL mapping enables significant regions related with a trait to be identified, functional genomic analysis, with the support of NGS, provides complete RNA profiles to determine the expression of QTL genes in specific biological conditions. The integration of these two strategies results in the detection of **expression quantitative trait loci** (eQTL) that enable the expression of complex traits governed by multiple QTLs/genes to be explained [56,57] (Figure 2). In the eQTL approach, segregating populations are both genotyped and phenotyped by expression profiling methods such as microarray or RNA-seq to collect the information of transcript abundance. Rather than microarray, RNA-seq is becoming the technique of choice in eQTL analyses because it can determine allele-specific expression and isoform-RNA expression [58]. Thousands of RNA expression levels are analyzed for linkage or association with genetic markers, leading to the detection of variations acting in *cis* or *trans* manners. *Cis*-acting factors are DNA variations located within or near a differentially expressed gene and regulating its transcription. *Trans*-acting factors are distantly mapped elsewhere in the genome and influence the activity of transcription factors that regulate the differentially expressed gene [58,59]. Using this approach in maize, a strong trans-acting eQTL has been successfully fine mapped to an interval of only 186 bp within a class I glutamine amidotransferase domain containing gene [60]. Under the effect of this eQTL, the transcription level of another gene encoding an ABA 8'-hydroxylase was upregulated to 6-fold greater in one parental genotype compared to the other. Although the regulatory mechanisms involving the glutamine amidotransferase protein on ABA 8'-hydroxylase gene expression remained unclear, the cloning of this trans-acting eQTL showed the efficiency of the eQTL approach to identify causative genes. Furthermore, coexpression network databases compiling a large number of microarray studies were developed to further help in identifying functionally related genes. For instance, RiceFREND (http://ricefrend.dna.affrc.go.jp) was helpful in detecting shared expression networks between candidate genes for panicle development in rice [61,62].

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

Figure 2

Polymorphism databases expedite the identification of candidate genes

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

SNP-Seek, the 3K project database (http://snp-seek.irri.org/), enabled immediate *in silico* access to sequence variations including SNPs and InDels for the target segment in rice. This resource allowed the validation of a QTL haplotype by identifying varieties that carried either contrasted haplotypes or recombinant haplotypes, phenotyping these varieties, and detecting which allelic variation was responsible for the QTL effect [70,71]. For instance, SNP-Seek facilitated the prediction of novel genes/alleles of resistance to rice blast disease based on sequence and structure variations between the resistant haplotypes and the susceptible ones [72]. In another example, SNP-seek was used to detect mutations in the Effector Binding Elements (EBE) of promoters of rice genes favorable to the proliferation of bacterial blight, making impossible the recognition of EBE by the bacteria Transcription Activator-like Effectors (TALE). Such mutations could improve plant resistance against the bacteria. The mining of such mutations in the 3K database combined with a rapid phenotyping for bacterial blight resistance is used to detect new sources of resistance [72].

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

Concluding remarks

With its broad applications, NGS is becoming an essential tool for crop geneticists to identify and characterize genomic variations associated with agronomical traits. WGR and transcription profiling that contribute to provide comprehensive information on genetic variability and their regulatory mechanisms are the most popular applications of NGS. QTL-seq, MutMap and their extended versions showed efficiency in narrowing down the position of QTLs and precisely detecting their causative variations. RNA-seq provided functional context to candidate genes. As such, a large number of QTLs/eQTLs were found in attempts to break down the genetic mechanisms regulating important agronomic traits.

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

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

ACKNOWLEDGEMENTS

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

References

1 Kole, C. *et al.* (2015) Application of genomics-assisted breeding for generation of climate resilient crops□: progress and prospects. *Front. Plant Sci.* 6, 1–16

2 Zuo, J. and Li, J. (2014) Molecular genetic dissection of quantitative trait loci regulating rice grain size. *Annu. Rev. Genet* 48, 99–118

3 Gupta, P.K. *et al.* (2017) QTL analysis for drought tolerance in wheat□: present status and future possibilities. *Agronomy* 7, 5

4 Moury, B. *et al.* (2017) Quantitative resistance to plant pathogens in pyramiding strategies for durable crop protection. *Front. Plant Sci.* 8, 1–9

5 Bailey-serres, J. *et al.* (2010) Submergence tolerant rice□: SUB1' s journey from landrace to modern cultivar. *Rice* 3, 138–147

6 Salvi, S. and Tuberosa, R. (2015) The crop QTLome comes of age. *Curr. Opin. Biotechnol.* 32, 179–185

7 Das, G. *et al.* (2017) Insight into MAS□: A molecular tool for development of stress resistant and quality of rice through gene stacking. *Front. Plant Sci.* 8, 1–9

8 Goffinet, B. and Gerber, S. (2000) Quantitative Trait Loci□: A Meta-analysis. *Genetics* 155, 463–473

9 Chardon, F. *et al.* (2004) Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics* 2185, 2169–2185

10 Zhang, X. *et al.* (2016) Meta-analysis of major QTL for abiotic stress tolerance in barley and implications for barley breeding. *Planta* 245, 283–295

11 Zhang, L. *et al.* (2010) Genomic distribution of quantitative trait loci for yield and yield-related traits in common wheat. *J. Integr. Plant Biol.* 52, 996–1007

12 Courtois, B. *et al.* (2009) Rice root genetic architecture: Meta-analysis from a drought QTL database. *Rice* 2, 115–128

13 Salvi, S. and Tuberosa, R. (2005) To clone or not to clone plant QTLs□: present and future challenges. *Trends Plant Sci.* 10, 297–304

14 Sanger, F. and Nicklen, S. (1977) DNA sequencing with chain-terminating. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467

15 Egan, A.N. *et al.* (2012) Applications of next - generation sequencing in plant biology. *Am. J. Bot.* 99, 175–185

16 Poland, J.A. and Rife, T.W. (2012) Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5, 92–102

17 He, J. *et al.* (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* 5, 484

18 Varshney, R.K. *et al.* (2014) Harvesting the promising fruits of genomics□: applying genome sequencing technologies to crop breeding. *Plos Biol.* 12, 6

19 Thomson, M.J. (2014) High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed. Biotechnol.* 2, 195–212

20 Spindel, J. *et al.* (2013) Bridging the genotyping gap: Using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor. Appl. Genet.* 126, 2699–2716

21 Huang, X. *et al.* (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19, 1068–1076

22 Elshire, R.J. *et al.* (2011) A robust , simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, 1–10

23 Chapman, J.A. *et al.* (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* 16, 1–26

24 Phung, N.T.P. *et al.* (2016) Genome-wide association mapping for root traits in a panel of rice accessions from Vietnam. *BMC Plant Biol.* 16, 64

25 Huang, B.E. *et al.* (2015) MAGIC populations in crops□: current status and future prospects. *Theor. Appl. Genet.* 128, 999–1017

26 Lu, F. *et al.* (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* 6, 6914

27 Varshney, R.K. *et al.* (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27, 522–530

28 Scheben, A. *et al.* (2017) Genotyping-by-sequencing approaches to characterize crop genomes□: choosing the right tool for the right application. *Plant Biotechnol. J.* 15, 149–161

29 Michelmore, R.W. *et al.* (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis - a rapid method to detect markers in specific genome regions by using segregating populations. *Proc. Natl. Acad. Sci.* 88, 9828–9832

30 Zou, C. *et al.* (2016) Bulked sample analysis in genetics , genomics and crop improvement. *Plant Biotechnol. J.* 14, 1941–1955

31 Thanda, K. *et al.* (2016) QTL mapping for downy mildew resistance in cucumber via bulked segregant analysis using next□generation sequencing and conventional methods. *Theor. Appl. Genet.* 130, 199–211

32 Wambugu, P. *et al.* (2017) Sequencing of bulks of segregants allows dissection of genetic control of amylose content in rice. *Plant Biotechnol. J.* 1, 1–11

33 Hayward, A. *et al.* (2015) Molecular marker applications in plants. *Methods Mol. Biol.* 1245, 101–18

34 Magwene, P.M. *et al.* (2011) The statistics of bulk segregant analysis using next generation sequencing. *PloS Comput. Biol.* 7, 1–9

35 Terauchi, R. *et al.* (2015) Whole genome sequencing to identify genes and QTL in rice. In *Advances in the understanding of biological sciences using next generation sequencing (NGS) approaches* 1pp. 33–42

36 Takagi, H. *et al.* (2013) QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations.

Plant J. 74, 174–183

37 Das, S. *et al.* (2014) Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Res.* 22, 193–203

38 Das, S. *et al.* (2015) MQTL-seq delineates functionally relevant candidate gene harbouring a major QTL regulating pod number in chickpea. *DNA Res.* 23, 53–65

39 Parida, S.K. *et al.* (2017) A genome-wide mQTL-seq scan identifies potential molecular signatures regulating plant height in chickpea. *Plant Mol. Biol. Report.* 35, 273–286

40 Abe, A. *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30, 174–178

41 Mascher, M. *et al.* (2014) Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biol.* 15, R78

42 Takagi, H. *et al.* (2011) MutMap accelerates breeding of a salt-tolerant rice cultivar. *Nat. Biotechnol.* 1, 1–5

43 Fekih, R. *et al.* (2013) MutMap+: Genetic Mapping and Mutant Identification without Crossing in Rice. *PLoS One* 8, 1–10

44 Takagi, H. *et al.* (2013) MutMap-Gap: Whole-genome resequencing of mutant F2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii. *New Phytol.* 200, 276–283

45 Gedil, M. *et al.* (2016) Perspectives on the application of next-generation sequencing to the improvement of Africa’ s staple food crops. In *Next generation sequencing - Advances, applications and challenges* pp. 287–321

46 Kujur, A. *et al.* (2015) A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. *Sci. reports* 5, 11166

47 Kloosterman, B. *et al.* (2010) From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. *BMC Genomics* 11, 158

48 Daware, A. *et al.* (2016) An efficient strategy combining SSR markers- and advanced QTL-seq-driven QTL mapping unravels candidate genes regulating grain weight in rice. *Front. Plant Sci.* 7, 1–17

49 Wang, Z. *et al.* (2010) RNA-Seq□: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63

50 Li, X. *et al.* (2012) Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.* 22, 2436–2444

51 Martin, J.A. and Wang, Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682

52 Kudo, T. *et al.* (2016) Identification of reference genes for quantitative expression analysis using large-scale RNA-seq data of Arabidopsis thaliana and model crop plants. *Genes Genet. Syst.* 91, 111–125

53 Gelli, M. *et al.* (2016) Mapping QTLs and association of differentially expressed gene transcripts for multiple agronomic traits under different nitrogen levels in sorghum. *BMC Plant Biol.* 16, 16

54 Jiang, Q. *et al.* (2016) Combining meta-QTL with RNA-seq data to identify

candidate genes of kernel row number trait in maize. *Crea J.* 61, 4

55 Qi, X. *et al.* (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nat. Commun.* 5, 4340

56 Majewski, J. and Pastinen, T. (2011) The study of eQTL variations by RNA-seq: From SNPs to phenotypes. *Trends Genet.* 27, 72–79

57 Westra, H. and Franke, L. (2014) From genome to function by studying eQTLs. *Biochim. Biophys. Acta - Mol. Basis Dis.* 1842, 1896–1902

58 Sun, W. and Hu, Y. (2012) eQTL mapping using RNA-seq data. *Stat. Biosci.* 5, 189–219

59 Cubillos, F.A. *et al.* (2012) Lessons from eQTL mapping studies: Non-coding regions and their role behind natural phenotypic variation in plants. *Curr. Opin. Plant Biol.* 15, 192–198

60 Holloway, B. *et al.* (2011) Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics* 12, 336

61 Sato, Y. *et al.* (2013) RiceFREND: A platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res.* 41, 1214–1221

62 Crowell, S. *et al.* (2016) Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. *Nat. Commun.* 7, 10527

63 Lipka, A.E. *et al.* (2015) From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24, 110–118

64 Feltus, F.A. (2014) Systems genetics: A paradigm to improve discovery of candidate genes and mechanisms underlying complex traits. *Plant Sci.* 223, 45–48

65 Lavarenne, J. *et al.* (2018) The Spring of Systems Biology-Driven Breeding. *Trends Plant Sci.* 23, 1–15

66 Xu, X. *et al.* (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–11

67 Huang, X. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967

68 Duitama, J. *et al.* (2015) Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. *PLoS One* 10, 1–20

69 Alexandrov, N. *et al.* (2015) SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.* 43, D1023–D1027

70 Mori, A. *et al.* (2016) The role of root size versus root efficiency in phosphorus acquisition in rice. *J. Exp. Bot.* 67, 1179–1189

71 Wissuwa, M. *et al.* (2016) From promise to application: root traits for enhanced nutrient capture in rice breeding. *J. Exp. Bot.* 67, 3605–3615

72 Leung, H. *et al.* (2015) Allele mining and enhanced genetic recombination for rice breeding. *Rice* 8, 34

73 Lyu, J. *et al.* (2013) Analysis of elite variety tag SNPs reveals an important allele in upland rice. *Nat. Commun.* 4, 2138

74 Doddamani, D. *et al.* (2015) CicArVarDB□: SNP and InDel database for

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Box1. NGS-based genotyping approaches used in crop genetics

Reduced-representation sequencing (RRS)

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

Whole genome resequencing (WGR)

This method supposes that a reference genome is available. Genomic DNA is sheared, ligated to adaptors and amplified. The amplified PCR products are then separated by size and purified to provide the sequencing libraries. Short-reads generated from sequencing are aligned on the reference genome. Skim-based genotyping by sequencing (SkimGBS) was develop for high-resolution whole genome resequencing of mapping populations [91]. After SNPs between the parents are called, the progeny reads are mapped on the same reference and compared to parental SNP data to determine the genotypes and recombination frequencies. In addition, a sliding window approach, which examines collectively consecutive SNPs instead of assessing SNPs individually, was proposed as a method to avoid erroneous SNP calling [21]. Compared to RRS approaches, WGR eliminates several steps in the preparation of sequencing libraries and provides a high-throughput genotyping with low cost per marker point. The polymorphisms detected by WGR are more comprehensive, including not only SNPs but also structural variations, gene conversions, recombination break points, etc. However, the cost per sample remains high depending on the chosen coverage and crop genome size.

Figure 1: Different approaches combining bulk segregant analysis and whole genome resequencing developed to identify genetic variations controlling valuable traits. Mapping populations are generated from biparental crosses (QTL-seq; mQTL-seq) or from crosses between a wild-type and its mutant (MutMap, MutMap-gap); In MutMap+, no cross is generated; The M_2 and M3 generations are obtained from M_1 and M2, respectively, by selfing; The portions of the population that are pooled as DNA bulks and sequenced are hatched. A SNP index is calculated to identify SNPs linked with the trait of interest. P: parent; WT: wild-type; LB: low-trait bulk; HB: high-trait bulk; MB: mutant bulk; WTB: wild-type bulk.

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

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

Glossary

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

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

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

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

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

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

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

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

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

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

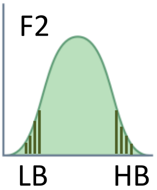
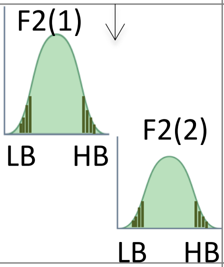
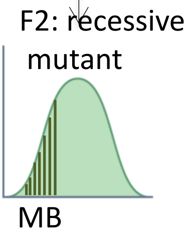
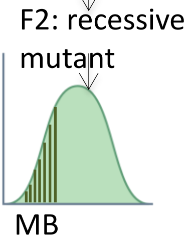
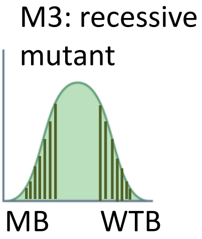
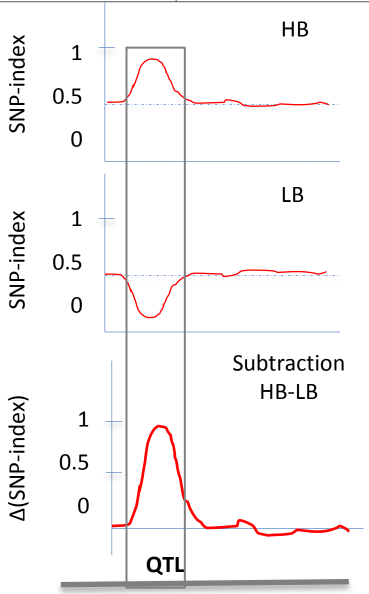
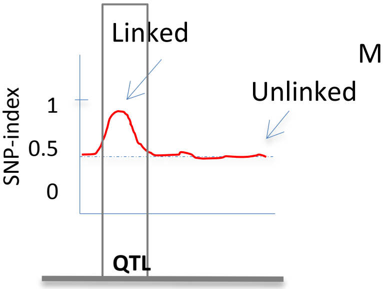
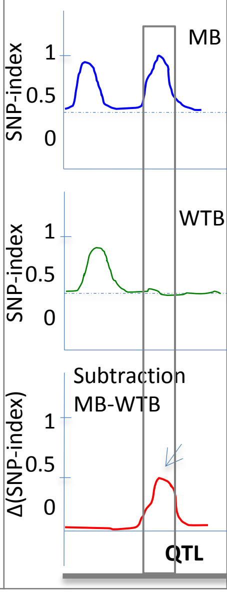
Recombinant inbred lines (RILs): homozygous lines derived from a biparental cross obtained by selfing plants during several generations up to fixation.

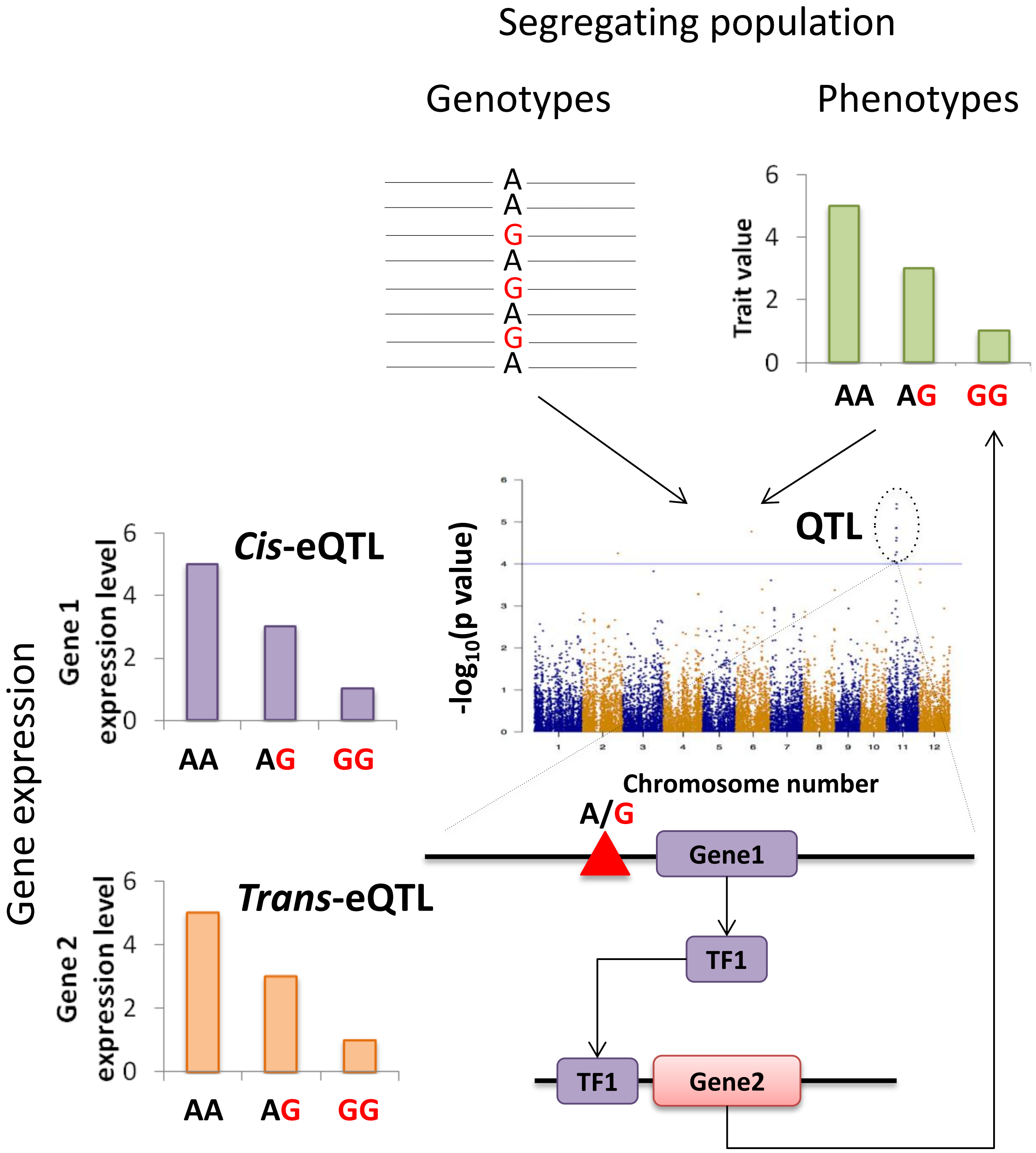
RNA-sequencing: a method to detect the presence and quantity of RNA in a given sample. The total RNA extracted from each sample is converted to cDNA, then sequenced by an NGS platform.

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

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

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

	(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 -> *De novo* sequencing -> WT-specific regions WT x Mutant F1 F2 ↓	WT-> Mutant ↓ M2 ↓ M3
Bulked samples					
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
A
A
G
A
G
A
G
A
Trait value
6
4
2
0
AA
AG
GG
QTL
-log$_{10}$(p value)
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
Chromosome number
Gene expression
Cis-eQTL
Gene 1 expression level
6
4
2
0
AA
AG
GG
Trans-eQTL
Gene 2 expression level
6
4
2
0
AA
AG
GG
A/G
Gene1
TF1
TF1
Gene2