the FEBS Journal

REVIEW ARTICLE

Genome walking in eukaryotes

Claudia Leoni[1], Mariateresa Volpicella[1], Francesca De Leo[2], Raffaele Gallerani[1] and Luigi R. Ceci[2]

1 Department of Biochemistry and Molecular Biology, University of Bari, Italy
2 Institute of Biomembranes and Bioenergetics, Italian National Research Council, Bari, Italy

Keywords
gene tagging; gene therapy; gene trapping; genome walking; insertional mutagenesis; multigene family; next generation sequencing; T-DNA; transposable elements; virus integration

Correspondence
L. R. Ceci, Institute of Biomembranes and
Bioenergetics, CNR, Via Amendola 165/A,
70126 Bari, Italy
Fax: +39 080 544 3317
Tel: +39 080 544 3311
E-mail: l.ceci@ibbe.cnr.it

(Received 1 April 2011, accepted 12 August 2011)

doi:10.1111/j.1742-4658.2011.08307.x

Genome walking is a molecular procedure for the direct identification of nucleotide sequences from purified genomes. The only requirement is the availability of a known nucleotide sequence from which to start. Several genome walking methods have been developed in the last 20 years, with continuous improvements added to the first basic strategies, including the recent coupling with next generation sequencing technologies. This review focuses on the use of genome walking strategies in several aspects of the study of eukaryotic genomes. In a first part, the analysis of the numerous strategies available is reported. The technical aspects involved in genome walking are particularly intriguing, also because they represent the synthesis of the talent, the fantasy and the intelligence of several scientists. Applications in which genome walking can be employed are systematically examined in the second part of the review, showing the large potentiality of this technique, including not only the simple identification of nucleotide sequences but also the analysis of large collections of mutants obtained from the insertion of DNA of viral origin, transposons and transfer DNA (T-DNA) constructs. The enormous amount of data obtained indicates that genome walking, with its large range of applicability, multiplicity of strategies and recent developments, will continue to have much to offer for the rapid identification of unknown sequences in several fields of genomic research.

Introduction

Identification of unknown nucleotide sequences flanking already characterized DNA regions can be pursued by a number of different PCR-based methods commonly known as genome walking (GW).

In times of high-throughput DNA sequencing technologies, when more than 1000 genomes have been completely sequenced, the development of GW strategies can appear as an out-of-date laboratory activity. Nevertheless, papers describing applications of GW methods and improvements of several available strategies continue to be published with a steady positive trend. Reasons for such constant interest can be found both in the relatively low difficulty of the different strategies, which do not require expensive equipment or highly trained personnel, and in the increasing possibilities of applying GW methods to eukaryotic genomes. Furthermore, in some highly sophisticated applications, GW strategies have recently been combined with pyrosequencing technology allowing the production of hundreds of thousands of sequences per

Abbreviations
ASLV, avian sarcoma-leukosis virus; blocked DLA, blocked digestion–ligation–amplification; EC-PCR, extension and cassette PCR; E-GW, extension-based GW; ET-PCR, extension and tailing PCR; FLEA-PCR, flanking sequence exponential anchored PCR; GW, genome walking; I-PCR, inverted-PCR; LAM-PCR, linear amplification mediated PCR; LM-PCR, ligation mediated PCR; MLV, murine leukaemia virus; P-GW, primer-based GW; RAGE, rapid amplification of genomic ends; R-GW, restriction-based GW; SHP-PCR, sequential hybrid primer PCR; TAIL-PCR, thermal asymmetric inter-laced PCR; T-DNA, transfer DNA; TVL-PCR, TOPO® vector-ligation PCR; UFW, universal fast walking.

single experiment and opening new application areas for GW.

Review articles on GW are not numerous. After a first paper by Hengen dated 1995 [1], where a limited number of strategies were compared, a complete survey of the available strategies was published by Hui *et al.* [2] more than a decade ago. Recently, two reviews have been dedicated to the description of GW strategies and their applications, but limited to microorganisms [3,4]. This review is intended to provide the missing information by describing the application of GW techniques to eukaryotic genomes. Numerous reports can be found in which such technology has been successfully used, avoiding in many cases the time-consuming process of the construction and screening of large genomic libraries.

A first section gives a general overview of the available GW methods, classified according to the basic strategy adopted. Most of these methods can be easily executed in any molecular biology laboratory. In addition, a list of commercial resources (kits and customer services) is also provided. A second part of the paper deals with the different applications of GW. Main areas of interest have been identified and the results obtained for specific applications are reported. Owing to the large diffusion of GW methods, this survey cannot by any means be complete. We have done our best to show the huge potential of these methods and we apologize to colleagues whose work has not been reported.

GW methods and resources

GW methods

GW methods differ in the strategies adopted to obtain the substrate for a final PCR step, in which a primer specific for the known sequence (sequence specific primer) is coupled with a primer dictated by the specific strategy of walking (walking primer).

In Table 1, the basic GW methods and related improvements that have been developed are listed. Methods are classified into three different groups according to their first (and sometimes conditioning) step: restriction-based (R-GW) methods, primer-based (P-GW) methods and extension-based (E-GW) methods. GW methods using a combination of two of the basic strategies are catalogued according to the first step of the procedure. Most of the methods listed in Table 1 have already been critically reviewed [1–4] and are not described further in this paper. Only more recent methods, together with those previously not reviewed, are examined in this report. These methods are highlighted in Table 1. Graphical representations of the strategies at the basis of GW methods are schematically reported in Fig. 1.

R-GW methods require a preliminary digestion of the genomic DNA by suitable restriction enzymes, whose sites must be located at a proper distance from the boundary between known and unknown sequences (not too far in order to allow subsequent PCR amplification; not too close to avoid amplification of short fragments). Restriction fragments can then be either self-circularized or ligated to specifically designed adaptors.

In the first case, the sub-group of the inverted-PCR (I-PCR) methods, first described by Triglia *et al.* [5], is obtained (Table 1). An improvement to this strategy, named rolling circle I-PCR, has recently been reported [8]. In this case circularized genomic DNA restriction molecules are subjected to rolling circle amplification by using random hexamers and employing the strand-displacement property of Phi29 polymerase.

In the latter case, a wide range of methods have been developed starting from the first strategy named single-specific-primer PCR [9]. These methods are catalogued according to Tonooka and Fujishima [3] as ‘cassette PCR’, for the use of double-stranded DNA linkers to ligate to genomic DNA restriction fragments. We prefer the term ‘cassette PCR’ instead of ‘ligation mediated PCR’, which is sometimes also used to indicate these methods ([7,21,22,28,44], for example), since the term ‘ligation mediated’ has been used since 1989 to indicate one of the first GW strategies, here classified among the E-GW methods (Table 1). Once the cassette has been ligated to genomic DNA restriction fragments, generating what is commonly known as the GW library, a PCR amplification of the region encompassing the boundary between the known and unknown sequences can be carried out using a sequence specific primer and a cassette specific primer. One major concern in the cassette PCR based methods is the background of non-specific products due to the cassette specific primer. To overcome this problem a number of tricks have been devised, such as those adopted in vectorette PCR [11], capture PCR [13] and other strategies reported below.

The strategy known as transfer DNA (T-DNA) fingerprinting PCR [20] adopted for GW an amplified fragment length polymorphism method developed for studying the number of *Agrobacterium tumefaciens* T-DNA insertions in transgenic plants. Amplified fragments corresponding to T-DNA/plant DNA junctions, identifiable thanks to a labelled T-DNA specific primer, are eluted from polyacrylamide gel, re-amplified and sequenced.

Table 1. GW strategies. GW methods are catalogued as R-GW, P-GW and E-GW. When available, the specific name of the method as given by the authors is reported. Otherwise the name of the first author is given. Methods in gray boxes are detailed in the GW methods section. P and E columns refer to the analysis of prokaryotic and eukaryotic genomes, respectively.

	Group	Sub-group	Name	References	P	E
1	R-GW	Inverted PCR	Inverted PCR	5		×
2			Long range inverted PCR	6		×
3			Bridged inverted PCR	7	×	
4			Rolling circle inverted PCR	8		×
5		Cassette PCR	Single-specific-primer PCR	9	×	
6			Fors *et al.*	10		×
7			Vectorette PCR	11		×
8			Cassette ligation	12		×
9			Capture PCR	13		×
10			Splinkerette PCR	14		×
11			Boomerang DNA amplification	1		×
12			Suppression PCR GW	15		×
13			Padegimas and Reichert	16		×
14			Step-down PCR	17		×
15			Simplified oligo-cassette	18	×	
16			Cottage *et al.*	19		×
17			T-DNA fingerprinting PCR	20		×
18			T-linker PCR	21		×
19			Versatile cassette	22	×	
20			Barcoding pyrosequencing	23,24		×
21			One-base excess adaptor ligation	25		×
22			Straight walk	26		×
23			Blocked DLA	27		×
24			Template-blocking PCR	28	×	×
25			TVL-PCR	29		×
26		Others	Panhandle PCR	30		×
27			Supported PCR	31		×
28			RAGE	32		×
29			Restriction site extension PCR	33		×
30	P-GW		Targeted gene-walking PCR	34		×
31			Restriction-site PCR	35		×
32			Restricted PCR (novel Alu PCR)	36		×
33			TAIL-PCR	37		×
34			Uneven PCR	38		×
35			Semi-random PCR	39	×	
36			Mishra *et al.*	40		×
37			UFW PCR	41,42		×
38			Lariat-dependent nested PCR	43		×
39			Site finding PCR	44		×
40			Touchdown PCR-based	45	×	
41			Walser *et al.*	46		×
42			Nested PCR-based	47	×	
43			Self-formed adaptor PCR	48	×	
44			Two-step gene walking PCR	49	×	
45			High-genome walking	50		×
46			SD-PCR	51	×	
47			SHP-PCR	52	×	
48	E-GW	EC-PCR	LM-PCR	53		×
49			LAM-PCR	54		×
50		ET-PCR	Long distance genome walking PCR	55		×
51			Leoni *et al.*	56,57		×
52		Others	Single-primer amplification	58		×
53			FLEA-PCR	59		×

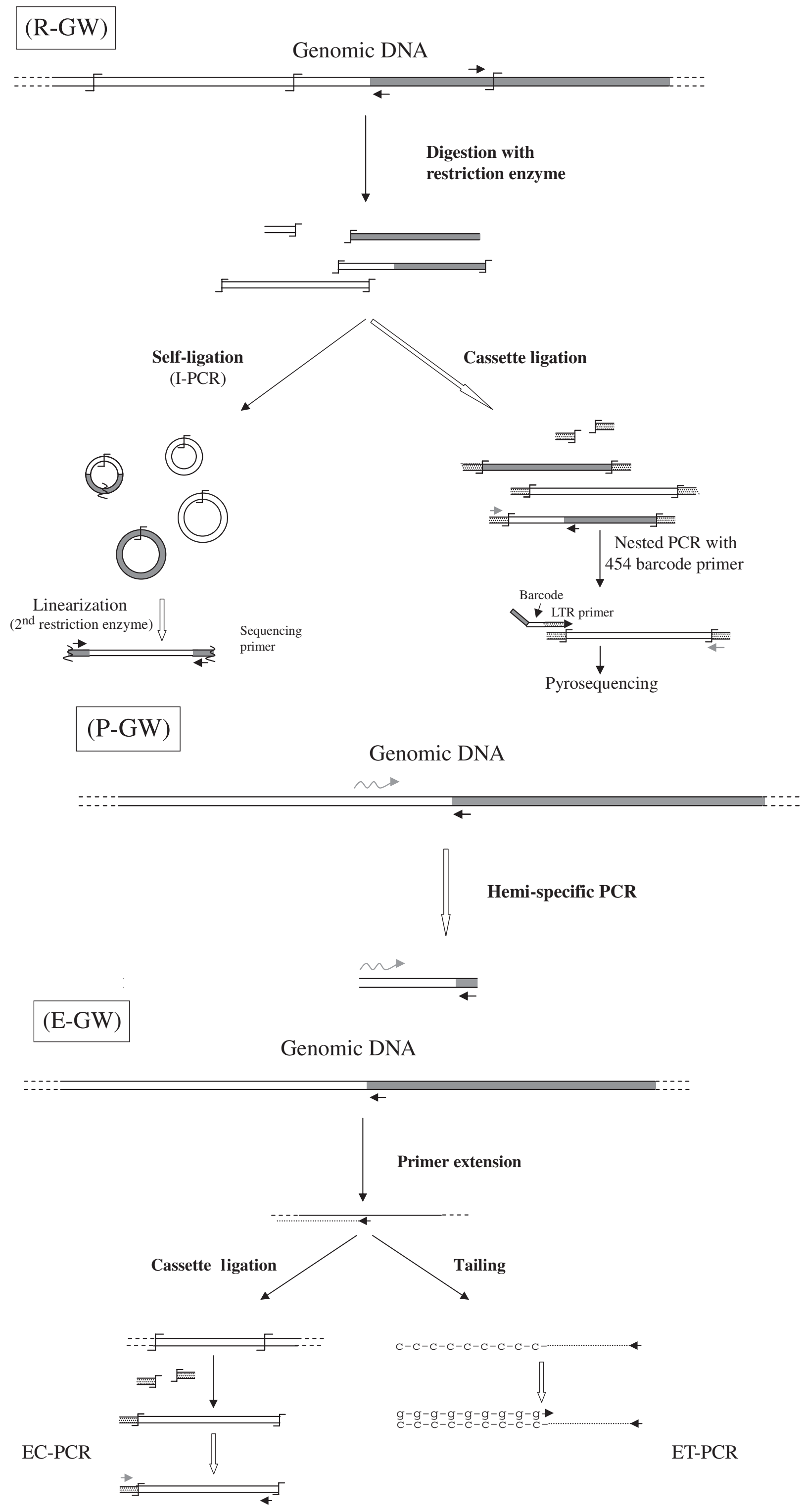

Fig. 1. Main GW strategies. R-GW, P-GW and E-GW strategies are schematically illustrated. Gray and white regions correspond to known and unknown DNA sequences, respectively. Horizontal black arrows correspond to sequence specific primers; gray arrows indicate cassette specific primers. A wavy gray arrow corresponds to a random (or degenerate) primer. Dotted lines indicate *in vitro* synthesized ssDNA. The symbols ⌐ and ≷ indicate restriction sites. White vertical arrows correspond to the penultimate step of the procedure, after which (with the exception of the pyrosequencing procedure) the obtained product can be subjected to PCR (nested PCR), cloning and sequencing.

The association of a cassette PCR GW method with the powerful pyrosequencing technology introduced by Wang *et al.* [24] is of great interest since it allows massive parallel sequencing of thousands and thousands of amplification fragments. A further improvement to this method was obtained by the same group, by application of a DNA barcoding strategy [23] (Fig. 1). By using cassette primers with different barcodes, authors were able to identify more than 160 000 integration sites for lentiviral and gamma-retroviral vectors in several tissue samples from mice. A similar approach was reported for the analysis of maize transposons by Liu *et al.* [27] in the blocked digestion–ligation–amplification (blocked DLA) method. The method adopts a single-strand adaptor provided with a 3′-terminus annealable with 3′-overhangs of genomic restriction fragments obtained by *NspI* (RCATG•Y) digestion. After the first step, the library fragments are blocked in 3′-termini by addition of dideoxynucleotide and then amplified with sequence specific and adaptor primers. In the so-called *Mu*Clone strategy, DLA is adapted to the identification of gene sequences flanking the highly active maize *mutator* (*Mu*) transposon, by using nested gene specific primers ending, downstream of the CATG *NspI* sequence, with all the possible three nucleotide tags starting with a C or T (so to be compatible with the complete *NspI* site). By combining the *Mu*Clone strategy with pyrosequencing technology the authors obtained about 965 000 reads [60].

Among the latest R-GW strategies, the straight walk method [26] is characterized by the use of a 3′-NH_2 blocked double-stranded linker, to avoid the fill-in reaction by the DNA polymerase and preventing extension of the walking primer in the first PCR cycle. The ssDNA produced from the sequence specific primer is then the substrate for the PCR amplification. More recently, in the template-blocking PCR method [28] genomic DNA restriction fragments are 3′-blocked by addition of a dideoxynucleotide before ligation with the DNA cassette, ensuring that subsequent PCR can start only from the specific sequence primer. Taking advantage of the ligation activity present on the pCR®4-TOPO® vector (Invitrogen, San Diego, CA, USA), Orcheski and Davis recently developed the TOPO® vector-ligation PCR (TVL-PCR) method [29], an improved strategy derived from the T-linker PCR [21] which overcomes the addition of a ligase enzyme.

Under the R-GW group, four additional methods can be catalogued in which genomic DNA restriction fragments are not directly ligated to the double-strand DNA cassette. In panhandle PCR [30] a single-strand oligonucleotide is ligated to allow subsequent PCR amplification of the unknown region.

The supported PCR method [31] combines restriction digestion of the genomic DNA with linear amplification of a sequence specific primer. A biotinylated DNA fragment is synthesized by elongation of a gene specific primer using Taq DNA polymerase and bio-11-dUTP, starting with denatured DNA restriction fragments. The streptavidin-purified molecule is then ligated to a double-stranded cassette, allowing the assembly of a suitable substrate for PCR.

The rapid amplification of genomic ends (RAGE) method [32] is unique among the R-GW methods since it does not need a ligation step. In this method, restriction fragments are polyadenylated with terminal transferase before the final PCR amplification, carried out in the presence of oligo-dT and sequence specific primers.

In the restriction site extension PCR [33], a walking primer ending with a 3′-sequence corresponding to a restriction site is annealed to an ssDNA molecule obtained by elongation of a gene specific primer on restricted genomic DNA fragments. A quick elongation (5 s) of the restricted ssDNA allows a suitable template to be generated for a subsequent PCR amplification.

P-GW methods are characterized by the use of variously designed walking primers containing either degenerate or random sequences. These primers are coupled to sequence specific primers in a number of different PCR strategies. In addition to the methods already described in previous review papers [2–4], some other strategies have to be mentioned (Table 1).

The restricted PCR method [36] was used for the identification of human DNA sequences flanked by highly repetitive elements. The authors improved the applicability of the Alu element mediated PCR by introducing in the reaction mix a competitor copy of the Alu primer, which, owing to the presence of 3′-deoxyadenosine, cannot be extended by the DNA polymerase. By finding the most appropriate ratio between the Alu primer and its competitor, amplification of portions of the retinoblastoma susceptibility gene and of the ribosomal protein SI7 gene have been obtained.

An improvement of their original universal fast walking (UFW) method was achieved by Myrick and Gelbart [41] by introducing an in-gel agarase digestion for the quantitative recovery of amplicons.

The Shine-Dalgarno PCR (SD-PCR) [51] takes advantage of the presence of the so-called Shine–Dalgarno sequences in prokaryotic genomes to identify sequences upstream of specific genes. The method is based on the use of hexameric degenerate primers, based on the Shine–Dalgarno sequences, which are coupled with prokaryotic gene specific primers in the amplification reaction.

The sequential hybrid primer PCR (SHP-PCR) method [52] relies on the use of two or three PCR amplifications carried out on the product of a first amplification in which a gene specific primer and a degenerate primer are used. Successive PCR amplifications are carried out by coupling sequence specific nested primers and 'hybrid' walking primers provided with 3′-ends complementary to a target sequence of the degenerate primer (or to the hybrid primers of the preceding round) and a 5′-end which constitutes the target for the next hybrid primer.

Among the E-GW methods (Table 1) it is possible to distinguish between methods in which the final PCR step is carried out on the ligation product between a DNA cassette and the DNA synthesized by the linear amplification of a specific primer (extension and cassette PCR, EC-PCR), and methods in which the final PCR has as substrate the product of a 3′-tailing reaction performed on an ssDNA (extension and tailing PCR, ET-PCR) (Fig. 1).

The ligation mediated PCR (LM-PCR) [53] is based on the primer extension of a specific oligonucleotide carried out on a chemically nicked DNA. A double-strand DNA cassette is then ligated to the obtained blunt-end DNA providing the substrate for the final PCR. The linear amplification mediated PCR (LAM-PCR) [54] differs in the strategy used to obtain the double-strand DNA fragment to ligate with the DNA cassette. In this case a 5′-biotinylated specific primer is extended on the genomic DNA. After capture by streptavidin beads of the extension product, a second strand is synthesized by random hexanucleotide priming. The double-strand DNA obtained is then digested with a four nucleotide recognizing restriction enzyme and ligated to a proper DNA cassette. The ligation product is subjected to a nested PCR amplification, whose products are further selected by gel electrophoresis analysis, eluted, re-amplified and sequenced.

In the ET-PCR methods the final PCR amplification adopts a tail-specific walking primer coupled with a sequence specific primer. The basic idea was settled in the so-called long distance genome walking PCR [55] in which sequences of 3–4 kb were obtained starting from the elongation of primers for the *hexamerin* and *hairy* genes from mosquito and *Drosophila*, respectively. More recently a slightly different strategy has been reported by Leoni *et al.* [56,57] for the simultaneous identification of members of the light harvesting protein *Lhcb1* multigene family in the spinach genome. This strategy, based on the optimization of experimental conditions for the primer extension reaction, gives the possibility of obtaining multiple information on the different members of a multigene family by using a single, highly conserved, sequence specific primer.

Two other E-GW methods cannot be catalogued as either EC-PCR or ET-PCR strategies. The single-primer amplification [58] introduces an E-GW method based on the capture of a biotinylated ssDNA molecule obtained by extension of a sequence specific primer, and its successive PCR amplification with a nested primer. The same nested primer must first misprime and extend on the ssDNA, allowing the formation of double-strand DNA substrate, amplifiable with a single primer. The amplification fragments obtained are screened by southern hybridization before sequencing.

The flanking sequence exponential anchored PCR (FLEA-PCR) method [59] is based on the use of a walking primer, provided with a known 5′-flanking sequence and a six nucleotide degenerate 3′-terminus, to couple with the sequence specific primer in the final PCR amplification of the ssDNA.

GW kits and customer services

As an alternative to the use of in-house assembled GW methods, a number of commercial kits are available from several companies (Table 2). The most used are the Genome Walker Kit (Clontech, Mountain View, CA, USA) and the Vectorette Genomic System (Sigma, St. Louis, MO, USA). These methods rely on the frequently employed suppression PCR GW and vectorette GW methods, respectively (Table 1). The DNA Walking SpeedUp Kit (Seegene, Seoul, Korea) is based on an E-GW strategy in which short oligomers are used in combination with sequence specific primers, under optimal conditions, in a series of nested PCRs (typically three). The final product can be either directly sequenced or cloned. The TOPO Walker Kit (Invitrogen) is based on the TVL-PCR strategy (Table 1).

Additionally, two companies offer a customer GW service. The APA Walking Service (BIO S&T, Montreal, Quebec, Canada) is based on the extension of a biotinylated specific primer and its capture on streptavidin paramagnetic beads. The immobilized ssDNA is then ligated to a so-called universal walking primer, forming the substrate for a successive PCR. The Genome Walking Service (Evrogen, Moscow, Russia) is based on the suppression PCR GW (Table 1).

GW patents

The development of so many GW methods gave rise to the application of numerous patents, claiming either

Table 2. Commercial resources for GW. Names of GW methods are as in Table 1. DW-ACP, DNA walking-annealing control primer.

Kit name	Customer service	GW technique	Company (website)
Genome Walker^{TM} Kit		Suppression PCR GW	Clontech (http://www.clontech.com)
Vectorette^{TM} Genomic Systems		Vectorette	Sigma Aldrich (http://www.sigmaaldrich.com)
TOPO^{TM} Walker Kit		TVL-PCR	Invitrogen (http://www.invitrogen.com)
DNA Walking SeedUp Premix^{TM} Kit		DW-ACP	Seegene (http://www.seegene.co.kr)
	Genome Walking Service	Suppression PCR GW	Evrogen (http://www.evrogen.com)
APAgene GOLD Genome Walking Kit	APA walking service	Bio-Primer extension/ligation UWP on ssDNA	Bio S&T (http://www.biost.com)

a methodological innovation of the process or the application of a GW strategy for the resolution of a specific problem. We thought it useful to collect them in this review (Table S1). Patent retrieval was performed by using the Orbit (http://www.orbit.com) platform (Questel, Paris), a web resource specialized in intellectual property. The search was executed by browsing the FAMPAT database (Comprehensive Worldwide Patent Family Database, Questel, Paris, France), which covers patents from more than 90 national offices, grouped in invention-based families. Patents were collected using the name of GW methods (Table 1) and subdivided into processes or applications (Table S1). Information on single patents can be retrieved by their patent number from the Esp@cenet portal (http://www.espacenet.com).

GW applications

GW finds application in topics where the immediate acquisition of genomic nucleotide sequences is necessary. They can be schematically catalogued in the two main areas of insertional mutagenesis, in which the integration of DNA of viral origin, transposons and T-DNA is studied, and *de novo* sequencing, in which several aspects of genes and genome sequencing can be considered (Fig. 2). Within these areas more specific subareas can be identified which have been indicated as specific molecular objectives in the third column of Fig. 2. Furthermore, molecular objectives can be associated with different applications (fourth column of Fig. 2). In the following sections specific applications of GW are reported in accordance with the scheme of Fig. 2.

Insertional mutagenesis

The genomes of both prokaryotes and eukaryotes are often the site of insertion of viral DNA and transposable elements. These natural events have been used as tools for genetic studies in medicine and biotechnologies. In order to report about the contribution and development of GW techniques in the field of insertional mutagenesis we divided the subject according to the molecular objective of the investigation, i.e. identification of the insertion sites of DNA of viral origin, transposons or T-DNA.

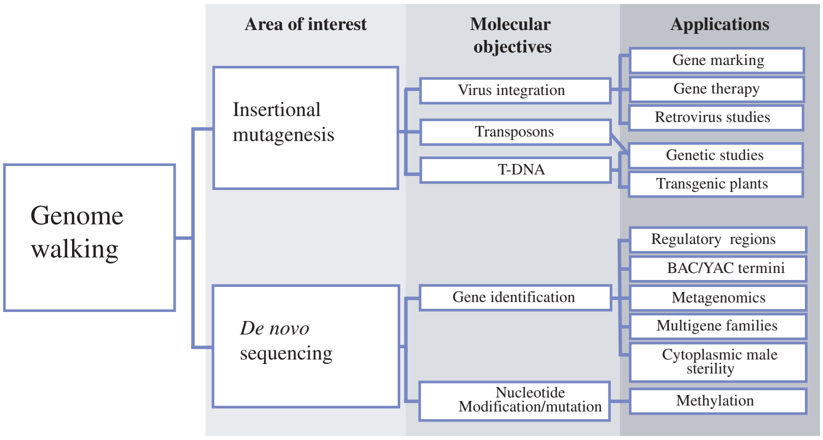

Fig. 2. Classification of GW applications. The scheme summarizes the development of the section GW applications.

Virus integration

A large number of investigations employing GW strategies have been developed to identify and characterize the insertion sites of retroviral cDNAs or retrovirus derived vectors in the human genome, furnishing precious data for understanding the mechanisms of retroviral replication and viral pathogenesis, and also contributing to designing safer retrovirus derived vectors. In this section, the description of the use of GW in the characterization of retroviral vectors in the human genome precedes the discussion about the analysis of the integration of retroviral cDNA.

Gene marking protocols were the first examples in the early 1990s of clinical gene transfer by retroviral vectors [61]. Gene marking studies have benefited considerably from GW techniques. The first analysis by GW of a retroviral-mediated gene marking of bone marrow cells used for autologous transplantion in patients with neuroblastoma was reported by Rill *et al.* [62]. By I-PCR it was possible to detect neuroblasts containing the marker gene in relapsing patients, establishing that malignant cells had been reintroduced in treated patients. The contribution of gene marking to cell and gene therapy has recently been reviewed by Tey and Brenner [63], who also described the advantage in this field of LAM-PCR and FLEA-PCR GW techniques (Table 1) (see the section on Critical evaluations of GW methods).

'Gene marking' efforts preceded the use of retroviral derived vectors in gene therapy as delivery vehicles of therapeutic genes. Unfortunately their use in gene therapy is not without serious drawbacks. Indeed, even if retroviral vectors are depleted of most of their genes in order to prevent dangerous infections, insertional mutagenesis can disrupt important genes, such as those involved in the control of cell growth and division, leading to cancer onset. Additionally, introduced viral promoters and enhancer can activate transcription of proto-oncogenes [64,65]. For these reasons, analysis of integration sites in the host genomes has to be considered as an essential step in gene therapy protocols. Emblematic was the case of patients with a form of severe combined immunodeficiency caused by a defect in a gene of the X chromosome, who first received a retrovirus derived vector engineered with the correct human gene [66]. Unfortunately, after 3 years, two patients developed a leukemia-like condition owing to the insertion of the retroviral vector in proximity to the *LMO-2* (LIM domain only 2) proto-oncogene, as demonstrated by LAM-PCR GW analysis [66].

The first extensive GW studies for the detection of integration sites of retrovirus derived vectors in mammalian chromosomes, which have been highly exploited by researchers interested in gene therapy, were those by Schmidt *et al.* [54,67]. In the first report the authors introduced a modification of the cassette PCR method by Fors *et al.* (Table 1) by using a biotinylated primer for the capture on streptavidin beads of the product of the extension reaction. By this approach they were able to identify the integration sites of retrovirus derived vectors both in transduced HeLa cells and in a murine transplant model. The second GW strategy is the already described LAM PCR (Table 1). By this approach, authors could analyze in two primate models the contribution of marked primitive cells to hematopoiesis.

A significant improvement in the analysis of integration sites of virus derived vectors has been obtained by applying to GW the powerful barcoding linked DNA pyrosequencing technique. This has been clearly illustrated in the report by Wang *et al.* [23] (see above) who were able to analyze more than 160 000 integration sites of gamma-retroviral and lentiviral vectors in mice.

As analysis of the use of retroviral vectors proceeded, a number of studies have also been dedicated to the integration of retroviruses in the human genome, since integration of proviral cDNA into chromosomes can influence subsequent latency or active transcription of viral genes. The first analysis focused on the integration of HIV, or HIV vectors [68], murine leukaemia virus (MLV) and avian sarcoma-leukosis virus (ASLV) [69,70]. All these studies employed the Genome Walker Kit (Clontech) for the identification of insertion sites. Results showed that gene sequences act as preferential integration sites of HIV cDNA, with a weak bias in favor of active genes. Interestingly, no favored integration sites could be detected in the analysis of control naked DNA, suggesting an influence of DNA interacting proteins (either histones or transcription factors) in the integration process. As for the integration of MLV and ASLV, it has been observed that, while MLV strongly favors integration near to transcription start regions with a weak preference for active genes, ASLV shows the weakest bias toward integration in active genes and has no preference for transcription start regions.

Also for the analysis of retrovirus integration Wang *et al.* [24] combined GW with pyrosequencing technology, being able to identify 40 569 unique HIV integration sites. Ontology analysis of HIV hosting genes revealed preferential integration in a group of housekeeping genes involved in metabolism, cell cycle and RNA metabolism. A detailed description of the pyrosequencing GW method applied to the detection of HIV integration sites in the human genome has recently been published [71].

Retroviruses have also been employed for 'gene trapping' approaches (see Transposons below for a more detailed description of gene trapping). Hansen *et al.* [72] developed an automated I-PCR procedure for a large-scale gene trapping analysis of mouse embryonic stem cells, in which all the steps were performed in 384-well plates. In such a way more than 10 000 mutated genes could be identified.

Transposons

Transposon insertional mutagenesis is a basic tool for addressing gene function through analysis of mutant phenotypes and identification of mutated genes in eukaryotic genomes [73–83]. Together with the classical 'transposon tagging' approach, other strategies based on insertional mutagenesis have been devised for the identification of genes that do not produce easily observable phenotypes when knocked-out. In the 'activation tagging' method [84], strong activating sequences, such as the cauliflower mosaic virus (CaMV) 35S enhancer, are inserted into transposon sequences. Tagged genes can consequently be overexpressed resulting in gain-of-function phenotypes. In 'gene trapping' and 'enhancer trapping', transposon constructs containing a reporter gene that can respond to *cis*-acting transcriptional signals at the site of insertion are employed. The host gene is identified by observing the reporter gene expression pattern and sequencing of tagged sites. Different trapping systems have been reviewed by Springer [85].

In such a complex scenario, GW has contributed significantly to the analysis of transposition events, providing valuable data for reverse genetic analysis. The first applications of GW for the identification of transposon integration sites in eukaryotes can be found in the review by Hui *et al.* [2]. In the following years, additional research was performed in which GW was employed not only for the identification of knocked-out genes but also for the characterization of specific transposons. To facilitate the description of the applications of GW to the study of transposable elements, this section has been divided into the subsections Plants, Invertebrates, Vertebrates and Yeast.

Plants

Gene inactivation by transposon insertion has been employed for functional genomics in several plant species. It is worth mentioning that also the T-DNA of the Ti-plasmid of *A. tumefaciens* has been used for this purpose. Although T-DNA mediated insertional mutagenesis is discussed separately in this review, a clear distinction between reports dealing with insertion analysis for transposons or T-DNA is not always possible since in numerous applications T-DNA also served as the launch-pad for transposons engineered in it [86]. Analysis of transposon tagging and characterization of insertion sites for specific transposons in plants are reported in Table 3.

Among the analyses reported in Table 3, of particular interest is the I-PCR based strategy for the identification of mutated genes in the case of high copy number transposable elements, which combines I-PCR with differential screening of amplification products [100]. An alternative strategy for identification of tagged genes in the case of high copy number transposable elements is the transposon display method, based on a cassette GW strategy [101]. In this approach, a *dTph1* specific primer is coupled with a cassette specific primer to amplify all the possible insertion sites. Amplified fragments are then analyzed through a high resolution polyacrylamide gel system.

Large-scale analyses of insertion sites have been performed by high-throughput modification of GW strategies. By means of the thermal asymmetric inter-laced PCR (TAIL-PCR) and suppression PCR GW methods more than 42 000 insertion sites of the *Tos17* retrotransposon in the rice genome were analyzed [107]. An automated TAIL-PCR approach was developed [96] for analysis of the collection of insertional mutants of *Ac/Ds* and *Ac* transposons in two cultivars of rice. The DLA GW strategy associated with pyrosequencing allowed the identification of about 965 000 sequences flanking the highly active maize *mutator* (*Mu*) transposon [60]. These analyses allowed the development of specific insertional mutant databases. In the case of *Arabidopsis thaliana* the ATIDB database (http://atidb.org) was established [110], which also contains information for T-DNA mutants. For rice, the OryGenesDB database (http://orygenesdb.cirad.fr/) was developed [111].

Invertebrates

Among invertebrates, *Drosophila melanogaster* has been the model organism on which the widest range of studies on transposon-based functional genomes has been developed. Insertional mutagenesis analysis can also be found for other insects and other invertebrate metazoa, such as the insects *Tribolium castaneum* and the worm *Caenorhabditis elegans*. An overview of the application of GW analysis for transposon characterization among these organisms, allowing us to appreciate the achievements reached in GW development, is reported in Table 4 and below.

Table 3. Transposon analysis in plant genomes by GW approaches. Transposons are listed in alphabetical order. Names of GW approaches and kits are as in Tables 1 and 2 respectively.

Transposon	Plant	GW approach/kit	Note	References
Ac/Ds	Transgenic tobacco	I-PCR	T-DNA launch-pad	74
	A. thaliana			87
	Tomato		Enhancer trapping[a]	88
	Maize		[a]	89
	Lotus japonicus		T-DNA launch-pad	90
	A. thaliana		Activation tagging	91
		TAIL-PCR	Enhancer trapping	92
			Gene trapping	93–95
	Rice		High-throughput	96
			Gene trapping	97
	Barley		Gene trapping	98
	Transgenic tobacco cells	LM-PCR		99
dTph1	*Petunia hybrida*	I-PCR	See text	100
		Cassette PCR		101
En	*A. thaliana*	I-PCR		102,103
		TAIL-PCR		103
En/Spm-like	*Zingeria biebersteiniana*	DNA Walking SpeedUp kit		104
	Antirrhinum majus	Genome Walker kit		105
Mu	Maize	DLA GW	High-throughput	60
Spm		I-PCR		74
Tos17	Rice		[a]	106
		TAIL-PCR Suppression PCR	High-throughput	107
Ty-1	Apple	Site finding PCR		108
		Genome Walker kit		109

[a] Analysis of transposon distribution.

With regard to *D. melanogaster*, the main tool of transposon insertional mutagenesis studies has been the *P-element*, employed in several strategies of gene tagging, enhancer trapping and gene trapping. I-PCR was first used as an alternative to the so-called 'site-selected mutagenesis' procedure for the selection of transformed lines [112]. By the same approach, large studies have been carried out allowing the analysis of *P-element* integration sites [113]. They showed that the insertion of *P-elements* is not a random process. Most of the insertions occur within a few hundred bases of the transcription start site of a gene. Databases of *Drosophila* insertional mutants are available at the Berkeley Drosophila Genome Project [130] and at the FlyBase Consortium [131]. A specific I-PCR protocol is available at http://www.fruitfly.org/about/methods/inverse.pcr.html.

A wide array of studies using the enhancer trap strategy, known as the GAL4 enhancer trap for the employment of the yeast transcriptional activator GAL4 [132], have also been carried out for *Drosophila*. In this strategy *P-elements* are used as vectors for specific genes engineered downstream of a yeast upstream activation sequence (UAS). A comprehensive overview on the GAL4/UAS enhancer trap system in *Drosophila* can be found in the special issue of *Genesis* [133]. I-PCR and more recently splinkerette PCR [115] have been used as GW procedures for the identification of GAL4 enhancer trap insertions. Vectorette PCR has been proposed as a GW method for investigating trapped genes directly from their mRNAs [134]. For the red flour beetle *T. castaneum* a large collection of *piggyBac* mutants has been produced recently by applying three different GW methods: I-PCR, restriction site PCR and vectorette PCR [121].

The UFW GW method has been used to study the diffusion and evolution of transposable elements in Darwinulidae, a family of non-marine ostracods (Crustacea), allowing two novel families of non long terminal repeat retrotransposons (*Syrinx* and *Daphne*) in *Darwinula stevensoni* to be characterized [124].

For insertional mutagenesis in *C. elegans* a first study adopted the vectorette PCR as the GW strategy to obtain a map of the *Tc1* transposon in the genome of a mutator strain [125]. The same approach has been further developed by coupling electrophoresis analysis of vectorette amplified fragment wild-type and mutant lines (transposon display). Co-segregating bands are excised from gel and further analyzed by nucleotide sequencing [126].

Table 4. Transposon analysis in invertebrate genomes by GW approaches. Organisms are listed according to an ascending taxonomic order, from Insecta (*D. melanogaster*, *Orseolia oryzae*, *T. castaneum*, *Bombyx mori*), back to Crustacea (*D. stevensoni*), Chelicerata (*Metaseiulus occidentalis*) and Pseudocoelomata (*C. elegans*, *Rotifera* sp.). Names of GW approaches are as in Table 1.

Organism	Transposon	GW approach/kit	Note	Reference
D. melanogaster	*P-element*	I-PCR	See text	112,113
		Vectorette PCR	See text	114
		Splinkerette PCR	See text	115
	piggyBac	I-PCR		116,117
	P-element/piggyBac	Splinkerette PCR	[a]	115
Orseolia oryzae	*Mariner*	I-PCR		118
T. castaneum	*Woot*	Restriction site PCR		119
	piggyBac	Restriction site PCR, I-PCR, vectorette-PCR		120,121
Bombyx mori	*piggyBac*	I-PCR	[a]	122
D. stevensoni	*Syrinx* *Daphne*	UFW		124
Metaseiulus occidentalis	*Mariner*	I-PCR		123
C. elegans	*Tc1*	Vectorette PCR		125
	Tc3	Vectorette PCR	See text	126
	Mos1	I-PCR		127,128
Rotifer sp.	*ITm, hAT, piggyBac, helitron, foldback*	UFW		129

[a] Enhancer trapping analysis.

The distributions of different transposon families (*ITm*, *hAT*, *piggyBac*, *helitron*, *foldback*) have been analyzed in several rotifer species by using the UFW method [129]. In these cases telomeric regions were found to be particularly rich in transposable elements, whereas gene-rich regions were transposon-free.

Vertebrates

Transposon-mediated insertional mutagenesis, originally developed for plants and invertebrates, has also been widely applied to vertebrates thanks to the identification of both reconstructed and naturally occurring active vertebrate transposon systems. Several GW methods have been employed for the development and characterization of such systems (Table 5).

Recent review papers on insertional mutagenesis strategies in vertebrates, where the application of GW methods is reported, are those by Izsvák *et al.* [144] and Hackett *et al.* [145] for the analysis of the *Sleeping Beauty* transposon in human cells, Yergeau and Mead [73] on the use of transposable elements in *Xenopus*, Clark *et al.* [146] dealing with the transposons in pig, and Friedel and Soriano [83] for gene trap mutagenesis in mouse. A specific database, containing insertion sequences obtained by I-PCR for the *piggyBac* transposon system in mice cells, was recently established [147].

Yeast

Transposon mutagenesis has also been applied to yeast (*Saccharomyces cerevisiae*). Transposition events of the yeast transposon *Ty1* have been analyzed by I-PCR, showing that insertions of this element occur only rarely (about 3%) in ORF regions [148]. For yeast genome-wide transposon mutagenesis, a more efficient shuttle mutagenesis strategy was developed [149] in which a library of yeast genomic DNA, mutagenized with a bacterial transposon, is first produced in *Escherichia coli*; mutant alleles are subsequently transferred into yeast for functional analysis. Kumar and Snyder [150] reported a detailed protocol for shuttle mutagenesis, in which vectorette PCR was the chosen GW strategy for the identification of insertion sites into the yeast genome. Reports on the use of either vectorette PCR [151] or I-PCR [152] for the analysis of insertional mutants obtained by screening shuttle libraries have been published recently.

T-DNA

The T-DNA of *A. tumefaciens* Ti-vector is a mobile element widely used either for 'plant transformation' with heterologous genes or for insertional mutagenesis analysis of plant genomes. In any case establishing the fate of the engineered T-DNA in the host genome stands as one of the classical GW applications. In the following discussion the two different topics will be treated separately.

T-DNA for gene transfer

Before field trials of genetically modified crops it is of primary importance both to identify T-DNA insertion

Table 5. Transposon analysis in vertebrate genomes by GW approaches. Transposons are listed in alphabetical order. Names of GW methods and kits are as in Tables 1 and 2 respectively.

Transposon	Organism	GW approach/kit	Note	References
Frog Prince	HeLa cells	Splinkerette PCR		139
Minos	HeLa cells	Vectorette PCR		140
NfCR1	Lungfish	GenomeWalker kit		141
piggyBac	Mouse cells	I-PCR	Gene trapping	80
Sleeping Beauty	HeLa cells	Splinkerette PCR		142
	Mouse embryonic stem cells	I-PCR		143
Tol2	Mouse embryonic stem cells	I-PCR		135
	Zebrafish	I-PCR	Gene trapping	136
		TAIL-PCR	Enhancer trapping	137
	X. tropicalis	GenomeWalker kit DNA Walking SpeedUp		138

sites in the host genome and to select transformed plants carrying a single T-DNA copy (necessary for avoiding possible transgene silencing processes activated by multiple T-DNA insertions [153]). Hence, it is not surprising that in a number of cases GW approaches were used to simply determine the number of T-DNA integration sites, without sequencing analysis. Spertini *et al.* [154] analyzed the complexity of the T-DNA integration pattern in transgenic *Arabidopsis* plants by simply analyzing the PCR pattern obtained by applying the suppression PCR GW method. Analogously, a T-DNA fingerprinting method for discrimination between multi-copy transgenic lines and single-copy transformants was developed on the basis of amplified length polymorphism of fragments encompassing the T-DNA/plant genome junctions [155].

Table 6 reports studies in which GW strategies have been employed for the identification of the integration site of the recombinant T-DNA in transgenic plants.

Devic *et al.* [156] published the first paper on the sequence identification of T-DNA insertion sites in *Arabidopsis* transgenic plants by adoption of Siebert's suppression PCR GW method. The same approach was subsequently used by different authors (Table 6). Analysis of integration sites of T-DNA in banana plants was carried out by using a modified version of the original subtractive PCR GW protocol [165]. In the new version, the adaptor cassette is characterized by an unphosphorylated 5′-end of the short strand, in order to favor its release during the first PCR denaturation step, thereby ensuring that only the longer adaptor strand remains ligated and avoiding unspecific amplifications.

The T-linker PCR method was used to compare biolistic and T-DNA transformation procedures in plants (rice and *Arabidopsis*). As expected, in the first case several gene copies were found integrated in the host genome, while in the T-DNA transformed plants GW analysis showed that a limited number of insertions (about 60%) seems guided by vector borders [21].

Table 6. Analysis of T-DNA transgenes in plant genomes by GW approaches. The names of the GW approaches are as in Table 1.

GW approach/kit	Plant	Note	References
Suppression PCR	Arabidopsis		156
	Potato		19,157,158
	Tobacco		19
	Shallot		159
	Maize		160,161
	Barley		162
	Grapefruit		163
	Cotton		164
Subtractive PCR	Banana	See text	165
T-DNA fingerprinting PCR	Soybean		20
	Arabidopsis		166
	Maize		167
	Canola		168
I-PCR	Tomato		169
	Maize		170
T-linker PCR	Rice	See text	21
	Arabidopsis		
TAIL-PCR[a]	Maize		171
APAgene GOLD	Potato	See text	158
DNA Walking SpeedUp	Potato	See text	158
Universal vectorette	Potato	See text	158

[a] Additional examples can be found in the original paper describing the strategy [37].

T-DNA for insertional mutagenesis

Several reports are also available on the use of T-DNA for genome-wide insertional mutagenesis with either gene tagging or gene trapping strategies. For these topics extended GW analyses have been done for the *Arabidopsis* and rice genomes, which are discussed separately below.

Large-scale analysis of T-DNA insertions in the *Arabidopsis* genome has been achieved by different groups who developed high-throughput versions of the original suppression PCR GW method [172–177]. Sequencing results are available at various databases (http://genoplante-info.infobiogen.fr [178], http://www.gabi-kat.de/ [177] and http://signal.salk.edu/cgi-bin/tdnaexpress [176]).

The TAIL-PCR strategy was also used for sequencing T-DNA insertions by various groups [179–182]. In particular, Sessions *et al.* [182] developed a high-throughput TAIL-PCR GW strategy for the analysis of about 100 000 T-DNA transformants of *Arabidopsis* plants. In this case, T-DNA insertion sites showed a higher presence in promoter regions (44%) than in transcribed regions (30%) and intergenic regions (26%). A library of the identified sites was developed and made available for external users (http://www.tmri.org.) [182]. Analysis of T-DNA insertion sites has also been carried out by applying both long inverted PCR and long tail PCR GW methods [183], resulting in fragments longer than 6 kb.

Large-scale T-DNA insertional mutagenesis has been applied to rice. Analysis of thousands of insertion sites has been achieved by both I-PCR [184] and suppression PCR GW strategies [185,186]. More recently a large-scale analysis of T-DNA insertions was performed by TAIL-PCR in about 63 000 transgenic plants. In all cases, inserts were found to be distributed all over the 12 chromosomes. Information on T-DNA transformed rice lines has been collected in the SHIP (Shangai T-DNA Insertion Population) collection and can be found at http://ship.plantsignal.cn [187].

The *Arabidopsis* genome has also been (and still is) a major field of application of several gene trapping T-DNA analyses. The first report about the production and analysis of a collection of *Arabidopsis* enhancer trap lines dates back to 1999 [188]. In this case TAIL-PCR and I-PCR were used to analyze flanking sequences of inflorescence related mutants. Similar protocols are still currently used [189–191]. A detailed protocol for promoter trapping and analysis of T-DNA flanking regions by TAIL-PCR can be found in the method paper by Blanvillain and Gallois [192]. Suppression PCR has been used for screening a gene trap T-DNA/*uidA* collection of about 10 000 transgenic *Arabidopsis* lines developed for the detection of GUS activity during seed germination [193]. Further applications of GW methods to gene trapping in *Arabidopsis* can be found in the review paper by Radhamony *et al.* [194].

In rice the first T-DNA gene trapping studies adopted TAIL-PCR for the analysis of constructs containing *Ac/Ds* transposable element and the *uidA* reporter gene [195,196]. Since then the numerous gene trapping studies in rice have been mostly accompanied by this sequencing procedure. A comprehensive description of gene trapping achievements in rice, including also the use of GW techniques, can be found in the paper by An *et al.* [197].

Analysis of large T-DNA gene trapping collections in rice have been carried out by I-PCR [198], high-throughput adaptation of the suppression PCR GW method [186] or, more recently, TAIL-PCR [199,200]. In these cases also specific databases have been developed: OTL (Oryza Tag Line) database (http://urgi.versailles.inra.fr/OryzaTagLine/) [201], TRIM (Taiwan Rice Insertion Mutants) database (http://trim.sinica.edu.tw) [202] and RMD (Rice Mutant Database) (http://rmd.ncpgr.cn) [203].

TAIL-PCR has also been chosen as the GW method for the analysis of gene trapping in barley [98] and banana genomes [204].

De novo sequencing

A basic application of GW is the identification of nucleotide sequences in the course of the characterization of genes and genomes. This can be aimed either at the characterization of unknown sequences or at the identification of nucleotide modifications and mutations.

Identification of unknown sequences

Browsing the literature in this area, it can be seen that most efforts have been devoted to the identification of 'regulatory regions', while a minor number of reports deal with the use of GW in 'gene identification', 'sequencing of BAC and YAC clones', 'cytoplasmic male sterility' (large modifications occurring in plant nuclear/mitochondrial genomes, which are at the basis of the cytoplasmic male sterility phenotypic trait) and 'multigene families'. Table 7 summarizes studies conducted by GW for the analysis of regulatory regions, while the other applications have been reported in Table 8. Additionally, it is worth mentioning the application of GW to metagenomics analysis, even if this topic is outside the scope of this review. Related information is available in the review paper by Singh *et al.* [205].

Siebert *et al.* [15] developed the well-known suppression PCR GW method to walk upstream of the 5′-end coding regions of the human TPA (tissue-type plasminogen activator) and transferrin genes for a valuable distance (4.5 and 6 kb, respectively). This technique is one of the most used GW methods, finding application in several different cases.

Padegimas and Reichert [16] succeeded in isolating promoter regions from three different maize peroxidase genes, improving the splinkerette strategy with the introduction of 3′-blocked adaptors and removal of unligated genomic DNA by *Exo*III digestion.

The identification of regulatory regions of multigene families can be pursued by different approaches. Leoni *et al.* [56] adopted an ET-GW strategy (Table 1) in which highly conserved regions of a multigene family were chosen to design common primers to be used as gene specific primers. In this way it has been possible to simultaneously identify regulatory elements of the spinach multigene family coding for isoforms of the light harvesting protein Lhcb1. Additionally, novel gene members of the same family could be detected by this approach. In a second case, the TVL-PCR GW method, applied to identify the regulatory regions of strawberry SUPERMAN-like genes [29], has been used for the identification of multiple members of a gene family using degenerate primers based on conserved sequences as priming sites. It must be noted also that the Universal GW kit (Clontech) has been employed for the analysis of a sugarcane BAC clone containing multiple copies of the sugarcane *DIRIGENT* gene [206].

Analysis of nucleotide modifications and mutations

The application of GW to the analysis of nucleotide modifications and mutations is essentially based on the LM-PCR strategy due to Mueller and Wold [53]. Indeed, although originally presented as a footprinting technique, LM-PCR has also been successfully regarded as a GW technique. Pfeifer *et al.* [241] illustrated its application for both genome sequencing and methylation analysis of the human X-linked phosphoglycerate kinase (*PGK-1*) gene. An automated version of the LM-PCR usable for GW analysis of DNA methylation, DNA damage and protein-DNA footprints has also been developed [242]. More recently the method was further improved (see also [67]) and widely used for mapping DNA damage in carcinogenesis etiology [243,244]. LM-PCR applications for the analysis of mitochondrial DNA damage due to chemicals or ageing have also been reported [245,246].

Critical evaluations of GW methods

The issue about which GW method better fits the specific experimental conditions is not easy to deal with exhaustively because of the numerous methods available

Table 7. Regulatory regions identified in eukaryotes by GW approaches. Since more than 200 papers can be found in the literature dealing with this topic, mostly reporting the use of commercially available kits, here only papers which describe the development of a specific GW method are reported. Analyses are reported in chronological order. Names of GW methods are as in Table 1.

Gene	Organism	GW approach/kit	Note	References
Po	Shark	Cassette PCR		10
TPA	Man	Suppression PCR		15
Transferrin				
At23	Arabidopsis	RAGE-GW		32
PR-10	Parsley			
Peroxidase	Maize	Modified splinkerette	See text	16
Sucrose phosphate synthase	Banana	Single primer amplification		58
Actin	Sugarcane			
S15 ribosomal protein	*Dunaliella tertiolecta*			
Several cDNAs	*Brassica juncea*	Mishra *et al.*		40
	Pennisetum glaucum			
Gibberellin 20-oxidase	Rice	T-linker PCR		21
Squalene synthase	*Ganoderma lucidum*	Self-formed adaptor PCR		48
Ascorbate peroxidaseHsp70 Hsp10	*P. glaucum*	High-throughput genome walking		50
Gst	*Salicornia brachiata*			
Lhcb1 family	Spinach	Leoni *et al.*		56,57
LRDEF	Lily	Straight walk		26
OgGSTZ2	Rice			
SuRB	Tobacco			
PGK1	*Pichia ciferrii*	Template blocking PCR		28
SUPERMAN-like	Strawberry	TVL-PCR		29

Table 8. Genes identified in eukaryotes by GW approaches. Most of the data were obtained by using the Universal GW kit (Clontech) or other common GW approaches described in the text. Voices are listed in alphabetical order for main taxonomic groups. Applications are as in Fig. 2.

Species	Gene	Application	References
Animals			
Drosophila melanogaster	*Sup 4*	BAC⁄YAC termini	207
Homo sapiens	*FLEB14-14*	BAC⁄YAC termini	208
	Scyb11	BAC⁄YAC termini	209
	HPRT	Gene sequencing	210
	LRP1B	Gene sequencing	211
	ELF3	Gene sequencing	212
	Dystrophin	Gene sequencing	213
Macropus eugenii	*LTB*, *TNF* and *LTA*	BAC⁄YAC termini	214
Schistosoma mansoni	*SmHox1 SmHox4* and *SmHox4*	BAC⁄YAC termini	215
Salmo salar	*Hox* genes	BAC⁄YAC termini	216
Pogona vitticeps	Z and W chromosome fragment	Gene sequencing	217
Fungi			
Latimeria menadoensis	*Hox*	Gene sequencing	218
Penicillium pinophilum	Endo-β-1.4-glucanase gene 5	Gene sequencing	219
Phoma betae	Aphidicolan-16β-ol synthase	Gene sequencing	220
Phomopsis amygdali	*PbGGs, ACS, PbP450-1, PbP450-2, PbTP, PbTF*	Gene sequencing	220
	PaDC1 and *PaDC2*	Gene sequencing	221
	Diterpene hydrocarbon phomopsene	Gene sequencing	222
Pucciniomycotina	*RHA1*, *RHA2* and *RHA3*	Gene sequencing	223
Plants			
Allium cepa	*Orf725*	CMS	224
Capsicum annuum	*Rf* flanking region	CMS	225
Cicer arientinum	*Pi-ta-2* and *xa5*	Gene sequencing	226
Coffea arabica; C. canephora	*ManS1* and *GMGT1*	Gene sequencing	227
Malus domestica	*Mal D3* genes	Gene sequencing	228
	MdAGP1, *MdAGP2* and *MdAGP3* genes	Gene sequencing	229
Oryza sativa	Slender glume	BAC⁄YAC termini	230
	OsPE	Gene sequencing	231
Spinacia oleracea	*Lhcb1*	Multigene family	56
Sugar beet	Restorer-of-fertility	BAC⁄YAC termini⁄CMS	232
Sugarcane	*DIRIGENT*	Multigene family	206
Taxus media	Geranyl geranyl diphosphate	Gene sequencing	233
Triticum aestivum L	*LMW-GS* genes	BAC⁄YAC termini	234
		BAC⁄YAC termini	235
Protozoa			
Cryptobia salmositica	Adenosylmethionone synthetase	Gene sequencing	236
	Cathepsin L-like cysteine proteinase	Gene sequencing	237
	MSP-1	Gene sequencing	238
Neospora caninum	Nc*SAG4*	Gene sequencing	239
	Nc*BSR4*	Gene sequencing	240

and the multiplicity of variables in the different assays. Nevertheless some general comments can be made.

The first issue to consider is whether a single sequencing (as in the case of the study of a single gene) or multiple sequencing data (as in the case of large insertional mutagenesis analysis) are necessary. In the first case it can be assumed that most of the methods give satisfactory results. This is clearly shown in the case of the identification of gene regulatory regions (Table 7), where at least 12 different methods have been employed. In contrast, in the identification of multiple sequences only a limited number of methods have been successfully used (I-PCR, vectorette, splinkerette, suppression, TAIL-PCR), which can therefore be considered as first choice in planning GW experiments.

In any case, some differences clearly exist among the GW strategies, and at least three parameters can be considered for their critical evaluation: specificity, sensitivity and efficiency. As for specificity, it can generally be assumed that it mostly relies on the specificity of the gene specific primer used in the GW approach. I-PCR which adopts two specific primers should therefore be considered as the most specific method. Nevertheless, all the other methods that use a gene specific primer coupled with an adaptor/tail specific walking primer can be regarded as highly specific as well. Methods adopting random/degenerate primers, conversely, may show lower specificity. A precautionary note must be added for cassette PCR methods that do not take countermeasures to prevent the synthesis of non-specific PCR products deriving from the walking primer. The blocked DLA GW method properly addresses this point, showing that blocking the adaptor extension in the first cycle of the final PCR amplification can increase specificity of PCR products from 44% to 100% [27].

Blocked DLA was also compared with splinkerette PCR for sensitivity. The higher sensitivity of the first method is clearly demonstrated by the relative intensity of the electrophoretic bands of amplification products. In the course of a screening experiment for the identification of *P-elements* in *Drosophila*, Eggert *et al.* [114] combined in several ratios flies carrying or not a defined transposon. They showed that vectorette PCR can be more sensitive than I-PCR in the identification of transgenic *P-elements*, allowing detection of a specific insertion in a ratio of 1 : 6000–10 000.

It must noted, however, that some technical improvements have undoubtedly improved the general sensitivity of GW methods. This is the case of introducing biotinylation of adaptors and primers. Nielsen *et al.* [247] showed the possibility that, when adopting solid-phase purification of biotinylated fragments, GW can reach a very high sensitivity, able to detect about two copies of a target sequence in a DNA background of 25 ng.

Recently three commercial kits [APAgene GOLD Genome Walking Kit (BIO S&T), DNA Walking SpeedUp Kit II (Seegene) and Universal Vectorette System (Sigma)] and the suppression PCR GW method (as modified by Spertini *et al.* [154]) were compared for the identification of T-DNA flanking regions in transgenic potato. In this analysis, the two methods based on the extension of gene specific primers and PCR amplification with degenerated primers (APAgene™ and DNA Walking SpeedUp™ II) showed higher success rates than the two cassette PCR methods, which identified a lower number of flanking regions [158].

As for efficiency, some strategies have to be mentioned for the reported capacity to read more than 3 kb for single walk, as in the case of LD-GW PCR [55], suppression PCR [15], long I-PCR and TAIL-PCR [183].

A last issue to consider for the choice of a GW method is the possibility of its scale-up for high-throughput analysis, if needed. This has been demonstrated to be possible for suppression PCR [107,172–177], TAIL-PCR [96,107,182,199], I-PCR [72,88], high-throughput GW [248], LM-PCR [242], straight walk high-throughput [26] and restriction site extension PCR [33].

Conclusions and perspectives

GW encompasses an array of easy-to-use strategies for the identification of genome nucleotide sequences, useful for both insertional mutagenesis analysis and *de novo* sequencing. In the first case it has largely contributed to advances in reverse genetic analysis, and to the development of databases of mutants of many eukaryotic genomes. In the second case, GW is particularly advantageous for the identification of specific sequences in cases where whole genome sequencing projects have not been undertaken. It is noteworthy to observe that most of the different GW strategies or improvements have been developed in the course of *de novo* sequencing approaches (see Identification of unknown sequences, for example).

The extreme flexibility of GW strategies makes its application possible in every standardly equipped research laboratory. In addition, the possibility of merging GW strategies to next generation sequencing approaches will undoubtedly extend the future application of this by now basic technique of molecular biology.

Acknowledgement

This work has been partly supported by a grant from the Italian Ministry of Instruction University and Research: PRIN 20087ATS57 ‘Food allergens’.

References

1 Hengen PN (1995) Vectorette, splinkerette and boomerang DNA amplification. *Trends Biochem Sci* **20**, 372–373.

2 Hui EK, Wang PC & Lo SJ (1998) Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. *Cell Mol Life Sci* **54**, 1403–1411.

3 Tonooka Y & Fujishima M (2009) Comparison and critical evaluation of PCR-mediated methods to walk along the sequence of genomic DNA. *Appl Microbiol Biotechnol* **85**, 37–43.

4 Kotik M (2009) Novel genes retrieved from environmental DNA by polymerase chain reaction: current

genome-walking techniques for future metagenome applications. *J Biotechnol* **144**, 75–82.

5 Triglia T, Peterson MG & Kemp DJ (1988) A procedure for *in vitro* amplification of DNA segments that lie outside the boundaries of known sequences. *Nucleic Acids Res* **16**, 8186.

6 Benkel BF & Fong Y (1996) Long range-inverse PCR (LR-IPCR): extending the useful range of inverse PCR. *Genet Anal Biomolec Eng* **13**, 123–127.

7 Kohda T & Taira K (2000) A simple and efficient method to determine the terminal sequences of restriction fragments containing known sequences. *DNA Res* **7**, 151–155.

8 Tsaftaris A, Pasentzis K & Argiriou A (2010) Rolling circle amplification of genomic templates for inverse PCR (RCA-GIP): a method for 5′- and 3′-genome walking without anchoring. *Biotechnol Lett* **32**, 157–161.

9 Shyamala V & Ames GF (1989) Genome walking by single-specific-primer polymerase chain reaction: SSP-PCR. *Gene* **84**, 1–8.

10 Fors L, Saavedra RA & Hood L (1990) Cloning of the shark Po promoter using a genomic walking technique based on the polymerase chain reaction. *Nucleic Acids Res* **18**, 2793–2799.

11 Riley J, Butler R, Ogilvie D, Finniear R, Jenner D, Powell S, Anand R, Smith JC & Markham AF (1990) A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res* **18**, 2887–2890.

12 Rosenthal A & Jones DS (1990) Genomic walking and sequencing by oligo-cassette mediated polymerase chain reaction. *Nucleic Acids Res* **18**, 3095–3096.

13 Lagerstrom M, Parik J, Malmgren H, Stewart J, Pettersson U & Landegren U (1991) Capture PCR: -efficient amplification of DNA fragments adjacent to a known sequence in human and YAC DNA. *PCR Methods Appl* **1**, 111–119.

14 Devon RS, Porteous DJ & Brookes AJ (1995) Splinkerettes – improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res* **23**, 1644–1645.

15 Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA & Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* **23**, 1087–1088.

16 Padegimas LS & Reichert NA (1998) Adaptor ligation-based polymerase chain reaction-mediated walking. *Anal Biochem* **260**, 149–153.

17 Zhang Z & Gurr SJ (2000) Walking into the unknown: a ‘step down’ PCR-based technique leading to the direct sequence analysis of flanking genomic DNA. *Gene* **253**, 145–150.

18 Kilstrup M & Kristiansen KN (2000) Rapid genome walking: a simplified oligo-cassette mediated polymerase chain reaction using a single genome-specific primer. *Nucleic Acids Res* **28**, E55.

19 Cottage A, Yang A, Maunders H, de Lacy RC & Ramsay NA (2001) Identification of DNA sequences flanking T-DNA insertions by PCR-walking. *Plant Mol Biol Rep* **19**, 321–327.

20 Windels P, Taverniers I, Depicker A, Van Bockstaele E & De Loose M (2001) Characterisation of the Roundup Ready soybean insert. *Eur Food Res Technol* **213**, 107–112.

21 Yuanxin Y, Chengcai A, Li L, Jiayu G, Guihong T & Zhangliang C (2003) T-linker-specific ligation PCR (T-linker PCR): an advanced PCR technique for chromosome walking or for isolation of tagged DNA ends. *Nucleic Acids Res* **31**, e68.

22 Nthangeni MB, Ramagoma F, Tlou MG & Litthauer D (2005) Development of a versatile cassette for directional genome walking using cassette ligation-mediated PCR and its application in the cloning of complete lipolytic genes from Bacillus species. *J Microbiol Meth* **61**, 225–234.

23 Wang GP *et al.* (2008) DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res* **36**, e49.

24 Wang GP, Ciuffi A, Leipzig J, Berry CC & Bushman FD (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* **17**, 1186–1194.

25 Tonooka Y, Mizukami Y & Fujishima M (2008) One-base excess adaptor ligation method for walking uncloned genomic DNA. *Appl Microbiol Biotechnol* **78**, 173–180.

26 Tsuchiya T, Kameya N & Nakamura I (2009) Straight walk: a modified method of ligation-mediated genome walking for plant species with large genomes. *Anal Biochem* **388**, 158–160.

27 Liu S, Dietrich CR & Schnable PS (2009) DLA-based strategies for cloning insertion mutants: cloning the gl4 locus of maize using Mu transposon tagged alleles. *Genetics* **183**, 1215–1225.

28 Bae JH & Sohn JH (2010) Template-blocking PCR: an advanced PCR technique for genome walking. *Anal Biochem* **398**, 112–116.

29 Orcheski BB & Davis TM (2010) An enhanced method for sequence walking and paralog mining: TOPO(R) vector-ligation PCR. *BMC Res Notes* **3**, 61.

30 Jones DH & Winistorfer SC (1992) Sequence specific generation of a DNA panhandle permits PCR amplification of unknown flanking DNA. *Nucleic Acids Res* **20**, 595–600.

31 Rudenko GN, Rommens CM, Nijkamp HJ & Hille J (1993) Supported PCR: an efficient procedure to amplify sequences flanking a known DNA segment. *Plant Mol Biol* **21**, 723–728.

32 Cormack RS & Somssich IE (1997) Rapid amplification of genomic ends (RAGE) as a simple method to clone flanking genomic DNA. *Gene* **194**, 273–276.

33 Ji J & Braam J (2010) Restriction site extension PCR: a novel method for high-throughput characterization of tagged DNA fragments and genome walking. *PLoS ONE* **5**, e10577.

34 Parker JD, Rabinovitch PS & Burmer GC (1991) Targeted gene walking polymerase chain reaction. *Nucleic Acids Res* **19**, 3055–3060.

35 Sarkar G, Turner RT & Bolander ME (1993) Restriction-site PCR: a direct method of unknown sequence retrieval adjacent to a known locus by using universal primers. *PCR Methods Appl* **2**, 318–322.

36 Puskas LG, Fartmann B & Bottka S (1994) Restricted PCR: amplification of an individual sequence flanked by a highly repetitive element from total human DNA. *Nucleic Acids Res* **22**, 3251–3252.

37 Liu YG & Whittier RF (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics* **25**, 674–681.

38 Chen X & Wu R (1997) Direct amplification of unknown genes and fragments by uneven polymerase chain reaction. *Gene* **185**, 195–199.

39 Ge Y & Charon NW (1997) Identification of a large motility operon in *Borrelia burgdorferi* by semi-random PCR chromosome walking. *Gene* **189**, 195–201.

40 Mishra RN, Singla-Pareek SL, Nair S, Sopory SK & Reddy MK (2002) Directional genome walking using PCR. *BioTechniques* **33**, 830–834.

41 Myrick KV & Gelbart WM (2007) A modified universal fast walking method for single-tube transposon mapping. *Nat Protoc* **2**, 1556–1563.

42 Myrick KV & Gelbart WM (2002) Universal fast walking for direct and versatile determination of flanking sequence. *Gene* **284**, 125–131.

43 Park DJ (2005) LaNe RAGE: a new tool for genomic DNA flanking sequence determination. *Electron J Biotechnol* **8**, 218–225.

44 Tan G, Gao Y, Shi M, Zhang X, He S, Chen Z & An C (2005) SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. *Nucleic Acids Res* **33**, e122.

45 Levano-Garcia J, Verjovski-Almeida S & da Silva AC (2005) Mapping transposon insertion sites by touchdown PCR and hybrid degenerate primers. *BioTechniques* **38**, 225–229.

46 Walser JC, Evgen'ev MB & Feder ME (2006) A genomic walking method for screening sequence length polymorphism. *Mol Ecol Notes* **6**, 563–567.

47 Guo H & Xiong J (2006) A specific and versatile genome walking technique. *Gene* **381**, 18–23.

48 Wang S, He J, Cui Z & Li S (2007) Self-formed adaptor PCR: a simple and efficient method for chromosome walking. *Appl Environ Microbiol* **73**, 5048–5051.

49 Pilhofer M, Bauer AP, Schrallhammer M, Richter L, Ludwig W, Schleifer KH & Petroni G (2007) Characterization of bacterial operons consisting of two tubulins and a kinesin-like gene by the novel two-step gene walking method. *Nucleic Acids Res* **35**, e135.

50 Reddy PS, Mahanty S, Kaul T, Nair S, Sopory SK & Reddy MK (2008) A high-throughput genome-walking method and its use for cloning unknown flanking sequences. *Anal Biochem* **381**, 248–253.

51 Ping L, Vogel H & Boland W (2008) Cloning of prokaryotic genes by a universal degenerate primer PCR. *FEMS Microbiol Lett* **287**, 192–198.

52 Martin-Harris MH, Bartley PA & Morley AA (2010) Gene walking using sequential hybrid primer polymerase chain reaction. *Anal Biochem* **399**, 308–310.

53 Mueller PR & Wold B (1989) *In vivo* footprinting of a muscle specific enhancer by ligation mediated PCR. *Science* **246**, 780–786.

54 Schmidt M *et al*. (2002) Polyclonal long-term repopulating stem cell clones in a primate model. *Blood* **100**, 2737–2743.

55 Min GS & Powell JR (1998) Long-distance genome walking using the long and accurate polymerase chain reaction. *BioTechniques* **24**, 398–400.

56 Leoni C, Volpicella M, Placido A, Gallerani R & Ceci LR (2010) Application of a genome walking method for the study of the spinach Lhcb1 multigene family. *Journal of Plant Physiology* **167**, 138–143.

57 Leoni C, Gallerani R & Ceci LR (2008) A genome walking strategy for the identification of eukaryotic nucleotide sequences adjacent to known regions. *BioTechniques* **44**, 229, 232–235.

58 Hermann SR, Miller JA, O'Neill S, Tsao TT, Harding RM & Dale JL (2000) Single-primer amplification of flanking sequences. *BioTechniques* **29**, 1176–1178, 1180.

59 Pule MA, Rousseau A, Vera J, Heslop HE, Brenner MK & Vanin EF (2008) Flanking-sequence exponential anchored-polymerase chain reaction amplification: a sensitive and highly specific method for detecting retroviral integrant-host-junction sequences. *Cytotherapy* **10**, 526–539.

60 Liu S *et al*. (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet* **5**, e1000733.

61 Brenner M (1996) Gene marking. *Hum Gene Ther* **7**, 1927–1936.

62 Rill DR *et al*. (1994) Direct demonstration that autologous bone marrow transplantation for solid tumors can return a multiplicity of tumorigenic cells. *Blood* **84**, 380–383.

63 Tey SK & Brenner MK (2007) The continuing contribution of gene marking to cell and gene therapy. *Mol Ther* **15**, 666–676.

64 Bushman FD (2007) Retroviral integration and human gene therapy. *J Clin Invest* **117**, 2083–2086.

65 Modlich U & Baum C (2009) Preventing and exploiting the oncogenic potential of integrating gene vectors. *J Clin Invest* **119**, 755–758.

66 Hacein-Bey-Abina S *et al.* (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415–419.

67 Schmidt M *et al.* (2001) Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. *Hum Gene Ther* **12**, 743–749.

68 Schroder AR, Shinn P, Chen H, Berry C, Ecker JR & Bushman F (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529.

69 Wu X, Li Y, Crise B & Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**, 1749–1751.

70 Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, Ecker JR & Bushman FD (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**, E234.

71 Ciuffi A & Barr SD (2010) Identification of HIV integration sites in infected host genomic DNA. *Methods* **53**, 39–46.

72 Hansen GM *et al.* (2008) Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. *Genome Res* **18**, 1670–1679.

73 Yergeau DA & Mead PE (2007) Manipulating the Xenopus genome with transposable elements. *Genome Biol* **8**(Suppl 1), S11.

74 Earp DJ, Lowe B & Baker B (1990) Amplification of genomic sequences flanking transposable elements in host and heterologous plants: a tool for transposon tagging and genome characterization. *Nucleic Acids Res* **18**, 3271–3279.

75 Walbot V (1992) Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. *Plant Physiol* **43**, 49–82.

76 Izawa T *et al.* (1997) Transposon tagging in rice. *Plant Mol Biol* **35**, 219–229.

77 Feschotte C, Jiang N & Wessler SR (2002) Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* **3**, 329–341.

78 Candela H & Hake S (2008) The art and design of genetic screens: maize. *Nat Rev Genet* **9**, 192–203.

79 Ryder E & Russell S (2003) Transposable elements as tools for genomics and genetics in Drosophila. *Brief Funct Genomic Proteomic* **2**, 57–71.

80 Wu S, Ying G, Wu Q & Capecchi MR (2007) Toward simpler and faster genome-wide mutagenesis in mice. *Nat Genet* **39**, 922–930.

81 Ivics Z, Li MA, Mates L, Boeke JD, Nagy A, Bradley A & Izsvák Z (2009) Transposon-mediated genome manipulation in vertebrates. *Nat Methods* **6**, 415–422.

82 Kumar A (2008) Multipurpose transposon insertion libraries for large-scale analysis of gene function in yeast. *Methods Mol Biol* **416**, 117–129.

83 Friedel RH & Soriano P (2010) Gene trap mutagenesis in the mouse. *Methods Enzymol* **477**, 243–269.

84 Tani H *et al.* (2004) Activation tagging in plants: a tool for gene discovery. *Funct Integr Genomics* **4**, 258–266.

85 Springer PS (2000) Gene traps: tools for plant development and genomics. *Plant Cell* **12**, 1007–1020.

86 Muskett PR, Clissold L, Marocco A, Springer PS, Martienssen R & Dean C (2003) A resource of mapped dissociation launch pads for targeted insertional mutagenesis in the Arabidopsis genome. *Plant Physiol* **132**, 506–516.

87 Long D, Martin M, Sundberg E, Swinburne J, Puangsomlee P & Coupland G (1993) The maize transposable element system Ac/Ds as a mutagen in Arabidopsis: identification of an albino mutation induced by Ds insertion. *Proc Natl Acad Sci USA* **90**, 10370–10374.

88 Meissner R, Chague V, Zhu Q, Emmanuel E, Elkind Y & Levy AA (2000) Technical advance: a high throughput system for transposon tagging and promoter trapping in tomato. *Plant J* **22**, 265–274.

89 Kolkman JM *et al.* (2005) Distribution of Activator (Ac) throughout the maize genome for use in regional mutagenesis. *Genetics* **169**, 981–995.

90 Tirichine L, Herrera-Cervera JA & Stougaard J (2005) *Ds* Gene-Tagging. In *Lotus japonicus Handbook* (Màrquez AJ ed.), pp 211–215. Springer, Amsterdam.

91 Wilson K, Long D, Swinburne J & Coupland G (1996) A dissociation insertion causes a semidominant mutation that increases expression of TINY, an Arabidopsis gene related to APETALA2. *Plant Cell* **8**, 659–671.

92 Klimmek F, Sjodin A, Noutsos C, Leister D & Jansson S (2006) Abundantly and rarely expressed Lhc protein genes exhibit distinct regulation patterns in plants. *Plant Physiol* **140**, 793–804.

93 Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H & Martienssen R (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev* **9**, 1797–1810.

94 Gu Q, Ferrandiz C, Yanofsky MF & Martienssen R (1998) The Fruitfull MADS-box gene mediates cell differentiation during Arabidopsis fruit development. *Development* **125**, 1509–1517.

95 Klimyuk VI, Nussaume L, Harrison K & Jones JD (1995) Novel GUS expression patterns following transposition of an enhancer trap Ds element in Arabidopsis. *Mol Gen Genet* **249**, 357–365.

96 van Enckevort LJ *et al.* (2005) EU-OSTID: a collection of transposon insertional mutants for functional genomics in rice. *Plant Mol Biol* **59**, 99–110.

97 Upadhyaya NM, Zhu QH & Bhat RS (2011) Transposon insertional mutagenesis in rice. *Methods Mol Biol* **678**, 147–177.

98 Lazarow K & Lutticke S (2009) An Ac/Ds-mediated gene trap system for functional genomics in barley. *BMC Genomics* **10**, 55.

99 Ott T, Nelsen-Salz B & Doring HP (1992) PCR-aided genomic sequencing of 5′ subterminal sequences of the maize transposable element Activator (Ac) in transgenic tobacco plants. *Plant J* **2**, 705–711.

100 Souer E, Quattrocchio F, de Vetten N, Mol J & Koes R (1995) A general method to isolate genes tagged by a high copy number transposable element. *Plant J* **7**, 677–685.

101 Van den Broeck D, Maes T, Sauer M, Zethof J, De Keukeleire P, D'Hauw M, Van Montagu M & Gerats T (1998) Transposon Display identifies individual transposable elements in high copy number lines. *Plant J* **13**, 121–129.

102 Aarts MG, Dirkse WG, Stiekema WJ & Pereira A (1993) Transposon tagging of a male sterility gene in Arabidopsis. *Nature* **363**, 715–717.

103 Marsch-Martinez N, Greco R, Van Arkel G, Herrera-Estrella L & Pereira A (2002) Activation tagging using the En-I maize transposon system in Arabidopsis. *Plant Physiol* **129**, 1544–1556.

104 Altinkut A, Raskina O, Nevo E & Belyayev A (2006) En/Spm-like transposons in Poaceae species: transposase sequence variability and chromosomal distribution. *Cell Mol Biol Lett* **11**, 214–230.

105 Roccaro M, Li Y, Sommer H & Saedler H (2007) ROSINA (RSI) is part of a CACTA transposable element, TamRSI, and links flower development to transposon activity. *Mol Genet Genomics* **278**, 243–254.

106 Yamazaki M *et al.* (2001) The rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. *Mol Genet Genomics* **265**, 336–344.

107 Miyao A *et al.* (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell* **15**, 1771–1780.

108 Zhao G, Zhang Z, Sun H, Li H & Dai H (2007) Isolation of Ty1-copia-like retrotransposon sequences from the apple genome by chromosome walking based on modified SiteFinding-polymerase chain reaction. *Acta Biochim Biophys Sin (Shanghai)* **39**, 675–683.

109 Zhao G, Dai H, Chang L, Ma Y, Sun H, He P & Zhang Z (2009) Isolation of two novel complete Ty1-copia retrotransposons from apple and demonstration of use of derived S-SAP markers for distinguishing bud sports of *Malus domestica* cv. Fuji. *Tree Genetics Genomes* **6**, 149–159.

110 Pan X, Liu H, Clarke J, Jones J, Bevan M & Stein L (2003) ATIDB: *Arabidopsis thaliana* insertion database. *Nucleic Acids Res* **31**, 1245–1251.

111 Droc G *et al.* (2006) OryGenesDB: a database for rice reverse genetics. *Nucleic Acids Res* **34**, D736–D740.

112 Sentry JW & Kaiser K (1994) Application of inverse PCR to site-selected mutagenesis of *Drosophila*. *Nucleic Acids Res* **22**, 3429–3430.

113 Liao GC, Rehm EJ & Rubin GM (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **97**, 3347–3351.

114 Eggert H, Bergemann K & Saumweber H (1998) Molecular screening for P-element insertions in a large genomic region of *Drosophila melanogaster* using polymerase chain reaction mediated by the vectorette. *Genetics* **149**, 1427–1434.

115 Potter CJ & Luo L (2010) Splinkerette PCR for mapping transposable elements in *Drosophila*. *PLoS ONE* **5**, e10168.

116 Thibault ST *et al.* (2004) A complementary transposon tool kit for *Drosophila melanogaster* using P and piggyBac. *Nat Genet* **36**, 283–287.

117 Bonin CP & Mann RS (2004) A piggyBac transposon gene trap for the analysis of gene expression and function in *Drosophila*. *Genetics* **167**, 1801–1811.

118 Behura SK, Nair S & Mohan M (2001) Polymorphisms flanking the mariner integration sites in the rice gall midge (*Orseolia oryzae* Wood–Mason) genome are biotype-specific. *Genome* **44**, 947–954.

119 Beeman RW & Stauth DM (1997) Rapid cloning of insect transposon insertion junctions using 'universal' PCR. *Insect Mol Biol* **6**, 83–88.

120 Lorenzen MD, Berghammer AJ, Brown SJ, Denell RE, Klingler M & Beeman RW (2003) piggyBac-mediated germline transformation in the beetle *Tribolium castaneum*. *Insect Mol Biol* **12**, 433–440.

121 Trauner J *et al.* (2009) Large-scale insertional mutagenesis of a coleopteran stored grain pest, the red flour beetle *Tribolium castaneum*, identifies embryonic lethal mutations and enhancer traps. *BMC Biol* **7**, 73.

122 Uchino K *et al.* (2008) Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori*. *Insect Biochem Mol Biol* **38**, 1165–1173.

123 Jeyaprakash A & Hoy MA (1995) Complete sequence of a mariner transposable element from the predatory mite *Metaseiulus occidentalis* isolated by an inverse PCR approach. *Insect Mol Biol* **4**, 31–39.

124 Schon I & Arkhipova IR (2006) Two families of non-LTR retrotransposons, Syrinx and Daphne, from the Darwinulid ostracod, *Darwinula stevensoni*. *Gene* **371**, 296–307.

125 Korswagen HC, Durbin RM, Smits MT & Plasterk RH (1996) Transposon Tc1-derived sequence-tagged sites in *Caenorhabditis elegans* as markers for gene mapping. *Proc Natl Acad Sci USA* **93**, 14680–14685.

126 Wicks SR, de Vries CJ, van Luenen HG & Plasterk RH (2000) CHE-3, a cytosolic dynein heavy chain, is required for sensory cilia structure and function in *Caenorhabditis elegans*. *Dev Biol* **221**, 295–307.

127 Bessereau JL, Wright A, Williams DC, Schuske K, Davis MW & Jorgensen EM (2001) Mobilization of a *Drosophila* transposon in the *Caenorhabditis elegans* germ line. *Nature* **413**, 70–74.

128 Granger L, Martin E & Segalat L (2004) Mos as a tool for genome-wide insertional mutagenesis in *Caenorhabditis elegans*: results of a pilot study. *Nucleic Acids Res* **32**, e117.

129 Arkhipova IR & Meselson M (2005) Diverse DNA transposons in rotifers of the class Bdelloidea. *Proc Natl Acad Sci USA* **102**, 11781–11786.

130 Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, Misra S & Rubin GM (1999) The Berkeley Drosophila Genome Project gene disruption project: single P-element insertions mutating 25% of vital *Drosophila* genes. *Genetics* **153**, 135–177.

131 Tweedie S *et al.* (2009) FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res* **37**, D555–D559.

132 Brand AH & Perrimon N (1993) Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415.

133 Various Authors (2002) Special Issue: GAL4/UAS in *Drosophila*. *Genesis* **34**, 1–173.

134 Lukacsovich T & Yamamoto D (2001) Trap a gene and find out its function: toward functional genomics in *Drosophila*. *J Neurogenet* **15**, 147–168.

135 Kawakami K & Noda T (2004) Transposition of the Tol2 element, an Ac-like element from the Japanese medaka fish *Oryzias latipes*, in mouse embryonic stem cells. *Genetics* **166**, 895–899.

136 Kawakami K, Takeda H, Kawakami N, Kobayashi M, Matsuda N & Mishina M (2004) A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell* **7**, 133–144.

137 Parinov S, Kondrichin I, Korzh V & Emelyanov A (2004) Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes *in vivo*. *Dev Dyn* **231**, 449–459.

138 Hamlet MR, Yergeau DA, Kuliyev E, Takeda M, Taira M, Kawakami K & Mead PE (2006) Tol2 transposon-mediated transgenesis in *Xenopus tropicalis*. *Genesis* **44**, 438–445.

139 Miskey C, Izsvák Z, Plasterk RH & Ivics Z (2003) The Frog Prince: a reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells. *Nucleic Acids Res* **31**, 6873–6881.

140 Klinakis AG, Zagoraiou L, Vassilatis DK & Savakis C (2000) Genome-wide insertional mutagenesis in human cells by the *Drosophila* mobile element Minos. *EMBO Rep* **1**, 416–421.

141 Sirijovski N, Woolnough C, Rock J & Joss JMP (2005) nfCR1, the first non-LTR retrotransposon characterized in the Australian lungfish genome, *Neoceratodus forsteri*, shows similarities to CR1-like elements. *J Exp Zool Pt B: Molec Develop Evol* **304**, 40–49.

142 Ivics Z, Hackett PB, Plasterk RH & Izsvák Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501–510.

143 Luo G, Ivics Z, Izsvák Z & Bradley A (1998) Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc Natl Acad Sci USA* **95**, 10769–10773.

144 Izsvák Z, Chuah MK, Vandendriessche T & Ivics Z (2009) Efficient stable gene transfer into human cells by the Sleeping Beauty transposon vectors. *Methods* **49**, 287–297.

145 Hackett PB, Largaespada DA & Cooper LJ (2010) A transposon and transposase system for human application. *Mol Ther* **18**, 674–683.

146 Clark KJ, Carlson DF & Fahrenkrug SC (2007) Pigs taking wing with transposons and recombinases. *Genome Biol* **8**(Suppl. 1), S13.

147 Sun LV *et al.* (2008) PBmice: an integrated database system of piggyBac (PB) insertional mutations and their characterizations in mice. *Nucleic Acids Res* **36**, D729–D734.

148 Ji H, Moore DP, Blomberg MA, Braiterman LT, Voytas DF, Natsoulis G & Boeke JD (1993) Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell* **73**, 1007–1018.

149 Burns N, Grimwade B, Ross-Macdonald PB, Choi EY, Finberg K, Roeder GS & Snyder M (1994) Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev* **8**, 1087–1105.

150 Kumar A & Snyder M (2001) Genome-wide transposon mutagenesis in yeast. In *Curr Protoc Mol Biol*, Genome-Wide Transposon Mutagenesis in Yeast, chapter 13, pp.13.3.1–13.3.15, John Wiley & Sons, New York.

151 Estruch F, Peiro-Chova L, Gomez-Navarro N, Durban J, Hodge C, Del Olmo M & Cole CN (2009) A genetic screen in *Saccharomyces cerevisiae* identifies new genes that interact with mex67-5, a temperature-sensitive allele of the gene encoding the mRNA export receptor. *Mol Genet Genomics* **281**, 125–134.

152 Hontz RD, Niederer RO, Johnson JM & Smith JS (2009) Genetic identification of factors that modulate ribosomal DNA transcription in *Saccharomyces cerevisiae*. *Genetics* **182**, 105–119.

153 Flavell RB (1994) Inactivation of gene expression in plants as a consequence of specific sequence duplication. *Proc Natl Acad Sci USA* **91**, 3490–3496.

154 Spertini D, Beliveau C & Bellemare G (1999) Screening of transgenic plants by amplification of unknown genomic DNA flanking T-DNA. *BioTechniques* **27**, 308–314.

155 Theuns I, Windels P, De Buck S, Depicker A, Van Bockstaele E & De Loose M (2002) Identification and characterization of T-DNA inserts by T-DNA fingerprinting. *Euphytica* **123**, 75–84.

156 Devic M, Albert S, Delseny M & Roscoe TJ (1997) Efficient PCR walking on plant genomic DNA. *Plant Physiol Biochem* **35**, 331–339.

157 Cote MJ, Meldrum AJ, Raymond P & Dollard C (2005) Identification of genetically modified potato (*Solanum tuberosum*) cultivars using event specific polymerase chain reaction. *J Agric Food Chem* **53**, 6691–6696.

158 Cullen D, Harwood W, Smedley M, Davies H & Taylor M (2011) Comparison of DNA walking methods for isolation of transgene-flanking regions in GM potato. *Mol Biotechnol* **49**, 19–31.

159 Zheng SJ, Henken B, Sofiari E, Jacobsen E, Krens FA & Kik C (2001) Molecular characterization of transgenic shallots (*Allium cepa* L.) by adaptor ligation PCR (AL-PCR) and sequencing of genomic DNA flanking T-DNA borders. *Transgenic Res* **10**, 237–245.

160 Holck A, Va M, Didierjean L & Rudi K (2002) 5′-Nuclease PCR for quantitative event-specific detection of the genetically modified Mon810 MaisGard maize. *Eur Food Res Technol* **214**, 449–453.

161 Collonnier C *et al.* (2005) Characterization and event specific detection by quantitative real-time PCR of T25 maize insert. *J AOAC Int* **88**, 536–546.

162 Salvo-Garrido H, Travella S, Bilham LJ, Harwood WA & Snape JW (2004) The distribution of transgene insertion sites in barley determined by physical and genetic mapping. *Genetics* **167**, 1371–1379.

163 Rai M (2006) Refinement of the Citrus tristeza virus resistance gene (Ctv) positional map in *Poncirus trifoliata* and generation of transgenic grapefruit (*Citrus paradisi*) plant lines with candidate resistance genes in this region. *Plant Mol Biol* **61**, 399–414.

164 Akritidis P, Pasentsis K, Tsaftaris AS, Mylona PV & Polidoros AN (2008) Identification of unknown genetically modified material admixed in conventional cotton seed and development of an event-specific detection method. *Electron J Biotechnol* **11**, 76–83.

165 Perez-Hernandez JB, Swennen R & Sagi L (2006) Number and accuracy of T-DNA insertions in transgenic banana (*Musa* spp.) plants characterized by an improved anchored PCR technique. *Transgenic Res* **15**, 139–150.

166 Windels P, De Buck S, Van Bockstaele E, De Loose M & Depicker A (2003) T-DNA integration in *Arabidopsis* chromosomes. Presence and origin of filler DNA sequences. *Plant Physiol* **133**, 2061–2068.

167 Windels P, Bertrand S, Depicker A, Moens W, Bockstaele E & Loose M (2003) Qualitative and event-specific PCR real-time detection methods for StarLink maize. *Eur Food Res Technol* **216**, 259–263.

168 Taverniers I, Windels P, Vaitilingom M, Milcamps A, Van Bockstaele E, Van den Eede G & De Loose M (2005) Event-specific plasmid standards and real-time PCR methods for transgenic Bt11, Bt176, and GA21 maize and transgenic GT73 canola. *J Agric Food Chem* **53**, 3041–3052.

169 Knapp S, Larondelle Y, Rossberg M, Furtek D & Theres K (1994) Transgenic tomato lines containing Ds elements at defined genomic positions as tools for targeted transposon tagging. *Mol Gen Genet* **243**, 666–673.

170 Ronning SB, Vaitilingom M, Berdal KG & Holst-Jensen A (2003) Event specific real-time quantitative PCR for genetically modified Bt11 maize (*Zea mays*). *Eur Food Res Technol* **216**, 347–354.

171 Yang L, Xu S, Pan A, Yin C, Zhang K, Wang Z, Zhou Z & Zhang D (2005) Event specific qualitative and quantitative polymerase chain reaction detection of genetically modified MON863 maize based on the 5′-transgene integration sequence. *J Agric Food Chem* **53**, 9312–9318.

172 Balzergue S *et al.* (2001) Improved PCR-walking for large-scale isolation of plant T-DNA borders. *Biotechniques*, **30**, 496–498, 502, 504.

173 Brunaud V *et al.* (2002) T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. *EMBO Rep* **3**, 1152–1157.

174 Alonso JM & Ecker JR (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nat Rev Genet* **7**, 524–536.

175 O'Malley RC, Alonso JM, Kim CJ, Leisse TJ & Ecker JR (2007) An adapter ligation-mediated PCR method for high-throughput mapping of T-DNA inserts in the *Arabidopsis* genome. *Nat Protoc* **2**, 2910–2917.

176 O'Malley RC & Ecker JR (2010) Linking genotype to phenotype using the *Arabidopsis* unimutant collection. *Plant J* **61**, 928–940.

177 Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K & Weisshaar B (2003) An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol Biol* **53**, 247–259.

178 Samson F, Brunaud V, Balzergue S, Dubreucq B, Lepiniec L, Pelletier G, Caboche M & Lecharny A (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. *Nucleic Acids Res* **30**, 94–97.

179 McElver J *et al.* (2001) Insertional mutagenesis of genes required for seed development in *Arabidopsis thaliana*. *Genetics* **159**, 1751–1763.

180 Budziszewski GJ *et al.* (2001) *Arabidopsis* genes essential for seedling viability: isolation of insertional mutants and molecular cloning. *Genetics* **159**, 1765–1778.

181 Krysan PJ, Young JC, Jester PJ, Monson S, Copenhaver G, Preuss D & Sussman MR (2002) Characterization of T-DNA insertion sites in *Arabidopsis thaliana* and the implications for saturation mutagenesis. *Omics* **6**, 163–174.

182 Sessions A *et al.* (2002) A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell* **14**, 2985–2994.

183 Szabados L *et al.* (2002) Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. *Plant J* **32**, 233–242.

184 An S *et al.* (2003) Generation and analysis of end sequence database for T-DNA tagging lines in rice. *Plant Physiol* **133**, 2040–2047.

185 Sallaud C *et al.* (2003) Highly efficient production and characterization of T-DNA plants for rice (*Oryza sativa* L.) functional genomics. *Theor Appl Genet* **106**, 1396–1408.

186 Sallaud C *et al.* (2004) High throughput T-DNA insertion mutagenesis in rice: a first step towards *in silico* reverse genetics. *Plant J* **39**, 450–464.

187 Fu FF, Ye R, Xu SP & Xue HW (2009) Studies on rice seed quality through analysis of a large-scale T-DNA insertion population. *Cell Res* **19**, 380–391.

188 Campisi L *et al.* (1999) Generation of enhancer trap lines in *Arabidopsis* and characterization of expression patterns in the inflorescence. *Plant J* **17**, 699–707.

189 Gardner MJ, Baker AJ, Assie JM, Poethig RS, Haseloff JP & Webb AA (2009) GAL4 GFP enhancer trap lines for analysis of stomatal guard cell development and gene expression. *J Exp Bot* **60**, 213–226.

190 Kim SG, Lee S, Kim YS, Yun DJ, Woo JC & Park CM (2010) Activation tagging of an *Arabidopsis* SHI-related sequence gene produces abnormal anther dehiscence and floral development. *Plant Mol Biol* **74**, 337–351.

191 Kuroha T, Okuda A, Arai M, Komatsu Y, Sato S, Kato T, Tabata S & Satoh S (2009) Identification of *Arabidopsis* subtilisin-like serine protease specifically expressed in root stele by gene trapping. *Physiol Plant* **137**, 281–288.

192 Blanvillain R & Gallois P (2008) Promoter trapping system to study embryogenesis. *Methods Mol Biol* **427**, 121–135.

193 Dubreucq B, Berger N, Vincent E, Boisson M, Pelletier G, Caboche M & Lepiniec L (2000) The *Arabidopsis* AtEPR1 extensin-like gene is specifically expressed in endosperm during seed germination. *Plant J* **23**, 643–652.

194 Radhamony RN, Prasad AM & Srinivasan R (2005) T-DNA insertional mutagenesis in *Arabidopsis*: a tool for functional genomics. *Electron J Biotechnol* **8**, 82–106.

195 Chin HG *et al.* (1999) Molecular analysis of rice plants harboring an Ac/Ds transposable element-mediated gene trapping system. *Plant J* **19**, 615–623.

196 Jeon JS *et al.* (2000) T-DNA insertional mutagenesis for functional genomics in rice. *Plant J* **22**, 561–570.

197 An G, Jeong DH, Jung KH & Lee S (2005) Reverse genetic approaches for functional genomics of rice. *Plant Mol Biol* **59**, 111–123.

198 Ryu CH *et al.* (2004) Generation of T-DNA tagging lines with a bidirectional gene trap vector and the establishment of an insertion-site database. *Plant Mol Biol* **54**, 489–502.

199 Hsing YI *et al.* (2007) A rice gene activation/knockout mutant resource for high throughput functional genomics. *Plant Mol Biol* **63**, 351–364.

200 Zhang J *et al.* (2007) Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13 804 T-DNA flanking sequences from an enhancer-trap mutant library. *Plant J* **49**, 947–959.

201 Larmande P *et al.* (2008) Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library. *Nucleic Acids Res* **36**, D1022–D1027.

202 Chern CG *et al.* (2007) A rice phenomics study – phenotype scoring and seed propagation of a T-DNA insertion-induced rice mutant population. *Plant Mol Biol* **65**, 427–438.

203 Zhang J, Li C, Wu C, Xiong L, Chen G, Zhang Q & Wang S (2006) RMD: a rice mutant database for functional analysis of the rice genome. *Nucleic Acids Res* **34**, D745–D748.

204 Santos E, Remy S, Thiry E, Windelinckx S, Swennen R & Sagi L (2009) Characterization and isolation of a T-DNA tagged banana promoter active during *in vitro* culture and low temperature stress. *BMC Plant Biol* **9**, 77.

205 Singh J, Behal A, Singla N, Joshi A, Birbian N, Singh S, Bali V & Batra N (2009) Metagenomics: concept, methodology, ecological inference and recent advances. *Biotechnol J* **4**, 480–494.

206 Damaj MB *et al.* (2010) Isolating promoters of multigene family members from the polyploid sugarcane genome by PCR-based walking in BAC DNA. *Genome* **53**, 840–847.

207 Wesley CS, Myers MP & Young MW (1994) Rapid sequential walking from termini of cosmid, P1 and YAC inserts. *Nucleic Acids Res* **22**, 538–539.

208 Asakawa S *et al.* (1997) Human BAC library: construction and rapid screening. *Gene* **191**, 69–79.

209 Meyer M, Erdel M, Duba HC, Werner ER & Werner-Felmayer G (2000) Cloning, genomic sequence, and chromosome mapping of Scyb11, the murine homologue of SCYB11 (alias betaR1/H174/SCYB9B/I-TAC/IP-9/CXCL11). *Cytogenet Cell Genet* **88**, 278–282.

210 Nelson DL, Ledbetter SA, Corbo L, Victoria MF, Ramirez-Solis R, Webster TD, Ledbetter DH & Caskey CT (1989) Alu polymerase chain reaction: a method for rapid isolation of human-specific sequences

from complex DNA sources. *Proc Natl Acad Sci USA* **86**, 6686–6690.

211 Liu CX, Musco S, Lisitsina NM, Yaklichkin SY & Lisitsyn NA (2000) Genomic organization of a new candidate tumor suppressor gene, LRP1B. *Genomics* **69**, 271–274.

212 Kaplan MH, Wang XP, Xu HP & Dosik MH (2004) Partially unspliced and fully spliced ELF3 mRNA, including a new Alu element in human breast cancer. *Breast Cancer Res Treat* **83**, 171–187.

213 Zhong M, Pan SY, Cai LD, Lu BX & Zhang GZ (2006) Strategy of localizing the deletion mutation point in large introns of Dystrophin gene. *Chinese Journal of Clinical Rehabilitation* **10**, 120–121.

214 Cross JGR, Harrison GA, Coggill P, Sims S, Beck S, Deakin JE & Marshall Graves JA (2005) Analysis of the genomic region containing the tammar wallaby (*Macropus eugenii*) orthologues of MHC class III genes. *Cytogenet Genome Res* **111**, 110–117.

215 Pierce RJ *et al.* (2005) Evidence for a dispersed Hox gene cluster in the platyhelminth parasite *Schistosoma mansoni*. *Mol Biol Evol* **22**, 2491–2503.

216 Mungpakdee S, Seo HC, Angotzi AR, Dong X, Akalin A & Chourrout D (2008) Differential evolution of the 13 Atlantic salmon Hox clusters. *Mol Biol Evol* **25**, 1333–1343.

217 Quinn AE, Ezaz T, Sarre SD, Graves JM & Georges A (2010) Extension, single-locus conversion and physical mapping of sex chromosome sequences identify the Z microchromosome and pseudo-autosomal region in a dragon lizard, *Pogona vitticeps*. *Heredity* **104**, 410–417.

218 Koh EGL, Lam K, Christoffels A, Erdmann MV, Brenner S & Venkatesh B (2003) Hox gene clusters in the Indonesian coelacanth, *Latimeria menadoensis*. *Proc Natl Acad Sci USA* **100**, 1084–1088.

219 Jeya M, Joo AR, Lee KM, Sim WI, Oh DK, Kim YS, Kim IW & Lee JK (2010) Characterization of endo-beta-1,4-glucanase from a novel strain of *Penicillium pinophilum* KMJ601. *Appl Microbiol Biotechnol* **85**, 1005–1014.

220 Toyomasu T *et al.* (2004) Cloning of a gene cluster responsible for the biosynthesis of diterpene aphidicolin, a specific inhibitor of DNA polymerase alpha. *Biosci Biotechnol Biochem* **68**, 146–152.

221 Toyomasu T *et al.* (2008) Identification of diterpene biosynthetic gene clusters and functional analysis of labdane-related diterpene cyclases in *Phomopsis amygdali*. *Biosci Biotechnol Biochem* **72**, 1038–1047.

222 Toyomasu T *et al.* (2009) Biosynthetic gene-based secondary metabolite screening: a new diterpene, methyl phomopsenonate, from the fungus *Phomopsis amygdali*. *J Org Chem* **74**, 1541–1548.

223 Coelho MA, Rosa A, Rodrigues N, Fonseca A & Goncalves P (2008) Identification of mating type genes in the bipolar basidiomycetous yeast *Rhodosporidium toruloides*: first insight into the MAT locus structure of the Sporidiobolales. *Eukaryot Cell* **7**, 1053–1061.

224 Kim S, Lee ET, Cho DY, Han T, Bang H, Patil BS, Ahn YK & Yoon MK (2009) Identification of a novel chimeric gene, orf725, and its use in development of a molecular marker for distinguishing among three cytoplasm types in onion (*Allium cepa* L.). *Theor Appl Genet* **118**, 433–441.

225 Min WK, Lim H, Lee YP, Sung SK, Kim BD & Kim S (2008) Identification of a third haplotype of the sequence linked to the Restorer-of-fertility (Rf) gene and its implications for male-sterility phenotypes in peppers (*Capsicum annuum* L.). *Molec Cells* **25**, 20–29.

226 Iruela M, Piston F, Cubero JI, Millan T, Barro F & Gil J (2009) The marker SCK13(603) associated with resistance to ascochyta blight in chickpea is located in a region of a putative retrotransposon. *Plant Cell Rep* **28**, 53–60.

227 Pre M, Caillet V, Sobilo J & McCarthy J (2008) Characterization and expression analysis of genes directing galactomannan synthesis in coffee. *Ann Bot* **102**, 207–220.

228 Gao ZS *et al.* (2005) Linkage map positions and allelic diversity of two Mal d 3 (non-specific lipid transfer protein) genes in the cultivated apple (*Malus domestica*). *Theor Appl Genet* **110**, 479–491.

229 Choi YO *et al.* (2010) Isolation and promoter analysis of anther-specific genes encoding putative arabinogalactan proteins in Malus x domestica. *Plant Cell Rep* **29**, 15–24.

230 Teraishi M, Hirochika H, Okumoto Y, Horibata A, Yamagata H & Tanisaka T (2001) Identification of YAC clones containing the mutable slender glume locus slg in rice (*Oryza sativa* L.). *Genome* **44**, 1–6.

231 Puri A, Basha PO, Kumar M, Rajpurohit D, Randhawa GS, Kianian SF, Rishi A & Dhaliwal HS (2010) The polyembryo gene (OsPE) in rice. *Funct Integrative Genom* **10**, 359–366.

232 Hagihara E, Matsuhira H, Ueda M, Mikami T & Kubo T (2005) Sugar beet BAC library construction and assembly of a contig spanning Rf1, a restorer-of-fertility gene for Owen cytoplasmic male sterility. *Mol Genet Genomics* **274**, 316–323.

233 Liao Z *et al.* (2005) An intron-free methyl jasmonate inducible geranylgeranyl diphosphate synthase gene from *Taxus media* and its functional identification in yeast. *Molec Biol* **39**, 11–17.

234 Huang XQ & Cloutier S (2007) Hemi-nested touchdown PCR combined with primer-template mismatch PCR for rapid isolation and sequencing of low molecular weight glutenin subunit gene family from a hexaploid wheat BAC library. *BMC Genet* **8**, 18.

235 Huang XQ & Cloutier S (2008) Molecular characterization and genomic organization of low molecular weight glutenin subunit genes at the Glu-3 loci in

hexaploid wheat (*Triticum aestivum* L.). *Theor Appl Genet* **116**, 953–966.

236 Jesudhasan PR & Woo PTK (2007) An S-adenosylmethionine synthetase gene from the pathogenic piscine hemoflagellate, *Cryptobia salmositica*. *Parasitol Res* **100**, 1401–1406.

237 Jesudhasan PR, Tan CW, Hontzeas N & Woo PT (2007) A cathepsin L-like cysteine proteinase gene from the protozoan parasite, *Cryptobia salmositica*. *Parasitol Res* **100**, 881–886.

238 Jesudhasan PR, Tan CW & Woo PTK (2007) A metalloproteinase gene from the pathogenic piscine hemoflagellate, *Cryptobia salmositica*. *Parasitol Res* **100**, 899–904.

239 Fernandez-Garcia A, Risco-Castillo V, Zaballos A, Alvarez-Garcia G & Ortega-Mora LM (2006) Identification and molecular cloning of the *Neospora caninum* SAG4 gene specifically expressed at bradyzoite stage. *Mol Biochem Parasitol* **146**, 89–97.

240 Risco-Castillo V, Fernandez-Garcia A, Zaballos A, Aguado-Martinez A, Hemphill A, Rodriguez-Bertos A, Alvarez-Garcia G & Ortega-Mora LM (2007) Molecular characterisation of BSR4, a novel bradyzoite-specific gene from *Neospora caninum*. *Intl J Parasitol* **37**, 887–896.

241 Pfeifer GP, Steigerwald SD, Mueller PR, Wold B & Riggs AD (1989) Genomic sequencing and methylation analysis by ligation mediated PCR. *Science* **246**, 810–813.

242 Dai SM, Chen HH, Chang C, Riggs AD & Flanagan SD (2000) Ligation-mediated PCR for quantitative *in vivo* footprinting. *Nat Biotechnol* **18**, 1108–1111.

243 Besaratinia A & Pfeifer GP (2009) DNA-lesion mapping in mammalian cells. *Methods* **48**, 35–39.

244 Besaratinia A & Pfeifer GP (2006) Investigating human cancer etiology by DNA lesion footprinting and mutagenicity analysis. *Carcinogenesis* **27**, 1526–1537.

245 LeDoux SP, Druzhyna NM, Hollensworth SB, Harrison JF & Wilson GL (2007) Mitochondrial DNA repair: a critical player in the response of cells of the CNS to genotoxic insults. *Neuroscience* **145**, 1249–1259.

246 Lezza AM, Fallacara FP, Pesce V, Leeuwenburgh C, Cantatore P & Gadaleta MN (2008) Localization of abasic sites and single-strand breaks in mitochondrial DNA from brain of aged rat, treated or not with caloric restriction diet. *Neurochem Res* **33**, 2609–2614.

247 Nielsen CR, Berdal KG & Holst-Jensen A (2008) Anchored PCR for possible detection and characterisation of foreign integrated DNA at near single molecule level. *Eur Food Res Technol* **226**, 949–956.

248 Reddy MK, Nair S & Sopory SK (2002) A new approach for efficient directional genome walking using polymerase chain reaction. *Anal Biochem* **306**, 154–158.

Supporting information

The following supplementary material is available:
Table S1. List of patents related to GW methods.

This supplementary material can be found in the online version of this article.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.