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.
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.

<table>
<thead>
<tr>
<th>Group</th>
<th>Sub-group</th>
<th>Name</th>
<th>References</th>
<th>P</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R-GW</td>
<td>Inverted PCR</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Long range inverted PCR</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Bridged inverted PCR</td>
<td>7</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Rolling circle inverted PCR</td>
<td>8</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>Cassette PCR</td>
<td>Single-specific-primer PCR</td>
<td>9</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Fors et al.</td>
<td>10</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Vectorette PCR</td>
<td>11</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Cassette ligation</td>
<td>12</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Capture PCR</td>
<td>13</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Splinkerette PCR</td>
<td>14</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Boomerang DNA amplification</td>
<td>1</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Suppression PCR GW</td>
<td>15</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Padegimas and Reichert</td>
<td>16</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>Step-down PCR</td>
<td>17</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Simplified oligo-cassette</td>
<td>18</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>Cottage et al.</td>
<td>19</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>F-DNA fingerprinting PCR</td>
<td>20</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>T-linker PCR</td>
<td>21</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>Versatile cassette</td>
<td>22</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>Barcoding pyrosequencing</td>
<td>23,24</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>One-base excess adaptor ligation</td>
<td>25</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>Straight walk</td>
<td>26</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>Blocked DLA</td>
<td>27</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>Template-blocking PCR</td>
<td>28</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>TVL-PCR</td>
<td>29</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>26</td>
<td>Others</td>
<td>Panhandle PCR</td>
<td>30</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>Supported PCR</td>
<td>31</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>RAGE</td>
<td>32</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>Restriction site extension PCR</td>
<td>33</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>30</td>
<td>P-GW</td>
<td>Targeted gene-walking PCR</td>
<td>34</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td>Restriction-site PCR</td>
<td>35</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>Restricted PCR (novel Alu PCR)</td>
<td>36</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>33</td>
<td></td>
<td>TAIL-PCR</td>
<td>37</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td>Uneven PCR</td>
<td>38</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>Semi-random PCR</td>
<td>39</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>Mishra et al.</td>
<td>40</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td>UFW PCR</td>
<td>41,42</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td>Lariat-dependent nested PCR</td>
<td>43</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>39</td>
<td></td>
<td>Site finding PCR</td>
<td>44</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>Touchdown PCR-based</td>
<td>45</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td>Walser et al.</td>
<td>46</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>42</td>
<td></td>
<td>Nested PCR-based</td>
<td>47</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td>Self-formed adaptor PCR</td>
<td>48</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td>Two-step gene walking PCR</td>
<td>49</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>High-genome walking</td>
<td>50</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td>SD-PCR</td>
<td>51</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td>SHP-PCR</td>
<td>52</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>E-GW</td>
<td>EC-PCR</td>
<td>53</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>49</td>
<td></td>
<td>LM-PCR</td>
<td>54</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>LAM-PCR</td>
<td>55</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>51</td>
<td></td>
<td>ET-PCR</td>
<td>56,57</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>52</td>
<td>Others</td>
<td>Single-primer amplification</td>
<td>58</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>53</td>
<td></td>
<td>FLEA-PCR</td>
<td>59</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
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 \( \mathcal{f} \) and \( \mathcal{i} \) 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 MuClone 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 site, 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 MuClone 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′-NH₂ blocked double-stranded linker, to avoid the fall-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 S17 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
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 sub-areas 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.

<table>
<thead>
<tr>
<th>Kit name</th>
<th>Customer service</th>
<th>GW technique</th>
<th>Company (website)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Walker™ Kit</td>
<td></td>
<td>Suppression PCR GW</td>
<td>Clontech (<a href="http://www.clontech.com">http://www.clontech.com</a>)</td>
</tr>
<tr>
<td>Vectorette™ Genomic Systems</td>
<td></td>
<td>Vectorette</td>
<td>Sigma Aldrich (<a href="http://www.sigmaaldrich.com">http://www.sigmaaldrich.com</a>)</td>
</tr>
<tr>
<td>TOPO™ Walker Kit</td>
<td></td>
<td>TVL-PCR</td>
<td>Invitrogen (<a href="http://www.invitrogen.com">http://www.invitrogen.com</a>)</td>
</tr>
<tr>
<td>DNA Walking SeedUp Premix™ Kit</td>
<td>Genome Walking Service</td>
<td>Suppression PCR GW</td>
<td>Seegene (<a href="http://www.seegene.co.kr">http://www.seegene.co.kr</a>)</td>
</tr>
<tr>
<td>Genome Walking Kit</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 transplantation 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.b.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.
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 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.

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

*Analysis of transposon distribution.*
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 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.

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

*Enhancer trapping analysis.
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].

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...
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.infoobiogen.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 (Shanghai 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 ExoIII 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-I) 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.

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

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene Description</th>
<th>Application</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macropus eugenii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schistosoma mansoni</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salmo salar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pogona viticeps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fungi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latimeria menadoensis</td>
<td>Hox</td>
<td>Gene sequencing</td>
<td>218</td>
</tr>
<tr>
<td>Penicillium pinophilum</td>
<td>Endo-β-1,4-glucanase gene 5</td>
<td>Gene sequencing</td>
<td>219</td>
</tr>
<tr>
<td>Phoma betae</td>
<td>Aphidicolan-16β-ol synthase</td>
<td>Gene sequencing</td>
<td>220</td>
</tr>
<tr>
<td>Phomopsis amygdali</td>
<td>PbGGS, ACS, PbP450-1, PbP450-2, PbTP, PbTF</td>
<td>Gene sequencing</td>
<td>220</td>
</tr>
<tr>
<td>Pucciniomycotina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allium cepa</td>
<td>Orf725</td>
<td>CMS</td>
<td>224</td>
</tr>
<tr>
<td>Capsicum annuum</td>
<td>Rf flanking region</td>
<td>CMS</td>
<td>225</td>
</tr>
<tr>
<td>Cicer arietinum</td>
<td>Pi-ta-2 and xa5</td>
<td>Gene sequencing</td>
<td>226</td>
</tr>
<tr>
<td>Coffea arabica; C. canephora</td>
<td>ManS1 and GMGT1</td>
<td>Gene sequencing</td>
<td>227</td>
</tr>
<tr>
<td>Malus domestica</td>
<td>Mal D3 genes</td>
<td>Gene sequencing</td>
<td>228</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>MdAGP1, MdAGP2 and MdAGP3 genes</td>
<td>Gene sequencing</td>
<td>229</td>
</tr>
<tr>
<td>Spinacia oleracea</td>
<td>Lhcb1</td>
<td>Gene sequencing</td>
<td>231</td>
</tr>
<tr>
<td>Sugar beet</td>
<td>Restorer-of-fertility</td>
<td>BAC/YAC termini/CMS</td>
<td>232</td>
</tr>
<tr>
<td>Sugarcane</td>
<td>DIRIGENT</td>
<td>Multigene family</td>
<td>206</td>
</tr>
<tr>
<td>Taxus media</td>
<td>Geranyl geranyl diphosphate</td>
<td>Gene sequencing</td>
<td>233</td>
</tr>
<tr>
<td>Triticum aestivum L</td>
<td>LMW-GS genes</td>
<td>BAC/YAC termini</td>
<td>234</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>MdAGP1, MdAGP2 and MdAGP3 genes</td>
<td>Gene sequencing</td>
<td>229</td>
</tr>
<tr>
<td>Prototina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptobia salmositica</td>
<td>Adenosylmethionone synthetase</td>
<td>Gene sequencing</td>
<td>236</td>
</tr>
<tr>
<td>Neospora caninum</td>
<td>NcSA64</td>
<td>Gene sequencing</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>NcBSR4</td>
<td>Gene sequencing</td>
<td>240</td>
</tr>
</tbody>
</table>

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.
ally 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
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.


128 Granger L, Martin E & Segalat L (2004) Mos as a tool for genome-wide insertional mutagenesis in 
Natl Acad Sci USA 102, 11781–11786.
Berkeley Drosophila Genome Project gene disruption project: single P-element insertions mutating 25% of 
D555–D559.
133 Various Authors (2002) Special Issue: GAL4/UAS in 
134 Lukacsovich T & Yamamoto D (2001) Trap a gene and find out its function: toward functional genomics in 
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.
cells by the Drosophila mobile element Minos. EMBO Rep 1, 416–421.
141 Sirijovski N, Woonough C, Rock J & Joss JMP (2005) nCR1, the first non-LTR retrotransposon 
characterized in the Australian lungfish genome, Neoceratodus forsteri, shows similarities to CR1-like 
142 Ivics Z, Hackett PB, Plasterk RH & Ivics Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-
mouse embryonic stem cells. Proc Natl Acad Sci USA 95, 10769–10773.
Mol Ther 18, 674–683.
Genome Biol 8(Suppl. 1), S13.
147 Sun LV et al. (2008) PBmice: an integrated database system of piggyBac (PB) insertional mutations and 
Cell 73, 1007–1018.
Genome-Wide Transposon Mutagenesis in Yeast, chapter 13, pp.13.3.1–13.3.15, John Wiley & Sons, New 
York.
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 
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.


**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.