Published October 6, 2016

ORIGINAL RESEARCH

Hierarchical Map of Orthologous Genomic Regions Reconstructed from Two Closely Related Genomes: Cucumber Case Study

Huifen Cao, Erli Pang, and Kui Lin*

Abstract

Accurate identification of orthologous genomic regions (OGRs) between two closely related genomes is crucial for the reliable detection of genomic changes, which range from small-scale changes (e.g., single nucleotide or small nucleotides) to large-scale structural changes. Although diverse OGRs inferred at different levels have been successfully applied to address various biological questions, a limited number of studies have simultaneously integrated OGRs from different levels. Here, we report on a new approach to construct a hierarchical map of OGRs. Using different types of genomic markers, this approach was applied to two very closely related cucumber genomes [*Cucumis sativus* L. var. *sativus* and *C. sativus* L. var. *hardwickii* (Royle) Alef.]. We identified two different levels of OGRs using Mugsy (denoted as dnaOGRs) and i-ADHoRe (denoted as proOGRs). Using information regarding the anchored chromosomes of the two genomes, a third level of OGRs (denoted chrOGRs) could be built at the chromosomal level. Together, these OGRs could be organized into a hierarchical map that represented the parent–child relationships (chrOGR:proOGRs:dnaOGRs) between the two genomes. For this case study, the map consisted of seven chrOGRs, 540 proOGRs, and 22,321 dnaOGRs. Based on this map, we designed different methods to detect both small-scale and large-scale genomic changes. Surprisingly, many genomic changes were detected at each OGR level despite the very short divergence time between the two subspecies. Together, our results show that a hierarchical map of OGRs and their related genomic changes are useful resources for elucidating the diversity and evolution of cucumber genomes and phenotypes.

Core Ideas

- A strategy for reconstructing a hierarchical map of orthologous genomic regions from two genomes.
- Both large- and small-scale genomic changes could be more accurately detected from this map.
- Many large-scale genomic changes are inferred between two cucumber subspecies diverged shortly.

A LARGE CHALLENGE in genome biology is to correlate potential genomic changes with observed differences in heritable phenotypes (Edwards and Batley, 2004). During the course of evolution, genomes accumulate various changes, which represent important sources of genetic diversity (Altshuler et al., 2012; Huang et al., 2012; Qi et al., 2013). Genomic changes range from small-scale changes, such as single-nucleotide variants (SNVs) and small insertion–deletion (InDel) markers, to large-scale structural changes, such as inversions, transpositions, and gene and segmental gains or losses, which are difficult to detect based on short reads. The identification of high-quality OGRs between two closely

H. Cao, E. Pang, and K. Lin, MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal Univ., Beijing 100875, China. Received 2 Mar. 2016. Accepted 10 July 2016. *Corresponding author (linkui@bnu.edu.cn). Assigned to Associate Editor Gerald Miller.

Abbreviations: chrOGR, chromosome-level orthologous genomic region; dnaOGR, orthologous genomic region detected using conserved DNA sequences as markers; GO, gene ontology; InDel, insertion–deletion; OGR, orthologous genomic region; proOGR, orthologous genomic region detected using conserved protein-coding genes as markers; SNV, single-nucleotide variant; SSR, simple-sequence repeat; UTR, untranslated region; WGA, whole-genome alignment.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2015.10.0099

related genomes is a reliable and indispensable prerequisite for the accurate detection of genomic changes. Various OGRs inferred at different levels have been successfully applied to address different biological problems. High-level, large-size OGRs have been used to construct ancestral genomes, infer whole-genome duplications, and detect large-scale genomic changes (Hu et al., 2011; Van de Peer et al., 2009), whereas low-level OGRs have been used to detect orthologous genes, construct phylogenetic trees, and infer small-scale genomic changes (Angiuoli and Salzberg, 2011; Zhang and Lin, 2012). The number and length of OGRs correlate with evolutionary distance. For very distantly related genomes, most, if not all, OGRs are limited to very short conserved elements or complete single protein-coding genes. By contrast, longer OGRs, which vary from adjacent gene clusters to an entire chromosome, may be observed or reconstructed from very closely related genomes.

To date, various whole-genome alignment (WGA) methods that compare different types of genomic markers can be used to detect different levels of OGRs between closely related genomes. Most alignment tools, such as Mugsy (Angiuoli and Salzberg, 2011), ProgressiveMauve (Darling et al., 2010), and TBA (Blanchette et al., 2004), use conserved DNA sequences as markers to detect OGRs. Among these tools, Mugsy uses segmental DNA sequences as markers and is one of the fastest and most efficient tools for detecting dnaOGRs among a set of closely related genomes (Earl et al., 2014). In addition to conserved DNA markers, some alignment tools use conserved protein-coding genes as markers to detect collinear genomic regions, namely proOGRs, which intuitively have much larger size than dnaOGRs, among related annotated genomes. These alignment tools use the protein sequences from orthologous protein-coding genes and their relative positions rather than nucleotides. Of this class of WGA tools, i-ADHoRe is one of the most useful tools for proOGR reconstruction (Proost et al., 2012). If two genomes are closely related, genetic markers, such as simple-sequence repeats (SSRs), can be used to identify very large OGRs at the chromosome level (chrOGRs) between two very closely related genomes. In the last decade, these different WGA strategies have been successfully applied to address various biological and evolutionary questions (Ren et al., 2009; Van de Peer et al., 2009; Zhang and Lin, 2012). However, few studies have simultaneously integrated the various OGRs from different levels (multiple-resolution maps). With the increasing number of well-annotated, very closely related genome sequences becoming available in public databases, a comprehensive hierarchical map of OGRs from different levels should provide insights into genome evolution.

The program Sibelia (Minkin et al., 2013) produces a so-called multiple-resolution map of OGRs via the addition of an iterative refinement procedure, which provides a range of granularity for the blocks by individually increasing *k*-mer sizes. However, this program is purely based on DNA sequence *k*-mers and may not detect large size OGRs such as proOGRs. In this paper, we propose a strategy to construct a hierarchical map from two very closely related genomes by developing a pipeline based on two existing WGA tools: Mugsy (Angiuoli and Salzberg, 2011) and i-ADHoRe (Angiuoli and Salzberg, 2011). Mugsy was used for dnaOGR detection, whereas i-ADHoRe was used for proOGR delineation. Because the average length of higher-level OGRs is longer than lower-level OGRs, each proOGR should contain one or more nonoverlapping dnaOGRs. This observation thus leads to the intuitive construction of a two-level hierarchical map of OGRs that represents the parent–child relationships between the proOGRs and dnaOGRs (Fig. 1). Here, we report on the performance of our pipeline using two published very closely related cucumber genomes: *C. sativus* var. *sativus* (a cultivated cucumber) (Huang et al., 2009) and *C. sativus* var. *hardwickii* (a wild cucumber) (Qi et al., 2013). Fortunately, based on previous work showing that the contigs and scaffolds are anchored onto their respective chromosomes (Ren et al., 2009), a third level of OGRs (chrOGRs) could be directly added to a previous map to form tree relationships of the OGRs (chrOGR:proOGRs:dnaOGRs) between the two cucumber genomes. The lack of information regarding the chromosomal positions for many sequenced genomes necessitates that this third level of OGRs can only be constructed among scaffolds or contigs rather than chromosomes. Based on this hierarchical OGR map, a corresponding map of the genomic changes between the two genomes was more directly and accurately determined. Large-scale genomic changes were identified by comparing the differences in the arrangements of the child OGRs between the two parent OGRs (the pair of OGRs from the two genomes) while treating their respective child OGRs as invariant elements. Intrachromosomal rearrangements were detected by comparing the structural differences in each chrOGR, whereas gene and segmental gains or losses were inferred from the comparison of the difference in the structure of each proOGR. The conventional comparison was used to compare each dnaOGR and to detect small-scale genomic changes such as SNVs and short InDels. This integrated three-level hierarchical map of OGRs identified by different methods with different genomic markers, together with the two scales of genomic changes detected by comparing each OGR at different levels, was robustly established for the two cucumber genomes. This map of OGRs and genomic changes with multiple resolutions provides an initial and useful resource to obtain insights into cucumber diversity and the evolution of genomes and phenotypes.

Materials and Methods

Global Gene Family Classification

Protein-coding genes from a cultivated cucumber, a wild cucumber, and three sequenced green plants, melon (*Cucumis melo* L.) (Garcia-Mas et al., 2012), watermelon [*Citrullus lanatus* (Thunb.) Matsum. & Nakai] (Guo et

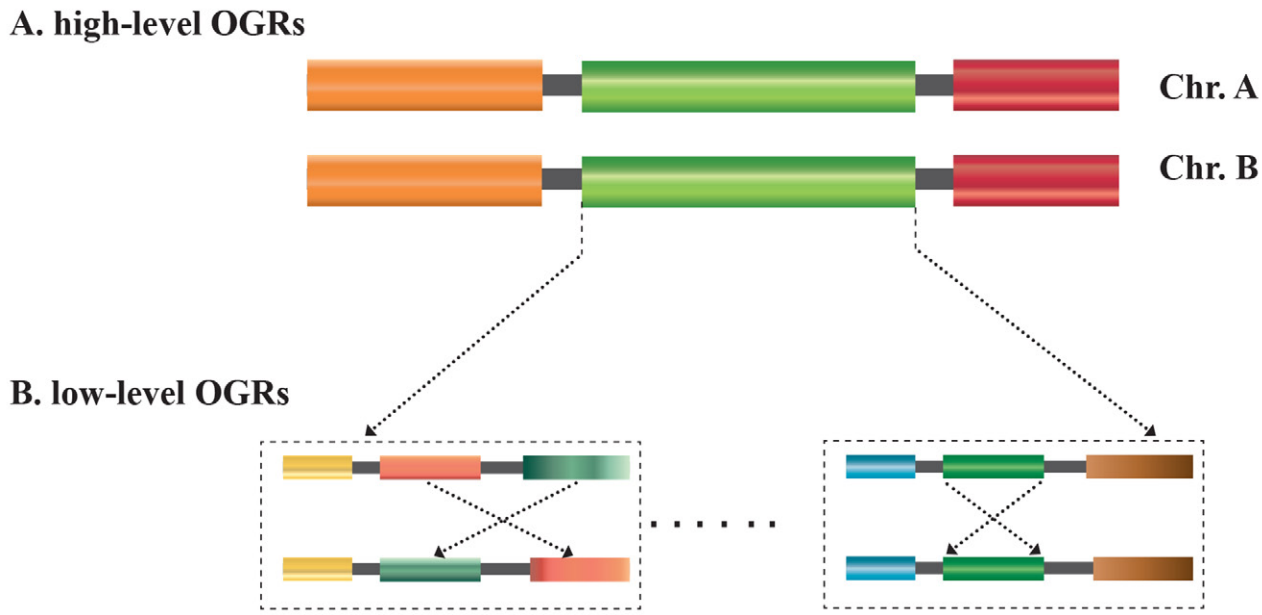

Fig. 1. The schematic diagram of hierarchical map of orthologous genomic regions (OGRs) between two genomes. The OGRs from different levels were detected through different types of markers. Since the average size of high-level OGRs would be much longer than low-level OGRs, each high-level OGR would contain one or more low-level OGRs, representing a parent–child structure. (A) A schematic diagram of high-level OGRs; (B) A schematic diagram of low-level OGRs, which position or orientation changes could represent the structural changes at the high-level OGR. The left one shows an intrachromosomal transposition and right one shows an inversion. The blue color indicates the regions are located at OGRs, and the yellow color indicates the regions are not located at OGRs.

al., 2013), and tomato (*Solanum lycopersicum* L.) (Tomato Genome Consortium, 2012), were analyzed. Detailed information is provided in Supplemental Table S1. For protein-coding genes with alternatively spliced isoforms, only the longest predicted protein sequence was maintained as a representative. A total of 133,704 protein-coding genes from the five genomes were used to assign gene family clusters using OrthoMCL software version 2.0.5 (Li et al., 2003). Based on pairwise sequence similarities between all input protein sequences, which were calculated using BLASTp with an *e*-value cutoff of 1×10^{-5} and an inflation value (*–I*) of 1.5, OrthoMCL classified 111,405 protein-coding genes into 21,488 families, of which, 18,877 families contain cucumber genes. These cucumber families were used to detect the homologous genomic regions.

Identification of Orthologous Genomic Regions at Different Levels

We first identified three levels of OGRs, namely chrOGRs, proOGRs, and dnaOGRs (Supplemental Fig. S1A–C), to construct the hierarchical map of the OGRs. Second, we mapped the proOGRs onto the chrOGRs to construct parent–child relationships called chr-proOGRs (Supplemental Fig. S1D). Similarly, we constructed pro-dnaOGRs (Supplemental Fig. S1E). Finally, a three-level hierarchical map describing the parent–child relationships among the detected OGRs was constructed by integrating the two parent–child relationship (D and E) between the OGRs (Supplemental Fig. S1F).

To investigate the chrOGRs, two publications reporting that the two cucumber genomes shared the same number of chromosomes (Ren et al., 2009; Yang et al., 2012) were used as the basis to construct seven pairs of homologous chromosomes between the two cucumber genomes (Supplemental Fig. S1A). The i-ADHoRe (version 3.0) was used to detect the proOGRs using the protein-coding genes as markers (Supplemental Fig. S1B). The software can be used to detect inversions and nested inversions, which are based on a conserved gene order and content, to identify homologous regions. The i-ADHoRe was run using the following parameters: alignment_method gg2, gap_size2, cluster_gap 3, q_value 0.75, prob_cutoff 0.01, multiple_hypothesis_correction FDR, anchor_points 3, and level_2_only false. Tandem gene duplicates were also determined using i-ADHoRe. Mugsy (version 2.0), a computationally efficient tool that can align closely related whole genomes, was used to detect the dnaOGRs (Supplemental Fig. S1C). Mugsy was run using the following parameters: -distance = 1000 and -minlength = 30, which specify the maximum genomic distance between adjacent anchors and the minimum block length, respectively. A previous investigation on cucumbers identified ancestral whole-genome duplications (Huang et al., 2009), which would add noise to the identification of OGRs and complicate the downstream analysis. Here, the identification of different types of genomic changes was limited to positional OGRs for simplification (Dewey, 2011), which could provide contextual information when paralogous genomic regions are present, and only the one-to-one orthologous regions were retained for further analysis.

The average length of high-level (large size) OGRs should be longer than the low-level (small size) OGRs; thus, each high-level OGR should contain one or more nonoverlapping low-level OGRs. The parent–child relationships between the OGRs were determined by mapping the low-level OGRs onto the adjacent high-level OGRs. Thus, two types of parent–child relationships among the OGRs were detected: pro-dnaOGRs by mapping dnaOGRs onto proOGRs and chr-proOGRs by mapping proOGRs onto chrOGRs (Supplemental Fig. S1D–E). A three-level hierarchical map of the OGRs was constructed from two very closely related genomes based on the integration of the two parent–child relationships between the OGRs (Supplemental Fig. S1F).

Identification of Rearrangements at the chrOGR and proOGR Levels

Because large-scale genomic changes can be detected from the parent–child OGR relationships, we can also detect large-scale genomic changes from the chr-proOGRs and pro-dnaOGRs (Supplemental Fig. S1D–E). For example, for one parent–child chr-proOGR relationship, we treated its proOGRs as invariant elements and then compared the structural differences in its corresponding chromosomal segments from the two genomes (Supplemental Fig. S2). Analogously, each pro-dnaOGR underwent the same computational procedure to detect putative large-scale genomic changes.

Supplemental Fig. S2 illustrates the method we used to detect different types of large-scale genomic changes. Inversions were detected using the orientations of all maintained child OGRs (Supplemental Fig. S2A). If the orientations were opposite from two segments of the child OGR, this child OGR would be treated as an inversion candidate. (B) If the orientations of the child OGRs between its respective chromosomal segments from the parent OGR were different and the child OGR and the two segments of the child and parent were from the same chromosomes, this child OGR would be treated as an intrachromosomal transposition (Supplemental Fig. S2B). If one chromosome from the child OGR was different from that of the parent OGR, this child OGR would be designated as an interchromosomal transposition (Supplemental Fig. S2C). If the region of the parent OGR was not overlapping with the child OGRs and the length was $\geq$30 bp, the region would be treated as a segmental gain or loss (Supplemental Fig. S2D).

Verification of the Detected Large-Scale Genomic Changes

Because systematic bias from the assembly process would add noise to the detection of large-scale genomic changes, two sets of raw sequencing reads were applied to filter out the raw candidates and to verify robust, candidate, large-scale genomic changes. These raw sequencing reads consisted of the original reads, which were used to assemble the two reference genomes. We mapped the paired-end reads onto the two cucumber genomes and then calculated the coverage of each OGR to ensure that all regions were covered by the reads with the exception of the segmental gains or losses. Based on the coverage and specific information about the orders and orientations from the paired-end reads, we validated the detected large-scale genomic changes. The following details were used to verify the large-scale genomic changes.

Using BWA software with default settings (Li and Durbin, 2010), each set of raw reads was mapped to two reference genomes. For each pair of OGRs (e.g., one segment is from genome A and the other segment is from B), the coverage of every segment was calculated: (i) the average coverage of the region from genome A was calculated using the raw sequencing reads (denoted as coverage_AA), (ii) the average coverage of the region from genome A was calculated by mapping the original sequencing reads from genome B (denoted as coverage_AB), (iii) the average coverage of the region from genome B was mapped using its own reads (denoted as coverage_BB), and (iv) the average coverage of the region from genome B was mapped using the original sequencing reads from genome A (denoted as coverage_BA). Only the OGRs with coverage_AA $\geq$ 1 and coverage_BB $\geq$ 1 were retained. The candidates were verified based on the paired-end reads. The criteria for each type of structural genomic change are as follows: (i) if the coverage_AA $\geq$ 1 and coverage_AB < 1, these segments were treated as segmental gains or losses; (ii) if the coverage_AA $\geq$ 1, coverage_AB $\geq$ 1, coverage_BB $\geq$ 1, coverage_BA $\geq$ 1, and at least one pair of paired-end reads were mapped to two segments from one OGR and the adjacent OGR, as shown in Supplemental Fig. S3A, these segments were treated as intrachromosomal transpositions; (iii) interchromosomal transpositions required coverage_AA $\geq$ 1, coverage_AB $\geq$ 1, coverage_BB $\geq$ 1, coverage_BA $\geq$ 1, and at least one sequencing read mapped onto both regions A and B; and (iv) inversions required coverage_AA $\geq$ 1, coverage_AB $\geq$ 1, coverage_BB $\geq$ 1, coverage_BA $\geq$ 1, and at least two pairs of reads mapped to both regions A and B; each read that was mapped onto the two segments were in opposite orientations, and the orders of these two reads were inverted at the two segments (Supplemental Fig. S3B).

Identification of Genomic Changes at the dnaOGR Level

The alignment of each dnaOGR detected by Mugsy (Angiuoli and Salzberg, 2011) was analyzed to detect the genomic changes, including SNVs, small InDels, and segmental gains or losses, at the dnaOGR level. By traversing each aligned position, we identified raw SNVs, small InDels, and segmental gains or losses. An unaligned region was marked as a segmental gain or loss candidate if its length was no less than 30 bp, otherwise it was marked as a small InDel. The raw SNVs and small InDels were filtered out to avoid bias from the assemblies and sequencing using the following criteria: no more than three variants among each sliding window with a window size of 100 bp and a step size of 1 bp; otherwise, the variant was filtered out.

Comparison of the Detected Small-Scale Genomic Changes with the Counterparts Detected by Read Mapping

In general, the small-scale genomic changes detected using different methods will not be the same. To evaluate the accuracy of the SNVs and small InDels identified based on the dnaOGR alignments, we also used the pipeline from SAMtools (version 0.1.18) (Li et al., 2009) to detect SNVs and small InDels based on read mapping and then compared these two datasets.

Two original sets of reads, which were used to assemble the reference genomes, were used to detect the SNVs and small InDels. One set was from a wild cucumber library created with 760 bp inserts, whereas the other set was from a cultivated cucumber library created with 788 bp inserts. The reads for each sample were mapped against the reference genome using BWA software (Li and Durbin, 2010). SAMtools (version 0.1.18) (Li et al., 2009) was applied to detect the SNVs and small InDels. As a result, up to 575,779 SNVs (accounting for 95%) and 64,760 small InDels (accounting for 41%) were shared between methods, as detected by SAMtools.

Annotation of the Genomic Changes at Different Scales

A detailed functional annotation is an important step in interpreting genomic changes. All genomic changes were mapped onto the genomic features and subsequently associated with the functional annotations to gain detailed insights into these changes. The identified SNVs and small InDels were further classified based on the gene annotation of the reference genomes. The SNVs and small InDels were categorized according to their locations, namely, in intergenic regions, 5′ untranslated regions (UTRs), coding sequences, introns, or 3′ UTRs. The SNVs in coding regions were further classified as synonymous SNVs that did not cause amino acid changes or nonsynonymous SNVs that caused amino acid changes. Small InDels in coding regions were further grouped as frame shift InDels that did or did not cause open reading frame shifts.

Gene Ontology Analysis

The mapping of the gene ontology (GO) terms and annotation analysis of the cultivated cucumber reference genome were performed by Blast2GO (v 2.5) using the default settings, which was based on the results of BLASTx against Swiss-Prot (2014-12-06). Blast2GO is a research tool designed with the main purpose of enabling GO-based data mining of sequence sets for which no GO annotation is yet available (Conesa et al., 2005). The GO enrichment analysis was performed with Ontologizer (Bauer et al., 2008), which can be used to perform a statistical analysis of the overrepresentation of GO terms in sets of genes.

Results

A Pipeline for Identifying Putative OGRs and Genomic Changes

In this study, we developed a strategy to integrate two different types of OGRs, dnaOGRs and proOGRs, from two very closely related genomes with well-annotated, protein-coding genes. The dnaOGRs and proOGRs were identified using Mugsy (Angiuoli and Salzberg, 2011) and i-ADHoRe (Proost et al., 2012), respectively. At the chromosomal level, a third type of OGRs (chrOGRs) could be easily constructed between these two genomes using genetic markers such as SSRs. Because each higher-level OGR should contain one or more lower-level OGRs, one or more three-level hierarchical maps of OGRs (chrOGRs:proOGRs:dnaOGRs) could be constructed (Supplemental Fig. S1F) by mapping the low-level OGRs onto the high-level OGRs (Supplemental Fig. S1D–E). Large-scale genomic changes, such as inversions, interchromosomal transpositions, intrachromosomal transpositions, and segmental gains or losses, and small-scale genomic changes could be detected using the hierarchical map of the OGRs. Supplemental Fig. S2 illustrates the large-scale genomic changes identified at the chrOGR and proOGR levels (refer to Materials and Methods section for more details). Using the hierarchical map of the OGRs and their related genomic changes, we can relate these changes to their neighboring annotated genomic elements (such as protein-coding genes, noncoding RNA genes, or repeats) to provide interpretations regarding the characteristics of the putative genomic change. Furthermore, although the current analysis focused on its application to two cucumber genomes, this pipeline can also be applied to other very closely related well-annotated genomes.

Three-Level Hierarchical Maps of the OGRs from Two Cucumber Genomes

We applied our pipeline to two very closely related cucumber genomes. Because of their short time of divergence and high-quality assemblies, the cucumber genomes were good resources to construct OGR maps. Figure 2 shows the three-level hierarchical map: the dnaOGR level is located in the outer circle, the proOGR level is located in the inner circle, and the chrOGR level is located in the innermost circle. These findings indicate that both the proOGRs and dnaOGRs contained most regions of the two cucumber genomes. Based on the conserved DNA sequences, 22,634 homologous genomic regions were detected by Mugsy. After filtering out duplications, 22,321 of the 22,634 regions were designated as dnaOGRs. With a minimum length of 30 bp and a maximum length of 615 kb, the dnaOGRs accounted for 96% of the domestic cucumber genome and 93% of the wild cucumber genome. Based on the gene family identified by OrthoMCL (Li et al., 2003) (Supplemental Table S2), 574 homologous genomic regions from two genomes were identified by i-ADHoRe. Of these 574 regions, 540 regions were designated as proOGRs and used for the

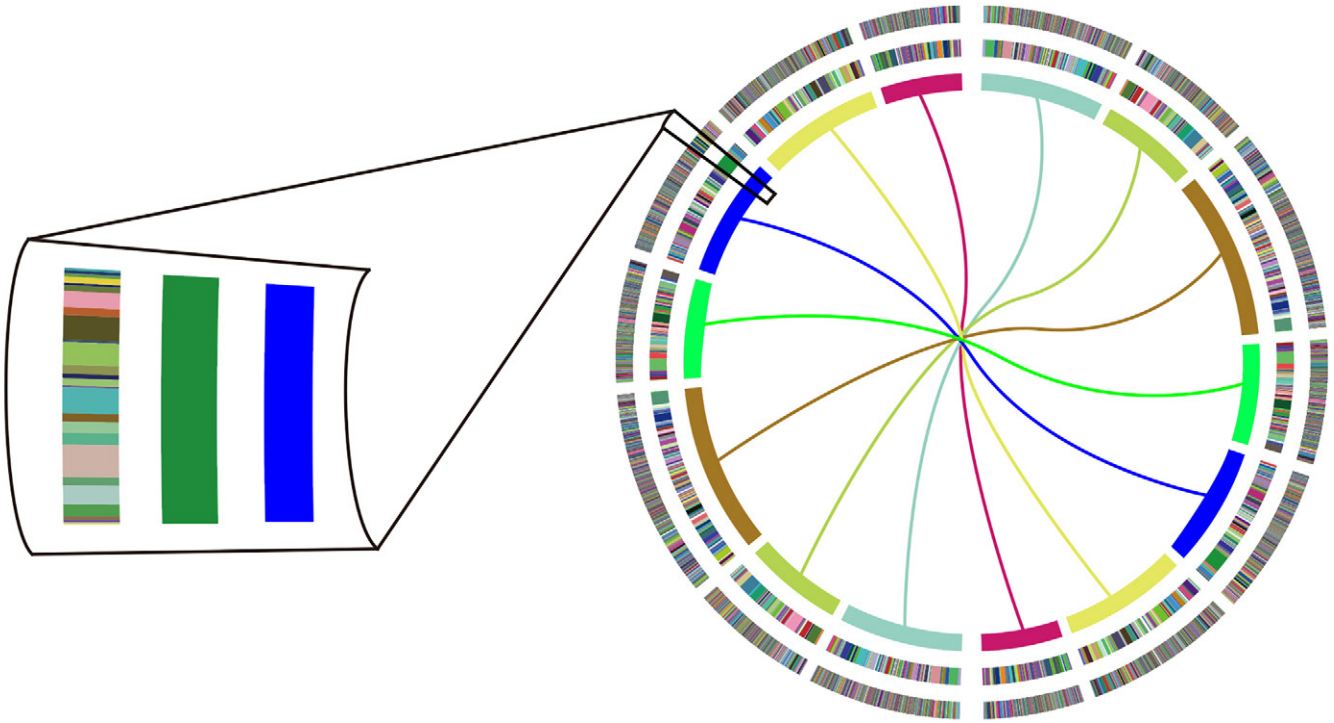

Fig. 2. The hierarchical map of orthologous genomic regions (OGRs) between two cucumbers. The hierarchical map with three levels was constructed including chrOGRs, proOGRs, and dnaOGRs. The chrOGRs were previously constructed based on simple-sequence repeats (Ren et al., 2009; Yang et al., 2012) and the proOGRs were detected by i-ADHoRe and the dnaOGRs by Mugsy (Angiuoli and Salzberg, 2011). Finally, in cucumber, seven chrOGRs (innermost circle), 540 proOGRs (inner circle), and 22,312 dnaOGRs (outmost circle) were detected.

subsequent analysis. These 540 proOGRs consisted of 86% of the cultivated cucumber genes and 85% of the wild cucumber genes, which covered 86.3 and 84.9% of the cultivated and wild cucumber reference genomes, respectively. The length distribution showed that the proOGR lengths ranged from 7576 to 3,648,515 bp, and the average length was 314,041 bp for the cultivated cucumber. At the chrOGR level, the length distributions of the assembled chromosomes from two genomes were similar (Ren et al., 2009; Yang et al., 2012). Seven chrOGRs accounted for 97.6 and 96.9% of the nucleotides from the cultivated and wild cucumber genomes, respectively. The detailed length distributions of the three-level OGRs are listed in Table 1.

As mentioned above, a tree structure can be used to organize the hierarchical relationships of the three-level OGRs. Among the pro-dnaOGRs, 80.7% (18,276 of 22,634) of the dnaOGRs overlapped with the 540 proOGRs. A total of 97.7 and 96.2% of the genomic regions of the 540 proOGRs were covered by the dnaOGRs in the cultivated cucumber and wild cucumber, respectively. There was substantial deviation among the numbers of dnaOGRs that each proOGR contained, which ranged from 1 to 252, with an average of 34.4. These findings implied that different proOGRs are likely to experience different rates of local genomic changes. For the chr-proOGRs, with the exception of the eight proOGRs that were not anchored on any chromosome, the remaining 532 (98.5%) proOGRs were mapped onto the chrOGRs. The number of proOGRs that each chrOGR contained varied from 53 to 114. Similarly, the chrOGRs may have also incurred local genomic changes at different rates (Supplemental Table S3–4).

Table 1. The length distribution of orthologous genomic regions (OGRs) at three different levels.

Level	No. of OGRs	Min.	Mean	Max.	Total	Coverage†
		bp				%
chrOGR	7	19,230,000	27,410,000	39,780,000	191,859,024	98
proOGR	540	7576	314,041	3649,000	169,582,133	86
dnaOGR	22,321	30	8448	615,600	188,567,808	96

† The coverage of cultivated cucumber genomes mapped by OGRs.

Detection of the Genomic Changes in Two Cucumber Genomes

Both large-scale and small-scale genomic changes between the two genomes could be more accurately detected within each chrOGR or proOGR. Because of the relatively short time of divergence between the two genomes, many large- and small-scale genomic changes were detected, indicating that there may be extremely complex relationships between the genotypes and phenotypes of the two cucumber genomes (Fig. 3).

At the chrOGR level, we identified 14 intrachromosomal transpositions with sizes that ranged from 28,752 to 167,847 bp, three interchromosomal transpositions

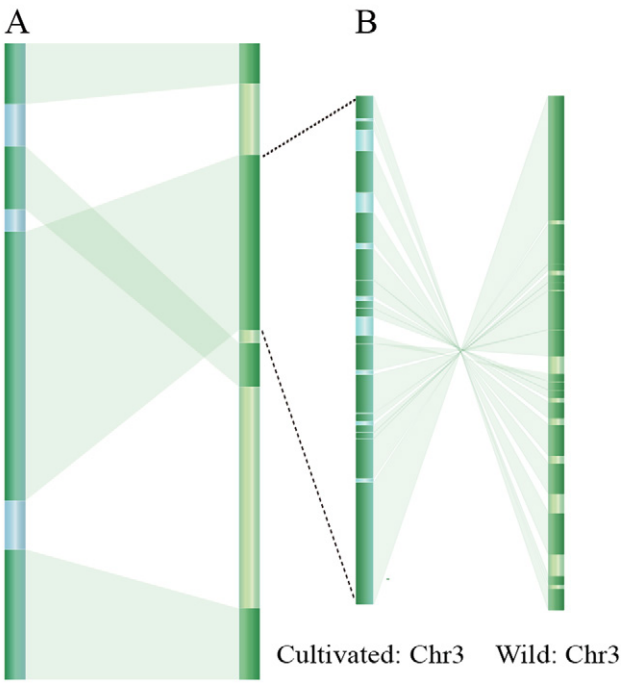

Fig. 3. An example of the large-scale genomic changes detected at the orthologous genomic regions (OGRs; chrOGR and proOGR levels) between two cucumbers. (A) There are four proOGRs present in the third chrOGR of the map (its identification is 3). Thus, according to their orders between the two chromosomes, there must have occurred an intrachromosomal transposition; (B) the third proOGR seems to be an inversion suggested by the orders of its 16 children dnaOGRs.

with sizes from 29,689 to 53,882 bp, and 542 segmental gains and losses that varied from 666 to 493,561 bp, which jointly affected ~23 Mb of sequences. Higher-resolution (lower-level) OGRs led to the detection of finer genomic changes; thus, additional rearrangements were detected at the proOGR level than the chrOGR level. Overall, we detected 386 intrachromosomal transpositions with sizes that ranged from 30 to 316,528 bp, 5304 interchromosomal transpositions with sizes varying from 30 to 12,826 bp, 2480 inversions with sizes ranging from 30 to 73,631 bp, and 5939 segmental gains or losses from 30 to 15,567 bp (Table 2; Supplemental Table S5).

At the dnaOGR level, 606,269 SNVs, 158,941 small InDels, and 38,620 segmental gains or losses were detected (Supplemental Table S6–7). Among the SNVs, the proportion of SNVs located in genes was 37% (222,682 of 606,269), whereas the other 63% of the SNVs were located in intergenic regions. Of the 222,682 SNVs, 24% (52,717 of 222,682) of the SNVs were located in the coding regions from 15,344 genes, which consisted of 8.7% of all detected SNVs. We further annotated the nonsynonymous and synonymous variants to investigate the functional annotations of the putative SNVs located in the coding DNA sequence regions. As a result, 44,158 nonsynonymous variants from 14,548 genes were detected (Supplemental Table S8). All detected genomic changes from the different levels heavily relied on the

Table 2. The summary of identified large-scale genomic changes at proOGR level.

Type	No. of genomic changes	No. of verified genomic changes	No. of genes located at verified changes
Intrachromosomal transposition	386	23	18
Interchromosomal transposition	5304	2000	147
Inversion	2840	175	105
Segmental gain or loss	5939	1726	141

assembled reference genomes; however, it is difficult to distinguish the large-scale genomic changes from the sequencing and assembling bias (Gurevich et al., 2013).

To reduce the systematic bias from the assembly process, we further used the original paired-end reads, which were used to assemble the reference genomes, to achieve more robust and reliable genomic rearrangements at different levels. However, because we were limited regarding the library insert sizes, only some of the large-scale genomic changes could be tested using the paired-end reads. In our case, we verified the large-scale genomic changes at the proOGR level. A total of 2000 interchromosomal transpositions, 23 intrachromosomal transpositions, 175 inversions, and 1726 segmental gains or losses were verified with average lengths of 1139, 374, 19,686, and 5736 bp, respectively (Supplemental Table S9).

We compared the detected large- and small-scale genomic changes with previous studies (Qi et al., 2013; Zhang et al., 2015). Because the previous data contained many accessions, here, we only focused on the variants from the accession CG0002, which was sequenced as the wild cucumber reference genome. The result revealed that 99% of the previously detected segmental gains or losses overlapped with our results. Of the seven previously detected inversions, five were identified in our results, and the other two inversions were missed because they were located outside the chrOGRs we detected. For the small-scale genomic changes, 86% of the previously detected SNVs overlapped with ours (Supplemental Table S10).

The putative functional characteristics of these more reliable large-scale genomic changes were annotated based on the genome structural annotation from the cultivated cucumber. At the proOGR level, 139 genes overlapped with the regions of 177 segmental gains or losses, 146 genes were mapped onto 259 interchromosomal transpositions, 18 genes were anchored on 16 intrachromosomal transpositions, and 105 genes were mapped on 64 inversions. Thus, there were 365 genes that could be related to large-scale genomic changes at the proOGR level (Table 2). Based on the GO analysis, eight of the 365 genes overlapping with the regions of the validated large-scale genomic changes were significantly enriched in functions related to flowering regulation (GO:0009909), and five of the 365 genes were significantly enriched in functions related to trichome morphogenesis (GO:0010090).

Discussion

The accurate identification of OGRs between two closely related genomes have proven to be important in our understanding of the diversity and evolution of genomes and phenotypes. Here, we developed a novel strategy for constructing a three-level hierarchical map of OGRs from two very closely related genomes. The primary advantage of our approach is its ability to build a comprehensive and general OGR map with multiple resolutions. Both small-scale and large-scale genomic changes could be accurately identified using these OGR maps. We tested our pipeline on two very closely related cucumber genomes, and the results provided an initial and useful resource to better understand cucumber diversity and the evolution of genomes and phenotypes.

Our method does have several potential limitations. Our work relies heavily on two WGA tools, the DNA-based aligner Mugsy (Angiuoli and Salzberg, 2011) and the protein-coding gene-based alignment approach i-ADHoRe (Proost et al., 2012), which are efficient tools for detecting dnaOGRs and proOGRs (Earl et al., 2014; Ghiurcuta and Moret, 2014), respectively. The qualities of both types of identified OGRs are heavily dependent on the quality of the two assembled genomes and the accuracy of their annotation of the protein-coding genes. In this study, two cucumber reference genomes have been assembled and anchored onto chromosomes with high quality using deep coverage sequencing (Huang et al., 2009; Qi et al., 2013). Moreover, both genomes were annotated using the same annotation pipeline that we built, which was mainly based on EVidenceModeler (Guo et al., 2013; Haas et al., 2008; Lin et al., 2014). The same annotation pipeline should greatly reduce the systematic bias of the accuracy of gene predictions among different annotation systems with the same genome sequences (Allen et al., 2004). With these foundations, our pipeline worked well for comparisons between the two cucumber genomes; thus, the construction of the three-level map of the OGRs and the identification of the related genomic changes should be accurate.

Based on the robust hierarchical OGRs, the detected genomic changes are reliable, particularly for large-scale genomic changes. However, a systematic detection bias still remains for large-scale genomic changes. We used the original paired-end reads, which were applied to assemble the reference genomes, to achieve more solid and reliable genomic changes and to reduce the systematic bias. In this process, only some of the large-scale genomic changes could be evaluated because of limitations in the insert size of the DNA libraries. Together, we verified 2023 transpositions, 175 inversions, and 1726 segmental gains or losses from the two cucumber genomes at the proOGR level. According to the results of the GO enrichment analysis, some of the verified genomic changes detected in this study were close to the genes associated with the regulation of the phenotypes of interest.

Our pipeline was designed to construct a three-level hierarchical map from two very closely related genomes, which have a short divergence time and share most of their genomes. For more divergent genomes that consist of very different chromosome structures, most OGRs may be limited to very short conserved elements or complete single protein-coding genes, which makes it difficult to reconstruct the type of hierarchical map constructed using our pipeline. With an increasing number of markers, such as ultraconserved elements (Siepel et al., 2005), we will attempt to extend our pipeline to include these markers to construct similar maps.

Acknowledgments

We thank three anonymous reviewers and Prof. Gary Muehlbauer for their valuable comments and suggestions. We are grateful to Zhonghua Zhang for his helpful discussion and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 31171235).

References

Allen, J.E., M. Pertea, and S.L. Salzberg. 2004. Computational gene prediction using multiple sources of evidence. Genome Res. 14:142–148. doi:10.1101/gr.1562804

Altshuler, D.M., R.M. Durbin, G.R. Abecasis, D.R. Bentley, A. Chakravarti, A.G. Clark, et al. 2012. An integrated map of genetic variation from 1092 human genomes. Nature 491:56–65. doi:10.1038/nature11632

Angiuoli, S.V., and S.L. Salzberg. 2011. Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342. doi:10.1093/bioinformatics/btq665

Bauer, S., S. Grossmann, M. Vingron, and P.N. Robinson. 2008. Ontologizer 2.0: A multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24:1650–1651. doi:10.1093/bioinformatics/btn250

Blanchette, M., W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715. doi:10.1101/gr.1933104

Conesa, A., S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676. doi:10.1093/bioinformatics/bti610

Darling, A.E., B. Mau and N.T. Perna. 2010. progressiveMauve: Multiple Genome Alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. doi:10.1371/journal.pone.0011147

Dewey, C.N. 2011. Positional orthology: Putting genomic evolutionary relationships into context. Brief. Bioinform. 12:401–412. doi:10.1093/bib/bbr040

Earl, D., N. Nguyen, G. Hickey, R.S. Harris, S. Fitzgerald, K. Beal, et al. 2014. Alignathon: A competitive assessment of whole-genome alignment methods. Genome Res. 24:2077–2089. doi:10.1101/gr.174920.114

Edwards, D., and J. Batley. 2004. Plant bioinformatics: From genome to phenome. Trends Biotechnol. 22:232–237. doi:10.1016/j.tibtech.2004.03.002

Garcia-Mas, J., A. Benjak, W. Sanseverino, M. Bourgeois, G. Mir, V.M. Gonzalez, et al. 2012. The genome of melon (*Cucumis melo* L.). Proc. Natl. Acad. Sci. USA 109:11872–11877. doi:10.1073/pnas.1205415109

Ghiurcuta, C.G., and B.M.E. Moret. 2014. Evaluating synteny for improved comparative studies. Bioinformatics 30:9–18. doi:10.1093/bioinformatics/btu259

Guo, S., J. Zhang, H. Sun, J. Salse, W.J. Lucas, H. Zhang, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat. Genet. 45:51–58. doi:10.1038/ng.2470

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. doi:10.1093/bioinformatics/btt086

Haas, B.J., S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biology 9:R7. doi:10.1186/gb-2008-9-1-r7.

Hu, T.T., P. Pattyn, E.G. Bakker, J. Cao, J.F. Cheng, R.M. Clark, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. 43:476–481. doi:10.1038/ng.807

Huang, S., R. Li, Z. Zhang, L. Li, X. Gu, W. Fan, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. 41:1275–1281. doi:10.1038/ng.475

Huang, X.H., N. Kurata, X.H. Wei, Z.X. Wang, A. Wang, Q. Zhao, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501. doi:10.1038/nature11532

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595. doi:10.1093/bioinformatics/btp698

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Li, L., C.J. Stoeckert, and D.S. Roos. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189. doi:10.1101/gr.1224503

Lin, K., E. Limpens, Z. Zhang, S. Ivanov, D.G.O. Saunders, D. Mu et al. 2014. Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. PLoS Genetics 10:e1004078. doi:10.1371/journal.pgen.1004078.

Minkin, I., H. Pham, E. Starostina, N. Vyahhi, and S. Pham. 2013. C-Sibelia: An easy-to-use and highly accurate tool for bacterial genome comparison. F1000 Res. 2:258–258. doi:10.12688/f1000research.2-258.v1

Proost, S., J. Fostier, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer et al. 2012. i-ADHoRe 3.0: Fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. 40:e11. doi:10.1093/nar/gkr955

Qi, J., X. Liu, D. Shen, H. Miao, B. Xie, X. Li, et al. 2013. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat. Genet. 45:1510–1515. doi:10.1038/ng.2801

Ren, Y., Z. Zhang, J. Liu, J.E. Staub, Y. Han, Z. Cheng, et al. 2009. Integrated genetic and cytogenetic map of the cucumber genome. PLoS One 4:e5795. doi:10.1371/journal.pone.0005795.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M.M. Hou, K. Rosenbloom, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050. doi:10.1101/gr.3715005

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. doi:10.1038/nature11119

Van de Peer, Y., J.A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele. 2009. The flowering world: A tale of duplications. Trends Plant Sci. 14:680–688. doi:10.1016/j.tplants.2009.09.001

Yang, L., D.H. Koo, Y. Li, X. Zhang, F. Luan, M.J. Havey, et al. 2012. Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. Plant J. 71:895–906. doi:10.1111/j.1365-313X.2012.05017.x

Zhang, Y. and K. Lin. 2012. A phylogenomic analysis of *Escherichia coli*/*Shigella* group: Implications of genomic features associated with pathogenicity and ecological adaptation. BMC Evolutionary Biology 12:174. doi:10.1186/1471-2148-12-174

Zhang, Z., L. Mao, H. Chen, F. Bu, G. Li, J. Sun, et al. 2015. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. Plant Cell 27:1595–1604. doi:10.1105/tpc.114.135848