# A hierarchical map of orthologous genomic regions reconstructed from two very closely related genomes: A case study of the cucumber (*Cucumis sativus*)

Huifen Cao, Erli Pang and Kui Lin*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal University, Beijing 100875, China.

*Corresponding author (linkui@bnu.edu.cn)

## Supplementary Material

**Supplementary Figures:**

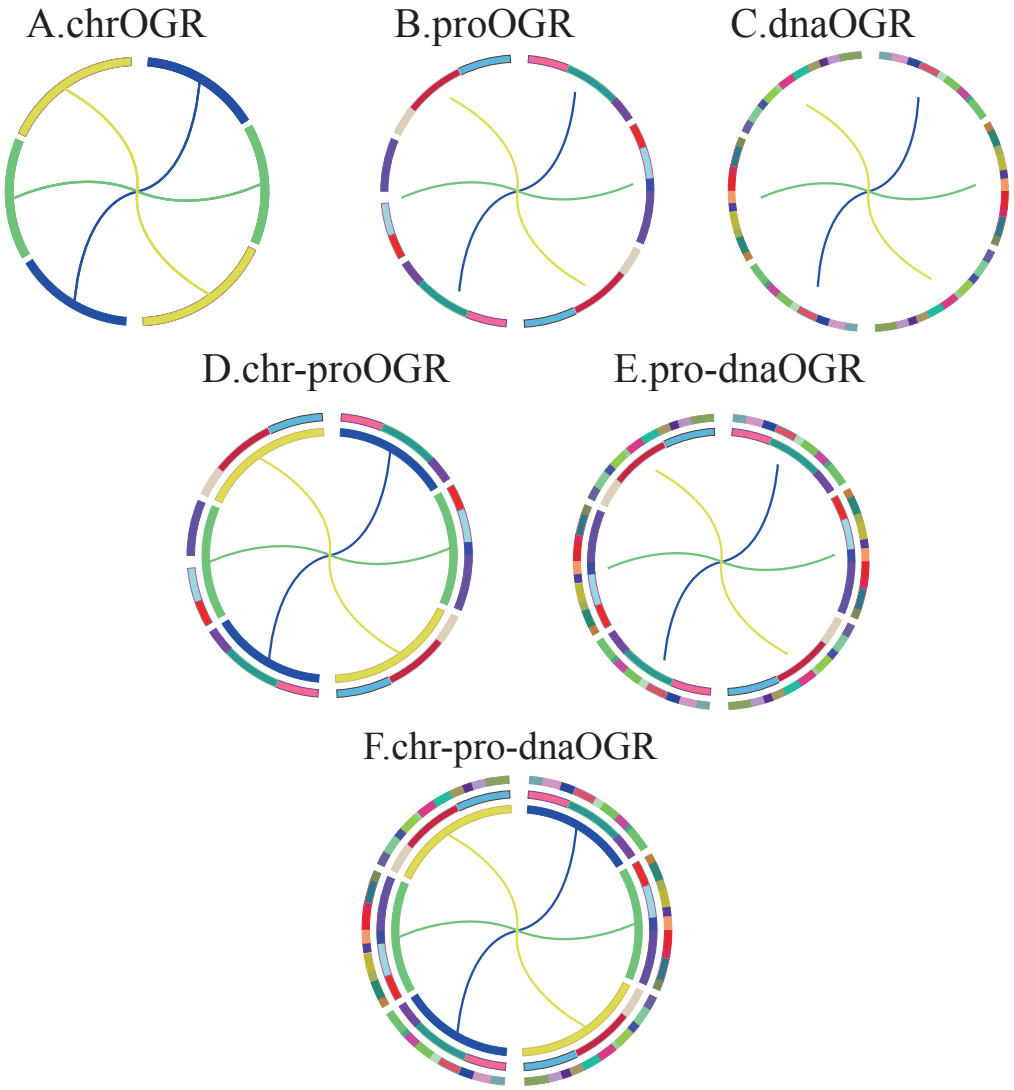


**Figure S1. The schematic diagram of constructing hierarchical map of OGRs between closely related genomes.** The OGRs from different levels were based on different types of markers. Each high-level OGR would contain one or more low-level OGRs. Based on the observation, we constructed a hierarchical map of OGRs. (A) The chrOGRs were detected by the previous work (SSRs as markers). (B) The proOGRs were detected by i-ADHoRe (protein-coding genes as markers). (C) The dnaOGRs were detected by Mugsy (conserved DNA sequences as markers). (D) The chr-proOGRs were constructed by mapping the proOGRs onto the chrOGRs. (E) The pro-dnaOGRs were constructed by mapping the dnaOGRs onto the proOGRs. (F) The three-level OGRs (chrOGRs:proOGRs:dnaOGRs) were constructed by integrating the chr-proOGRs and pro-dnaOGRs. Finally the three-level hierarchical map of OGRs were constructed.

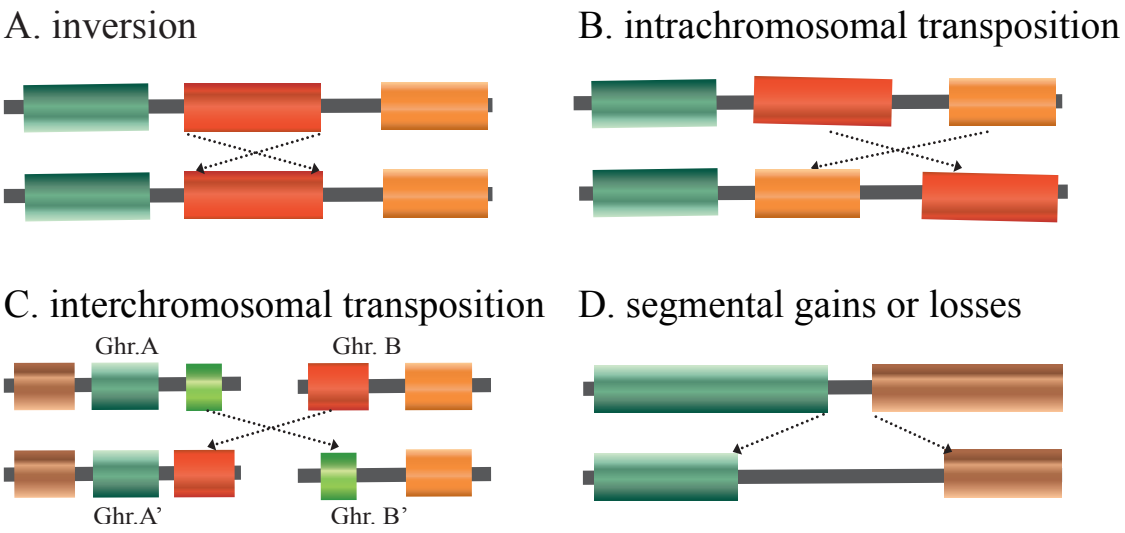


**Figure S2. Configurations of large-scale genomic changes detected in the hierarchical map of OGRs.** The large-scale genomic changes were detected by comparing the structural difference between its respective chromosomal segments from the two genomes, while treating their respective child OGRs as invariant markers. (A) When the orientations of two segments of the child OGR were opposite, this child OGR would be assigned as an inversion candidate. (B) If the orders of the child OGRs between its respective chromosomal segments from the parent OGR were different and the two segments of the child and parent were from the same chromosome, respectively, this child OGR would be treated as an intrachromsomal transposition. (C) If one chromosome from the child OGR was different with the chromosome of the parent OGR, this child OGR would be assigned as interchromsomal transposition. (D) If the regions of the parent OGR were not overlapped with the child OGRs, these regions would be treated as segmental gains or losses. The colorful bar represents the regions located at OGRs and the grey bar represents the regions not located at OGRs.

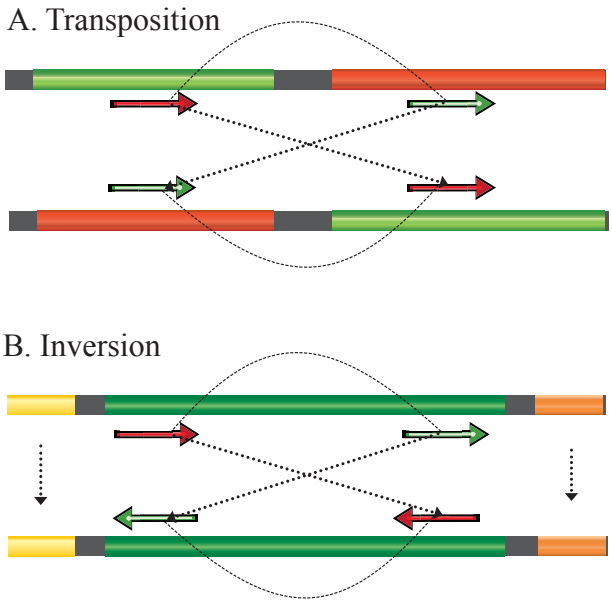


**Figure S3. Configurations of verified large-scale genomic changes by pair-end reads.** We used the pair-end reads to verify detected large-scale genomic changes: (a) Transposition, where the orders of two pairs of reads are different but the orientations are same; (b) inversion, where the orders from two pairs of reads are different and the orientations are opposite. The colorful bar represents the regions are located at OGRs and the grey bar represents the regions that are not located at OGRs.

## Supplementary Tables:

**Table S1. The websites and versions of the five genomes**

| Genomes | Version | The Date of Download | Websites |
|---|---|---|---|
| ***Cucumis sativus L.*** | V2 | 2014/08/10 | http://cmb.bnu.edu.cn/Cucumis_ sativus_v20/ |
| ***C. sativus* var. *hardwickii*** | V1 | 2014/08/10 | http://cmb.bnu.edu.cn/Cucumis_ sativus_v20/resequence/ |
| ***Cucumis melo L.*** | V3.5 | 2012/10/17 | https://melonomics.net/files/Genome/ |
| ***Citrullus lanatus*** | V1 | 2013/09/21 | ftp://www.icugi.org/pub/genome/watermelon/ |
| ***Solanum lycopersicum*** | V2.3 | 2013/08/08 | ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.3_release/ |

**Table S2. Numbers of gene families (and genes) in five plant genomes**

| Genomes | Singletons[†] | Two-genes[‡] | >=3[§] | Total gene families[¶] |
|---|---|---|---|---|
| ***Cucumis sativus L.*** | 16,948(16,948) | 1,032(2,064) | 435(2,008) | 18,415(21,020) |
| ***C. sativus* var. *hardwickii*** | 16,840(16,840) | 1,018(2,036) | 507(2,724) | 18,365(21,600) |
| ***Cucumis melo L.*** | 14,703(14,703) | 1,291(2,582) | 795(5,248) | 16,789(22,533) |
| ***Citrullus lanatus*** | 14,221(14,221) | 1,109(2,582) | 652(4,446) | 15,982(20,885) |
| ***Solanum lycopersicum*** | 9,065(9,065) | 3,076(6,152) | 1,675(10,150) | 13,816(25,367) |
| **All five genomes** | 71,777(71,777) | 7,526(15,052) | 4,064(24,576) | 21,488(111,405) |

[†]The number of singletons.

[‡]The number of gene families (genes) with two genes.

[§]The number of gene families (genes) contain at least three genes.

[¶]The number of total gene families (genes).

**Table S3. The statistics of dnaOGRs overlapped with proOGRs (in supported excel)**

**Table S4. The statistics of dnaOGRs overlapped with chrOGRs (in supported excel)**

**Table S5. The statistics of length distributions of large-scale genomic changes at proOGR level**

| Type | No. of genomic changes | Min. of length | Max. of length | Mean of length | Total length |
|---|---|---|---|---|---|
| Intrachromsomal transposition | 386 | 30 | 316,528 | 2,997 | 1,156,715 |
| Interchromsomal transposition | 5,304 | 30 | 12,826 | 240 | 1,273,820 |
| Inversion | 2,480 | 30 | 73,631 | 675 | 1,673,537 |
| Segmental gain or loss | 5,939 | 30 | 15,567 | 643 | 3,820,578 |

**Table S6. The distribution of the SNVs and Indels at dnaOGR level**

| Feature | No. of SNVs | No. of indels | No. of kept SNVs | No. of kept indels | Ratio of kept SNVs | Ratio of kept indels |
|---|---|---|---|---|---|---|
| Gene | 395,546 | 81,716 | 222,682 | 43,429 | 36.73% | 27.32% |
| Exon | 108,837 | 13,464 | 70,859 | 7,862 | 11.69% | 4.95% |
| CDS | 79,077 | 4,159 | 52,717 | 2,098 | 8.70% | 1.32% |
| 5'utr | 9,738 | 3,107 | 6,201 | 2,021 | 1.02% | 1.27% |
| 3'utr | 20,022 | 6,198 | 11,941 | 3,743 | 1.97% | 2.35% |
| Intron | 275,312 | 65,056 | 145,449 | 33,752 | 23.99% | 21.24% |
| Intergenetic | 1,628,026 | 399,422 | 383,373 | 115,450 | 63.23% | 72.64% |
| ncRNA | 1,162 | 362 | 214 | 62 | 0.04% | 0.04% |
| Total | 2,024,734 | 481,500 | 606,269 | 158,941 | 100.00% | 100.00% |

**Table S7. The statistics of the length distributions of large-scale genomic changes at chrOGR and dnaOGR levels**

| Type | No. of genomic changes | Min. of length | Max. of length | Mean of length | Total length |
|---|---|---|---|---|---|
| Intrachromsomal transposition† | 14 | 28,752 | 167,847 | 77,386 | 1,083,399 |
| Interchromsomal transposition† | 3 | 29,689 | 53,882 | 42,447 | 127,340 |
| Segmental gain or loss† | 542 | 666 | 493,561 | 42,190 | 22,867,006 |
| Segmental gain or loss‡ | 38,620 | 30 | 2,279 | 170 | 6,551,635 |

† Genomic changes at chrOGR level

‡ Genomic changes at dnaOGR level

**Table S8. The number of synonymous and nonsynonymous SNVs**

| Type | Synonymous | Nonsynonymous | Stop-codon SNVs |
|---|---|---|---|
| SNVs | 12,163 | 66,914 | 3,082 |
| Kept SNVs | 8,559 | 44,158 | 2,117 |

**Table S9. The statistics of length distributions of verified large-scale genomic changes at proOGR level**

| Type | No. of genomic changes | Min. of length | Max. of length | Mean of length | Total length |
|---|---|---|---|---|---|
| Intrachromsomal transposition | 23 | 70 | 316,528 | 19,686 | 452,789 |
| Interchromsomal transposition | 23 | 70 | 316,528 | 19,686 | 452,789 |
| Inversion | 174 | 32 | 73,631 | 5,736 | 998,150 |
| Segmental gain or loss | 1,726 | 30 | 15,567 | 1,129 | 1,948,044 |

**Table S10. The comparison of previous detected genomic changes and our detected genomic changes**

| Type | No. of Genomic Changes | No. of Overlapped Genomic Changes | Ratio of Overalapped Genomic Changes |
|---|---|---|---|
| Segmetnal gains or losses[†] | 1,936 | 1,922 | 99% |
| Inversions[†] | 7 | 5 | 71% |
| SNVs[‡] | 1,421,881 | 1,220,768 | 86% |
| Small Indels[‡] | 18,420 | 2,814 | 15% |

[†]The large-scale genomic changes were from the work of Zhang (Zhang, Mao, et al., 2015).

[‡] The small-scale genomic changes were unpublished and from the work of Qi (Qi, Liu, et al., 2013).

**Table S11. The number of frameshift indels**

| Type | Frameshift | Non-frameshift |
|---|---|---|
| Indels | 2, 330 | 1, 829 |
| Kept indels | 693 | 534 |

**Table S12. Statistics of overlapped proportions of detected SNVs and small indels at dnaOGR level with results detected by reads mapping**

| | No. of SNVs | No. of Indels | No. of shared SNVs | No. of shared indels | Ratio of SNVs overlap | Ratio of indel overlap |
|---|---|---|---|---|---|---|
| Raw genomic changes | 2,024,734 | 481,500 | 1,623,617 | 138,709 | 80% | 29% |
| Kept genomic changes | 606,269 | 158,941 | 575,779 | 64,760 | 95% | 41% |

**Table S13. The top 20 GO terms of verified large-scale genomic changes (in supported excel)**

**Table S14. The top 20 GO terms of verified small-scale genomic changes (in supported excel)**

**Table S15. The coordinates in the two cucumber draft genome assemblies of the chrOGRs (in supported excel)**

**Table S16. The coordinates in the two cucumber draft genome assemblies of the proOGRs (in supported excel)**

**Table S17. The coordinates in the two cucumber draft genome assemblies of the dnaOGRs (in supported excel).**

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. Nature Genet. 45: 1510-U1149. doi:10.1038/ng.2801.

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.