Title: Tandem repeats contribute to coding sequence variation in bumblebees (Hymenoptera: Apidae)

Authors: Xiaomeng Zhao[1#], Long Su[1#], Sarah Schaack[2], and Cheng Sun[1,*]

[1]Key Laboratory of Pollinating Insect Biology of the Ministry of Agriculture, Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing 100093, China

[2]Reed College, Department of Biology, Portland OR, 97202, USA

[#]Contribute equally.

[*]**Corresponding author**:

Cheng Sun
Key Laboratory of Pollinating Insect Biology of the Ministry of Agriculture
Institute of Apicultural Research
Chinese Academy of Agricultural Sciences
Beijing 100093, China
+86 131-2688-6081
Email: transposable.element@gmail.com

Abstract

Tandem repeats (TRs) are highly dynamic regions of the genome. Mutations occurred at these loci represent a significant source of genetic variation. Bumblebees are important pollinating insects occupying a wide range of habitats. However, to date, molecular mechanisms underlying the adaptation of bumblebees to such a diverse array of habitats are largely unknown. In the present study, we investigate how TRs contribute to genetic variation, thus potentially facilitating adaptation in bumblebees. We identified 26,595 TRs in the buff-tailed bumblebee (*Bombus terrestris*) genome, 66.7% of which reside in genic regions. We also compared TRs found in *B. terrestris* with those present in the whole genome sequence of a congener, *B. impatiens*. We found a total of 1,137 TRs were variable in length between the two sequenced bumblebee species, and further analysis reveals that 101 of them are located within coding regions. Variable-length TRs in coding regions were confirmed by PCR. Functional classification of bumblebee genes where coding sequences include variable-length TRs suggests that a majority of those genes are related to transcriptional regulation. Our results show that TRs contribute to coding sequence variation in bumblebees and TRs may facilitate the adaptation of bumblebees through diversifying proteins involved in controlling gene expression.

Introduction

Tandem repeats (TRs) are DNA tracts in which a short DNA sequence, dubbed a repeat unit, is repeated several times in tandem, and they are ubiquitous in the genomes of diverse species[1,2,3]. Most mutations in TRs are due to the variation in repeat unit number that occurs when one or more repeat units are added or deleted via a variety of different mutational mechanisms (e.g., polymerase slippage[4]). Because they are known to be highly variable, TRs are also known as VNTRs (variable number of tandem repeats)[1]. A number of cellular processes (for example, replication, recombination, DNA damage repair, and other aspects of DNA metabolism) and external factors are known to cause mutations in TRs, thus the frequency of mutations at these loci is thought to be 100 to 10,000 times higher than point mutations[4,5,6,7,8]. Mutations in TRs can have phenotypic consequences. Firstly, mutations in TRs residing in coding regions can impact the function or processing of messenger RNAs or proteins. Several neurodegenerative diseases have been linked to the repeat unit number variation of TRs located in coding regions, the most famous case being the abnormal expansion of a CAG repeat in exon 1 of the *IT15* gene leading to Huntington's disease (HD). Repeat numbers ranging from 6 to 35 are found in healthy individuals, whereas alleles with 40 repeats or more cause HD[9,10].

In addition to their role in disease, TRs in coding regions also confer phenotypic variability without major fitness costs. The repeat unit number variation in TRs located in *FLO1* gene in *Saccharomyces cerevisiae* generates plasticity in adherence to substrates[11]. In canines, variable TRs located in developmental genes confer variability to skeletal morphology[12]. Further, mutations in TRs located in non-coding regions can also have functional roles. Tandem repeats in promoters change gene expression in yeast[3], and contribute to gene expression variation in humans[13]. Therefore, given that TRs are highly mutable regions in the genome and therefore represent a significant source of variation, in cases where this variation is at loci influencing morphological, physiological and behavioral traits, it could facilitate the adaptation of the organism to different environments[1,3,11,14,15,16,17,18].

Bumblebees (Hymenoptera: Apidae) are a genus of pollinating insects that play an important role in agriculture production and ecosystem balance[19,20,21]. They are distributed widely across the globe, from Greenland to the Amazon Basin and from sea level to altitudes of 5800 m in the Himalayas[22]. Bumblebees occupy a remarkably

wide diversity of habitats, from alpine meadows to lowland tropical forest[23]. However, to date, molecular mechanisms underlying the adaptation of bumblebees to such a diverse array of habitats are largely unknown. Genetic variation is important for adaptation to new environments[24,25,26], however, little is known about sources or levels of genetic variation in bumblebees (but see[27,28]).

In the present study, we performed a systematic examination of TRs in the bumblebee genome and investigate their contribution to genetic variation in bumblebees. We further examine the functional significance of the genetic variation introduced by TRs to bumblebee genes. Lastly, we discuss the potential significance of the added genetic variation, especially as it may influence the regulation of gene expression.

Results

The identification of TRs in bumblebee genome

We used the whole genome sequence of *Bombus terrestris*[29] as a reference in which we identified TRs in the bumblebee genome using Tandem Repeats Finder[30]. After redundancy elimination (see Methods), a total of 26,595 TRs were identified. Considering the estimated genome size of *B. terrestris* is 249 Mb[29], the average density of TRs in *B. terrestris* is 106 TRs/Mb.

Molecular features of TRs in bumblebee

The distribution of repeat unit lengths of TRs in the bumblebee genome is summarized in Fig. 1A. In general, the number of TR loci decrease with increasing repeat unit length. However, there are exceptions: two peaks occur when the repeat unit length is 12 and 15 nt long. The top 10 most abundant repeat unit sequences, all either dinucleotide or trinucleotide, were determined (Fig. 1B) with the repeat unit "AG" as the most abundant in the bumblebee genome.

Most of the TR loci in the bumblebee genome are relatively short and 90% of TR loci have a length that is equal to or shorter than 111 base pair (bp) (Fig. 2A). To characterize the genome-wide distribution of TRs, the coordinates of TR loci were compared with the genome annotation for *B. terrestris*. Our results indicate that 66.7% (17,739 out of 26,595) of TRs in bumblebee genome were located within predicted genes (Fig. 2B).

TRs contribute to genetic variations in bumblebee

To understand the contribution of TRs to genetic variation in bumblebees, TRs identified in the non-repetitive regions of the *B. terrestris* genome were used as queries to find the orthologous loci in another sequenced bumblebee genome, *Bombus impatiens*. Based on the pairwise alignments between the TR array sequences from the two bumblebee species, we identified variable TRs (see Methods). Results suggest that a total of 2,862 TRs were located within the non-repetitive regions of the *B. terrestris* genome, and 1,137 of them are variable-length TRs (Supplementary Data 1). To understand if there are certain repeat unit lengths of TRs that are most likely to be sequence length variable in bumblebee, we calculated the ratio between the number of TRs showing variable in length between the two bumblebee species and the number of TRs that do not exhibit variable in length for each repeat unit length and plotted the ratio against the repeat unit length of TRs (Fig. 3). Generally, TRs with repeat unit length ranging from 2 to 10 bp are more likely to be sequence length variable than longer TRs (Fig. 3).

TRs contribute to coding sequence variation in bumblebee

To identify TRs generating coding sequence variation in bumblebee, we compared the genomic coordinates of the 1,137 variable TRs identified from the above step with those annotated as coding sequence (CDS) in the *B. terrestris*. We constructed pairwise alignments between protein sequences containing variable-length TRs to identify TRs generating protein sequence length variation between the two bumblebee species (see Methods for details). Based on this analysis, 101 of the 1,137 variable TRs exhibit coding sequence variation in bumblebee (Supplementary Data 2), and correspondingly, the genes harboring these TRs generate different length protein sequences (Supplementary Data 3).

In Figure 4, we show one example of a TR generating coding sequence variation; the focal TR has a repeat unit of CAG (encodes glutamine), and there are five more repeat units in *B. terrestris* than in *B. impatiens* (Fig. 4A). As a result, there are five more Qs (Q = glutamine residue in the one-letter code) in the protein sequence encoded by the TR-containing gene in *B. terrestris* than in *B. impatiens* (Fig. 4B). To further confirm that TRs promote coding sequence variation in bumblebee, we designed PCR primers that span the identified variable TRs in coding sequences and use them to amplify the genomic DNA extracted from 17 bumblebee species whose specimens are available in our lab (see Methods). Here, we show two examples of the PCR amplification results for variable TRs within the coding regions of bumblebee

genes (Fig. 5). Results suggest that there is a great deal of length polymorphism for the amplified bands between bumblebee species, indicating that TRs contribute to coding sequence variation in bumblebees more generally.

We took a closer look at the repeat unit length of the 101 variable-length TRs found in coding sequences. We observed 35 of them have a repeat unit length of 3, with all the other variable TRs having a repeat unit length of multiples of three (Supplementary Table S1). This finding is consistent with previous research in other species, which indicates that selection should favor or tolerate mutations that avoid high impact frameshift mutations[31,32,33,34].

Protein-coding gene sequence variation driven by TRs in bumblebee

The identified 101 variable TRs that contribute to coding sequence variation in bumblebee reside in 85 protein-coding genes. We performed a functional classification using PANTHER based on which 74 of them could be functionally classified. Over half of the classified genes (26 out of the 48 genes that could be assigned a molecular function) are involved in binding (Fig. 6A). The second most frequent molecular function is catalytic activity, with 15 genes falling in this category. Other molecular functions of classified genes include structural molecular activity, receptor activity, and transporter activity (Fig. 6A).

Proteins encoded by those genes containing variable-length TRs were assigned to 18 protein class categories, and the top 9 categories (categories having two or more genes) are shown in Fig. 6B. The most frequent protein class category represented is transcription factors and a total of 11 genes were found to encode them (Fig. 6B). Based on the recent release of KEGG BRITE database (http://www.genome.jp/kegg/brite.html, last updated on July 13, 2017), there are 10,581 protein-coding genes in *B. terrestris* genome, and 259 of them encode transcription factors. Thus, while ~2.45% of bumblebee genes encode transcription factors, 12.94% (11 out of 85) of the classified genes containing variable-length TRs are transcription factors-- a five-fold overrepresentation in this category. Other identified protein class categories include transferase and enzyme modulators (Fig. 6B).

Bumblebee genes where coding sequences contained variable-length TRs are involved in a variety of biological processes (Fig. 6C). The most frequent biological process categories are cellular and metabolic processes, each with 26 classified genes. Other

biological processes represented include biological regulation, developmental process, and response to stimulus. Significantly, genes containing variable-length TRs were involved in 8 known pathways, namely, Wnt signaling, Nicotinic acetylcholine receptor signaling pathway, Apoptosis signaling pathway, Alzheimer disease-presenilin pathway, 5HT2 type receptor mediated signaling, p38 MAPK pathway, Heterotrimeric G-protein signaling pathway, and Huntington's disease pathway. Interestingly, one bumblebee gene where the coding sequence contains variable-length TRs has the same tri-nucleotide repeat expansion (CAG) as that which causes Huntington's disease in humans (Fig. 4) and was determined to be involved in Huntington's disease pathway by PANTHER.

Discussion

Tandem repeats (TRs) are ubiquitous in the genomes of diverse species, where they represent highly dynamic regions of mutation and can thus facilitate the evolution of coding and regulatory sequences[1]. However, to date, little is known about TRs in bumblebees despite their importance as pollinator species and their wide range of habitats[22,23]. The present study represents the first systematic analysis of TRs in bumblebees. Our results indicate that TRs are abundant in bumblebee genome, where a total of 26,595 TRs were identified in *B. terrestris*, 1,137 of which are polymorphic when compared to a closely-related species, *B. impatiens*. Our analysis likely underestimates the true number of variable-length TRs among species of bumblebee because we only included TRs in non-repetitive regions (2,862) for subsequent analysis (see Method). Furthermore, variable-length TRs were identified based on a comparison of only two bumblebee species. There are 38 subgenera of bumblebees, and *B. terrestris* and *B. impatiens* only represent two[35-36]. Because genetic variation is an essential starting point for adaptation to new environments[24,25,26], we postulate TRs may contribute to adaptation of bumblebees across the many niches in which they are found.

Both changes in protein sequences and changes in gene expression could drive adaptation, although the relative importance of these two molecular mechanisms has long been controversial[12,37,38,39,40]. To understand the possible molecular mechanisms employed by TRs to facilitate adaptation in bumblebees, we focus on changes in protein sequences rather than changes in gene expression because even *cis*-regulatory

sequences, which are directly related to changes in gene expression[38], have not been extensively annotated in bumblebee genome yet. In this study, we searched for TRs that generate coding sequence variation, which in turn produce proteins of varying lengths (Supplementary Data 3). For the 101 variable-length TRs identified, all the repeat units have a length of multiples of three (Supplementary Table S1), which is consistent with findings in other species suggesting that natural selection may favor mutations that avoid frame-shifts[31,32,33,34]. Instead, mutations in TRs altering the length of protein sequences without introducing frame-shifts have the potential to majorly increase the functional diversity of host genes[1,11,12,41,42].

To understand the functional roles of genes affected by the 101 variable-length TRs, we did a functional classification. The genes could be assigned to 18 protein class categories, and we ranked them by their frequency (Fig. 6B). Results indicate that the most frequent protein class category is transcription factor with a total of 11 genes (Fig. 6B； Supplementary Table S1), which is ~five-fold overrepresentation than expected (see Results). Our initial goal of this study was to characterize the contribution of changes in protein-coding sequence driven by TRs in order to gain insight into the role of variable-length TRs in the adaptation of bumblebees. Interestingly, the most frequent protein class category identified, transcription factor, is directly related to changes in levels of gene expression[42,43,44]. Organisms can adapt to new environments by regulating gene expression at multiple stages of mRNA biogenesis, a process governed by many different proteins, such as transcription factors, chromatin-remodeling factors, signaling molecules, and receptors[43,44]. The second and the third most frequent protein class categories, transferases and enzyme modulators, respectively, are also involved in gene expression regulation (Fig. 6B). We checked all these protein class categories manually, and identified a total of 34 genes (out of the 39 genes that could be assigned to a protein class by PANTHER) involved in regulating gene expression (Supplementary Table S1). Altogether, our results indicate that TRs in bumblebee drive potentially functional variability at loci involved in gene expression regulation and other biological functions. As a result, length variation of TRs may facilitate the adaptation of bumblebees through diversifying bumblebee proteins, particularly those which regulate gene expression as has been previously hypothesized[37,38,40].

Conclusions

In the present study, we performed a comprehensive investigation of TRs in bumblebees. Our results indicate that TRs are abundant in bumblebee genome and a majority of them reside within genic regions. We found out that TRs represent a significant source of genetic variation in bumblebees. They promote coding sequence variation and influence the functional diversity of bumblebee genes. The functional roles of genes whose coding sequences contain variable-length TRs were analyzed, and our results indicate that a majority of those genes are related to transcriptional regulation. Given the importance of gene expression changes for adaptation, our observation that loci encoding transcription factors are enriched for variable-length TRs may suggest an important role for expanded repeats in the evolution of bumblebees.

Methods

Genomic sequences, annotation and predicted proteins

The genomic sequences, genome annotation, and predicted protein sequences of *Bombus terrestris* were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genome/2739, last accessed on April 5, 2016; GenBank assembly accession of GCF_000214255.1 [Bter_1.0]). The genomic sequences and predicted protein sequences of *Bombus impatiens* were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genome/3415, last accessed on April 5, 2016; GenBank assembly accession of GCA_000188095.2 [BIMP_2.0]).

Bumblebee genomic DNA

DNA was extracted from a single drone for each bumblebee species using Genomic tips and the blood and cell culture DNA kit (Qiagen). The drone specimen of *Bombus polaris* was provided by Paul Williams (Natural History Museum, London, England), and drones of all the other bumblebee species were collected in the summer of 2016 in China.

Identify TRs in *B.terrestris* genome

Each of the 18 chromosome sequences of *B. terrestris* was uploaded to Tandem Repeats Database (TRDB)[45]. First, the sequence of each chromosome was analyzed using Tandem Repeats Finder (TRF) using default parameters : 2, 7, 7, 50 (match, mismatch, indels, minimal alignment score)[30]. Then, TRs with repeat unit length less

than 2 or array length less than 30 bp were discarded. Finally, redundant TRs reported for the same loci were excluded using the Redundancy Elimination tool at TRDB. For redundancy elimination, if TRs overlapped by more than 50% of their length, the repeat with the longer array was retained, or in the case of ties, the repeat with the shorter repeat unit length was retained. Manual correction was done when necessary.

Characterizing the molecular features of TRs

The molecular features of TRs in *B. terrestris*, including repeat unit and repeat unit length distribution, TR array length distribution and genomic locations, were gleaned from the non-redundant TRs obtained from the above step by using a set of in-house Perl scripts, which are available upon request.

Mining variable-length TRs between *B. terrestris* and *B. impatiens*

The sequence of each TR array, along with 100 bp of upstream and downstream flanking sequence, was extracted from the soft-masked *B. terrestris* genomic sequences (GCF_000214255.1). Second, if there were continuous lower-case letters longer than 10 bp in either flanking sequence, indicating that the TR may reside in a repetitive region, the TR locus was excluded from further analysis. The sequences of the remaining TR loci, along with their 100 bp flanking regions, were used as queries to do BLASTn searches against the genomic sequence of *Bombus impatiens*, with an e-value cutoff of 1e-10. For each query, we retained the best hit (based on e-value) that included both the TR array sequence and more than 95 bp of flanking sequences on both sides (because these hits likely represent the query's orthologous locus in the *B. impatiens* genome). Finally, the pairwise alignments between the sequences of the TR arrays in *B. terrestris* and their best hits in *B. impatiens* were parsed to check if sequence length variation was observed within the TR array.

Identify TRs contributing to coding sequences variation

The coordinates of the identified variable-length TRs from the above step were used to search against the genome annotation of *B. terrestris* to identify those that resided in predicted CDS (coding DNA) sequence. Then, whenever one variable-length TR was found in the coding sequence of one *B. terrestris* gene, the full-length protein sequence encoded by this *B. terrestris* gene was used as a query to do BLASTp search against the protein database of *B. impatiens* to find the best hit from *B. impatiens*. Finally, based on the pairwise alignments between the protein sequences of the query and its best hit, we checked for amino acid sequence variation caused by the variable-length TR (e.g., if one or more amino acid residues were added or deleted from one

bumblebee species). If there was variation in the amino acid sequence, the variable-length TR was believed to contribute to the coding sequence variation in bumblebee.

PCR amplification of variable TRs in coding sequences

The sequences of variable-length TRs residing in coding sequences, along with 200 bp of flanking sequences, were extracted from the genomic sequence of *B. terrestris*. Then, polymerase chain reaction (PCR) was used with primers spanning the variable-length TRs. Primers were designed based on the obtained sequences using Primer 3[44]. Primers designed for the variable TR residing in the gene that encodes XP_012167698.1 are: Forward, 5'-CTGATGGCATCGTAGCTGGT-3'; and Reverse, 5'-GCTACCCTCAAAAGC
CGGAT-3'. Primers designed for the variable TR residing in the gene that encodes XP_012169902.1 are: Forward, 5'GTCGCGCAGTAGCTAGAAGT-3'; and Reverse, 5'-CCCCTCTCTGAAGCGTCTTC-3'.

A 15 μL reaction mixture composed of 50 ng of template DNA, 0.3 μL of 10 mM each deoxynucleotide triphosphate (dNTP), 0.4 units of *Taq* DNA polymerase (Sangon Biotech, Shanghai, China), 1.5 μL of 10× PCR buffer with Mg^{2+}, and 1.2 μL of 10 μmol/L forward and reverse PCR primers was prepared.

Amplification was carried out using the following reaction conditions: initial denaturation at 94°C for 5 min, followed by 35 cycles of 30 s at 94°C, 30 s at 56°C, and 30 s at 72°C, with a final extension at 72°C for 10 min. PCR products were separated on 8% polyacrylamide denaturing gels, and the bands were revealed by silver-staining[47].

Functional classification of genes containing variable TRs

We used the predicted protein sequences of *B. terrestris* genes containing variable-length TRs as queries to do local BLASTp against the downloaded Swiss-Prot database (http://www.uniprot.org/uniprot/, last accessed on September 1, 2016), with an e-value cutoff of 1e-10. The UniProt accession of the best hit was used to represent this gene. The collected UniProt accessions were uploaded onto the PANTHER server (http://pantherdb.org/) and classified by PANTHER system[48]. If a TR-containing gene did not get significant hit from Swiss-Prot database or the obtained UniProt accession could not be mapped using PANTHER, we used the protein sequence encoded by the *B. terrestris* gene as query to do search against the PANTHER library Version 12.0 (http://pantherdb.org/) with default settings to get a UniProt accession which could be recognized by PANTHER system to represent the *B. terrestris* gene.

Data Availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

References

1. Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet*. **44**, 445-477 (2010).
2. Melters, D.P. *et al*. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol*. **14**, R10 (2013).
3. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K.J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213-1216 (2009).
4. Tachida, H. & Iizuka, M. Persistence of repeated sequences that evolve by replication slippage. *Genetics* **131**, 471-478 (1992).
5. Paques, F., Leung, W.Y. & Haber, J.E. Expansions and contractions in a tandem repeat induced by double-strand break repair. *Mol*. *Cell*. *Biol*. **18**, 2045-2054 (1998).
6. Schmidt, A.L. & Mitter, V. Microsatellite mutation directed by an external stimulus. *Mutat*. *Res*. **568**, 233-243 (2004).
7. Rando, O.J. & Verstrepen, K.J. Timescales of genetic and epigenetic inheritance. *Cell* **128**, 655-668 (2007).
8. Lopez Castel, A., Cleary, J.D. & Pearson, C.E. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol*. **11**, 165-170 (2010).
9. Duyao, M. *et al*. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat Genet*. **4**, 387-392 (1993).
10. Gatchel, J.R. & Zoghbi, H.Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat*. *Rev*. *Genet*. **6**, 743-755 (2005).
11. Verstrepen, K.J., Jansen, A., Lewitter, F. & Fink, G.R. Intragenic tandem repeats generate functional variability. *Nat*. *Genet*. **37**, 986-990 (2005).
12. Fondon, J.W. 3rd & Garner, H.R. Molecular origins of rapid and continuous

morphological evolution. *Proc. Natl. Acad. Sci. USA*. **101**, 18058-18063 (2004).

13. Gymrek, M. *et al*. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet*. **48**, 22-29 (2016).
14. Tautz, D., Trick, M., Dover, G.A. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322**, 652-656 (1986).
15. Fidalgo, M., Barrales, R.R., Ibeas, J.I. & Jimenez, J. Adaptive evolution by mutations in the FLO11 gene. *Proc. Natl Acad. Sci. USA*. **103**, 11228-11233 (2006).
16. Fonville, N.C., Ward, R.M. & Mittelman, D. Stress-induced modulators of repeat instability and genome evolution. *J Mol Microbiol Biotechnol*. **21**, 36-44 (2011).
17. Zhou, K., Aertsen, A. & Michiels, C.W. The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol Rev*. **38**, 119-141 (2014).
18. Feliciello, I., Akrap, I., Brajković, J., Zlatar, I. & Ugarković, Đ. Satellite DNA as a driver of population divergence in the red flour beetle Tribolium castaneum. *Genome Biol Evol*. **7**, 228-239 (2014).
19. Fontaine, C., Dajoz, I., Meriguet, J. & Loreau, M. Functional diversity of plant-pollinator interaction webs enhances the persistence of plant communities. *Plos Biol*. **4**, e1 (2006).
20. Velthuis, H.H.W. & van Doorn, A. A century of advances in bumblebee domestication and the economic and environmental aspects of its commercialization for pollination. *Apidologie* **37**, 421–451(2006).
21. Garibaldi, L.A. *et al*. Wild pollinators enhance fruit set of crops regardless of honey bee abundance. *Science* **339**, 1608-1611 (2013).
22. Williams, P.H. A preliminary cladistic investigation of relationships among the bumble bees (Hymenoptera, Apidae). *Systematic Entomology* **10**, 239–255 (1985).
23. Sakagami, S.F. Specific differences in the bionomic characters of bumblebees: a comparative review. Journal of the Faculty of Science, *Hokkaido University Series VI, Zoology* **20**, 390–447 (1976).
24. Lande, R. & Shannon, S. The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution* **50**, 434-437 (1996).

25. Barrett, R.D. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol Evol*. **23**, 38-44 (2008).
26. Paaby, A.B. & Rockman, M.V. Cryptic genetic variation: evolution's hidden substrate. *Nat Rev Genet*. **15**, 247-258 (2014).
27. Lozier, J.D., Strange, J.P., Stewart, I.J. & Cameron, S.A. Patterns of range-wide genetic variation in six North American bumble bee (Apidae: Bombus) species. *Mol Ecol*. **20**, 4870-4888 (2011).
28. Maebe, K. *et al*. A century of temporal stability of genetic diversity in wild bumblebees. *Sci Rep*. **6**, 38289 (2016).
29. Sadd, B.M. *et al*. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol*. **16**, 76 (2015).
30. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. **27**, 573-580 (1999).
31. Young, E.T., Sloan, J.S. & Van Riper, K. Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. *Genetics* **154**, 1053-1068 (2000).
32. Richard, G.F. & Dujon, B. Molecular evolution of minisatellites in hemiascomycetous yeasts. *Mol Biol Evol*. **23**, 189-202 (2006).
33. Legendre, M., Pochet, N., Pak, T. & Verstrepen, K.J. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res*. **17**, 1787-1796 (2007).
34. Mularoni, L., Ledda, A., Toll-Riera, M. & Albà, M.M. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res*. **20**, 745-754 (2010).
35. Cameron, S.A., Hines, H.M. & Williams, P.H. A comprehensive phylogeny of the bumble bees (Bombus). *Biol J Linn Soc*. **91**, 161–188 (2007).
36. Hines, H.M. Historical biogeography, divergence times, and diversification patterns of bumble bees (Hymenoptera: Apidae: Bombus). *Syst Biol*. **57**, 58-75 (2008).
37. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).
38. Wray, G.A. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*. **8**, 206-216 (2007).
39. Hancock, A.M. *et al*. Adaptations to climate-mediated selective pressures in

humans. *PLoS Genet*. **7**, e1001375 (2011).
40. Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Res*. **23**, 1089-1096 (2013).
41. Caburet, S., Cocquet, J., Vaiman, D. & Veitia, R.A. Coding repeats and evolutionary "agility". *Bioessays* **27**, 581-587 (2005).
42. Radó-Trilla, N. *et al*. Key Role of Amino Acid Repeat Expansions in the Functional Diversification of Duplicated Transcription Factors. *Mol Biol Evol*. **32**, 2263-2272 (2015).
43. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat Rev Genet*. **12**, 833-845 (2011).
44. Kadonaga, J.T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116**, 247-257 (2004).
45. Gelfand, Y., Rodriguez, A. & Benson, G. TRDB-The Tandem Repeats Database. *Nucleic Acids Res*. **35**, D80-87 (2007).
46. Untergasser, A. *et al*. Primer3-new capabilities and interfaces. *Nucleic Acids Res*. **40**, e115 (2012).
47. Panaud, O., Chen, X. & McCouch, S.R. Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (Oryza sativa L.). *Mol. Gen. Genet*. **252**, 597-607 (1996).
48. Mi, H., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc*. **8**, 1551-1566 (2013).

Acknowledgements

We thank members of Dr. Jiandong An's group (Institute of Apicultural Research, Chinese Academy of Agricultural Sciences) for providing excellent assistance with bumblebee collection in China and for helpful discussion of the manuscript. We thank Dr. Paul Williams (Natural History Museum, London, England) for providing the specimen of *Bombus polaris*. This work was supported by the Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences [CAAS-ASTIP-2017-IAR], the Elite Youth Program of Chinese Academy of Agricultural Sciences [to CS], and National Science Foundation [MCB-1150213] funding [to SS].

Authors' contributions

CS conceived of the study; all authors contributed to study design; XZ, LS, and CS performed analyses with input from SS; LS performed the PCR verification of variable-length TRs in coding sequences; CS and SS wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Figure legends

Figure 1. Repeat unit features for TRs identified in bumblebee. (A) Repeat unit length distribution of TRs in bumblebee. Only repeat unit lengths, at which there are more than 100 TR loci in bumblebee, were shown. (B) The top 10 most abundant repeat unit sequences in bumblebee.

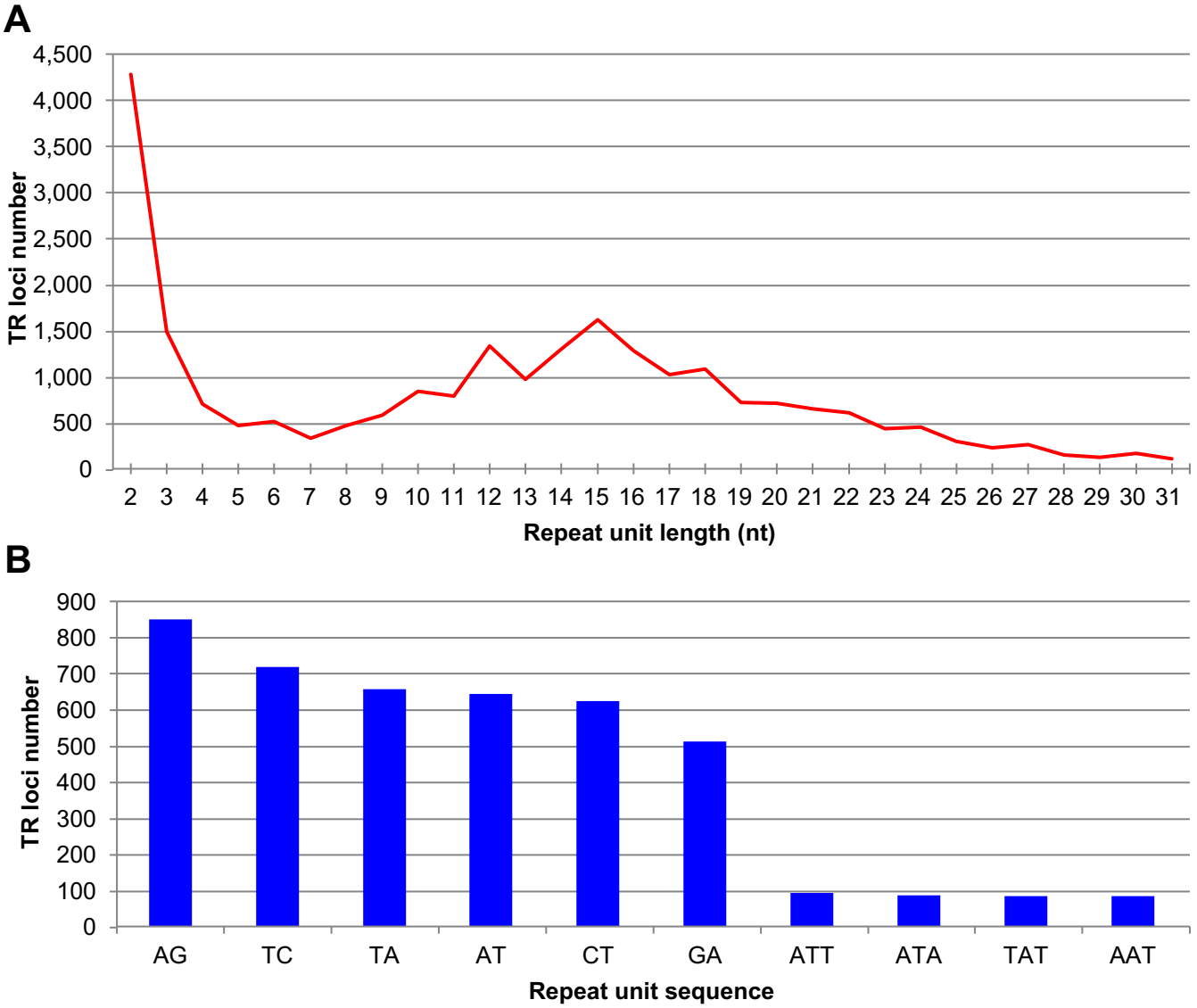
Figure 2. Distribution features for TR loci in bumblebee. (A) TR locus length distribution in bumblebee. (B) The distance between TRs and predicted genes. As shown in the figure, a majority of TRs in bumblebee reside within genes.

Figure 3. The relationship between repeat unit length and mutation propensity for TRs in bumblebee. The ratio between the number of TRs showing variable in length and the number of TRs that do not exhibit variable in length was plotted against the repeat unit length of TRs. Generally, TRs with repeat unit length ranging from 2 to 10 bp are more likely to be sequence length variable than longer TRs.

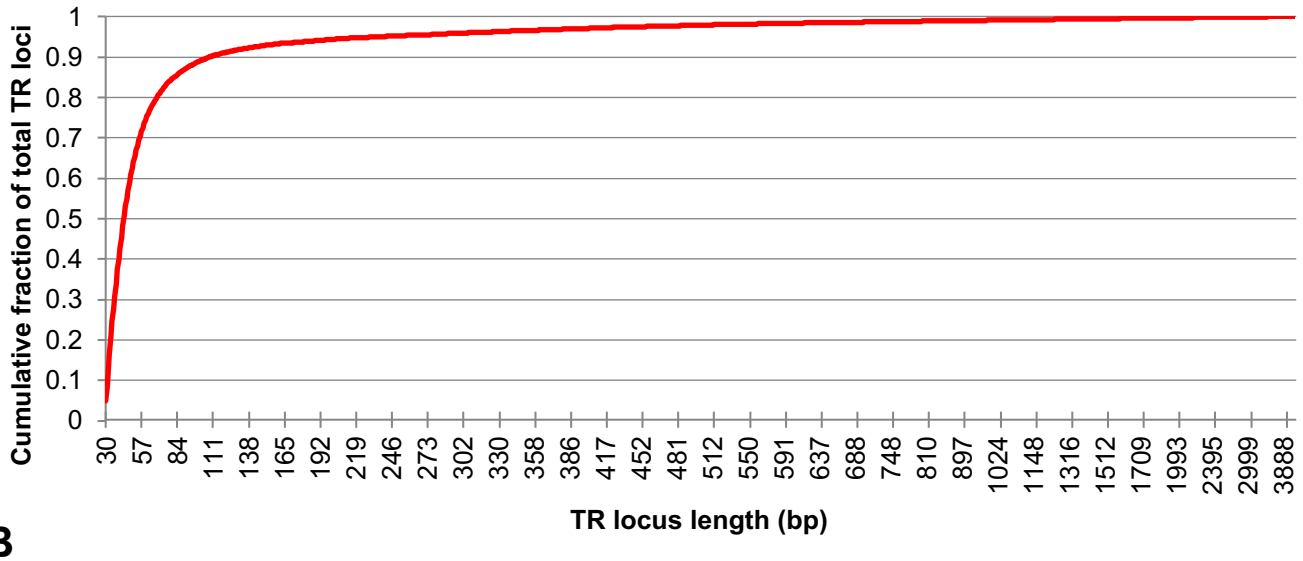
Figure 4. An example of TRs contributing to coding sequence variation in bumblebee. (A) Pairwise alignments of TR arrays between *B. terrestris* and *B. impatiens*. Colored letters indicate TR array sequences, while black letters show their flanking sequences. The TR array has a repeat unit of CAG, and there are five more repeat units in *B. terrestris* than in *B. impatiens*. The coordinate for the variable TR is NC_015770.1:2190704-2190753 in *B. terrestris*. (B) Pairwise alignments of protein sequences encoded by genes containing the variable TR. Colored letters indicate TR array sequences, while black letters show their flanking sequences. There are five more glutamine residues (Q) in *B. terrestris* than in *B. impatiens*. Genes containing this variable TR encode XP_012166765.1 and XP_012249688.1 in *B. terrestris* and *B. impatiens*, respectively.

Figure 5. PCR amplification of variable-length TRs residing in coding sequences across species of bumblebee. (A) PCR amplification of the variable-length TRs residing in the gene that encodes XP_012167698.1. (B) PCR amplification of the variable-length TRs residing in the gene that encodes XP_012169902.1. BPO: B. polaris; BPI: B. picipes; BBR: B. breviceps; BOP: B. opulentus; BCO: B. consobrinus; BIG: B. *ignitus*; BHA: *B. haemorrhoidalis*; BSU: *B. superbus*; BPY: *B. pyrosoma*; BDI: *B. difficillimus*; BSK: *B. skorikovi*; BSO: *B. soroeensis*; BTU: *B. turneri*; BWA: *B. waltoni*; BSI: *B. sibiricus*; BCU: *B. cullumanus*; BCF: *B. confuses*.

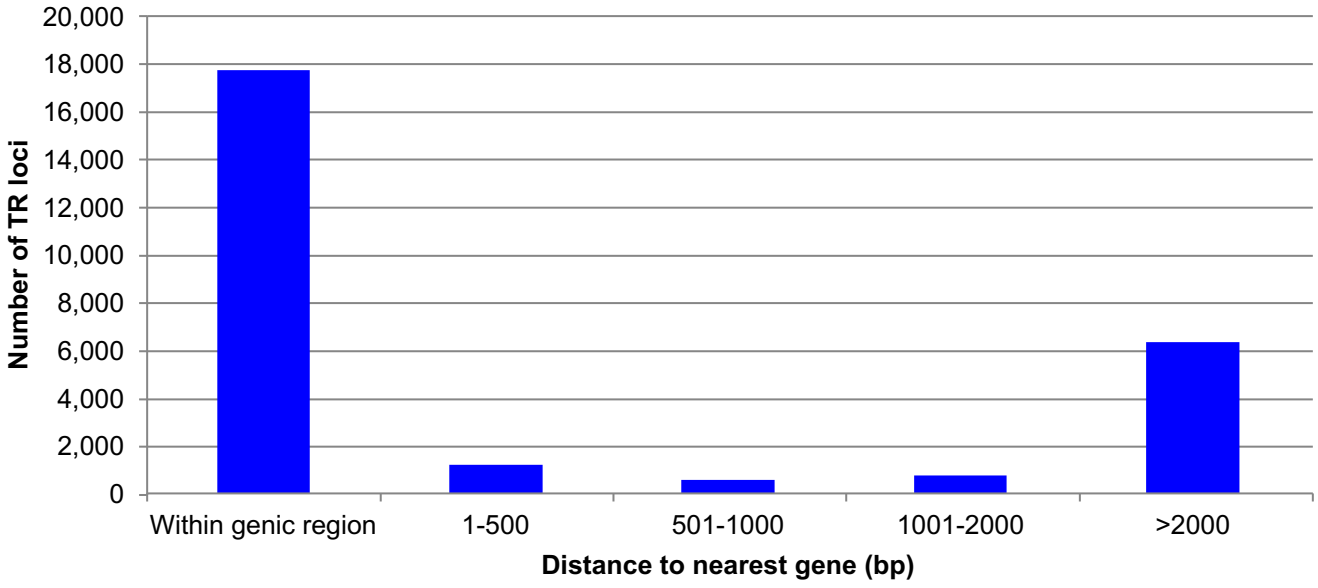
Figure 6. Functional classification of genes that include variable-length TRs. (A) The number of genes classified in each molecular function category. (B) The number of genes classified in each protein class. The gene number shown in the nucleic acid binding category excludes transcription factors. (C) Biological processes that genes including variable-length TRs are involved in.

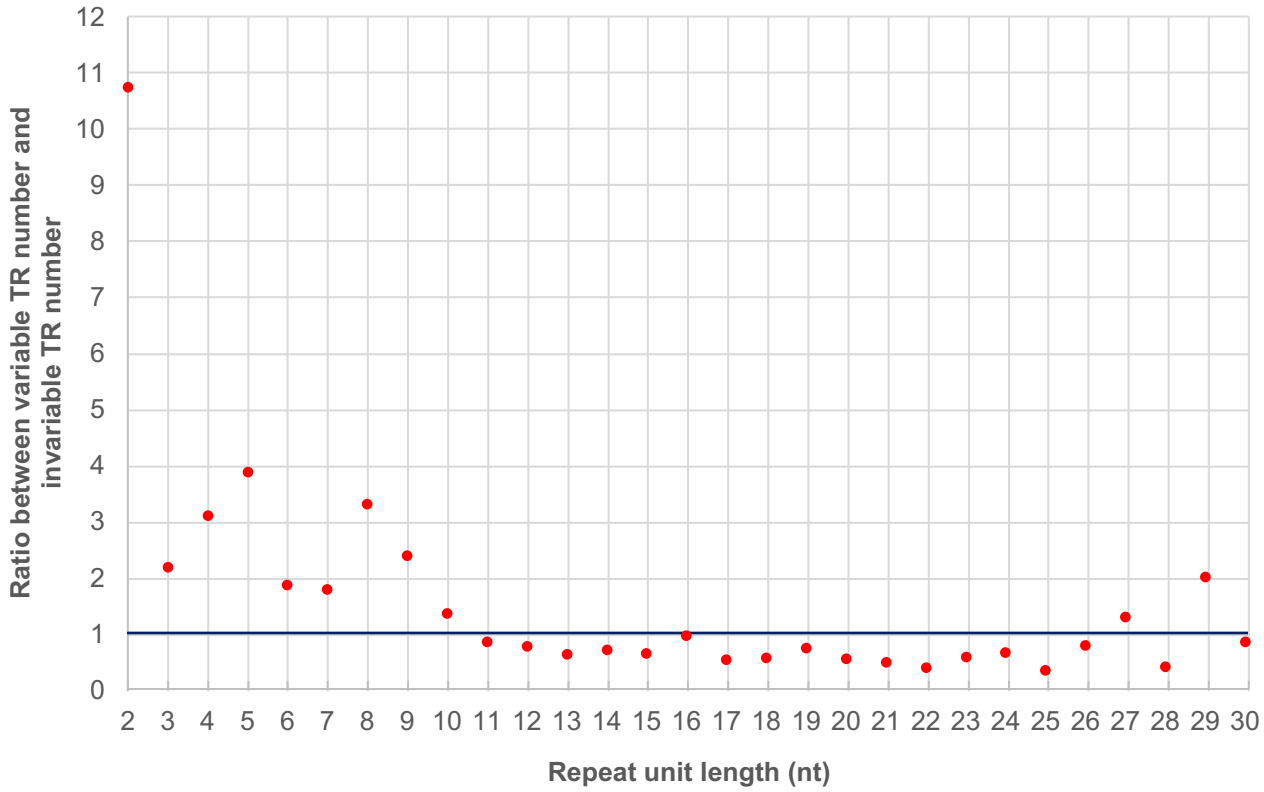

A
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
TR loci number
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Repeat unit length (nt)
B
900
800
700
600
500
400
300
200
100
0
TR loci number
AG
TC
TA
AT
CT
GA
ATT
ATA
TAT
AAT
Repeat unit sequence

A


Cumulative fraction of total TR loci
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
30
57
84
111
138
165
192
219
246
273
302
330
358
386
417
452
481
512
550
591
637
688
748
810
897
1024
1148
1316
1512
1709
1993
2395
2999
3888
TR locus length (bp)

B


Number of TR loci
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Within genic region
1-500
501-1000
1001-2000
>2000
Distance to nearest gene (bp)


Ratio between variable TR number and invariable TR number
0
1
2
3
4
5
6
7
8
9
10
11
12
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Repeat unit length (nt)

A

B.terrestris ……ACAATCGCAACAGCAGCAGCAGCAGCAGCAGCAACAGCAACAGCAGCAGCAG……

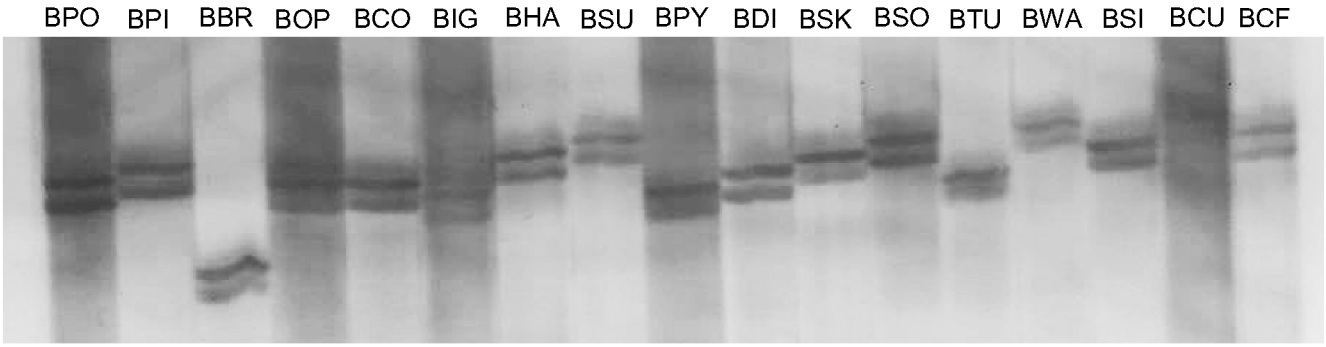
B.impatiens ……ACAATCACAA----------CAGCAGCAGCAACAGCAA---CAGCAGCAG……

B

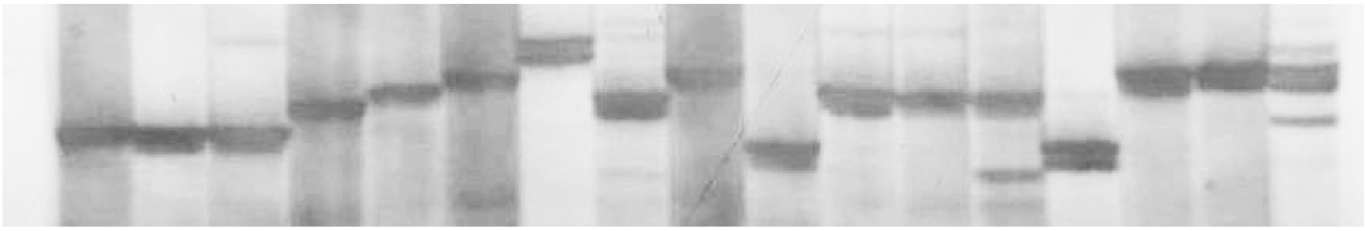
B.terrestris ……AGDERQIIRVTQQQSQQQQQQQQQQQQQQQQHQHQHQHQHQHQQPASSVGK……

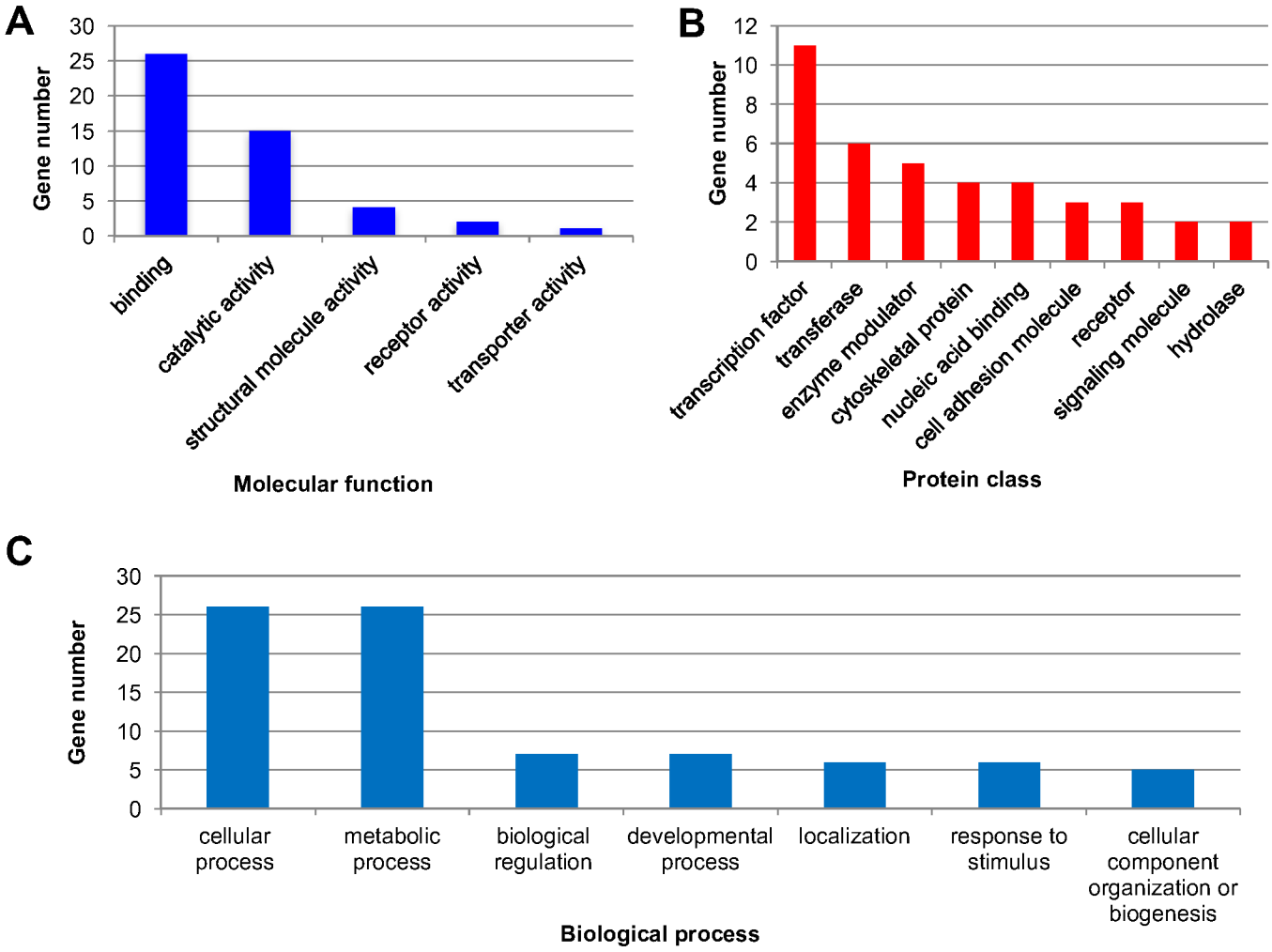
B.impatiens ……AGDERQIIRVTQQQS-----QQQQQQQQQQQQHQHQHQHQHQHQQPASSVGK……

A


BPO BPI BBR BOP BCO BIG BHA BSU BPY BDI BSK BSO BTU BWA BSI BCU BCF

B


BPO BPI BBR BOP BCO BIG BHA BSU BPY BDI BSK BSO BTU BWA BSI BCU BCF


A
Gene number
30
25
20
15
10
5
0
binding
catalytic activity
structural molecule activity
receptor activity
transporter activity
Molecular function
B
Gene number
12
10
8
6
4
2
0
transcription factor
transferase
enzyme modulator
cytoskeletal protein
nucleic acid binding
cell adhesion molecule
receptor
signaling molecule
hydrolase
Protein class
C
Gene number
30
25
20
15
10
5
0
cellular process
metabolic process
biological regulation
developmental process
localization
response to stimulus
cellular component organization or biogenesis
Biological process