1 **Corresponding author**

2 Seunggill Kim

3 Research Assistant Professor, Department of Plant Science and Plant Genomics and Breeding Institute,

4 College of Agriculture and Life Sciences, Seoul National University

5 Seoul, 151-921, Republic of Korea

6 Tel: +82-2-880-4946

7 E-mail: ksi2204@gmail.com

8

9 Doil Choi

10 Professor, Department of Plant Science and Plant Genomics and Breeding Institute,

11 College of Agriculture and Life Sciences, Seoul National University

12 Seoul, 151-921, Republic of Korea

13 Tel: +82-2-880-4568

14 Fax: +82-2-873-2056

15 E-mail: doil@snu.ac.kr

16

17

Abstract

Whole genome annotation errors that omit essential protein-coding genes hinder further research. We developed Target Gene Family Finder (TGFam-Finder), an optimal tool for structural annotation of protein-coding genes containing target domain(s) of interest in eukaryotic genomes. Large-scale re-annotation of 100 publicly available eukaryotic genomes led to the discovery of essential genes that were missed in previous annotations. An average of 117 (346%) and 148 (45%) additional FAR1 and NLR genes were newly identified in 50 plant genomes. Furthermore, 117 (47%) additional C2H2 zinc finger genes were detected in 50 animal genomes including human and mouse. Accuracy of the newly annotated genes was validated by RT-PCR and cDNA sequencing in human, mouse and rice. In the human genome, 26 newly annotated genes were identical with known functional genes. TGFam-Finder along with the new gene models provide an optimized platform for unbiased functional and comparative genomics and comprehensive evolutionary study in eukaryotes.

1 Introduction

2 The recent emergence of long read sequencing approaches such as single molecule real-time
3 sequencing and nanopore technology enables near-perfect assemblies of even enormous genomes like
4 that of the salamander¹ (genome size = 32 Gb, contig N50 = 300 kb) and the improvement of existing
5 assemblies for complex genomes²⁻⁴. Structural gene annotation of protein-coding sequences is a post-
6 assembly process that is essential for further research⁵. To date, hundreds of plant and animal genome
7 resources have been released in public databases and used for subsequent functional genomics
8 analyses, evolutionary analyses, and biotechnology applications. However, there are continuous
9 reports of annotation errors, including imperfect gene models and missing functional genes⁶⁻⁸. These
10 reports demonstrate that the accidental omission of essential genes that are correctly located in
11 assembled genomes but not annotated in the gene model can ultimately generate biases in downstream
12 studies.

13 Pre-existing gene models are continuously being improved using manual, computational, and
14 experimental analyses for model species such as human, mouse, and Arabidopsis. Since the Gene
15 Encyclopedia of DNA Elements (GENCODE) project was initiated, a total of 25 and 17 updates for
16 whole gene models of human and mouse have been accomplished, respectively (see URLs). In plant,
17 The Arabidopsis Information Resource (TAIR) is maintaining genomic resources for the model plant
18 *A. thaliana*, currently providing the 11th updated gene model, (Araport 11, see URLs). In addition,
19 groups conducting genome sequencing projects of major animal and plant species continuously
20 improve the quality of their gene models (see URLs). However, the majority of published gene
21 models remain inaccurate and likely incomplete, and have not been updated beyond the initial version.

22 In general, the potential function of a gene is predicted based on the identification of conserved
23 domains or motifs. Studies focusing on specific genes or families often begin working with annotated
24 gene models by identifying those genes of interest that contain the appropriate target domain(s) or
25 motifs^{8,9}. Performing new annotation and improving existing gene models require huge inputs of
26 human labor and computational resources. Therefore, researchers have designed novel approaches to
27 identify specific genes or gene families^{7, 9-11}. For example, Jupe et al. developed resistance gene

enrichment and sequencing (RenSeq), a high-throughput sequencing method for the selective capture and sequencing of nucleotide-binding and leucine-rich-repeat (NLR) genes without whole genome sequencing⁹. Although these methods enable detection of candidate regions containing target genes, further annotation to determine accurate gene structure in candidate regions remains a bottleneck.

Here, we present Target Gene Family Finder (TGFam-Finder), an optimal tool that allows automatic annotation of all protein-coding genes containing specific target domain(s) in assembled genomes. We evaluated TGFam-Finder through the massive re-annotation of far-red-impaired response 1 (FAR1) transcription factor and nucleotide-binding and leucine-rich-repeat (NLR) gene families in 50 plant genomes, as well as Cys2-His2 zinc finger (C2H2 zinc finger) and homeobox transcription factor gene families in 50 animal genomes. A large number of missing target genes in the pre-existing gene models were newly identified during re-annotation with TGFam-Finder. In particular, 26 newly annotated genes that were omitted in the existing human gene model had identical sequences with known functional genes. We validated the accuracy and expression of these newly annotated genes in human, mouse, and rice by performing RT-PCR and cDNA sequencing analyses. Our analyses demonstrate the efficiency of TGFam-Finder for users, including bench-based researchers given notably reduced annotation time using a desktop computer. TGFam-Finder, a domain search-based gene annotation tool could provide an optimal solutions for target-gene family annotation in functional, comparative, and evolutionary studies.

Results

Conceptual overview of TGFam-Finder

We designed TGFam-Finder as an unbiased annotation tool to identify any target-gene family of interest in assembled genomes. TGFam-Finder was developed for an audience including novice bioinformaticians, incorporating ease of use from installation to completion of structural annotation. To this, we provide additional tool packages enabling the automatic installation of prerequisite tools for further structural gene annotation using TGFam-Finder without any manual configuration (Supplementary Fig. 1 and Online Methods)

1 An automatic annotation process using TGFam-Finder consists of the following three steps : (1)
 2 genome-wide identification of target regions containing specific target-gene sequences of interest, (2)
 3 structural annotation of the target regions using available proteins, transcriptomes and *ab-initio*
 4 prediction, and (3) construction of the final gene model (Fig. 1). One of the distinct features of
 5 TGFam-Finder is the extraction of target regions containing sequences of target-gene family. To
 6 reduce annotation time and unnecessary computation, TGFam-Finder identifies all genomic regions
 7 containing domain(s) of the target genes using HMMER¹² from six-frame translated genome
 8 sequences. The target regions are determined after masking unnecessary sequences as 'X', except for
 9 the identified genomic regions and their flanking sequences (Fig. 1). Then, structural annotation of
 10 target regions is performed to generate the initial gene model through serial processes of protein
 11 mapping, transcriptome annotation, and *ab-initio* prediction (Fig. 1). On the basis of the evidence
 12 gathered in the previous steps, TGFam-Finder combines the initial gene models and determines the
 13 final gene model of target families (Fig.1).

14 TGFam-Finder improves upon pre-existing gene models through identification of missing essential
 15 genes, providing an unbiased model of target-gene families. To evaluate TGFam-Finder, we collected
 16 genomic data for 50 plants and 50 animals, including assemblies, annotated genes, and transcriptome
 17 data from public databases (Supplementary Table 1-2). In the plant genomes, we searched for the
 18 FAR1 transcription factor family that modulates phytochrome A signaling¹³ and NLR gene family that
 19 typically contains plant cytoplasmic immune receptor genes¹⁴ as target-gene families. In the animal
 20 genomes, we searched for the C2H2 zinc finger proteins, the largest transcription factor family, which
 21 is involved in functions such as sequence-specific DNA-binding and protein-protein interaction¹⁵, and
 22 the homeobox transcription factor family which primarily induces cellular differentiation by
 23 transcriptionally controlling co-regulated gene expression cascades¹⁶.

24

25 **FAR1 and NLR annotation in plant genomes**

26 We re-annotated FAR1 and NLR genes in 50 plant genomes using TGFam-Finder (Fig. 2,
 27 Supplementary Fig. 2-3 and Online Methods). Only 1.3% and 1.9% of the plant genomes (average

1 genome length 1,127 Mb) were determined as target regions for the re-annotation of FAR1 and NLR,
2 respectively (Supplementary Fig. 2 and Supplementary Table 3). On average, 34 FAR1 and 327 NLR
3 genes considering representative loci were identified in previously annotated gene models of the 50
4 plant genomes (Supplementary Table 3). In addition to these, we identified 117 (346%) and 148
5 (45%) new FAR1 and NLR genes using TGFam-Finder, respectively, indicating that only 23% (34 of
6 151) of FAR1s and 69% (327 of 475) of NLRs were annotated in previous studies (Fig. 2a,
7 Supplementary Fig. 3a, and Supplementary Table 3). Specifically, 104 (89%) and 108 genes (73%) of
8 the newly annotated FAR1 and NLR genes were located in genomic regions without any genes in the
9 existing models (Supplementary Fig. 4 and Supplementary Table 4). In addition, TGFam-Finder found
10 intact gene structures in regions containing previously annotated partial FAR1 and NLR genes that
11 have no start or stop codons (Supplementary Table 3). TGFam-Finder newly annotated 31 FAR1s
12 (26%) and 69 NLRs (43%) based on protein or transcriptome evidences in the new gene models
13 (Supplementary Table 5).

14 Compared to the number of genes in the existing gene models, we found a large number of new
15 FAR1s in 18 of the 50 genomes, including 12 dicot and 5 monocot plants, which was more than a 3-
16 fold increase, or more than 100 new FAR1s (Fig.2a). Specifically, 7 and 19 FAR1s were annotated in
17 the existing model of carrot and maize genomes, but TGFam-Finder identified 142 and 383 new
18 FAR1s in carrot and maize, respectively, a more than 20-fold increase relative to the number of
19 existing genes (Fig. 2a and Supplementary Table 3). In the wheat genome, we detected more than
20 3,000 new FAR1 genes. For the new gene model of NLRs, we also found a number of genes in 14
21 plant genomes (9 dicot species), which is a more than 1-fold increase or more than 200 genes
22 (Supplementary Fig. 3a and Supplementary Table 3). Although only three NLRs were annotated in
23 existing gene model of *Selaginella* genome, 56 (>18-fold increase) more NLRs were correctly
24 detected by TGFam-Finder in the same genome sequences (Supplementary Fig. 3a and
25 Supplementary Table 3). Furthermore, we identified over 1,000 more NLRs in genome sequences of
26 *Eucalyptus* and wheat, respectively. These results indicate that gene models for FAR1 and NLR were
27 greatly improved using TGFam-Finder compared with the extremely biased existing models omitting

1 many of those genes.

2 To study phylogenetic relationships of FAR1 and NLR in the 50 plant genomes, we performed gene
3 clustering and phylogenetic analyses of the new gene models containing the newly annotated genes
4 and the existing gene models (Fig. 2b, Supplementary Fig. 3b and Supplementary Table 6). The gene
5 clustering analyses among the new gene models of 50 plant genomes revealed that an average of 48
6 (41%) new FAR1 genes and 20 (13%) new NLR genes did not cluster with existing genes
7 (Supplementary Table 6). Overall, the number of FAR1s in new clusters was significantly larger in
8 monocots than in dicots (Fig. 2b). These results indicate that a large number of the newly annotated
9 genes, especially FAR1 in monocots, belonged to new subfamilies distinguished from the other
10 subfamilies of the existing gene models. Furthermore, phylogenetic analyses of FAR1 and NLR in
11 various plant genomes revealed that a large number of newly annotated genes were distinctly grouped
12 in specific lineages. This means that certain gene clades were never annotated in earlier studies (Fig.
13 2c, Supplementary Fig. 3c and 4).

14

15 **Annotation of C2H2 zinc finger and homeobox genes in animal genomes.**

16 We performed re-annotation of C2H2 zinc finger and homeobox gene families in 50 animal
17 genomes (Fig. 3 and Supplementary Fig. 5-6). A total of only 0.6% and 1.7% of the animal genomes
18 (average genome size 1,652Mb) were used as target regions for the re-annotation of C2H2 zinc finger
19 and homeobox genes, respectively (Supplementary Fig. 5 and Supplementary Table 7). An average of
20 117 (47%) and 22 (14%) additional C2H2 zinc finger and homeobox genes, respectively, were newly
21 annotated in the 50 animal genomes (Supplementary Table 7). Specifically, 75% (88 and 17) of the
22 new C2H2 zinc finger and homeobox genes did not overlap with any previously annotated genes,
23 indicating that majority of the newly annotated genes were located in new chromosomal regions
24 where no genes had been identified in earlier studies (Supplementary Table 8). We observed many
25 partial C2H2 zinc finger and homeobox genes in existing gene models of several invertebrate
26 genomes in contrast to the result of re-annotation of plant genomes. In that regions containing those
27 partial genes, TGFam-Finder identified new or their intact gene structures (Supplementary Table 7).

1 Our analyses revealed that over half of the new C2H2 zinc finger (54%) and homeobox genes (66%)
 2 were annotated based on protein or transcriptome evidence, indicating that a great number of new
 3 genes were identified by obvious evidence (Supplementary Table 9). We found a number of new
 4 C2H2 zinc finger (> 1-fold increase compared to the number of existing genes or 100 additional
 5 copies) in 24 of the 50 animal genomes including 17 genomes of mammalian species (Supplementary
 6 Table 7-8). In particular, we observed 133 (23%) and 227 (47%) newly identified C2H2 zinc finger
 7 genes in human and mouse genomes, respectively, and 77 (33%) new homeobox genes in the human
 8 genome (Supplementary Table 7-8). Together with the re-annotation of FAR1 and NLR in plant
 9 genomes, these results confirm that TGFam-Finder significantly improved the existing gene models of
 10 C2H2 zinc finger and homeobox genes in the animal genomes by identifying new gene structures
 11 absent in the existing gene models.

12 Phylogenetic comparisons of C2H2 zinc finger and homeobox gene families using the new gene
 13 models revealed that the many (36%) newly annotated C2H2 zinc finger genes were clustered as
 14 unique clades, suggesting that those were derived from newly identified lineages that were absent in
 15 existing models (Fig. 3b, Supplementary Fig. 6b and Supplementary Table 10). Moreover,
 16 phylogenetic trees of C2H2 zinc finger and homeobox genes in the human genome revealed
 17 significantly expanded lineages containing a small number of previously annotated genes and a large
 18 number of newly annotated genes (Fig. 3c and Supplementary Fig. 6c). We also found remarkably
 19 expanded or newly constructed lineages of C2H2 zinc finger and homeobox gene families in other
 20 animal genomes including mouse (Supplementary Fig. 7).

21

22 **Validation and expression of the new gene model**

23 We tested the accuracy of the new gene models by comparing the new gene models to UniProt¹⁷
 24 and NR¹⁸ databases (Supplementary Fig. 8 and Supplementary Table 11-12). We determined that
 25 sequences in NR and UniProt databases matched a percentage of the newly annotated genes ranging
 26 from, 56% (C2H2 zinc finger) to 61% (NLR), and 39% (FAR1) to 57% (NLR), respectively, covering
 27 more than 80% of the sequences in those databases (Supplementary Fig. 8). For all gene families, the

1 matched percentage of newly annotated genes was significantly lower than that of previously
2 annotated genes because of relatively low coverage of *ab-initio* genes (Supplementary Fig. 8).
3 However, we also found a large portion of previously annotated genes perfectly matched to the
4 sequences in NR and UniProt, suggesting that many of the previously annotated genes were already
5 present in those databases and these made higher coverage of the previously annotated genes.
6 (Supplementary Fig. 9).

7 Swiss-Prot of UniProt is a database consisted of experimentally validated gene sequences¹⁷. In total,
8 we detected that 25 and 1 newly annotated homeobox and C2H2 zinc finger genes in the human
9 genome had identical sequences in Swiss-Prot, respectively (Supplementary Table 13). This may
10 indicate that functional protein-coding genes were recorded in Swiss-Prot based on experimental
11 evidences but yet annotated their gene structures from the genome sequences. When we compared
12 previously annotated genes to SwissProt, the majority of perfectly matched genes to sequences of
13 SwissProt was enriched in specific genomes such as human, mouse and Arabidopsis, suggesting that
14 SwissProt primarily contained sequences of those species (Supplementary Fig. 10). Thus, we
15 performed identification of new genes having strong homology to sequences in the SwissProt (>95%
16 identity and 100% coverage) and found many newly annotated genes (65 and 34 of homeobox and
17 C2H2 zinc finger, respectively) in animal genomes (Supplementary Fig. 11).

18 To validate the expression of the newly annotated genes, we performed RT-PCR and sequencing
19 analyses using human, mouse and rice RNA samples (Fig. 4, Supplementary Fig. 12 and
20 Supplementary Table 14). We designed 47, 43, and 79 primer sets for newly annotated genes in
21 human, mouse and rice, respectively. Of these, 19, 15, and 50 genes (40, 35, and 63%) were
22 confirmed expression and accurately validated by comparison between sequenced PCR products and
23 annotated gene models. These results verify that the newly annotated genes are truly present and
24 expressed as mRNA in each genome (Fig. 4 and Supplementary Table 14).

25

26 **Annotation run-time using TGFam-Finder**

27 The full annotation of a large genome can take weeks to complete and requires enormous

1 computational resources¹⁹. To evaluate the performance of TGFam-Finder, we estimated the actual
 2 annotation run-time using TGFam-Finder for representative plant and animal genomes ranging from
 3 ~200 Mb to ~3 Gb using a desktop computer (32 Gb memory and 4 CPU). In total, it took 2 hours for
 4 NLRs of the *Selaginella* (~200 Mb) to 44 hours for C2H2 zinc finger of the human (~3 Gb) for
 5 annotation, indicating that completion of gene model construction could be finished within 2 days for
 6 the large genome such as human using a standard desktop computer (Table 1). Specifically, we
 7 verified that annotation of FAR1 in maize (~2 Gb) and C2H2 zinc finger in chicken (~1 Gb) required
 8 less than 12 hours. This indicates that users can efficiently annotate their target genes within half a day,
 9 with the exception of several huge genomes.

10

11 **Genomic features of the newly annotated genes in plant and animal genomes**

12 The genomic positions of the newly annotated genes could be classified into the following three
 13 categories: (1) non-overlapping, (2) overlapping with existing genes without target domain(s), and (3)
 14 overlapping with existing partial target genes (Supplementary Fig. 13). For non-overlapping genes, a
 15 large portion of those were overlapped with repetitive sequences (Supplementary Fig. 14a-b).
 16 Interestingly, the non-overlapping FAR1s in plant genomes were remarkably resided in the regions
 17 containing DNA-transposons. We observed that a significant number of NLRs in plant genomes co-
 18 localized with LTR-retrotransposons (Supplementary Fig. 14c). We also found that many C2H2 zinc
 19 finger were located in regions consisting of unclassified transposable elements in animal genomes.
 20 Considering previous reports describing annotation processes^{19, 20}, our results suggest that repeat
 21 masking before gene annotation could have a crucial impact in generating imperfect gene models. In
 22 the case of newly annotated genes overlapping with existing genes without target domain(s), we
 23 observed that the newly annotated gene families were primarily overlapped with hypothetical genes
 24 without known domain(s). This result suggests that several newly annotated genes were ignored in
 25 previous annotations due to the presence of uncharacterized genes in the same region (Supplementary
 26 Fig. 15).

27

1 Conclusion

2 The construction of accurate gene models after genome assembly is a critical step in genomic and
 3 functional analysis. Previous methods have been shown to be ineffective, hampered by imperfect
 4 methodologies, resources, and knowledge. Here, we described TGFam-Finder, a highly efficient tool
 5 to implement automatic structural annotation for target-gene families of interest. We evaluated and
 6 demonstrated the competitiveness of TGFam-Finder through the re-annotation of FAR1 and NLR
 7 gene families in 50 plants and C2H2 zinc finger and homeobox families in 50 animals. We only used
 8 publicly available resources, and identified large numbers of newly annotated genes that were omitted
 9 in the existing gene models (346, 45, 47, and 14% more FAR1, NLR, C2H2 zinc finger, and
 10 homeobox genes, respectively). The newly annotated genes were significantly supported by protein or
 11 transcriptome evidences. A total of 25 newly annotated homeobox and 1 C2H2 zinc finger genes in
 12 the human genome were identical to known functional gene sequences. We performed RT-PCR and
 13 sequencing analyses to confirm and validate the expression and accuracy of newly annotated genes in
 14 human, mouse, and rice genomes.

15 TGFam-Finder is easy to use and requires considerably less run-time and computing power than
 16 full annotation. The estimated annotation run-time of TGFam-Finder using a desktop PC was 2 hours
 17 for *Selaginella* (~200 Mb) NLRs and 44 hours for human (~3 Gb) C2H2 zinc finger. Compared to the
 18 long run-time of several weeks and intensive computational power required for full annotation, our
 19 results demonstrate that TGFam-Finder enables even novice users to obtain their target gene models
 20 within several days.

21 In summary, TGFam-Finder is designed to detect all protein-coding genes containing target
 22 domain(s) of interest and provide annotation evidence in eukaryotic genomes. TGFam-Finder enables
 23 users who study gene functions to determine their experimental priorities based on the annotation
 24 evidence and exact copy number of the genes of interest. Moreover, large-scale comparative studies of
 25 gene families will not be biased by missing genes, which are frequently encountered in previous
 26 annotations. Our approach provides an optimal solution for the identification and characterization of
 27 target-gene families, accelerating accurate functional, comparative and evolutionary analyses.

1 **Additional information**

2 **Acknowledgements**

3 This study was supported by the Basic Science Research Program through the National Research
4 Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1A6A3A04004014) to
5 S.K. and by a grant from the Agricultural Genome Center of the Next Generation Biogreen 21
6 Program of RDA (Project No. PJ013153) and by the National Research Foundation of Korea (NRF)
7 grant funded by the Korea government(MSIT) (No. 2018R1A5A1023599, SRC) to D.C. We
8 appreciate the assistance from the KOBIC Research Support Program. We also acknowledge to the
9 following researchers who performed iterative beta testing of TGFam-Finder: Ho-Sub Shin, Myung-
10 Shin Kim, and Jun-Ki Lee in Seoul National University; Namjin Koo in Korea Research Institute of
11 Bioscience and Biotechnology; and Eunyoung Seo in University of California, Berkeley.

12

13 **Author contributions**

14 S.K. and D.C. conceived the project, designed the content, and organized the manuscript. S.K., J.P.,
15 M.-S.K. and K.C. developed TGFam-Finder and annotation of the gene families. S.K., M.-S.K., Y.-
16 M.K., and N.K. collected resources of 50 plant and 50 animal genomes. J.-H.K., S.-H.K., K.-S.K.,
17 N.O., S.-K.Y., and K.-S.P. prepared the RNA samples. J.-H.K., K.-S.K., N.O., and K.-S.P.
18 implemented the RT-PCR analysis. S.K., M.-K.S., K.-T.K., J.J., H.K., Y.-Y.L., K.-H.S., H.C.M., and
19 Y.-H.L performed phylogenetic analyses and validation of the new gene models. S.K. and H.K.
20 designed and constructed the figures. S.K. and D.C. wrote the manuscript.

21

22 **URLs**

23 GENECODE (human and mouse), <https://www.gencodegenes.org/>
24 TAIR (Arabidopsis), <https://www.arabidopsis.org/>
25 GenBank (Zebrafish), https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Danio_rerio/105/
26 GenBank (Pig), https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Sus_scrofa/106/
27 iTAG (Tomato), <https://solgenomics.net/>
28 RepeatModeler, <http://www.repeatmasker.org/RepeatModeler/>
29 RepeatMasker, <http://www.repeatmasker.org/>

30

31

32

1 **Table 1. Annotation run-time of TGFam-Finder using a desktop PC**

Species	<i>S. moellendorffii</i>	<i>G. gallus</i>	<i>Z. mays</i>	<i>H. sapiens</i>	<i>H. sapiens</i>
Gene family	NLR	C2H2 zinc finger	FAR1	Homeobox	C2H2 zinc finger
Genome size (Mb)	213	1,046	2,067	3,209	3,209
Run-time (Total)	2h 16m	11h 25m	6h 26m	8h 49m	43h 54m
0.Sixframe translation	3m	13m	27m	44m	55m
1.Domain identification	1m	5m	9m	16m	6h 54m
2.Protein mapping	17m	4h 23m	1h 1m	1h 5m	16h 59m
3.Transcriptome annotation	1h 40m	6h 12m	3h 36m	6h 9m	17h 17m
4. <i>Ab-initio</i> prediction	13m	28m	1h 10m	31m	1h 44m
5.Final gene model generation	2m	4m	3m	4m	5m

2 Notes: h, hour; m, minutes.

3

Figure legends

Figure 1. Annotation process of TGFam-Finder. An automatic process for new identification of target-gene families using TGFam-Finder is depicted. The diagram shows serial processes starting from six-frame translation to integration of the gene model. The gray block of the diagram describes determination of target regions containing target domain(s) and their flanking sequences for further annotation. The blue and pink blocks indicate structural annotation using proteins and transcriptomes, and the *ab-initio* method, respectively. Names of representative tools²¹⁻²⁴ for structural annotation are given in the blue and pink blocks. Initial gene models are integrated from the structural annotation as depicted in the white block.

Figure 2. Re-annotation of FAR1 genes in 50 plant genomes. (a) Heat map indicates the number of existing FAR1 genes in the representative loci of 50 plant genomes. Bar graphs show the fold-increase of newly annotated genes compared with the number of existing genes (left) and the number of newly annotated genes (right). Colors in the bar represent the number of newly annotated genes from protein or transcriptome evidence (orange) and *ab-initio* model (navy blue). (b) Boxplot shows the distribution of the number of newly annotated genes in clusters without any existing genes. (c) Phylogenetic tree of newly annotated genes in the rice genome. Red and blue circles indicate newly annotated and previously annotated genes, respectively. Red branches indicate sublineages with multiple newly annotated genes including no copies or only a small number of previously annotated genes.

Figure 3. Newly annotated C2H2 zinc finger genes in 50 animal genomes. (a) Heat map and bar graphs represent the number of previously annotated genes in representative loci and the fold-increase and number of newly annotated genes, respectively. Orange bars indicate the number of genes generated from protein or transcriptome evidence. Navy blue bars represent *ab-initio* prediction. (b) The number of newly annotated genes in clusters without any previously identified genes is depicted. (c) Phylogenetic tree of new gene model in the human genome shows the phylogenetic relationship of those genes. The newly annotated and previously annotated genes are marked as red and blue circles, respectively.

Figure 4. Validation and expression of newly annotated genes. Bar graphs show the numbers of primer-designed (yellow), expressed (blue) and validated genes (red) in human, mouse, and rice genomes, respectively, identified by RT-PCR and sequencing analyses.

1 **Methods**

2 **Annotation overview of TGFam-Finder**

3 TGFam-Finder was developed to run in the Linux OS environment. For novices in bioinformatics-
 4 based analyses, we constructed an install package that allow auto-installation of prerequisite tools to
 5 run TGFam-Finder (Supplementary Fig. 1). Through the install package, prerequisite tools, including
 6 Bowtie2-2.3.1²⁵, HMMER-3.1b2¹², BLAST 2.6.0+²⁶, InterproScan-5.22-61.0²⁷, Exonerate-2.2.0²¹,
 7 Blat v35²⁸, Tophat-2.1.1 and Cufflinks-2.2.1²⁴, Augustus-3.2.3²², Scipio-1.4²⁹, and ClustalW-2.1³⁰ are
 8 provided for further annotation using TGFam-Finder. To run TGFam-Finder, users need to configure
 9 the location information of genomic resources and the prerequisite programs in ‘RESOURCE.config’
 10 and ‘PROGRAM_PATH.config’. Basically, ‘PROGRAM_PATH.config’ is automatically generated
 11 through the auto-installation process. Whereas, users should enter the location of the target genome,
 12 peptide sequences of target or allied species, and peptide sequences including target domains in
 13 various species as minimum resources. To classify specific proteins having target domain(s) of
 14 interest, TGFam-Finder requires the location of functional annotation information of target or allied
 15 species formatted as tsv and target domain ID(s) in ‘RESOURCE.config’. Moreover, users can input
 16 ‘EXTENSION_LENGTH’ to determine target regions for further annotation, and
 17 ‘MAX_INTRON_LENGTH’ for alignment processes using proteins. For extra configuration, users
 18 can also register the location of transcriptome, genomic position of genes and coding DNA sequences
 19 of target species in ‘RESOURCE.config’.

20 The annotation pipeline of TGFam-Finder consists of three steps. (1) Determination of target
 21 regions using ‘0.SixFrameTranslation.pl’ and ‘1.Domain_Identification.pl’, (2) Gene prediction in the
 22 target regions via ‘2.Auto_ProteinMapping.pl’, ‘3.Auto_ISGAP.pl’ and ‘4.Auto_Augustus.pl’, and (3)
 23 Generation of final gene model through ‘5. Generating_FinalGeneModel.pl’ (Fig. 1). To identify the
 24 position of target domains in an assembled genome, TGFam-Finder generates six-frame translated
 25 genome sequences. Then, a hidden Markov model (hmm) matrix is constructed through alignments
 26 among target domain(s) in protein sequences of target or allied species using ClustalW³⁰. After
 27 identification of genomic regions containing target domain(s) using HMMER¹², target regions
 28 including the target domain(s) and their flanking sequences are determined.

29 Structural annotation for the target regions is conducted via processes of protein mapping,
 30 transcriptome-based annotation and *ab initio* prediction. For efficient protein mapping, TGFam-Finder
 31 detects proteins with homology to target regions in the resource peptide sequences using BLAST+,
 32 and aligns between the proteins and matched target regions using Exonerate²¹. ‘Transcriptome-based
 33 annotation is implemented in the order of reference-guided transcriptome assembly using Tophat and
 34 Cufflinks²⁴, and annotation via ISGAP pipeline²³. For *ab-initio* gene prediction, the training set is
 35 constructed using the protein sequences of target or allied species having target domain(s), and the
 36 gene models generated from protein mapping and transcriptome-based annotation. Then, Augustus²²

generates the gene model based on the training set. Finally, the final gene model is generated by combining the initial gene models in order from the transcriptome-based annotation, protein mapping, and *ab initio* prediction.

Structural annotation of target-gene families in plant and animal genomes

FAR1 and NLR were re-annotated in 50 plant genomes, and C2H2 zinc finger and homeobox genes were re-annotated in 50 animal genomes using TGFam-Finder (Fig. 2-3, Supplementary Fig. 3 and 6, and Supplementary Table 3 and 7). We used assembled genomes and proteins described in Supplementary Table 1 and 2 as 'TARGET_GENOME' and 'PROTEINS_FOR_DOMAIN_IDENTIFICATION' in 'RESOURCE.config'. After performing functional annotation using InterproScan-5²⁷ for the proteins, generated tsv files for the proteins were used as 'TSV_FOR_DOMAIN_IDENTIFICATION'. PF03101 (FAR1), PF00931 (NLR), PF00096 (C2H2 zinc finger), and PF00046 (homeobox) were selected as 'TARGET_DOMAIN_ID' for classification of target-gene families. 'EXTENSION_LENGTH' and 'MAX_INTRON_LENGTH' were determined as 30kb. We extracted target genes in the existing gene models of plants or animals having the Pfam IDs in the tsv files, and then merged them to use as 'RESOURCE_PROTEIN'. Location of transcriptome, gff3, and coding DNA sequences of the plant and animal genomes were also recorded in 'RESOURCE.config' (Supplementary Table 1-2).

Phylogenetic analyses of the new gene models.

To perform phylogenetic comparison of new gene models for FAR1 and NLR in plant genomes, and C2H2 zinc finger and homeobox genes in animal genomes, we implemented gene clustering using OrthoMCL³¹ among the new models of each gene family (Supplementary Table 6 and 10). Then, the amino-acid sequences of target domains in each genome were aligned using ClustalW2³⁰, and the phylogenetic trees of each gene family in the specific plant and animal genomes were constructed with MEGA7³² using the neighbor-joining method (Supplementary Fig. 4 and 7).

Comparison of newly annotated genes and sequences in NR and UniProt

We compared new gene models of each gene family as query to sequences in NR and UniProt databases as subject using BLASTP. We counted the number of newly annotated genes that matched to NR and UniProt databases considering an e-value score of $1e^{-5}$ and more than 80% subject coverage (Supplementary Fig. 8 and Supplementary Table 11-12). To verify the number of newly annotated genes with strong homology to sequences in those databases, we used cut-off values greater than 98% identity and 100% subject coverage (Supplementary Fig. 9). We estimated the number of newly annotated genes with strong homology to functional genes in Swiss-Prot using cut-off values greater than 95% sequence similarity and 100% subject coverage (Supplementary Fig. 11).

1

2 **Validation of newly annotated genes using RT-PCR and sequencing analyses**

3 MCF-7 and MCF-10A cell lines (ATCC, Teddington, UK) with LEC and ADSC (LONZA, Basel,
4 Switzerland) were used for gene-expression verification and sequence validation for newly annotated
5 genes in the human genome. Mouse embryonic fibroblasts (MEFs) (Stemgent, MA, USA) and two
6 tissues (brain and spleen) from 6 to 8-week-old C57BL/6J mice (Koatech, Gyunggi-do, Korea) were
7 used for analyses of the mouse genome. C57BL/6J mice were maintained according to the policies
8 approved by CHA University. Seedlings of the rice (*Oryza sativa*) cultivar Nakdong were used for
9 validation of rice FAR1 and NLR genes. Total RNA was extracted from rice seedlings, cultured
10 human cells, and each mouse tissue using TRIzol[®] reagent (Invitrogen, USA). The total RNA was
11 treated with DNase I and reverse-transcribed with Oligo (dT) primers and Superscript II (Invitrogen),
12 according to the manufacturer's instructions. Subsequently, RT-PCR was conducted to analyze gene
13 expression using the designed gene primers for each family. We used *GapdH* of human and mouse
14 and *Actin* of rice as controls. PCR products were purified with AccuPrep[®] PCR purification kit
15 (Bioneer, Korea) or gel elution. Finally, the PCR products were sequenced using ABI3730XL
16 (Applied Biosystems) as described in Supplementary Table 14.

17

18 **Identification of genes overlapping with repeat sequences**

19 To annotate genomic sequences containing non-overlapping genes along with previously annotated
20 genes described in Supplementary Figure 13, we performed repeat annotation for the genomic regions
21 using RepeatModeler and RepeatMasker (see URL). *De novo* repeat libraries of each plant and animal
22 genome were constructed using RepeatModeler, and then RepeatMasker was used to repeat masking
23 on the repeat libraries. We considered that if specific repeat sequences covered more than 50% of a
24 non-overlapping gene, the gene resided in the genomic region containing the specific repeat sequence
25 (Supplementary Fig. 14).

26

27 **Computational resources used to run TGFam-Finder**

28 To estimate annotation run-time of TGFam-Finder, we performed annotation using a desktop
29 computer (Intel Core i7-4770 CPU @ 3.40GHz, 8 processors, and 32Gb memory) for NLR families in
30 *S. moellendorffii*, C2H2 zinc finger gene families in *G. gallus*, FAR1 genes in *Z. mays*, and C2H2 zinc
31 finger and homeobox genes in *H. sapiens* (Table 1). For efficient test, we randomly extracted and used
32 ~10Gb of whole transcriptome data of *G. gallus*, *Z. mays*, and *H. sapiens* as described in
33 Supplementary Table 1 and 2. For the annotation of FAR1 and NLR genes in 50 plant genomes, and
34 C2H2 zinc finger and homeobox genes in 50 animal genomes, we used two of our computer servers
35 with the following specifications (1) Intel Xeon CPU E5-2697 v2 @ 2.70GHz, 48 processors, and
36 264Gb memory, and (2) Intel Xeon CPU E5-4650 v2 @ 2.40GHz, 80 processors, and 512Gb memory.

1 The re-annotation of each gene family was finished within one week using those servers.

2

3 **Data availability**

4 The newly generated gene models including peptide and coding DNA sequences with gff3 and tsv are
5 accessible at: <http://TGFam-Finder.snu.ac.kr>.

6

7 **Code availability**

8 The TGFam-Finder program package including auto-installation and annotation scripts with sample
9 data is accessible at: <https://github.com/tgfam-finder> and <http://TGFam-Finder.snu.ac.kr>.

10

References

1. Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50-55 (2018).
2. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338-345 (2018).
3. Jiao, Y.P. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524-527 (2017).
4. Schmidt, M.H.W. et al. *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* **29**, 2336-2348 (2017).
5. Jones, S.J.M. Prediction of genomic functional elements. *Annu. Rev. Genom. Hum. G.* **7**, 315-338 (2006).
6. Johnston, J.J. et al. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am. J. Hum. Genet.* **86**, 743-748 (2010).
7. Lagarde, J. et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731-1740 (2017).
8. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8**, 995-1005 (2007).
9. Jupe, F. et al. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* **76**, 530-544 (2013).
10. Li, Y.I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151-158 (2018).
11. Teer, J.K. & Mullikin, J.C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145-R151 (2010).
12. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
13. Hudson, M., Ringli, C., Boylan, M.T. & Quail, P.H. The FAR1 locus encodes a novel nuclear protein specific to phytochrome A signaling. *Genes Dev.* **13**, 2017-2027 (1999).
14. Jacob, F., Vernaldi, S. & Maekawa, T. Evolution and conservation of plant NLR functions. *Front. Immunol.* **4**, 297 (2013).
15. Pabo, C.O., Peisach, E. & Grant, R.A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.* **70**, 313-340 (2001).
16. Cillo, C., Cantile, M., Faiella, A. & Boncinelli, E. Homeobox genes in normal and malignant cells. *J. cell. Physiol.* **188**, 161-169 (2001).
17. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
18. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61-D65 (2007).
19. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**,

1 329-342 (2012).

2 20. Bennetzen, J.L. & Park, M. Distinguishing friends, foes, and freeloaders in giant genomes. *Curr. Opin.*

3 *Genet. Dev.* **49**, 49-55 (2018).

4 21. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison.

5 *BMC Bioinformatics* **6**, 31 (2005).

6 22. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and

7 genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **7**, S11-18 (2006).

8 23. Kim, S. et al. Integrative structural annotation of de novo RNA-Seq provides an accurate reference

9 gene set of the enormous genome of the onion (*Allium cepa* L.). *DNA Res.* **22**, 19-27 (2015).

10 24. Ghosh, S. & Chan, C.K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods Mol. Biol.*

11 **1374**, 339-361 (2016).

12 25. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359

13 (2012).

14 26. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

15 27. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-

16 1240 (2014).

17 28. Kent, W.J. BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

18 29. Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. Scipio: Using protein sequences to

19 determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC*

20 *Bioinformatics* **9**, 278(2008).

21 30. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

22 31. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic

23 genomes. *Genome Res.* **13**, 2178-2189 (2003).

24 32. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0

25 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870-1874 (2016).

26

27