Title

Genome organization of the vg1 and nodal3 gene clusters in the allotetraploid frog Xenopus laevis

Journal

Developmental Biology, 426(2)

ISSN

0012-1606

Authors

Suzuki, Atsushi

Uno, Yoshinobu

Takahashi, Shuji

et al.

Publication Date

2017-06-01

DOI

10.1016/j.ydbio.2016.04.014

Peer reviewed

Genome organization of the *vg1* and *nodal3* gene clusters in the allotetraploid frog *Xenopus laevis*

Atsushi Suzuki[1*], Yoshinobu Uno[2], Shuji Takahashi[1], Jane Grimwood[3], Jeremy Schmutz[3], Shuji Mawaribuchi[4], Hitoshi Yoshida[1], Michihiko Ito[4], Yoichi Matsuda[2], Daniel Rokhsar[5, 6, 7], and Masanori Taira[8]

[1]Institute for Amphibian Biology, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima 739-8526, Japan
[2]Department of Applied Molecular Biosciences, Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
[3]HudsonAlpha Institute of Biotechnology, 601 Genome Way, Huntsville, AL 35806, USA
[4]Kitasato Institute for Life Sciences, Kitasato University, 5-9-1 Shirokane Minato-ku Tokyo 108-8641 Japan
[5]University of California, Berkeley, Department of Molecular and Cell Biology and Center for Integrative Genomics, Life Sciences Addition #3200, Berkeley California 94720-3200, USA
[6]US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA;
[7]Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan
[8]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Abbreviated Title: *vg1* and *nodal3* gene clusters in *X. laevis*

*Corresponding Author
Contact info. for corresponding author:
Tel: 81-82-424-7103, FAX: 81-82-424-0739
E-mail: asuzuki@hiroshima-u.ac.jp
Postal address: Institute for Amphibian Biology, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan

Abstract

Extracellular factors belonging to the TGF-β family play pivotal roles in the formation and patterning of germ layers during early *Xenopus* embryogenesis. Here, we show that the *vg1* (*gdf1*) and *nodal3* genes of *X. laevis* are present in gene clusters on chromosomes 1L and 3L, respectively and that both gene clusters have been completely lost from the syntenic S chromosome regions. The presence of gene clusters and chromosome-specific gene loss were confirmed by cDNA FISH analyses. Sequence and expression analyses revealed that paralogous genes in the *vg1* and *nodal3* clusters on the L chromosomes were also altered compared to their *X. tropicalis* orthologs. *X. laevis vg1* and *nodal3* paralogs have potentially become pseudogenes or sub-functionalized genes and were expressed at different levels. These findings provide insights into function and molecular evolution of TGF-β family genes in response to allotetraploidization.

Keywords: *vg1*, *gdf1*, *nodal3*, *Xenopus*, TGF-β, genomic structure

Highlights:

(1) *vg1* and *nodal3* form gene clusters on L chromosomes of allotetraploid *Xenopus laevis*.

(2) Both *vg1* and *nodal3* gene clusters were deleted from the *X. laevis* S chromosome.

(3) Some *X. laevis vg1* and *nodal3* paralogs are pseudogenes or sub-functionalized genes.

(4) Previously reported *vg1* and *nodal3* paralogs belong to gene clusters identified here.

Introduction

The transforming growth factor-beta (TGF-β) family of polypeptide growth factors consists of more than 40 members and plays multifunctional roles in development, cell differentiation, organogenesis and tissue homeostasis (Derynck and Miyazono, 2008; Wu and Hill, 2009). In *Xenopus*, several ligands of this family regulate induction and patterning of germ layer formation during embryogenesis and a wealth of information is available on their roles in signal transduction and transcriptional regulation pathways that produce diverse cellular responses.

Transcripts of the *vg1* (*gdf1*) gene are vegetally localized in mature *Xenopus* oocytes and encode a TGF-β family ligand that has a potent mesoderm inducing activity when processed into a mature dimeric form of the C-terminal subunits (Rebagliati et al., 1985; Weeks and Melton, 1987; Dale et al., 1993; Thomsen and Melton, 1993; Kessler and Melton, 1995). Although the vegetal localization of maternal *vg1* mRNAs suggests that Vg1 is a prime candidate as an endogenous mesoderm inducer emanating from the vegetal pole of *Xenopus* blastulae, the initially identified *vg1* cDNA was not processed efficiently and did not show mesoderm inducing activity (Dale et al., 1989; Tannahill and Melton, 1989). Therefore, the precise role of endogenous Vg1 in mesoderm induction has long been controversial. Disruption of Vg1 has been found to reduce mesoendodermal marker gene expression and to cause abnormalities in gastrulation, formation of mesodermal and neural tissues, and axis patterning (Joseph and Melton, 1998). Moreover, a Vg1 allele (S20), which has a serine residue at position 20 compared to proline in the first identified sequence, is processed and can rescue defects observed by Vg1 disruption (Birsoy et al., 2006). While this compelling evidence is helping to resolve long-standing questions in *Xenopus* embryology, the genomic organization of the two allelic versions of *vg1* genes remains uncertain. The genes have extremely high identity in their nucleotide sequences but it is not clear whether this is the result of tandem gene duplication or the allotetraploid nature of *Xenopus laevis*. In the diploid frog *X. tropicalis* and most of other vertebrates, *vg1* is present as a single copy gene and has axis specification functions (Hellsten et al., 2010; Seleiro et al., 1996; Shah et al., 1997). This suggests that the *X. laevis vg1* locus may have undergone unique changes in genomic organization and expression during tetraploidization; this possibility can now be investigated through analysis of the whole genome sequence.

nodal3, one of six *nodal*-related genes in *X. laevis*, was originally found in an expression screen to identify genes that could rescue dorsal axis formation in ventralized embryos (Smith et al., 1995). The expression of *nodal3* is localized in the Spemann's organizer region of early gastrulae; Nodal3 functions as a neural inducer by

antagonizing bone morphogenetic proteins (Hansen et al., 1997; Haramoto et al., 2004). Nodal3 also regulates the expression of the *xbra* (*t*) gene and is involved in convergent extension movements during gastrulation and neurulation (Yokota et al., 2003; Morita et al., 2013). cDNA cloning analyses in *X. tropicalis* and *X. laevis* have shown that both have more than one copy of *nodal3,* and that in *X. tropicalis* the *nodal3* genes are tandemly duplicated in the genome (Hellsten et al., 2010).

Following the completion of the draft genome sequence of *X. laevis*, we decided to investigate and compare the genomic organizations of the *vg1* and *nodal3* genes in *X. laevis* and *X. tropicalis*. In addition to shotgun sequencing and assembly, we performed BAC and fosmid sequencing of *vg1* and *nodal3* loci, and then analyzed genomic structures and gene expression in detail. We found that *vg1* is specifically amplified on the *X. laevis* L chromosome and that the amplified copies contain two types of *vg1* allele that have different biological activities, resolving the origin of the two *vg1* copies found in *X. laevis*. In contrast to the tandem duplications on the L chromosome, the *vg1* gene was found to be absent from the corresponding region of the S chromosome although there was a highly degenerate pseudogene at this position. Similarly to *X. tropicalis*, the *X. laevis nodal3* genes form a tandemly duplicated gene cluster on the L chromosome and include pseudogenes within this cluster. Interestingly, *X. laevis nodal3* was found to be lost from the S chromosome, similarly to *X. laevis vg1*. These results establish the genomic organization of multi-copied TGF-β ligands that play essential roles in embryogenesis. They also provide insight into the genomic structures of the allelic *vg1* copies that encode proteins with different characteristics and which have been under study for some time.

Materials and Methods

Sequencing and gene expression of vg1 and nodal3 loci

Xenopus laevis J strain shotgun sequencing, preparation/sequencing of the BAC and fosmid library clones including XLFIC-005N07, and RNA-seq analysis have been described in detail by Session et al. (submitted). Three BAC clones (XLB1-052I23, XLB1-308L17 and XLB1-126A10) covering the *vg1* locus were additionally sequenced using long read Pacific BioSciences technology. Briefly BAC clones were prepped using Qiagen maxiprep columns and the resulting DNA pooled into pools of 12 clones. One Pacific Biosciences library was then constructed, shearing the DNA with a g-tube to 20kb average size (Covaris). Sequence was generated using one SMRT cell on a 4 hour run. Data was assembled with RS_HGAP_Assembly.3 protocol implemented in SMRT Portal version 2.3.0 using default parameters. Consensus was integrated with existing Illumina fragment reads assembled with phrap 0.990319 and the combined assemblies computational finished with Consed, version 26.33 (Gordon, D., Abaijian, C., and Green, P. (1988). Consed: A graphical tool for sequence finishing. Genome Research 8: 195-202.)

The sequences were deposited in GenBank (AP014675.1 for XLFIC-005N07; KU558979 for a contig made from XLB1-052I23, XLB1-308L17 and XLB1-126A10).

Gene nomenclature

In order to distinguish the multi-copy genes revealed by the *X. laevis* genome analysis, a new form of gene nomenclature is used throughout this manuscript (described by Session et al., submitted). Briefly, paralogs were indicated by "- (hyphen)" (*e.g. gene_name-1* and *-2*). *vg1* and *nodal3* are tandemly multiplicated (expanded) genes having "one-to-multi" or "multi-to-multi" relationship with their *X. tropicalis* orthologs. These paralogous genes were indicated by "-e" (e stands for expansion; *e.g. gene_name-e1* and *-e2*).

Cell culture, chromosome preparation, and FISH

One adult female of the J strain of *X. laevis* was used for cell culture. After pithing, heart, lung, and kidney tissues were collected, minced, and cultured, as previously described (Uno et al., 2013). All experimental procedures involving animals conformed to the guidelines established by the Animal Care Committee of Nagoya University, Nagoya, Japan.

For replication-banded chromosome preparation, 5-bromodeoxyuridine (BrdU) was added during late S phase. BrdU (25 μg/ml) was added to cell cultures during the log

phase, after which cell culture was continued for 6 h, including 1 h of colcemid treatment (0.17 µg/ml), before harvesting. Chromosome slides were made following a standard air-drying method. After staining with Hoechst 33258 (1 µg/ml) for 5 min, the slides were heated to 65°C for 3 min on a hotplate and then exposed to ultraviolet light for an additional 5–6 min at 65°C (Matsuda and Chapman, 1995).

FISH mapping of cDNA fragments was performed as previously described (Matsuda and Chapman, 1995; Uno et al., 2013). A plasmid containing either *vg1* or *nodal3* cDNA was labeled with biotin-16-dUTP (Roche Diagnostics, Basel, Switzerland) using a nick translation kit (Roche Diagnostics) following a standard protocol. After hybridization with biotin-labeled cDNA fragments, the hybridized probes were incubated with a goat anti-biotin antibody (Vector Laboratories, Burlingame, CA, USA), and stained with Alexa Fluor® 488 rabbit anti-goat Immunoglobulin G (H+L) conjugate (Thermo Fisher Scientific/Molecular Probes, Inc., Eugene, OR, USA). The preparations were counterstained with 0.75 µg/ml propidium iodide (PI). Identification of each chromosome and determination of the subchromosomal localization of the hybridization signals were performed using the Hoechst 33258-banded ideogram constructed in our previous study (Matsuda et al., 2015).

Results and Discussion

Genomic organization and expression of vg1 genes

We initially analyzed *X. laevis* assembly v7.1 and identified the complete sequence of the *vg1* exon1 on the end of scaffold 70461 (3.34 Mbp). Synteny upstream of exon 1 is in accord with that of other vertebrates including *X. tropicalis* and human (not shown). We selected five of the BAC clones that map to the region surrounding exon 1 and sequenced these fully using Pacific Biosciences sequencer (see Materials and Methods). Although the repetitive nature of this locus hampered sequence assembly, we successfully assembled three BAC clones, XLB1-052I23, XLB1-308L17 and XLB1-126A10, and produced a single contig overlapping the end of scaffold 70461 (Fig. 1A). We found that this contig contained three copies of the full-length *vg1* gene indicating that the *vg1* locus formed a tandemly duplicated gene cluster (*vg1-e1.L*, *-e2.L* and *-e3.L*) in *X. laevis*. In *X. laevis* assembly v9.1, this gene cluster and a homeologous region were present on chromosomes 1L and 1S, respectively. In addition, synteny around the *vg1* gene cluster was well conserved with *X. tropicalis* except for *gdf3.L* gene which was missing from the shotgun sequences and BAC clones analyzed in this study.

vg1 is a single copy gene in other vertebrates including *X. tropicalis*, human and mouse (Hellsten et al., 2010; Rankin et al., 2000). Therefore, the presence of at least three tandem copies of *vg1* in *X. laevis* strongly suggested the occurrence of gene amplification of this TGF-β ligand gene in the course of allotetraploidization. Moreover, we also found that in contrast to the gene amplification on the L chromosome, the *vg1* copy was largely deleted from the corresponding region of the S chromosome and was a pseudogene (*vg1p.S*) (Fig. 1A). As shown in Fig. 2, *vg1p.S* only contained a mutated fragment of exon 1 and lacked the exon-intron junction sequence at the 3' region. To confirm the deletion of *vg1* on the S chromosome, we performed a cDNA FISH analysis using chromosome samples prepared from an *X. laevis* cell culture. As shown in Fig. 1B, fluorescent signals were observed on the 1L chromosomes, but not on the 1S chromosomes, indicating that *vg1* was deleted on the latter. We also analyzed the sizes of the intergenic regions around the *vg1* loci in both *X. tropicalis* and *X. laevis* (Fig. 1A) and found that the distance between the *cers1* and *gdf3* genes on the S chromosome (~9.6 kb) was shorter than that in *X. tropicalis* (~39 kb). These results clearly demonstrate the deletion of sequences from the *vg1* gene in the S chromosome.

There are two *vg1* copies in *X. laevis*; one has a proline residue at position 20, the other has a serine residue. The origin of these allelic forms is uncertain (Birsoy et al., 2005, 2006). Sequence analysis of the three tandem copies on the L chromosome showed that two were serine alleles (*vg1-e1.L* and *vg1-e3.L*) and one was a proline allele

(*vg1-e2.L*) (Fig. 1A). The S chromosome copy, *vg1p.S* contained a serine at position 20 similarly to *vg1* genes of *X. tropicalis* and other vertebrates. Overexpression of *X. laevis* Vg1(S20), but not Vg1(P20), results in a mesoderm-inducing activity and the rescue of defects caused by inhibition of maternal Vg1 function (Birsoy et al., 2006). Therefore, the *vg1(S20)* gene is likely to be the ancestral version of vertebrates and we postulate that *X. laevis vg1(P20)* was produced after a tandem duplication of the *vg1(S20)* allele.

We next examined the expression of *vg1* genes by RNA-seq analysis and found that the three *vg1* copies were maternally expressed during oogenesis and early embryogenesis in a similar pattern to that reported previously (Fig. 1C; Rebagliati et al., 1985). Similarly, Birsoy et al. (2006) reported that equal numbers of EST sequences were present for *vg1(S20)* and *vg1(P20)* genes. A detailed comparison of the nucleotide and amino acid sequences of the three *vg1* copies with those of published *vg1* cDNAs deposited in the NCBI GenBank database showed that *vg1-e2.L* and *-e3.L* corresponded to BC090232.1 (IMAGE cDNA clone: 6870372) and AY838794.1 (Birsoy et al., 2006), respectively (Fig. 3). Interestingly, *vg1-e1.L* was grouped into the Vg1(S20) type similarly to *vg1-e3.L* and AY838794.1 but was distinct from them, suggesting that *vg1-e1.L* is a novel form of the *vg1(S20)* gene. We confirmed the expression of *vg1-e1.L* not only by RNA-seq but also by searching EST records in the NCBI GenBank database (not shown). The first reported *vg1* cDNA (GenBank M18055.1; Rebagliati et al. 1985) was a Vg1(P20) type and differed slightly from *vg1-e2.L* and BC090232.1. Therefore, this transcript may originate from a genomic region that was not cloned and sequenced in this study. The precise location of this gene awaits future studies; however, it is possible that the *vg1* gene cluster on the L chromosome contains more than three copies of *vg1* genes.

vg1 was one of the first identified maternally localized transcripts during early *Xenopus* embryogenesis (Rebagliati et al., 1985). The localization of *vg1* transcripts depends on the 3'-untranslated sequence which contains consensus localization motifs (Mowry and Melton, 1992; Gautreau et al, 1997; Deshler et al., 1997). Therefore, we compared the corresponding regions among the *vg1* paralogs identified here on the *X. laevis* L chromosome. As shown in Fig. 4, the 340 bp 3'-UTR regions containing the localization elements were well conserved among *vg1-e1.L*, *-e2.L* and *-e3.L*, except for a few nucleotide changes that located outside of the consensus motifs.

Genomic organization and expression of nodal3 genes

X. laevis nodal3 genes are located on chromosome 3L, adjacent to two other *nodal* genes *nodal1* and *nodal2*, and the synteny of this region is similar to that of *X. tropicalis*

(Fig. 5A). On the *X. laevis* L chromosome, four complete copies of *nodal3* were identified (*nodal3-e1.L* ~ *-e4.L*) compared to five copies (*nodal3-e1* ~ *-e5*) on the *X. tropicalis* genome assembly. It was not possible to construct a gene model for *nodal3-e1.L* from the *X. laevis* v9.1 assembly as this was built from shotgun sequences and there is a short gap region (Ns) in the *nodal3-e1.L* region. However, we successfully filled the gap by sequencing a fosmid clone (XLFIC-005N07) and obtained a complete sequence of *nodal3-e1.L*. Comparison with cDNA records deposited in the NCBI GenBank database revealed that *nodal3-e1.L* and *nodal3-e2.L* corresponded to *nodal3.1* (NM_001085790.1; Smith et al., 1995) and *nodal3.2* (NM_001085596.1; Ecochard et al., 1995), respectively. *nodal3-e3.L* and *nodal3-e4.L* were similar to *nodal3-e2.L*, but distinct from *nodal3.2* (not shown).

We found a *nodal3* pseudogene candidate (*nodal3p1.L*) located between *nodal3-e1.L* and *nodal3-e2.L*; this sequence has truncations at both N- and C-terminal portions of its translated protein product (Fig. 6A). While *nodal3p1.L* was similar to *nodal3-e1.L*, the initiation methionine corresponded to the 5th methionine of the Nodal3-e1.L protein and there was no significant homology in the upstream nucleotide sequence of the methionine between *nodal3p1.L* and *nodal3-e1.L* genes (not shown). Importantly, the lack of an N-terminal portion resulted in deletion of the signal peptide, suggesting that Nodal3p1.L was unlikely to be secreted and functional. The truncations at both ends of *nodal3p1.L* in the *X. laevis* assembly were confirmed by sequencing of the fosmid clone. In addition to *nodal3p1.L*, we also identified two highly degenerate pseudogenes (*nodal3p2.L* and *nodal3p3.L*) between *nodal3-e4.L* and *nodal2.L* genes (Fig. 5A). These genes had accumulated mutations in exons 1 and 2 as well as in the splice junctions and completely lacked exon 3 (not shown).

The mature peptide region of the Nodal3 protein shares five of the seven conserved cysteine residues found in TGF-β family proteins (Smith et al., 1995; Ecochard et al., 1995). All copies of *nodal3* on the L chromosome, including *nodal3p1.L*, contained these cysteines residues. They also contained the proteolytic processing site (RRXXR; Fig. 6A). Interestingly, phylogenetic analysis of Nodal3 proteins revealed that the clustered *nodal3* paralogs were grouped exclusively in *X. laevis* and *X. tropicalis*, rather than showing a one-to-one relationship between these species as is the case for *nodal1* and *nodal2* genes (Fig. 6B). This suggested that clustered *nodal3* genes might have evolved in concert and resulted in a more homogeneous sequence within the cluster, as has been proposed for rRNA and histone genes (Nei and Rooney, 2005).

Similarly to the *vg1* gene cluster, *nodal3* genes were deleted from the S chromosome; there was no pseudogene between the flanking genes *tacc1.S* and *nodal2.S* (Fig. 5A). In

agreement with this result, the intergenic distance between *tacc1.S* and *nodal2.S* was significantly shorter (~28 kb) than in *X. tropicalis* (~69 kb). We also confirmed the gene loss by cDNA FISH (Fig. 5B).

We analyzed expression of four complete copies of *nodal3* (*nodal3-e1~-e4.L*) and the truncated pseudogene (*nodal3p1.L*) using RNA-seq. At least two of the copies (*nodal3-e1.L* and *nodal3-e2.L*) were expressed (Fig. 5C), and this result was consistent with the fact that the *nodal3* copies correspond to cDNAs deposited in the NCBI GenBank (see above). Moreover, a search of the EST database also identified sequences corresponding to *nodal3-e1.L* and *nodal3-e2.L*, confirming that these *nodal3* genes were expressed (not shown). We have been unable to find EST clones corresponding to *nodal3* genes (*nodal3p1.L*, *nodal3-e3.L* and *nodal3-e4.L),* which is unsurprising in view of the low expression levels found here for these genes.

Conclusions

The deletion of tandemly clustered TGF-β family genes from the S chromosome has been observed for *vg1*, *nodal3* and *nodal5* genes (Suzuki et al. this issue; Takahashi et al. this issue). Although the biological significance of this observation is not understood, this gene loss might be the result of complex functional interactions between L and S gene products after tetraploidization and selective pressure to avoid having too many genes for key signaling molecules in the allotetraploids. The pseudogenization of *nodal3p1.L* is consistent with this interpretation as it indicates the possibility of reducing *nodal3* copy numbers in *X. laevis* compared to *X. tropicalis*. However, it is equally possible that amplification of *vg1* genes on the L chromosome might be a compensatory response to the loss of *vg1* on the S chromosome. Alternatively, the amplification could be the source of novel functionality as has been proposed for the role of gene duplication (Ohno, 1970; Holland, 1999). While the function of these paralogous genes must be carefully analyzed by specific gene knockouts, the appearance of the vg1(P20) variant might indicate such a functional differentiation. Nevertheless, the gene loss/amplification observed for these TGF-β family proteins indicates their functional importance in development as well as in homeostasis. In addition, the differences in copy number changes between *vg1* and *nodal3* might reflect differences in the biological functions of the proteins (Vg1 activates, while Nodal3 inhibits in TGF-β receptor/Smad-mediated cellular signaling). Advances in genome editing technology will enable us to address the biological significance of these clustered gene sequences and to investigate the evolutionary history of copy number changes. The genomic structures of the *vg1* and *nodal3* loci have not been addressed previously despite the extensive

functional analysis of these genes. Therefore, our results open a new window in studies of function and molecular evolution of the TGF-β family genes that play essential roles in animal development.

Acknowledgements

This work was supported by Japan Society for the Promotion of Science KAKENHI Grant Numbers, 25460245 (A. S. and K. T.-S.), 23113004 (Y. M.), and 25251026 (M. T.) 25460245. We would like to thank National Bio-Resource Project of Ministry of Education, Culture, Sports, Science and Technology, Japan, and National Sci- ence Center for Basic Research and Development of Hiroshima University. We would also like to thank Drs. Richard Harland, Taejoon Kwon, Adam Session, Gert Veenstra, Simon van Heerin- gen, Mariko Kondo and Hajime Ogino for helpful discussions. The generation of BAC sequence data was supported by National In- stitutes of Health, USA, Grant number 1R01HD080708. RNA-seq data was obtained by a collaborative effort of S. T., Atsushi Toyoda, Yutaka Suzuki, Taejoon Kwon, Naoto Ueno, A. S., and M. T.. We would like to thank them and all the members of the Xenopus laevis genome consortium for helpful discussions. We would also like to thank Takafumi Ikeda for the annotation of derrière gene.

References

Birsoy, B., Berg, L., Williams, P.H., Smith, J.C., Wylie, C.C., Christian, J.L., Heasman, J., 2005. XPACE4 is a localized pro-protein convertase required for mesoderm induction and the cleavage of specific TGFβ proteins in *Xenopus* development. Development 132, 591–602.

Birsoy, B., Kofron, M., Schaible, K., Wylie, C., Heasman, J., 2006. Vg 1 is an essential signaling molecule in *Xenopus* development. Development 133, 15–20.

Dale, L., Matthews, G., Colman, A., 1993. Secretion and mesoderm-inducing activity of the TGF-β-related domain of *Xenopus* Vg1. EMBO J. 12, 4471–4480.

Dale, L., Matthews, G., Tabe, L., Colman, A., 1989. Developmental expression of the protein product of Vg1, a localized maternal mRNA in the frog *Xenopus* laevis. EMBO J. 8, 1057–1065.

Derynck, R., Miyazono, K., 2008. TGF-ß and the TGF-ß Family, in: The TGF-ß Family. Cold Spring harbor laboratory press, pp. 29–43.

Deshler, J.O., Highett, M.I., Schnapp, B.J., 1997. Localization of *Xenopus* Vg1 mRNA by Vera Protein and the Endoplasmic Reticulum. Science 276, 1128–1131.

Ecochard, V., Cayrol, C., Foulquier, F., Zaraisky, A., Duprat, A.M., 1995. A novel TGF-β-like gene, *fugacin*, specifically expressed in the Spemann organizer of *Xenopus*. Dev. Biol. 172, 699–703.

Gautreau, D., Cote, C.A., Mowry, K.L., 1997. Two copies of a subelement from the Vg1 RNA localization sequence are sufficient to direct vegetal localization in *Xenopus* oocytes. Development 124, 5013–5020.

Hansen, C.S., Marion, C.D., Steele, K., George, S., Smith, W.C., 1997. Direct neural induction and selective inhibition of mesoderm and epidermis inducers by Xnr3. Development 124, 483–492.

Haramoto, Y., Tanegashima, K., Onuma, Y., Takahashi, S., Sekizaki, H., Asashima, M., 2004. *Xenopus* tropicalis nodal-related gene 3 regulates BMP signaling: an essential role for the pro-region. Dev. Biol. 265, 155–168.

Hellsten, U., Harland, R., Gilchrist, M., Hendrix, D., Jurka, J., V, K., Ovcharenko, I., Putnam, N., Shu, S., Taher, L., Blitz, I., Blumberg, B., Dichmann, D., Dubchak, I., Amaya, E., Detter, J., Fletcher, R., Gerhard, D., Goodstein, D., Graves, T., Grigoriev, I., Grimwood, J., Kawashima, T., Lindquist, E., Lucas, S., Mead, P., Mitros, T., Ogino, H., Ohta, Y., Poliakov, A., Pollet, N., Robert, J., Salamov, A., Sater, A., Schmutz, J., Terry, A., Vize, P., Warren, W., Wells, D., Wils, A., Wilson, R., Zimmerman, L., Zorn, A., Grainger, R., Grammer, T., Khokha, M., Richardson, P., Rokhsar, D., 2010. The genome of the western clawed frog *Xenopus* tropicalis.

Science 328, 633–636.

Holland, P.W.H., 1999. Gene duplication: Past, present and future. Semin. Cell Dev. Biol. 10, 541–547.

Joseph, E.M., Melton, D.A., 1998. Mutant Vg1 ligands disrupt endoderm and mesoderm formation in *Xenopus* embryos. Development 125, 2677–2685.

Kessler, D.S., Melton, D.A., 1995. Induction of dorsal mesoderm by soluble, mature Vg1 protein. Development 121, 2155–2164.

Matsuda, Y., Chapman, V.M., 1995. Application of fluorescencein situ hybridization in genome analysis of the mouse. Electrophoresis 16, 261–272.

Matsuda, Y., Uno, Y., Kondo, M., Gilchrist, M.J., Zorn, A.M., Rokhsar, D.S., Schmid, M., Taira, M., 2015. A new nomenclature of *Xenopus laevis* chromosomes based on the phylogenetic relationship to *Silurana*/*Xenopus tropicalis*. Cytogenet. Genome Res. 8601, 1–5.

Morita, M., Yamashita, S., Matsukawa, S., Haramoto, Y., Takahashi, S., Asashima, M., Michiue, T., 2013. Xnr3 affects brain patterning via cell migration in the neural-epidermal tissue boundary during early *Xenopus* embryogenesis. Int. J. Dev. Biol. 57, 779–786.

Mowry, K.L., Melton, D.A., 1992. Vegetal messenger RNA localization directed by a 340-nt RNA sequence element in *Xenopus* oocytes. Science 255, 991–994.

Nei, M., Rooney, A.P., 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39, 121–152.

Ohno, S., 1970. Evolution by gene duplication. Springer Science & Business Media, LLC.

Rankin, C.T., Bunton, T., Lawler, A.M., Lee, S., 2000. Regulation of left-right patterning in mice by growth/differentiation factor-1. Nature 24, 262–265.

Rebagliati, M.R., Weeks, D.L., Harvey, F.L.R., Melton, D.A., 1985. Identification and cloning of localized maternal RNAs from *Xenopus* eggs. Cell 42, 769–777.

Seleiro, E.A.P., Connolly, D.J., Cooke, J., 1996. Early developmental expression and experimental axis determination by the chicken Vg1 gene. Curr. Biol. 6, 1476–1486.

Shah, S.B., 1997. Misexpression of chick Vg1 in the marginal zone induces primitive streak formation. Development 124, 5127–5138.

Smith, W.C., McKendry, R., Ribisi, S., Harland, R.M., 1995. A nodal-related gene defines a physical and functional domain within the Spemann organizer. Cell 82, 37–46.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular

evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Tannahill, D., Melton, D.A., 1989. Localized synthesis of the Vg1 protein during early *Xenopus* development. Development 106, 775–785.

Thomsen, G.H., Melton, D.A., 1993. Processed Vgl protein is an axial mesoderm inducer in *Xenopus*. Cell 74, 433–441.

Uno, Y., Nishida, C., Takagi, C., Ueno, N., Matsuda, Y., 2013. Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. Heredity 111, 430–436.

Weeks, D.L., Melton, D.A., 1987. A maternal mRNA localized to the vegetal hemisphere in *Xenopus* eggs codes for a growth factor related to TGF-β. Cell 51, 861–867.

Wu, M.Y., Hill, C.S., 2009. TGF-ß Superfamily signaling in embryonic development and homeostasis. Dev. Cell 16, 329–343.

Yokota, C., Kofron, M., Zuck, M., Houston, D.W., Isaacs, H., Asashima, M., Wylie, C.C., Heasman, J., 2003. A novel role for a nodal-related protein; Xnr3 regulates convergent extension movements via the FGF receptor. Development 130, 2199–2212.

Figure legends

Figure 1. Structure and expression of the *vg1* gene cluster.

(A) Schematic diagrams of the *vg1* gene cluster in *X. tropicalis* (XTR) and *X. laevis* (XLA). Arrowheads indicate the direction and positions of genes. Distances in the schematic are proportional to the nucleotide lengths of the genes, except for *vg1p.S* which contains a partial exon 1. Magenta and blue arrowheads indicate complete *vg1* genes and the *vg1* pseudogene, respectively. Orange and blue bars indicate the region covered by BAC full-sequences and scaffold70461 which were originally constructed in the v7.1 assembly. (B) cDNA FISH analysis of the *vg1* gene. Hybridization signals (upper) are shown along with Hoechst staining images (lower) of chromosomes 1L and 1S. The position of the *vg1* gene alongside the chromosome banding pattern of XLA1L. (C) RNA-seq expression profile of *vg1* genes during oogenesis and early development. oo12, oo34 and oo56 indicate oocyte stage 1/2, 3/4 and 5/6, respectively.

Figure 2. Structure of the vg1 pseudogene *vg1p.S*.

(A) Schematic diagram of the *vg1p.S* structure in comparison with that of *vg1-e3.L*. *vg1p.S* lacks exon 2 and the splice donor/acceptor sites. The exon 1 region of *vg1p.S* shows 92% identity with that of *vg1-e3.L*. (B) Comparison of the exon 1 nucleotide sequences of *vg1-e3.L* and *vg1p.S*. *vg1p.S* contains a stop codon near the end of exon 1 that resulted from an insertion and a deletion (blue boxes). The intron 1 of *vg1-e3.L* is underlined. *vg1p.S* also contains a serine residue at position 20 of the deduced protein product.

Figure 3. Alignment of vg1 protein sequences.

Deduced protein sequences of vg1-e1.L~-e3.L, vg1_AY838794.1 (Birsoy et al., 2006), vg1_BC090232.1 (IMAGE cDNA clone) and vg1_M18055.1 (Rebagliati et al., 1985) were aligned by the Clustal W method. Serine or proline residues at positon 20 of the deduced protein products are indicated. The proteolytic processing site is shown by a double underline. Asterisks indicate seven conserved cysteine residues in the mature domain which are shared with other TGF-β family proteins.

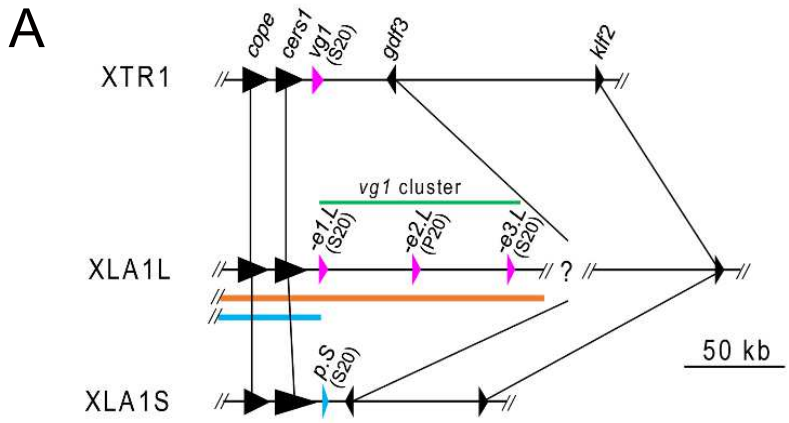
Figure 4. Alignment of the 3'-untranslated regions of *vg1* genes.

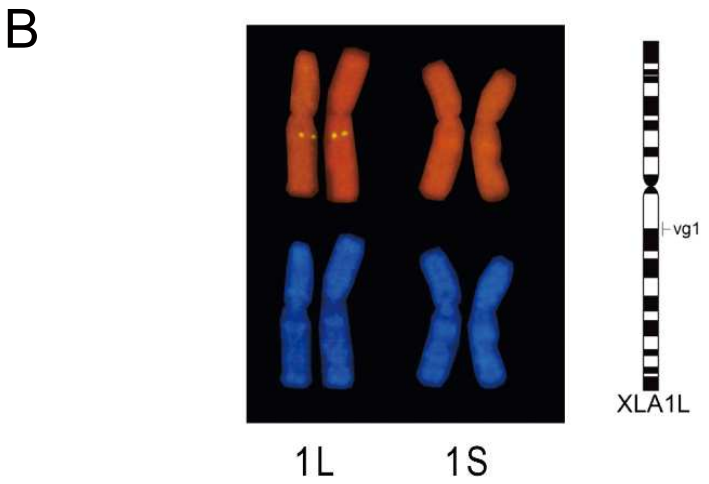
The 340 bp sequence of the 3'-untranslated regions of *vg1* genes were aligned by the Clustal W method. Boxes indicate consensus repeat motifs important for vegetal localization of *vg1* transcripts. Three E1 motifs were originally identified by Gautreau et al (1997). In addition, other motifs (E2~E4) involved in the process of *vg1* RNA

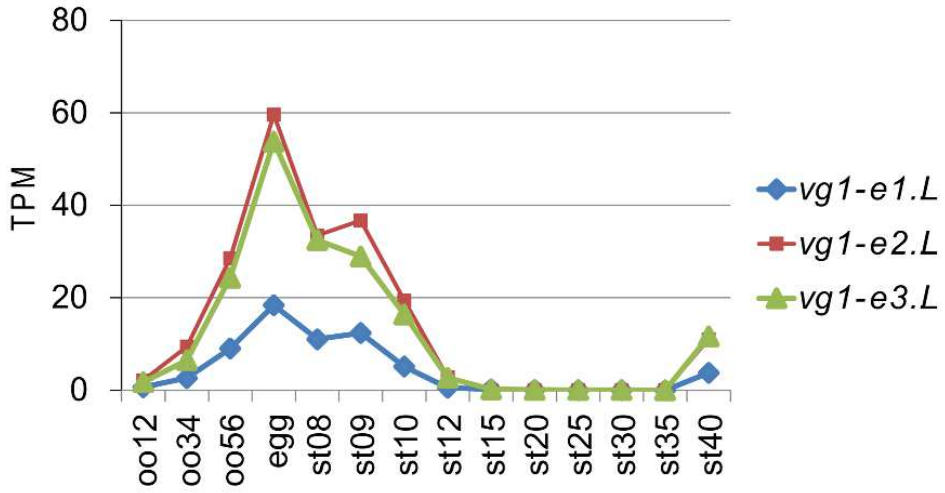
localization are also shown (Deshler et al., 1997).

Figure 5. Structure and expression of the *nodal3* gene cluster.
(A) Schematic diagrams of the *nodal3* gene clusters of *X. tropicalis* (XTR) and *X. laevis* (XLA). Arrowheads indicate the direction and positions of genes. Distances in the schematic are proportional to the nucleotide lengths of genes. Magenta and blue arrowheads indicate complete *nodal3* genes and *nodal3* pseudogenes, respectively. An orange bar indicates the region covered by the fosmid full-sequence. (B) cDNA FISH analysis of the *nodal3* gene. Hybridization signals (upper) are shown in comparison with Hoechst staining images (lower) of chromosome 3L and 3S. The position of *nodal3* is shown alongside the chromosome banding pattern of XLA3L. (C) RNA-seq expression profile of *vg1* genes during oogenesis and early development. oo12, oo34 and oo56 indicate oocyte stage 1/2, 3/4 and 5/6, respectively.

Figure 6. Alignment of Nodal3 protein sequences.
(A) *X. laevis* Nodal3 protein sequences were aligned by the Clustal W method. The box indicates the deduced signal peptide region. Note that Nodal3p1.L lacks this signal peptide and is unlikely to be functional. The proteolytic processing site is shown by a double underline. Asterisks indicate five conserved cysteine residues in the mature domain which are shared with other TGF-β family proteins. (B) Phylogenetic tree of *X. laevis* and *X. tropicalis* Nodal3 proteins. Nodal1 and Nodal2 proteins were also included in the analysis. Evolutionary history was inferred using the Neighbor-Joining method. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013).

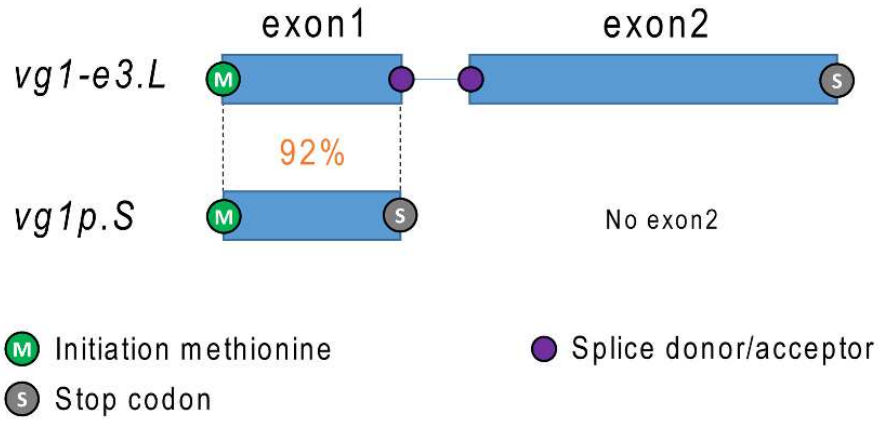






Suzuki et al. Fig. 1

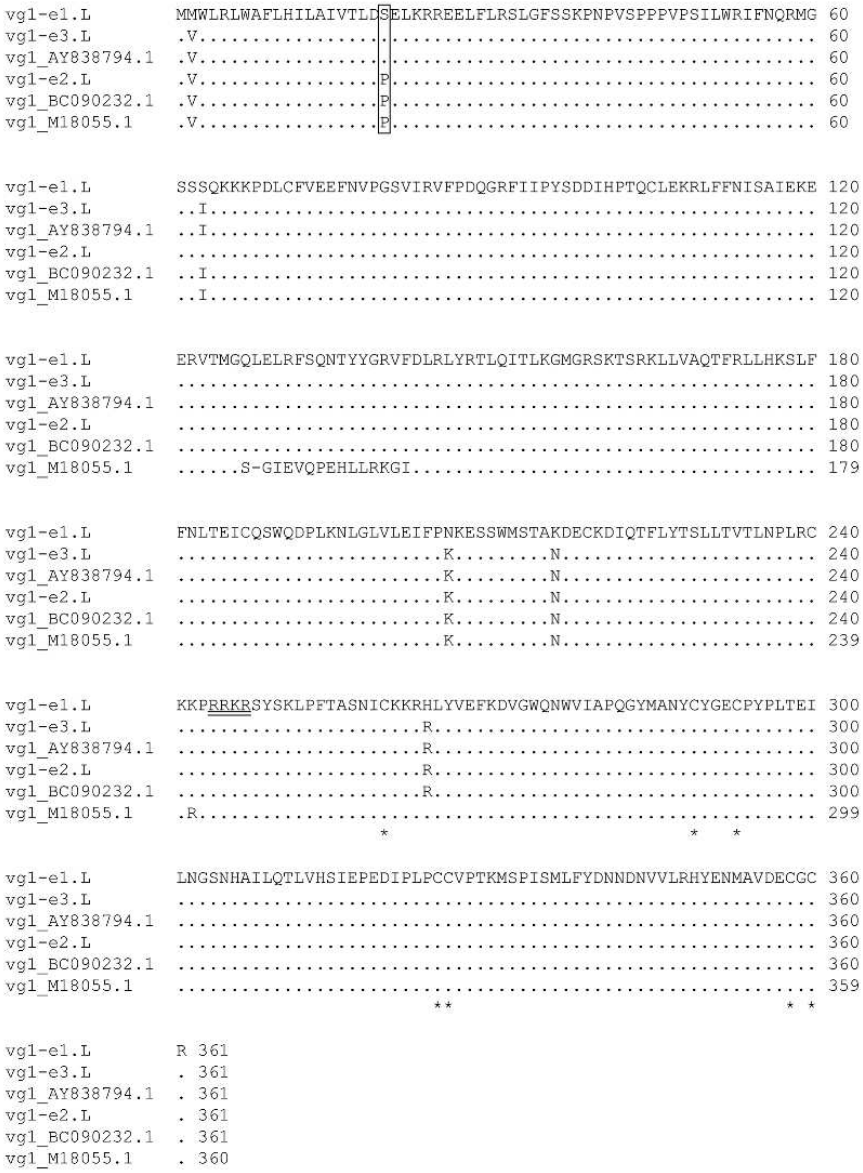
A



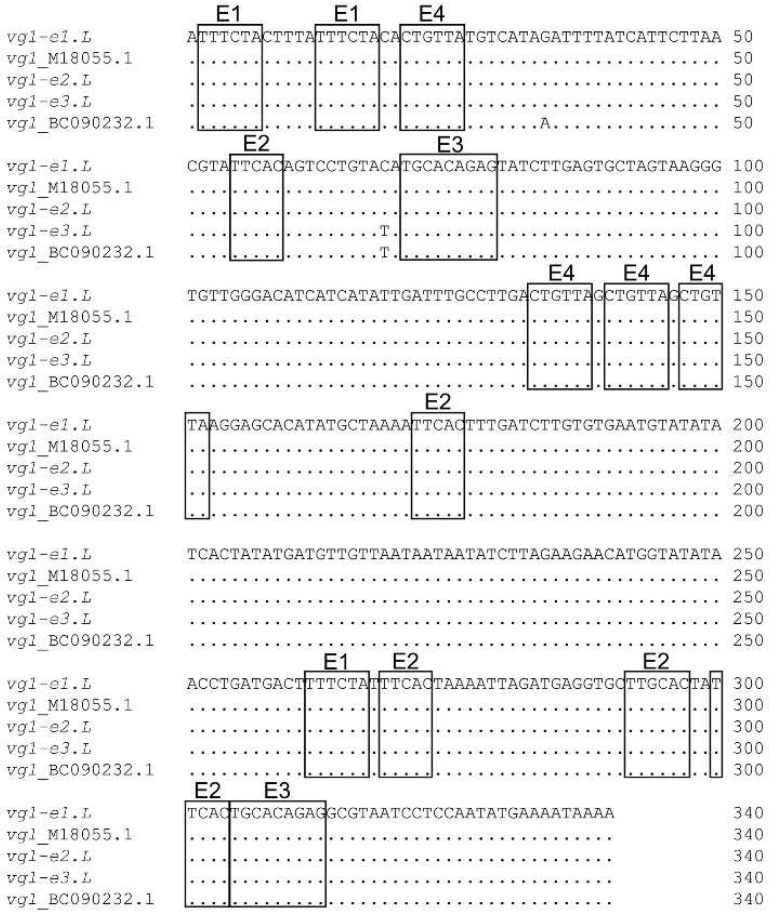
B

```
           Met                                                                                                  Ser20
vg1-e3.L   ATGGTGTGGCTGAGACTGTGGGCTTTCCTGCATATACTTGCTATTGTAACTTTGGATTCA  60
vg1p.S     ...................A......A......C.....................T  60

vg1-e3.L   GAGCTCAAGAGGCGGGAAGAGCTGTTTCTGAGGAGTTTGGGTTTCTCTTCCAAGCCTAAC 120
vg1p.S     ........A...T.......T............C.......................... 120

vg1-e3.L   CCGGTATCTCC---TCCTCCTGTCCCTTCTATACTTTGGAGGATATTCAACCAAAGGATG 177
vg1p.S     ..T........TAT.....................C.......................... 180

vg1-e3.L   GGGAGCTCCATTCAGAAAAAGAAACCAGACTTGTGCTTTGTGGAGGAATTTAATGTTCCT 237
vg1p.S     A..GCT..TG...................C..........AA................A 240

vg1-e3.L   GGTAGTGTTATCAGAGTGTTCCCTGATCAAGGTGAGCT 275
vg1p.S     ..A.....--..............TAA.GTAA.CTGA 276
                                          Stop
```

Suzuki et al. Fig. 2

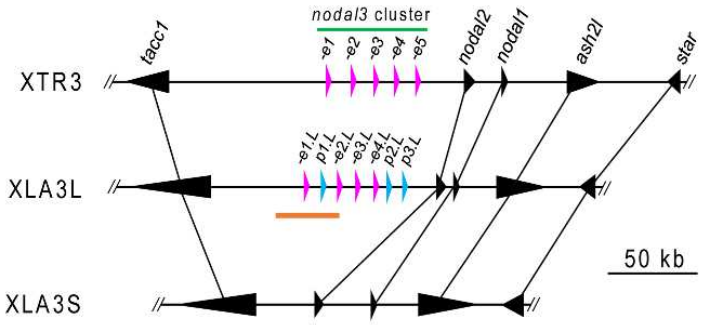


Suzuki et al. Fig. 3

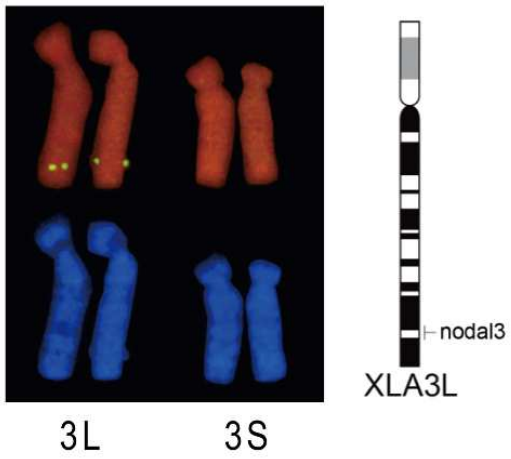


Suzuki et al. Fig. 4

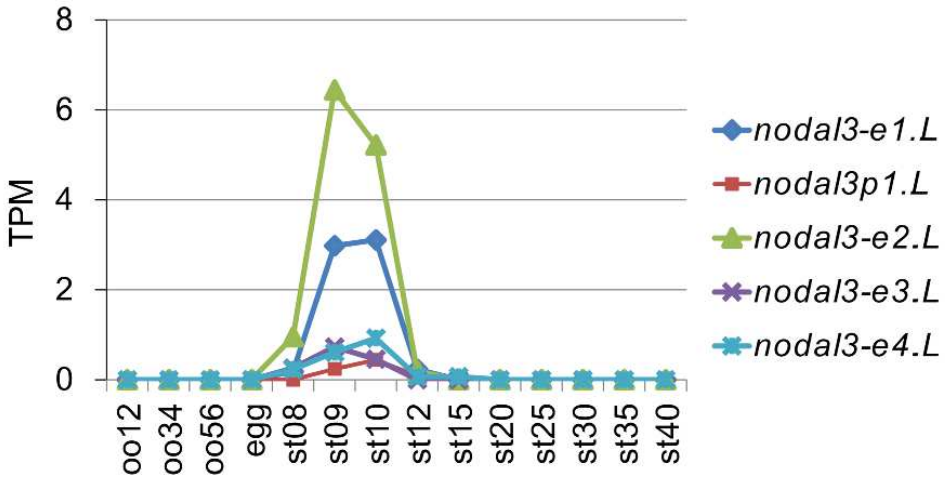
A



B



C

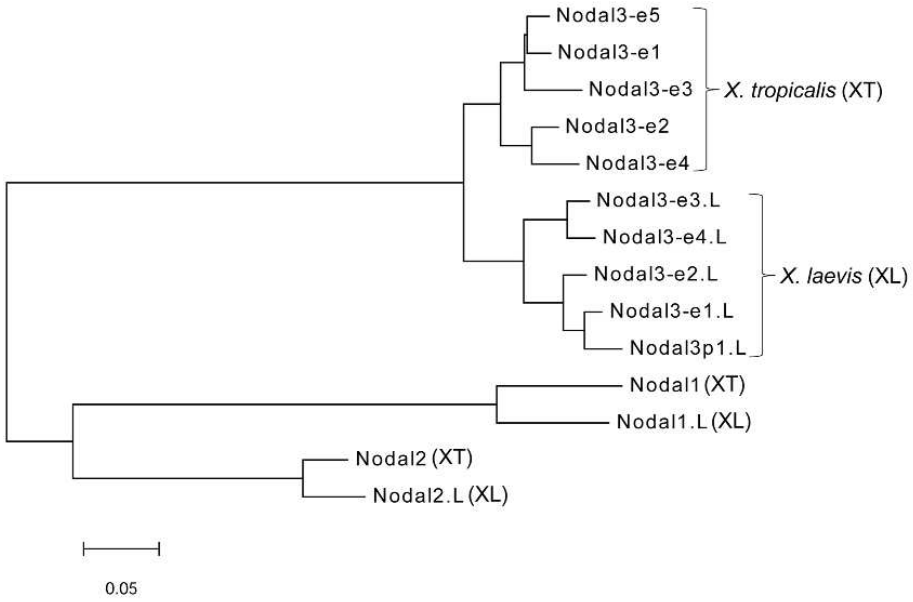


Suzuki et al. Fig. 5

A

```
Nodal3-e1.L     MAFLNLFFCLVFSSPLMAMPPILQGRMSISPDSILKDTSTDIGAREFQGRKFPNFMMQLY 60
Nodal3p1.L      ------------------------------------------------------------  5
Nodal3-e2.L     ..L.................V...K.M...........D...D.......Q...... 60
Nodal3-e3.L     ..L.................V...K.M...........D.V.D.......Q...... 60
Nodal3-e4.L     ....................V...K.M...........D...D.......Q...... 60

Nodal3-e1.L     QNIIRGRDKDLSNLEHPTLQESDTVQSFIAKSYTTVGNRWTLFFDMSSISRSNELKLAEL 120
Nodal3p1.L      ........N.............I.............C.....................  65
Nodal3-e2.L     ............................T........H..................... 120
Nodal3-e3.L     ................................I.K..H.............I....... 120
Nodal3-e4.L     ................................K..H.............I....V.. 120

Nodal3-e1.L     RICLPSFRKSHSVTVDIYHTNDGKEKLFMGSFKTKPSSALDSDCKVFNLTILLQNFLTRG 180
Nodal3p1.L      ......................S............L....................... 125
Nodal3-e2.L     ....................................A...A.................. 180
Nodal3-e3.L     .......G.F.T...E....K............S.L.....A...........E...I.. 180
Nodal3-e4.L     .......G.F....IE....K.D..........S.L.....A...........E...I.. 180

Nodal3-e1.L     KRLIKDEYIQAKGLHLKDLEKSATEKDTENVDTMKQDQYHVSDFAAERIMLVVFAKEQSH 240
Nodal3p1.L      ..................P.........................V.............. 185
Nodal3-e2.L     ........................G......L........................... 240
Nodal3-e3.L     ...............R.....VI.IRA.....L...RH.....T..K.I.......E.. 240
Nodal3-e4.L     ...............R.....VI.ISA.....L...KH.....T..K.I.......E.. 240

Nodal3-e1.L     AKPDPPSLGQKLFPSKYGIDDNANKVNGFRRLRRNKKEKTQIHVSTVPPKPIEEIKPKCK 300
Nodal3p1.L      .........K.................................................. 245
Nodal3-e2.L     ..S......K.....................................G............. 300
Nodal3-e3.L     .........K.................................................. 300
Nodal3-e4.L     ...ET....K.................................................. 300
                                                                          *

Nodal3-e1.L     KVDMFVDFQKIGWGSWIIYPKAYNAYRCESTCAVPQNETENATNHSYIKSLLPLSDMEKK 360
Nodal3p1.L      ...........................................................R 305
Nodal3-e2.L     ................V.................L.....................R. 360
Nodal3-e3.L     ................V.....................................R. 360
Nodal3-e4.L     ................V............A....L..M..................R. 360
                                          *   *

Nodal3-e1.L     ECPSCVPMKMMSMSMLYYENEDFILRHHEEMIVEECGFKDM------- 401
Nodal3p1.L      .......V.............................----HVTLTVS 349
Nodal3-e2.L     .......V.....................................------- 401
Nodal3-e3.L     .......V.....................................------- 401
Nodal3-e4.L     .......V.....................................------- 401
                    *                               *
```

B



Suzuki et al. Fig. 6