www.sciencemag.org/cgi/content/full/314/5801/952/DC1

Supporting Online Material for

Genomic Insights into the Immune System of the Sea Urchin

Jonathan P. Rast,* L. Courtney Smith, Mariano Loza-Coll, Taku Hibino, Gary W. Litman

*To whom correspondence should be addressed. E-mail: jrast@sri.utoronto.ca

Published 10 November 2006, *Science* **314**, 952 (2006).
DOI: 10.1126/science.1134301

This PDF file includes

Supporting Online Material

Materials and Methods

Homolog identification. Numbers of innate recognition gene homologs for human, *Ciona intestinalis*, *Drosophila melanogaster* and *Caenorhabditis elegans* were determined from literature cited in the paper, searches of genes containing relevant domains in the PFAM (*S1*) and SMART (*S2*) databases, and by targeted searches of each genome. Homologs were identified from the *Strongylocentrotus purpuratus* genome using combinatorial HMMER searches (*S3*) with relevant PFAM profiles targeting gene model data bases (see Science Sea Urchin Genome Main Paper), and whole genome translations using shotgun sequence and various assembly releases. Searches also employed the TBLASTN and BLASTP programs (*S4*) carried out with heterologous and sea urchin sequences as queries.

Phylogenetic analysis. Neighbor joining analyses of *Strongylocentrotus purpuratus* TLR TIR domain sequences were made with the MEGA 3.1 program (*S5*). Amino acid alignment positions that include gaps were completely excluded from the analysis. A Poisson correction was used to calculate genetic distances. Bootstrap values were calculated from 1000 replicates. Gene models with incomplete TIR domains were not used in this analysis.

Determination of TLR expression in sea urchin coelomocytes. Quantitative PCR was carried out as described previously (*S6*, *S7*) in order to determine the relative levels of expression of TLR families in sea urchin coelomocytes taken from a single animal. Primer sets were designed to target subfamilies of TLR genes. RNA was isolated from coelomocytes using Trizol (Invitrogen) and an RNA Easy Micro Kit (Qiagen, Valencia, CA). All RNA samples were DNase treated prior to reverse transcription. cDNA was synthesized from random primers using TaqMan Reverse Transcription Reagents (Applied Biosystems). Measurements were made in triplicate using an ABI7000 real-time PCR machine and SYBR green chemistry (Applied Biosystems). Each reaction used cDNA derived from 7.5 ng RNA as template.

Resulting values were normalized to measurements of 18*S* ribosomal RNA as an estimate of input RNA. 18*S* measurements were made on 1/1000 diluted cDNA. For TLR measurements, cycle threshold values for positive sample ranged 20-32 cycles with a distinct dissociation curve (TLR family members were highly conserved within the amplicon region). No-template controls samples did not yield detectable product. Average TLR subclass C_t values (*a*) were subtracted from 18*S* average C_t (*b*). The normalized ordinate value (*Y*) was calculated as $Y = 10000 \times 1.95^{(b-a)}$.

Table S1. Quantitative PCR primers for TLR subgroup expression measurements.

TLR GROUP	Forward Primer	Reverse Primer
IA	CCATGGAYCACGTGAGTGA	AGRTACAACCTYGCCAAGAA
IB	CCGAATTYGTAGTCGTGGTCTT	GTAGGGTCTGCCRTCACTTAAA
IC	TCCTCGGGTACAACGAGCTA	TTCGCGWATCCAATGTTCA
ID	CAATGATTATGAGTTCGACATGAA	GGGAGTCTTTCTTCGAGAGTTG
IE	AGTGACGGCAGACCTTACCT	AGGTTGATGGTCAGATTCTTTG
IIa-1	GCCATCGACAACAGCTTCAA	CAGGAACAACCTGACAAGGTA
IIa-2	CATCTRCAGAACATCATCTACG	GCCAWACGRAGCTTGGTCA
IIb-1	CTGCARGAGAGGTTTCCACAT	GTCAAGAACCARGCATCATC
IIb-2	CTGCAGGAGAGGTTTCCACAT	GTCAGAAACCAGGCYTCATC
IIIa	ATATTTGGYGATGCCGATCT	CTTGTGGCTGTTTTCGATGA
IIIb	CCGTGGAGCATATGAATGAC	GACTCAGCAACAGCCKTACAA
IV	TGAYCTTCCGTTRGGTATGC	GGAARATCACCACCATGTTCTC
V	CGAATCGTTTGTGGTGATGA	CATGAACCACTGGTCCTGTAA
VI	ACCTYCCGCTTGGAATGTACT	GCCCATCAGGAATGTCTTC
VII	GACTCKCCATTTTGGGCTACA	GTTKGTCCTGAATCCAATTCTC
Prot.like1	TTGTGGATGTCATTGGATGTTT	GCGCCTTACAAGTCCAAGTC
Prot.like2	ACTTGCGACTGCCGTCTTAC	ATTCAGGTTGGGTGGTGAAA
Intron	GGAGAAGGAAGWACCAGCGTATC	CTAGGTKGCGAACCAGWTCATT
Short1	ACTTTCAACGCCGTGCTAAT	TTCGTGATGGCTGACATAGG
Short2	CCTAAGTCCTCCTCCAGGAAAG	GATCAGGCCTGGGTAGACAA
Short3	GCACGATAGGCCTACCCATA	CTCCGACAGCAATGAGAACA
18*S* **Ribosomal**	CAGGGTTCGATTCCGTAGAG	CCTCCAGTGGATCCTCGTTA

SOM Text: Reliability of Gene Multiplicity Estimates

Gene counts were made from the 07/18/05 whole genome shotgun genome (WGS) assembly, the 6/15/06 BAC+WGS V. 2.0 assembly and from analysis of approximately 5X coverage of unassembled whole genome shotgun (WGS) sequence (*S8*). The sequence was derived from DNA prepared from sperm of a single animal so inter-individual polymorphisms are absent. A caveat when analyzing these assemblies is the possibility for some genes of accidental inclusion of both haplotypes. For the TLR genes our independent estimates indicate that this problem is minor. Because the TLRs are almost entirely encoded by single exon genes, they are amenable to this type of analysis and enumeration from unassembled sequence. Our independent estimate of gene number is based on the number of unique TIR domain sequences and completely coincides with the assembly-based estimate.

In this analysis we identified 357 unique TIR unique sequences that are present at least twice in 5X coverage of whole genome shotgun trace sequence. Assuming that all loci are polymorphic in the TIR region and taking into account the 25 non-TLR TIR-domain containing proteins that we identified, a minimum of 150 genes are estimated to be present. This estimate was dependent on intact TIR domains since it relied on a HMMER PFAM search of ORFs > 200 nt. Accounting for pseudogenes with disrupted TIR domains (about 30% of the TLR sequences appear to be pseudogenes) and allowing for alleles that do not differ between haplotypes in this region, the two estimates (i.e., from the assembly and from the raw genomic sequence) are entirely consistent. The NLR estimates similarly rely on the number of identified unique NACHT domain sequences of this subfamily. These genes are more complex and estimates may be somewhat less precise than for TLRs, but even in the most unlikely extreme scenario in which every allele is sampled as a separate gene, the sea urchin NLR genes are many times more prevalent than their counterparts in mammals. Notably for nearly all genes known to be single copy, only a single haplotype is included in the assembly.

Supplementary Figure Legends

Fig. S1. A diversity of TLR gene subfamilies are expressed in sea urchin coelomocytes. A quantitative PCR scan of TLR subfamily expression using primer sets that are designed to specifically amplify all members of TLR subfamily expression using primer sets are designed to specifically amplify all members of each subfamily (see supplementary materials and methods). Expression level (ordinate) is given in arbitrary units normalized to 18*S* ribosomal RNA measurements carried out in the same experiment to serve as an estimate of total input RNA. Notably, expression is not strictly correlated with gene family size. A similar measurement from a second animal produced nearly identical results, while measurements from larval message suggest that a different but overlapping suite of TLRs is expressed at this stage (data not shown).

Fig. S2. Three representative V-type Ig domains from a multigene family encoded in the sea urchin genome. A family of approximately 45 genes from this subgroup is distributed in small clusters in 25 Scaffolds of the V2.0 assembly. One cluster is linked in Scaffold_V2_74946 to a large cluster of TLR genes although this association will need to be verified by independent genomic methods (see paper Fig. 2). Each V region is associated with a signal peptide with no intervening intron (boxed) and an open reading frame that continues downstream and is sometimes associated with a potential transmembrane region. Conserved cysteine and tryptophan residues are highlighted in yellow. Sequence names include the Scaffold in which they are encoded and the first nucleotide position and orientation. Characterization of putative transcript sequence beyond the V region (shown here) will need to be verified in cDNA analyses.

References

S1. A. Bateman *et al.*, *Nucleic Acids Res* **32 Database issue**, D138 (2004).

S2. I. Letunic *et al.*, *Nucleic Acids Res* **32**, D142 (2004).

S3. S. R. Eddy, *Bioinformatics* **14**, 755 (1998).

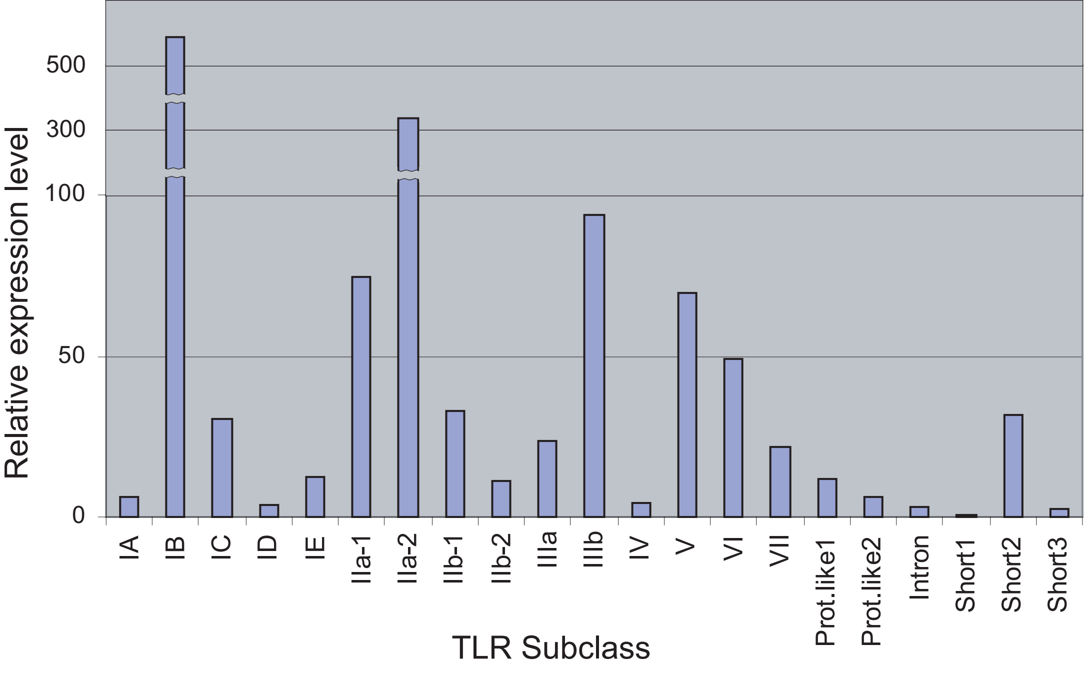
S4. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (1997).

S5. S. Kumar, K. Tamura, M. Nei, *Brief Bioinform* **5**, 150 (2004).

S6. J. P. Rast, R. A. Cameron, A. J. Poustka, E. H. Davidson, *Dev Biol* **246**, 191 (2002).

S7. S. D. Fugmann, C. Messier, L. A. Novack, R. A. Cameron, J. P. Rast, *Proc Natl Acad Sci U S A* **103**, 3728 (2006).

S8. Sea Urchin Genome Sequencing Consortium, *Science* **314**, 941 (2006).


Relative expression level
500
300
100
50
0
IA
IB
IC
ID
IE
IIa-1
IIa-2
IIb-1
IIb-2
IIIa
IIIb
IV
V
VI
VII
Prot.like1
Prot.like2
Intron
Short1
Short2
Short3
TLR Subclass

```
Scaffold_v2_32378 (6,447 RC)    MALVNASTIILLCLAFRTYFCNG-AISMTRI
Scaffold_v2_74946 (513,821 F)   MGIFSLSTIFLICVTWSTIFCAGSALFVKKA
Scaffold_v2_20857 (2,018 F)     MGIFSLSTTFLLCVTWSTIFSGRSALSVKKT

LDVGDNMTLDCVTDRDQDIDMHWIQRKNGSP--QYVAKAGKSFPNGKIYENIRH
LNVGDDITLDCVTDWNPNITLHWLQRKIGSSRHSYIATAGPQNEKGRIWQRFRD
VNIGDDITLDCVTKWNPNITLRWLQRKIGSPRHRYIATAGPQNEKGHMLQRFRD

DPRISVSFVNVSSDKKRINLALSISKIAEQDSAVYTCASSNLLNNKTTNLYFYD
DSRFSYSYEIVSSSEMKIALTIYISNITEEDSAYYTCISVEKGSYNPKTLSTYD
DSRFSYSYENVSSSEMKIALKIYISNITEEDSANYTCIAVEKGSYNPNTLSTYD
```