Alien invasion in the kingdom of insects

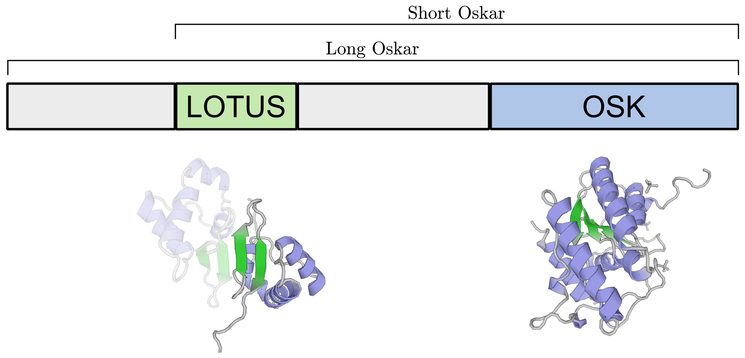
Leo Blondel[a], Cassandra G. Extavour[a]

Introduction

To acquire new functions, evolution often relies on the creation of new genes. Well understood mechanisms of new gene creation include duplication, local mutation, and domain rearrangements. In some cases, however, genetic material can be acquired from exogenous DNA, a phenomenon known as horizontal gene transfer (HGT). *oskar* is a gene found only in the insect lineage that has evolved to be absolutely necessary for germ cell formation, and thus species survival, in some clades, including the fly Drosophila. Its evolutionary history, however, remains a mystery, as no homologs appear to exist outside the insect lineage. Here we elucidate the evolutionary origins of *oskar*, and show that *oskar* likely arose through a novel gene formation history. We used numerous recently published insect genomes and transcriptomes to analyze over 100 *oskar* sequences. We provide strong evidence that while one of *oskar*'s two conserved domains, the LOTUS domain, was present in an ancestral insect genome, the second one, the OSK domain, was acquired through a horizontal domain transfer from a bacterial GDSL-like lipase. Both domains would then have fused via conversion of intervening genomic DNA from non-coding to coding sequence,creating a new unique gene, which we hypothesize would have been the ancestral *oskar* sequence. Finally, we show that the OSK domain is related to GDSL-like lipases from a bacteria of a clade including to the genus Wolbachia, suggesting that bacterial endosymbionts could have provided source material for the HGT event that led to the genesis of *oskar*.

oskar

The gene *oskar* is necessary and sufficient in some insect lineages to establish the germ line [1]. It is composed of two protein domains separated by an unfolded region of variable length. The LOTUS domain is a canonical domain found across Eukaryotes, and the OSK domain is of unknown origin and found only in this protein [2]. In drosophilids there exist two isoforms of the protein, long and short.



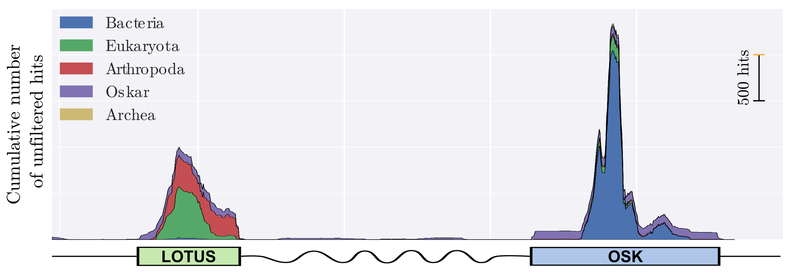
Above is the schematic structure of the gene *oskar*, along with the two crystal structures of the LOTUS and OSK domains.

An intriguing sequence identity

In order to analyze the proportion of sequences each domain matched with in the databases, we created a small algorithm [3] that performs a sliding window HMMER search (based on the 100 *oskar* alignment) on the Trembl [4] database. For each position, we computed the proportion of the best matching sequences from each domain of life. One would expect to find a very similar distribution for each protein domain, or at least one that has similar domains proportions. Interestingly we found that the LOTUS domain is almost exclusively present in Eukaryotes, and the OSK domain in Bacteria (fig1). The question then becomes:

Is *oskar* the product of a Horizontal Domain Transfer ?

Figure 1: Proportion of HMMER hits against Uniprot Trembl database



LOTUS and OSK Phylogenies

In order to infer a Horizontal Gene Transfer event, phylogeny reconstruction is the most robust method to date. In our case, we first gathered the closest 100 non oskar sequences in Trembl [4] for the LOTUS and OSK domains. As there were no Eukaryotic sequences in the closest 100 sequences, and in order not to bias our OSK tree, we manually included all Eukaryotic sequences in the set, and the reverse (bacterial sequences) for the LOTUS set. We then performed a phylogenetic reconstruction using RaxML [5] and MrBayes [6]. Both approaches revealed that the LOTUS domain is most closely related to domains in the eukaryotic Tudor protein family (fig2). Interestingly the Dipteran LOTUS domain seems to have followed a distinct evolutionary trajectory compared to that of the other insect orders, which may be due to the unique germ-plasm role of oskar that has been found in this order [7] (fig2). However, when we looked at the OSK domain, we found that it branched among bacterial GDSL-like lipase, rather than eukaryotic sequences (fig3). Furthermore, the OSK domain was nested within a bacterial clade known to be germ cell symbionts, one of which (*Firmicutes*) contains the Wolbachia clade [8] (fig3). Accordingly, we hypothesize that the *oskar* gene arose from a Horizontal Domain Transfer event from a bacterial to an insect genome.

Figure 2: LOTUS phylogeny

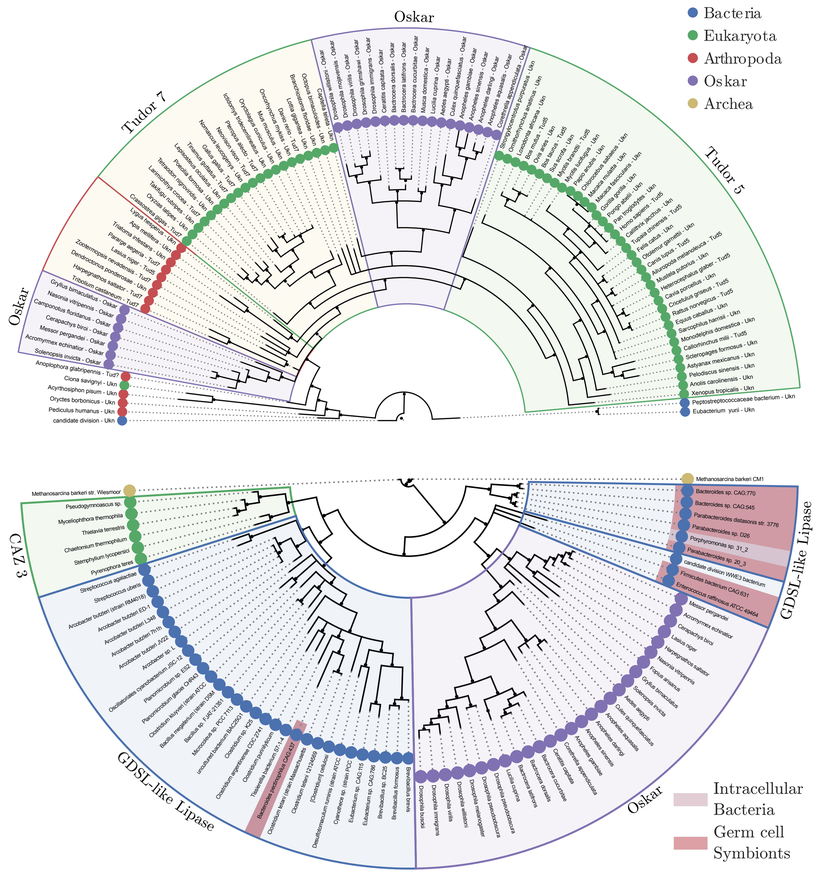
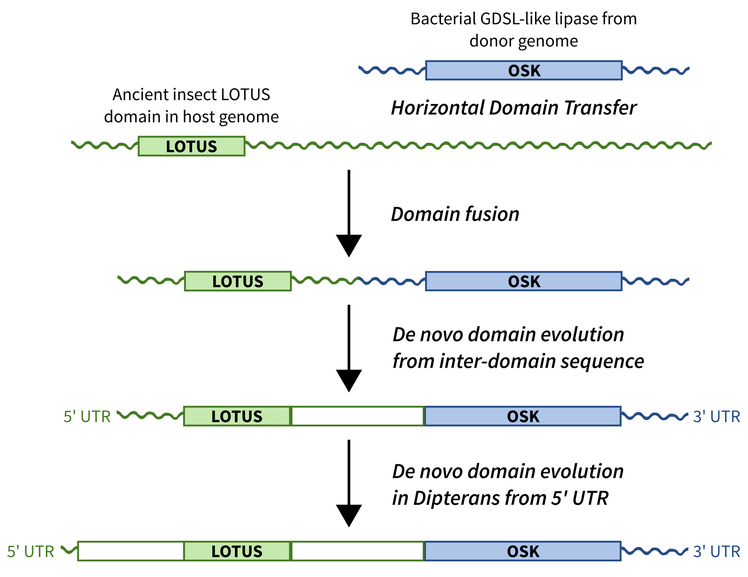


Figure 3: OSK phylogeny

oskar genesis hypothesis



Codon Analysis

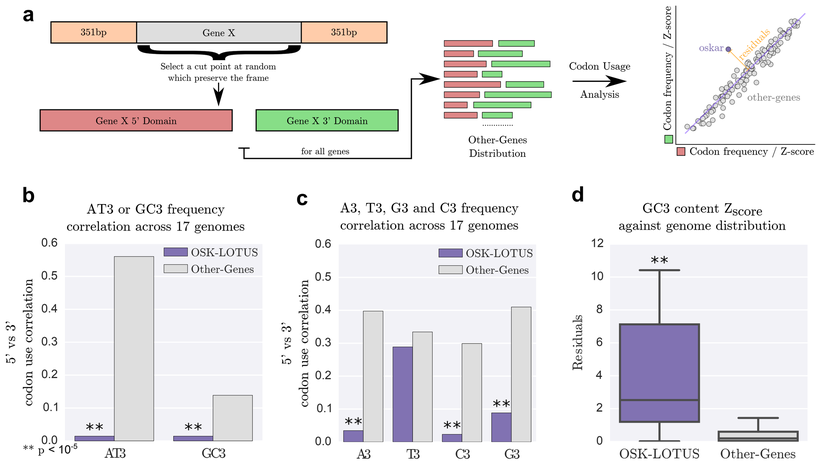


Figure 4: Codon usage of *oskar*

Every organism has a fine tuning of its codons. Indeed, tRNA concentration varies, and sequence mutations on the wobble base (3rd base of a codon, ie A3, G3, ...) can be subjected to selective pressure [9]. Thus, variation from the organism codon usage distribution can be a signal of a Horizontal Gene Transfer [10]. In our case, we must compare the codon usage of two domains within the same gene. Thus, for a group of 17 well-annotated Dipteran genomes, we created a null distribution of intra-gene codon use by randomly splitting each gene into two parts while preserving the frame of each part (fig4a). We then compared in (fig4b) and (fig4a), the codon use between OSK and LOTUS domains to the Intra-Gene distribution. Interestingly, oskar does not follow the expected distribution. In other words, the two domains of Oskar proteins show different patterns of codon use, whereas other genes in the genome show, on average, the same patterns of codon use across the length of the gene sequence. Finally, we fit a linear model to the Intra-Gene GC3 Zscore distribution (see schematic) and looked at the distance (residuals) of all genes and of oskar (fig4d) to the fit. Once again, we found that oskar is much more distant from the pattern of codon use than expected by the pattern observed for most genes in the genome.We conclude that those results are consistant with the Horizontal Domain Transfer origin for the OSK domain of the oskar gene.

Conclusion

Here, we have shown multiple lines of evidences in favor of a Horizontal Domain Transfer origin of the OSK domain of *oskar*. First, the two protein domains of the Oskar protein show similarities to proteins from distinct Domains of life. Second, the phylogenetic affinity of the LOTUS domain is as expected based on the species tree, with the closest sequences being eukaryotic LOTUS domains from the TUDOR 5 and 7 proteins. In contrast, the phylogenetic affinity of the OSK domain suggests that this domain is most closely related to GDSL-like lipases present only in bacteria. Third, the organisms whose GDSL-like lipases most closely resemble the OSK domain come from clades known to be germ cell symbiots (notably *Firmicutes*, of which Wolbachia is a member), providing a biologically plausible mechanism for the hypothesized Horizontal Transfer. Fourth, the codon usage between the LOTUS and OSK domain are more different than one would expect based on the analysis of 17 Dipteran insect genomes.

Based on those multiple lines of evidence, we conclude that the most likely hypothesis for the genesis of *oskar* is a Horizontal Domain Transfer of an ancestral GDSL-like lipase to a genomic position close to a LOTUS domain, in an ancestral insect genome.

We propose that following this event, an extension of the open reading frame of the gene containing the LOTUS domain occurred, allowing for the OSK domain to be translated in frame. Finally, in some dipterans the 5'UTR evolved into a coding sequence.

References

[1] Ephrussi, A., & Lehmann, R. (1992). Induction of germ cell formation by oskar. Nature, 358(6385), 387–392. https://doi.org/10.1038/358387a0

[2] Quan, H., & Lynch, J. A. (2016). The evolution of insect germline specification strategies. Current Opinion in Insect Science, 13, 99–105. https://doi.org/10.1016/j.cois.2016.02.013

[3] https://github.com/Xqua/Iterative-HMMER

[4] The UniProt Consortium; UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017; 45 (D1): D158-D169. doi: 10.1093/nar/gkw1099

[5] Alexandros Stamatakis; RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30 (9): 1312-1313. doi: 10.1093/bioinformatics/btu033

[6] Ronquist, F. and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

[7] Ahuja, A., & Extavour, C. G. (2014). Patterns of molecular evolution of the germ line specification gene oskar suggest that a novel domain may contribute to functional divergence in Drosophila. Development Genes and Evolution, 224(2), 65–77. https://doi.org/10.1007/s00427-013-0463-7

[8] Dunning Hotopp, J. C., Clark, M. E., Oliveira, D. C., Foster, J. M., Fischer, P., Munoz Torres, M. C., … Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science, 317(5845), 1753–1756. https://doi.org/10.1126/science.1142490

[9] Li, J., Zhou, J., Wu, Y., Yang, S., & Tian, D. (2015). GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage. G3 (Bethesda, Md.), 5(10), 2027–36. https://doi.org/10.1534/g3.115.019877

[10] Tuller, T. (2011). Codon bias, tRNA pools and horizontal gene transfer. Mobile Genetic Elements, 1(1), 75–77. http://doi.org/10.4161/mge.1.1.15400

a) Research performed in the Extavour laboratory, Organismic and Evolutionary Biology (OEB) and Mollecular and Cellular Biology (MCB), Harvard University

lblondel@g.harvard.edu

mco molecules, cells and organisms

HARVARD UNIVERSITY