2013-1

MonarchBase: the monarch butterfly genome database

Shuai Zhan
University of Massachusetts Medical School

Et al.

D758–D763 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 9 November 2012*
doi:10.1093/nar/gks1057

MonarchBase: the monarch butterfly genome database

Shuai Zhan and Steven M. Reppert*

Department of Neurobiology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, MA 01605, USA

Received July 29, 2012; Revised October 9, 2012; Accepted October 11, 2012

ABSTRACT

The monarch butterfly (*Danaus plexippus*) is emerging as a model organism to study the mechanisms of circadian clocks and animal navigation, and the genetic underpinnings of long-distance migration. The initial assembly of the monarch genome was released in 2011, and the biological interpretation of the genome focused on the butterfly's migration biology. To make the extensive data associated with the genome accessible to the general biological and lepidopteran communities, we established MonarchBase (available at http://monarchbase.umassmed.edu). The database is an open-access, web-available portal that integrates all available data associated with the monarch butterfly genome. Moreover, MonarchBase provides access to an updated version of genome assembly (v3) upon which all data integration is based. These include genes with systematic annotation, as well as other molecular resources, such as brain expressed sequence tags, migration expression profiles and microRNAs. MonarchBase utilizes a variety of retrieving methods to access data conveniently and for integrating biological interpretations.

INTRODUCTION

The eastern North American monarch butterfly (*Danaus plexippus*) undergoes a spectacular long-distance migration in the fall. The monarch has emerged as an excellent model for investigating the general molecular and neural basis of long-distance migration (1,2). The remarkable navigational capabilities of monarchs are part of a genetic program that is initiated in migrants; the butterflies that travel south to Mexico are at least two generations away from the previous generation of fall migrants (3). Fundamental to decoding the genetic basis of the long-distance migration has been the construction of the draft sequence of the monarch genome (4).

The monarch genome and its transcriptome were sequenced *de novo* using next-generation sequencing technologies (4). The difficulty of assembling the genome from wild-caught butterflies with potentially high heterozygosity was overcome, thus allowing the construction of the initial version of the monarch genome assembly (v1) which consisted of 273 Mb with 16 866 protein-coding genes (4).

Although the original assembly was quite complete for gene coverage, its quality was hindered because of small scaffold size (N50 of 53 kb) and high redundancy (~10%). By implementing new assembling strategies and new libraries, these difficulties have been largely overcome, resulting in a substantial improvement of the monarch butterfly assembly (named v3): 90% of the 249 Mb assembled sequence is now represented by 366 major scaffolds whose minimum length is 160 kb. The improved organization of the monarch genome should allow more precise annotation work. Furthermore, it provides a high quality reference that will facilitate future population genetic studies. For example, researchers now can re-sequence other monarch populations or non-migratory *Danaus* species to help identify migratory genes.

MonarchBase was developed as a public database for readily accessing the monarch genome, its proteome and related biological processes. The growing amount of genomic data and its continuous qualitative improvement necessitated a centralized database to coordinate the inflow of monarch genomic resources. Compared with public data repository, organism-specific databases provide the community with specialized data sets, powerful retrieving interfaces, a platform for extensive biological interpretations and a site for the integration of a variety of previously dispersed data types. MonarchBase serves not only researchers interested in monarch butterfly biology and the biology of the migration but also the wider lepidopteran community.

*To whom correspondence should be addressed. Tel: +1 508 856 6148; Fax: +1 508 856 6233; Email: Steven.Reppert@umassmed.edu

We report here the development of MonarchBase, its components and the latest version of monarch genome assembly and its corresponding geneset.

RESULTS AND DISCUSSION

Data content

The current data content in MonarchBase is summarized in Table 1.

Genome assembly

Assembling genomes with potential high levels of polymorphism has remained a challenge, as haplotypes are assigned to allelic variants, which results in residual redundancy. The occurrence of residual redundancy in the initial assembly has been reported in several studies (8, 12). To remove redundancy from the initial monarch v1 assembly (4), we used both automated and manual methods. In brief, the shorter one of a duplicated pair of sequences was discarded; this was done by considering sequence identity and sequencing depth. Suspicious sequences that were only detected in one sequencing library were also excluded. Paired-end sequencing libraries, from 200 bp to 20 kb (4), were aligned to the non-redundant sequences, step by step, using BOWTIE2 (13). Local alignment mode of BOWTIE2 helped us effectively map Roche 454 libraries (8 and 20 kb), which were not as rigorously analyzed previously (4). Scaffolds were subsequently constructed based on mapped linkages using SSPACE v2.0 (14). The resulting assembly (v3) consists of 5397 scaffolds spanning ~249 Mb (Table 1). The monarch genome was previously estimated to be 0.29 pg by Feulgen image analysis (15). However, the actual assembled genome size for many species is smaller than their early estimated size (7,16,17), partly because of the presence of heterochromatin, which is near impossible to sequence and assemble (12). Compared with the previous version, the latest monarch assembly has a substantial improvement in connectedness (Table 2). Gene coverage in the new geneset (OGS2.0) is also increased, although our previous, initial version showed good quality of gene coverage (Table 2). The monarch whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AGBW00000000. The version described in this paper (v3) is the second version, AGBW02000000.

Genome annotation

We identified 25 Mb of sequence as repetitive sequences and transposable elements for the v3 assembly, as described for the v1 assembly (4). We applied a variety of prediction methods to annotate repeat-masked scaffolds and provide accurate gene models (Table 1). Five *ab initio* prediction sets, including AUGUSTUS (23), GeneMark (24), Genscan (25), GlimmerHMM (26) and SNAP (27), were independently generated as described earlier (4). Importantly, we added data from the recently released geneset of the passion-vine butterfly *Heliconius melpomene* (8) to help identify butterfly specific genes.

Table 1. Data content in current version of MonarchBase

Genome reference	
Assembly (v3)	5397 scaffolds spanning 248.6 Mb genome with 6.7 Mb as gaps
Repeat	121 269 repetitive elements spanning 25.3 Mb genome
Gene repertoire	
Official geneset (OGS2.0)	15 130
GLEAN consensus set	16 216
Maker consensus set	13 744
AUGUSTUS *ab initio* set	14 550
GeneMark *ab initio* set	27 256
Genscan *ab initio* set	12 921
Glimmer *ab initio* set	23 898
SNAP *ab initio* set	25 758
RNAseq assembly	18 563 genes with 23 543 alternative transcripts
Annotation for OGS2.0	
Public databases[a]	12 943
Lepidoptera genesets[b]	13 572
GO term	8120 genes assigned with 1539 GO terms
InterPro domain	10 034 genes assigned with 5069 domains
KO	8157 genes assigned with 3856 KO terms 4708 genes assigned into 264 pathways
Ortholog group	198 021 proteins from 15 species assigned into 34 392 ortholog groups
Non-coding RNAs	
MicroRNA	116
Transfer RNA[c]	379
Ribosome RNA[d]	127
Other resources	
Brain ESTs	9484
ESTs with microarray data	9417

[a]Public databases used for annotating monarch genes include RefSeq (5), UniRef50 (6) and non-redundant database of NCBI.
[b]Lepidopteran genesets include *Bombyx* geneset (7) and *Heliconius* geneset (8).
[c]tRNAs were predicted by tRNAscan-SE (9).
[d]rRNAs were predicted by RNAmmer (10) or Rfam scan pipeline (11) following the default settings.

Table 2. Quality control of the latest monarch assembly v3 compared with v1 and the other lepidopterans

	Danaus v3	*Danaus* v1	*Heliconius* v1.1[a]	*Bombyx*[b]
Assembly statistics[c]				
L50 (bp)	715 606	53 032	194 302	3 998 728
N50	101	1138	345	38
L90 (bp)	160 499	6262	38 051	60 675
N90	366	7140	1634	260
CEGMA analysis for 248 ultra-conserved CEGs present in genome[d]				
# Complete	230	229	214	195
# Partial	243	241	237	241
Homologs in *Drosophila* geneset[e]				
# Recovered	9655	9653	9539	9524
Average coverage	55.2%	54.5%	53.0%	52.8%
Homologs in *Tribolium* geneset[f]				
# Recovered	11 015	11 017	10 915	10 983
Average coverage	63.8%	63.0%	61.9%	61.3%
Homologs in *Bombyx* geneset[g]				
# Recovered	13 010	12 996	12 820	—
Average coverage	84.3%	83.1%	82.4%	—
Homologs in *Heliconius* geneset[h]				
# Recovered	12 860	12 840	—	—
Average coverage	86.5%	84.9%	—	—

[a]The *Heliconius* assembly used here is the latest version available for downloading from http://butterflygenome.org/, date to June 1, 2012, though a better N50 value (277 kb) was reported on a linkage-based improved version (8), which was not available to us.
[b]The *Bombyx* assembly (7) was downloaded from SilkDB 2.0 (18).
[c]For quantitative statistics of assembly, N50 indicates that half of the total sequence in the assembly is presented by a total of N50 scaffolds of length more than or equal to the L50 size; in a similar way, N90 and L90 indicates how 90% of sequence is presented in the assembly.
[d]Statistics of the complete and partial presence of 248 ultra-conserved CEGs were calculated by CEGMA pipeline v2.4 following the default settings (19).
[e]*Drosophila* geneset r5.36 is from FlyBase (20) and only the longest protein per gene was used for analysis. Recovered queries were automatically calculated by GenBlastA (21) as follows: genblasta_v1.0.4_linux_x86_64 -P blast -pg tblastn -p T -e 1e-5 -g T -f F -a 0.5 -r 1 -c 0.5, output then was processed by a custom Perl script to sort out coverage on a single scaffold.
[f]*Tribolium* geneset 3.0 is from BeetleBase (22) and analyzed as described earlier.
[g]*Bombyx* geneset is from SilkDB 2.0 (18) and analyzed as described earlier.
[h]*Heliconius* geneset 1.1 (8) is from http://butterflygenome.org/ and analyzed as described earlier.

All these predicted genesets and the evidence of monarch cDNAs and insect homology were selected by GLEAN (28) to generate a consensus geneset. In addition, we used the MAKER annotation pipeline (29) to build another consensus geneset using the same inputs as used for GLEAN. As a result, GLEAN and MAKER identified 16 216 and 13 969 genes, respectively. According to the evaluation of 389 manually curated gene models and 20 cloned monarch genes, we chose the non-redundant GLEAN set as our new reference geneset, though we kept both GLEAN and MAKER, as well as all other independent prediction genesets, that are available in MonarchBase for browsing (Table 1).

A total of 15 130 of 16 216 GLEAN genes whose existence was supported from either monarch cDNAs or insect homologs were selected as the new official geneset (OGS2.0) for comprehensive annotation (Table 1). We performed BLASTP against both RefSeq (5) and UniRef50 (6) databases to report annotation information. We also performed both BLASTP and BLASTX against the non-redundant NCBI database to help annotate those uncommon genes and pseudogenes.

We used several methods to annotate genes into families and pathways. A local InterProScan (30) was run against the InterPro domain database (31) to map domains and GeneOntology (GO) terms (32) to monarch genes. KEGG is well-known for their collection of manually delineated pathway maps representing the current state of knowledge on the molecular interactions and reactions (33). We queried monarch proteins against KEGG orthology (KO) using BLASTP (1e-5) and assigned them to biological pathways. In addition, we used an OrthoMCL algorithm (34) to analyze gene orthology among 15 species, as described (4), and clustered genes into ortholog groups representing monarch-specific genes, butterfly specific genes (monarch and *Heliconius*) and lepidopteran-specific genes (monarch, *Heliconius* and *Bombyx*), as well as universal genes. For comparative analysis, we performed multiple alignment for each ortholog group using MUSCLE (35) and selected well-aligned blocks using Gblocks (36).

Functional resources

By mapping monarch brain-derived expressed sequence tags (ESTs) (37) to the geneset, previously identified transcripts associated with the oriented flight behavior of migratory butterflies (38) have all been annotated (4). In addition, more than 7000 monarch genes have expression data for comparison between summer and migratory monarchs (38). Using an integration approach, we also found an unexpected sexually dimorphic pattern within the monarch juvenile hormone biosynthesis regulatory pathway (4). RNAseq reads, representing multiple monarch tissues and developmental stages (4), were aligned back to the new assembly using Cufflinks (39) to present alternative splicing patterns. Universal expression

value for each gene was calculated based on the normalized transcriptome coverage, as described (4). Small non-coding RNA sequencing data for both summer and migratory butterflies (4) were also integrated with the new assembly.

Database organization

We store and manage data for MonarchBase using MySQL (http://www.mysql.com). Several Common Gateway Interface scripts were developed to process users' input to search the database, connect to third-party application, parse the result and generate pages for retrieved data. A schematic diagram of database organization is shown in Figure 1.

Genome browser

MonarchBase utilizes a genome browser, implemented with GBrowse 2.0 (40), to navigate annotation along with the genome assembly. GBrowse is a well-known browser that integrates database and interactive web pages for displaying annotations of genomes, and has been applied to a variety of databases (18,22,41). Through GBrowse of MonarchBase, researchers can access data representing consensus genesets, independent genesets, alternative splicing patterns, homolog and cDNA alignments, repeat content, non-coding RNAs and other genomic features.

Accurate prediction of gene models is the most important task of genome annotation work. For consistency among users, we provide, as already indicated, an official reference geneset, OGS2.0, which is superior in overall quality to each of the independent genesets. Because each gene prediction program currently in use has both strengths and weaknesses, displaying all prediction sets is useful to optimize gene models when there are conflicting overlaps between sets.

Retrieved data

MonarchBase has been designed with several entry sites and accepts entry ID, key words or sequence as input to retrieve data for either a single gene or a group of genes (Figure 1). Gene page is the core of MonarchBase, at which researchers can access all related information for each OGS2.0 gene, including gene symbol, genomic position, evidence of monarch cDNA or insect homology, gene family, biological pathway, ortholog group and nucleotide and deduced protein sequence (Figure 1). Each entry in the gene page links to informative web page. MonarchBase can also return a list of monarch genes, coupled with biological interpretation, for retrieving entries of GO, InterPro, KO, ortholog groups or pathways. In addition, users can browse a list of differentially expressed ESTs and expanded/contracted gene families.

BLAST server

Local Basic Local Alignment Search Tool (BLAST) is one of the most useful entrance sites for a genomic database. At MonarchBase, users can search against a variety of monarch genome-wide data, including scaffolds, contigs, genes and ESTs. We also packed 332 930 proteins from genesets of 20 insect species as a single database, which facilitates search for homologs of most insect orders. We used html4blast, a Bioperl module (42), to customize BLAST output. Through extended links, users can click on identifiers to retrieve relevant information conveniently.

Broad application

As monarchs are famous for their long-distance migration, the biological interpretation of the genome has focused on genes potentially involved in the migration. We have manually annotated more than 1000 genes of biological interest for monarch migration biology and curated more than 100 chemoreception genes (4). With the new assembly, we have updated these gene inventories with OGS2.0 gene models; these are available for browsing in MonarchBase. MonarchBase also includes data from other insect species, which are integrated with appropriate links to other databases. We also provided lepidopteran-specific genes, microRNAs and contracted or expanded gene families based on our analysis. Users from other fields can also download multiple datasets for use in their local comparative analyses. Detailed instructions about how to use each component can be checked in the help file of MonarchBase.

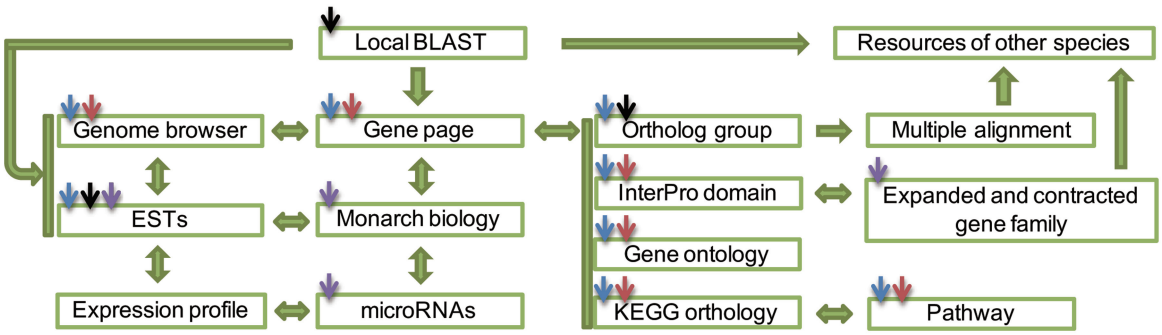

Figure 1. Schematic view of the components of MonarchBase and their connections. The green arrows represent the clickable connections between the components. Thin arrows represent the major entrances of MonarchBase accepting users' input to retrieve data: black arrows indicate the sequence inputs; blue arrows indicate ID inputs; red arrows indicate keyword inputs; and purple arrows indicate browsing menus.

FUTURE DIRECTIONS

Population genomic studies for monarchs and other *Danaus* species should be forthcoming. Identifying variations will be useful for analyzing population substructure and distribution rates, dating the migration of the eastern North American population and eventually uncover candidate migratory genes.

The completeness and contiguity of the monarch genome assembly will be continuously improved as more genomic sequences become available. In addition, the manual curation of additional genes is ongoing and will be updated in MonarchBase. We encourage other research groups to contribute annotations, curations and related datasets via Email (steven.reppert@umassmed.edu). Suggestions and requests for additional functions are also welcome.

ACKNOWLEDGEMENTS

We thank Jeffrey L. Boore for help with initial aspects of the monarch v1 assembly; Alan Ritacco and David Lapointe for assistance with security issue and public access; the *Heliconius* Genome Consortium for early access to the *Heliconius* geneset; and Christine Merlin for discussions and comments.

FUNDING

Funding for open access charge: National Institutes of Health [GM086794-02S1].

Conflict of interest statement. None declared.

REFERENCES

1. Reppert,S.M., Gegear,R.J. and Merlin,C. (2010) Navigational mechanisms of migrating monarch butterflies. *Trends Neurosci.*, **33**, 399–406.
2. Reppert,S.M. (2006) A colorful model of the circadian clock. *Cell*, **124**, 233–236.
3. Brower,L.P. (1996) Monarch butterfly orientation: missing pieces of a magnificent puzzle. *J. Exp. Biol.*, **199**, 93–103.
4. Zhan,S., Merlin,C., Boore,J.L. and Reppert,S.M. (2011) The monarch butterfly genome yields insights into long-distance migration. *Cell*, **147**, 1171–1185.
5. Pruitt,K.D., Tatusova,T., Browse,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
6. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
7. International Silkworm Genome Consortium. (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Molec.*, **38**, 1036–1045.
8. Dasmahapatra,K.K., Walters,J.R., Briscoe,A.D., Davey,J.W., Whibley,A., Nadeau,N.J., Zimin,A.V., Hughes,D.S., Ferguson,L.C., Martin,S.H. *et al.* (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.
9. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
10. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
11. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
12. Chapman,J.A., Kirkness,E.F., Simakov,O., Hampson,S.E., Mitros,T., Weinmaier,T., Rattei,T., Balasubramanian,P.G., Borman,J., Busam,D. *et al.* (2010) The dynamic genome of *Hydra*. *Nature*, **464**, 592–596.
13. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
14. Boetzer,M., Henkel,C.V., Jansen,H.J., Butler,D. and Pirovano,W. (2010) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.
15. Hebert,P.D.N. and Gregory,T.R. (2003) Genome size variation in lepidopteran insects. *Can. J. Zool.*, **81**, 1399–1405.
16. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
17. International Chicken Genome Sequencing Consortium. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.
18. Duan,J., Li,R., Cheng,D., Fan,W., Zha,X., Cheng,T., Wu,Y., Wang,J., Mita,K., Xiang,Z. *et al.* (2010) SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.*, **38**, D453–D456.
19. Parra,G., Bradnam,K. and Korf,I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.
20. McQuilton,P., St Pierre,S.E. and Thurmond,J. (2012) FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res.*, **40**, D706–D714.
21. She,R., Chu,J.S., Wang,K., Pei,J. and Chen,N. (2009) GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.*, **19**, 143–149.
22. Kim,H.S., Murphy,T., Xia,J., Caragea,D., Park,Y., Beeman,R.W., Lorenzen,M.D., Butcher,S., Manak,J.R. and Brown,S.J. (2010) BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res.*, **38**, D437–D442.
23. Stanke,M., Keller,O., Gunduz,I., Hayes,A., Waack,S. and Morgenstern,B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, **34**, W435–W439.
24. Borodovsky,M. and Lomsadze,A. (2011) Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr. Protoc. Bioinform.*, **35**, 46.1–4.6.10.
25. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
26. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
27. Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.
28. Elsik,C.G., Mackey,A.J., Reese,J.T., Milshina,N.V., Roos,D.S. and Weinstock,G.M. (2007) Creating a honey bee consensus gene set. *Genome Biol*, **8**, R13.
29. Cantarel,B.L., Korf,I., Robb,S.M., Parra,G., Ross,E., Moore,B., Holt,C., Sanchez Alvarado,A. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.
30. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
31. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
32. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.

33. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
34. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
35. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
36. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
37. Zhu,H., Casselman,A. and Reppert,S.M. (2008) Chasing migration genes: a brain expressed sequence tag resource for summer and migratory monarch butterflies (*Danaus plexippus*). *PloS One*, **3**, e1345.
38. Zhu,H., Gegear,R.J., Casselman,A., Kanginakudru,S. and Reppert,S.M. (2009) Defining behavioral and molecular differences between summer and migratory monarch butterflies. *BMC Biol.*, **7**, 14.
39. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
40. Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinform.*, **28**, 9.9.1–9.9.25.
41. Cameron,R.A., Samanta,M., Yuan,A., He,D. and Davidson,E. (2009) SpBase: the sea urchin genome database and web site. *Nucleic Acids Res.*, **37**, D750–D754.
42. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.