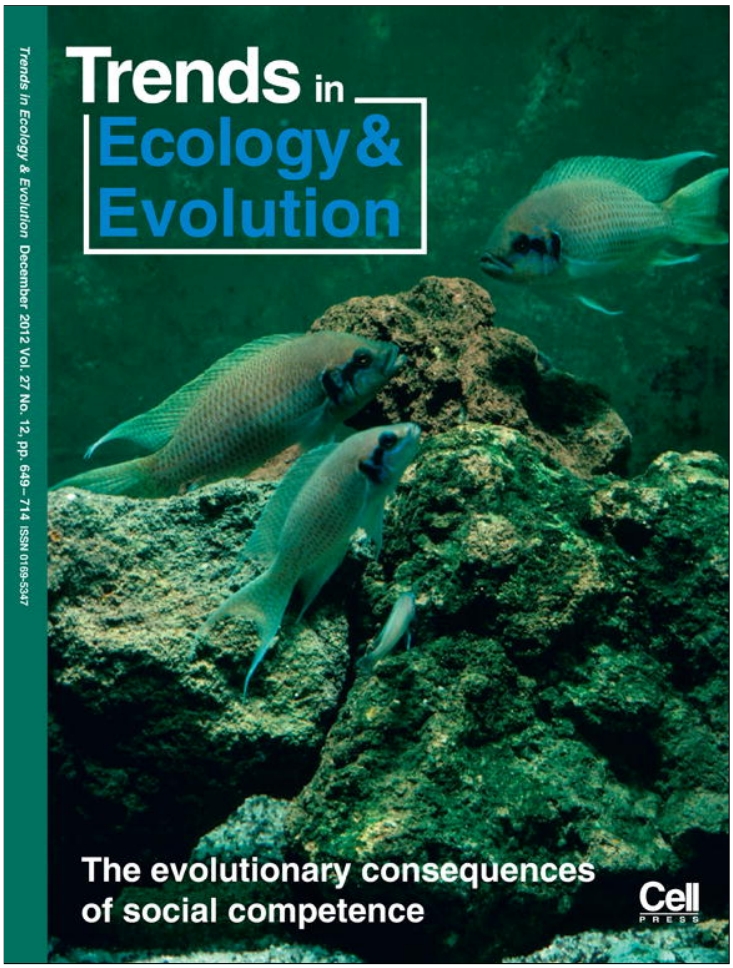

Opinion

Cell PRESS

What is needed for next-generation ecological and evolutionary genomics?

Scott A. Pavey, Louis Bernatchez, Nadia Aubin-Horth and Christian R. Landry

Département de Biologie & Institut de Biologie Intégrative et des Systèmes (IBIS), Pavillon Charles-Eugène-Marchand, 1030 Avenue de la Médecine, Université Laval, QC, G1V 0A6, Canada

Ecological and evolutionary genomics (EEG) aims to link gene functions and genomic features to phenotypes and ecological factors. Although the rapid development of technologies allows central questions to be addressed at an unprecedented level of molecular detail, they do not alleviate one of the major challenges of EEG, which is that a large fraction of genes remains without any annotation. Here, we propose two solutions to this challenge. The first solution is in the form of a database that regroups associations between genes, organismal attributes and abiotic and biotic conditions. This database would result in an ecological annotation of genes by allowing cross-referencing across studies and taxa. Our second solution is to use new functional techniques to characterize genes implicated in the response to ecological challenges.

Ecological and evolutionary genomics

Understanding how the genome of an organism evolves and copes with environmental challenges is the main goal of EEG [1–3]. This involves the integration and detailed knowledge of processes acting at different spatial and temporal scales, spanning from molecules (i.e., genes, genomes, transcriptomes, proteomes, and metabolomes), to the integrated phenotype and ecosystem function of the organism, over both short and long timescales [4]. During the early development of the field, EEG investigations were limited to species for which tools such as DNA microarrays or genetic linkage maps were available, with the consequence that only a few species qualified for this type of investigation. Recent developments of next-generation sequencing (NGS), genotyping, and proteomics technologies have opened the door for species in which these original tools were not readily available [5]. For instance, NGS alleviates the absolute need for taxa-specific genomics tools or a reference genome of the same species to measure quantitative and qualitative gene expression variation [6], to genotype individuals densely [7], or to sequence and assemble a complete genome *de novo* [8]. These technologies also allow for the thorough characterization of the microbiomes found in specific environments or that are associated with plants and animals [9]. The availability of the coding sequences and predicted proteome of organisms by direct cDNA sequencing (RNA-seq) with the same technologies also makes proteomic analysis by mass spectrometry possible [10].

Corresponding authors: Aubin-Horth, N. (Nadia.Aubin-Horth@bio.ulaval.ca); Landry, C.R. (Christian.Landry@bio.ulaval.ca).

These technological advancements make the study of key questions in the field more amenable, including the molecular bases of phenotypic convergence, the contribution of the various cellular regulatory mechanisms to adaptation, the link between the plasticity of a trait and its long-term evolution, and the contribution of genes from symbiotic microbes to plant and animal adaptation (Box 1). However, these technological breakthroughs do not alleviate one of the major challenges in the field, which is to establish mechanistic links between genes, phenotypes, and biotic and abiotic factors, as a large fraction of genes remains without any annotation. These challenges must be overcome if the questions listed above and many others are to be answered fully.

The burden of 'unknown' and poorly annotated genes for data interpretation

The ability to acquire genomic data has exploded in recent years, whereas capacity to interpret this abundance of data is lagging far behind. This disconnect stems from the fact that interpretation of these results largely rests on the establishment of homologies between genomic features of interest and features of traditional model organisms, such as yeast, worms, flies, and mice. It is now possible to use new tools to study many different taxa, but interpretation of the results is limited by the need to refer to these few, well-known models that were established and developed for non-ecological purposes. Even if it is possible to sequence the entire genome and transcriptome of a species, the annotation of the genes will remain the limiting factor when the time comes to interpret the results mechanistically. This is even the case for many genes in model organisms, such as the budding yeast *Saccharomyces cerevisiae*, which remain without any phenotypes in standard laboratory settings [11]. Relating the expression of these genes to the ability to thrive in specific natural conditions may be the only way to annotate completely the genome of model organisms.

Therefore, there is a foreseeable bottleneck in the field, despite the immense opportunities provided by new technologies. There is a consensus in the community that this is an obstacle to moving the field forward, and here we present two reachable solutions for this issue.

Solution 1: better data integration to enable systematic ecological associations of genes and genomes

Without a way of putting together knowledge amassed from the study of various ecological models, researchers

0169-5347/$ – see front matter http://dx.doi.org/10.1016/j.tree.2012.07.014 Trends in Ecology and Evolution, December 2012, Vol. 27, No. 12

Box 1. Major long-standing and emerging questions to be addressed by EEG

Major questions enabled or facilitated by technological developments in ecological genomics and that would benefit from better data integration among studies and from functional analysis include the following:

(i) What are the molecular bases of phenotypes generated by convergent evolution? Do they involve the same genes and/or the same molecular pathways? What is the taxonomic scope of conservation in function? Did these phenotypes evolve by independent mutational events or from ancestral polymorphism?
(ii) What are the roles of post-transcriptional regulatory mechanisms (e.g., small regulatory RNAs or protein modification) in phenotypic diversity, plasticity, adaptation, and evolution?
(iii) Is there variation in the architecture of gene networks (i.e., which genes regulate which other genes) within and between closely related species and does it have a role in adaptive variation and phenotypic plasticity?
(iv) Are the same genes and molecular pathways involved in plastic responses within species and phenotypic divergence for the same traits among populations or species?
(v) What are the roles of genes from symbiotic and endosymbiotic microbes in animal and plant adaptation?
(vi) Are the molecular and cellular functions of genes studied in model organisms conserved in different species and are they environment and species dependent?

Box 2. The Gene Ontology

The Gene Ontology project is a bioinformatics initiative aiming at developing standard descriptions of gene functions (and their products) that would apply to all species. These descriptions are associated with evidence codes, which describe how a given GO annotation is supported by the literature (ex. 'Inferred from mutant phenotype' or 'Traceable author statement'). Direct evidence of gene function is always favored over indirect evidence.

There are three main domains that are covered by current ontologies:

(i) cellular component: the parts of the cell where the protein is found, such as in the nucleus, cytoplasm, or mitochondrion;
(ii) molecular function: the activity of the gene products, such as protein kinase activity, endopeptidase activity, or receptor activity; and
(iii) biological process: an ensemble of molecular interactions or reactions that have an identified beginning and end. These include, for instance, regulation of meiotic cell cycle, response to DNA damage, or peptide maturation.

Within these three domains, annotations are hierarchically organized from general to more specific terms. This organized architecture enables different genes to be grouped together based on shared ontologies. This systematic annotation also enables the comparison of different processes and molecular functions across species, even in cases when the genes under study are not orthologous and when gene orthologies are not well established.

More details are available at: http://www.geneontology.org.

deprive themselves of a tremendous amount of information with which to understand biodiversity. A first, immediately implementable solution would be to develop a standard approach of reporting results that would allow investigators to associate genes or genomic features with organismal traits and ecological conditions linked to the ecology of an organism. For instance, a database could be used to classify the data obtained when deciphering the genomic bases of thermal adaptation along a latitudinal gradient, or of rapid adaptive evolution driven by pollutants, other environmental stressors, or life-history traits. One major question in ecology and evolution relates to the evolutionary convergence of phenotypes in different taxa. If there was a compendium of functional genomics data [gene expression, quantitative trait loci (QTL), or landscape genomics data] from different species that found similar solutions for diversification and occurrence in particular environments or clines (e.g., altitude, latitude, temperature, salinity, water current, precipitation, etc.), then existing studies, coupled with data mining, could be used to help answer this long-standing question. Furthermore, it could serve to generate hypotheses regarding the functions of completely unknown genes that show similar patterns across environments in different species, and to test them by focusing on the candidate genes obtained through ecological annotation.

Towards an ecological association ontology

Ecological associations would be the basis for the creation of a systematic ontology similar to the Gene Ontology (GO; Box 2) but which would instead characterize the ecological function and association of genes. Current annotations provided by GO focus specifically on suborganismal phenotypes, such as biological process, molecular function, and cellular component. Thus, as stated on the GO website, their annotation does not address 'environment, evolution and expression, anatomical or histological features above the level of cellular components, including cell types.' Our proposed 'ecological annotation' [12] would complement the three current types of ontology in GO (Box 2). This might not mean that the mechanisms of action of the genomic feature of interest (e.g., gene, miRNA, or transcription factor binding sites) could be determined, but the systematic association of this feature with particular conditions can help narrow down the particular mechanisms or help to select candidate genes for traits of interest.

Currently, all of the available evidence codes (Box 2) are specific and primarily applicable only to the few traditional model organisms. This stringent criterion is important for GO, but not conducive for EEG to get started on moving from describing genes as 'unknown' to anything more descriptive. The Ecological Association Ontology (EAO) would have a different, less stringent source of evidence codes, such as transplant and/or common garden experiment, environmental gradient, laboratory manipulation and/or reaction norm, replicated parallel divergence, or outlier genetic divergence. Many studies in ecological model organisms have reported genes of unknown function associated with the expression of important phenotypes (e.g., [13–18]). These studies would have greatly benefited from such a database. These 'unknown genes' might share ecologically important functions in different species facing a similar environmental challenge. In such a case, they could be annotated according to their link with a particular ecologically relevant trait. For instance, genes with unknown functions have been identified as being upregulated in the brain of migrating Atlantic salmon compared with non-migratory salmon [16]. These same genes could have been associated with this behavior or another in studies on other vertebrate species, but there is currently no tool available to identify this association. In the currently available annotation systems, these genes remain without

any functional identity, whereas they would become annotated as 'migratory behavior genes' in an ecological annotation system. In another transcriptomic investigation in salmon, Roberge *et al.* [19] showed that domestication of European and North American strains led to the rapid, parallel evolution of similar transcription profiles at many genes, but most of them were not annotated and were labeled as 'unknown' [19]. Moreover, some of those same genes were associated with domestication in another salmonid species [20]. These genes would be annotated as 'domestication genes' in an ecological annotation system.

A system of ecological association would also harness the power of comparative genomics to better define the function of genes that have almost exclusively been functionally characterized in model species. For instance, Star *et al.* [21] found in Atlantic cod (*Gadus morhua*) that the major histocompatibility complex (MHC) II immune system pathway has been lost, and that a massive diversification in the MHC I pathway has filled the void. Thus, through genome sequencing and comparative genomics, the authors discovered a natural 'knock-out' in the adaptive immune system that shows that MHC II genes are not essential for some vertebrates.

Example of an ecological association

Contrasting the existing GO annotation with two possible ecological EAOs for a gene found in *Drosophila sechellia* (Figure 1) illustrates the advantages of an EEG database. This fruit fly species is unique because it reproduces on the 'vomit fruit', *Morinda citrifolia*, which is both toxic and undesirable to other *Drosophila* species [22,23]. Through ecological genomic methods, the gene implicated in the olfactory preference shown by the flies was identified as being an odorant-binding protein (*Obp57e* reference GenBank accession number: **EDW48711.1**) [24]. The existing GO of this gene provides good functional information about this protein from a cellular and systems perspective; however, the pivotal role this gene has in the reproductive ecology of this species is not present and would be highlighted by the EAO (Figure 1).

The need for the systematic reporting of results and gene–ecology associations

We propose that the solution lies in implementing a full-text literature search database of EEG journals, and incorporating the yet-to-be created, vocabulary-controlled EAO in addition to other ontologies. For instance, the Textpresso Project provides flexible open-source software for full-text database creation and browsing of the literature (http://www.textpresso.org). Databases created so far mainly focus on a specific model organism or human diseases. The user selects a combination of terms from a list of hierarchical search terms, including GO, and the literature database is searched for the occurrence of selected terms in the same sentence. This comprehensive full-text search engine, accompanied by a sequence-based 'EcoBLAST' that

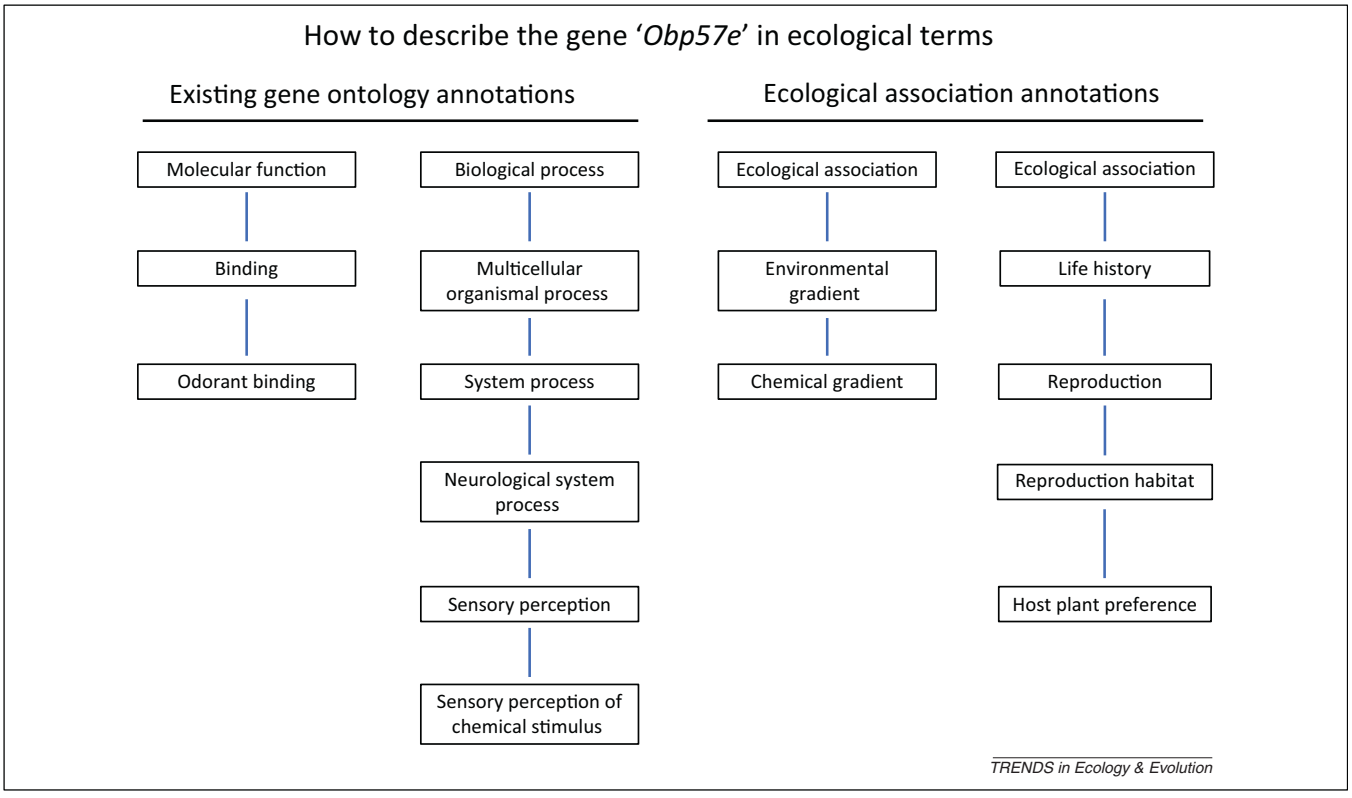

Figure 1. The existing Gene Ontologies (GO) as well as the proposed Ecological Association Ontologies (EAO) for the gene *Odorant binding protein 57e* (*Obp57e* reference Genbank accession number: **EDW48711.1**). The two existing annotations of this gene, both obtained from the AmiGO website (http://amigo.geneontology.org) (the two ontologies on the left), provide a good description of the function of this gene from a cellular and neurological perspective, but the pivotal ecological findings present in the literature are not present in the existing GOs. Complementing the current GOs with an example of our proposed EAO would greatly facilitate cross-referencing genes among studies of ecological and evolutionary genomics (EEG). Many genes of 'unknown' function currently have no GO description. Many of these genes might first be given an annotation with ecological association long before rigorous functional analyses necessary for establish GO are performed.

encompasses EEG data depositions, is what we envision for this system. Such a database would take advantage of the abundant archived data. As a curated database and system of annotation will take time and resources, an interim non-curated database would be far better than the total lack of tools, as it is currently the case. However, a Textpresso server focusing on select EEG journals with the existing options would be relatively easy to implement and would be a huge step forward for EEG resources.

The need for EEG databases

Databases and resources do currently exist that are helpful for EEG researchers (Box 3). The Gene Expression Atlas (http://www.ebi.ac.uk/gxa/) has an excellent format for viewing transcription profile data across studies for particular gene names, linking keywords, expression-level direction, and citation, all on one color-coded chart. The Information Hyperlinked Over Proteins (http://www.ihop-net.org [25]) database relates protein name, physiology, pathology, and phenotype. However, these two resources are both dedicated to model organisms and human disease. There are databases that do search via protein sequences, such as the Protein Families database [26] and Interpro (http://www.ebi.ac.uk/interpro), but these are dedicated largely to molecular functions. These tools also have the ability to dissect unknown proteins and recognize common motifs and domains. However, the main limitation with existing databases is that they are either based exclusively on traditional model organisms or the potential ecological results are overwhelmed by the vast amounts of annotations and papers about traditional models. Thus, EEG needs a database that not only focuses on studies of EEG, but also includes the vast amount of ecologically relevant information from medical, agricultural, and domestic animal genomics (Box 3). To facilitate both of these ventures, it is crucial that EEG investigators publish all genes of interest along with the associated sequence features that allow the inference of homologies, including those without current detailed annotations.

Solution 2: functional analyses

Once traits of interest have been associated with specific genes or genomic features, the final annotation and causal relation would be established by directly perturbing the gene in experimental conditions. This is a *sine qua none* step towards answering another central question: do the candidate mutations act to change gene regulation, protein functions, or the relations of genes within particular molecular pathways (Box 1)? Therefore, more organisms might be needed in which targeted genetic experiments can be performed, such as gene knockouts, knockdown, and gene replacement. This has been taking place in many different fields and standard established protocols are being developed for a wide range of emerging organisms (http://cshprotocols.cshlp.org/cgi/collection/emerging_model_organisms), in a way that previously could only be considered for a handful of organisms (*Drosophila*, yeast, nematodes, and mice) in the recent past. This new group of EEG functional genetic models includes, for instance, *Paramecium*, *Daphnia*, wasps, crickets, cavefishes, and stickleback. EEG researchers will need to tap into these resources even more or develop new ones from their ecological model system to test and functionally characterize candidate genes lacking annotations in an ecological context. This would help answer another key question in the field (Box 1): do genes that are well known and described in model species have the same role in other species or in

Box 3. Functional genomics databases for model and non-model species

The ecological annotation of genes will necessitate the integration of all information related to these genes, coming from various species, obtained at different levels of organization, and for different conditions, using DNA or RNA sequences as the anchoring point. All this information will need to be gathered in databases that are searchable and linked to databases of knowledge on model species and to the literature. There are several existing databases aimed at gathering information in a contextual and highly searchable format that can be used as models for the creation of an ecological annotation database.

Bird base (http://birdbase.arizona.edu/birdbase)
Bird base is a gateway focusing on avian genes and genomes that is a repository of several individual tools that can be used to study different bird species. It offers genome browsers (chicken, turkey, and zebra finches) and functional information can be obtained using a tool (eGIFT) that searches abstracts of publications from PubMed. This database is also connected to GEISHA, a repository of Gallus Expression In Situ Analysis and to GallusReactme, a hand-curated database of metabolic pathways.

AgBase (http://agbase.msstate.edu)
This resource enables one to obtain functional information on genes in animals (from *Daphnia* to cat and guinea pig), agricultural plants (rice and soybean), microbes, and parasites, using, for example, gene names or GO terms as query. Among the tools available focused on GO annotations, GOanna allows one to add GO annotations to a sequence from a species of interest using sequence homology to model systems.

iPlant (http://www.iplantcollaborative.org)
This collaborative portal funded by the National Science Foundation provides tools aiming to uncover the link between genotypes and phenotypes in plants. In addition to several computational tools and data storage space, it provides a collaborative space to connect the community of scientists working in this field. A similar web resource is planned for animals (iAnimal, http://genepro.cshl.edu/ianimal/)

Comparative genomics of cichlid fishes (BouillaBase.org)
This site offers genome browsers for all the non-model cichlid fish species, BLAST services, a comparative genetic map viewer, and physical map viewers.

LumbriBASE (http://xyala.cap.ed.ac.uk/Lumbribase/lumbribase_php/lumbribase.shtml)
This site provide access to genomic and transcriptome datasets for four species of earthworm and is linked with functional information from the model species *Caenorhabditis elegans*. The transcriptome data can be queried in various ways, including by selecting specific libraries dedicated to a certain life stage or environmental condition.

Tomato functional genomics database and the Sol genomics network (http://ted.bti.cornell.edu/ and http://solgenomics.net)
These portals offer data on different tomato species, including transcriptomes, metabolites, and small RNAs, and tools to analyze this type of data. Wild tomatoes are models in ecological genomics [33] and information is also available on maps and markers, genes, phenotypes, pathways, genomes, and specific sequences.

Phenoscape (http://www.phenoscape.org)
This database supported by NESCent contains phenotype statements retrieved from the literature and linked with taxa or genes.

other contexts? Historically, traditional studies aim to find common and shared functions among organisms to better understand human biology, whereas ecological genomics is trying to find what could be the differences in gene functions among species and environments. Therefore, functional analyses need to be done, as much as technologies allows, in the native context. In the same way, studying model systems, such as yeast and *Caenorhabditis elegans*, in ecologically relevant conditions will be key to annotating their entire genome [27].

What to do with recalcitrant species?

We are fully aware that, even in the best conditions, it will remain impossible to perform these types of experiment in many species for which one would like to carry out EEG investigations. For instance, the species might be impossible to breed or maintain in captivity or have a long generation time. However, there are some manipulative techniques that are possible in these situations. One option is transfecting divergent gene variants along with reporter genes into cell cultures to test for gene–environment interactions (or reaction norms). In the Atlantic cod [21], functional polymorphisms were found in promoter regions of hemoglobin genes, and adaptive differences in reaction norms with temperature and oxygen affinity were demonstrated by transfecting different alleles into cell cultures. Another technique is simulating ecological conditions and measuring life-history traits and environmental conditions identified in EEG studies using traditional models where gene manipulation is easier [27]. A third is to start by identifying ecologically relevant orthologous phenotypes or 'phenologs' of EEG species in traditional model organisms [28]. This will lead to unique experiments in traditional models where corresponding orthologous genes can be manipulated, shedding light on the function of ecologically important genes.

The difficulty in performing direct gene manipulations does not prevent the drawing of functional links between genes and phenotypes of interest. We can use *Homo sapiens* as an example of a species that cannot be experimentally bred or genetically manipulated, is long lived, and has a large genome. In short, humans do not *a priori* qualify as an ideal model system to investigate links between genes and functions. Yet, very detailed knowledge of family pedigree and population structure, coupled with meticulous phenotyping, genotyping, statistical methods, and functional studies on orthologous genes have resulted in detailed mapping of the genes underlying Mendelian and quantitative phenotypes (see [28–31]). In the same way, EEG researchers need to follow the example of the biomedical, agricultural, and domestication fields (Box 3) and exploit the explosion of technological developments in sequencing, genotyping, and bioinformatics, along with developing better sampling strategies, computational methods, and functional analyses in the field, to uncover causal links between genotypes, phenotypes, and ecological functions.

Concluding remarks

It is an exciting time for EEG research. Revolutionary technological advances give unprecedented information with which to understand the mechanistic basis of diversity [32]. To use these data at their fullest and reach the main goal of EEG, a common language must be created to connect data obtained from different EEG model systems to create an ecological annotation of genes and genomes, resulting in an amount of usable functional information that will be bigger than the sum of its parts. Finally, the full power of available genetic and molecular studies, including those that have been applied to humans, must be harnessed to study the best EEG models, thus enabling the field to truly enter the era of next-generation EEG.

Acknowledgments

L.B., N.A.H., and C.R.L.'s research in ecological genomics is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). S.A.P. is supported by a NSERC CREATE fellowship from the Réseau Aquaculture Québec (RAQ). We thank Katie Peichel, Tom Mitchell-Olds, John Colbourne, and three anonymous reviewers for comments on earlier versions of the manuscript.

References

1 Elmer, K.R. and Meyer, A. (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends Ecol. Evol.* 26, 298–306
2 Mitchell-Olds, T. *et al.* (2008) Evolutionary and ecological functional genomics. *Heredity* 100, 101–102
3 Rice, A.M. *et al.* (2011) A guide to the genomics of ecological speciation in natural animal populations. *Ecol. Lett.* 14, 9–18
4 Song, B.H. and Mitchell-Olds, T. (2011) Evolutionary and ecological genomics of non-model plants. *J. Syst. Evol.* 49, 17–24
5 Ekblom, R. and Galindo, J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107, 1–15
6 Meyer, E. *et al.* (2011) Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Mol. Ecol.* 20, 3599–3616
7 Hohenlohe, P.A. *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* 6, e1000862
8 Zhan, S. *et al.* (2011) The monarch butterfly genome yields insights into long-distance migration. *Cell* 147, 1171–1185
9 Taberlet, P. *et al.* (2012) Environmental DNA. *Mol. Ecol.* 21, 1789–1793
10 Diz, A.P. *et al.* (2012) Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Mol. Ecol.* 21, 1060–1080
11 Pena-Castillo, L. and Hughes, T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics* 176, 7–14
12 Landry, C.R. and Aubin-Horth, N. (2007) Ecological annotation of genes and genomes through ecological genomics. *Mol. Ecol.* 16, 4419–4421
13 Williams, E.A. *et al.* (2009) Widespread transcriptional changes pre-empt the critical pelagic–benthic transition in the vetigastropod *Haliotis asinina*. *Mol. Ecol.* 18, 1006–1025
14 Sapir, Y. *et al.* (2007) Patterns of genetic diversity and candidate genes for ecological divergence in a homoploid hybrid sunflower, *Helianthus anomalus*. *Mol. Ecol.* 16, 5017–5029
15 Pavey, S. *et al.* (2011) Ecological transcriptomics of lake-type and riverine sockeye salmon (*Oncorhynchus nerka*). *BMC Ecol.* 11, 31
16 Aubin-Horth, N. *et al.* (2009) Gene-expression signatures of Atlantic salmon's plastic life cycle. *Gen. Comp. Endocrinol.* 163, 278–284
17 Collin, H. *et al.* (2010) Response of the Pacific oyster *Crassostrea gigas*, Thunberg 1793, to pesticide exposure under experimental conditions. *J. Exp. Biol.* 213, 4010–4017
18 Franssen, S.U. *et al.* (2011) Transcriptomic resilience to global warming in the seagrass *Zostera marina*, a marine foundation species. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19276–19281
19 Roberge, C. *et al.* (2006) Rapid parallel evolutionary changes of gene transcription profiles in farmed Atlantic salmon. *Mol. Ecol.* 15, 9–20
20 Sauvage, C. *et al.* (2010) Fast transcriptional responses to domestication in the brook charr *Salvelinus fontinalis*. *Genetics* 185, U105–U168
21 Star, B. *et al.* (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477, 207–210

22 Jones, C.D. (1998) The genetic basis of *Drosophila sechellia*'s resistance to a host plant toxin. *Genetics* 149, 1899–1908
23 Whiteman, N.K. and Pierce, N.E. (2008) Delicious poison: genetics of *Drosophila* host plant preference. *Trends Ecol. Evol.* 23, 473–478
24 Matsuo, T. *et al.* (2007) Odorant-binding proteins *OBP57d* and *OBP57e* affect taste perception and host-plant preference in *Drosophila sechellia*. *PLoS Biol.* 5, 985–996
25 Hoffmann, R. and Valencia, A. (2004) A gene network for navigating the literature. *Nat. Genet.* 36, 664
26 Finn, R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222
27 Coolon, J.D. *et al.* (2009) *Caenorhabditis elegans* genomic response to soil bacteria predicts environment-specific genetic effects on life history traits. *PLoS Genet.* 5, e1000503
28 McGary, K.L. *et al.* (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6544–6549
29 Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99
30 Yi, X. *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78
31 Faro, A. *et al.* (2012) Combining literature text mining with microarray data: advances for system biology modeling. *Brief. Bioinform.* 13, 61–82
32 Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J. Heredity* 100, 659–674
33 Moyle, L.C. (2008) Ecological and evolutionary genomics in the wild tomatoes (*Solanum Sect. Lycopersicion*). *Evolution* 62, 2995–3013