REVIEW PAPER
Genomic approaches with natural fish populations

M. F. OLEKSIAK
Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149, U.S.A.

(Received 13 May 2009, Accepted 15 October 2009)

Natural populations v. inbred stocks provide a much richer resource for identifying the effects of nucleotide substitutions because natural populations have greater polymorphism. Additionally, natural populations offer an advantage over most common research organisms because they are subject to natural selection, and analyses of these adaptations can be used to identify biologically important changes. Among fishes, these analyses are enhanced by having a wide diversity of species (>28 000 species, more than any other group of vertebrates) living in a huge range of environments (from below freezing to >46°C, in fresh water to salinities >40 ppt.). Moreover, fishes exhibit many different life-history and reproductive strategies and have many different phenotypes and social structures. Although fishes provide numerous advantages over other vertebrate models, there is still a dearth of available genomic tools for fishes. Fishes make up approximately half of all known vertebrate species, yet <0.2% of fish species have significant genomic resources. Nonetheless, genomic approaches with fishes have provided some of the first measures of individual variation in gene expression and insights into environmental and ecological adaptations. Thus, genomic approaches with natural fish populations have the potential to revolutionize fundamental studies of diverse fish species that offer myriad ecological and evolutionary questions.

Key words: adaptation; Fundulus heteroclitus; genomics; microarrays; natural populations; population genomics.

INTRODUCTION

In general, there have been two methods to study adaptation: experimental evolution (Gibbs et al., 1997; Treves et al., 1998; Ferea et al., 1999; Wichman et al., 2000; Riehle et al., 2001; Archer et al., 2007) and studies of natural populations (Koehn et al., 1976; Watt, 1977; Pierce & Crawford, 1997). Experimental evolution studies produce clear demonstrations of adaptation because one can demonstrate a derived change that has arisen independently and affects fitness relative to the original population or a non-selected population. Even though these studies are powerful tools for understanding adaptive mechanisms, they rely on relatively small populations or little standing genetic variation. Natural populations can be much larger and tend to have much more genetic variation. Additionally, natural populations suffer from
the effects of migrations, stochastic variation in population size and unknown gene
by environmental effects. These traits can enhance natural selection by providing
greater genetic variation, but also make adaptive changes more difficult to disting-
guish from stochastic differences. Thus, populations exposed to natural selection in
an appropriate ecological context or natural setting frequently exhibit more realistic
responses than organisms exposed in the laboratory. Moreover, similar to labora-
tory replicates, comparisons among populations or species can provide the means to
discern biologically relevant adaptations.

One of the reasons that inbred organisms are so often used in medical studies is
that it is easier to study cause and effect when you can account for genetic varia-
tion. Yet, most populations of interest (endangered species, ecologically important
species, species heavily affected by anthropogenic pollutions and humans) have large
amounts of genetic variation (on average 33–50% of enzymes are polymorphic (Mit-
ton, 1997)), and individuals have many heterozygotic loci [c. 15% of loci are in a
heterozygotic state (Mitton, 1997)]. Among humans, it is estimated that >3 million
nucleotide differences occur between any two humans (Li, 1997), and the analysis
of whole genome sequences for individual humans supports this number (Bentley
et al., 2008; Wang et al., 2008). A compelling question throughout biology, how-
ever, is how much interindividual variation is biologically important? In other words,
how much of this variation affects longevity, reproductive fitness or probability of
survival? The huge challenge in studying variation is that most of the DNA vari-
tions within and between populations and species are effectively neutral and have
no phenotypic or fitness consequences. Moreover, most traits are influenced by mul-
tiple genes and environmental factors so that a particular trait will only manifest
in the right condition (e.g. human height is a heritable trait that will only manifest
with proper nutrition and sufficient food). Thus, identification of genetic variation
underlying phenotypic traits is one of the most challenging tasks in biology, and
finding causative polymorphisms that affect ecologically relevant traits in free-living
organisms is even more challenging.

This review focuses on genomic approaches used with natural fish populations to
discern biologically important variation. The power of genomic tools is just beginning
to be harnessed for adaptive studies of natural fish populations. Two complementary
approaches include microarrays and population genomics. Although the vast major-
ity of microarray studies have examined physiological or dose responses, far fewer
microarray studies have looked at adaptive changes in gene expression in natural
fish populations. Population-genomic approaches more directly target the underlying
DNA variation found among individuals and populations and are just now being used
with natural fish populations to identify adaptively important sequence variations.
Together, these two genomic approaches (adaptive gene expression and population
genomics) have the potential to leverage studies of diverse fish species that offer
many different ecological and evolutionary questions.

WHY FISH?

Fish are the most specious vertebrate [28 600 extant species v. 4629 in mam-
mals and 9946 in birds (Wilson & Reeder, 1993)]. Although the mammalian lineage
separated from fishes c. 410 million years before present (B.P.), divergence times
among fish species can be <10 000 years, and the diversity of fish species, many of them closely related, provides for powerful phylogenetic comparisons (Felsenstein, 1985; Garland et al., 1992; Garland & Carter, 1994; Crawford et al., 1999; Fig. 1). As vertebrates, fish share many development pathways, physiological mechanisms
and organ systems with humans and other mammals. Thus, studies with different fish species should further our understanding of vertebrate embryogenesis and physiological processes in general. This is exemplified by the plethora of gene knock-out and knock-in studies in zebrafish *Danio rerio* (Hamilton) development aimed at understanding the mechanistic basis of different diseases as well as the many studies that target particular physiological processes such as respiration, cardiovascular metabolism, osmoregulation and endocrinology. Moreover, as experimental animals, fish are generally much less expensive to maintain in the laboratory than any other vertebrates.

While fish have many advantages for understanding fundamental vertebrate biological processes, they also comprise a highly diverse group. They inhabit a huge range of environmental conditions (Nelson, 1984), and the intimate contact that all fish species have with their aquatic environment (i.e. the direct contact of water in their environment with many tissues and internal compartments) has prompted the view that fish represent excellent models for ‘environmental genomics’ (Cossins & Crawford, 2005). Fish also exhibit a variety of phenotypes, reproductive modes, behaviours and morphologies.

Finally, many fish species have large populations. Thus, evolution by natural selection only requires relatively small selection coefficients \(s > 1/N_e\) to dominate evolution by random drift [where \(N_e\) is the effective population size (Li, 1997)]. Many fish species also (at least in the laboratory) can have very large family sizes. These large family sizes allow association studies and heritability studies to be performed even without inbred strains. For instance, Atlantic salmon *Salmo salar* L. infected with the bacterium *Aeromonas salmonicida* suffer high mortality due to furunculosis. A study using 40 full siblings from each of 120 families of *S. salar* showed that survival was associated with particular MHC alleles, and the relative fitness difference between individuals carrying different MHC alleles was as high as 0.5 (Langefors et al., 2001; Lohm et al., 2002). This type of association study was only possible with large family sizes. Similarly, multiple large families were used to measure heritability of gene expression in *Fundulus heteroclitus* (L.). This study used 13 families of sibling–parent analyses to show that 6.5% of genes had significant heritability in gene expression with a median \(h^2\) of 0.86, similar to studies using inbred strains (D. Crawford, pers. comm.). Thus, fish have many attributes that make them good model organisms for both vertebrates in general as well as ecological, environmental and behavioural questions.

**GOOD NATURAL POPULATIONS**

Attributes of a good species include species with a known population structure and phylogeny so that similarities and differences due to relatedness can be separated from those due to a stress, environment or treatment. Similarly, a known population structure is needed to gauge the relatedness of individuals within species. Thus, one powerful approach to studying natural variation is to combine measures of genetic distance such as microsatellite analyses with measures of variation. For example, to identify genes evolving by natural selection in *F. heteroclitus*, Whitehead & Crawford (2006) used a phylogenetic comparative method to correct for genetic relatedness; they allocated among-population variation in gene expression to genetic distance and then examined the remaining variation relative to temperature
[see Felsenstein (1985) and Harvey & Pagel (1991) for the reasoning behind phylogenetic comparative methods]. Although much of the variation in gene expression fit a null model of neutral drift, the variation in expression for 22% of the temperature related genes (4% of all genes examined) regressed with habitat temperature and was far greater than could be accounted for by genetic distance alone suggesting evolution by natural selection (Whitehead & Crawford, 2006). Similarly, Larsen et al. (2007) combined gene expression and microsatellite analyses to explore adaptive differences in gene expression between North Sea and Baltic Sea flounders maintained in a long-term reciprocal transplantation experiment mimicking natural salinities. Several of the differentially regulated genes could be directly linked to fitness traits (Larsen et al., 2007).

In addition to the ability to measure relatedness within and among species, good natural populations require an observable ecology with environmental attributes likely to affect the fitness of individuals. For example, alcohol dehydrogenase polymorphism of Drosophila melanogaster has been studied in relation to environmental ethanol (within a winery), ethanol tolerance and alcohol dehydrogenase activity (Gibson & Wilks, 1988), armour in stickleback Gasterosteus aculeatus L. and Daphnia has long been studied in association with predation (Harvell, 1990; Peichel et al., 2001a), Saccharomyces paradoxus, the closest known relative of the model eukaryote Saccharomyces cerevisiae (bakers’ yeast), has been isolated mainly from the bark of oak trees and surrounding soil, whereas S. cerevisiae is commonly obtained from vineyards (Koufopanou et al., 2006), and F. heteroclitus has long been studied to understand temperature adaptation among populations (Powers et al., 1991). This is exemplified by work that has explored temperature adaptations in Fundulus populations that inhabit the East coast of North America where they experience a steep thermal cline (Mied & Powers, 1978; Place & Powers, 1979, 1984; DiMichele & Powers, 1982a, b; Van Beneden & Powers, 1985; Crawford & Powers, 1989, 1992; DiMichele et al., 1991; Bernardi et al., 1993; Quattro et al., 1993; Segal et al., 1996; Schulte et al., 2000).

Ideal model species have been further defined as those with a large community of investigators, available genomic sequence and chromosomal maps, forward and reverse genetic tools and ability measure transcript and protein levels (Feder & Mitchell-Olds, 2003). By definition, this limits ideal model species to species with known genomic sequences. Among fishes, G. aculeatus has the most characteristics of an ideal model species. With the advent of new sequencing technologies and decreasing costs, however, many more natural fish populations have the potential to join G. aculeatus as ideal model species.

CAVEATS FOR STUDIES OF NATURAL POPULATIONS

Studying adaptation in natural populations is not trivial primarily due to two interacting factors: genetic variation and environmental variation. Thus, an organism (its genotype) responds to its environment, and differing genotypes in heterogeneous environments probably will respond differently. The genetic contribution to variation can be directly measured by using controlled breeding designs (if not inbred organisms), by measuring relatedness using genetic markers or by mapping the genetic factors using linkage or association studies. These approaches, however, often are
not possible with natural fish populations that lack crosses, pedigrees and sufficient genetic markers to measure relatedness and linkage. Thus, a common strategy to differentiate genetic and environmental effects in natural populations is to hold the environment constant and vary the genetics (*e.g.* common or garden experiments where individuals from two or more populations are maintained in a common environment, most often a controlled laboratory environment). The opposite also can be done: reciprocal transplantations or one genetically related population placed in different environments. Endler (1986) provides a comprehensive summary of methods for detecting natural selection in the wild that is beyond the scope of this paper and also discusses many practical problems in detecting natural selection. These problems lead to three broad outcomes: the lack of detection of natural selection when it is present (type II error), apparent detection of natural selection when it is not present (type I error) and misleading demonstrations of natural selection. Many of the problems derive from a lack of knowledge of the ecology and biology of the species or from poorly designed sampling and statistics (Endler, 1986).

**GENOMIC RESOURCES FOR FISH**

Currently, there are five fish nuclear genome projects (*v.* mitochondrial genomes). These are *D. rerio*, Japanese rice fish or medaka *Oryzias latipes* (Temminck & Schlegel), two pufferfish *Takifugu rubripes* (Temminck & Schlegel) and *Tetraodon nigroviridis* Marion de Procé and *G. aculeatus*. Atlantic salmon *Salmo salar* L. and rainbow trout *Onchorhyncus mykiss* (Walbaum) genome projects are underway and due to be completed by 2011. Although none of these fish genomes has been completely assembled (sequences joined together to reassemble chromosomal order), fish genomic sequences were used to identify many human genes and were used as one of the primary pieces of evidence that the vertebrate genome had 30 000–40 000 genes and not 70 000–140 000 (Roest Crollius et al., 2000). Although nuclear genome projects are lacking for most fish species, a wealth of sequence data exists for many more fish species. These sequence data are largely based on expressed sequence tag (EST) projects. ESTs represent a partial sequence of the much longer RNA expressed in a cell. Because the mRNAs have been processed and edited in the cell, ESTs encode genes that are actively transcribed without intervening intron sequences and so can be more informative about the ultimate function of the gene. Even though ESTs are transcribed genes and thus we should expect 30 000–40 000, however, there are hundreds of thousands of different ESTs in many EST databases. Alternative splicing is partially responsible for the hundreds of thousands of ESTs in different databases; *e.g.* while humans have between 20 000 and 30 000 genes, the Gene Index Project human database (TGI) lists 1 083 935 total unique sequences for humans. The other reason for the multitude of ESTs is that the clustering of mRNA from the same gene is often incomplete, yielding more than one unigene per locus (Paschall et al., 2004).

In teleost fishes, three-dozen species in diverse orders have EST collections that contain more than 10 000 sequences (Table I). Even when all fishes with listed ESTs are included (Table II), only half a dozen more species are added. This is a small number compared to the total number of fish species. Not surprisingly, *D. rerio* have the most ESTs, followed by *O. latipes*, then the salmoniformes (*S. salar* and...
Table I. Teleosts with large (>9000) expressed sequence tag (EST) collections

<table>
<thead>
<tr>
<th>Super order, order and species</th>
<th>Common name</th>
<th>Number of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acanthopterygii (1 304 844)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scorpaeniformes (23 668)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Sebastes caurinus</em></td>
<td>Copper rockfish</td>
<td>23 668</td>
</tr>
<tr>
<td>Tetraodontiformes (26 069)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Takifugu rubripes</em></td>
<td>Torafugu</td>
<td>26 069</td>
</tr>
<tr>
<td>Perciformes (270 476)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Sparus aurata</em></td>
<td>Gilthead seabream</td>
<td>50 930</td>
</tr>
<tr>
<td><em>Dicentrarchus labrax</em></td>
<td>European seabass</td>
<td>44 363</td>
</tr>
<tr>
<td><em>Dissostichus mawsoni</em></td>
<td>Antarctic toothfish</td>
<td>37 104</td>
</tr>
<tr>
<td><em>Paralabidochromis chilotes</em></td>
<td>African cichlid</td>
<td>21 652</td>
</tr>
<tr>
<td><em>Hippoglossus hippoglossus</em></td>
<td>Atlantic halibut</td>
<td>20 836</td>
</tr>
<tr>
<td><em>Lipochromis</em> ‘matumbi hunter’</td>
<td>African cichlid</td>
<td>16 795</td>
</tr>
<tr>
<td><em>Ptyochromis</em> ‘redtail sheller’</td>
<td>African cichlid</td>
<td>14 073</td>
</tr>
<tr>
<td><em>Psetta maxima</em></td>
<td>Turbot</td>
<td>12 427</td>
</tr>
<tr>
<td><em>Sebastes rastrelliger</em></td>
<td>Grass rockfish</td>
<td>11 207</td>
</tr>
<tr>
<td><em>Solea senegalensis</em></td>
<td>Senegalese sole</td>
<td>10 631</td>
</tr>
<tr>
<td><em>Haplochromis burtoni</em></td>
<td>African cichlid</td>
<td>10 312</td>
</tr>
<tr>
<td><em>Thunnus thynnus</em></td>
<td>Northern bluefin tuna</td>
<td>10 163</td>
</tr>
<tr>
<td><em>Paralichthys olivaceus</em></td>
<td>Olive flounder</td>
<td>9 983</td>
</tr>
<tr>
<td>Gasterosteiformes (276 922)</td>
<td><em>Gasterosteus aculeatus</em></td>
<td>276 992</td>
</tr>
<tr>
<td>Cyprinodontiformes (90 970)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Fundulus heteroclitus</em></td>
<td>Killifish</td>
<td>74 755</td>
</tr>
<tr>
<td><em>Poecilia reticulata</em></td>
<td>Guppy</td>
<td>16 215</td>
</tr>
<tr>
<td>Beloniformes (616 739)</td>
<td><em>Oryzias latipes</em></td>
<td>616 739</td>
</tr>
<tr>
<td>Paracanthopterygii (206 507)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Gadus morhua</em></td>
<td>Atlantic cod</td>
<td>206 507</td>
</tr>
<tr>
<td>Protacanthopterygii (867 412)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Esociformes (32 833)</td>
<td><em>Esox lucius</em></td>
<td>32 833</td>
</tr>
<tr>
<td>Salmoniformes (834 579)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Salmo salar</em></td>
<td>Atlantic salmon</td>
<td>494 146</td>
</tr>
<tr>
<td><em>Oncorhynchus mykiss</em></td>
<td>Rainbow trout</td>
<td>281 335</td>
</tr>
<tr>
<td><em>Oncorhynchus tsawytscha</em></td>
<td>Chinook salmon</td>
<td>13 965</td>
</tr>
<tr>
<td><em>Oncorhynchus nerka</em></td>
<td>Sockeye salmon</td>
<td>11 389</td>
</tr>
<tr>
<td><em>Thymallus thymallus</em></td>
<td>Grayling</td>
<td>10 975</td>
</tr>
<tr>
<td><em>Coregonus clupeaformis</em></td>
<td>Lake whitefish</td>
<td>10 312</td>
</tr>
<tr>
<td><em>Salvelinus fontinalis</em></td>
<td>Brook trout</td>
<td>10 047</td>
</tr>
<tr>
<td>Osmeriformes (36 926)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Osmerus mordax</em></td>
<td>Rainbow smelt</td>
<td>36 028</td>
</tr>
<tr>
<td>Otocephala (2 307 644)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siluriformes (493 853)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ictalurus punctatus</em></td>
<td>Channel catfish</td>
<td>354 377</td>
</tr>
<tr>
<td><em>Ictalurus furcatus</em></td>
<td>Blue catfish</td>
<td>139 475</td>
</tr>
</tbody>
</table>
Table I. Continued

<table>
<thead>
<tr>
<th>Super order, order and species</th>
<th>Common name</th>
<th>Number of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cypriniformes (1,813,791)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Misgurnus anguillicaudatus</em></td>
<td>Oriental weatherfish</td>
<td>22,169</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>Zebrafish</td>
<td>1,473,159</td>
</tr>
<tr>
<td><em>Pimephales promelas</em></td>
<td>Fathead minnow</td>
<td>249,941</td>
</tr>
<tr>
<td><em>Cyprinus carpio</em></td>
<td>Common carp</td>
<td>32,046</td>
</tr>
<tr>
<td><em>Rutilus rutilus</em></td>
<td>Roach minnow</td>
<td>18,470</td>
</tr>
<tr>
<td><em>Carassius auratus</em></td>
<td>Goldfish</td>
<td>9,230</td>
</tr>
</tbody>
</table>

*Numbers in parentheses represent the total number of ESTs in a particular super order or order.

Table II. Other fishes with significant expressed sequence tag (EST) collections (>8,000 ESTs)*

<table>
<thead>
<tr>
<th>Class, order and species</th>
<th>Common name</th>
<th>Number of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chondrichthyes (76,075)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Squaliformes (32,562)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Squalus acanthias</em></td>
<td>Spiny dogfish</td>
<td>32,562</td>
</tr>
<tr>
<td>Rajiformes (31,167)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Leucoraja erinacea</em></td>
<td>Little skate</td>
<td>31,167</td>
</tr>
<tr>
<td>Torpediniformes (10,185)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Torpedo californica</em></td>
<td>Pacific electric ray</td>
<td>10,185</td>
</tr>
<tr>
<td>Chimaeriformes (27,944)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Callorhinchus milii</em></td>
<td>Elephant fish</td>
<td>27,944†</td>
</tr>
<tr>
<td>Petromyzontida (129,520)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petromyzontiformes (129,520)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Petromyzon marinus</em></td>
<td>Lamprey</td>
<td>120,731</td>
</tr>
<tr>
<td><em>Lethenteron japonicum</em></td>
<td>Arctic lamprey</td>
<td>8,788</td>
</tr>
<tr>
<td>Myxini (23,886)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myxiniformes (23,886)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Eptatretus burgeri</em></td>
<td>Inshore hagfish</td>
<td>23,886</td>
</tr>
</tbody>
</table>

*Numbers in parentheses represent the total number of ESTs in a particular class or order.
†*Callorhinchus milii* sequences are genome survey sequences (GSS) rather than ESTs. GSS are genomic in origin, rather than cDNA (mRNA). The *C. milii* sequences were derived from a whole genome shotgun library.

O. mykiss) and finally three-spine stickleback. Three of these species also have whole genome sequences. The two pufferfish with whole genome sequences do not have large EST collections because the whole genomes were sequenced first. Among the Acanthopterygii (spiny finned fishes), which comprise 55% of all teleosts and over one-third of all vertebrates, only half of the orders have species with EST collections. At the species level, only 0.13% (19/14,797) of the species in Acanthopterygii has significant EST collections. There is a higher percentage of species with ESTs in the less diverse Protacanthopterygii (2.5%, 9/367), mostly salmoniformes, but these small percentages have yet to exploit the full potential of diverse fish species (Fig. 1). The conundrum remains: should resources be directed to select ideal model species...
or more widely dispersed among species which are perhaps more relevant to the question at hand? The hope is that cheaper sequencing costs will make this question moot.

While ESTs often are sequenced with the end goal of using them for gene expression analyses, ESTs also are a rich source for discovering microsatellites and SNPs. For instance, Ju et al. (2005) mined EST sequences from 49,430 unigenes representing a total of 692,654 ESTs from 4 model fish (O. latipes, Fundulus, D. rerio and Xiphophorus) for their potential use in developing simple sequence repeats (SSRs), or microsatellites. They identified a total of 3,018 EST-derived SSRs and simulation analysis suggested that a majority of these EST-SSRs have sufficient flanking sequences for polymerase chain reaction (PCR) primer design (Ju et al., 2005). Another approach to find polymorphic markers is to compare ESTs within any one species that has multiple ESTs representing the same contig. If these multiple sequence were derived from more than one individual, then the differences among sequences can be used to identify putative SNPs. Caution needs to be used, however, because one cannot always be certain that a particular SNP in an EST is due to true polymorphism v. sequence error or clustering of different loci (i.e. sequence paralogues). EST-derived microsatellites have been used for linkage mapping in P. maxima, S. salar, O. mykiss (Rexroad et al., 2005; Bouza et al., 2008; Moen et al., 2008), and recently, Kucuktas et al. (2009) combined both microsatellites and SNPs derived from ESTs, to construct a genetic linkage map of the Ictalurus punctatus (Rafinesque) genome. Other uses for these EST-derived microsatellites and SNPs include population-genomic analyses (Vasemagi et al., 2005) and whole genome association studies.

EST SEQUENCING

For EST-sequencing projects, typically mRNA is reverse transcribed into cDNA, which is made into a cDNA library (Oleksiak et al., 2001). Clones from the library then are randomly sequenced. Often, either the mRNA or cDNA library is normalized to reduce the copies of highly expressed genes and increase the efficiency of sequencing novel transcripts. Normalization is necessary because a few (10–20) genes make up at least 20% of the total mRNA mass (i.e. there are many thousands of transcripts for a few genes) and a few hundred genes make up 40–60% of the total mRNA mass (Bonaldo et al., 1996). Unless the copies of transcripts for each gene are equalized or normalized, the same gene will be sequenced many times, lessening the discovery of new gene sequences. For example, in the Fundulus EST collection, 33% of the cDNAs was redundant from the non-normalized cDNA library, but only 11% were redundant from the normalized library (Oleksiak et al., 2001). Thus, normalization decreased the probability of sampling the same cDNA by three-fold, and was an effective method for increasing productivity. This approach, of making a cDNA library, normalizing it and sequencing random clones, provides both sequence and the DNA in a plasmid. To be useful, the clones must be organized and stored for later retrieval. This requires considerable space and methods to precisely locate clones. These clones, however, can be used to create DNA for printing microarrays, can be re-sequenced to verify gene identification or get greater sequence depth or for any future manipulations.
A faster sequencing method is 454 sequencing. Using 454 sequencing obviates the cloning step, and the cDNAs are directly sequenced using a highly parallel sequencing platform. Using 454 sequencing, up to 400,000 transcripts can be sequenced in a few hours. What took months just 5 years ago now can be accomplished in days. Among other uses, these sequences can and have been used to design microarrays for studies of natural populations (Vera et al., 2008). This opens the door for transcriptomic studies in virtually any species. Note that with 454 technology, one only gains sequences, and thus, oligonucleotide arrays have to be synthesized in order to perform transcriptomic studies.

Illumina and SOLiD are two other high-throughput, sequencing technologies. These technologies are widely used in species with sequenced genomes because the read lengths are relatively short. While the average read length for 454 sequencing is c. 400 base pairs (bp), Illumina sequences are typically 35–100 bp and SOLiD sequences are 35–50 bp. The longer 454 sequences can be more readily assembled and annotated; however, a genome scaffold is needed to most easily assemble and identify the shorter Illumina and SOLiD reads. (The large number of reads, however, makes Illumina appropriate for de novo transcriptome studies.) Thus, although 454 sequencing is more expensive per megabase of sequence and suffers from greater sequencing error (mostly at the end of polynucleotide stretches) than either Illumina or SOLiD sequencing, it currently is the most viable option for high-throughput, de novo sequencing.

MEASURING GENE EXPRESSION WITH MICROARRAYS

To enhance our understanding of genetic differences among populations or species, we need to be able to survey much of the genome. One of the more powerful ways to achieve this is to quantify most or all of gene expression. Currently, microarrays are the most cost-effective means to measure all of gene expression in both model and non-model species. Although feasible, quantitative transcriptome sequencing using high-throughput sequencing technologies is too expensive for more than a few individuals.

Microarrays are thousands of 150–250 micron spots of DNA bound to microscope slides in a precise and known pattern (Ramsay, 1998; Schena et al., 1998). Each DNA spot quantitatively hybridizes to a specific mRNA so that expression of thousands of individual genes can be measured simultaneously. Importantly, microarray techniques are sensitive: typically two-fold differences in mRNA concentration are determined, and each gene–DNA spot has a sensitivity of 15 attomoles (Schena et al., 1995) or approximately one of 300,000 transcripts can be measured (Hill et al., 2000). Application of statistical analyses with appropriate replication has improved these analyses such that <1.5-fold differences are readily discerned (Jin et al., 2001; Wolfinger et al., 2001; Oleksiak et al., 2002).

Microarray studies of mRNA expression ignore other processes that affect protein activity. Variation in protein activities due to allosteric activators, phosphorylation or other post-translational modifications is not measured. Although enzyme activities or protein concentrations in a cell are not regulated solely by mRNA levels, often changes in cell type, development or physiological states are affected by changes in mRNA. For example, most cancers have altered gene expression, and this has
been used with microarray analyses as a way to classify cancers (Zhang et al., 1997; Golub et al., 1999; Elek et al., 2000; Forozan et al., 2000; Ono et al., 2000). As well, stress responses resulting from infection by pathogenic organisms, exposure to xenobiotics or altered environments (e.g. hypoxia) often result in altered gene expression (Lam et al., 2006; Ju et al., 2007; Oleksiak, 2008; Peatman et al., 2008).

The huge advantage of microarrays is that they can be used to measure thousands of gene transcripts simultaneously. The basic idea is that knowledge of where, when and under what conditions a gene or set of genes is induced or repressed will give insight into biological mechanisms. Gene expression studies have huge potential for non-model species by using 454 sequencing to quickly sequence thousands of transcripts: these sequences are used to direct the synthesis of oligonucleotide microarrays (Vera et al., 2008).

STUDIES WITH FISH MICROARRAYS

This Journal recently published a special volume devoted to fish microarrays [Journal of Fish Biology: Fish Microarrays (2008), edited by K. M. Miller and N. Maclean] and described arrays from 12 fish species: largemouth bass Micropterus salmoides (Lacépède), goby Gillichthys mirabilis Cooper, fathead minnow Pimephales promelas Rafinesque, I. punctatus and the closely related blue catfish Ictalurus furcatus (Valenciennes), European flounder Platichthys flesus (L.), rainbow trout O. mykiss, S. salar, common carp Cyprinus carpio L., Atlantic halibut Hippoglossus hippoglossus (L.), G. aculeatus and Kryptolebias marmoratus (Poey) (Rivulus marmoratus Poey until 2004). The array for the largemouth bass used 454 sequencing to quickly sequence transcripts for the array (Garcia-Reyero et al., 2008). Interestingly, von Schalburg et al. (2008) investigated oligonucleotide designs based on distance from the polyA tail, the effects of mismatches and cross-species hybridization specificity. Notable for comparisons of different natural populations, they found that even three mismatches (95% identity) in a 70-mer oligonucleotide probe decreased the signal by 40% (von Schalburg et al., 2008). Thus, apparent differences among populations may simply be due to variable sequences among individuals. While this should not be a problem for cDNA arrays given the longer probe sizes, this can be mitigated for oligo arrays by averaging signal for each gene across multiple probes. Yet, this adds both to cost and complexity.

Microarrays also have been used with other fish species. In addition to the species mentioned above, there are multiple microarray studies in the model species, D. rerio and O. latipes. As well, gene expression studies have examined hypoxia in Xiphophorus maculatus (Günther) using a heterologous medaka array (Boswell et al., 2009), osmoregulation in the European eel Anguilla Anguilla (L.) (Kalujinaia et al., 2007), dominance in cichlids (Aubin-Horth et al., 2007), limnetic niche use among Coregonine fishes (Coregonus spp., Salmonidae) (St-Cyr et al., 2008) and both adaptive and physiological responses in F. heteroclitus (Oleksiak et al., 2002, 2005; Whitehead & Crawford, 2006; Fischer & Oleksiak, 2007; Roling et al., 2007; Oleksiak, 2008; Everett & Crawford, in press).

The first microarray experiments to measure interindividual variation in gene expression in any species were done in F. heteroclitus (Oleksiak et al., 2002). A subsequent experiment measured an astonishing amount of individual variation.
in metabolic gene expression: 95% of the metabolic genes varied significantly in expression among individuals (Oleksiak et al., 2005). It is hard to believe that this frequency of interindividual difference in gene expression is biologically interesting. Yet, the differences in gene expression are associated with changes in cardiac metabolism: the patterns of gene expression explain up to 81% of the variation in cardiac metabolism. Surprisingly, the patterns of gene expression that explain metabolism differ among individuals. Thus, for glucose metabolism, in one-third of the individuals, variation in glycolytic gene expression is most important (explains >80% of the variation in metabolism) but in other individuals it is the variation in oxidative phosphorylation genes that explain metabolism (Oleksiak et al., 2005; Crawford & Oleksiak, 2007).

The surprisingly large amount of individual variation that explains much of a physiological trait (substrate-specific metabolism) has four critically important implications for human and other studies. First, examining one or a few inbred individuals could be misleading because a pathway or mechanism that is important for one individual is not necessarily important for another one. Second, many genes or different pathways can affect substrate specific metabolism. While enzymes involved in glycolysis might explain metabolic variation for a subset of individuals, enzymes involved in oxidative phosphorylation might be important for metabolic variation in a different subset of individuals. Thus, we cannot measure one or a few enzymes and expect it to always explain metabolic variation among many different individuals. Third, it is unlikely that one gene or one set of genes is responsible for the phenotypic variation among all individuals. The same phenotype might have different underlying causes. Thus, we should not expect a magic bullet that will cure everyone of a specific disease. Instead, different cures will be necessary for different individuals. Finally, the lack of correlation between gene expression and phenotype is not necessarily due to the lack of importance of gene expression. Instead, if the importance of gene expression is context dependent, then a significant relationship will only be discernable in a specific context.

Most gene expression studies measure treatment effects rather than individual variation. Additionally, the vast majority of microarray studies have examined physiological or dose responses such as responses to temperature, salinity, pollution, specific chemical treatments, stress, disease and age to name a few. Far fewer microarray studies have looked at adaptive changes in gene expression in natural fish populations (or in any natural population, Table III).

**EVOLUTIONARY GENOMICS**

For adaptive studies, one can use an evolutionary or phylogenetic perspective to reveal changes that are produced by natural selection and therefore indicate whether the variation affects longevity, reproductive fitness or probability of survival, *i.e.* whether the variation is biologically important. Specifically, if the variation is neutral, the pattern of variation will be random and thus will be correlated with evolutionary distance among taxa rather than significantly correlated with important environmental variables (Pierce & Crawford, 1997; Crawford et al., 1999; Oleksiak et al., 2002). In contrast, if the pattern of variation is non-random and most probably due to evolution by natural selection, it will correlate significantly with important
Table III. Examples of studies using microarrays or ESTs to identify adaptively important changes in gene expression

<table>
<thead>
<tr>
<th>Species</th>
<th>Array type</th>
<th>Tissue</th>
<th>Comparison</th>
<th>Major findings</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Fundulus heteroclitus</em> (killfish)</td>
<td>cDNA</td>
<td>Heart</td>
<td>Northern population v. a southern population and sister taxa from the Gulf, <em>Fundulus grandis</em></td>
<td>Gene expression for 20% of genes varied significantly between individuals. 15 genes (1.5%) with adaptive signature perhaps due to temperature</td>
<td>Oleksiak <em>et al.</em> (2002)</td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>Brain and liver</td>
<td>Three chronically polluted populations, each with two flanking reference populations</td>
<td>Gene expression for 8–32% of genes showed potentially adaptive changes in gene expression. Few significant genes shared among different polluted populations</td>
<td>Fisher &amp; Oleksiak (2007) and Oleksiak (2008)</td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>Heart</td>
<td>Clinal variation among five populations</td>
<td>Much of variation in gene expression was neutral, but 22% of gene expression was potentially adaptive</td>
<td>Whitehead &amp; Crawford (2006)</td>
</tr>
<tr>
<td><em>Platichthys flesus</em> (European flounder)</td>
<td>cDNA</td>
<td>Liver</td>
<td>Baltic Sea v. North Sea population</td>
<td>5% (158) of genes were significantly differently expressed in North Sea and Baltic Sea flounders subjected to the same salinities under controlled conditions</td>
<td>Larsen <em>et al.</em> (2007)</td>
</tr>
<tr>
<td><em>Salmo salar</em> (Atlantic salmon)</td>
<td>cDNA</td>
<td>Brain</td>
<td>Sneaker males v. age-matched females and immature males</td>
<td>15% of genes differed between sneaker and immature males</td>
<td>Aubin-Horth <em>et al.</em> (2005)</td>
</tr>
<tr>
<td><em>Neolamprologus pulcher</em> (Tanganyikan cichlid)</td>
<td>cDNA</td>
<td>Brain</td>
<td>Dominant males and females v. subordinate males and females</td>
<td>Dominant breeder females were masculinized at the molecular and hormonal level</td>
<td>Aubin-Horth <em>et al.</em> (2007)</td>
</tr>
<tr>
<td><em>Coregonus</em> sp. (whitefish)</td>
<td>cDNA</td>
<td>Liver</td>
<td>Dwarf v. normal whitefish populations</td>
<td>Significant overexpression of genes potentially associated with enhanced activity, and downregulation of genes associated with growth in dwarf whitefish</td>
<td>St-Cyr <em>et al.</em> (2008)</td>
</tr>
<tr>
<td><em>Dissostichus mawsoni</em> (Antarctic toothfish)</td>
<td><em>In silico</em></td>
<td>Brain, liver, head kidney and ovary</td>
<td>Antarctic fish v. five temperate–tropical species</td>
<td>177 notothenioid protein families highly expressed compared to the temperate species. Significant Antarctic specific gene duplications</td>
<td>Chen <em>et al.</em> (2008)</td>
</tr>
</tbody>
</table>
environmental variables and be independent of evolutionary distance between taxa. Thus, studying signatures of selection in natural populations has been used to discern biologically important variation in studies from flies to fish to men (Endler, 1986).

ADAPTIVE MICROARRAY STUDIES IN NATURAL FISH POPULATIONS

*Fundulus heteroclitus* provide a powerful system to explore evolved changes in gene expression because many populations show adaptations (Burnett et al., 2007), $N_e$ in local population sizes is generally $>10^5$, allowing evolution by natural selection with relatively small selection coefficients to dominate evolution by random drift (Li, 1997), and migration among populations is sufficient to minimize neutral divergence but does not appear to overcome local adaptation (Brown & Chapman, 1991; Powers et al., 1993; Fisher & Oleksiak, 2007). *Fundulus heteroclitus* also has a well-described phylogeny, and this was used to explore adaptive changes in gene expression among taxa. Oleksiak et al. (2002) compared heart gene expression in a northern population of *F. heteroclitus* to that in a southern population of *F. heteroclitus* and the sister taxa *Fundulus grandis* Baird & Girard. The expectation is that the northern and southern populations should share more similarities with each other than either has to the sister taxa, because they are more closely related. A certain subset of genes, however, had expression patterns that were significantly different in the northern population from both the southern population and *F. grandis*, suggesting adaptive differences in gene expression, perhaps due to temperature (Oleksiak et al., 2002).

A similar approach was taken with *F. heteroclitus* populations that inhabit highly polluted Superfund sites (Superfund sites contain high levels of a variety of lipophilic, persistent and toxic contaminants worthy of being remediated using U.S. Federal funds.) and have adapted to the pollutants in their environment. Instead of using phylogeny, this experiment used triads of comparisons: each polluted site was compared to two flanking reference sites (Fisher & Oleksiak, 2007; Oleksiak, 2008). This is experimentally robust because the genetic distance between the two clean reference populations is greater than the genetic distance between the polluted population and either reference population. This triad comparison also controls for clinal effects. Because *F. heteroclitus* inhabit a temperature cline along the Atlantic seaboard, if only two populations were compared, differences between the two populations could be due to temperature differences rather than pollution or genetic drift. By comparing the polluted population to both a northern and a southern population, effects of pollution are no longer confounded with either temperature or genetic drift. Thus, divergence in a polluted population compared to both paired reference populations suggests that pollution might be causative. That is, if the differences are shared with both reference populations (*i.e.* polluted *v.* both reference 1 and reference 2), those shared differences are more probably to be due to pollution. This experimental design was used to quantify gene expression in both brain and liver from *F. heteroclitus* collected from polluted sites and found a number of significantly altered gene expression patterns that probably evolved due to chronic exposure to pollution (Fisher & Oleksiak, 2007; Oleksiak, 2008).

Even relatively weak signals of population structure have been suggested to represent evolutionary units able to adapt to the local environment causing fine-scale
spatial and temporal divergence (Conover et al., 2006), and adaptive differences in gene expression were found among P. flesus with genetically weakly differentiated populations ($F_{ST}$ ranging from 0·005 to 0·006). Platichthys flesus populations distributed across the salinity gradient stretching from the brackish Baltic Sea to the high saline North Sea showed that 5% of genes expressed in the liver were significantly differently expressed in Baltic Sea and North Sea P. flesus under common or garden conditions (Larsen et al., 2007). This study was followed by tissue specific (gill and kidney) gene expression of four candidate genes. Hsp70 was upregulated six-fold in kidney of North Sea P. flesus following long-term acclimation to Baltic Sea salinities, and significant differences in expression of 5-aminolevulinic acid synthase (ALAS) in response to different salinities in gill and liver tissue were found between North Sea and Baltic Sea P. flesus (Larsen et al., 2008). The identification of adaptive differences even among high gene flow marine organisms has important implications for management and protection of marine biodiversity.

Microarrays also have been used to explore life-history traits in S. salar, cichlids and Coregonus clupeaformis (Mitchill). Male S. salar either attain sexual maturity precociously (1–3 years) and become sneaker males or migrate out to sea and return 3–7 years later as large, anadromous animals (Fleming, 1998). A sneaker male invades the larger male’s territory during spawning and ejects sperm towards the female’s newly extruded eggs while the larger male is distracted by his own spawning. Although this tactic fertilizes eggs poorly in comparison to the large males, sneaker males have a greater probability of surviving to reproduce such that the lifetime fitness of smaller males is similar to that of the large, anadromous males (Gross, 1985). This developmental plasticity of sneaker males was explored by comparing gene expression profiles in mature sneaker males and age-matched immature males and females. Fifteen percent of genes were significantly differently expressed in the brain of mature sneaker males compared with immature males. Not surprisingly, reproduction associated genes were almost exclusively sneaker biased. The importance of the gene expression differences was further evidenced by a cluster analysis that separated the different phenotypes (Aubin-Horth et al., 2005).

In cichlids, microarrays were used with 14 established social groups of the Tanganyikan cichlid Neolamprologus pulcher (Trewavas & Poll) to profile brain gene expression of dominant females. Gene expression in dominant breeder females was most similar to that of the males (independent of social rank), and dominant breeder females also were masculinized at the hormonal level (Aubin-Horth et al., 2007). Thus, while cichlids are an emerging evolutionary genomic model system for fundamental questions on the origin of phenotypic diversity, cooperatively breeding species such as N. pulcher provides a powerful model system to further our understanding of the evolution of social behaviour.

Dwarf and normal C. clupeaformis populations have morphological, behavioural, ecological and life-history traits that most probably represent adaptations towards exploiting distinct trophic resources (Fenderson, 1964; Bernatchez et al., 1999; Lu & Bernatchez, 1999; Rogers et al., 2002). These reproductively isolated, C. clupeaformis phenotypes occurred in many lakes following ice cap retreat after the Wisconsin glaciations 15 000 years ago. Gene expression was explored in dwarf and normal C. clupeaformis from two natural lakes, as well as from populations reared in controlled environments. Dwarf C. clupeaformis consistently showed significant overexpression of genes potentially associated with survival through
enhanced activity (energy metabolism, iron homeostasis, lipid metabolism, detoxification) and downregulation of genes associated with growth (protein synthesis, cell cycle, cell growth) in dwarf relative to normal Coregonus sp. (St-Cyr et al., 2008). The authors suggest that enhanced survival via a more active swimming, necessary for increased foraging and predator avoidance, entails energetic costs that result in slower growth rate and reduced fecundity in dwarf relative to normal coregonids. These three studies illustrate the power of microarrays for exploring questions of life-history evolution.

Finally, a recent genomic, gene expression study was done in silico. Gene expression of the Antarctic notothenioid fish Dissostichus mawsoni (Norman) was based on transcript abundance from an EST-sequencing project of non-normalized cDNA libraries and compared to same tissue transcript abundances of five temperate–tropical fish species in the database. In all, 177 notothenioid protein families were highly expressed compared to the temperate species. The genomic basis for these upregulated genes was explored using comparative genomic hybridization of DNA from four pairs of Antarctic and basal non-Antarctic notothenioids. Importantly, significant Antarctic specific gene duplications occurred for 118 genes, many corresponding to the upregulated genes identified in silico (Chen et al., 2008).

MARKERS FOR POPULATION GENETICS–GENOMICS

Most genomic and genetic studies in natural populations use one or a few molecular markers to assess population parameters (Palumbi, 1994; McMillan & Palumbi, 1995; Williams et al., 2001; Barber et al., 2002; Crandall et al., 2008). Allozymes have long been used due to the ease of use across species (Nevo, 1990). The limited number of loci and low variability of allozymes, however, limit the statistical power of the resulting data. Mitochondrial DNA (mtDNA) was the first widely used DNA marker and has provided many insights into the demography of natural populations (Avise et al., 1986; Saunders et al., 1986; Bowen et al., 2007), but because mtDNA is a single locus, its ability to resolve population structure is relatively limited (Avise, 1994). Most recent genetic studies of natural populations have used microsatellites. The high variation in many microsatellites provides high statistical power for population genetics, but they suffer from two drawbacks. First, they require species-specific marker development, and second, they suffer from a high potential for null alleles and are prone to genotyping errors due to their size-based nature (homoplasy). These shortcomings and the fact that analysis of microsatellite loci requires experience and expertise make data comparisons among laboratories and with published data problematic (Hoffman & Amos, 2005).

Amplified fragment length polymorphisms (AFLP) have been widely used since first described (Vos et al., 1995) due to their ease of use in species with no prior sequence information: many AFLP markers can be easily amplified and scored. AFLP analyses, however, require high-quality DNA and provide dominant markers so that heterozygotes cannot be directly measured. Additionally, AFLPs are anonymous, and thus loci can be scored incorrectly due to their size-based nature (the same sized band can represent multiple loci).

Single nucleotide polymorphisms (SNP) have recently attracted the interest of geneticists due to their potential for automation and high-throughput as well as
lower genotyping errors than microsatellites (Hauser & Seeb, 2008). In addition, SNP analyses can use DNA of varying quality. Furthermore, as they provide sequence data, the data are directly transferable among different laboratories. Yet, like microsatellites, SNPs require species-specific development for non-model species. Thus, although SNPs are the most abundant marker in most genomes, discovery of useful SNPs can be challenging in organisms with no genomic information, such as many fish species. One approach to identify SNPs is to mine large, expressed sequence tag (EST), data collections for putative SNPs.

**POPULATION-GENOMIC APPROACHES**

Although microarray approaches can identify significant changes in gene expression that are potentially adaptively important, a thorough understanding of the genomic basis underlying phenotypic adaptive divergence is still emergent (MacCallum & Hill, 2006). Population genomics is one approach to better understand the genomic basis of adaptive divergence. Population genomics is more than population genetic analyses on a large number of loci (Stinchcombe & Hoekstra, 2008) because this large number of loci makes statistical tests for non-random patterns possible (Nielsen, 2001; Schlotterer, 2002; Luikart et al., 2003; Storz, 2005). The basic tenet is that local adaptation and directional selection should reduce genetic variability within populations and increase variation among populations at the locus under selection. One can use a variety of statistical tests to ask whether an individual locus acts as an outlier compared to loci in the rest of the genome (Storz, 2005). The application of these approaches is useful for understanding adaptation in genetically uncharacterized natural populations (Pogson et al., 1995; Beaumont & Nichols, 1996; Yan et al., 1999; Allendorf & Seeb, 2000; Bradshaw et al., 2000; Frary et al., 2000; Parsons & Shaw, 2001; Peichel et al., 2001b; Mock et al., 2002; Kohn et al., 2003; Storz & Nachman, 2003; Storz & Dubach, 2004) and should yield insights into long-standing evolutionary questions.

Population genomics first requires genotyping a large number of loci. These loci might be microsatellites, AFLPs, SNPs or sequences. Without a genome, AFLPs are the easiest and cheapest method to genotype large numbers of loci. With new sequencing technologies, however, other markers are likely to be as available. The advantages of population-genomics approaches are the ease of generating large numbers of genetic markers, the ability to scan the genome without measuring phenotypes and the ability to sample individuals without knowledge of their breeding history (Stinchcombe & Hoekstra, 2008). A big disadvantage of AFLPs as well as many SNPs and microsatellites used in population genomics is that they are anonymous or in non-functional sequence; hence, linking a candidate locus to a causative gene or mutation is improbable without further analyses or experiments.

AFLP analyses have long been used to explore genetic diversity, but it is only recently that they have been used to identify loci indicative of adaptation in natural fish populations. To date, only a handful of studies using anonymous genome-wide markers have analyzed multiple populations to test for conserved outlier loci (Wilding et al., 2001; Campbell & Bernatchez, 2004; Vasmagi et al., 2005; Bonin et al., 2006; Williams & Oleksiak, 2008). Three of these were in natural fish populations: _S. salar_ from fresh, brackish and saltwater habitats in the Barents and White Seas,
dwarf and normal ecotypes of *C. clupeaformis* and pollution-adapted populations of *F. heteroclitus*. Although shared differences among populations are suggestive of non-random patterns, they still could be demographic (e.g. due to a bottleneck). Population genomics allows one to test whether the loci are significantly different in population A v. population B and also are significantly different from all other loci in population A. That is, with hundreds of loci, one makes the assumption that a majority of differences among loci are either non-functional (a polymorphism with no phenotypic effect) or neutral (variation whose phenotypic effect has no or little fitness consequences, *i.e.* selective advantage is \(<1/2N_e\)). Thus one looks for outliers among these loci: loci that have significantly different divergence v. a random permutation or modelling of the data.

Genomic and EST-derived microsatellites were screened for patterns of divergent selection among eight *S. salar* populations from fresh, brackish and saltwater habitats in the Barents and White Seas. This study found a high percentage of differentiated loci (12%, nine of 78 EST-linked loci) suggesting that EST-associated markers could improve the efficiency of whole genome scans (Vasemagi *et al.*, 2005). AFLPs were used to compare dwarf and normal ecotypes in *C. clupeaformis*, and found 1.4% differentiated loci (Campbell & Bernatchez, 2004). AFLPs also were used to compare pollution-adapted populations of *F. heteroclitus* in a similar experimental design to that used for microarray studies discussed earlier. After correcting for multiple comparisons, 1–6% of loci were implicated as being under selection or linked to areas of the genome under selection (Williams & Oleksiak, 2008).

Note, however, that even when both phenotypes and selective environments are very similar among populations, this does not mean that the same loci will be fixed in response to similar environmental conditions (Hoekstra & Nachman, 2003; Hoekstra, 2006). On an evolutionary scale, this is similar to the physiological results where different patterns of gene expression explain substrate specific metabolism (Oleksiak *et al.*, 2005). This suggests a biological complexity that we are only beginning to understand.

**CONCLUSIONS AND FUTURE DIRECTIONS**

Lewontin & Hubby (1966) first quantified genetic variation in different fly populations and were surprised to find 30% of loci had two or more alleles. This level of polymorphism was unexpected. With the advent of whole genome sequencing projects, we can more definitively quantify genetic variation and do it in a diversity of species to reveal evolutionary adaptive patterns. Even among organisms without sequenced genomes, genomic approaches including microarray studies and SNP identification can reveal substantial interindividual variation. Again, a critical question is: how much of this variation is biologically important? The first vertebrate studies using microarrays to measure individual variation in gene expression found 20–95% of genes had significant variation among individuals (Oleksiak *et al.*, 2002, 2005). A recent study in flies corroborated this amount of individual variation in gene expression and showed further, that 68% of it was heritable (Ayroles *et al.*, 2009). With natural populations, natural selection is one way to assess biological importance, and discerning evolution by natural selection is more readily accomplished.
using the rich species diversity among teleost fishes and the depth of environmental knowledge about these species.

Having a full-genome sequence provides many resources to better understand the genetics and molecular mechanisms affecting change in gene frequency or gene expression. Yet, with the advances of newer high-throughput sequencing approaches, genomic approaches can be applied to most species to yield genome-wide information. As discussed, high-throughput sequencing technologies have been used to sequence transcripts for gene expression studies but also can be used for ampli-con resequencing for cost-effective SNP discovery (Bundock et al., 2009), genomic DNA SNP discovery (Wiedmann et al., 2008), microsatellite discovery by deep sequencing of microsatellite-enriched genomic libraries (Santana et al., 2009), mitochondrial sequencing (Cui et al., 2009) and targeted resequencing of particular genes or gene regions (Hodges et al., 2007; Okou et al., 2007). Resulting tools coupled with genomic analyses can give insights into how individuals and populations evolve. Thus, although most fishes do not have full-genome sequences and will not until the U.S. National Institute of Health’s $1000 genome goal becomes reality, the potential for researchers to exploit powerful genomic tools in their studies of fish species is ever expanding.

The author thanks D. L. Crawford at the University of Miami for helpful comments on this manuscript. Funding for this work was received from NIH 5 RO1 ES011588, 2P42 ES010356 and 2 P42 ES007381.

References


**Electronic Reference**