Part 1

Broodstock Improvement
Chapter 1
Genomic Tools for Understanding the Molecular Basis of Production-Relevant Traits in Finfish

Marije Booman and Matthew L. Rise

OVERVIEW
Significant genomic resources (e.g., expressed sequence tag (EST) databases, DNA microarrays, single nucleotide polymorphism (SNP) genotyping platforms, bacterial artificial chromosome (BAC) libraries and BAC end sequences, genetic linkage maps, and physical maps) have been generated for several finfish species of importance to global aquaculture. Over the last few years, numerous articles (e.g., Cerdá et al. 2008; Koop et al. 2008), reviews (e.g., Douglas et al. 2006; Canario et al. 2008; Goetz and MacKenzie et al. 2008; Martin et al. 2008), and book chapters (e.g., Palti 2009; Rise et al. 2009; and several chapters in book Aquaculture Genome Technologies, 2007, edited by Z. Liu) have been published on the creation and application of finfish genomics resources. With the advent of next-generation sequencing (NGS) technologies, it is anticipated that finfish genomic resources will continue to expand.

For species of key importance to global aquaculture and fisheries (e.g., Atlantic salmon (Salmo salar), rainbow trout (Oncorhynchus mykiss), Atlantic cod (Gadus morhua), channel catfish (Ictalurus punctatus), and common carp (Cyprinus carpio)), complete "genomics toolboxes" will be needed in order to take full advantage of the power of genomics in aquaculture research (e.g., for marker-assisted selection (MAS) of superior broodstock, development of optimal and sustainable feed formulations, development of maximally effective vaccines and therapeutants, etc.). Complete, high-quality reference genome sequences are critically important components of these toolboxes, and whole-genome sequencing projects are already underway for Atlantic salmon (Davidson et al. 2010) and catfish (Lu et al. 2011). As aquaculture finfish species' genomes are sequenced and assembled, and as we move into a postgenomics era for these species, bioinformatics will undoubtedly play an ever-increasing role in the success of aquaculture genomics research.

Because of the aforementioned publications on aquaculture finfish genomic resources, we do not attempt to provide an exhaustive summary of the field. Rather, the aim of this chapter is to cover new territory. Since we have been involved in the Atlantic Cod Genomics and Broodstock Development Project (CGP, http://codgene.ca), we use CGP
examples to illustrate different aspects of finfish functional genomics tools. In Section “Targeted, Trait-Relevant Gene Discovery,” we discuss CGP examples of how suppression subtractive hybridization (SSH) cDNA libraries were employed for discovery of Atlantic cod genes that were involved in defense responses. In Section “The Application of Microarray Technology in Finfish Aquaculture and Research,” we discuss microarray technology, including brief overviews of current finfish microarray platforms, the construction of the CGP Atlantic cod oligonucleotide microarray focusing on trait-relevant transcripts, and software for the analysis of microarray data. We discuss how these functional genomics tools will likely play key roles in genomics toolboxes by allowing the identification of the best targets for MAS of superior aquaculture broodstock. We end the chapter with recommendations and future directions for finfish microarray (and other genomics) research.

**TARGETED, TRAIT-RELEVANT GENE DISCOVERY**

Normalized (e.g., Rise et al. 2004b; Gahr et al. 2007), unnormalized (e.g., Goetz et al. 2009), and SSH (e.g., Rise et al. 2008, 2010; Feng et al. 2009; Hori et al. 2010) cDNA libraries have been utilized in large-scale gene discovery efforts for several aquaculture finfish and shellfish species (see Chapters 2 (by Robalino et al.), 7 (by Douglas), and 8 (by Dixon and Becker) of this book for information on targeted gene discovery in aquaculture species). The numbers of ESTs currently in GenBank for selected aquaculture fish species are shown in Table 1.1.

For each aquaculture-relevant finfish species, ESTs are generally contributed from several researchers or genomics projects (e.g., see Rise et al. 2007 for information on origins of salmonid EST resources). Each of the large-scale aquaculture finfish genomics research projects to date has employed a particular strategy for cDNA library

### Table 1.1. Expressed Sequence Tag (EST) Collections of Selected Aquaculture Fish Species.

<table>
<thead>
<tr>
<th>Selected Orders of Teleosts</th>
<th>Examples of Aquaculture Species With Genomic Resources</th>
<th>Number of ESTs in GenBank*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superorder Acanthopterygii</td>
<td>Nile tilapia (<em>Oreochromis niloticus</em>)</td>
<td>117,222</td>
</tr>
<tr>
<td>Perciformes</td>
<td>Gilthead seabream (<em>Sparus aurata</em>)</td>
<td>67,670</td>
</tr>
<tr>
<td></td>
<td>European seabass (<em>Dicentrarchus labrax</em>)</td>
<td>55,835</td>
</tr>
<tr>
<td></td>
<td>Yellow perch (<em>Perca flavescens</em>)</td>
<td>21,968</td>
</tr>
<tr>
<td>Superorder Ostariophysi</td>
<td>Channel catfish (<em>Ictalurus punctatus</em>)</td>
<td>354,466</td>
</tr>
<tr>
<td>Siluriformes</td>
<td>Blue catfish (<em>Ictalurus furcatus</em>)</td>
<td>139,475</td>
</tr>
<tr>
<td></td>
<td>Common carp (<em>Cyprinus carpio</em>)</td>
<td>34,067</td>
</tr>
<tr>
<td>Superorder Paracanthopterygii</td>
<td>Atlantic cod (<em>Gadus morhua</em>)</td>
<td>229,090</td>
</tr>
<tr>
<td>Gadiformes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superorder Protacanthopterygii</td>
<td>Atlantic salmon (<em>Salmo salar</em>)</td>
<td>498,212</td>
</tr>
<tr>
<td>Salmoniformes</td>
<td>Rainbow trout (<em>Oncorhynchus mykiss</em>)</td>
<td>287,967</td>
</tr>
<tr>
<td></td>
<td>Chinook salmon (<em>Oncorhynchus tshawytscha</em>)</td>
<td>14,173</td>
</tr>
<tr>
<td></td>
<td>Sockeye salmon (<em>Oncorhynchus nerka</em>)</td>
<td>11,389</td>
</tr>
<tr>
<td></td>
<td>Coho salmon (<em>Oncorhynchus kisutch</em>)</td>
<td>4,942</td>
</tr>
</tbody>
</table>

*Numbers of ESTs were acquired from NCBI Entrez Taxonomy on November 4, 2010.
construction and gene discovery. For example, the CGP generated over 138,000 ESTs from a total of 23 normalized libraries and over 20,000 ESTs from a total of 19 SSH libraries (Bowman et al. 2011). Of the ~158,000 ESTs generated by the CGP, approximately 45,000 were from cDNA libraries that were designed for the discovery of cod transcripts that responded to pathogens or pathogen-associated molecular patterns (PAMPs) (Bowman et al. 2011); of these, approximately 10,000 ESTs came from cod SSH libraries that were enriched for transcripts that respond to nodavirus and/or polyribinosinic polyribocytidylic acid (pIC, a synthetic double-stranded RNA that induces interferon pathway genes; Rise et al. 2008, 2010; Bowman et al. 2011), and over 4300 ESTs came from cod SSH libraries enriched for transcripts that respond to bacterial antigens (Feng et al. 2009; Fig. 1.1). Many of the transcripts

![Figure 1.1](image-url)

**Figure 1.1.** Functional genomics workflow in the Atlantic Cod Genomics and Broodstock Development Project (CGP). Suppression subtractive hybridization (SSH) libraries were designed for the targeted discovery of cod transcripts that respond to virus (Rise et al. 2008, 2010), bacterial antigens (Feng et al. 2009), or heat stress (Bowman et al. 2011; Hori et al. 2010). Many trait-relevant genes identified in SSH libraries were included on the CGP 20K oligonucleotide microarray (Booman et al. 2011), making this a valuable new resource for defense-relevant functional genomics research on Atlantic cod. Quantitative reverse transcription – polymerase chain reaction (QPCR) assays were developed and used to study transcript expression of many SSH-identified cod genes (e.g., Feng et al. 2009; Rise et al. 2008, 2010; Booman et al. 2011; Hori et al. 2010). In addition, several SSH-identified defense-relevant transcripts (e.g., interferon regulatory factor 1 (IRF1) and anti-apoptotic Bcl-2 family genes NR-13, Mcl-1, and Bcl-X1) were further characterized using bidirectional rapid amplification of cDNA ends (RACE) and other molecular techniques. (From Feng et al. (2009) and Feng and Rise (2010).)
identified in the CGP’s cDNA libraries designed for targeted discovery of immune-relevant genes were included on the CGP 20,000 gene (20K) oligonucleotide microarray (Booman et al. 2011; see section of this chapter entitled “Design of a New Microarray for Aquaculture Research: The Atlantic Cod 20K Oligonucleotide Microarray”), making this microarray platform a valuable new tool for functional genomics research on Atlantic cod defense responses.

An obvious goal of any large-scale, gene discovery project is to identify all of the genes in the species of interest. Although the EST collections of some aquaculture finfish species are quite large (e.g., >200,000 ESTs for Atlantic salmon, channel catfish, rainbow trout, and Atlantic cod; Table 1.1), it is likely that they are still missing important, trait-relevant transcripts. For example, if high-quality cDNA libraries enriched for gill transcripts that are responsive to infection by a fungal pathogen are not incorporated into a finfish species’ genomics research plan, then that species’ EST collection may be missing defense-relevant transcripts that respond specifically to fungi. The same may be said of species for which cDNA libraries are quite large (e.g., salmonids).

THE APPLICATION OF MICROARRAY TECHNOLOGY IN FINFISH AQUACULTURE AND RESEARCH

The establishment of large EST collections (e.g., Table 1.1) has facilitated the development of finfish DNA microarrays that have become important tools in finfish aquaculture genomics research. In this section, we give an overview of the microarray platforms that are currently described in the literature. We briefly discuss the development of the CGP 20K Atlantic cod oligonucleotide microarray as an example of a platform focusing on trait-relevant transcripts for aquaculture research. Next, we discuss tools for microarray data analysis. Finally, we end this section with future perspectives and some recommendations for future microarray research.

Overview of Microarray Platforms and Their Use in Aquaculture

A search through NCBI’s PubMed and Gene Expression Omnibus (GEO) databases shows that there are over 40 finfish species that have been studied using microarrays and that research areas vary from biomedicine and ecotoxicology to nutrigenomics and immunology. The largest numbers of microarray platforms are available for zebrafish and salmonids. Some microarray platforms are most useful in specific research areas as they are limited to transcripts from specific tissues or developmental stages, or contain only transcripts involved in specific functions or transcripts responsive to a specific treatment or condition. Other microarray platforms represent the transcriptome of multiple tissues, developmental stages, treatments, and conditions and are useful for a wider range of research areas. These platforms are often made available either through companies like Agilent, Affymetrix, and NimbleGen or through research institutes, as with the consortium for Genomics Research on All Salmonids (cGRASP) arrays (Rise et al. 2004b; von Schalburg et al. 2005; Koop et al. 2008) and the CGP 20K Atlantic cod oligonucleotide microarray (Booman et al. 2011). Table 1.2 gives an overview of selected finfish species and microarray platforms on the basis of their economic importance and their usefulness in aquaculture research.

Microarrays have successfully been used to study multiple aspects of finfish biology that are of interest to aquaculture. Examples are the response to stress such as hypoxia (e.g., van der Meer et al. 2005), temperature changes (e.g., Vornanen et al. 2005; Chou et al. 2008), or handling (e.g., Krasnov et al. 2005; Wiseman et al. 2007); the immune response to different types of pathogens such as bacteria (e.g., Rise et al. 2004a; Gerwick et al. 2007), viruses (e.g.,
Table 1.2. Overview of Selected Finfish Microarray Platforms.

<table>
<thead>
<tr>
<th>Order</th>
<th>Species</th>
<th>Technology</th>
<th>Microarray Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cypriniformes</td>
<td><em>Cyprinus carpio</em> (common carp)</td>
<td>Spotted cDNA</td>
<td>CarpARRAY v5 26K (Williams et al. 2008)</td>
</tr>
<tr>
<td></td>
<td><em>Danio rerio</em> (zebrafish)</td>
<td>Spotted cDNA</td>
<td>Compugen/Sigma-Genosys 16K (e.g., Malek et al. 2004; Rawls et al. 2004); MWG 14K (e.g., van der Meer et al. 2005; Chou et al. 2008); Compugen/Sigma-Genosys/MWG/Operon 35K (e.g., Pei et al. 2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In situ oligo</td>
<td>Leiden custom Agilent 44K (Stockhammer et al. 2009; GPL7735); Agilent zebrafish 21K (GPL7244); Agilent zebrafish 44K (GPL3701); Affymetrix zebrafish 15K (GPL1319); Nimblegen zebrafish 32K (GPL5746)</td>
</tr>
<tr>
<td>Gadiformes</td>
<td><em>Gadus morhua</em> (Atlantic cod)</td>
<td>Spotted cDNA</td>
<td>CodStress 0.7K (Lie et al. 2009); IMR Atlantic cod 16K (Edvardsen et al. 2011)</td>
</tr>
<tr>
<td>Perciformes</td>
<td><em>Dicentrarchus labrax</em> (European sea bass)</td>
<td>Spotted cDNA</td>
<td>CGP Atlantic cod 20K v1.0 (Booman et al. 2011; GPL 10532)</td>
</tr>
<tr>
<td></td>
<td><em>Sparus aurata</em> (gilthead seabream)</td>
<td>Spotted cDNA</td>
<td>INRA-Agenae rainbow trout 9K (Govoroun et al. 2006; Darias et al. 2008; GPL3650)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In situ oligo</td>
<td>Padova custom Agilent 40K (Ferraresso et al. 2008; GPL6467)</td>
</tr>
<tr>
<td>Pleuronectiformes</td>
<td><em>Hippoglossus hippoglossus</em> (Atlantic halibut)</td>
<td>Spotted cDNA</td>
<td>Pleurogene halibut 10K (Douglas et al. 2008; GPL6361)</td>
</tr>
<tr>
<td></td>
<td><em>Paralichthys olivaceus</em> (Japanese flounder)</td>
<td>Spotted cDNA</td>
<td>Immune-related 0.9K (Kurobe et al. 2005); immune-related v2 1.1K (Matsuyama et al. 2007); custom digital genomics 0.3K (Nakayama et al. 2008)</td>
</tr>
<tr>
<td></td>
<td><em>Psetta maxima</em> (turbot)</td>
<td>Spotted cDNA</td>
<td>Virus-stimulated 2K (Park et al. 2009)</td>
</tr>
<tr>
<td></td>
<td><em>Solea senegalensis</em> (Senegalese sole)</td>
<td>Spotted cDNA</td>
<td>Pleurogene Sole 5K (Cerdá et al. 2008)</td>
</tr>
<tr>
<td>Salmoniformes</td>
<td><em>Oncorhynchus mykiss</em> (rainbow trout)</td>
<td>Spotted cDNA</td>
<td>Stress-responsive SFA salmonid V1 1.3K (Krasnov et al. 2005; GPL1212)<em>; Immunochip SFA salmonid V2 1.8K (Jørgensen et al. 2008; GPL6154)</em>; cGRASP salmonid 3.6K (Rise et al. 2004b; GPL966)<em>; cGRASP salmonid v2 16K (von Schalburg et al. 2005; GPL2716)</em>; cGRASP salmonid 32K (Koop et al. 2008; GPL8904)*; INRA-Agenae rainbow trout 9K (Govoroun et al. 2006; GPL3650); stress-responsive UWVJLAB RT 150 0.1K (Wiseman et al. 2007; GPL3713)</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Order</th>
<th>Species</th>
<th>Technology</th>
<th>Microarray Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Oncorhynchus tshawytscha</em> (Chinook salmon)</td>
<td>Spotted cDNA</td>
<td>Stress-responsive OSUrbt v2 1.6K (Tilton et al. 2005; GPL5478); Custom rainbow trout 21.5K (Olohan et al. 2008)</td>
</tr>
<tr>
<td></td>
<td><em>Salmo salar</em> (Atlantic salmon)</td>
<td>In situ oligo</td>
<td>Custom Agilent rainbow trout 37K (Salem et al. 2008; GPL8205); Gothenburg rainbow trout 15.5K (Gunnarsson et al. 2009; GPL8254)</td>
</tr>
<tr>
<td></td>
<td><em>Oncorhynchus tshawytscha</em> (Chinook salmon)</td>
<td>In situ oligo</td>
<td>cGRASP salmonid v2 16K (von Schalburg et al. 2005; GPL2716)†</td>
</tr>
<tr>
<td></td>
<td><em>Salmo salar</em> (Atlantic salmon)</td>
<td>In situ oligo</td>
<td>Agilent Atlantic salmon 44K (GPL7303)</td>
</tr>
<tr>
<td></td>
<td><em>Ictalurus punctatus</em> (channel catfish)</td>
<td>In situ oligo</td>
<td>Custom Nimblegen channel catfish 19K (Li and Waldbieser 2006; GPL2814); Custom Nimblegen channel catfish 28K (Peatman et al. 2007; GPL4476)†</td>
</tr>
<tr>
<td></td>
<td><em>Ictalurus furcatus</em> (blue catfish)</td>
<td>In situ oligo</td>
<td>Custom Nimblegen catfish 28K (Peatman et al. 2007; GPL4476)†</td>
</tr>
<tr>
<td></td>
<td><em>Siluriformes</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


*Note: This table lists the microarray platforms that are of interest to aquaculture research. Microarray platforms are grouped by order, species, and technology (spotted cDNA, spotted oligonucleotide, or in situ synthesized oligonucleotide). References include the publication describing the microarray design and, where available, the GEO platform accession number (in “GPLxxxx” format).

†Combined *I. furcatus* and *I. punctatus*.
Functional Genomic Tools for Aquaculture

MacKenzie et al. 2008; Workenhe et al. 2009), or parasites (e.g., Morrison et al. 2006; Skugor et al. 2008; Young et al. 2008); the influence of nutrition (e.g., Leaver et al. 2008; Murray et al. 2010); and gene expression profiles related to egg quality (e.g., Bonnet et al. 2007) and growth (e.g., Rise et al. 2006; Gahr et al. 2008; Devlin et al. 2009). As previously mentioned, there have been several reviews and chapters on the applications of finfish DNA microarrays (e.g., Douglas 2006; Rise et al. 2007, 2009; Goetz and MacKenzie et al. 2008). Other chapters in this book review literature pertaining to the use of finfish microarrays to study fish egg and embryonic gene expression (Chapter 11, Traverso et al.) and fish immune responses (Chapter 6, Johnson and Brown). We refer the reader to these sources for reviews of the microarray-related aquaculture literature.

Design of a New Microarray for Aquaculture Research: The Atlantic Cod 20K Oligonucleotide Microarray

For most aquaculture-relevant finfish species, a whole-genome sequence is unavailable, and sequences for microarray construction are provided by EST collections. Therefore, the possibilities for application of a newly developed microarray platform in aquaculture research largely depend on the characteristics of the underlying EST collection and the process of selection of sequences from that collection.

An example of a microarray platform that was specifically designed for aquaculture research is the CGP Atlantic cod 20K oligonucleotide microarray platform (Booman et al. 2011). It is based on the CGP EST collection, discussed earlier in this chapter, which contains over 150,000 ESTs (Booman et al. 2011). This EST collection represents 42 cDNA libraries from 12 tissue types and 4 developmental stages; some of the libraries were constructed with tissues sampled from fish that were exposed to stressors or to bacterial or viral immunostimulants. This diversity, together with its size, made the CGP EST collection very well suited for the construction of a microarray that contains a wide range of trait-relevant transcripts for aquaculture research. To prevent redundancy on the microarray and maintain the sequence diversity, the ESTs were first clustered and assembled into unique transcripts, which were then used to design oligonucleotide probes. For the selection of the final set of 20,000 probes, sequences were given priority if they had an informative annotation, if they were represented by a large number of ESTs, or if they were unannotated but represented by a large number of ESTs from the SSH libraries enriched for stress- or immune-responsive transcripts. This selection process ensured not only that the microarray contains enough annotated genes to enable functional interpretation of results but also that it contains unannotated, and therefore possibly novel, genes that have a possible role in traits relevant to aquaculture, such as resistance to stress and pathogens. Within the CGP, this microarray is currently being used in a number of different projects to identify transcripts associated with egg quality, response to heat stress, and response to viral and bacterial immunostimulation.

Microarray Data Analysis Tools

Most data acquisition programs such as Imagene (BioDiscovery) or GenePix (MDS Analytical Technologies) offer basic functionality for data normalization and visualization. For a more in-depth analysis of the acquired microarray data, specialized commercial software packages are available (e.g., Agilent’s GeneSpring GX), which provide full data analysis functionality, including data normalization, identification of differentially expressed genes, and pathway analysis. Most commercial software is relatively easy to use with a good graphical user interface, but it is also expensive and often after the release of a new version, the older versions are no longer supported. Part of the functionality of these packages, such as functional analysis or pathway analysis, relies on the annotation options provided by the package, which are usually limited to model species.

There are noncommercial, free alternatives for most commercial programs (summarized in Table 1.3), such as Significance Analysis of Microarrays (SAM) (Tusher et al. 2001) for class comparison and Gene Set Analysis, and Genesis (Sturn et al. 2002) for clustering. A large source of noncommercial microarray analysis tools is the Bioconductor project, an open source software project “to provide tools for the analysis and comprehension of genomic data” (Gentleman et al. 2004). These tools
Table 1.3. Microarray Data Analysis Noncommercial Software.

<table>
<thead>
<tr>
<th>Program</th>
<th>Reference</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data visualization and processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>marray</td>
<td>Yang and Paquet 2005</td>
<td><a href="http://www.bioconductor.org">www.bioconductor.org</a></td>
</tr>
<tr>
<td>limma</td>
<td>Smyth 2005</td>
<td><a href="http://www.bioconductor.org">www.bioconductor.org</a></td>
</tr>
<tr>
<td>Statistical analysis of gene expression differences</td>
<td>Smyth 2005</td>
<td><a href="http://www.bioconductor.org">www.bioconductor.org</a></td>
</tr>
<tr>
<td>limma</td>
<td>Smyth 2005</td>
<td><a href="http://www.bioconductor.org">www.bioconductor.org</a></td>
</tr>
<tr>
<td>siggenes</td>
<td>Schwender et al. 2006</td>
<td><a href="http://www.bioconductor.org">www.bioconductor.org</a></td>
</tr>
<tr>
<td>multtest</td>
<td>Pollard et al. 2005</td>
<td><a href="http://www.bioconductor.org">www.bioconductor.org</a></td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster/TreeView</td>
<td>Eisen et al. 1998</td>
<td><a href="http://rana.lbl.gov/eisen/?page_id=42">http://rana.lbl.gov/eisen/?page_id=42</a></td>
</tr>
<tr>
<td>Cluster 3.0</td>
<td>de Hoon et al. 2004</td>
<td><a href="http://bonsai.ims.u-tokyo.ac.jp/~mddehoon/software/cluster/">http://bonsai.ims.u-tokyo.ac.jp/~mddehoon/software/cluster/</a></td>
</tr>
<tr>
<td>Java TreeView</td>
<td>Saldanha 2004</td>
<td><a href="http://jtreeview.sourceforge.net/">http://jtreeview.sourceforge.net/</a></td>
</tr>
<tr>
<td>Genesis</td>
<td>Sturn et al. 2002</td>
<td>genome.tugraz.at/genesisclient/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>genesisclient_description.shtml</td>
</tr>
<tr>
<td>Functional analysis of gene lists</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene set enrichment analysis</td>
<td>Subramanian et al. 2005</td>
<td><a href="http://www.broadinstitute.org/gsea/">www.broadinstitute.org/gsea/</a></td>
</tr>
<tr>
<td>DAVID</td>
<td>Huang et al. 2009</td>
<td>david.abcc.ncifcrf.gov/</td>
</tr>
<tr>
<td>GenMAPP</td>
<td>Salomonis et al. 2007</td>
<td><a href="http://www.genmapp.org/">www.genmapp.org/</a></td>
</tr>
<tr>
<td>Blast2GO</td>
<td>Götz et al. 2008</td>
<td><a href="http://www.blast2go.org">www.blast2go.org</a></td>
</tr>
</tbody>
</table>

are based on the statistical computing language R (R Development Core Team 2010). Since most do not have a graphical user interface, they have a steeper learning curve than most commercial software, but there is very strong community-based support. Also, Bioconductor and R tools are customizable, so they could be adapted to work with custom array formats and nonmodel species. Following is a short description of the programs listed in Table 1.3 by category.

**Data Visualization and Processing**

The Bioconductor packages marray (Yang and Paquet 2005) and limma (Smyth 2005) both contain functions for data visualization and normalization. Both packages share the same origins and so there is some overlap, but their functions for data visualization and normalization are mainly complementary. Limma has an optional graphical user interface (limmaGUI; Wettenhall and Smyth 2004). A typical first step in quality control of array data would be to plot several measurements as a color range onto a spatial image (such as signal or background intensity, spot size or shape, or quality flags; the marray package has more options for spatial image plotting than limma). In a spatial image, the layout of the array is shown as a collection of squares, each representing a single spot location. Plotting, for instance, the background signal intensities onto this spatial image gives a quick overview of spatial artifacts present on the array. There are also functions for creating box plots and scatter plots, such as the typical M–A plots, with optional smoothed fits.

Both packages have several options for data normalization. The results from the various plots will help determine what type of normalization is necessary (e.g., a global normalization or a normalization per print-tip).

The limma package has additional functionality for different types of background correction and the
assignment of quality weights to spot measurements. It also has gene expression analysis functionality (discussed in the next section).

**Statistical Analysis of Gene Expression Differences**

When determining which genes are differentially expressed between two groups in a microarray experiment, performing a simple t-test on each gene is not allowed. Because a high number of genes are tested at the same time, more sophisticated statistics are needed, e.g., to correct for multiple testing. Statistical analysis can also be complicated by the experimental design, such as when analyzing a time series. Several software packages are available for microarray analysis. Some provide only basic functionality and require considerable statistical knowledge from the researcher, while others are very user-friendly and guide the researcher through the process.

The Bioconductor package multtest (Pollard et al. 2005) is an example of a basic statistical package. When comparing gene expression levels between two groups, this package can be used to run parametric or nonparametric tests, and it provides several methods for applying multiple testing correction to the results. However, the package does not give much guidance as to which test statistic and which correction procedure are best suited for the specific dataset, and there is a large number of parameters that must be set correctly for each test. Therefore, help from a statistician is recommended when using this package.

The limma package (Smyth 2005) can be used to determine differentially expressed genes in a variety of experimental setups such as one-color arrays, universal reference designs, and direct comparisons. It uses linear model fitting, empirical Bayes smoothing, and multiple testing correction. It also includes methods to deal with arrays that contain duplicate spots. As mentioned, it includes an optional graphical user interface for more convenient analysis.

One of the most widely used software packages for gene expression analysis is SAM (Tusher et al. 2001). It is supplied as an add-in for Microsoft Excel, which makes it very easy to use, even for inexperienced researchers. It can correlate gene expression differences to a variety of response variables, including two-class comparisons (e.g., control versus treatment), multiclass comparisons (e.g., different types of treatment), quantitative variables (e.g., weight), and time-course experiments. It can also handle one-class experiments, which are used in case of a direct comparison experiment (e.g., where control and treatment samples are hybridized together on the same array). The user has control over which test statistic is used (parametric or nonparametric). SAM uses repeated permutations of the response variables to determine q-values (which are representations of p-values that are corrected for multiple testing) and false discovery rates (FDRs), and the list of genes that are significantly associated with the response variable is determined by a user-defined cutoff value for the FDR. The effect of changing the FDR cutoff on the list of significant genes is visualized by a SAM plot, which allows the user to make the best decision to acquire desired results. Apart from the control over the FDR, the user can also add an optional fold change requirement for the list of significant genes. When very large datasets must be analyzed, the Excel add-in might become very slow or even unresponsive. For analysis of these large datasets, the Bioconductor package siggenes (Schwender et al. 2006) is available, which enables users to run the SAM algorithm in R.

**Clustering**

Clustering of microarray data is used to identify groups of samples or genes with related gene expression patterns. Often, the whole dataset is used in a process called unsupervised hierarchical clustering, where there is no information provided to the clustering algorithm regarding specific sample classes or gene classes. Using unsupervised hierarchical clustering, one can gain information on whether there are specific subgroups within the total sample set, and on which genes show similar expression patterns over all samples. Clustering of genes is often used as a starting point for finding genes that are involved in specific pathways or share similar functions. Clustering can also be performed using a list of differentially expressed genes. This can provide more information on, e.g., the heterogeneity of sample subgroups. The results of hierarchical clustering are often displayed as a heat map, where the gene expression values are represented by colors.

Apart from hierarchical clustering, there are a number of other clustering algorithms such as k-means clustering (where the user specifies a desired
number of clusters), self-organizing maps (SOM), and principal component analysis (PCA). In order to cluster the data, the user has to choose a distance or similarity measurement (e.g., Pearson correlation or Euclidian distance) and a clustering algorithm.

The most well-known clustering programs for microarray analysis are Michael Eisen’s Cluster (for clustering) and TreeView (for visualization of the clustering results) (Eisen et al. 1998); the original programs have been continued as open-source versions Cluster 3.0 (de Hoon et al. 2004) and Java TreeView (Saldanha 2004). These programs can perform hierarchical, SOM, k-means, and PCA clustering using a number of parametric and nonparametric distance measurements. The output of the Cluster program is numerical and the output files can be loaded into TreeView to visualize the results in a heat map.

A second program that is recommended for clustering is Genesis (Sturn et al. 2002). This is a Java-based program that can perform a large number of different clustering algorithms using a variety of distance or similarity measurements. Advantages of Genesis are detailed control over the visualization of the results and convenient options for marking samples or genes and for selecting a specific cluster for further analysis, e.g., to turn it into a separate gene list or to investigate behavior of genes in this cluster in other clustering algorithms.

**Functional Annotation of Gene Lists**

The final step in microarray data analysis is to make biological sense of the acquired data. When a significant gene list has been determined using gene expression analysis, as described in Section “Statistical Analysis of Gene Expression Differences,” it is often not easy to discover the biological meaning behind the selected genes.

There are programs available that can annotate a list of differentially expressed genes and determine if certain annotations are overrepresented by comparing the proportion of genes with a specific annotation in the list of significant genes to the proportion of genes with the same annotation in the background gene list (often representing either the whole genome or the complete list of genes present on the array platform). One such program is the Database for Annotation, Visualization and Integrated Discovery (DAVID; Dennis et al. 2003; Huang et al. 2009), which can be used online. It uses a modified Fisher-exact p-value to determine overrepresentation of annotations in the list of significant genes. It can also analyze overrepresentation of gene sets. Furthermore, it is a very useful program to easily annotate a list of genes with information from a large number of different databases.

Another option is to perform a gene set enrichment analysis. This differs from the previously discussed overrepresentation analysis because it does not require that a list of significant genes be determined first; instead, it uses the complete dataset. In a gene set enrichment analysis, instead of identifying single genes that are differentially expressed between two groups, the genes are first grouped into gene sets that can be provided by the user (e.g., based on biological functions, pathways, or transcription factor-binding sites) and then the analysis determines which of these gene sets show significant concordant differences between two sample groups. The advantage is that, although multiple genes from one pathway may not reach statistical significance on their own, they might be significant when regarded as a group. Examples of programs that can perform gene set enrichment analysis are GSA (Efron and Tibshirani 2007), which is available as a module in SAM (Tusher et al. 2001), and GSEA (Mootha et al. 2003; Subramanian et al. 2005), which is available as a Java-based program (Subramanian et al. 2007) or as an R package.

A third group of functional annotation programs are the pathway analysis programs. These programs use databases of known pathways, such as Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.jp/kegg) and Biocarta (www.biocarta.com), and project the gene expression data onto these pathways to give a visualization of the regulation of all the genes within one pathway. A good example is GenMAPP (Doniger et al. 2003; Salomonis et al. 2007). This program can also be used to construct new pathways and to perform overrepresentation analysis.

There is one drawback to annotation-based analysis programs for aquaculture research. They all work with a certain “standard” annotation that they use to couple gene expression data to the gene sets and pathways. Most programs use Unigene or Entrez human gene symbols. Some have additional
databases for other model organisms such as mouse, and some also contain databases to convert, for instance, Affymetrix probeset IDs to their standard annotation. If these tools are to be used for aquaculture research, the researcher needs to be able to convert the ID used on the aquaculture microarray platform to one of these standard identifiers. In most cases, this will probably mean that the array must be annotated with the closest human putative orthologs. Therefore, these tools must be used with caution, as putative orthologs do not always share the same function.

For researchers working on nonmodel species, there is a program, Blast2GO, that can provide automated Gene Ontology (GO) annotation for a large set of sequences (Conesa et al. 2005; Götz et al. 2008). This program uses the NCBI Blast database to annotate the sequences and then uses the Blast results to map the sequences to GO terms and KEGG pathways. Functional analysis is provided by the program in the form of charts, graphs, and over-representation analysis. In addition, the annotations provided by Blast2GO can be used for further analysis by the tools discussed.

**Recommendations and Future Directions**

There are some recommendations to make for future microarray research. First of all, the research community would benefit greatly if the platforms that are developed would be made publicly available. This would reduce the redundancy of platforms that are produced from the same oligo sets, such as is currently seen with the zebrafish oligo arrays. Having a small number of “standard,” widely applicable arrays per species would also make comparison of results between studies (metadata analysis) more feasible. Second, although we are aware that not all journals currently require it, all published microarray data should be entered into a database such as NCBI’s GEO. The standardization of information that is required and provided by a database like GEO ensures that there is unambiguous and detailed information available about microarray experiments and that they are compliant to the Minimum Information About a Microarray Experiment (MIAME) standard (Brazma et al. 2001). In the case of platform entries, this would also benefit researchers searching for a suitable platform for their experiments.

The development of well-annotated whole transcriptome microarrays for all aquaculture finfish species would be very beneficial for aquaculture research. Currently, the finfish species with the most complete microarray platforms is the zebrafish, because this species is well characterized both on a genomic and transcriptomic level. (For example, NCBI’s Taxonomy Browser shows 1,481,936 publicly available *Danio rerio* ESTs.) To reach the same goal for other finfish species, large EST databases must be established. To ensure that these databases cover the entire transcriptome, they should include ESTs from a variety of cDNA libraries (e.g., normalized tissue- and developmental stage-specific cDNA libraries; SSH libraries enriched for transcripts of interest, e.g., immune- and stress-responsive transcripts), subjected to deep sequencing using a NGS platform such as Roche/454’s GS FLX Titanium, Illumina/Solexa’s GA, or Life/APG’s SOLiD 3 (reviewed by Metzker 2010). Combined with high-throughput NGS-based characterization of aquaculture finfish whole genomes, this should enable the development and annotation of whole transcriptome microarrays for species important in aquaculture research.

When high-quality reference genome sequences are available for aquaculture finfish species, high-throughput NGS-based transcriptome analysis (i.e., RNA-seq) will be utilized for some global transcript expression studies. However, DNA microarrays will likely continue to be utilized for global gene expression studies for finfish species lacking high-quality reference genome sequences. In addition, microarrays will likely continue to be utilized for complex experimental designs (e.g., multiple treatments, sampling time points, etc.), as they currently allow global transcript expression studies to be performed on large numbers of biological replicates at reasonable costs using established data analysis methods (e.g., Booman et al. 2011).

The development of functional genomics resources such as targeted gene discovery tools (e.g., SSH libraries) and gene expression microarrays, combined with complementary technologies (e.g., NGS and quantitative reverse transcription–polymerase chain reaction (QPCR)), are helping researchers to take a large step toward identifying genes that are involved in traits that are
important in aquaculture, such as resistance to disease and stress and good growth characteristics. For example, suites of trait-relevant genes (e.g., upregulated in pathogen-resistant Atlantic cod spleen compared to pathogen-susceptible Atlantic cod spleen) identified using functional genomics experiments can be validated and further studied at the transcript expression level in numerous individuals and families using high-throughput QPCR. In addition, trait-relevant candidate genes (exons, introns, and regulatory regions) can be subjected to NGS-based targeted resequencing with pooled templates from individuals in families enrolled in broodstock development programs to identify SNPs that may be used in MAS of finfish with superior production traits (Fig. 1.2). Aquaculture genomics projects can increase the likelihood of identifying useful, trait-correlated SNPs for MAS if high-throughput SNP

<table>
<thead>
<tr>
<th>Examples of experimental approaches to identify genes contributing to traits of interest</th>
<th>Enhanced growth rate</th>
<th>Resistance to heat stress</th>
<th>Resistance to pathogens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare fast-growing and slow-growing family/individual transcriptomes using genomic techniques such as: 1. SSH library construction and characterization; 2. DNA microarray hybridization; and 3. Next-generation sequencing.</td>
<td>Compare heat stress-resistant and heat stress-susceptible family/individual gene expression responses to heat stress using genomic techniques such as: 1. SSH library construction and characterization; 2. DNA microarray hybridization; and 3. Next-generation sequencing.</td>
<td>Compare pathogen-resistant and pathogen-susceptible family/individual gene expression responses to pathogen or PAMP using genomic techniques such as: 1. SSH library construction and characterization; 2. DNA microarray hybridization; and 3. Next-generation sequencing.</td>
<td></td>
</tr>
</tbody>
</table>

**Goal of gene expression studies**
- Identify a comprehensive set of growth-relevant genes
- Identify a comprehensive set of heat stress-relevant genes
- Identify a comprehensive set of immune-relevant genes

**Next steps**
- SNP discovery
  - Map all growth-relevant genes
  - QTL studies with high-throughput genotyping
- SNP discovery
  - Map all heat stress-relevant genes
  - QTL studies with high-throughput genotyping
- SNP discovery
  - Map all immune-relevant genes
  - QTL studies with high-throughput genotyping

**Outcome**
- SNP markers for selecting fast-growing broodstock
- SNP markers for selecting heat stress-resistant broodstock
- SNP markers for selecting pathogen-resistant broodstock

**Figure 1.2.** Examples of experimental approaches to (1) identify genes contributing to aquaculture-relevant traits (e.g., rapid growth, or resistance to pathogens or environmental stressors); and (2) identify single nucleotide polymorphisms (SNPs) associated with trait-relevant candidate genes (e.g., exonic, intronic, or in regulatory regions) that may be useful in marker-assisted selection (MAS) of finfish with superior production traits.
genotyping (e.g., using SNP platforms by companies such as Illumina) is employed to screen large numbers of candidate SNPs with large numbers of individuals.

In addition to previously mentioned applications, functional genomics studies (e.g., using gene expression microarrays) will also be helpful in the development of nonlethal diagnostic tests for pathogen infection and carrier state. To help diagnosis of pathogen infection and to increase our understanding of biological processes that take place during infection of finfish, some pathogen-specific microarrays have been developed (Chen et al. 2006; Nash et al. 2006).

Most current finfish microarray platforms are used for gene expression analysis. However, regulation takes place on many other levels, such as at the protein level or by epigenetic mechanisms. In the future, other array technologies such as antibody arrays or ChIP-on-chip arrays should be able to give a more complete view of gene and protein regulation associated with important traits.

REFERENCES


Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, and Conklin BR. 2003. MAPMFinder: using Gene Ontology and GenMAPP to create a

Functional Genomic Tools for Aquaculture
Aquaculture Biotechnology

16

Goveroum M, Le Gac F, and Guiguen Y. 2006. Generation of a large scale repertoire of expressed sequence tags (ESTs from normalized rainbow trout cDNA libraries. BMC Genomics. 7: 196.
challenged differences between individuals with early, intermediate and late mortality. BMC Genomics. 9: 179.


Aquaculture Biotechnology


