Transcriptional analysis of endocrine disruption using zebrafish and massively parallel sequencing

Michael E Baker and Gary Hardiman
1,2
Department of Medicine, University of California San Diego, 9500 Gilman Drive 0605, La Jolla, California 92093-0605, USA
1CSRC and BIMRC, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-7720, USA
2Department of Medicine, Medical University of South Carolina, 135 Cannon Street, Suite 303 MSC 835, Charleston, South Carolina 29425, USA

Abstract

Endocrine-disrupting chemicals (EDCs), including plasticizers, pesticides, detergents, and pharmaceuticals, affect a variety of hormone-regulated physiological pathways in humans and wildlife. Many EDCs are lipophilic molecules and bind to hydrophobic pockets in steroid receptors, such as the estrogen receptor and androgen receptor, which are important in vertebrate reproduction and development. Indeed, health effects attributed to EDCs include reproductive dysfunction (e.g. reduced fertility, reproductive tract abnormalities, and skewed male:female sex ratios in fish), early puberty, various cancers, and obesity. A major concern is the effects of exposure to low concentrations of endocrine disruptors in utero and post partum, which may increase the incidence of cancer and diabetes in adults. EDCs affect transcription of hundreds and even thousands of genes, which has created the need for new tools to monitor the global effects of EDCs. The emergence of massive parallel sequencing for investigating gene transcription provides a sensitive tool for monitoring the effects of EDCs on humans and other vertebrates, as well as elucidating the mechanism of action of EDCs. Zebrafish conserve many developmental pathways found in humans, which makes zebrafish a valuable model system for studying EDCs, especially on early organ development because their embryos are translucent. In this article, we review recent advances in massive parallel sequencing approaches with a focus on zebrafish. We make the case that zebrafish exposed to EDCs at different stages of development can provide important insights on EDC effects on human health.

Key Words
- endocrine disruptors
- steroid receptors
- zebrafish
- massive parallel sequencing

Endocrine disruptors

There is increasing concern about the effects on the health of humans and wildlife of organic and inorganic contaminants that have been introduced into the environment by manufacturing industries (Colborn et al. 1993, Calafat et al. 2008, Diamanti-Kandarakis et al. 2009, Heindel & vom Saal 2009, Rubin 2011, Baker & Chandsawangbhuwana 2012, Baker et al. 2012). Many of these chemicals are used as plasticizers, pesticides, detergents, and pharmaceuticals, providing important benefits for modern industrial societies. However, some of these chemicals disrupt a variety of hormone-regulated physiological pathways, including reproductive responses mediated by the estrogen receptor (ER) and androgen receptor (AR) in humans (Kelce et al. 1998, Sonnenschein & Soto 1998, Calafat et al. 2008, Soto et al. 2008, 2009,
In humans, exposure to endocrine-disrupting chemicals (EDCs) may lead to premature puberty in females, and decreased reproductive ability in men. Of major concern is transient exposure to EDCs in utero and in newborns, which can have toxic effects on reproduction and development, as well as causing some endocrine-related cancers and heart disease later in life (Grun & Blumberg 2006, Henley & Korach 2006, Soto et al. 2008, Swan 2008).

Estrogens, androgens, other steroids (Fig. 1), and many EDCs (Fig. 2) are small molecules, lipophilic in nature, which are thought to explain binding of EDCs to the ER, AR, and other steroid receptors. For example, chemicals, such as 4-nonylphenol, which contain a phenolic group that mimics the A ring on estradiol (E2), and bisphenol A (BPA), which mimic the A and D rings on E2, can disrupt physiological responses mediated by the ER (Sonnenschein & Soto 1998, Diamanti-Kandarakis et al. 2009, Rubin 2011, Baker & Chandsawangbhuwana 2012). The ER, AR, and other steroid receptors are nuclear receptors, a large and diverse family of transcription factors (Escriva et al. 2000, Bertrand et al. 2004, Baker 2005, Markov et al. 2009, Bertrand et al. 2011, Krasowski et al. 2011, Sladek 2011) that are activated by a variety of lipophilic molecules (Chawla et al. 2001, Ingraham & Redinbo 2005, Markov et al. 2009, Huang et al. 2010, Baker 2011, Sladek 2011). Nuclear receptors regulate differentiation, development, homeostasis, and responses to stress in humans and other vertebrates. Thus, disruption of nuclear receptor signaling by EDCs poses a threat to many physiological responses in humans, fish, and other animals (Grun & Blumberg 2006, Swan 2008, Diamanti-Kandarakis et al. 2009, Heindel & vom Saal 2009).
Soto et al. 2009, Vandenberg et al. 2009, Nelson et al. 2013). However, nuclear receptors are not the only target for EDCs. As more and more diverse chemicals are synthesized and dispersed into the environment, it is likely that additional physiological responses will be disrupted. As a result, government agencies and environmental scientists have a critical need for methods to monitor both short-term and long-term effects of exposure to low concentrations of EDCs on human health.

Zebrafish: a model for studying endocrine disruption

Zebrafish (Danio rerio) have emerged as an important tool for studying the biological effects of hormones and EDCs. Zebrafish is a small tropical fresh-water fish indigenous to northern India, northern Pakistan, Nepal, and Bhutan in South Asia. Zebrafish are relatively inexpensive to maintain compared with rats and mice and are easily bred in large numbers. Zebrafish have short life spans (60 days to maturity). Thus, three generations of zebrafish can be easily generated within a 12-month period, making it possible to carry out immediate studies in response to exposure to endocrine disruptors, in addition to longitudinal transgenerational studies analysis with F1 and F2 progeny derived from the exposed fish. Zebrafish embryos are optically transparent, allowing microscopic observations of organs as they develop (Lieschke & Currie 2007, Segner 2009). Thus, one can study the onset and course of exposure to EDCs on organ development in vivo and in real time. At this point, there is no marine vertebrate model that is as well characterized as the zebrafish. The highly inbred, lab-dependent nature of these animals and the availability of the complete genome sequence make them an ideal model for toxicology studies and rapid phenotypic assessment (Lieschke & Currie 2007).

Zebrafish genome

The draft genome of the zebrafish genome has been publicly available for over a decade, facilitating many novel discoveries, including positional cloning of hundreds of genes from mutations. Recent sequencing efforts using massively parallel sequencing have focused on generating a high-quality reference genome similar to those available for the human and mouse genomes. The Zv9 assembly reported recently by Howe et al. (2013) is a hybrid of high-quality finished clone sequence (83%) and...
whole-genome shotgun sequence (17%), with a total size of 1.412 Gb. This reference sequence is linked to a high-resolution, high-density meiotic map. It greatly facilitates the identification of mutations, because it enables direct comparison of both mutated and normal sequences (Schier 2013). Analysis of the reference has revealed that more than 75% of human genes implicated in disease have counterparts in zebrafish. An advantage of the annotated reference genome is possibilities for exome-enrichment deep sequencing experiments analogous to those now routinely carried out with human genome samples. This will accelerate positional cloning projects and novel genome-wide mutation discovery efforts (Howe et al. 2013). This further highlights the use of the zebrafish model for elucidation of gene function and studies of the effects of endocrine disruption.

**Genetic modification**

Zebrafish embryos and adults can be genetically modified by microinjection, chemical mutagenesis and transgenesis. Moreover, the zebrafish database ZFIN (http://zfin.org; Bradford et al. 2011) provides up-to-date information on all aspects of zebrafish. Most important, many of the developmental processes in zebrafish are conserved in humans (Barbazuk et al. 2000, Dooley & Zon 2000, Woolfe et al. 2004). The relevance to humans of many genes and pathways for development in zebrafish makes these fish an excellent system to study possible effects of EDCs on human health. In addition, data from zebrafish exposed to samples taken from sewage outfalls (Vajda et al. 2008, Lange et al. 2009), rivers, drinking water, soil, food, etc. can be used to determine their toxicity to humans of EDCs from different sources.

*An ere-zvtg1: gfp* transgenic zebrafish line for environmental monitoring of endocrine disruptors was developed recently for environmental monitoring of endocrine disruptors. In this transgenic model, under control conditions, green fluorescent protein (GFP) was exclusively expressed in the liver of mature adult female fish. Male and larval transgenic fish did not express GFP but could be induced to express GFP in the liver after exposure to 17-α-ethynylestradiol (Chen et al. 2010, Gorelick & Halpern 2011, Lee et al. 2012).

**Monitoring exposure to endocrine disruption**

Monitoring pollution at environmental sites typically involves, sampling sediment or water from that location, determination of the constituents present and comparison of the levels of contaminants to reference levels associated with adverse exposure outcomes. In many cases, only a percentage of contaminants are bio-available, making this approach problematic. Further confounding this issue is the synergistic effects provided by compound mixtures, which can yield outcomes that differ significantly from single compound exposures. There is a growing need to develop alternate approaches to biomonitoring, which encompass chemical analysis and bioavailability studies of pollutants end-point phenotypic and biological assays (Celiz et al. 2009, Letcher et al. 2010).

An advantage of using zebrafish to monitor exposure to EDCs is that an oral method can be used to administer EDCs to zebrafish. Zebrafish are raised in glass aquaria, at a constant temperature of 29 °C with a light:darkness cycle ratio of 1410 h. The fish typically are acclimated for 1 week before commencing the exposure experiments. EDC exposures use a continuous flow-through system to maintain constant concentrations of the chemicals during the duration of the exposure, after which the fish can be selected for phenotypic and genomic analyses. Recent advances in molecular methods, such as microarrays, have substantially improved the sensitivity for assessing the effects of EDCs. Indeed, the ability to screen a wide range of endocrine responses in organisms from polluted waters and monitor harmful effluents entering the ecosystem is facilitated by microarray technology.

**DNA microarray technology**

The multiplicity of genes that will have altered transcription in humans and wildlife due to exposure to EDCs requires comprehensive assays for analyzing gene transcription. DNA microarrays have been useful for this purpose (Brown & Botstein 1999, Hardiman 2004). Microarrays facilitate the measurement of expression levels of hundreds to thousands of genes simultaneously from a single tissue sample. This technology has been widely adopted by the biomedical research community and for the past decade it has been a key molecular tool in basic, translational, and clinical research. Microarray technology has found widespread use by the zebrafish community, in applications as diverse as rod photoreceptor degeneration and regeneration (Morris et al. 2011), the toxic effects of nanoparticulate silver in zebrafish (Griffitt et al. 2013), and differences in response to teratogenic and non-teratogenic exposure concentrations of BPA and 17β-E2 (Salli et al. 2013). Although their primary use has been for gene expression analyses, microarrays have also found the use in genotyping and
re-sequencing applications (Hardiman 2004, Trachtenberg et al. 2012). We have successfully employed microarray technology in the context of environmental monitoring (Baker et al. 2009, 2013).

DNA microarrays, however, rely on the kinetics of probe binding and consequently low abundance transcripts may not be detected owing to poor signal-to-noise ratios. A pitfall is that many commercial catalog microarrays are frozen in time, in that their design may be based on a draft or early genome build for an organism. It is important, when working with zebrafish microarrays, to verify that current annotation has been used in the array design. Older array designs derived from draft genome builds, when the genome annotation lacked the sophistication and refinement of the current reference sequence, may contain probes that are no longer relevant and also may lack critical probes (Hardiman 2004, Trachtenberg et al. 2012).

Massively parallel sequencing

A major technological shift in the research community in the past 5-year period has been the adoption of high-throughput sequencing (HTS) technologies to facilitate whole-genome and transcriptome sequencing (ten Bosch & Grody 2008, Tucker et al. 2009). Conventional sequencing approaches were both labor intensive and costly. High molecular DNA fragments were broken down into smaller pieces, which were amplified and sequenced individually using fluorescent ddNTPs as DNA chain terminators. Capillary-based electrophoresis separated these DNA fragments and facilitated the detection and recording of fluorescence, one base or fragment at a time, yielding short reads 100–1000 bp in length, which were assembled into longer contiguous reads. The main advantage of ‘next-generation’ or ‘second-generation’ technologies is that complex DNA cloning and library construction are avoided and shotgun approaches are implemented where DNA is randomly fragmented and many DNA fragments are sequenced simultaneously (Hardiman 2008, Bhasker & Hardiman 2010). A detailed description of the technologies and the inherent biases associated with each of these new technologies is beyond the scope of this review. However, we guide the reader to the following articles (Glenn 2011, Meacham et al. 2011, Minoche et al. 2011, Pareek et al. 2011, Suzuki et al. 2011, Carneiro et al. 2012, Lam et al. 2012, Liu et al. 2012, Quail et al. 2012, Boland et al. 2013, Frey et al. 2014).

To date the majority of the published studies combining RNA-sequencing (RNAseq) and zebrafish have focused on immunity (Hagedus et al. 2009, Soares et al. 2009, Aday et al. 2011). Other biological questions are being addressed using HTS include development, microRNA discovery, and chromatin modification (Soares et al. 2009, Aanes et al. 2011, Aday et al. 2011). The advantages and applications of these technologies are discussed below. The salient features of these key HTS technologies are summarized in Table 1. Applications and the strengths and weaknesses of these platforms are summarized in Table 2.

454 Sequencing

The first commercial next generation sequencing platform was introduced in 2005 by 454 Life Sciences (Branford, CT, USA), a subsidiary of Roche. Since then, there has been phenomenal growth and improvement of HTS technologies. 454 Sequencing employs a large-scale parallel pyro-sequencing system routinely capable of sequencing ~400–600 Mb of DNA per 10-h run on the Genome Sequencer FLX with GS FLX Titanium series reagents (Voelkerding et al. 2009). More recent advances with the GS FLX+ System facilitate read lengths up to 1000 bp in length. 454 sequencing has been used for whole-genome sequencing, transcriptome sequencing, amplicon sequencing, capture-based sequencing, and metagenomics analyses of complex environmental samples (Table 1).

454 Sequencing has been used in metagenomic studies where 16S rRNA gene sequencing was performed to investigate EDC-degrading microbial enrichment cultures (Villemur et al. 2013). 454 Sequencing was also used to examine the impact of a known endocrine disruptor (tributyltin or TBT) on the transcriptome of the dog whelk, Nucella lapillus (Chapman & Guillette 2013, Pascoal et al. 2013). This revealed that TBT mimics the endogenous ligand of the nuclear retinoid X receptor (RXR) and/or peroxisome proliferator-activated receptor disrupting pathways.

454 Sequencing was used for miRNA discovery in zebrafish. Twenty-five novel miRNAs were predicted. In addition, 107 miRNA star sequences and 41 candidate miRNA targets were identified. A miRNA expression profile built on the basis of pyrosequencing read numbers demonstrated high expression of most miRNAs throughout zebrafish development and identified tissue-specific miRNAs (Soares et al. 2009).

To date, 454 sequencing technology has not been used for the study of endocrine disruption in zebrafish. The strength of this sequencing approach is the long read data. This permits optimization of de novo transcriptome
assembly from next-generation sequencing data in the absence of a scaffold reference in zebrafish strains exposed to endocrine disruptors. This technology has been useful to date in the analysis of transcriptomes from the Atlantic cod (*Gadus morhua*) including RNA transcripts in cells and tissues from various life stages, tissue types, physiological states, and environmental conditions (Johansen et al. 2011). It has also facilitated de novo assembly of the guppy (*Poecilia reticulata*) transcriptome (Fraser et al. 2011).

The competitive landscape of HTS and the disruptive nature of this rapidly evolving technology are highlighted by the fact that 454 sequencing will soon be obsolete. The platform will cease to be supported by Roche by mid-2016.

**Illumina platforms**

The HiSeq 2500 (Illumina, San Diego, CA, USA) employs a proprietary clonal array approach, where individual DNA molecules are fixed to a flat flow-cell surface and are amplified in situ, and serve as templates for synthetic sequencing with novel fluorescent reversible terminator chemistry (Bentley et al. 2008). With the Illumina approach, isothermal amplification of DNA constructs is carried out creating clonal template clusters (~1000 copies each). The Illumina HiSeq instruments directly sequence the resulting high-density array of template clusters on the flow cell using sequencing by synthesis. Four proprietary fluorescently labeled, reversible terminator nucleotides are used to sequence the millions of clusters base-by-base in parallel. The surface is subsequently imaged to generate massive amounts of DNA sequence data (Bentley et al. 2008).

This platform yields high-quality short-sequence tags (up to 150 bp in length), outputting up to 600 Gb sequence per instrument run, and has enabled high-throughput genome and transcriptome analyses for thousands of organisms. A related instrument, the Illumina desktop MiSeq sequencer enables longer sequence tags (up to 300 bp in length) and an output of up to 15-Gb sequence per instrument run.

In early 2014, Illumina presented two new sequencing platforms, the NextSeq 500 and the HiSeq X Ten. The NextSeq 500 represents a novel sequencing-by-synthesis approach based on a two channel approach, single dyes for the bases adenine and cytosine, two dyes for the base thymine, and a dark state for the base guanine. This desktop instrument yields sequence tags (up to 150 bp in

---

**Table 1** High-throughput sequencing platforms; chemistry, read lengths, and run times

<table>
<thead>
<tr>
<th>Company</th>
<th>Platform</th>
<th>Chemistry</th>
<th>Read length (bp)</th>
<th>Per full run</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche</td>
<td>GS junior system</td>
<td>Pyrosequencing</td>
<td>400</td>
<td>35 Mb</td>
<td>10 h</td>
</tr>
<tr>
<td></td>
<td>GS FLX + system</td>
<td></td>
<td>Up to 1000 (with GS FLX Titanium XL+)</td>
<td>700 Mb</td>
<td>23 h</td>
</tr>
<tr>
<td>Illumina</td>
<td>MiSeq</td>
<td>Sequencing by synthesis</td>
<td>2 × 300 (MiSeq Reagent kit v3)</td>
<td>13.2–15 Gb</td>
<td>65 h</td>
</tr>
<tr>
<td></td>
<td>Illumina HiSeq2500 Dual flow cell High output run mode Rapid run mode</td>
<td></td>
<td>2 × 100</td>
<td>540–600 Gb</td>
<td>11 days</td>
</tr>
<tr>
<td></td>
<td>5500xl System</td>
<td>Sequencing by ligation</td>
<td>2 × 150</td>
<td>150–180 Gb</td>
<td>40 h</td>
</tr>
<tr>
<td>ABI SOLiD</td>
<td></td>
<td></td>
<td>Mate-paired: 2 × 60 bp Paired-end: 75 bp × 35 bp Fragment: 75 bp</td>
<td>10–15 Gb/day</td>
<td>2–7 days</td>
</tr>
<tr>
<td>Ion torrent</td>
<td>Ion PGM</td>
<td>Semiconductor sequencing technology</td>
<td>400 bp (Ion 318 Chip v2) Up to 200-base fragment reads (with Ion PI Chip)</td>
<td>600 Mb–1 Gb</td>
<td>7.3 h</td>
</tr>
<tr>
<td></td>
<td>Ion proton</td>
<td></td>
<td></td>
<td>Up to 10 Gb</td>
<td>2–4 h</td>
</tr>
<tr>
<td>Pacific biosciences</td>
<td>Pacific biosciences PacBio RS</td>
<td>Single-molecule real time</td>
<td>4200 bp mean mapped read length (with XL Binding/C2 Sequencing Chemistry)</td>
<td>220 Mb</td>
<td>2 × 55 m (per SMRT cell)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7600 bp mean mapped read length (with P5/C3 Sequencing Chemistry)</td>
<td>608 Mb</td>
<td>180 m (per SMRT cell)</td>
</tr>
</tbody>
</table>
length), outputting up to 120-Gb sequence per instrument run. The X Ten has been designed with one focused application, whole-human genome sequencing. One instrument run generates 600 Gb of sequence per day, which is the equivalent of five human genomes sequenced at 40X genome coverage.

Over 90% of all sequencing data generated to date has been with Illumina sequencers (Hardiman 2008, Voelkerding et al. 2009) and based on continuing innovations in chemistry, optics, flow-cell design, and bioinformatics approaches for data analyses this platform will continue to be a major contributor to large-scale sequencing projects for the foreseeable future. In addition, the cost per bp of DNA sequence will continue to drop.

This technology has been successful in sequencing human genomes (Levy et al. 2007). The short individual read lengths have led to primary applications in re-sequencing, where an established reference genome exists rather than de novo sequencing. Cancer genomes are routinely sequenced to great depth using Illumina sequencing. The effects of tobacco smoke on a small-cell lung cancer cell line, NCI-H209 was investigated using Illumina sequencing, providing in depth views of mutational processes, cellular repair pathways, and gene networks associated with cancer (Pleasance et al. 2010). Genome wide location studies have been extensively carried out with Illumina sequencing. Chromatin immunoprecipitation (ChIP)-seq combines ChIP with HTS to uncover the binding sites of DNA-associated proteins. (Barski et al. 2007, Johnson et al. 2007).

Illumina sequencing has been used to investigate the effects and mechanisms of toxicity of silver nanoparticles in zebrafish. In recent years, nanoparticles have been increasingly used in several industrial, consumer, and medical applications because of their unique physiochemical properties. Low concentrations of silver nanoparticles can cause significant disruptions to natural ecosystems (Colman et al. 2013). The detrimental health effects of nanoparticles remain to be fully elucidated, but there is increasing interest in the role as potential endocrine disruptors (Iavicoli et al. 2013). The effects of exposures to silver in different forms (nano, bulk, and ionic forms) were investigated by exposing zebrafish embryos and carrying out transcriptomic analysis using High-Throughput SuperSAGE (van Aerle et al. 2013). This approach is an advanced form of conventional serial analysis of gene expression technology (SAGE). A specific sequence tag from each

<table>
<thead>
<tr>
<th>Company</th>
<th>Platform</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche</td>
<td>GS Junior system</td>
<td>Short run time, long reads</td>
<td>Low sequencing depth</td>
<td>Amplicon</td>
</tr>
<tr>
<td></td>
<td>GS FLX+ system</td>
<td></td>
<td></td>
<td>Sequence capture</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Whole genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Metagenomics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td>Illumina</td>
<td>MiSeq</td>
<td>High throughput</td>
<td>DNA sequence amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>Illumina HiSeq2500 Dual flow cell</td>
<td>Ultra-high throughput (up to 3 billion single reads or 6 billion paired-end reads)</td>
<td>Amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>High output run mode</td>
<td></td>
<td></td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>Rapid run mode</td>
<td></td>
<td></td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td>ABI SOLiD</td>
<td>5500xl system</td>
<td>Ultra-high throughput</td>
<td>Amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td>Ion torrent</td>
<td>Ion PGM</td>
<td>High throughput</td>
<td>Amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>Ion proton</td>
<td>High throughput (60–80 million reads passing filter)</td>
<td>Amplification bias</td>
<td>Targeted sequencing, copy number analysis, de novo small genome sequencing, metagenomics, ChIP sequencing, methylation analysis Genome, exome, gene, ChIP, methylation, transcriptome sequencing</td>
</tr>
<tr>
<td>Pacific biosciences</td>
<td>PacBio RS</td>
<td>Long read sequencing</td>
<td>Low single-pass accuracy ~85%</td>
<td>Small genome sequencing, targeted sequencing, DNA base modification analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Platform</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche</td>
<td>GS Junior system</td>
<td>Short run time, long reads</td>
<td>Low sequencing depth</td>
<td>Amplicon</td>
</tr>
<tr>
<td></td>
<td>GS FLX+ system</td>
<td></td>
<td></td>
<td>Sequence capture</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Whole genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Metagenomics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td>Illumina</td>
<td>MiSeq</td>
<td>High throughput</td>
<td>DNA sequence amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>Illumina HiSeq2500 Dual flow cell</td>
<td>Ultra-high throughput (up to 3 billion single reads or 6 billion paired-end reads)</td>
<td>Amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>High output run mode</td>
<td></td>
<td></td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>Rapid run mode</td>
<td></td>
<td></td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td>ABI SOLiD</td>
<td>5500xl system</td>
<td>Ultra-high throughput</td>
<td>Amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td>Ion torrent</td>
<td>Ion PGM</td>
<td>High throughput</td>
<td>Amplification bias</td>
<td>De novo genomes, transcriptomes, metagenomes</td>
</tr>
<tr>
<td></td>
<td>Ion proton</td>
<td>High throughput (60–80 million reads passing filter)</td>
<td>Amplification bias</td>
<td>Targeted sequencing, copy number analysis, de novo small genome sequencing, metagenomics, ChIP sequencing, methylation analysis Genome, exome, gene, ChIP, methylation, transcriptome sequencing</td>
</tr>
<tr>
<td>Pacific biosciences</td>
<td>PacBio RS</td>
<td>Long read sequencing</td>
<td>Low single-pass accuracy ~85%</td>
<td>Small genome sequencing, targeted sequencing, DNA base modification analysis</td>
</tr>
</tbody>
</table>

Table 2. High-throughput sequencing platforms; applications, strengths, and weaknesses
transcribed gene is mapped, identified, and counted. By sequencing and counting as many tags as possible, a transcriptional profile is provided. Significant alterations in gene expression were uncovered for all forms of silver and many of the gene pathways affected were associated with oxidative phosphorylation and protein synthesis. There was a strong overlap between the different forms of silver. This indicates similar mechanisms of toxicity between the different forms of silver and suggests a requirement for bioavailability of silver ions.

Arsenic is a worldwide and mobile metalloid pollutant in environment. It is highly toxic to most species, although a few species of bacteria use arsenic compounds as respiratory metabolites. Arsenic contamination of groundwater is a problem that affects millions of people worldwide (Fendorf et al. 2010). RNA–SAGE (serial analysis of gene expression) using zebrafish has been used to investigate the molecular mechanism of arsenate toxicity (Xu et al. 2013). Using 12 million SAGE tags mapped to the zebrafish genome, transcriptional profiles revealed that differentially expressed genes were significantly enriched in several major biological processes including oxidation reduction, translation, iron ion transport, cell redox, and homeostasis. Accordingly, the main pathways disturbed by arsenic treatment include metabolic pathways, protease, oxidative phosphorylation, and cancer. Pathway analysis identified a network with four important hub genes, including Jun, Kras, APOE, and Nr2f2.

RNAseq was used to study the transcriptome of the zebrafish pineal gland and the effects on circadian clockwork by light exposure (Ben-Moshe et al. 2014). The authors focused on light-induced genes that encode transcription factors. They noted that two factors, dec1, in addition to the core clock gene per2, were critical for the light-entrainment of rhythmic locomotor activity.

**Applied Biosystems SOLiD DNA sequencing**

Agencourt Personal Genomics introduced ‘supported oligoligation detection’ (SOLiD) DNA sequencing technology, which was subsequently developed and commercialized by Life Technologies (Applied Biosystems). It uses a different approach to Illumina ‘sequencing by synthesis’ termed ‘sequencing-by-ligation’. With SOLiD sequencing, a DNA library is prepared and used to generate clonal magnetic bead populations. Each bead contains a unique fragment species. Emulsion PCR generates DNA template which is subsequently covalently attached to a glass surface. Sequencing employs four fluorescent tags and a two-base readout system. Every ligation step interrogates a pair of adjacent nucleotides. As each base is effectively interrogated twice, this results in high accuracy of the sequence calls. Read lengths are mate-paired: 2 × 60 bp; paired-end: 75 × 35 bp; and fragment: 75 bp. Applications to date have included the analyses of sequence variations including, copy number variations, single-base duplications, insertions, insertions, and deletions in addition to single-nucleotide polymorphisms. A high-resolution, nucleosome position map of the nematode Caenorhabditis elegans was generated using SOLiD. This technology facilitated a global view of the chromatin architecture of a multicellular animal at extremely high density and resolution (Valouev et al. 2008).

Owing to the different metabolic requirements for male and female reproduction, the liver is one of the most sexually dimorphic organs in terms of its transcriptomes. The liver has been the primary focus of many studies on endocrine disruption as it is the key organ involved in detoxification and responses to estrogens and other hormones. In oviparous species, such as zebrafish, the female liver is the main organ for the production of yolk protein precursors (vitellogenins) and some zona pellucida proteins. HTS using SAGE was carried out to delineate the transcriptome profiles in male and female zebrafish (Xu et al. 2013). Construction of SAGE libraries and sequencing were carried out using SOLiD Analyzer 4 (Applied Biosystems). In zebrafish, sexual dimorphism in xenobiotic metabolism and anti-oxidation gene expression was reported. This indicated that RXR and liver X receptor, important targets for endocrine disruption, play central roles in the regulation of the sexual differences of lipid and cholesterol metabolisms.

**Ion semiconductor sequencing**

Ion semiconductor sequencing (Ion Torrent Systems, Thermo Fisher Scientific, Carlsbad, CA, USA) represents another platform that has become popular in recent years. Rothberg and colleagues described this post-light sequencing approach that was unique in that it was not dependent on conventional fluorescence, complex optics, or modified nucleotides (Merriman et al. 2012). Low-cost semiconductor manufacturing techniques were used to generate an integrated circuit capable of direct non-optical DNA sequencing using natural nucleotides. The sequencer contains an electronic reader board which interfaces with the sequencing chip, a signal processing microprocessor, and a fluids system to control reagents flow over the chip. The ions generated by the DNA polymerase catalyzed synthesis are detected via an ion sensor. Instrument cost is...
diminished as there are no optical components. Two instruments are currently available, the Ion PGM and Ion Proton with maximal yields of 600 Mb–1 and 10 Gb respectively.

Ion sequencing has evolved as a rapid and cost-effective tool for developing microsatellite markers for non-model fish species. One recent effort sequenced the genome of an endemic fish species (*Schizothorax biddulphi*) classified as an extremely endangered freshwater fish in China. Sequencing was carried out using the Ion torrent PGM instrument to obtain a large number of micro-satellites for *S. biddulphi* (Luo et al. 2012). Ion sequencing has found use in classifying individual bacterial species that comprising complex, polymicrobial patient specimens and has become a tool that complements classical culture-based and molecular microbiology methods. The sequencing data can be used for accurate genus- or species-level taxonomic assignment of the metagenomic communities (Salipante et al. 2013).

At this time, there are no records in the literature of genomic or transcriptomic zebrafish sequencing using Ion Torrent. This is clearly due to the short read lengths and lower depth compared with other technologies. Another challenge with this sequencing has been distinguishing homopolymer repeats of the same nucleotide. However, the technology has undergone much development since its initial description and the increased read lengths of 400 bp/run, and greater chip densities for the Ion PGM have improved the technology considerably. The relatively low upfront costs, modest operating costs, and rapid sequencing speed will see this technology gain traction in zebrafish transcriptional profiling in response to endocrine disruptors.

**Single molecule real time sequencing**

Distinguished from the approaches described above is long-read Pacific Biosciences Single Molecule Real Time (SMRT) sequencing, a complementary technology which permits observation of natural DNA synthesis by a DNA polymerase as it occurs. SMRT DNA sequencing technology uses millions of 10 nm holes (zeromode waveguides (ZMWs)) in a silicon dioxide substrate. Each opening has a zeptoliter volume and contains a single DNA polymerase molecule attached to the surface. Incorporation of fluorescent nucleotides into a single strand of DNA is monitored in these ZMWs using a high numerical aperture lens and single photon-sensitive CCD array. The SMRT system generates relatively large numbers (100 million) of long sequence reads with average lengths of 3 kb with some sequences extending up to 16 kb, coupled with a very fast read time of 8 h for an entire run (Eid et al. 2009, Korlach et al. 2010).

The recent global health efforts to sequence and analyze the genomes of the strains of *Vibrio cholerae* and *Escherichia coli* in outbreaks in Haiti and Germany respectively have highlighted the considerable power of SMRT sequencing in global medicine (Chin et al. 2011, Rasko et al. 2011). The ability to sequence microbial genomes in hours (rather than days) with long read lengths is the emerging characteristic of third-generation DNA sequencing technologies such as the Pacific Biosciences platform. Longer reads facilitate easier and faster genome-assembly efforts and provide greater insights into structural variations.

In addition to providing genomic sequences, the Pacific Biosciences technology is unique amongst all others in that it allows the mapping of DNA modifications as an integral part of the sequencing protocol. DNA base modifications are important for understanding of biological processes such as gene expression, DNA damage, and DNA repair. The SMRT Analysis pipeline determines the rate of DNA base incorporation during sequencing and transforms the kinetic information for each nucleotide addition into base calls. This permits the platform to accurately distinguish between modified and native bases (Korlach & Turner 2012). Currently, this approach is applicable to only smaller microbial genomes, but advances in capture technologies will open this up to larger genomes, such as the zebrafish. At this time, there are no records in the literature of genomic or transcriptomic sequencing of zebrafish using Pacific Biosciences technology. Although Pacific Biosciences generates extremely long reads (up to 16 kb), the single-pass error rates are high and have been widely publicized (85% accuracy – 15% error) (Koren et al. 2012, Zhang et al. 2012). Coverage at a depth of 20×, however, guarantees a consensus accuracy of 99.999%, (Q50; Koren et al. 2012). As a consequence, the platform has been used to date exclusively for studies of small genomes.

Recent advances in long insert (~20 kb) DNA library preparation methods and Pacific Biosciences P5-C3 sequencing chemistry yielded the first shotgun human genome sequence dataset comprised entirely of long reads from a human DNA sample, the cell line (CHM11hert). The average read length was >7.6 kb, the longest read was 42.7 kb in size with 54× genome coverage. This long read approach will provide insights into structural genomic variation that are difficult to assess using short-read technologies.
All of the sequencing technologies described possess positive attributes. Illumina data, for example, are characterized by depth, high base accuracy, compensating for the relatively low-single molecule accuracy of the Pacific Biosciences data (Ribeiro et al. 2012). Conversely, Pacific Biosciences sequence data are generated without amplification, and can provide coverage in regions under-represented or absent in Illumina data due to amplification bias introduced during the sample preparation. In addition, Pacific Biosciences sequence reads, although lacking the depth of Illumina sequencing, yield read lengths up to 42 kb in length, with a median read length of > 7 kb, which complement the shorter 50–250 bp Illumina reads (Ribeiro et al. 2012). Pacific Bioscience sequencing has the potential to facilitate higher quality de novo assemblies of complex genomes including zebrafish, without reliance on conventional reference-guided assemblies. Future sequencing efforts in zebrafish will use a combination of approaches to delineate transcriptomic responses in response to EDC exposure.

Transcriptome sequencing
Massively parallel sequencing of the RNA content of a cell, tissue, or organism (RNAseq) is an innovative technique to interrogate global RNA expression profiles. The transcriptome reflects cellular activity within a tissue at a given point in time. Genome-wide expression studies, which are not influenced by deductive assumptions, can provide an unbiased approach for investigating the effects of endocrine disruption. RNAseq is rapidly becoming the method of choice for detecting and quantifying all the genes expressed in a cell (Maher et al. 2009, Wang et al. 2009, Ozsolak & Milos 2011).

The zebrafish represents an attractive model for the study of transcriptome perturbations in response to EDCs, as its reference genome is well characterized (Howe et al. 2013). Zebrafish, by virtue of the conservation of many key transcription factors, kinases and other regulatory proteins in its genome and the human genome, is an excellent proxy for assessing the effects of contaminants in the marine environment and for health risk assessment. Coupling high-throughput transcriptomic approaches to the zebrafish model has the potential to provide a powerful tool for comprehensive toxicology studies to properly dissect the immediate and transgenerational elements of EDC action. Deep transcriptome sequencing can potentially uncover the minimum required dosage of an EDC needed to obtain a phenotype, and uncover the mechanism of action of the EDC. This approach has the potential to reveal what dosage levels of EDC pose a threat to human health. In addition, RNAseq of zebrafish in response to EDCs has the potential to provide potentially critical diagnostic markers and targets for therapeutic intervention.

RNAseq allows short fragments of cDNA to be sequenced and mapped onto the latest build of the zebrafish reference genome. Unlike microarray data analysis, RNAseq enables the identification of transcription initiation sites and new splicing variants, and it permits precise quantitative determination of exon and splicing isoform expression. Another advantage of RNAseq over microarray-based methods of detection is in its ability to identify novel transcripts. With most RNAseq approaches, RNA is extracted and PolyA+ mRNA molecules enriched using poly-T oligo-nucleotide attached magnetic beads. The mRNA is fragmented and converted to double-stranded cDNA. Illumina (or other platform) platform-specific adaptors are ligated to the DNA and HTS is carried out. With Illumina sequencing, short reads of 35–150 bp in length are generated and subsequently aligned to the reference genome. RNAseq provides much greater sensitivity compared with microarray analysis, particularly when interrogating low abundance transcripts. In addition, the improved ability to discriminate regions of high sequence identity and the more accurate digital measurements makes RNAseq an attractive proposition.

Bioinformatics considerations
The bioinformatics step in the analysis of data from massive parallel sequencing is a major task with many pitfalls. A schema for transcriptome sequencing and analysis is outlined in Fig. 3. Fortunately, a series of new software packages and bioinformatics pipelines optimized for the computational challenges of short-read sequencing have emerged in the past few years (Pepke et al. 2009, Trapnell & Salzberg 2009, Robertson et al. 2010, Wolf 2013). Two of these programs include Maq and Bowtie, which use a computational strategy known as ‘indexing’ to speed up their mapping algorithms. Maq is based on spaced seed indexing where a sequence read is broken into four segments or seeds of equal length (Li et al. 2008). These seeds are paired and stored in a lookup table. Each read is fragmented and seed pairs of seeds are used to query matching positions in the reference. Bowtie uses the Burrows–Wheeler transform to store a memory-efficient representation of the reference genome (Langmead et al. 2009).

After the sequence tags have been mapped, the next step typically involves transcript-level quantification. An analytical tool Cufflinks has found the widespread use for...
transcript analysis (Trapnell et al. 2010, Roberts et al. 2011a,b). Cufflinks does not make use of existing gene annotations during assembly of transcripts, but instead constructs a minimum set of transcripts that best describe the sequence reads present in the dataset. This approach allows Cufflinks to uncover alternative transcription and splicing that are not described by pre-existing gene models. A related program Cuffdiff employs the Cufflinks transcript quantification method and facilitates gene and transcript expression levels across different transcriptomes (Trapnell et al. 2013). The accuracy of this approach was tested via differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOX1A, which alters the expression levels of thousands of individual transcripts, in addition to isoform switching events in key cell cycle regulators. Cuffdiff searches for differentially expressed genes by estimating how many fragments derive from each isoform and subsequently converting these counts into isoform expression levels. Cuffdiff thereby facilitates testing for significant differences in isoforms. Cuffdiff relies on a β-negative binomial model to estimate the variance of the RNAseq data for transcript analysis using t-like statistics from FPKM (fragments per kilobase of exon per million fragments mapped) values.

A different approach to RNAseq analyses uses count-based approaches. These methods consider the total output of a locus, without regard to any isoform diversity that may be present. Many tools have been developed for differential expression of read counts, but two of these have gained in popularity, edgeR and DESeq (Anders & Huber 2010, Robinson et al. 2010, Anders et al. 2013). The catalyst for development of these tools has been the need for robust calls of differential transcript expression when working with low numbers of biological replicate samples. Early analysis of RNA employed the Poisson distribution to test for differential expression. The rationale behind this approach was that if sequence reads were independently sampled from a population with given, fixed fractions of genes, the read counts would follow a multinomial distribution, which could be approximated using the Poisson distribution (Marioni et al. 2008, Wang et al. 2010). The problem with the application of this approach is that the Poisson distribution is too restrictive, predicting smaller variation that actually existed with the data. In addition type I errors (the probability of false discoveries) were not adequately controlled. Both EdgeR and DEseq employ a negative binomial distribution which better addresses this issue. A limitation with EdgeR compared with DEseq is that weakly expressed transcripts may be over-represented and highly expressed transcripts may not be called as differentially expressed (Anders & Huber 2010).

Platform evaluation for zebrafish transcriptomic analysis

At this point in time, a hybrid approach is recommended for the analyses of zebrafish transcriptomes. The sequencing platform should be evaluated on the basis of the goals of the sequencing experiments. If the objective is to understand how environmental pollutants perturb gene expression and alter endocrine signaling pathways, an approach that yields very deep sequencing coverage is recommended (Illumina/Ion Torrent). Biomarkers for pollutant exposures using embryonic, adult male and female zebrafish exposed to different pollutants across a concentration gradient can be performed in parallel. One lane of an Illumina flow cell can generate >200 million reads and this readily permits multiplexing of different samples. The sequencing depth can be optimized to facilitate the analysis of low expressed transcripts. In our experience, 50 million reads permits the analysis of low abundance transcripts. The rationale behind these experiments would be to expose fish to concentrations sufficient to obtain a strong signal for biomarker discovery, guided initially by levels that have been used in the literature to provide sufficient regulation of gene expression and ultimately assessing physiological contaminant levels. If an approach that requires de novo transcriptome assembly or an examination of mRNA variant expression in response to environmental pollutants, a mix of Illumina deep sequencing and Pacific Biosciences long read sequencing should be evaluated.
Conclusions

Ongoing developments promise further advances in the application of RNAseq, particularly direct RNAseq and approaches that allow RNA quantification from very small amounts of cellular materials (Maher et al. 2009, Wang et al. 2009, Ozsolak & Milos 2011). High-throughput genome and transcriptome sequencing of zebrafish exposed to EDCs will provide important data sets to better comprehend the risks posed to human health. The zebrafish model is attractive because its short life-span facilitates longitudinal studies. These HTS data sets will provide a better understanding of the number of impacted future generations and their potential for recovery. HTS data sets from embryos and tissues from adults exposed to physiological levels of EDC will provide insights on how developmental mechanisms that integrate genetic and epigenetic interactions are perturbed by EDCs. The development and validation of epigenomic classifiers and biomarkers using zebrafish will provide an important foundation for understanding the molecular basis of EDC toxicity in humans, yielding crucial data on the multigenerational effects of transient exposure during critical stages of development. The data obtained from these studies will permit development of diagnostic and prognostic tests for determining the epigenetic impact of EDCs on human health and for regulating exposure.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this review.

Funding

G H gratefully acknowledges support from NIH grant numbers DK063491, CA023100, and DK080506 and UC Senate grant number RK126HHARDIMAN.

References


McEwen BS 2012 Brain on stress: how the social environment gets under the skin. PNAS 109 (Suppl 2) 17180–17185. (doi:10.1073/pnas.1121254109)


Review

M E BAKER and G HARDIMAN

Transcriptional analysis of endocrine disruption

52:3 R255


Received in final form 7 March 2014

Accepted 28 March 2014