Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Toxicology

Next-generation sequencing data for use in risk assessment

Bruce Alexander Merrick

Abstract

Next-generation sequencing (NGS) represents several powerful platforms that have revolutionized RNA and DNA analysis. The parallel sequencing of millions of DNA molecules can provide mechanistic insights into toxicology and provide new avenues for biomarker discovery with growing relevance for risk assessment. The evolution of NGS technologies has improved over the last decade with increased sensitivity and accuracy to foster new biomarker assays from tissue, blood, and other biofluids. NGS technologies can identify transcriptional changes and genomic targets with base pair precision in response to chemical exposure. Furthermore, there are several exciting movements within the toxicology community that incorporate NGS platforms into new strategies for more rapid toxicological characterizations. These include the Tox21 *in vitro* high-throughput transcriptomic screening program, development of organotypic spheroids, alternative animal models, mining archival tissues, liquid biopsy, and epigenomics. This review will describe NGS-based technologies, demonstrate how they can be used as tools for target discovery in tissue and blood, and suggest how they might be applied for risk assessment.

Addresses

Molecular and Genomic Toxicology Group, Biomolecular Screening Branch, Division National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC, 27709, USA

Corresponding author: Merrick, B. Alex (merrick@niehs.nih.gov)

Current Opinion in Toxicology 2019, **18**:18–26

This review comes from a themed issue on **Genomic Toxicology**

Available online 8 March 2019

For a complete overview see the Issue and the Editorial

https://doi.org/10.1016/j.cotox.2019.02.010

2468-2020/Published by Elsevier B.V.

Keywords

Next generation sequencing, RNA-seq, DNA-seq, Risk assessment, Biomarkers, High throughput transcriptomics, Liquid biopsy, Transcriptomics.

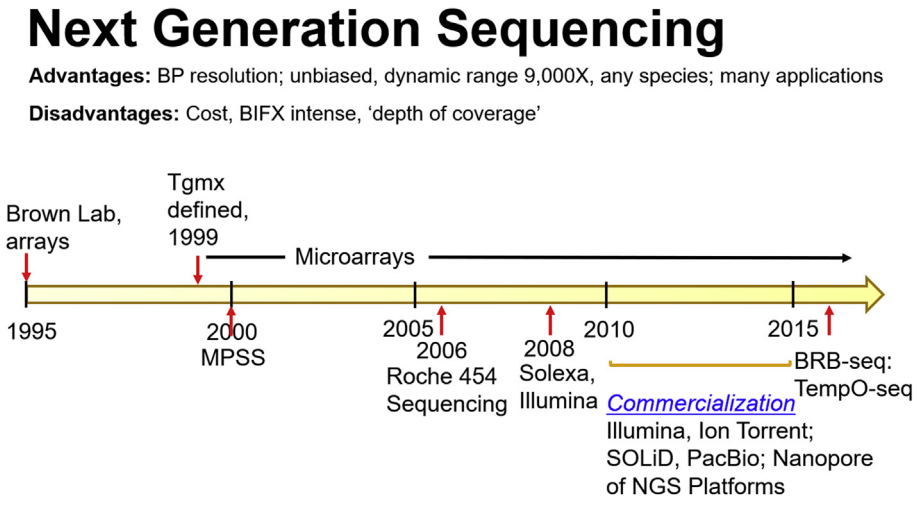
Introduction

The Sanger sequencing method was developed in the late 1970s to analyze DNA sequence using 32P-labeled nucleotides separated by polyacrylamide gels for autoradiograms [1]. Radionuclides were replaced by fluorescently labeled nucleotides and capillary electrophoresis that gave rise to automated sequencing instruments such as the Applied Biosystems, Inc. model 370A sequencers and others, on the basis of Sanger chemistries [2]. Read length per each sample was at 500–800 nucleotides, and sample throughput was limited.

Sanger-based DNA sequencing instruments are considered first-generation platforms. Instruments that perform multiple sequencing reactions simultaneously in a 'massively parallel' fashion have been dubbed, 'NextGeneration' or NextGen Sequencing [3]. How NGS came about compared with other genomic platforms is interlinked with microarray technology (Figure 1). Both platforms can provide whole genomic approaches to research problems. Microarrays are a fluorescent probe hybridization-based technology with origins in the mid-1990s that are now a mature genomic platform with a well-established data analysis pipeline. Downsides of microarrays are that a prior genomic knowledge is needed to generate probes which are species-specific with a limited dynamic range for differential expression.

Development of the second wave of sequencing technologies, termed NextGen sequencing (NGS) technologies, has overlapped with microarray platforms (Figure 1). NGS began in the new millennium as exemplified by the massively parallel signature sequencing (MPSS) system that came from university research. Improvements in sequencing chemistries, detection, and automation over the next decade promoted a rapid development of NGS platforms. From 2010 to 2015, many NGS instruments became commercially available that could produce millions of reads from 100 to 1000 bases in length. A read is a short piece of sequence (e.g. 100 nucleotides) that can be aligned to a transcript, and it also serves as a quantitative measure of a transcript when summed up with other aligning reads. These second-generation sequencers include the Roche '454 FLX,' Life Technologies 'Ion Torrent,' Applied Biosystems, Inc. 'SOLiD', and Illumina family of sequencers including the HiSeq 2000 series, MiSeq, X-Ten, and NovaSeq [4]. Further advances, such as single-molecule real-time sequencing, in NGS sequencing technology have led to longer read sequencers such as the Pacific Biosciences 'PacBio RS II' instrument that produces reads greater than 10,000 bases [3].

Figure 1



Timeline for development of microarray and next-generation sequencing (NGS) technology platforms. Microarray developments are above the timeline and NGS activities are below. For microarray development, Brown's laboratory at Stanford was one of the first to develop a multigene expression measurement system using fluorescent detection. The term, toxicogenomics (Tgmx) was first defined by Nuwaysir et al., in 1999 [67]. Commercialized platforms such as Affymetrix, Agilent, and NimbleGen matured through 2010. For NGS, the massively parallel signature sequencing (MPSS) was developed in 2000 by the Brenner lab at Lynx Therapeutics. 454 Life Sciences developed a massively parallel pyrosequencing method in 2006 followed by a commercial instrument put out by Roche. The Solexa short-read platform was acquired by Illumina in 2008 and has undergone continued development and improvement. Commercialization of NGS platforms continues with various speeds of analysis, read lengths, and sequencing capacities. More recent developments include BRB-seq or 'Bulk RNA Barcoding and Sequencing' and TempO-Seq by BioSpyder as a library of bar-coded probes that hybridize to representative gene transcripts as a targeted NGS approach to transcript expression. Advantages and disadvantages are summarized and discussed further in the text.

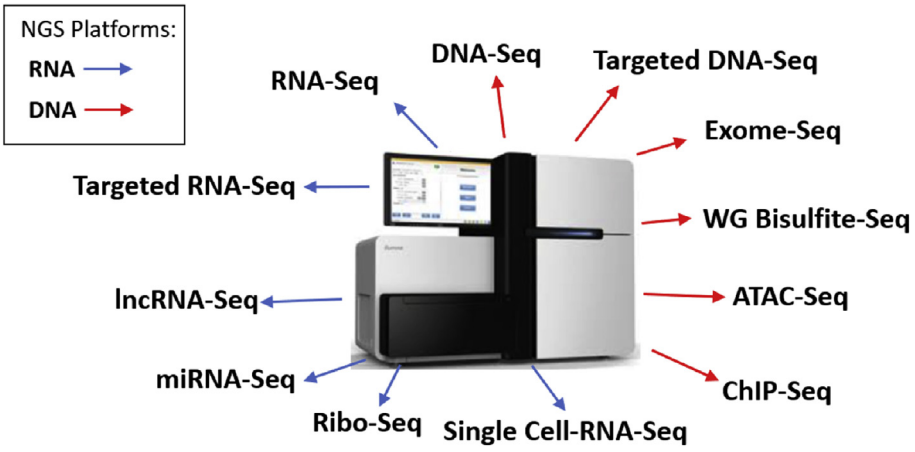
A more recent sequencing technology has been advanced by Oxford Nanopore Technologies. Nanopore instruments read bases directly from single DNA or RNA molecules through a biological nanopore channel—a nanoscale biological tube that sequences by sensing changes in ionic current as the nucleic acid molecule passes through [5]. The sequencing devices can provide rapid analysis (hours), and some units are portable (size of an USB flash drive) that can be readily applied to teaching laboratories, medical offices, and field work. Reads lengths can be in the tens to hundreds of kb in length. A primary advantage of long read length is to reduce the ambiguity of highly homologous genes, splice variants, and repetitive regions in the genome where alignment is inherently more difficult using short reads. The high sequence resolution of NGS instruments has come at the expense of relatively low sample throughput. This issue has been addressed by creating libraries of targeted probe sets that analyze the complete transcriptome (e.g. TempO-Seq [6] and bulk RNA coding-seq [BRB-seq], reviewed later), rapidly and at relatively low cost. A brief depiction of NGS applications is shown in Figure 2.

NGS and risk assessment

Traditional risk assessment often involves identifying hazard(s) in a dose—response manner after chemical or test article exposure in animal models or human data if available [7]. Test articles can include chemicals and many other agents, including pharmaceuticals, drugs, natural products, particles such as asbestos, nanoparticles, physical factors such as radiation, metals, and many others. The type of toxicity or hazard can be widely defined as macroscopic or microscopic lesions and pathologies, altered pharmacologic, immunologic, functional, and behavioral reactions, changes in biochemistry and physiology, or any measurable response that is considered adverse or outside of normal health. Study of the underlying molecular changes contributing to toxicity has been greatly facilitated by omics technologies, particularly transcriptomics, while standardization of data analysis and interpretation continue to be refined [8]. There are many *in vitro* assays and screens (e.g. anticholinesterase activity or bacterial mutagenesis) that support the mode of action in the risk assessment process, but new initiatives such as Tox21 aim to develop new assays incorporating NGS platforms for a larger role in risk determination [9].

The dynamic nature of gene expression (transcriptomics) in response to a chemical or test article exposure makes it well suited as part of the hazard identification and dose-setting process for risk assessment [10,11]. There are approximately 15,000 coding genes and probably an equal number of noncoding genes expressed at any one time in a specific cell type. Splice

Figure 2



Many applications for measuring RNA and DNA in toxicogenomics are supported by NGS platforms. Whole genome or transcriptome analysis or targeted portions of each can be measured by NGS. seq, sequencing; WG, whole genome; ATAC, assay for transposase-accessible chromatin; ChIP, chromatin immunoprecipitation; Ribo, ribosome; miRNA, microRNA. RNA analysis platforms are indicated by blue arrows, and DNA analysis platforms are shown by red arrows. Further description of these applications is provided in the text.

variants also add more complexity to response. Of those expressed genes, only a proportion may change in response to chemical exposure. A robust transcriptomic response may number in the thousands of altered transcripts, while a low level of response may differ by a few hundred transcripts or less. The field has been greatly assisted by genomic dose–response analysis software (e.g. BMDExpress) to facilitate use of transcriptomic data in toxicology and risk assessment [12].

RNA-seq

RNA-seq is the principal NGS platform for transcriptome analysis [13,14]. Unlike microarrays, transcript sequencing can occur without prior genomic knowledge, although accurate alignment is greatly enhanced by genome data assemblies. RNA-seq performs tens or hundreds of thousands of small-scale DNA sequencing reactions (cDNA converted from RNA) that produce relatively short sequences (reads) of 100–400 bases in length that in aggregate represent the transcriptome after alignment to a reference genome. About 95% of total RNA isolates are ribosomal RNA (rRNA), and because it provides little value, rRNA must be removed either by using poly(A) enrichment or by rRNA depletion strategies [15]. In most tissues, the transcriptome is primarily composed of mRNA that is translated into protein; however, there is a substantial portion of RNA that is not protein coding, expressed as noncoding RNA (ncRNA), that can be detected by RNA-seq. Such ncRNA includes microRNA, long non-coding RNA (lncRNA), and other specialized small RNAs (Table 1). mRNA or ncRNA are reverse transcribed into cDNA, and then, a library of cDNA fragments is constructed for each sample with short adaptor sequences attached to either fragment end. RNA libraries can be sequenced in one direction single-end reads and also from the opposite direction (paired-end reads). Paired-end reads provide a much more accurate alignment but at more expense.

Table 1 Transcriptome transcript classification[a].

Transcript type	Genomic number	Mature size	Examples
mRNA – coding	20–25,000	500–15,000nt	TP53, GAPDH
ncRNA – miRNA	2–5000	22nt	miRNA-29, miR-122
ncRNA - lncRNA	>30,000	>200nt	HOTAIR, PVT1
small ncRNA - regulatory	1000	20–100nt	tRNA, rRNA, siRNA

HOTAIR, HOX antisense intergenic RNA; PVT1, plasmacytoma variant translocation 1; GAPDH, glyceraldehyde-3-phosphate dehydrogenase.
[a] The transcriptome comprises coding and noncoding transcripts (ncRNA), including microRNAs (miRNAs); long noncoding transcripts (lncRNA); and regulatory small ncRNAs (e.g. transfer RNA, ribosomal RNA, small interfering RNA).

The number of reads per sequencing lane or sequencing run varies with each NGS platform, and the number of samples for RNA-seq analysis can be mixed and distinguished by a multiplexing process called 'bar coding.' Statistical confidence in differential transcript expression is increased by devoting a requisite number of reads per sample for adequate 'depth of coverage' of the transcriptome [14]. Sequence coverage is the number of times each base within the transcriptome is sequenced; so, if each base within the transcriptome on average was sequenced 10 times, the coverage would be 10-fold. Coverage needs for differential expression using RNA-seq vary from 15- to 30-fold depending upon the organism and reference genome. However, the detection of rare transcripts or inferring single-nucleotide polymorphisms (SNPs) from RNA-seq data may require significantly more reads from 100- to 200-fold coverage [16]. It should be acknowledged that the expense, depth of transcriptome coverage, greater data complexity, greater demands for computational analysis, and computational infrastructure are considerations in RNA-seq analysis.

One of the landmark studies in comparing RNA-seq with the microarray platform for chemical exposure studies involved the Sequencing Quality Control project that examined thousands of microarrays and RNA-seq analyses to compare differential expression profiling and quality metrics for each platform [17]. In general, the consortium found good agreement between RNA-seq and microarray relative to gene expression, despite some data variability in low-expression genes that could be attributed to differences in expression platforms and data analysis pipelines. A follow-up study by this working group compared transcriptomes from RNA-seq and microarray data on 498 primary neuroblastomas and showed that RNA-seq outperforms microarrays in terms of overall transcript characterization, but both platforms show similar results in clinical endpoint prediction [18].

The sensitivity and discovery potential of RNA-seq has found applications in biomarker discovery and environmental monitoring that can be relevant for many stages of risk assessment. For example, nine chemical pollutants were screened in undifferentiated mouse embryonic stem cells in a cell-based toxicity assay in which RNA-seq identified novel RNA biomarkers including ncRNAs that showed substantial response to *in vitro* chemical exposure [19]. Bis-(2-ethylhexyl)-tetrabromophthalate, a widely used commercial chemical, was screened *in vitro* in a fish embryo system (Atlantic killifish) by RNA-seq that related transcriptional and pathway changes to developmental endpoints as part of environmental risk assessment [20]. A very innovative study profiled liver expression responses from mice in conventional and germ-free conditions exposed to the persistent environmental contaminants, polybrominated diphenyl ethers, to determine effect of presence or absence of gut microbiomes on toxicity and gene expression [21]. These authors found several protein-coding lncRNA pairs that may serve as specific biomarkers to distinguish various polybrominated diphenyl ether congeners and shed light on chemical effects and toxicity related to changes in the microbiome. In another RNA study, livers from rats subchronically exposed to aflatoxin B1 (AFB1) in diet showed differential expression of 25 new lncRNAs to exposure that were discovered as candidate predictive biomarkers of hepatocellular carcinomas [22]. RNA-seq can also detect microRNAs, as recently described in the rat microRNA body atlas [23], for detecting specific biomarkers such as miR122 that is released from hepatocytes into biological fluids after chemically induced liver injury [24]. Other recent studies have similarly used the sensitivity and base-pair resolution power of RNA-seq for differential expression and pathway changes to suggest modes of action for toxicity with estradiol [25] or pharmaceutical water contaminants [26] in zebrafish, with diesel fractions [27] or the flame-retardant contaminant, tetrabromobisphenol A [28], in zebrafish embryos *in vitro*.

High-throughput transcriptomics

Even though RNA-seq can provide a detailed measurement of the transcriptome, it is not a high-throughput platform. Tox21 is interagency program to develop and encourage high throughput *in vitro* screening and advanced computation methods to better predict toxicological effects of chemical exposure [29]. In the past few years, high-throughput transcriptomics or 'HTT' has been developed to measure thousands of transcripts for differential expression after chemical exposure, in a highly multiplexed fashion that accommodates thousands of samples to establish concentration—response relationships at the gene and pathway level [6,30]. How was this accomplished? A library of transcript-specific probes can be synthesized that bind to RNA in a hybridization-ligation reaction. Indexing sequences are unique to each transcript and sample. Sample libraries can be mixed for simultaneous analysis on an NGS sequencing instrument with a sensitivity of detection in the picogram range for RNA. This sensitivity for RNA transcripts means that transcriptional changes can be measured using thousands or just hundreds of cells. RNA-seq methods have been developed to profile single-cell transcriptomes of known and novel cell types in complex tissues such as kidney [31]. The sensitivity of NGS methods has been adapted for toxicant screening in 96- or 384-well plates for high-resolution concentration—response assessment to chemical exposures [32]. Several articles have demonstrated its application in toxicity screening. For example, six compounds were screened in differentiated kidney primary renal proximal tubule epithelial cells (RPTEC) or liver HepaRG cells in a time- and concentration-related

manner using a 2800 transcript panel to discriminate compound and cell type–specific responses [33]. The development of liver spheroids comprising 1000 to 2000 cells has an increased metabolic capacity over two-dimensional cells in flat culture. The high sensitivity of HTT can be exploited for rapid, high-throughput concentration–response studies with multiple chemicals [32,34].

While the whole transcriptome can be screened in each HTT analysis, considerable sequencing must be performed to deliver a statistical level of confidence for gene expression, especially for low copy number transcripts. As a result, a strategy of selecting a transcriptomic subset of genes has been developed into a platform called the S1500+, or 'Sentinel' 1500 [30]. This platform evolved as a hybrid approach of combining (1) the L1000 platform; (2) transcripts using a toxicogenomic data-driven method from public databases for selecting the most responsive transcripts; and (3) expert contributed genes [30]. The S1500 + platform represents a biological space reflecting a diverse pharmacologic and toxicity gene expression that represents all known canonical pathways from the Molecular Signature Database and can infer changes from the remainder of the transcriptome. A study comparing the S1500 + gene set with RNA-seq and microarray rat liver mode of action samples demonstrated that the S1500 + platform results are consistent with findings performed with genome-wide platforms (e.g. microarray, RNA-seq) for measuring genome-wide transcriptional responses [35]. Another aspect of the HTT approach for risk assessment can be the comparative screening of specific cell types *in vitro* (e.g. liver, kidney, heart, neurons) to test articles during the same experiment. For animal testing, it will eventually be possible to monitor transcriptional changes in all tissues of test article–exposed animals to accompany histopathologic and clinical chemistry evaluations.

Archival Transcriptomics

Toxicology studies with archival specimens such as formalin-fixed and paraffin-embedded (FFPE) tissues comprise an invaluable resource for linking histopathologic diagnosis to gene expression profiles. Establishing molecular and pathologic relationships can provide a basis for risk assessment on the basis of linking established molecular pathways with chemical pathologies and disease. Advanced procedures for deparaffinization and enzymatic digestion have been developed to release nucleic acids for extraction and purification from slices of paraffin blocks after years in storage [36].

The nonspecific hybridization and higher background from microarray analysis of archival blocked samples have encouraged researchers to use RNA-seq for transcriptomic profiling. Several studies have successfully used NGS transcript profiling technologies on archival samples. One study showed genomic signatures and gene set analysis in AFB1 differentially expressed transcripts that were highly comparable for matched fresh frozen and FFPE tissues [37]. Subsequent studies have shown a conservation of gene expression patterns in FFPE and frozen tissue samples, especially when rRNA depletion procedures were used [38]. A comprehensive analysis of archival liver sample sets involving di(2-ethylhexyl) phthalate or dichloroacetic acid, varying from 2 to 20 years storage, showed remarkably high correlation in dose–response of differentially expressed genes, despite challenges of lower read counts from the older study [39]. In particular, the more recently (2 year) archived FFPE sample data were highly similar to frozen sample transcriptional data regarding sequencing quality metrics, differential expression, and dose–response relationships [39].

DNA sequencing

The base-pair resolution of NGS platforms is uniquely positioned to detect mutations and single-nucleotide variations (SNPs) related to genomic changes and health hazards posed by chemical exposure. Targeted sequencing of molecular sensors of genotoxic and cellular stress such as TP53 [40]; multiple Omic analyses involving genomic, epigenomic, and transcriptomic characterization of the mutagen, 1,3-butadiene, in mouse strains [41]; and research on genomic interrogation of oxidative DNA damage [42] are representative studies that have applied multiple NGS platforms to more completely describe adverse effects of chemical exposure across the genome. Unlike the dynamic nature of RNA transcription to hazardous substance exposures, the stability of DNA in human, rodent, and other animal model systems does not as readily lend itself to rapid chemical screening for genomic changes in risk assessment. Genetic toxicology uses a range of screening tests for DNA damage including the Ames assay, the micronucleus test, the Comet assay, and several chromosomal aberration and DNA damage and repair assays [43].

There is increasing interest in the role that high-throughput screening and NGS platforms might play in creating a new generation of tests for genotoxicity and genetic susceptibility to disease and chemical hazards [44,45]. An important part of NGS approaches is the ability to interrogate the whole genome or the exome, where changes in coding regions of genes could alter translational protein products and damage regulatory processes. In the past, human genetic and epidemiological studies were limited to a candidate gene approach to establish genomic disease and toxicity relationships. Now, whole-genome sequencing can survey SNPs, copy number variation, and chromosomal aberrations with increasing accuracy. Sequencing studies of humans and experimental species (e.g. mouse) provide publicly available data to better estimate 'normal'

sequence variation (normal phenotype) which is critical for distinguishing those variants leading to environmental disease. Inherent in this task of genetic variation is discerning germline variants (heritable sequences) from somatic variants acquired in DNA of tissues during life. NGS in forward genetic screens in mice is one experimental approach to help sort out candidate genes and mutations that correspond to specific phenotypes [46]. The dbSNP public database maintained by the NCBI is a catalog of single-nucleotide variants, small-scale deletions or insertions, and short tandem repeats such as microsatellites [47,48]. As of 2017, the NCBI will only accept human data variant submissions, whereas EBI's European Variation Archive will continue to accept data and host the collection of nonhuman data variants. dbGaP is the NCBI public database that archives genotype–phenotype associations from many sources such as genome-wide association study data, short read archive, molecular diagnostic assays, and others.

Whole-genome sequencing

The cost of whole-genomic sequencing has crossed below the one thousand dollar barrier. In the wake of decreasing costs of NGS platforms, considerable public resources are now being devoted to sequencing thousands of human genomes to benefit personalized medicine and understanding genetic susceptibility. For example, the 'All of Us' project sponsored by the National Institutes of Health aims to fund several centers for complete genomic sequencing on one million or more people to interrelate effects of genetics, environment, and life style [49]. As public sequencing projects get underway, the varying levels of sequencing depth are required for how genomic information can be used with confidence for distinguishing germline variants and rare variants from somatic variation, use in clinical decisions, surgery and therapeutics; use in genetic counseling; or use in advising patients with risk factors, either known or suspect [50]. Not all variants will have risk; the risks may yet be undiscovered; variants may be multifactorial in disease risk; or variants may be protective, counteracting the disease potential posed by other variants. Pharmacogenetics and pharmacogenomics are research areas where NGS may inform both clinical and research efforts for therapeutic and chemical exposure risks [51].

Genomic material for sequencing is generally collected from blood or oral swabs, and such procedures offer a noninvasive ease of collection [52] but may have limitations. First is that many disease phenotypes occur in tissues or organs far away from the collection site so critical sequence variants may not be readily observed when obtained from blood or pharynx. Second, initiation of early-stage disease may begin in a single cell or small cluster of cells so that detection of sequence variants may require cell enrichment or greater depths of DNA sequencing. Despite these concerns, there is enthusiasm for use of noninvasive or minimally invasive (e.g. blood) DNA or RNA for NGS and genotyping [53,54].

Exome sequencing

Because coding regions comprise about 2% of genomes, designing probe sets to capture and sequence only those coding genomic regions provides an efficient way to examine sequence variants in the most consequential regions of the genome where changes may be linked to abnormal phenotypes and disease without sequencing the entire genome. The other advantages to this approach are the greater depth of sequencing possible with each sample (e.g. 50- to 100-fold or more) compared with whole-genome sequencing and the greater number of samples possible for analysis in a sequencing run [55]. Thus far, use of exome sequencing for assessment of risk has been clinically focused on prenatal and reproductive medicine [56] as well as cancer diagnosis and treatment [57]. Use of exome sequencing in environmental risk assessment lies in the future.

Duplex sequencing

Duplex sequencing is a highly specific NGS method for detecting rare sequence variants and mutation with frequencies as low as one in 10 million [58]. Specific adapters and tags can uniquely identify reads from each strand of DNA. Although there is considerable sensitivity in NGS platforms, sample preparation and polymerase amplification error rate contribute to substantial noise that obscures low-frequency mutations. For example, duplex sequencing was used to determine TP53 mutations in peritoneal fluid samples from women with ovarian carcinomas and control individuals without cancer [59]. Findings showed nearly all patients with and without cancer (35/37 total) had low-frequency TP53 mutations that were more abundant with cancer, clustered in hotspots, and increased with age. Widespread, age-associated somatic TP53 mutations in noncancerous tissue suggests overall mutational burden even in normal individuals. Such TP53 mutations could also be detected in peripheral blood samples. Another study identified a characteristic mutation spectrum for the liver carcinogen, AFB1, months before tumors were detectable in a mouse model using duplex sequencing [60]. The AFB1 spectrum proved clinically useful in accurately identifying a subset of cancers associated with AFB1 exposure from a larger set of human liver tumors. Use of duplex sequencing as a measure of mutational load in sensitive genomic regions due to environmental chemical exposure remains to be explored.

Liquid biopsy

Liquid biopsy refers efforts to detect and monitor disease or toxicity in accessible biofluids, notably in blood,

because of the relative ease of sampling that can be performed repeatedly overtime. Release of cell-specific miRNAs during chemically induced toxicity can be exploited by NGS for a liquid biopsy approach to assess risk to chemical exposure overtime [61]. For example, miRNA-seq analysis of urine in rats sampled after 1 week exposure to the renal toxicant, gentamicin, found 227 unique miRNAs of which 146 were differentially expressed, with nine being novel miRNAs not found on a primer-designed qPCR platform [62]. In addition to circulating miRNA, NGS analysis of plasma DNA used for tumor diagnosis in oncology might be similarly applied in toxicology. Circulating, cell-free DNA (ccfDNA) comprises short fragments extracellular DNA ($\sim$180bp) that normally circulate in blood at low levels (e.g. 1–5 ng/ml) in healthy individuals. ccfDNA is derived from leukocytes and tissue apoptosis and cell turnover from all tissues. However, in many tumors, the amount of ccfDNA is increased (10–100 ng/ml), and a portion may harbor diagnostic cancer mutations [63]. Use of ccfDNA has gained attention in the diagnosis, staging, and biomarker discovery for many types of tumors [64] and also many other diseases [65] (e.g. autoimmune and infectious diseases) as a novel, minimally invasive form of 'liquid biopsy', an attractive alternative to needle biopsy. Another exciting development in this field is the epigenetic analysis of ccfDNA for cancer diagnostics and determining tissue of origin along with somatic mutations [66]. To date, ccfDNA has been little studied in environmental health sciences. However, for those exposures that leave a somatic mutation or epigenetic pattern, NGS analysis of ccfDNA at the exome, whole genomic, or epigenomic level could provide new data on the amounts and types of environmental exposures overtime in experimental and epidemiological settings for improved risk assessments.

Summary

Transitioning from the clinic to environmental carcinogenesis research, the impact of biomarker research is shifting from therapeutics and companion diagnostic development toward risk assessment and early detection of chemical exposure and disease-related changes. Biomarkers evaluating risk assessment directly relate to environmental regulation and can be used to help define amounts and the type of environmental chemical exposures, biomarkers of effect, and biomarkers of susceptibility, all of which reflect the interactions between the environment and the population. Thus, NGS data can contribute to an understanding of the environmental and/or genetic factors that could lead to potential adverse health effects and have a very positive impact on the future of chemical risk assessment.

Acknowledgements

This work was supported by the Division of the National Toxicology Program at the National Institute of Environmental Health Sciences under grant Z99 ES999999.

Conflicts of interest

Nothing declared.

Funding

No funding was received for this work.

Disclaimer

This article is the work product of employees of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH); however, the statements, opinions, or conclusions contained therein do not necessarily represent the statements, opinions, or conclusions of NIEHS, NIH, or the United States government.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**: 5463–5467.

2. Heather JM, Chain B: **The sequence of sequencers: the history of sequencing DNA**. *Genomics* 2016, **107**:1–8.

3. • Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies**. *Nat Rev Genet* 2016, **17**:333–351.

A current overview of NGS technologies and applications, including a comparison against microarrays, and subgenomic platforms like NanoString, qPCR and Optical mapping. Different NGS instrument and sequencing platforms are described ranging from DNA-seq, RNA-seq, ATAC-seq and others.

4. Mutz KO, Heilkenbrinker A, Lonne M, Walter JG, Stahl F: **Transcriptome analysis using next-generation sequencing**. *Curr Opin Biotechnol* 2013, **24**:22–30.

5. Feng Y, Zhang Y, Ying C, Wang D, Du C: **Nanopore-based fourth-generation DNA sequencing technology**. *Genom Proteom Bioinform* 2015, **13**:4–16.

6. Yeakley JM, Shepard PJ, Goyena DE, VanSteenhouse HC, McComb JD, Seligmann BE: **A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling**. *PLoS One* 2017, **12**:e0178302.

7. McCarty LS, Borgert CJ, Posthuma L: **The regulatory challenge of chemicals in the environment: toxicity testing, risk assessment, and decision-making models**. *Regul Toxicol Pharmacol* 2018, **99**:289–295.

8. Sauer UG, Deferme L, Gribaldo L, Hackermuller J, Tralau T, van Ravenzwaay B, Yauk C, Poole A, Tong W, Gant TW: **The challenge of the application of 'omics technologies in chemicals risk assessment: background and outlook**. *Regul Toxicol Pharmacol* 2017, **91**(Suppl 1):S14–S26.

9. • Tice RR, Austin CP, Kavlock RJ, Bucher JR: **Improving the human hazard characterization of chemicals: a Tox21 update**. *Environ Health Perspect* 2013, **121**:756–765.

A review of the Tox21 interagency program, the goals and accomplishments for Phase I and Phase II is detailed. High throughput screening assays in 1536 well format are described to form 15 concentration-response curves to a 10K (10,000 environmental, pesticide and pharmaceutical compounds) in a variety of nuclear receptor and stress reponse reporter assay systems.

10. Bourdon-Lacombe JA, Moffat ID, Deveau M, Husain M, Auerbach S, Krewski D, Thomas RS, Bushel PR, Williams A, Yauk CL: **Technical guide for applications of gene expression**

profiling in human health risk assessment of environmental chemicals. *Regul Toxicol Pharmacol* 2015, **72**:292–309.

11. Wilson VS, Keshava N, Hester S, Segal D, Chiu W, Thompson CM, Euling SY: **Utilizing toxicogenomic data to understand chemical mechanism of action in risk assessment**. *Toxicol Appl Pharmacol* 2013, **271**:299–308.

12. Phillips JR, Svoboda DL, Tandon A, Patel S, Sedykh A, Mav D, Kuo B, Yauk CL, Yang L, Thomas RS, *et al.*: **BMDExpress 2: enhanced transcriptomic dose-response analysis workflow**. *Bioinformatics* 2018. https://doi.org/10.1093/bioinformatics/bty878. bty878.

13. Chen G, Shi T, Shi L: **Characterizing and annotating the genome using RNA-seq data**. *Sci China Life Sci* 2017, **60**:116–125.

14. Hrdlickova R, Toloue M, Tian B: **RNA-Seq methods for transcriptome analysis**. *Wiley Interdiscip Rev RNA* 2017, **8**.

15. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D: **Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion**. *Sci Rep* 2018, **8**:4781.

16. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data**. *Am J Hum Genet* 2013, **93**:641–651.

17. •• Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, *et al.*: **The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance**. *Nat Biotechnol* 2014, **32**:926–932.
Large consortium study comparing the RNA-seq and microarray platforms iwht 27 chemicals representing multiple modes of action (MOA). Cross platform concordance was very high but differentially expressed genes and pathways was affected by transcript abundance and biological complexity of MOA. The bioinformatic methods and data viscualization are very informative for analyzing large datasets and platform comparison.

18. Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, Wang J, Furlanello C, Devanarayan V, Cheng J, *et al.*: **Comparison of RNA-seq and microarray-based models for clinical endpoint prediction**. *Genome Biol* 2015, **16**:133.

19. Tani H, Takeshita JI, Aoki H, Nakamura K, Abe R, Toyoda A, Endo Y, Miyamoto S, Gamo M, Sato H, *et al.*: **Identification of RNA biomarkers for chemical safety screening in mouse embryonic stem cells using RNA deep sequencing analysis**. *PLoS One* 2017, **12**:e0182032.

20. Huang W, Bencic DC, Flick RL, Nacci DE, Clark BW, Burkhard L, Lahren T, Biales AD: **Characterization of the Fundulus heteroclitus embryo transcriptional response and development of a gene expression-based fingerprint of exposure for the alternative flame retardant, TBPH (bis (2-ethylhexyl)-tetrabromophthalate)**. *Environ Pollut* 2019, **247**:696–705.

21. Li CY, Cui JY: **Regulation of protein-coding gene and long noncoding RNA pairs in liver of conventional and germ-free mice following oral PBDE exposure**. *PLoS One* 2018, **13**:e0201387.

22. Merrick BA, Chang JS, Phadke DP, Bostrom MA, Shah RR, Wang X, Gordon O, Wright GM: **HAfTs are novel lncRNA transcripts from aflatoxin exposure**. *PLoS One* 2018, **13**:e0190992.

23. Smith A, Calley J, Mathur S, Qian HR, Wu H, Farmen M, Caiment F, Bushel PR, Li J, Fisher C, *et al.*: **The Rat microRNA body atlas; Evaluation of the microRNA content of rat organs through deep sequencing and characterization of pancreas enriched miRNAs as biomarkers of pancreatic toxicity in the rat and dog**. *BMC Genomics* 2016, **17**:694.

24. Kullak-Ublick GA, Andrade RJ, Merz M, End P, Benesic A, Gerbes AL, Aithal GP: **Drug-induced liver injury: recent advances in diagnosis and risk assessment**. *Gut* 2017, **66**:1154–1164.

25. Zheng Y, Yuan J, Meng S, Chen J, Gu Z: **Testicular transcriptome alterations in zebrafish (Danio rerio) exposure to 17beta-estradiol**. *Chemosphere* 2019, **218**:14–25.

26. Wu M, Liu S, Hu L, Qu H, Pan C, Lei P, Shen Y, Yang M: **Global transcriptomic analysis of zebrafish in response to embryonic exposure to three antidepressants, amitriptyline, fluoxetine and mianserin**. *Aquat Toxicol* 2017, **192**:274–283.

27. Mu X, Liu J, Yang K, Huang Y, Li X, Yang W, Qi S, Tu W, Shen G, Li Y: **0# Diesel water-accommodated fraction induced lipid homeostasis alteration in zebrafish embryos**. *Environ Pollut* 2018, **242**:952–961.

28. Chen J, Tanguay RL, Xiao Y, Haggard DE, Ge X, Jia Y, Zheng Y, Dong Q, Huang C, Lin K: **TBBPA exposure during a sensitive developmental window produces neurobehavioral changes in larval zebrafish**. *Environ Pollut* 2016, **216**:53–63.

29. Merrick BA, Paules RS, Tice RR: **Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective**. *Int J Biotechnol* 2015, **14**:7–27.

30. Mav D, Shah RR, Howard BE, Auerbach SS, Bushel PR, Collins JB, Gerhold DL, Judson RS, Karmaus AL, Maull EA, *et al.*: **A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics**. *PLoS One* 2018, **13**:e0191105.

31. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, Li M, Barasch J, Susztak K: **Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease**. *Science* 2018, **360**:758–763.

32. •• Ramaiahgari SC, Waidyanatha S, Dixon D, DeVito MJ, Paules RS, Ferguson SS: **Three-dimensional (3D) HepaRG spheroid model with physiologically relevant xenobiotic metabolism competence and hepatocyte functionality for liver toxicity screening**. *Toxicol Sci* 2017, **160**:189–190.
Human liver spheroid model demonstrates enhanced xenobiotic (CYP1A2, CYP2B6and CYP3A4/5) and functional capabilities. HepaRG liver spheroids can be screened in 384 well plates and show liver enzyme inducibility with activators of hepatic receptors to AhR, CAR and PXR. The 3D spheroids have a longevity in culture needed for repeated exposures and lab-to-lab and year-to-year repeatability for toxicology screening. Generated data could rapidly provide concentration-response data for risk assessment.

33. Limonciel A, Ates G, Carta G, Wilmes A, Watzele M, Shepard PJ, VanSteenhouse HC, Seligmann B, Yeakley JM, van de Water B, *et al.*: **Comparison of base-line and chemical-induced transcriptomic responses in HepaRG and RPTEC/TERT1 cells using TempO-Seq**. *Arch Toxicol* 2018, **92**:2517–2531.

34. Ramaiahgari SC, den Braver MW, Herpers B, Terpstra V, Commandeur JN, van de Water B, Price LS: **A 3D in vitro model of differentiated HepG2 cell spheroids with improved liver-like properties for repeated dose high-throughput toxicity studies**. *Arch Toxicol* 2014, **88**:1083–1095.

35. Bushel PR, Paules RS, Auerbach SS: **A comparison of the TempO-seq S1500+ platform to RNA-seq and microarray using rat liver mode of action samples**. *Front Genet* 2018, **9**:485.

36. Wehmas LC, Wood CE, Gagne R, Williams A, Yauk C, Gosink MM, Dalmas D, Hao R, O'Lone R, Hester S: **Demodifying RNA for transcriptomic analyses of archival formalin-fixed paraffin-embedded samples**. *Toxicol Sci* 2018, **162**:535–547.

37. Auerbach SS, Phadke DP, Mav D, Holmgren S, Gao Y, Xie B, Shin JH, Shah RR, Merrick BA, Tice RR: **RNA-Seq-based toxicogenomic assessment of fresh frozen and formalin-fixed tissues yields similar mechanistic insights**. *J Appl Toxicol* 2015, **35**:766–780.

38. Webster AF, Zumbo P, Fostel J, Gandara J, Hester SD, Recio L, Williams A, Wood CE, Yauk CL, Mason CE: **Mining the archives: a cross-platform analysis of gene expression profiles in archival formalin-fixed paraffin-embedded tissues**. *Toxicol Sci* 2015, **148**:460–472.

39. • Hester SD, Bhat V, Chorley BN, Carswell G, Jones W, Wehmas LC, Wood CE: **Editor's highlight: dose-response analysis of RNA-seq profiles in archival formalin-fixed paraffin-embedded samples**. *Toxicol Sci* 2016, vol. 154:202–213.
Demonstration of dose–response from RNA-seq transcript profiled after RNA extraction from paired frozen and FFPE samples from two archival studies in mice at 2 years and 20 years of storage. Extraction procedures, RNA-seq methods and data analysis are exemplary for

use of archival materials that connect histopathology and NGS-based transcriptomics.

40. Zerdoumi Y, Kasper E, Soubigou F, Adriouch S, Bougeard G, Frebourg T, Flaman JM: **A new genotoxicity assay based on p53 target gene induction**. *Mutat Res Genet Toxicol Environ Mutagen* 2015, **789–790**:28–35.

41. Israel JW, Chappell GA, Simon JM, Pott S, Safi A, Lewis L, Cotney P, Boulos HS, Bodnar W, Lieb JD, *et al.*: **Tissue- and strain-specific effects of a genotoxic carcinogen 1,3-butadiene on chromatin and transcription**. *Mamm Genome* 2018, **29**:153–167.

42. Poetsch AR, Boulton SJ, Luscombe NM: **Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis**. *Genome Biol* 2018, **19**:215.

43. Cimino MC: **Comparative overview of current international strategies and guidelines for genetic toxicology testing for regulatory purposes**. *Environ Mol Mutagen* 2006, **47**:362–390.

44. Maslov AY, Quispe-Tintaya W, Gorbacheva T, White RR, Vijg J: **High-throughput sequencing in mutation detection: a new generation of genotoxicity tests?** *Mutat Res* 2015, **776**:136–143.

45. Gomy I, Diz Mdel P: **Hereditary cancer risk assessment: insights and perspectives for the Next-Generation Sequencing era**. *Genet Mol Biol* 2016, **39**:184–188.

46. Schneeberger K: **Using next-generation sequencing to isolate mutant genes from forward genetic screens**. *Nat Rev Genet* 2014, **15**:662–676.

47. Deng JE, Sham PC, Li MX: **SNPTracker: a swift tool for comprehensive tracking and unifying dbSNP rs IDs and genomic coordinates of massive sequence variants**. *G3 (Bethesda)* 2015, **6**:205–207.

48. Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z: **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine**. *Bioinformatics* 2018, **34**:80–87.

49. Scherr CL, Aufox S, Ross AA, Ramesh S, Wicklund CA, Smith M: **What people want to know about their genes: a critical review of the literature on large-scale genome sequencing studies**. *Healthcare (Basel)* 2018, **6**.

50. Chen HZ, Bonneville R, Roychowdhury S: **Implementing precision cancer medicine in the genomic era**. *Semin Canc Biol* 2019, **55**:16–27.

51. Schwarz UI, Gulilat M, Kim RB: **The role of next-generation sequencing in pharmacogenetics and pharmacogenomics**. *Cold Spring Harb Perspect Med* 2019, **9**.

52. Woo JG, Martin LJ, Ding L, Brown WM, Howard TD, Langefeld CD, Moomaw CJ, Haverbusch M, Sun G, Indugula SR, *et al.*: **Quantitative criteria for improving performance of buccal DNA for high-throughput genetic analysis**. *BMC Genet* 2012, **13**:75.

53. Bellairs JA, Hasina R, Agrawal N: **Tumor DNA: an emerging biomarker in head and neck cancer**. *Cancer Metastasis Rev* 2017, **36**:515–523.

54. Yin Y, Lan J, Zhang Q: **Application of high-throughput next-generation sequencing for HLA typing on buccal extracted DNA**. *Methods Mol Biol* 2018, **1802**:101–113.

55. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M: **Exome sequencing: current and future perspectives**. *G3 (Bethesda)* 2015, **5**:1543–1550.

56. Normand EA, Alaimo JT, Van den Veyver IB: **Exome and genome sequencing in reproductive medicine**. *Fertil Steril* 2018, **109**:213–220.

57. Kamps R, Brandao RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok MJ, Romano A: **Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification**. *Int J Mol Sci* 2017, **18**.

58. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA: **Accuracy of next generation sequencing platforms**. *Next Gener Seq Appl* 2014, **1**.

59. • Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, Loeb LA, Swisher EM, Risques RA: **Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues**. *Proc Natl Acad Sci U S A* 2016, **113**:6005–6010.

Use of novel NGS methods for ultradeep sequencing of TP53 mutations applied to ovarian cancer. The implications are that similar queries on environmentally important genes could be useful for determining risk in environmentally contaminated regions or lifestyles. Mutational load on critical genes may be determinants of risk and environmental disease.

60. Fedeles BI, Chawanthayatham S, Croy RG, Wogan GN, Essigmann JM: **Early detection of the aflatoxin B1 mutational fingerprint: a diagnostic tool for liver cancer**. *Mol Cell Oncol* 2017, **4**:e1329693.

61. • Harrill AH, McCullough SD, Wood CE, Kahle JJ, Chorley BN: **MicroRNA biomarkers of toxicity in biological matrices**. *Toxicol Sci* 2016, **152**:264–272.

Biofluid-based miRNA studies reviewed as a “liquid biopsy” by sampling extracellular fluids like blood for chemically-induced toxicity in liver, kidney, heart and pancreas. Description of unique factors in miRNA analysis including biogenesis and baseline circulating expression levels, normalization, spike-in oligonucleotides and potential interference by erythrocyte lysis.

62. Nassirpour R, Mathur S, Gosink MM, Li Y, Shoieb AM, Wood J, O'Neil SP, Homer BL, Whiteley LO: **Identification of tubular injury microRNA biomarkers in urine: comparison of next-generation sequencing and qPCR-based profiling platforms**. *BMC Genomics* 2014, **15**:485.

63. Parsons HA, Beaver JA, Park BH: **Circulating plasma tumor DNA**. *Adv Exp Med Biol* 2016, **882**:259–276.

64. Petit J, Carroll G, Gould T, Pockney P, Dun M, Scott RJ: **Cell-free DNA as a diagnostic blood-based biomarker for colorectal cancer: a systematic review**. *J Surg Res* 2018, **236**:184–197.

65. Ghosh RK, Pandey T, Dey P: **Liquid biopsy: a new avenue in pathology**. *Cytopathology* 2019, **30**:138–143.

66. Gai W, Sun K: **Epigenetic biomarkers in cell-free DNA and applications in liquid biopsy**. *Genes (Basel)* 2019, **10**.

67. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA: **Microarrays and toxicology: the advent of toxicogenomics**. *Mol Carcinog* 1999, **24**:153–159.