COMMITTEE ON MUTAGENICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT (COM)

TOXICOGENOMICS AND RISK ASSESSMENT: APPLICATION OF TRANSCRIPTOMICS AND NEXT-GENERATION SEQUENCING TO GENOTOXICITY AND CARCINOGENICITY ASSESSMENT

Background

1. The potential application of toxicogenomics to risk assessment has been raised as an area of interest for COC, COM and COT. At a joint COC/COM meeting in November 2020 it was agreed that a document addressing this topic would be progressed, with COM best placed to lead on its development.

2. At the COM meeting in February 2021, the Committee discussed some preliminary literature on 'toxicogenomics and risk assessment' (MUT/2021/06). Members noted that this field could at present be considered to comprise two different major elements; the more highly established field of transcriptomics, and the newer area of next-generation sequencing technologies. It was felt that it would be useful for a document to be prepared providing a preliminary overview of these two areas and their potential applications to risk assessment in the fields of mutagenicity and carcinogenicity.

3. This paper (MUT/2021/04) provides an overview based on some recently published literature sources.

Literature searches

4. The PubMed database was searched on 08/03/2021 with combinations of the following search terms: transcriptomics, next-generation sequencing, genotoxicity or mutagenicity, and carcinogenicity. Searches were limited to review articles published from 2018 onwards. Full search details are provided at Annex A.

5. A total of 91 citations were identified, of which 15 were selected for secondary screening. These 15 citations are listed at Annex A and those of most relevance are discussed in detail in the sections below.

Terminology

6. The term 'toxicogenomics' is often used as a general term to describe the application of 'omics' techniques to toxicological studies. This encompasses numerous technologies; for example, genomics, transcriptomics, epigenomics, proteomics, and metabolomics. However, the development of this field, which has taken place over the last couple of decades, has mostly been driven by studies

focusing on microarray-based analysis of gene expression profiles, termed 'transcriptomics'. For this reason, rather confusingly, the term 'toxicogenomics' is often used specifically to refer to transcriptomics studies, and the two terms are often used interchangeably to describe studies and development in the area of mRNA expression profiling using microarrays (see discussion in the review articles by Schmitz-Spanke, 2019; David, 2020; Hartwig, et al., 2020).

7. Additionally, in a more recent sense, the term 'transcriptomics' is also used to describe the study of the transcriptome in total, encompassing different techniques such as microarray analysis and also high-throughput RNA sequencing to study all RNA transcripts (mRNA, non-coding RNAs, etc).

8. For the purpose of this paper, the term 'transcriptomics' is used in the narrower sense, to represent the area of study of mRNA expression profiling by microarray analysis that developed from the early 2000s onwards. As explained in paragraph 6 above, the term 'toxicogenomics' is sometimes also used in this sense, depending on the terminology adopted by the authors.

mRNA expression profiling ('transcriptomics', 'toxicogenomics')

9. Consideration of the application of toxicogenomics to genotoxicity and carcinogenicity risk assessment has been discussed in some recent review articles summarised below.

Schmitz-Spanke (2019): Toxicogenomics – What added value do these approaches provide for carcinogen risk assessment?

10. Schmitz-Spanke (2019) published a review summarising the application of toxicogenomic approaches to carcinogen risk assessment, comparing these with established methodologies. The main scope of this article covers the potential and challenges of transcriptomics (mRNA expression profiling) for carcinogen risk assessment in terms of hazard identification and dose-response assessment. An overview is given in the following paragraphs (11 - 15). Full citations for the references shown in italics can be found in the bibliography of the review article by Schmitz-Spanke (2019).

11. The first part of the review addresses the application of transcriptomics to hazard identification. Predictive transcriptomics uses gene expression signatures (transcriptional signatures) to distinguish different chemical carcinogens (genotoxic or non-genotoxic) by assuming that different gene signatures indicate different carcinogenic modes of action (MOAs). Genotoxic compounds predominantly regulate pathways related to p53 signalling, whereas the assessment of non-genotoxic carcinogens is less clear as they act via a wider range of mechanisms. Transcriptional signatures are set using a 'training set' of toxicants belonging to a certain carcinogenic class (genotoxic carcinogens, non-genotoxic carcinogens, non-carcinogens), operating through various MOAs. Subsequent statistical and bioinformatic analysis allows for the extraction of gene signatures, and

this then allows for the classification of compounds with unknown toxicity. Samples of the validation set will be classified by the extracted signature to assess the classification performance (*Romer et al., 2014*). This approach has been further developed through a pan-laboratory project to identify and prioritise potentially genotoxic agents, coordinated by the Health and Environmental Sciences Institute (HESI) Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment. A gene signature that can distinguish DNA damage-inducing (DDI) from non-DDI agents has been developed (*Li et al., 2017*), and this work is described in the following paragraphs (12 - 13).

12. For practical and ethical reasons, *in vitro* models are preferred over animal tests for screening purposes; cells should be p53 competent, capable of DNA repair, ideally of human origin, and capable of metabolic activation (*Li et al. 2015; Li 2016*). The non-tumour derived human lymphoblastoid cell line, TK6, is preferred, with the addition of rat liver S9 homogenate (*Buick and Yauk 2016*); other suitable models include human hepatic cell lines. The importance of assay optimisation with regard to exposure time and dose is emphasised. Genes involved in p53-mediated DNA damage response induce early gene expression changes and the HESI technical committee selected 4 h post treatment as the optimum exposure time, taking this into account. Dose optimisation is conducted by analysis of quantitative expression changes in well-characterised stress-response genes.

13. The HESI technical committee established a signature comprising 65 genes, known as TGx-DDI (DNA damage-inducing; previously known as TGx-28.65) (*Li et al. 2015, 2017; Corton et al. 2018*). Validation of this signature has been conducted using sets of compounds, including non-genotoxins. Ultimately, although TGx-DDI has been shown to successfully classify compounds that produced conflicting results using standard *in vitro* test batteries[1] (therefore avoiding the requirement for follow-up studies *in vivo*), the signature screening method is labour-intensive and complex and not considered to be suitable for first-line screening. However, new online tools are being developed that may help towards the development of TGx-DDI for use in a regulatory aspect, including new approaches to data interpretation (*Jackson et al. 2017; Cho et al. 2019*). The narrative of Schmitz-Spanke (2019) comments that studies have shown good transferability of *in vitro* data to *in vivo* relevance, but the approach needs quantitative development before it could be used routinely in risk assessment.

14. Following from hazard identification, the review by Schmitz-Spanke (2019) addresses the application of toxicogenomics (transcriptional analysis) to dose-response assessment. Gene-expression analysis enables qualitative and quantitative assessment of molecular events or pathways to provide information to define the shape of the dose-response relationship. Transcriptional analysis has

[1] The cited study (*Li et al. 2017)* , used a validation set of 11 compounds known to have irrelevant positive results in *in vitro* genotoxicity assays: phenobarbital, esomeprazole, donepezil, cyclohexamide, 2,4-dinitrophenol (2,4-DNP), olmesartan, exemastan, rabeprazole-NA, rotigotin, dexamethasone, staurosporine. All but one were shown to be non-DDI.

been integrated into dose-response assessment, and a benchmark dose (BMD) approach has been applied to derive transcriptional-based point-of-departure (POD) values (*reviewed by Thomas and Waters 2016*). Examples include comparison of BMDs from traditional cancer bioassays with transcriptional BMDs derived from cancer-associated pathways in tissues exposed to benzo[a]pyrene (*Moffat et al. 2015*), and multi-walled carbon nanotube (MWCNT)-induced lung fibrosis in mice (*Labib et al. 2016*). Gene fold-changes or numbers of enriched pathways have been correlated well with apical PODs. Time of exposure is important, and studies found that transcriptional BMDs after five days of exposure were closely correlated with non-cancer BMDs for organ weight and histology after 14, 28, and 90 days of exposure. For cancer-related endpoints, the transcriptional BMD values at 90 days were the most representative and were within three-fold of the cancer BMD values (*Farmahin et al. 2017*). It is not yet clear whether BMDs from *in vitro* assays correlate with those obtained *in vivo* (*Thomas et al. 2017*). This will require incorporation of knowledge of toxicokinetic parameters. The International Cooperation on Cosmetics Regulations (ICCR) has developed recommendations for risk assessment of cosmetic ingredients integrating *in silico*, *in chemico*, and *in vitro* approaches (*Dent et al. 2018; Mahony 2019*).

15. In summary, Schmitz-Spanke (2019) notes that gene expression profiling is currently the most advanced 'omics' technology. For hazard identification, the TGx-DDI is well advanced and shows promise in genotoxicity testing to gain mechanistic insights and reduce additional follow-up testing. The BMD concept is being used to analyse quantitative dose-response relationships and derive POD values, which show good agreement with traditional toxicological endpoints. Other 'omics' technologies, such as metabolomics, proteomics and epigenomics, have not been well developed with regard to risk assessment.

Hartwig et al. (2020): Mode of action-based risk assessment of genotoxic carcinogens.

16. A review article by Hartwig et al. (2020) on the topic of mode of action-based risk assessment of genotoxic carcinogens also addresses the use of toxicogenomics for hazard identification and risk assessment. The authors note that toxicogenomics is a confusing term that is on the one hand generally used to refer to 'omics' techniques applied to toxicological studies, and on the other more specifically to analysis of gene expression profiles (transcriptomics). The commentary is limited to a description of transcriptomics, given that transcriptomics is the most advanced technique in this field. An overview is given in the following paragraphs (17 – 23). Full citations for references shown below in italics can be found in the bibliography of the review by Hartwig *et al.* (2020).

17. Alterations in expression or regulation of biomolecules need to be linked to apical endpoints (phenotypic anchoring), to indicate adverse effects (*Paules 2003; Buesen et al. 2017*). The development and application of bioinformatics tools (principal component analysis, clustering, statistical comparison of classes, class prediction, mechanistic analysis) enables the extraction of patterns or signatures

from differentially expressed toxicogenomic datasets (*Afshari et al. 2011*). A framework has been developed to incorporate bioinformatics procedures into the processes of data generation and storage, processing, and interpretation (*Gant et al. 2017*). This can be performed using the Gene Ontology (GO) tool (*Gaudet and Dessimoz 2017*), and development in this area is ongoing.

18. In line with information reported by Schmitz-Spanke (2019), Hartwig *et al*. (2020) outline the history of applying toxicogenomics to carcinogenicity hazard identification and to risk assessment, with the narrative limited to the area of transcriptomics, which is noted to be the most frequently used 'omics' technique with the most advanced quality standards.

19. The commentary first addresses the topic of hazard identification. Characteristic gene signatures elicited by model compounds have enabled the classification of other compounds with unknown toxicity, and of molecular mechanisms. Toxicogenomics methods have been used to classify and predict carcinogenic potential (*Ellinger-Ziegelbauer et al. 2004*; *Suenaga et al. 2013*; *Eichner et al. 2014*; *Schaap et al. 2015*; *Williams et al. 2015*; *Yauk et al. 2016*), and signature gene sets have been described to predict carcinogenic potential, including in *in vitro* systems (*Li et al. 2017*). The MOA for genotoxic compounds leads to activation of p53 pathways in response to DNA damage and a cascade of pathways including DNA damage response, DNA repair response, apoptosis, and cell cycle arrest. Non-genotoxic carcinogens act through different pathways, including cell proliferation, decreased apoptosis, energy depletion, and production of reactive oxygen species (*Deferme et al. 2015*). They may also act as tumour promoters, for example through peroxisome proliferation, endocrine disruption, receptor mediation or immunosuppression (*Rieswijk et al. 2015*).

20. Hartwig and colleagues then address the application of toxicogenomics to risk assessment. Separation of signatures into genotoxic or non-genotoxic is a qualitative method and there is increasingly a requirement for quantitative dose-response analysis, which is essential for the application of toxicogenomic data in risk assessment. There is a particular requirement to determine the low dose-response relationship and POD (*Johnson et al. 2015; Li et al. 2017*). The utilisation of toxicogenomic data in risk assessment is an approach that has been developed by a small number of groups (*McMullen et al. 2016; Farmahin et al. 2017; NTP 2018*).

21. Thomas and colleagues conducted studies on mice exposed to five different chemical carcinogens for 13 weeks (*Thomas et al. 2011, 2012*). BMDs for expression of individual genes were calculated to determine the point in the dose-response curve where most pathways became transcriptionally active. Gene enrichment analysis was performed to determine which GO categories were activated, and BMD and BMDL values were calculated for each GO category. The pathway with the lowest median transcriptional BMD/BMDL (disregarding the biological function) was selected for comparison with the corresponding value for apical endpoints (e.g. liver weight, histological changes in target tissues, tumour

incidence); transcriptional values correlated well with traditional endpoints, allowing the conclusion that the lowest transcriptional BMD/BMDL could be used as a suitable POD. Case studies were conducted with benzo[a]pyrene (*Moffat et al. 2015*), and subsequently with other substances including furan, MWCNT, naphthalene, nickel subsulfide, and cholestatic drugs (*Clewell et al. 2014; Efremenko et al. 2014; Dong et al. 2015; Labib et al 2016; Kawamoto et al. 2017*). Farmahin and colleagues evaluated eleven approaches comparing transcriptional BMD values with BMDs derived from apical endpoint changes: eight approaches for selecting genes for POD derivation, and three previously proposed approaches (the lowest pathway BMD, the mean and the median BMD of all genes). In most cases, the transcriptomic POD was within ten-fold of the apical endpoint-derived value, and four approaches produced BMDs showing a good concordance with apical BMD values across multiple time points (*Farmahin et al 2017*).

22. Some groups have focussed on the 'no observed transcriptional effect level (NOTEL) threshold, where no effect on the transcriptome is observed, with NOTEL values being substantially lower than corresponding NOAELS (*Zarbl et al. 2010; Pisani et al. 2015; Quercioli et al. 2018*). Other approaches include determination of the 'no observed genotoxic effect level' (NOGEL); in studies of rats exposed to methyl methanesulfonate (MMS) and methylnitrosourea (MNU), at exposure levels below the NOGEL (derived from analysis of blood reticulocyte micronucleus induction) no statistically significant changes in gene alterations on whole-genome transcript analysis of liver were seen (*Ji et al 2016*).

23. The review by Hartwig et al. (2020) concludes that toxicogenomic approaches can be useful as a complement to traditional hazard identification and risk assessment; in particular these techniques can provide information to improve mechanistic understanding of dose-response relationships and of biological thresholds, and a large amount of data can be generated in a timely and cost-effective manner, minimising the requirement for *in vivo* studies. However, some points are also noted by the authors that need to be addressed. Genes of unknown function are not included in pathway and enrichment analyses, which may introduce a bias against recognising potentially important pathways. The choice of metabolic activation is extremely important for pre-mutagens. Also, use of the lowest transcriptional BMDL is a very conservative approach that may not actually represent the apical adverse effect or dose-response relation. This may be addressed by distinguishing adaptive from adverse effects and defining pathways, doses and exposures times indicating transitions between different states. A final issue is that microarray analyses produce semi-quantitative outputs; RT-qPCR is an alternative, quantitative high-throughput approach that shows promise for development and application in risk assessment (*Fischer et al. 2016*).

COC (2019): Guidance statement G07c 'Omics, high-throughput screening technologies, and bioinformatics'

24. The COC guidance statement, G07 'Alternatives to the 2-year bioassay' (part c) addresses the topic of the application of 'omics' technologies to

carcinogenicity evaluation, noting that most work to date has focussed on the field of transcriptomics (COC 2019). The document summarises areas of study that have used omics technologies to predict outcomes of 2-year rodent bioassays by applying such methods to short-term studies *in vivo*, which have mostly focussed on mRNA profiling in rat liver, and the identification of gene signatures to discriminate between direct- and indirect-acting genotoxic carcinogens, non-genotoxic carcinogens, and non-carcinogens. Other concepts that are introduced in the document include: 'shared cancer biology'; 'profiling to the phenotype'; the application of 'omics' technologies *in vitro* and current obstacles to this approach (noting that such methods are useful in characterising toxicity pathways to elucidate modes of action); the 'comparison approach'; the availability of a large catalogue of (*in vivo* and *in vitro*) datasets, based on a large set of compounds, consistent study designs and standardised experimental protocols; the 'parallelogram approach' and 'concordance model'; progress in integrating 'omics' data into quantitative cancer risk assessments (derivation of PoDs, usually BMDs, which can be compared with PoDs from conventional/apical endpoints); and a framework for applying transcriptomic data to (non-cancer and cancer) risk assessment. It also mentions concepts of managing and evaluating large datasets (artificial intelligence, deep learning, data mining). Detailed review articles addressing many of the concepts relating to 'omics' that are introduced in G07 part c can be found in the book 'Toxicogenomics in Predictive Carcinogenicity' (Waters 2016).

Next-generation sequencing

25. The European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) online training site[2] notes the following with regard to next-generation sequencing:

- "In contrast to microarray methods, sequence-based approaches directly determine the nucleic acid sequence of a given DNA or cDNA molecule"
- "Compared to conventional Sanger sequencing using capillary electrophoresis, the short read, massively parallel sequencing technique is a fundamentally different approach that revolutionised sequencing capabilities and launched the second-generation sequencing methods – or next-generation sequencing (NGS) – that provide orders of magnitude more data at much lower recurring cost."
- "Next generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies. These technologies allow for sequencing of DNA and RNA much more quickly and cheaply than the

[2] https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/, accessed 15/04/2021.

previously used Sanger sequencing, and as such revolutionised the study of genomics and molecular biology."

26. A review article by David (2020) summarises the application of toxicogenomics to the field of genetic toxicology. The article first reviews the history of development of transcriptomics techniques, noting that transcriptomic profiles have been identified that can discriminate between genotoxic and non-genotoxic carcinogens (as discussed in the above section on 'mRNA expression profiling ('transcriptomics', 'toxicogenomics')', paragraphs 9-24), but concludes that this approach is limited in application and has not transformed the field of genetic toxicology in the way that had been predicted.

27. With regard to the development and potential application of next generation sequencing technologies, the author discusses several different next-generation sequencing platforms that have been developed (reviewed by *Van Dijk et al. 2014*) which allow the rapid and cheap sequencing of hundreds of billions of base pairs. The different next generation sequencing technologies and their applications to genetic toxicology discussed by David (2020) are summarised in the following paragraphs, 25-35. Full citations for the publications shown below in italics can be found in the bibliography of the article by David (2020).

28. *RNA sequencing (RNAseq).* This is a method for high-throughput deep sequencing of cDNA. Advantages include the capability to identify and quantify transcripts without prior knowledge of gene sequence, which is applicable when the MOA is not known. The method can identify sequence variations in transcribed regions (e.g. single nucleotide polymorphisms; SNPs). RNAseq can detect genes with low or high expression levels, due to the very low background signal and large dynamic range compared with DNA microarrays and can provide information about sequence variation and on alternative splicing in different biological conditions (*Wang et al. 2009*). In application to genetic toxicology, RNAseq has been used to detect transcriptional changes in pathways following treatment with DNA damaging agents. Treatment with etoposide indicated alteration in other pathways in addition to DNA damage response (*Wei et al. 2018*). Treatment of inducible pluripotent stem cell (iPSC)-derived cardiomyocytes with doxorubicin altered DNA damage and cell cycle genes (*Reyes et al. 2018*). This approach has been used to investigate DNA repair gene expression in prostate cancer (*Jividen et al. 2018*) and processes that occur at double-strand breaks during repair (*Bonath et al. 2018*). Modified RNAseq protocols have been developed to investigate DNA damage relating to DNA adducts; these include eXcision Repair (XR)-seq (single-nucleotide resolution of nucleotide excision repair of bulky DNA adducts (*Hu et al. 2019*) and Damage-seq (blocking of DNA polymerase by bulky DNA adducts and immunoprecipitation with antibodies (*Hu et al. 2016*). These two protocols in combination have been used to study genome-wide profiles of cisplatin damage (*Yimit et al. 2019*). A modified protocol called click-code-seq allows detection of oxidative damage on a genome-wide scale (*Wu et al. 2018*).

29. *Mutation signatures*. These are specific patterns of mutations associated with particular chemical classes or MOAs. Next generation sequencing has allowed the determination of genome-wide mutational signatures from controlled studies *in vitro* or *in vivo*. A recent study reported mutational signatures identified after exposure of human iPSCs to 79 known or suspected environmental carcinogens, in some cases allowing correlation of signatures between carcinogens and cancer types (e.g. benzo[a]pyrene, benzo[a]pyrene-7, and 8-dihydrodiol-9,10-oxide showed similar signatures to those from lung cancers in smokers) and providing information of potential mechanistic relevance (*Kucab et al. 2019)*. Thousands of cancer genomes have been sequenced and mutational signatures are stored in the COSMIC (Catalogue of Somatic Mutations in Cancer) database (*Bamford et al. 2004*). Signatures have in some cases been attributed to specific compounds or exposures, such as cigarette smoke including tobacco carcinogens (*Alexandrov et al. 2020*), UV radiation, aflatoxin B1, alkylating chemotherapy drugs, and aristolochic acid (*Phillips 2018*).

30. *Sub-clonal mutation detection.* This is the detection of sub-clonal mutations in mixed cell populations without selection or expansion of mutant clones, which could in theory allow for direct evaluation of mutagenicity without requirement for reporter or surrogate systems. However, the high error rate of next generation sequencing methods currently limits the detection of sub-clonal mutations (*Kennedy et al. 2014*).

31. *Genome-wide sequencing.* Several methods have been developed, mostly for detecting off-target mutations induced by nucleases such as CRISPR-Cas9 (*Lazzarotto et al. 2018*). Such methods could be applied to the detection of mutations following treatment with chemical carcinogens. However, applications are to some extent currently limited by lack of sensitivity of the techniques. One technique, CIRCLE-seq, is highly sensitive but the propagation of misincorporation events in the first DNA polymerase round may lead to incorrect identification of mutations.

32. *Single-cell DNA sequencing (scDNA-seq).* This method has been used to detect mutations in breast cancer (*Wang et al. 2014*) and renal carcinoma (*Xu et al. 2012*), while single-cell exome sequencing has been used to detect mutations in circulating lung tumour cells (*Ni et al. 2013*). Technical problems arise due to the low copy number of input DNA, although methods are being developed to overcome these issues.

33. *Single-cell RNA sequencing (scRNA-seq).* This method has been used to investigate single nucleotide polymorphisms (SNPs) in colon cancer (*Chen et al. 2016*). The technique currently has some limitations, including the requirement for mutations to be in exomes, good alignment and annotation, and for the mutation to be in a highly expressed gene and to be statistically detectable (*Vu et al. 2019*).

34. *Amplicon sequencing.* This method detects low-frequency genetic variants by deep sequencing[3] of the amplicon (PCR product). Mutations at a target locus can be identified by comparison with a reference sequence. This may have potential for the identification of treatment-induced mutations, but detection of true mutations is currently challenging due to the low frequency with which they occur.

35. *Duplex sequencing.* Fragmented DNA molecules are tagged with unique sequences, allowing the sequence of both original DNA strands to be determined. The method has high accuracy (detection of 1 mutation in 10^7 bases) compared with conventional next generation sequencing. The method has been applied to M13mp2 DNA, which has been used extensively in sensitive genetic mutation assays, and to mitochondrial DNA. It has also been used to detect TP53 mutations in ovarian cancer (*Salk et al. 2019*). Duplex sequencing could be used to validate assays such as the *Pig-a* assay, as well as providing information on the mutation context of the response in the *Pig-a* gene relative to the whole genome.

36. David (2020) also discusses the role of artificial intelligence and machine learning. The large datasets generated by next generation sequencing are amenable to machine learning and could be used to classify compounds and predict toxicity (*Wu and Wang 2018*). Several publications are noted that describe many examples with multiple chemicals (*Ding et al. 2012; Schwartz et al. 2015; Fange et al. 2015; Spinella et al. 2016; Yamane et al. 2016; Wood et al. 2018*). As algorithms improve and more useful data sets are generated, these could represent powerful techniques for mutation detection and allow the classification of different chemicals.

37. The future potential of the new sequencing technologies and techniques for the field of genetic toxicology are recognised. However, David (2020) highlights a number of important issues that must be taken into consideration, including:

- The cell type(s) used for mutational analysis. Genetic toxicity testing generally uses cancer cells, which have a different mutation profile to healthy cells, with DNA damage response (DDR) pathways that may be perturbed, leading to altered response to chemical carcinogens. It might be beneficial to use multiple cell types to establish if responses are cell-type specific.

- The requirement to establish a minimum number of cells for analysis, given that mutations are induced at low rates.

- The importance of sampling time-point. Traditional mutation assays (TK, HPRT) and the *Pig-a* assay require an expression period to fix the mutation. Although this aligns well with sampling for mutation detection

[3] Deep sequencing refers to sequencing a genomic region multiple times, sometimes hundreds or even thousands of times. This approach allows for the detection of rare clonal types, cells, or microbes comprising as little as 1% of the original sample (https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/deep-sequencing.html, accessed 16/03/2021).

by DNA sequencing, it is potentially then too late to detect changes by RNAseq. Thus multiple sampling points are probably required, with RNA collected at early time points to detect gene expression changes, and DNA collected at later time points for mutation analysis. This is practically not possible at present due to the large amount of sequencing that would be required.

- Automation and machine learning may facilitate the analysis of multiple doses and time points in the future. An example is DRUG-seq (Digital RNA with perturbation of Genes); automation applied to RNA-seq to allow high-throughput and cost-effectiveness, an approach which is being used to profile compounds and cluster them based on transcriptomic profile (*Ye et al. 2018*).

- In order to replace other genotoxicity tests, sequencing-based tests need to be able to detect aneuploidy. This may be achieved by analysis of copy number variations (CNVs); structural variations with abnormal copy number changes in fragments of DNA typically longer than 1 kb that lead to losses, gains or complex genome rearrangements. Bioinformatic approaches exist to detect CNVs from whole genome sequencing, but no single tool is yet capable of detecting the full range of DNA modifications (*Pirooznia et al. 2015*).

38. David (2020) concludes that there is the potential for sequencing to be able to provide all the information required to determine whether a compound is genotoxic, and by which mechanism. Data on mutation frequency and patterns, CNV, gene expression profiles and 'missed' data from machine learning could be used to build fingerprints of genotoxicity that could be used to interpret, and possible predict, responses to new compounds of unknown mechanism.

Summary

39. An introductory overview of the development and potential future applications of transcriptomics and next generation sequencing methodologies in the context of genetic toxicity and carcinogenicity is presented, based on recent review articles on this topic. The purpose of this paper is to initiate discussion and develop a workplan to address the potential application of these methodologies in genotoxicity assessment and risk assessment.

Questions for the Committee

Members are invited to comment on the information included in the discussion document, and address the following questions:

i. How should the COM progress the review of the use of transcriptomics and next-generation sequencing in the assessment of genotoxicity and carcinogenicity?
ii. Are members aware of any additional areas that should be included in the discussion?

IEH Consulting under contract supporting the PHE COM Secretariat.
Date: June 2021

Abbreviations

BMD	Benchmark dose
BMDL	Benchmark dose (lower confidence limit)
CNV	Copy number variant
COSMIC	Catalogue Of Somatic Mutations In Cancer
DDI	DNA damage inducing
DDR	DNA damage response
DRUG-seq	Digital RNA with perturbation of genes
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
GO	Gene ontology
HESI	Health and Environmental Sciences Institute
ICCR	International Cooperation on Cosmetics Regulations
MMS	Methyl methanesulfonate
MNU	Methylnitrosourea
MOA	Mode of action
NGS	Next generation sequencing
NOAEL	No observed adverse effect level
NOGEL	No observed genotoxic effect level
NOTEL	No observed transcriptional effect level
POD	Point of departure
RNAseq	RNA sequencing
scDNA-seq	Single cell DNA sequencing
SNP	Single nucleotide polymorphism
XR	eXcision repair

References

David, R. (2020) The promise of toxicogenomics for genetic toxicology: past, present and future. *Mutagenesis,* 35**,** 153-159.

Hartwig, A., M. Arand, B. Epe, S. Guth, G. Jahnke, A. Lampen, H. J. Martus, B. Monien, I. Rietjens, S. Schmitz-Spanke, G. Schriever-Schwemmer, P. Steinberg & G. Eisenbrand (2020) Mode of action-based risk assessment of genotoxic carcinogens. *Arch Toxicol,* 94**,** 1787-1877.

Schmitz-Spanke, S. (2019) Toxicogenomics - What added Value Do These Approaches Provide for Carcinogen Risk Assessment? *Environ Res,* 173**,** 157-164.

Waters, M. D., Thomas, R.S. 2016. Toxicogenomics in Predictive Carcinogenicity. In *Issues in Toxicology,* ed. W. M. Andersen D, Marrs TC. Cambridge UK: The Royal Society of Chemistry.

COMMITTEE ON MUTAGENICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT (COM)

Details of literature searches carried out by IEH Consulting Ltd.

Searches of the PubMed database were carried out on 08/03/2021 with combinations of the following search terms: transcriptomics, next generation sequencing, genotoxicity or mutagenicity, and carcinogenicity. Searches were limited to review articles published from 2018 onwards. Search strings are indicated below.

("Transcriptomics"[Title/Abstract] AND ("genotox*"[Title/Abstract] OR "mutagen*"[Title/Abstract]) AND (2018/01/01:2021/01/31[Date - Publication] AND "english"[Language]) AND (("review"[Publication Type] OR "systematic review"[Publication Type]) AND (2018/01/01:2021/01/31[Date - Publication] AND "english"[Language]))) AND ((2018/1/1:2021/1/31[pdat]) AND (english[Filter]))

("Transcriptomics"[Title/Abstract] AND "carcinogen*"[Title/Abstract] AND (2018/01/01:2021/01/31[Date - Publication] AND "english"[Language]) AND (("review"[Publication Type] OR "systematic review"[Publication Type]) AND (2018/01/01:2021/01/31[Date - Publication] AND "english"[Language]))) AND ((2018/1/1:2021/1/31[pdat]) AND (english[Filter]))

(("Next generation sequencing"[Title/Abstract]) AND (genotox*[Title/Abstract] OR mutagen*[Title/Abstract]) AND ((2018/1/1:2021/1/31[pdat]) AND (english[Filter]))) AND (review[Publication Type] OR systematic review[Publication Type] AND ((2018/1/1:2021/1/31[pdat]) AND (english[Filter])))

(("next generation sequencing"[Title/Abstract]) AND (carcinogen*[Title/Abstract]) AND ((2018/1/1:2021/1/31[pdat]) AND (english[Filter]))) AND (review[Publication Type] OR systematic review[Publication Type] AND ((2018/1/1:2021/1/31[pdat]) AND (english[Filter])))

A total of 91 citations were identified. Of these, 15 publications were selected as being of particular relevance to the topic of interest and these are listed below:

Beal, M. A., M. J. Meier, D. P. LeBlanc, C. Maurice, J. M. O'Brien, C. L. Yauk & F. Marchetti (2020) Chemically induced mutations in a MutaMouse reporter gene inform mechanisms underlying human cancer mutational signatures. Commun Biol, 3, 438. https://pubmed.ncbi.nlm.nih.gov/32796912/

Chakraborty, S., M. I. Hosen, M. Ahmed & H. U. Shekhar (2018) Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. Biomed Res Int, 2018, 9836256. https://pubmed.ncbi.nlm.nih.gov/30402498/

David, R. (2020) The promise of toxicogenomics for genetic toxicology: past, present and future. Mutagenesis, 35, 153-159. https://pubmed.ncbi.nlm.nih.gov/32087008/

Dertinger, S. D., Y. Totsuka, J. H. Bielas, A. T. Doherty, J. Kleinjans, M. Honma, F. Marchetti, M. J. Schuler, V. Thybaud, P. White & C. L. Yauk (2019) High information content assays for genetic toxicology testing: A report of the International Workshops on Genotoxicity Testing (IWGT). Mutat Res, 847, 403022. https://pubmed.ncbi.nlm.nih.gov/31699343/

Evans, S. J., B. Gollapudi, M. M. Moore & S. H. Doak (2019) Horizon scanning for novel and emerging in vitro mammalian cell mutagenicity test systems. Mutat Res, 847, 403024. https://pubmed.ncbi.nlm.nih.gov/31699342/

Harris, K. L., M. B. Myers, K. L. McKim, R. K. Elespuru & B. L. Parsons (2020) Rationale and Roadmap for Developing Panels of Hotspot Cancer Driver Gene Mutations as Biomarkers of Cancer Risk. Environ Mol Mutagen, 61, 152-175. https://pubmed.ncbi.nlm.nih.gov/31469467/

Hartwig, A., M. Arand, B. Epe, S. Guth, G. Jahnke, A. Lampen, H. J. Martus, B. Monien, I. Rietjens, S. Schmitz-Spanke, G. Schriever-Schwemmer, P. Steinberg & G. Eisenbrand (2020) Mode of action-based risk assessment of genotoxic carcinogens. Arch Toxicol, 94, 1787-1877. https://pubmed.ncbi.nlm.nih.gov/32542409/

Kennedy, S. R., Y. Zhang & R. A. Risques (2019) Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. Trends Cancer, 5, 531-540. https://pubmed.ncbi.nlm.nih.gov/31474358/

Li, H. H., C. L. Yauk, R. Chen, D. R. Hyduke, A. Williams, R. Frötschl, H. Ellinger-Ziegelbauer, S. Pettit, J. Aubrecht & A. J. Fornace, Jr. (2019) TGx-DDI, a Transcriptomic Biomarker for Genotoxicity Hazard Assessment of Pharmaceuticals and Environmental Chemicals. Front Big Data, 2, 36. https://pubmed.ncbi.nlm.nih.gov/33693359/

Li, W. & A. Sancar (2020) Methodologies for detecting environmentally induced DNA damage and repair. Environ Mol Mutagen, 61, 664-679. https://pubmed.ncbi.nlm.nih.gov/32083352/

Liu, Z., Z. Wang, E. Jia, T. Ouyang, M. Pan, J. Lu, Q. Ge & Y. Bai (2019) Analysis of genome-wide in cell free DNA methylation: progress and prospect. Analyst, 144, 5912-5922. https://pubmed.ncbi.nlm.nih.gov/31436778/

Phillips, D. H. (2018) Mutational spectra and mutational signatures: Insights into cancer aetiology and mechanisms of DNA damage and repair. DNA Repair (Amst), 71, 6-11. https://pubmed.ncbi.nlm.nih.gov/30236628/

Saini, N. & D. A. Gordenin (2018) Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. Environ Mol Mutagen, 59, 672-686. https://pubmed.ncbi.nlm.nih.gov/30152078/

Salk, J. J. & S. R. Kennedy (2020) Next-Generation Genotoxicology: Using Modern Sequencing Technologies to Assess Somatic Mutagenesis and Cancer Risk. Environ Mol Mutagen, 61, 135-151. https://pubmed.ncbi.nlm.nih.gov/31595553/

Schmitz-Spanke, S. (2019) Toxicogenomics - What added Value Do These Approaches Provide for Carcinogen Risk Assessment? Environ Res, 173, 157-164. https://pubmed.ncbi.nlm.nih.gov/30909101/