Title

Detection of Transgene Location in the CYP2A13/2B6/2F1-transgenic Mouse Model using Optical Genome Mapping Technology

Permalink

https://escholarship.org/uc/item/31x5j1rr

Journal

Drug Metabolism and Disposition, 51(1)

ISSN

0090-9556

Authors

Ding, Xinxin
Han, John
Van Winkle, Laura S
et al.

Publication Date

2023

DOI

10.1124/dmd.122.001090

Peer reviewed

1521-009X/51/1/46–53$35.00
DRUG METABOLISM AND DISPOSITION

dx.doi.org/10.1124/dmd.122.001090
Drug Metab Dispos 51:46–53, January 2023

Detection of Transgene Location in the CYP2A13/2B6/2F1-transgenic Mouse Model using Optical Genome Mapping Technology

Xinxin Ding, John Han,[1] Laura S. Van Winkle, and Qing-Yu Zhang

Department of Pharmacology and Toxicology, College of Pharmacy, University of Arizona, Tucson, Arizona (X.D., J.H., Q.-Y.Z.) and Center for Health and the Environment and Department of Anatomy Physiology and Cell Biology, School of Veterinary Medicine, UC Davis, Davis, California (L.S.V.W.)

Received August 18, 2022; accepted October 4, 2022

ABSTRACT

Most transgenic mouse models are generated through random integration of the transgene. The location of the transgene provides valuable information for assessing potential effects of the transgenesis on the host and for designing genotyping protocols that can amplify across the integration site, but it is challenging to identify. Here, we report the successful utility of optical genome mapping technology to identify the transgene insertion site in a CYP2A13/2B6/2F1-transgenic mouse model, which produces three human cytochrome P450 (P450) enzymes (CYP2A13, CYP2B6, and CYP2F1) that are encoded by neighboring genes on human chromosome 19. These enzymes metabolize many drugs, respiratory toxicants, and chemical carcinogens. Initial efforts to identify candidate insertion sites by whole genome sequencing was unsuccessful, apparently because the transgene is located in a region of the mouse genome that contains highly repetitive sequences. Subsequent utility of the optical genome mapping approach, which compares genome-wide marker distribution between the transgenic mouse genome and a reference mouse (GRCm38) or human (GRCh38) genome, localized the insertion site to mouse chromosome 14, between two marker positions at 4451324 base pair and 4485032 base pair. A transgene-mouse genome junction sequence was further identified through long-polymerase chain reaction amplification and DNA sequencing at GRCm38 Chr.14:4484726. The transgene insertion (~2.4 megabase pair) contained 5–7 copies of the human transgenes, which replaced a 26.9–33.4 kilobase pair mouse genomic region, including exons 1–4 of Gm3182, a predicted and highly redundant gene. Finally, the sequencing results enabled the design of a new genotyping protocol that can distinguish between hemizygous and homozygous CYP2A13/2B6/2F1-transgenic mice.

SIGNIFICANCE STATEMENT

This study characterizes the genomic structure of, and provides a new genotyping method for, a transgenic mouse model that expresses three human P450 enzymes, CYP2A13, CYP2B6, and CYP2F1, that are important in xenobiotic metabolism and toxicity. The demonstrated success in applying the optical genome mapping technology for identification of transgene insertion sites should encourage others to do the same for other transgenic models generated through random integration, including most of the currently available human P450 transgenic mouse models.

Introduction

The human cytochrome P450 (P450) *CYP2A13*, *CYP2B6*, and *CYP2F1* genes are located on chromosome (Chr.) 19, within a cluster of *CYP2* genes that also include *CYP2A6* and *CYP2S1* (Wang et al., 2003). All three genes are expressed in the respiratory tract, with *CYP2A13* and *CYP2F1* being selectively expressed in the lung and nasal mucosa, and *CYP2B6* being expressed more dominantly in the liver (Ding et al., 2018). The CYP2A13 enzyme metabolizes many respiratory toxicants and carcinogens, including various nitrosamines, such as the tobacco-specific lung carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, aromatic hydrocarbons, such as toluene, styrene, and naphthalene, indoles, aromatic and heterocyclic amines, and drug substrates, such as nicotine, phenacetin, and theophylline (Ding et al., 2018). CYP2F1 also metabolizes many respiratory toxicants, particularly naphthalene, styrene, trichloroethylene, and 3-methylindole (Ding et al., 2018). CYP2B6, although also important for the metabolism of environmental toxicants, such as polychlorinated biphenyls (Uwimana et al., 2019), is better known for its ability to metabolize clinical drugs, such as propofol, bupropion, methadone, cyclophosphamide, ifosfamide, nevirapine, and efavirenz (Hedrich et al., 2016; Li et al., 2018).

A CYP2A13/2B6/2F1-transgenic mouse model was previously generated to study the function and regulation of the three human *CYP* genes in vivo (Wei et al., 2012). The mouse was generated through random integration of a human genomic DNA clone containing the three *CYP* genes. The transgenic mouse colony, which has been continuously maintained for 10 years, does not exhibit any notable biologic phenotype, in terms of gross morphologic features, development, and fertility. The transgenic mouse shows human-like expression of the three CYPs, with

This work was supported in part by the National Institutes of Health [National Cancer Institute Grants CA023074 and CA092596, and National Institute of Environmental Health Sciences Grants ES006694, ES004940, ES014901, and ES020867]. The CYP2A13/2B6/2F1-transgenic and *Cyp2abfgs*-null strains are available from breeding colonies maintained at the University of Arizona.

The authors have no conflict of interest to declare.

[1] High school student intern from Basis Tucson North Charter School, Tucson, Arizona.

dx.doi.org/10.1124/dmd.122.001090.

S This article has supplemental material available at dmd.aspetjournals.org.

ABBREVIATIONS: bp, base pair; Chr., chromosome; P450, cytochrome P450; kbp, kilobase pair; Mbp, megabase pair; PCR, polymerase chain reaction; SV, structural variant; WT, wild-type.

CYP2A13 and CYP2F1 produced in the lung and nasal mucosa, and CYP2B6 produced in the liver (Wei et al., 2012). The CYP2A13/2B6/2F1-transgenic mouse model has been used in several published studies, e.g., to demonstrate the role of CYP2A13 in 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone-induced lung tumorigenesis (Megaraj et al., 2014), the role of CYP2A13 and CYP2F1 in naphthalene-induced lung toxicity (Li et al., 2017; Kovalchuk et al., 2019), the regulation of CYP2A13 by inflammation and lung tumorigenesis (Wu et al., 2013; Liu et al., 2015b), and the role of hepatic CYP2B6 in nicotine metabolism in vivo (Liu et al., 2015a). The mouse model is also used in a number of ongoing studies by several laboratories.

A limitation of the CYP2A13/2B6/2F1-transgenic mouse model, which is also a potential limitation for essentially all transgenic mouse models generated through random integration of the transgene into the mouse genome, is the uncertainty with whether the insertion of the transgene disrupted any important genes in the mouse genome. While general characterization of the mouse model can reveal gross abnormality, subtle changes will be difficult to detect but may confound study outcomes, particularly when complex biologic parameters are used as experimental end points. In addition, the lack of knowledge about where the transgene is inserted also makes it difficult to distinguish between hemizygotes and homozygotes, which would appear similar during genotypic analysis using primers internal to the transgene sequence. This practical limitation makes it time-consuming to identify homozygotes for establishing breeding pairs, and nearly impossible to produce hemizygous and homozygous littermates to study gene copy number effects on a complex pharmacological or toxicological outcome.

The aim of this study was to identify the region of the mouse genome where the human CYP2A13/2B6/2F1 transgenes are located. Initial attempts using whole genome DNA sequencing for the transgenic mouse was unsuccessful as the data represented only false positive results. Subsequently, an optical genome mapping technology was used, which revealed a single insertion site on mouse Chr. 14. Subsequent analysis of the gene mapping data and further experimental validation using long-polymerase chain reaction (PCR) amplification and DNA sequencing of putative mouse-human DNA sequence junctions yielded the precise location of the insertion site at one end and the proximate location of the insertion site at the other end of the transgene insert. Comparisons of the optical genome mapping data for the transgenic mouse and a reference human genome confirmed the presence of multiple copies of the human transgenes at this single location, provided a detailed structural view of each copy of the transgene insert (many were partial copies), and revealed the exact copy number for the three functional human *CYP* transgenes. A further analysis of the mouse genomic region that is replaced by the transgene insertion revealed the disruption of only a predicted mouse gene, which appears to be redundant and has no known function. Finally, utilizing the new knowledge of the transgene insertion site, a new genotyping method was devised, which can distinguish between hemizygous and homozygous CYP2A13/2B6/2F1-transgeic mice.

Materials and Methods

CYP2A13/2B6/2F1-Transgenic Mouse. The CYP2A13/2B6/2F1-transgenic mouse, generated by the random insertion of a human bacterial artificial chromosome clone (CTD-2535H15) containing the human *CYP2A13*, *CYP2B6*, and *CYP2F1* genes, into the mouse genome, has been described previously (Wei et al., 2012). Briefly, the ~210-kilobase pair (kbp) bacterial artificial chromosome DNA construct, without the vector, was microinjected into the pronuclei of fertilized eggs from the C57BL/6J strain, and the resulting transgenic mice were maintained on either the wild-type (WT) C57BL/6J background or crossbred with various knockout mouse models, including *Cyp2abfgs*-null (Li et al., 2014). The experiments in this study were conducted using CYP2A13/2B6/2F1$^{(tg/tg)}$/*Cyp2abfgs*$^{(-/-)}$ mice, on the C57BL/6 genetic background. All animal use protocols were approved by the University of Arizona Institutional Animal Care and Use Committee.

Whole Genome Sequencing. Genomic DNA was prepared from mouse liver tissue. DNA library was prepared for sequencing using TruSeq DNA PCR-Free library preparation kit (Illumina Inc., San Diego, CA). Sequencing of the CYP2A13/2B6/2F1-transgenic/*Cyp2abfgs*-null mouse genome was performed by Novogene Corporation Inc. (Sacramento, CA, USA) on an Illumina model 2000 sequencer using 2 × 150 nucleotide paired-end reads. The genome was sequenced to 40X coverage with 120 G raw data. Candidate insertion sites were identified with dual analyses. In one analysis, reads were aligned to the reference mouse genome (GRCm38). In the other analysis, reads were aligned to the reference human genome sequence (GRCh38, Chr.19: 40940526-41139581, a region containing the transgene sequence).

Optical Genome Mapping. Optical mapping of the CYP2A13/2B6/2F1-transgenic/*Cyp2abfgs*-null mouse genome was performed by Bionano Genomics (San Diego, CA). Ultra-high molecular weight DNA was extracted using the Animal Tissue DNA Isolation Kit (Bionano Genomics) from 30 mg of mouse liver, according to the manufacturer's protocol. Genomic DNA (750 ng) was fluorescently labeled at the CTTAAG motif using the DLE-1 direct labeling enzyme and DLS DNA Labeling Kit (Bionano Genomics). The labeled DNA was linearized in a SaphyrChip using NanoChannel arrays; individual molecules were imaged and the images were digitized. As molecules are uniquely identifiable by distinct distribution of sequence motif labels, they were then assembled by pairwise alignment into de novo genome maps using Bionano Solve version 3.6 (Bionano Genomics). Structural variants (SVs) (based on assembled maps) were called against the in-silico DLE-1-digested mouse (GRCm38) and human reference genome (GRCh38). Data were analyzed with Bionano Access and Bionano Tools on Saphyr Computer Servers (Bionano Genomics). Marker positions for the reference mouse and human genomes were assigned based on fully assembled chromosome maps, whereas marker positions for the transgenic mouse genome were assigned for each de novo assembled map before alignments were made with the reference genome maps.

Other Methods. To validate the insertion site obtained from optical genome mapping, breakpoint region was amplified with PCR. PCR was carried out using a GoTag Long PCR kit (Promega, Madison, WI). Genomic DNA (200 ng) was used as a PCR template. PCR conditions were: 95°C for 2 minutes followed by 30 cycles of 92°C for 30 seconds, 65°C for 15 minutes, with a final extension at 72°C for 10 minutes. PCR products were analyzed by agarose gel electrophoresis or were purified and then subjected to Sanger DNA sequencing. Primer sequences and coordinates are listed in Table 1. Primers were designed with Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). Sanger DNA sequencing was performed at the University of Arizona Genetics Core facility.

For PCR-based genotype analysis, PCR was carried out using a GoTag PCR kit (Promega) and ~200 ng genomic DNA as template. PCR conditions for the WT allele were: 94°C for 2 minutes followed by 35 cycles of 94°C for 30 seconds, 66°C for 30 seconds and 72°C for 1 minute, with a final extension at 72°C for 5 minutes; PCR conditions for the transgenic allele were the same, except that the annealing temperature was at 62°C. For agarose gel analysis of DNA, a 1-kb Plus DNA ladder (ThermoFisher Scientific, Waltham, MA) was used for size determination.

Results

Attempts to Identify Candidate Transgene Insertion Sites by Whole Genome Sequencing. Sequencing reads were aligned to the reference mouse genome (GRCm38), as well as to the reference human genome (GRCh38, Chr.19: 40991282-41128381). Those reads aligned to both mouse and human reference genome were identified as containing candidate insertion sites. Seventeen candidate insertion sites were identified, which are located on 12 different mouse chromosomes. Most of the candidate insertion sites had very low numbers of supporting reads (2–6) (Supplemental Table 1), which meant low confidence and high probability of false discovery. Attempts to validate these candidate sites using PCR, with primers flanking the putative mouse–human sequence junctions, were unsuccessful (data not shown).

TABLE 1

PCR primers used for the amplification of the human–mouse sequence junction or for genotyping

Primer name	Primer sequence	Position in mouse or human genome	PCR product size
TG-F1	5′-ggtcaggagatcgagaccatc-3′	GRCh38, Chr.19:41010646-41010626	>15 kbp
WT-R1	5′-aacctgagcctgtgagaagc-3′	GRCm38, Chr.14:4484774-4484754	
TG-F2	5′-gcatcatgcctccagctttgttctt-3′	GRCh38, Chr.19:41104777-41104752	391 bp
WT-R2	5′-gatgttcttgctggcctcct-3′	GRCm38, Chr.14:4484753-4484733	
WT-DF1	5′-aattagcaccgaggggacat-3′	GRCm38, Chr.14:4483642-4483661	872 bp
WT-DR1	5′-ccatgaacccctgacagtcc-3′	GRCm38, Chr.14:4484513-4484494	

Locating the Transgene Insertion Site Through Optical Genome Mapping. As an alternative strategy, optical genome mapping was performed for the CYP2A13/2B6/2F1-transgenic/*Cyp2abfgs*-null mouse genome to locate the transgene insertion site to a specific chromosomal region. The number of detected SVs between the de novo genome (CYP2A13/2B6/2F1-transgenic/*Cyp2abfgs*-null) and the reference genome (GRCm38) is dependent on the size of the SV filter used for the analysis. As shown in Fig. 1A, with a minimum size filter set to 200,000 base pair (bp) for insertion and 1,300,000 bp for deletion, a single deletion and a single insertion were detected in the de novo genome. Consistent with the use of the CYP2A13/2B6/2F1$^{(tg/tg)}$/*Cyp2abfgs*$^{(-/-)}$ mouse for this analysis, the deletion, of at least 1327862 bp (based on the positions of missing markers), was found on Chr.7, between nucleotide positions 25804872 and 27132734, which contains the *Cyp2abfgs* gene cluster in the WT genome. Notably, the gap between the two remaining markers was 85867 bp. Thus, the actual deletion due to Cre-mediated recombination of the floxed *Cyp2abfgs* gene cluster was between 1327862 bp and 1413729 bp, consistent with the 1.4 megabase pair (Mbp) size originally reported (Li et al., 2014).

The insertion, of ~2.4 Mbp, was located on Chr.14, displacing up to 33 kbp of reference mouse genome sequence, between nucleotide positions GRCm38, Chr.14:4451324 and 4485032 (Fig. 1B). The human transgene sequence was detected on the de novo genome sequence, between marker 43 at 273864 bp, which aligned with the marker at 4451324 on the reference Chr.14 sequence, and marker 336 at 2705889 bp, which aligned with the marker at 4485032 on the reference Chr.14 sequence. The size of the transgene insert should approximate the distance between the two flanking markers (43 and 336), which was 2432025 bp (2705889–273864).

When the de novo genome maps of the transgenic mouse were aligned with the in-silico DLE-1-digested human reference genome (GRCh38), six regions, clustered together, were aligned to the region on human Chr.19, between nucleotide positions 40940526 and 41139581, that contained the *CYP2A13, CYP2B6,* and *CYP2F1* genes and several *CYP* pseudogenes (Fig. 2). A detailed examination of each region that aligned with the human transgene sequence showed that some regions (e.g., region 2) contained nearly the entire sequence (~210 kbp) that was included in the original bacterial artificial chromosome clone used to generate the transgenic mouse, whereas others contained only partial copies (e.g., regions 1 and 4). Furthermore, regions 5 and 6 contained sub-regions (5A-D, 6A, and 6B) comprising varying lengths of the human transgene sequence. The human *CYP* genes detected in each region are summarized in Table 2, with the corresponding DLE-1 marker alignments shown in Supplemental Fig. 1 (Panels A–K). Collectively, the transgene sequence appeared to contain seven copies of *CYP2A13*, 6 copies of *CYP2B6*, and five copies of *CYP2F1*. These findings update the previous estimation of the presence of five to six copies of the human *CYP* gene cluster based on densitometry analysis of *CYP2A13* gene restriction fragments detected on southern blots (Wei et al., 2012).

Validation of the Transgene Insertion Site by PCR Amplification and DNA Sequencing. The predicted transgene insertion site is marked by two breakpoints (left and right) in marker alignment between the transgenic mouse genome and the reference mouse and human genomes (Fig. 3A). To verify the insertion site, PCR was performed with a series of primer combinations to amplify across putative junctions between mouse and human gene sequences.

The left breakpoint is expected to be between markers 43 (mouse, at 273864 bp) and 44 (at 280347 bp, which did not align with either mouse or human sequence) in the de novo transgenic mouse genome, which are 6483 bp apart (Fig. 3A). However, attempts to identify the precise junction at the left breakpoint, through sequencing of PCR products (directly or following subcloning), have so far been unsuccessful, due to the presence of large numbers of extensive repeat sequences that reduce PCR specificity.

The right breakpoint is between markers 335 (human, at 2690238, Supplemental Fig. 1K) and 336 (mouse, at 2705889, Fig. 1) in the de novo transgenic mouse genome, which are 15651 bp apart. A greater than 15 kbp PCR product was obtained (Fig. 3B) using the forward primer TG-F1, which was 292 bp upstream of marker 335, and the reverse primer WT-R1, which was 278 bp downstream of marker 336. Sanger sequencing, using WT-R1 as the sequencing primer, identified the precise junction between mouse and human gene sequence, which mapped to GRCh38, Chr.19:41104413 and GRCm38, Chr.14:4484726 (Fig. 3C). Thus, the human transgene sequence replaced a 26.9–33.4 kbp mouse genomic region on Chr.14 (maximally between nucleotides 4451324 and 4484726; minimally between 4457807 and 4484726) (Fig. 3A).

Designing a New Genotyping Method for Differentiating between Hemizygous and Homozygous CYP2A13/2B6/2F1-Transgenic Mice. The new knowledge of the precise transgene insertion site at the right side, the size of the human transgene cluster, and the mouse genome fragment that was replaced by the human transgene made it possible to design a new genotyping method using primers that are within the region of the mouse genome that is replaced by the transgenes. As shown in Fig. 3A, PCR primers WT-DF1 and WT-DR1, both located on the deleted sequence near the right breakpoint, would amplify an 872-bp PCR product from a WT allele, but not from the transgenic allele. Conversely, PCR primers TG-F2 and WT-R2, which flank the breakpoint, would amplify a 391-bp PCR product in the transgenic allele, but not from the WT allele. The specificity of the genotyping method was confirmed in experiments using genomic DNA from WT mice and hemizygous as well as homozygous transgenic mice (Fig. 3, D and E). The genotypes of the transgenic mice were previously verified by breeding test, in which a breeding pair between a homozygous transgenic mouse and a WT mouse would produce pups that are 100% positive for the presence of the transgene, detected using a previous genotyping method for the transgenic allele (Wei et al., 2012). Conversely, a breeding pair between a hemizygous transgenic mouse and a WT mouse would produce pups that are less than 100% positive for the presence of the transgene.

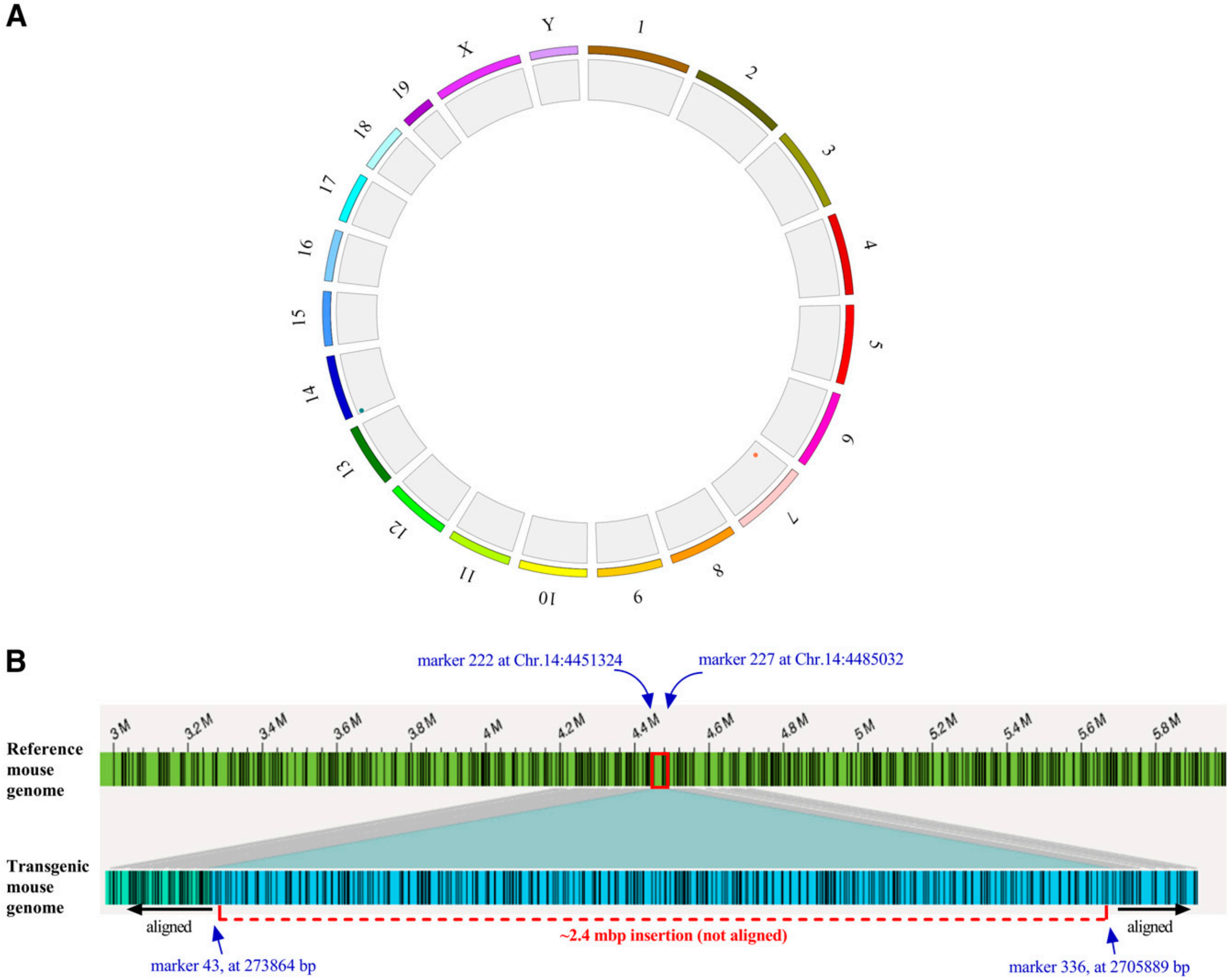

Fig. 1. Alignment of transgenic mouse genome with reference mouse genome. SVs were called by comparing maps of DEL-1 labels between the transgenic (de novo; experimentally determined) and reference mouse genomes (GRCm38; in-silico digested). (A) A global view of the locations of the transgene insertion and the mouse gene deletion detected by optical genome mapping. The outer circle is the cytoband, with the arrayed chromosomes annotated (1–19, X, Y). The inside circle of blocks is the SV track, displaying all SVs detected under the set conditions, including one insertion found on mouse Chr.14 (green dot) and one deletion found on mouse Chr.7 (orange dot). The “minimum size” SV filters were set to 1.3 Mbp for deletions and to 200,000 bp for insertions, which were just below the expected sizes of the target SVs sought (∼1.4 Mbp for the *Cyp2abfgs* cluster (Li et al., 2014) and ∼210 kbp for the *CYP2A13/2B6/2F1* transgene insert (Wei et al., 2012)). (B) Mapping analysis of the transgene insertion site on mouse Chr.14. Maps of DLE-1 labels (vertical black bars) in the de novo genome of the transgenic mouse and those in reference mouse genome (GRCm38) are shown for the transgene insertion site on Chr.14. Nucleotide positions in Chr. 14 of the reference sequence are marked in 0.2 Mbp intervals. The red box shows where the reference mouse sequence was missing. The positions of the DLE-1 markers that flanked the human transgene insert (and the missing mouse reference sequence) are shown.

Searching for Mouse Genes Located within the Region Displaced by the Human Transgenes. Informatics analysis to annotate the mouse genome sequence that was replaced by the human transgenes indicated that there are no known genes in this region, although the replaced region contained putative exons 1–4 of a computationally predicted gene, Gm3182 (GRCm38, Chr.14:4481808-4489857; GRCm39, Chr.14: 17979916-17987965; consisting of 5 exons total; https://www.ncbi.nlm.nih.gov/gene/100041177) (Supplemental Fig. 2). Nucleotide BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) analysis of Gm3182 against the mouse whole genome sequence (GRCm39) indicated that the Gm3182 gene sequence (NC_000080.7) was highly similar (>98.13–99.64% identity) to nine other predicted genes (Supplemental Table 2) and that its predicted mRNA sequence was highly similar (>98.53–99.02% identity) to that of five additional predicted genes (Supplemental Table 3). Protein Blast (https://blast.ncbi.nlm.nih.gov/house_mouse_blastP) analysis indicated proteins encoded by these five transcripts are also highly similar (>96.57–97.55% identity) to Gm3182 protein (Supplemental Table 4). These data suggested that Gm3182 is a highly redundant gene.

Discussion

Whole genome sequencing technology should be able to provide the precise location of a transgene, but it can be challenging if the transgene is located in highly repeated regions. Next-generation DNA sequencing platforms that provide relatively short reads (typically 150–300 bp) (e.g., Illumina) are becoming increasingly affordable, but the limited sequence lengths generated prevent the reads from spanning larger regions that are repetitive and complex, which leads to inaccurate genome assemblies (Salzberg and Yorke, 2005; Alkan et al., 2011; Treangen and Salzberg, 2011; Cameron et al., 2019). In our study, none of the putative insertion sites suggested by the whole genome sequencing data (Supplemental Table 1) was on Chr.14, the correct insertion site. The actual transgene

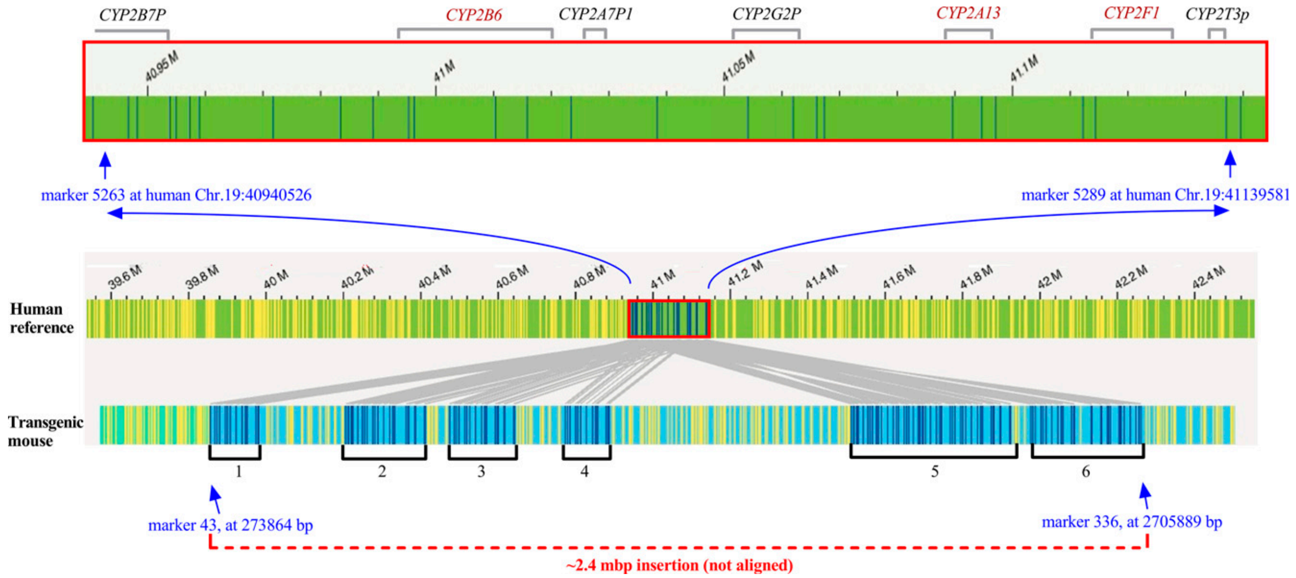

Fig. 2. Mapping analysis of the human transgene insertion. Maps of DLE-1 labels (vertical bars) at the transgene insertion site on Chr.14 of the de novo genome and corresponding sequences on Chr.19 (red box) of the reference human genome (GRCh38) are shown. The positions of human *CYP2A13*, *CYP2B6*, and *CYP2F1* are indicated on top. The 2.4 Mbp transgene insert consisted of six regions that aligned to the reference human genome maps. Results of a detailed alignment of each region or sub-region with the human gene sequence used for generation of the transgenic mouse are shown in Supplemental Figure 1, and the estimated copy number of each of the human *CYP* transgenes is shown in Table 2.

insertion site on Chr.14 is indeed in a highly repetitive region, as illustrated by the multiple homologous sequences identified for Gm3812 (Supplemental Table 2). The fact that the transgene itself was inserted in multiple copies spanning a long, 2.4-Mbp region (Fig. 2 and Supplemental Fig. 1) further adds to the complexity of the sequencing task.

The newer, long-read sequencing platforms (e.g., those provided by Pacific Biosciences and Oxford Nanopore Technologies) can sequence through most repeats and produce more complete genome assemblies (Goodwin et al., 2016; Yuan et al., 2017; Sedlazeck et al., 2018; Amarasinghe et al., 2020). However, their read lengths (∼15 kbp on average) are still insufficient to cover very large repetitive and complex genomic regions (Belser et al., 2018; Yuan et al., 2020).

Optical genome mapping uses a light microscope-based technique to physically locate specific enzyme-catalyzed labeling of sequence motifs to produce DNA sequence fingerprints (Schwartz et al., 1993). The resulting optical maps contain only the physical locations of selected enzyme recognition sites, rather than actual nucleotide sequence information. Optical genome mapping technology utilizes much longer molecules than sequencing does, with read lengths ranging up to 1 Mbp (Yuan et al., 2020). The average molecule length of optical maps (∼225 kbp) is substantially greater than the read length produced by short-read and long-read sequencing (Shelton et al., 2015), a feature that enables optical maps to cover genomic regions that are difficult to resolve by DNA sequencing. Optical genome mapping has found wide application in assisting genome assembly and characterization of complex structural variants (McCaffrey et al., 2017; Levy-Sakin et al., 2019; Young et al., 2020; Wong et al., 2021). Thus, although we have not found prior examples of utility of the optical genome mapping technology for identification of transgene insertion sites, the task is well within its capability.

As demonstrated in the present study of the CYP2A13/2B6/2F1-transgenic mouse, the mapping data provide information on the mouse genomic regions (based on marker position) that are deleted (Figs. 1 and 3), as well as an approximate size and location of the transgene insert (Fig. 2). A comparison of the marker maps of the transgene insert with a reference human genome or the source sequence used for the generation of the transgenic mouse can define the overall structure of the insert and copy numbers of the transgenes. However, optical genome mapping does not provide precise sequence information, which is needed to confirm the location of transgene insertion. The subsequent determination of the precise sequence junction between the transgene

TABLE 2

Human CYP genes identified in the transgenic insert

Regions	Length of the transgene region (bp)	Corresponding region on human Chr.19	*CYP* genes present
1	125,337	40971755-41097091	CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13
2	199,056	40940526-41139581	CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13, CYP2F1
3	167,827	40971755-41139581	CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13, CYP2F1
4	113,604	40983488-41097091	CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13
5A	54,750	40940526-40995275	5/9 exons of CYP2B7p, 1/9 exons of CYP2B6
5B	126,835	41067360-40940526	CYP2A7p1, CYP2G2p, 4/9 exons of CYP2B7p
5C	28,999	41038362-41067360	CYP2G2p
5D	156,094	40983488-41139581	CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13, CYP2F1
6A	113,604	40983488-41097091	CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13, CYP2F1
6B	129,228	41139581-41010354	3/9 exons of CYP2B6, CYP2A7p1, CYP2G2p, CYP2A13, CYP2F1

The human transgene sequence identified in the genome of the CYP2A13/2B6/2F1-transgenic mouse, as shown in Fig. 2, is divided into six regions, with region 5 consisting of four sub-regions and region 6 consisting of two subregions. Detailed alignments, based on DLE-1 marker distribution profile, of each region or sub-region with the human gene sequence used for generation of the transgenic mouse are shown in Supplemental Figure 1. Functional genes are underlined.

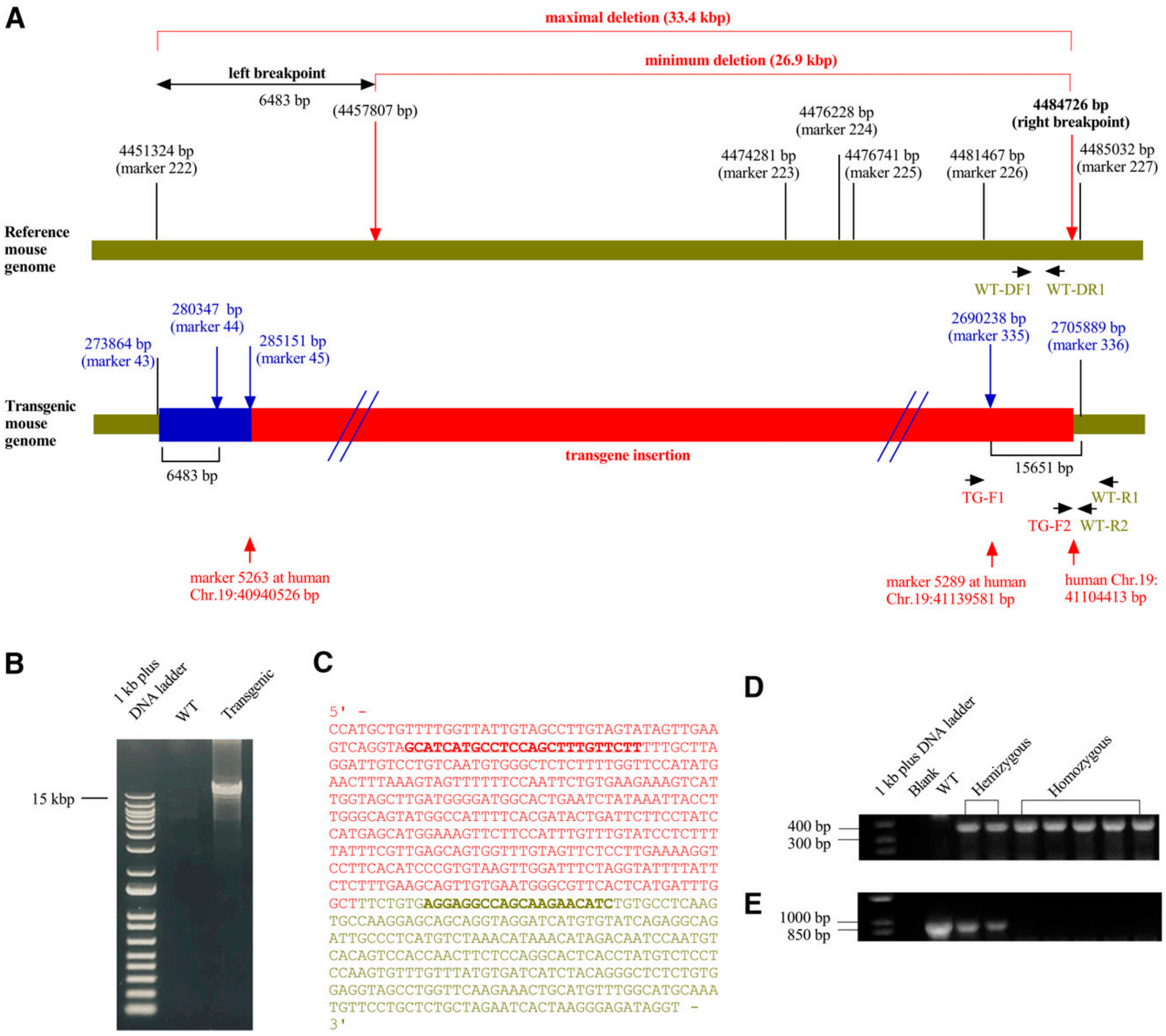
A
maximal deletion (33.4 kbp)
minimum deletion (26.9 kbp)
left breakpoint
6483 bp
(4457807 bp)
4476228 bp
(marker 224)
4484726 bp
(right breakpoint)
4451324 bp
(marker 222)
4474281 bp
(marker 223)
4476741 bp
(maker 225)
4481467 bp
(marker 226)
4485032 bp
(marker 227)
Reference mouse genome
WT-DF1
WT-DR1
280347 bp
(marker 44)
273864 bp
(marker 43)
285151 bp
(marker 45)
2690238 bp
(marker 335)
2705889 bp
(marker 336)
Transgenic mouse genome
transgene insertion
6483 bp
15651 bp
TG-F1
WT-R1
TG-F2
WT-R2
marker 5263 at human
Chr.19:40940526 bp
marker 5289 at human
Chr.19:41139581 bp
human Chr.19:
41104413 bp
B
1 kb plus DNA ladder
WT
Transgenic
15 kbp
C
5' -
CCATGCTGTTTTGGTTATTGTAGCCTTGTAGTATAGTTGAA
GTCAGGTAGCATCATGCCTCCAGCTTTGTTCTTTTTGCTTA
GGATTGTCCTGTCAATGTGGGCTCTCTTTTGGTTCCATATG
AACTTTAAAGTAGTTTTTTCCAATTCTGTGAAGAAAGTCAT
TGGTAGCTTGATGGGGATGGCACTGAATCTATAAATTACCT
TGGGCAGTATGGCCATTTTCACGATACTGATTCTTCCTATC
CATGAGCATGGAAAGTTCTTCCATTTGTTTGTATCCTCTTT
TATTTCGTTGAGCAGTGGTTTGTAGTTCTCCTTGAAAAGGT
CCTTCACATCCCGTGTAAGTTGGATTTCTAGGTATTTTATT
CTCTTTGAAGCAGTTGTGAATGGGCGTTCACTCATGATTTG
GCTTTCTGTGAGGAGGCCAGCAAGAACATCTGTGCCTCAAG
TGCCAAGGAGCAGCAGGTAGGATCATGTGTATCAGAGGCAG
ATTGCCCTCATGTCTAAACATAAACATAGACAATCCAATGT
CACAGTCCACCAACTTCTCCAGGCACTCACCTATGTCTCCT
CCAAGTGTTTGTTTATGTGATCATCTACAGGGCTCTCTGTG
GAGGTAGCCTGGTTCAAGAAACTGCATGTTTGGCATGCAAA
TGTTCCTGCTCTGCTAGAATCACTAAGGGAGATAGGT -
3'
D
1 kb plus DNA ladder
Blank
WT
Hemizygous
Homozygous
400 bp
300 bp
E
1000 bp
850 bp

Fig. 3. Validation of transgene insertion site by PCR and DNA sequencing. (A). An overview of the insertion site, including the estimated location of the left-side breakpoint, the location of the confirmed right-side breakpoint, and the maximal and minimal sizes of the region of the mouse genome that is displaced by the insertion of the transgene. Marker 43 and 336 in the transgenic mouse genome align with marker 222 and 227, respectively, in reference mouse genome (Chr.14), and marker 45 and 335 in the transgenic mouse genome align with marker 5263 and 5289, respectively, of the reference human genome (Chr.19). The positions of the PCR primers (TG-F1 and WT-R1) used to amplify the right-side breakpoint and those used for genotyping analysis (TG-F2, WT-R2, WT-DF1, and WT-DR1) are also shown. (B). PCR amplification of the right breakpoint using primers TG-F1 and WT-R1. Genomic DNA from the transgenic mouse was used as the template. Genome DNA from a WT mouse was used as template in the negative control. (C). DNA sequence surrounding the right breakpoint. The human genome sequence is shown in red (GRCh38, Chr.19:41104826-41104413), while the mouse genome sequence is shown in olive green (GRCm38, Chr.14:4484726-4485264). Locations of new PCR primers (TG-F2 and WT-R2) designed for detection of the breakpoint (for genotyping purpose) are shown in bold and underlined. (D). PCR amplification of the right breakpoint using primers TG-F2 and WT-R2. DNA samples from 1 WT, 2 hemizygous transgenic, and 5 homozygous transgenic mice were analyzed. Selected bands of the 1-kb Plus DNA ladder are shown. The expected product size was 391 bp. (E). PCR amplification of the WT allele using primers WT-DF1 and WT-DR1, which amplify the region of WT mouse genome that is deleted in the transgenic mouse. DNA samples analyzed were the same as in panel D. The expected product size was 872 bp.

and the mouse genome will most likely require the use of long-PCR, given the relatively long average distance between neighboring markers, and primer walking or nested PCR to bridge the gap between the nearest marker and the junction. The selection of PCR primers for the transgene side would benefit from a detailed mapping of the transgene insert, to aid in the selection of nearest regions and correct orientation of transgene sequence for primer design.

In the present study, the right-side junction was fortuitously within a short distance of the nearest mouse marker. However, the left-side junction could not be pinpointed to a specific sequence beyond the nearest markers due to highly repetitive nature of the DNA sequences in the region (Fig. 3). This result demonstrates a limitation of the current approach, that the task of finding the precise junction can be daunting if the insertion is within highly repetitive regions and/or when the distance between the two nearest markers is long.

One benefit of knowing where the junction is between the transgene and the mouse genome is the ability to design genotyping assays that can distinguish hemizygotes from homozygotes. Without a specific genotyping assay, the genotype of a given transgenic mouse is determined by a breeding test, in which the transgenic mouse mates with a WT mouse, and the distribution frequency of the transgene among the pups reveals whether the transgenic parent was a homozygote or a hemizygote. Alternatively, quantitative PCR is performed to determine the abundance of the transgene. These tests are time-consuming (breeding test) or at elevated risks of mistyping (quantitative PCR). In the present study, a new genotyping assay was designed and confirmed to be able to distinguish between hemizygous and homozygous CYP2A13/2B6/2F1-transgenic mice, as well as WT mice. Notably, the repetitive nature of the mouse genome sequence in the region displaced by the CYP2A13/2B6/2F1 transgenes also made it challenging to design a PCR assay specifically for the WT allele. The primers WT-DF1 and WT-DR1 (Table 1 and Fig. 3) are less than ideal because they had only one or two nucleotide mismatches with some other regions of the mouse genome, potentially producing non-specific PCR products that are similar in size to the target amplicon. Thus, a relatively stringent annealing temperature was required to ensure specificity.

The identification of mouse genome regions that are deleted by the insertion of the transgene allows a better assessment of potential functional impact of the transgenesis. In the CYP2A13/2B6/2F1-transgenic mouse, no biologic phenotype, such as changes in fertility, growth, or routine activity, has been observed to date. Thus, any impact of the transgenesis is expected to be subtle or dependent on other internal or external factors. Currently, no annotated gene is identified in the deleted region, and very little is known about the potential function of the computationally predicted gene, Gm3812 (Church et al., 2009; Church et al., 2011). Furthermore, several transcripts and proteins from other genes were found to have nearly identical sequences to those of Gm3812 (Supplemental Tables 3 and 4), which implies that, if Gm3182 is a functional gene, its disruption is unlikely to cause notable functional consequences, given the apparent redundancy.

The size of the transgene construct used for the generation of the transgenic mouse was ∼210 kbp (Wei et al., 2012), which was the basis for setting the minimum insertion size at 200 kbp to search for the transgene insertions. However, the genome mapping data indicated that the size of the transgene insert was ∼2.4 Mbp, which suggested the presence of ∼11 copies of the transgene. Interestingly, the total number of complete and partial transgene copies (Table 2 and Supplemental Fig. 1) was 10, close to the expected value. These accounted for ∼50% of the total amount of genomic sequence within the 2.4-Mbp insert (Fig. 2). The missing sequences from the partial copies were probably intermixed (or mixed with the displaced mouse genome sequences) so thoroughly during the transgene integration process that they are no longer recognized by the mapping algorithm. In that connection, only five to seven copies of the functional human *CYP* genes (*CYP2A13*, *CYP2B6*, *CYP2F1*) are detected by the mapping analysis (Table 2), a result consistent with the previous estimate that was based on abundance of a *CYP2A13*-containing restriction fragment detected on southern blots (Wei et al., 2012). The detailed information on the copy numbers and location of the functional *CYP* transgenes will be useful for future manipulation using gene editing technology.

The tendency to insert multiple transgene copies, as well as the randomness of the insertion site location, are two well-known disadvantages of the commonly used random-integration method for transgenic mouse production (Gordon and Ruddle, 1981; Brinster et al., 1992; Chicas and Macino, 2001; Lampreht Tratar et al., 2018). More sophisticated methods for studying human genes in mice are available, which can circumvent both issues. For example, syntenic replacement is a state-of-the-art method used recently to compare metabolism of arsenic by human and mouse arsenite methyltransferase in a transgenic mouse model (Koller et al., 2020). This and other gene-targeted transgenic approaches (Kumar et al., 2009) eliminate the need to crossbreed mice carrying human transgenes to null mouse lines, though the much increased technical difficult makes them less frequently used than random-integration methods.

Notably, the mapping analysis occasionally fails to detect a marker label when two markers are closely situated (usually less than 500 bp), as they could have aligned to either label in the reference. As a result, only one label was detected, where there are two labels in the reference, as shown in Supplemental Fig. 1. This "skipping" of a marker label may impact the ability to accurately measure distance between two labels, but is unlikely to affect the size of a region that is determined based on the label positions of multiple markers, as is the case for the regions and subregions shown in Supplemental Fig. 1.

In summary, this study provides further characterization of the genomic structure of a transgenic mouse model that expresses three human P450 enzymes important in xenobiotic metabolism and toxicity. The results indicate that the insertion of the human transgenes did not disrupt any known mouse gene, which supports the usefulness of the mouse model. The study also provides a new genotyping method. Furthermore, the study illustrates a novel application of the optical genome mapping technology for locating transgenes in transgenic mouse genome, a challenging task that should be routinely undertaken for all transgenic mouse models that were generated through random integration, which include most of the currently available human P450 transgenic mouse models (Bissig et al., 2018; Hannon and Ding, 2022).

Acknowledgments

We gratefully acknowledge the use of the services of the University of Arizona Animal Care and University of Arizona Genetics Core facility. We thank Dr. Weiguo Han for technical assistance and helpful discussions.

Authorship Contributions

Participated in research design: Ding, Van Winkle, Zhang.

Conducted experiments: Ding, Han.

Performed data analysis: Ding, Han.

Wrote or contributed to the writing of the manuscript: Ding, Han, Van Winkle, Zhang.

References

Alkan C, Sajjadian S, and Eichler EE (2011) Limitations of next-generation genome sequence assembly. *Nat Methods* **8**:61–65.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, and Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**:30.

Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, Genete M, Berrabah W, Chèvre AM, Delourme R et al. (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* **4**:879–887.

Bissig KD, Han W, Barzi M, Kovalchuk N, Ding L, Fan X, Pankowicz FP, Zhang QY, and Ding X (2018) P450-Humanized and Human Liver Chimeric Mouse Models for Studying Xenobiotic Metabolism and Toxicity. *Drug Metab Dispos* **46**:1734–1744.

Brinster RL, Chen HY, Trumbauer M, Senear AW, Warren R, and Palmiter RD (1992) Somatic expression of herpes thymidine kinase in mice following injection of a fusion gene into eggs. 1982. *Biotechnology* **24**:411–419.

Cameron DL, Di Stefano L, and Papenfuss AT (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10**:3240.

Chicas A and Macino G (2001) Characteristics of post-transcriptional gene silencing. *EMBO Rep* **2**:992–996.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M et al.; Mouse Genome Sequencing Consortium (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**:e1000112.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR et al. (2011) Modernizing reference genome assemblies. *PLoS Biol* **9**:e1001091.

Ding X, Li L, Van Winkle LS, and Zhang Q-Y (2018) Biochemical Function of the Respiratory Tract: Metabolism of Xenobiotics, in *Comprehensive Toxicology* (McQueen CA, ed) pp 171–193, Elsevier Ltd, Oxford.

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

Gordon JW and Ruddle FH (1981) Integration and stable germ line transmission of genes injected into mouse pronuclei. *Science* **214**:1244–1246.

Hannon SL and Ding X (2022) Assessing cytochrome P450 function using genetically engineered mouse models. *Adv Pharmacol* **95**:253–284.

Hedrich WD, Hassan HE, and Wang H (2016) Insights into CYP2B6-mediated drug-drug interactions. *Acta Pharm Sin B* **6**:413–425.

Koller BH, Snouwaert JN, Douillet C, Jania LA, El-Masri H, Thomas DJ, and Stýblo M (2020) Arsenic Metabolism in Mice Carrying a *BORCS7/AS3MT* Locus Humanized by Syntenic Replacement. *Environ Health Perspect* **128**:87003.

Kovalchuk N, Zhang QY, Kelty J, Van Winkle L, and Ding X (2019) Toxicokinetic Interaction between Hepatic Disposition and Pulmonary Bioactivation of Inhaled Naphthalene Studied Using

Cyp2abfgs-Null and CYP2A13/2F1-Humanized Mice with Deficient Hepatic Cytochrome P450 Activity. *Drug Metab Dispos* **47**:1469–1478.

Kumar TR, Larson M, Wang H, McDermott J, and Bronshteyn I (2009) Transgenic mouse technology: principles and methods. *Methods Mol Biol* **590**:335–362.

Lampreht Tratar U, Horvat S, and Cemazar M (2018) Transgenic Mouse Models in Cancer Research. *Front Oncol* **8**:268.

Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY et al. (2019) Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10**:1025.

Li L, Carratt S, Hartog M, Kovalchik N, Jia K, Wang Y, Zhang QY, Edwards P, Winkle LV, and Ding X (2017) Human CYP2A13 and CYP2F1 Mediate Naphthalene Toxicity in the Lung and Nasal Mucosa of CYP2A13/2F1-Humanized Mice. *Environ Health Perspect* **125**:067004.

Li L, Megaraj V, Wei Y, and Ding X (2014) Identification of cytochrome P450 enzymes critical for lung tumorigenesis by the tobacco-specific carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK): insights from a novel Cyp2abfgs-null mouse. *Carcinogenesis* **35**:2584–2591.

Li L, Zhang QY, and Ding X (2018) A CYP2B6-humanized mouse model and its potential applications. *Drug Metab Pharmacokinet* **33**:2–8.

Liu Z, Li L, Wu H, Hu J, Ma J, Zhang QY, and Ding X (2015a) Characterization of CYP2B6 in a CYP2B6-humanized mouse model: inducibility in the liver by phenobarbital and dexamethasone and role in nicotine metabolism in vivo. *Drug Metab Dispos* **43**:208–216.

Liu Z, Megaraj V, Li L, Sell S, Hu J, and Ding X (2015b) Suppression of pulmonary CYP2A13 expression by carcinogen-induced lung tumorigenesis in a CYP2A13-humanized mouse model. *Drug Metab Dispos* **43**:698–702.

McCaffrey J, Young E, Lassahn K, Sibert J, Pastor S, Riethman H, and Xiao M (2017) High-throughput single-molecule telomere characterization. *Genome Res* **27**:1904–1915.

Megaraj V, Zhou X, Xie F, Liu Z, Yang W, and Ding X (2014) Role of CYP2A13 in the bioactivation and lung tumorigenicity of the tobacco-specific lung procarcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone: in vivo studies using a CYP2A13-humanized mouse model. *Carcinogenesis* **35**:131–137.

Salzberg SL and Yorke JA (2005) Beware of mis-assembled genomes. *Bioinformatics* **21**:4320–4321.

Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, and Wang YK (1993) Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* **262**:110–114.

Sedlazeck FJ, Lee H, Darby CA, and Schatz MC (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19**:329–346.

Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, Sheth P, and Brown SJ (2015) Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* **16**:734.

Treangen TJ and Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**:36–46.

Uwimana E, Ruiz P, Li X, and Lehmler HJ (2019) Human CYP2A6, CYP2B6, AND CYP2E1 Atropselectively Metabolize Polychlorinated Biphenyls to Hydroxylated Metabolites. *Environ Sci Technol* **53**:2114–2123.

Wang H, Donley KM, Keeney DS, and Hoffman SM (2003) Organization and evolution of the Cyp2 gene cluster on mouse chromosome 7, and comparison with the syntenic human cluster. *Environ Health Perspect* **111**:1835–1842.

Wei Y, Wu H, Li L, Liu Z, Zhou X, Zhang QY, Weng Y, D'Agostino J, Ling G, Zhang X et al. (2012) Generation and characterization of a CYP2A13/2B6/2F1-transgenic mouse model. *Drug Metab Dispos* **40**:1144–1150.

Wong JS, Jadhav T, Young E, Wang Y, and Xiao M (2021) Characterization of full-length LINE-1 insertions in 154 genomes. *Genomics* **113**:3804–3810.

Wu H, Liu Z, Ling G, Lawrence D, and Ding X (2013) Transcriptional suppression of CYP2A13 expression by lipopolysaccharide in cultured human lung cells and the lungs of a CYP2A13-humanized mouse model. *Toxicol Sci* **135**:476–485.

Young E, Abid HZ, Kwok PY, Riethman H, and Xiao M (2020) Comprehensive Analysis of Human Subtelomeres by Whole Genome Mapping. *PLoS Genet* **16**:e1008347.

Yuan Y, Bayer PE, Batley J, and Edwards D (2017) Improvements in Genomic Technologies: Application to Crop Genomics. *Trends Biotechnol* **35**:547–558.

Yuan Y, Chung CY, and Chan TF (2020) Advances in optical mapping for genomic research. *Comput Struct Biotechnol J* **18**:2051–2062.

Address correspondence to: Dr. Xinxin Ding, Department of Pharmacology and Toxicology, College of Pharmacy, University of Arizona, Tucson, AZ 85721. E-mail: xding@pharmacy.arizona.edu; or Qing-Yu Zhang. E-mail: qyzhang@pharmacy.arizona.edu