Large-scale discovery of mouse transgenic integration sites reveals frequent structural variation and insertional mutagenesis

Leslie O. Goodwin, Erik Splinter, Tiffany L. Davis, et al.

Genome Res. 2019 29: 494-505 originally published online January 18, 2019
Access the most recent version at doi:10.1101/gr.233866.117

Resource

Large-scale discovery of mouse transgenic integration sites reveals frequent structural variation and insertional mutagenesis

Leslie O. Goodwin,[1] Erik Splinter,[2] Tiffany L. Davis,[1] Rachel Urban,[1] Hao He,[3] Robert E. Braun,[1] Elissa J. Chesler,[1] Vivek Kumar,[1] Max van Min,[2] Juliet Ndukum,[1] Vivek M. Philip,[1] Laura G. Reinholdt,[1] Karen Svenson,[1] Jacqueline K. White,[1] Michael Sasner,[1] Cathleen Lutz,[1] and Stephen A. Murray[1]

[1]*The Jackson Laboratory, Bar Harbor, Maine 04609, USA;* [2]*Cergentis B.V., 3584 CM Utrecht, The Netherlands;* [3]*The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA*

Transgenesis has been a mainstay of mouse genetics for over 30 yr, providing numerous models of human disease and critical genetic tools in widespread use today. Generated through the random integration of DNA fragments into the host genome, transgenesis can lead to insertional mutagenesis if a coding gene or an essential element is disrupted, and there is evidence that larger scale structural variation can accompany the integration. The insertion sites of only a tiny fraction of the thousands of transgenic lines in existence have been discovered and reported, due in part to limitations in the discovery tools. Targeted locus amplification (TLA) provides a robust and efficient means to identify both the insertion site and content of transgenes through deep sequencing of genomic loci linked to specific known transgene cassettes. Here, we report the first large-scale analysis of transgene insertion sites from 40 highly used transgenic mouse lines. We show that the transgenes disrupt the coding sequence of endogenous genes in half of the lines, frequently involving large deletions and/or structural variations at the insertion site. Furthermore, we identify a number of unexpected sequences in some of the transgenes, including undocumented cassettes and contaminating DNA fragments. We demonstrate that these transgene insertions can have phenotypic consequences, which could confound certain experiments, emphasizing the need for careful attention to control strategies. Together, these data show that transgenic alleles display a high rate of potentially confounding genetic events and highlight the need for careful characterization of each line to assure interpretable and reproducible experiments.

[Supplemental material is available for this article.]

Since the report of the production of the first germline-competent transgenic mouse more than 35 yr ago (Gordon and Ruddle 1981), transgenic mouse models have had an enormous impact on biomedical research, providing a range of tools from critical disease models to more broadly useful reporters and recombinase-expressing lines. The majority of transgenic lines are produced through microinjection of the desired DNA fragment into the pronucleus of a zygote, although lentiviral transgenesis and production through an ES cell intermediate has been reported and used to some extent (Pease et al. 2011). Typically, transgenes comprise engineered DNA fragments ranging in size from small plasmid-based constructs to much larger bacterial artificial chromosomes (BACs), which insert into the genome in a presumably random fashion, usually as a multicopy array. Founder lines are then examined for both transmission and for the desired expression levels and specificity, often leading to the rejection of many lines that fail to express the transgene properly. While the mechanism for this variation in outcome is unclear, it is presumed that genetic context of the integration locus plays some role in providing a transcriptionally permissive environment. There are many additional factors that could affect transgene expression, including copy number, and thus ultimately selection of founders is an empirical exercise and often only a single line is chosen for experiments and publication.

Of the 8012 transgenic alleles published in the Mouse Genome Database, only 416 (5.2%) have an annotated chromosomal location. For transgenic *cre* alleles, the number is even lower, with a known chromosomal location for 36/1631 (2.3%) lines, highlighting the challenge of identification of integration sites despite widespread acknowledgment that such information is useful and important. Low resolution mapping of transgenes can be achieved through FISH or linkage mapping, but these approaches offer little information about potential mutagenesis at the integration site. Inverse PCR can be used to clone the actual fusion sequence but has a high failure rate owing to the multicopy nature of most transgenes. More recently, high-throughput sequencing (HTS) has been employed to identify transgene insertion sites (Dubose et al. 2013), with improvements offered by the use of mate pair libraries (Srivastava et al. 2014). Despite the promise, HTS-based approaches have not seen widespread implementation, possibly due to the cost and/or complexity of the analysis.

The identification of transgene insertion sites is useful for a number of reasons. First, it allows the user to avoid experimental

Corresponding author: steve.murray@jax.org
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.233866.117. Freely available online through the *Genome Research* Open Access option.

29:494–505 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

designs that attempt to combine linked alleles (e.g., a conditional allele with a *cre* transgene), obviating a long and possibly fruitless breeding exercise. Second, it enables the design of allele-specific genotyping assays, which assist in colony management and determination of zygosity. Finally, it alerts the investigator to potential confounding effects of insertional mutagenesis through the direct disruption of the coding sequence of endogenous genes, indirect effects on the regulation of nearby genes, or complex structural variations (inversions or duplications) that can accompany the integration event. Cases of insertional mutagenesis with dramatic phenotypic consequences have been reported. For example: the Tg(TFAP2A-cre)1Will allele inserted into the *Hhat* gene, disrupting its function, leading to a variety of severe developmental abnormalities in homozygous embryos including holoprosencephaly with acrania and agnathia, reflecting a disruption of the hedgehog signaling pathway (Dennis et al. 2012). Given the utility of this line in targeting branchial arches of the developing face, this could confound the interpretation of experimental data if the correct breeding scheme and controls are not included. Because so few insertion sites have been mapped, the scale of this issue is unknown. A prior report using FISH found that transgenes tend to insert into G-positive band regions (Nakanishi et al. 2002), which typically have reduced gene density, but the mapped transgenes were not assessed for expression levels, so it is unclear if these data are representative of transgenes used in the wider scientific community. More recently, targeted locus amplification (TLA) (de Vree et al. 2014; Hottentot et al. 2017) has been employed to identify the insertion site for seven Cre driver lines (Cain-Hom et al. 2017), only one of which was found to insert into an annotated gene. However, because of the small sample size, it is not clear if this rate of mutagenesis is representative of the genome-wide rates in larger collections representing a variety of transgene types.

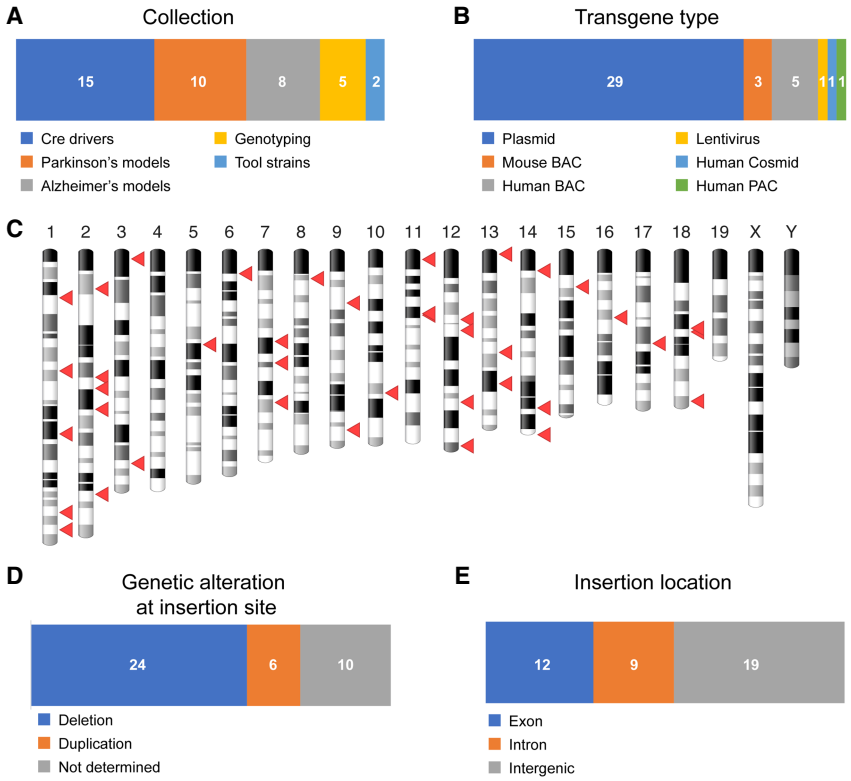

Figure 1. Discovery of the integration loci for 40 transgenic mouse lines. (*A*) Distribution of the categories of transgenes included in this study. (*B*) Distribution of transgenes by molecular type. (*C*) Ideogram showing the physical distribution of transgene insertion sites identified by TLA. (*D*) Types of genetic alterations that accompany transgene insertions. (*E*) Proportion of insertion sites that occur in genes (exon or intron) or nongene loci (intergenic).

Results

We selected a total of 40 transgenic lines from live colonies in The Jackson Laboratory (JAX) Repository for our study, including four lines distributed through the Mouse Mutant Research and Resource Center (MMRRC) at JAX. All lines are broadly utilized and thus represent important research tools that would benefit from insertion site identification. The list comprises 17 genetic tool strains, including 15 Cre drivers, many of which have demonstrated off-target or unexpected excision activity (Heffner et al. 2012). In addition, we included five lines that lack an allele-specific genotyping assay and 18 critical Alzheimer's or Parkinson's disease models (Fig. 1A). We selected lines that were generated through a variety of means (Fig. 1B), including standard small plasmid-based transgenes, human and mouse BAC transgenes, a human PAC, a human cosmid, and a transgene generated through lentiviral-mediated transgenesis. The BAC/PAC/cosmid vectors were included both to capture critical lines of interest and to test the feasibility of the TLA process on these larger constructs. The TLA process, depicted in Supplemental Figure S1, was performed essentially as previously described (de Vree et al. 2014; Hottentot et al. 2017; Methods) using primers specific for known elements of each transgenic line (Supplemental Table S1). Sequence reads from each TLA experiment were mapped to the appropriate reference genome (mouse, human, rat; additional based on predicted elements) and sorted to identify regions with the greatest sequence coverage, which indicates the likely insertion locus. Transgene insertion sites result in high sequencing coverage across the transgene and its insertion site(s), and at least one putative fusion fragment across the transgene-genome breakpoint was identified for all 40 lines (Table 1; Supplemental Table S1). In some cases, follow-up TLA analysis using additional primers designed based on the initial results was required to identify a fusion fragment at either junction. Insertions were found genome-wide on 17/19 autosomes, with five insertions each identified on Chromosomes 1 and 2 (Fig. 1C). Of note, we had one strain overlapping with the Cain-Hom study [Tg(Vil1-cre)997Gum], and we identified the same integration site (Cain-Hom et al. 2017). Structural variations accompanying the insertion were identified for a majority of lines (30/40), comprising 24 deletions and six duplications (Fig. 1D). As some of the fusion contigs were constructed using single reads, we used PCR and Sanger sequencing to verify fusion fragments identified by TLA. In a small number of cases, the confirmed sequence differed from the

Table 1. Summary of transgene insertion sites identified in this study

Allele	JAX Stock #	Target type	Insertion coordinates	Insertion mutation	RefSeq genes affected
Tg(Ins2-cre)25Mgn	3573	Transgene	Chr 7: 62,991,157–63,008,557	Duplication	None
Tg(Alb-cre)21Mgn	3574	Transgene	Chr 13: 3,172,116–3,172,120	4-bp deletion	*Speer6-ps1* (intron)
Tg(Nes-cre)1Kln	3771	Transgene	Chr 12: 90,524,592–90,524,609	17-bp deletion	None
Tg(Lck-cre)548Jxm	3802	Transgene	Chr 11: 41,490,714	Duplication	None
Tg(Tek-cre)12Flv	4128	Transgene	Chr 13: 68,459,931–68,701,276	241-kb deletion	*Mtrr*, *Fastkd3*, *1700001L19Rik*, *Adcy2*
Tg(Vil1-cre)997Gum	4586	Transgene	Chr 17: 55,326,957–55,341,510	14.6-kb deletion	None
Tg(Ddx4-cre)1Dcas	6954	Transgene	Chr 18: 85,696,612–86,794,868	1098-kb deletion	*Neto1*, *Cbln2*
Tg(UBC-cre/ERT2)1Ejb	8085	Lenti transgenic	Chr 2: 25,249,816–25,249,821	5-bp deletion	*Ndor1*
Tg(Cspg4-cre)1Akik	8533	Mouse BAC	Chr 1: 173,692,115	ND	*Ifi208* (intron)
Tg(Cspg4-cre/Esr1*)BAkik	8538	Mouse BAC	Chr 14: 106,654,779–106,655,407	628-bp deletion	None
Tg(Th-cre)1Tmd	8601	Transgene	Chr 9: 33,514,690–34,139,124	624-kb deletion	*7630403G23Rik*
Tg(Vav1-icre)A2Kio	8610	Transgene	Chr 18: 47,022,629	ND	*Commd10* (intron)
Tg(Wnt1-cre)11Rth	9107	Transgene	Chr 11: 6,425,500–6,456,783	31.2-kb deletion	*H2afv*
Tg(Sox2-cre)1Amc	14094	Transgene	Chr 13: 89,311,726	ND	*Edil3* (intron)
Tg(Wnt1-cre)2Sor	22501	Transgene	Chr 2: 154,561,346–154,561,603	257-bp deletion; complex inversion	*E2f1*
Tg(Itgax-DTR/EGFP)57Lan	4509	Transgene	Chr 1: 80,448,681–80,455,738	7057-bp deletion	*1700016L21Rik*
Tg(Camk2a-tTA)1Mmay	7004	Transgene	Chr 12: 116,101,154–116,609,271	508-kb deletion	*Vipr2*, *Wdr60*, *Esyt2*, *D430020J02Rik*, *Ncapg2*, *Ptprn2*
Tg(GFAP-APOE_i4)1Hol	4631	Transgene	Chr 15: 23,364,633–23,373,281	8648-bp deletion	*Cdh18*
Tg(APPSwe,tauP301L)1Lfa	4807	Transgenes	Chr 2: 87,862,466–87,862,463	3-bp deletion	None
Tg(MAPT)8cPdav	5491	Human PAC	Chr 7: 10,447,768	ND	*Trim30d* (intron)
Tg(APPswe,PSEN1dE9)85Dbo	5864	Transgenes	Chr 9: 113,003,660	Duplication	None
Tg(PDGFB-APPSwInd)20Lms	6293	Transgene	Chr 16: 43,086,322–43,127,049	40.7-kb deletion	*Zbtb20* (intron)
Tg(Prnp-MAPT*P301S)PS19Vle	8169	Transgene	Chr 3: 140,354,280–140,603,283	249-kb deletion	None
Tg(APPSwFlLon, PSEN1*M146L*L286V) 6799Vas	8730	Transgenes	Chr 3: 6,297,836	ND	None
Tg(tetO-MAPT*P301L) 4510Kha	15815	Transgene	Chr 14: 124,457,842–124,702,169	244-kb deletion	*Fgf14*
Tg(Prnp-SNCA*A53T)83Vle	4479	Transgene	Chr 12: 48,212,716	ND	None
Tg(Prnp-SNCA*A53T)23Mkle	6823	Transgene	Chr 10: 95,350,683–95,399,000	48.3-kb deletion	*2310039L15Rik*
Tg(LRRK2*R1441G)135Cjli	9604	Human BAC	Chr 1: 32,289,302–32,289,738	436-bp deletion	*Khdrbs2* (intron)
Tg(Lrrk2*G2019S)2Yue	12467	Mouse BAC	Chr 18: 44,968,085	ND	None
Tg(LRRK2)66Mjff	13725	Human BAC	Chr 6: 16,279,287–16,327,995	756-bp deletion	None
Tg(Thy1-SNCA)12Mjff	16936	Transgene	Chr 14: 14,719,103	Duplication	*Slc4a7* (intron)
Tg(Thy1-SNCA)15Mjff	17682	Transgene	Chr 11: 40,456,787–40,495,044	38.4-kb deletion	None
Tg(SNCA)129Mjff	18442	Human BAC	Chr 7: 77,604,164–77,605,062	898-bp deletion	None
Tg(LRRK2*G2019S)2AMjff	18785	Human BAC	Chr 1: 80,896,405	ND	None
Tg(LRRK2*R1441G)3IMjff	18786	Human BAC	Chr 1: 121,956,000–121,995,855	39.9-kb deletion	None
Tg(HLA-A/H2-D)2Enge	4191	Transgene	Chr 12: 41,759,331–41,760,601	1279-bp deletion	*Immp2l* (intron)
Tg(CAG-FCGRT)276Dcr	4919	Transgene	Chr 1: 185,129,377	Duplication	None
Tg(FXN*)1Sars	8586	Transgene	Chr 5: 61,755,638	ND	None
Tg(HLA-A2.1)1Enge	9617	Transgene	Chr 8: 18,736,683–18,757,058	Duplication	*Mcph1* (intron), *Angpt2*
Tg(FCGRT)32Dcr	14565	Human cosmid	Chr 2: 101,081,712	ND	None

reconstructed fusion contig, highlighting the critical need for independent confirmation of breakpoint sequences. These confirmation assays also provide an allele-specific genotyping assay for each transgenic line for unambiguous identification of hemizygous animals. Overall, we identified and confirmed both fusion breakpoints for 19/40 transgenes and a single fusion breakpoint for an additional 21, demonstrating the efficiency of the TLA process in identifying the precise insertion sites (Supplemental Tables S1, S2). For deletions where only one fusion fragment could be confirmed, a quantitative PCR loss-of-native-allele (LOA) (Valenzuela et al. 2003; Frendewey et al. 2010) assay was used to confirm the loss of either the genes within the deletion (see below) or a region close to the estimated insertion site (Supplemental Tables S1, S3). We found the transgene insertion event in 21 of 40 lines to be either a deletion of at least one exon of one or more RefSeq genes (12) or an insertion into an intron, likely affecting its normal transcription (9) (Fig. 1E). Transgene insertions and accompanying deletions also affected multiple noncoding features, including lincRNAs, miRNAs, and snRNAs (Supplemental Table S1). Overall, these data indicate a strong enrichment of transgene insertion events in genic regions of the genome, placing these lines at high risk for confounding phenotypes due to insertional mutagenesis.

A majority of insertion events discovered were accompanied by a deletion (Fig. 1D), which varied in size from a few base pairs to 1.1 Mb in the case of the Tg(Ddx4-cre)1Dcas (Fig. 2A). As noted above, among the 24 deletions, we identified a high rate of insertional mutagenesis, either deleting or disrupting between one and six mouse genes (Fig. 2B), each with potential phenotypic consequences. In addition, three genes are disrupted through duplication events that accompanied their respective insertion (Supplemental Table S1). Of the total 31 genes disrupted, 16 have a previously reported knockout (KO) phenotype, including

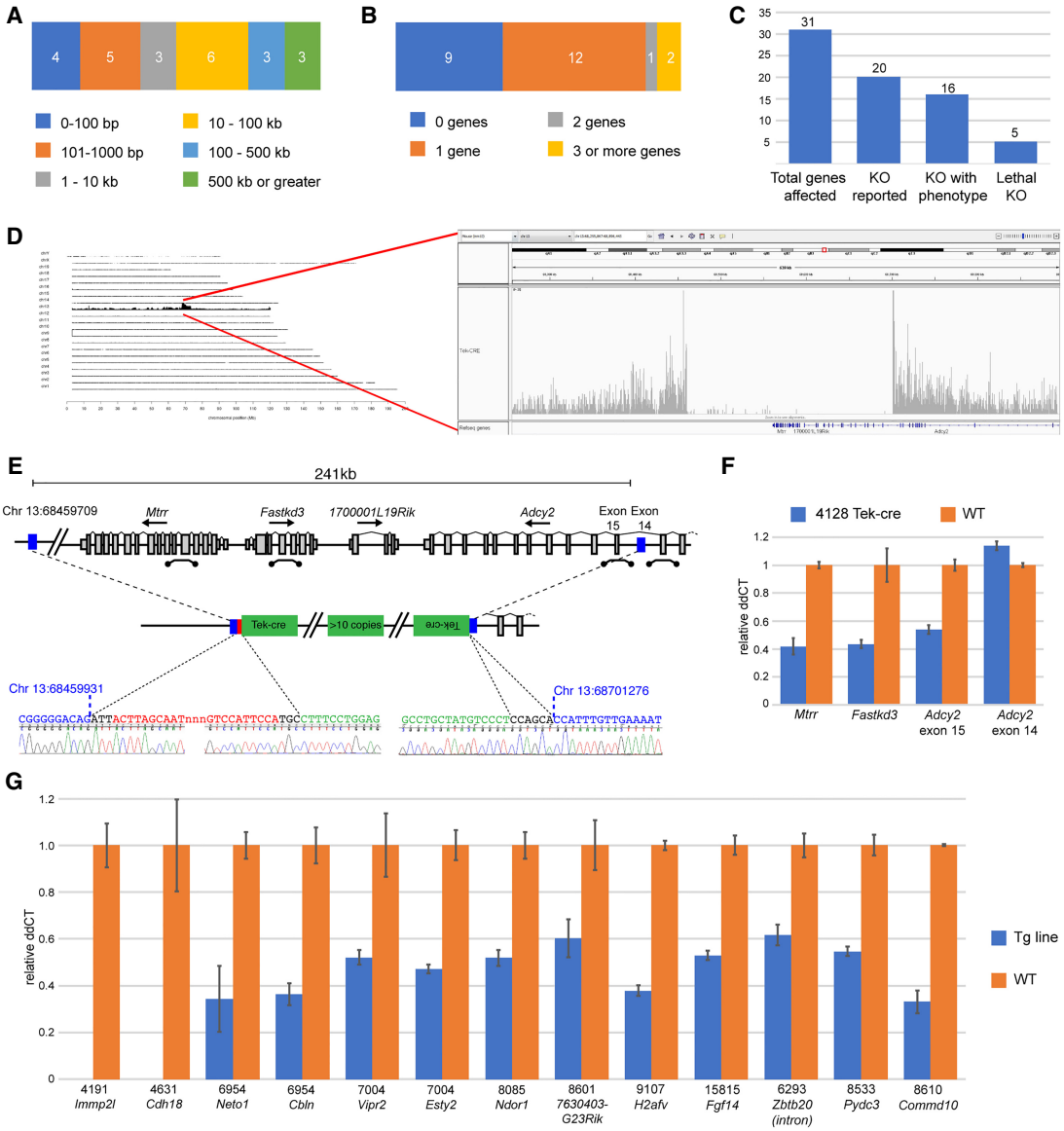

Figure 2. Deletions accompanying transgenic insertions. (*A*) Profile of sizes of deletions identified at integration loci. (*B*) For integrations that occur in genes, the profile of the number of genes affected by the insertion event. (*C*) Illustration of the potential impact of transgene insertions into genes, including the number of genes with reported knockout (KO) alleles, the number of KO alleles with a reported phenotype, and number of genes shown to be essential for life. (*D*) Genome-wide and zoomed Chr 13 view of TLA reads mapped to the mouse genome. (*E*) Schematic of the insertion locus in the Tek-cre [Tg(Tek-cre)12Flv] line. Blue bars indicate the 5′ and 3′ limits of the deleted region, with the relative orientation of transgene copies adjacent to the breakpoint as determined from sequence-confirmed fusion fragments. Locations of qPCR probes to confirm copy number are shown. (*F*) Results of LOA qPCR assays showing the expected loss of one copy of *Mtrr* and *Fastkd3* and exon 15 of *Adcy2*, which lie within the deletion. *Adcy2* exon 14, which lies outside of the deletion, has the expected two copies. WT copy number is arbitrarily set at 1, thus a value of 0.5 would indicate loss of one copy. (*G*) LOA assays for 13 other genes/loci deletions identified in this study. Strains are indicated by Stock # *above* the gene symbol for each test. For strains 4191 and 4631, the complete loss of *Immp2l* and *Cdh18*, respectively, is consistent with the homozygous maintenance of these lines.

five with embryonic lethal phenotypes (Fig. 2C), and multiple genes are deleted in three lines, highlighting the potentially confounding effect of the insertion event. For example, in the Tg (Tek-cre)12Flv transgene, we identified a 241-kb deletion in Chromosome 13 that includes four protein coding genes (*Mtrr, Fastkd3, 1700001L19Rik, Adcy2*) (Fig. 2D,E). We validated both breakpoints and confirmed the deletion of *Mtrr*, *Fastkd3*, and *Adcy2* by loss-of-native-allele qPCR assays, showing clear loss of one copy for all three genes and in addition confirming the breakpoint between exons 14 and 15 in *Adcy2* (Fig. 2F). In addition to the frequent widespread (off-target) activity seen in this line (Heffner et al. 2012), prior reports show that an *Mtrr* gene trap allele exhibits trangenerational epigenetic effects leading to severe developmental abnormalities when breeding from a female carrier (Padmanabhan et al. 2013). Given the common use of this line to analyze vascular development, the transgene itself could confound analysis depending on the breeding scheme, highlighting the need for proper controls (i.e., Cre-only) in studies using this transgene. Finally, we confirmed the deletion of an additional 13 genes (Fig. 2G; Supplemental Table S1), including two (*Immp2l* and *Cdh18*) that appear

null due to the maintenance of the line as a homozygote. One transgene [Tg(PDGFB-APPSwInd)20Lms] showed insertion into the gene *Zbtb20*, but a recent report clearly shows that despite the insertion of more than 10 copies of the transgene, expression of the protein in heterozygous transgenic mice is comparable to WT, suggesting other regulatory mechanisms to maintain a uniform level of expression (Tosh et al. 2017). Therefore, in some cases the consequences of intron insertion require independent validation. Together, these data show that transgene insertions are often associated with large mutagenic deletions, affecting one or more genes, potentially confounding interpretation of results unless the proper control strategies are employed.

TLA analysis also revealed additional structural variations around the insertion site, including six instances of duplications and one inversion accompanied by a duplication and large deletion. In many duplication cases, fusion fragments were only identified on one end of the transgene insertion, so the exact extent of the duplication could only be estimated through read depth. However, we were able to confirm additional copies of parts of the genes *Mcph1* and *Slc4a7* (Supplemental Fig. S2), although it is not clear how this might affect gene function. For Tg(Wnt1-cre)2Sor (Lewis et al. 2013), we observed a complex structural variation on Chromosome 2, involving a large 45-kb inverted segment inserted into exon 5 of the *E2f1* gene (Fig. 3A). The inversion itself contains all of exon 5 but deletes 23 kb including exons 6 and 7 of *E2f1* proximal to the transgene insertion location, all of *Necab3, 1700003F1Rik,* and a portion of the 3′ UTR of *Cbfa2t2*. As a result, the structural variation disrupts the *E2f1* gene, with the concomitant duplication of exon 5 in the opposite orientation. We used an LOA assay to confirm the disruption of exon 5 (Fig. 3B), but this does not capture the duplication of the inverted exon, as the inverted fragment is smaller than the amplicon of the qPCR probe. The copy number of *E2f1* exons 4 and 6 are unaffected, as they surround the structural variation. In addition, an LOA assay shows the duplication of exon 6 of *Cbfa2t2*, and an LOA for exon 2 of *1700003F12Rik*, which resides in the deleted portion of the duplicated fragment, shows the normal two copies as expected (Fig. 3B). Together, these data illustrate the potential complex structural variations that can occur with transgene integration.

Because TLA isolates all DNA fragments in close proximity to the transgene integration site, it is possible to identify components of the transgene itself, in addition to the surrounding mouse sequence. The only limitation is the selection of reference genomes for mapping. In this study, we typically mapped to genomes predicted to be part of the transgene, based on the published description of the transgene construction. While for the most part we were able to identify construct elements described in the original publications, unexpected components were seen in several transgenic lines. For example, we found that an entire human growth hormone (*GH1*, also known as *hGH*) minigene, described as a poly(A) sequence, was present in four lines [Tg(Ins2-cre)25Mgn, Tg(Alb-cre)21Mgn, Tg(Nes-cre)1Kln, and Tg(Lck-cre)548Jxm] (Fig. 4A), as previously reported for Tg(Nes-cre)1Kln (Declercq et al. 2015). Of note, publications for these lines reference a vector originally described in Orban et al. (1992), which clearly describes the minigene structure of the cassette. Similarly, TLA and PCR validation of the Tg(Vil1-cre)997Gum line reveals the presence of the entire mouse *Mt1* gene sequence (Fig. 4B), despite its description as a "metallothionein poly(A) signal" in the original publication (Madison et al. 2002). The source plasmid does indeed describe it as containing the poly(A) and several introns (Sauer and Henderson 1990). Although the impact of the presence of the *Mt1* minigene is unclear, there is evidence that the Nes-cre *hGH* minigene is expressed and that this expression is responsible for some of the metabolic phenotypes observed in mice carrying Tg(Nes-cre)1Kln (Galichet et al. 2010; Giusti et al. 2014). These data indicate that TLA has the added potential to expand on and confirm reported transgene composition, and in some cases can correct, clarify, and/or update the record for these strains.

In our initial analysis, we identified several cases where mapped sequences were fused to unknown sequences. Further analysis revealed that some of these fusions were with the *Escherichia coli* genome and not vector sequences, suggesting that fragments of contaminating *E. coli* DNA were cointegrating with the transgene. To assess the frequency of this phenomenon, we mapped all of the TLA data to the *E. coli* genome (K-12) and found evidence for cointegration in 10/40 strains, with total composition ranging from as little as 300 bp to more than 200 kb (Fig. 4C). Some of the small fragments identical between samples align to cloning vectors in the NCBI database, including pBACe3.6, which is documented to be the cloning vector for both the 12467-Lrrk2*G2019S and 18442-SCNA models. However, for lines with significant *E. coli* genome contribution, it is likely that this is the result of contamination in the microinjection preparation of the construct. While the impact of this finding for these specific lines is unclear, prior reports have shown that bacterial sequences can contribute to transgene silencing (Scrable and Stambrook 1997; Chen et al. 2004).

While some of the small number of transgene insertion sites currently known were discovered following the serendipitous identification of an unexpected transgene-specific phenotype, systematic phenotyping of transgenic lines to assess the impact of transgene insertion has not been reported. Taking advantage of the high-throughput KOMP2 Phenotyping platform at JAX (White et al. 2013; de Angelis et al. 2015; Dickinson et al. 2016; Karp et al. 2017; Meehan et al. 2017), we asked whether we could detect phenotypes in a selection of seven Cre driver lines from the lines examined above. Cohort genotypes varied according to colony maintenance strategy (see Methods), and we created a reference population by pooling WT C57BL/6J and control mice from lines maintained as hemizygotes. As shown in Figure 5 (full table of results in Supplemental Table S5), we identified 66 significant phenotypes among strains, with Tg(Nes-cre)1Kln displaying the most phenotype hits (21) and Tg(Vil1-cre)997Gum displaying the least (two). Physiology phenotypes were most common, with both Tg(Nes-cre)1Kln and Tg(Ins2-cre)25Mgn showing 19 and 13 abnormalities, respectively. As noted above, metabolic phenotypes in Tg(Nes-cre)1Kln mice have been reported by others (Galichet et al. 2010; Giusti et al. 2014), and a recent paper suggests these phenotypes are due to the presence of the *hGH* minigene (Declercq et al. 2015). The same *hGH* minigene is also found in the Tg(Ins2-cre)25Mgn allele, possibly explaining the metabolic phenotypes observed. This line is reported to develop age-related impaired glucose tolerance of unknown etiology (Lee et al. 2006), and thus the presence of the *hGH* minigene should be explored as a possible explanation. Both Tg(Alb-cre)21Mgn and Tg(Lck-cre)548Jxm also carry this minigene, but do not show the same number of phenotypic hits, suggesting that expression of the minigene varies between transgenes, and thus it cannot be assumed that its presence alone is necessarily confounding. It is interesting that Tg(Vav1-icre)A2Kio displayed only two hits despite landing in the *Commd10* gene, for which a KO allele is homozygous lethal. This is consistent, however, with the International Mouse Phenotyping Consortium (IMPC) phenotyping data

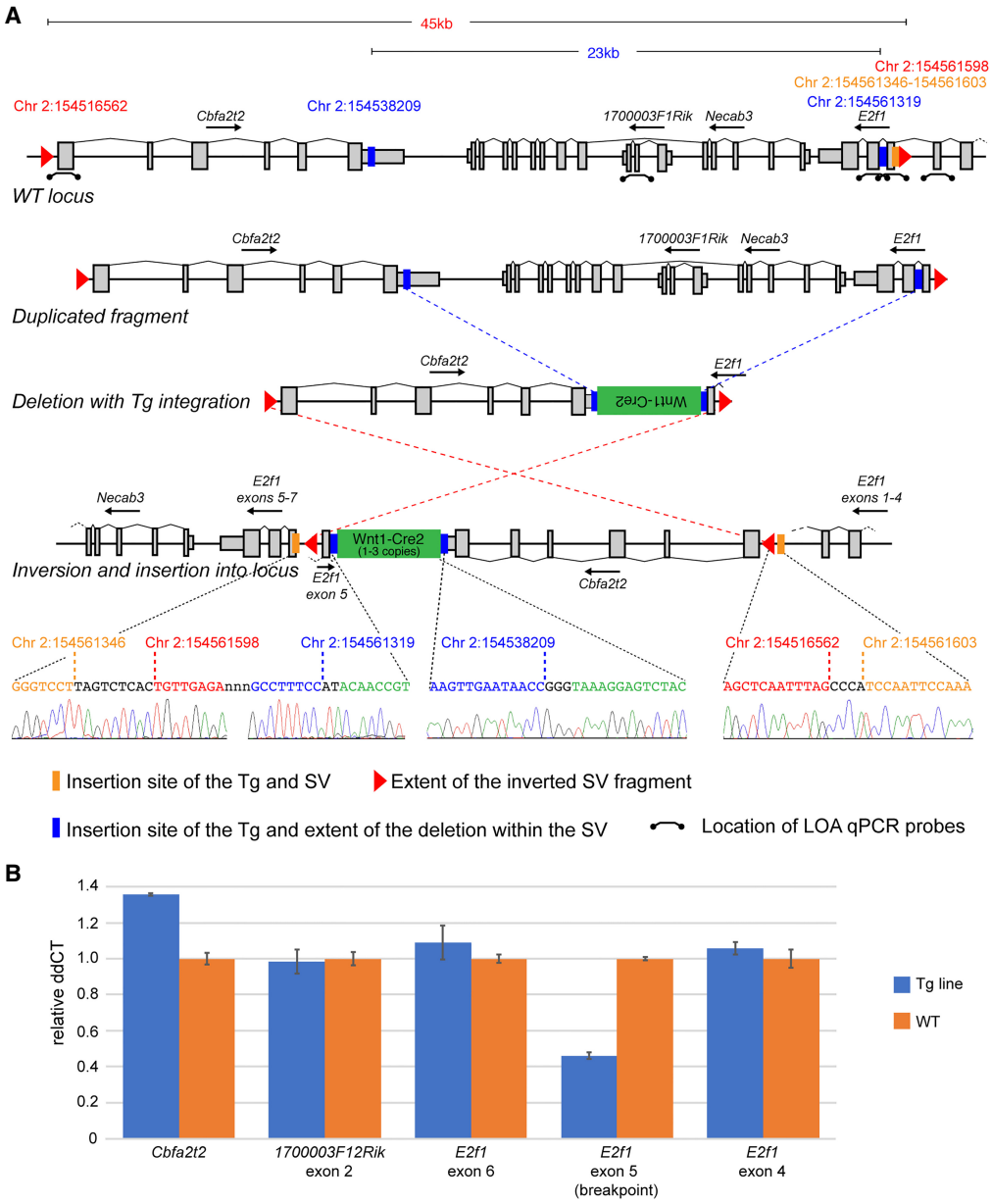

Figure 3. Complex structural variations (SVs) accompanying transgenic insertions. (*A*) Schematic of the SV accompanying the Wnt1-cre2 [Tg(Wnt1-cre)2Sor] transgene insertion. The locus includes a large duplication with a partial deletion that accompanies the transgene insertion. The entire duplicated interval is inverted and is inserted into exon 5 of the *E2f1* gene. The red triangles identify the extent of the entire SV that is inverted, the blue bars indicate the insertion site of the transgene and the extent of the deletion within the duplicated fragment, and the orange bars indicate the location of the SV insertion. qPCR probes are indicated on the WT locus. The qPCR probe for *E2f1* exon 5 spans the breakpoint of SV insertion. Confirmation of each fusion fragment that defines the SV by PCR-Sanger sequence is illustrated. (*B*) LOA confirmation of the expected copy number for each gene/exon affected by the SV.

(www.mousephenotype.org), which shows no significant phenotypes in *Commd10*$^{tm1a(EUCOMM)Wtsi/+}$ mice. In contrast, the Tg (Wnt1-cre)11Rth transgene inserts into the histone gene *H2afv* and shows 11 phenotypic hits, which span several domains (Fig. 5), including four significant behavioral phenotypes. Currently, there are no reports of targeted mutations of phenotypes for this gene.

Discussion

Despite widespread use of transgenic lines in the scientific community, the consequences of random transgene insertions in these lines is largely unknown. Here, expanding on prior work, we show that TLA represents a rapid and efficient means to precisely identify the site of insertion, and critically, the corresponding molecular consequences. These events are associated with structural variations, primarily deletions and duplications, including a deletion of greater than 1 Mb and a complex structural variation that includes a simultaneous duplication, deletion, and inversion. While TLA simplifies the discovery process, reconstruction of the full transgene structure is difficult due to the high copy concatemerization of the transgene, coupled with complex structural variation that can accompany insertion. Given that these contigs are

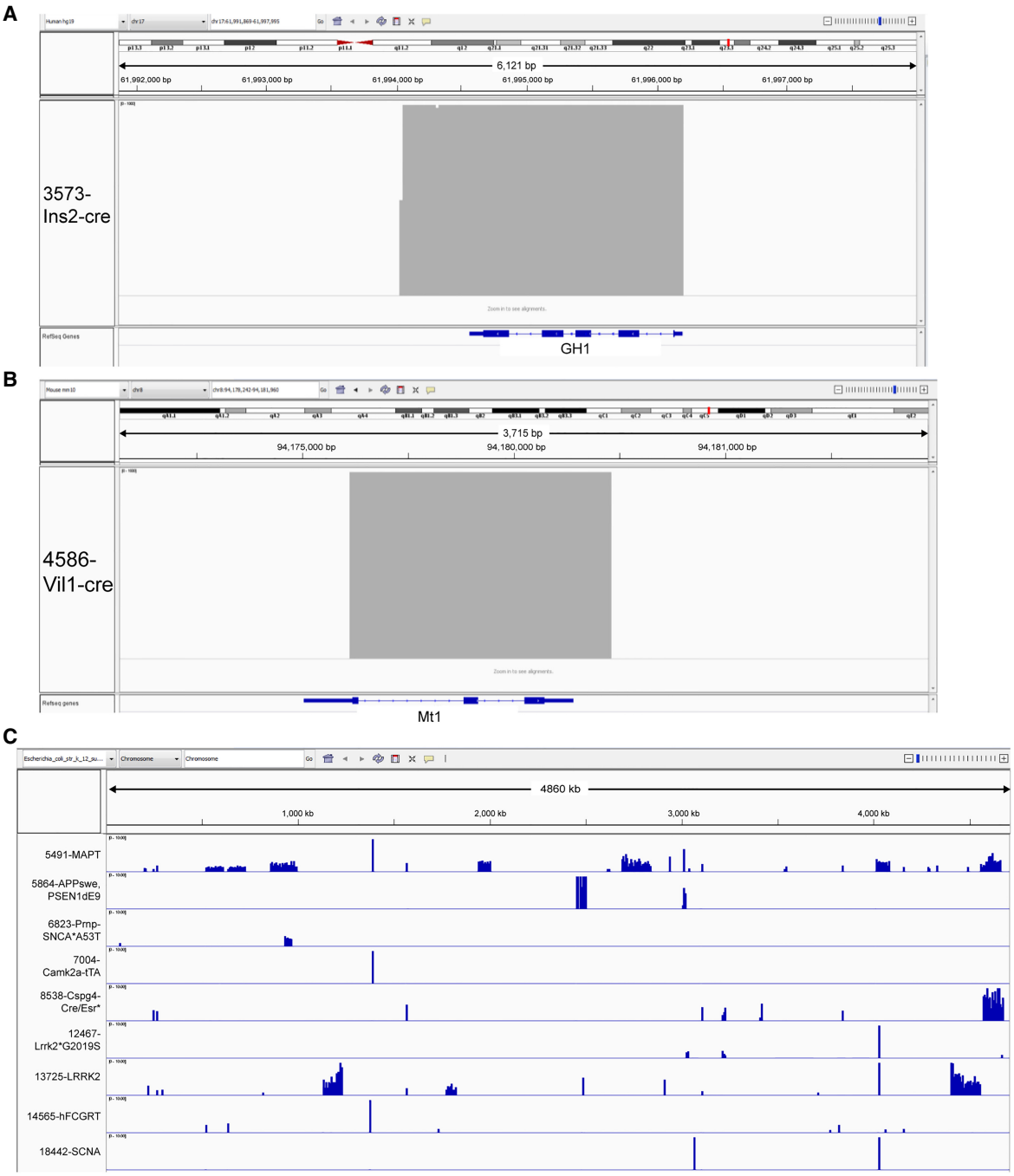

Figure 4. TLA reveals additional passenger cassettes and fragments in transgenes. (*A*,*B*) View of TLA reads (indicated in gray *above* the gene model) that map to the human growth hormone (*GH1*, also known as *hGH*) and the mouse metallothionein (*Mt1*) gene for two transgenes (Ins2-cre and Vil-cre, respectively), showing the inclusion of the entire gene structure, including coding exons. (*C*) Reads for nine transgenes mapped to the *Escherichia coli* genome indicating a variable level of coinsertion into the transgene integration site. Deep coverage for discrete loci shared between multiple lines indicates sequences that are part of the transgene vector. The amount of *E. coli* cointegration ranges from a few hundred bp to more than 200 kb. Short names for each transgene are used for readability and are defined in Supplemental Table S1.

built using relatively short-read sequencing technology, including several instances where the fusion was covered by one read, it is critical to validate each putative fusion fragment using PCR-sequencing, and we indeed found small differences in the actual fusion sequence in a few lines. Thus, TLA should be considered a "first pass" tool for integration locus discovery, and full reconstruction of both the integration consequence and the transgene itself require substantial follow-up effort. Targeted long-read sequencing approaches (Pacific Biosciences, Oxford Nanopore), might prove to be a useful complement to TLA for full characterization of transgenic alleles. However, such reconstruction is not necessary for typical use, as the key elements of chromosome location,

Figure 5. Physiology and behavioral testing of the *cre* transgenic lines in the KOMP pipeline. Mice were tested in 12 phenotypic domains spanning behavior and physiology (color-coded, *right* bar). Each test is further grouped broadly into behavior (peach) or physiology (gray) domains. Significant differences from controls are shown in the heatmap (FDR-corrected *P*-values). Individual output parameters are listed and color-coded on the *left y*-axis. Tests with no data are shown in gray. (ABR) Auditory brainstem response; (HB) hole board; (LD) light/dark transition; (OFA) open field assay; (PPI) pre-pulse inhibition; (RR) rotarod; (SLEEP) piezoelectric sleep/wake; (TST) tail suspension test; (BC) body composition; (CBC) clinical biochemistry; (ECT) electroconvulsive seizure threshold; (HEM) hematology. Homozygous transgenic lines: Alb-cre [B6.Cg-Tg(Alb-cre)21Mgn/J], Ins2-cre [B6.Cg-Tg(Ins2-cre)25Mgn/J], Lck-cre [B6.Cg-Tg(Lck-cre)548Jxm/J]. Hemizygous transgenic lines: Nes-cre [B6.Cg-Tg(Nes-cre)1Kln/J], Vav1-cre [B6.Cg-Tg(Vav1-icre)A2Kio/J], Vil1-cre [B6.Cg-Tg(Vil1-cre)997Gum/J], Wnt1-cre [B6.Cg-Tg(Wnt1-cre)11Rth Tg(Wnt1-GAL4)11Rth/J].

breakpoint, and structural variation are easily obtained with TLA and PCR validation alone.

In contrast to a recent similar screen on a small number of Cre driver lines (Cain-Hom et al. 2017), we found that a high percentage (50%) disrupt annotated genes, the majority of which are protein coding. Many of these genes, when mutated, are known to result in observable phenotypes, including four that are homozygous lethal. At face value, this is unexpected as only 3% of the genome is protein coding, and thus we would expect a similar hit rate with truly random insertion. However, many of these insertions

are accompanied by substantial deletions, which would increase the odds of hitting a gene. In addition, our set of transgenic lines is not an unbiased selection of random transgenic animals; they are lines selected for robust transgene expression and activity. Genic regions of the genome are likely to support active transgene expression, as opposed to intergenic stretches and regions of heterochromatin. While we do observe that larger transgenes derived from BAC (or similar) constructs, which contain a larger complement of genetic elements required for proper expression, can insert into and/or delete coding genes [e.g., Tg(Cspg4-cre)1Akik], our sample size is too small to determine if the rate is different than that of smaller transgenic constructs.

TLA-based discovery of transgenic insertion sites provides a number of practical benefits that should improve quality control for both public repositories and the end user. For example, allele-specific assays can be developed at the integration site to distinguish all genotype classes, allowing for homozygous mating strategies unless precluded by insertional mutagenesis. End users of Cre lines can use knowledge of the genetic locus before attempting to mate to a floxed target allele that is linked to the Cre line, selecting an alternative unlinked line, or scaling their breeding to assure identification of rare recombinants.

Our findings further illustrate the need to use proper controls in all experiments that include transgenic lines. Several studies have shown that expression of Cre itself can have phenotypic or toxic effects (Loonstra et al. 2001; Naiche and Papaioannou 2007; Bersell et al. 2013; Lexow et al. 2013). Given these findings, it is clear that animals/embryos expressing Cre alone must be included as a control, and our results provide additional evidence that this control strategy is essential. Therefore, the potentially confounding impact of frequent insertional mutagenesis in Cre driver lines can be managed with the proper use of controls, depending on the research question and phenotype. For transgenic lines that are employed as disease models, a Tg-only control is not possible. In many cases, the original publication included results from multiple founders corroborating the findings of the line that ultimately became the "standard" for subsequent studies. It is interesting, however, that most studies published with an "established" model do not include the same level of independent corroboration, despite significant differences in study design, including analysis of additional phenotypes, inclusion of additional mutant alleles, and/or the use of a distinct genetic background. It is plausible that in those scenarios, effects of insertional mutagenesis not seen in the original publication might manifest, confounding the interpretation of the data. Typically, multiple alleles are not deposited in a public biorepository for distribution or retained at all, and thus reproduction of results with independent transgenic lines is impossible, notwithstanding the practical and financial challenge of reproducing every study with multiple transgenic lines. Thus, it seems prudent, and now feasible, for investigators to determine the insertion site of the transgenic line used in their study if independent corroboration is not possible.

One potential use of TLA is to confirm the content of a given transgenic line, providing a level of quality control not available through other means. This includes clarification of the specific details of the constructs components (e.g., *hGH* or *Mt1* minigenes) that are either omitted or reported incorrectly. It is worth noting that in the case of both "poly(A)" signals reported here, a careful review of the literature clearly shows that the original content of the vectors is correctly reported (Sauer and Henderson 1990; Orban et al. 1992), but this information was omitted or incorrectly cited in subsequent descriptions of the construct or mouse strain. Given the number of years and hands involved, this type of "information mutation" is not surprising. Indeed, we have seen that ~20% of all lines submitted to the JAX Repository carry alleles or have been bred to mouse strains not reported by the donating investigator. TLA provides an additional tool for assuring the content and nature of the allele for both investigators sharing their strain and for repositories distributing strains to scientists around the world.

Primary phenotyping of a subset of Cre drivers included in this study demonstrates the potential scope of "endogenous" phenotypes in transgenes in common use. While the impact of insertional mutagenesis is clear, for most KO alleles only homozygous mutants are carefully phenotyped, and thus potential confounding heterozygous phenotypes are unclear. Moreover, some transgenes delete multiple genes, and the combinatorial effect on phenotype would require independent evaluation. Finally, the transgene-specific caveats of passengers (minigenes, genes on BACs, *E. coli* genome, etc.) require specific testing. As noted above, the use of proper controls can mitigate most concerns arising from insertional mutagenesis, passenger cassettes, or transgene toxicity, assuming it does not directly impact the phenotype of interest. However, with the emergence of high-throughput phenotyping pipelines (Dickinson et al. 2016; Karp et al. 2017; Meehan et al. 2017), it is now feasible to broadly characterize the phenotypes of a larger collection of transgenic tool lines, perhaps in parallel to insertion site discovery. The high efficiency of insertion site identification enabled by TLA suggests the possibility that random transgenesis, and its consequence, could be used as a tool for discovery. Indeed, despite the challenges, many transgene insertion discovery efforts were inspired by the presence of unexpected phenotypes in these lines. The frequent complex structural variation and impact on multiple genes and noncoding features provide an alternative to single-gene targeted mutagenesis efforts such as the IMPC, while the presence of the transgene simplifies discovery versus spontaneous mutants with similar complex alleles. Moreover, insertion sites that do not damage gene features but support tissue-specific expression could be useful loci to establish as "safe harbors" for targeted transgenesis.

Given our findings, and the potential caveats implied for the use of transgenes, it is tempting to suggest that the community should move away from making, and ultimately using, lines generated by random transgenesis. For Cre lines, knock-in alleles targeting the endogenous locus of a desired driver gene have the added potential advantage of providing greater specificity, desirable given the high rate of off-target activity seen in many transgenic lines (Heffner et al. 2012). To this end, the EUCOOMTOOLs program has produced hundreds of new Cre driver lines using this strategy (Murray et al. 2012; Rosen et al. 2015). However, this typically comes at the cost of haploinsufficiency at the driver locus, often a gene that is part of a pathway critical to the development of the cell type or tissue in question. Expression levels in a knock-in might be lower than that of a multicopy transgene, thus sacrificing effectiveness for specificity. The use of neutral locus docking sites or targeted transgenesis facilitated by CRISPR can avoid the mutagenic risk associated with random insertion, but the former is typically a single-copy event and the latter is relatively untested. Thus, while alternatives to random transgenesis exist, they come with their own caveats and do not necessarily provide a suitable alternative. Rather, given the impact of discoveries enabled by transgenic lines, knowledge of the transgenic insertion site is best viewed as one of many critical pieces of information that should be considered in an experimental design.

Methods

Mice

All strains used for the TLA analysis were obtained from the Jackson Laboratory Repository, four of which are distributed from the JAX Mouse Mutant Research and Resource Center (MMRRC). The specific mouse strains and JAX Stock # (and MMRRC Stock # if applicable) are available in Supplemental Table S1. All procedures and protocols (see "Phenotyping" section below) were approved by the Jackson Laboratory Animal Care and Use Committee and were conducted in compliance with the National Institutes of Health Guideline for Care and Use of Laboratory Animals.

TLA: isolation of splenocytes

The use of live cells provides greater sensitivity and results in a higher success rate than archived DNA samples. Splenocytes were isolated from each line as previously described (de Vree et al. 2014).

In brief, the spleens were dissected and stored on ice. A single-cell suspension was made using a 40-µm mesh filter suspending the cells in 10% fetal calf serum (FCS)/ phosphate buffered saline (PBS). Following centrifugation at 4°C at 500*g* for 5 min, the supernatant was discarded, and the pellet was dissolved in 1 mL 1× Pharm Lyse (BD Biosciences) and incubated at room temperature for 3 min to lyse splenic erythrocytes. To terminate the lysis reaction, 0.5 mL phosphate buffered saline was added, followed by centrifugation at 4°C, 500*g* for 5 min. The supernatant was discarded and the pellet resuspended in 0.5 mL PBS. After one final centrifugation step for 2 min, the supernatant was discarded and the cell pellet resuspended in 1 mL freeze medium (PBS with 10% dimethyl sulfoxide and 10% fetal calf serum). The samples were stored at −80°C until shipment for TLA processing.

TLA: amplification and isolation of linked chromatin

Targeted locus amplification was performed as previously described (de Vree et al. 2014). In brief, spleen cells were crosslinked using formaldehyde, after which the DNA was digested using the restriction enzyme NlaIII (CATG recognition sequence). Subsequently, the sample was ligated, crosslinks were reversed, and the DNA was purified. To obtain circular chimeric DNA molecules for PCR amplification, the DNA molecules were trimmed with NspI and ligated at a DNA concentration of 5 ng/µL to promote intramolecular ligation. NspI has a RCATGY recognition sequence that encompasses the CATG recognition sequence of NlaIII, which ensures only a subset of NlaIII sites were (re-)digested, generating DNA fragments of ~2 kb and allowing the amplification of entire restriction fragments. After ligation, the DNA was purified, and eight 25-µL PCR reactions, each containing 100 ng template, were pooled for sequencing. Sequences of the inverse primers, which were designed using Primer3 software (Untergasser et al. 2012), can be found in Supplemental Table S4.

TLA: sequencing, mapping, and sequence alignment

The primer sets were used in individual TLA amplifications. PCR products were purified and library-prepped using the Illumina Nextera XT protocol and sequenced on an Illumina MiSeq sequencer.

Reads were mapped using BWA-SW, which is a Smith-Waterman alignment tool (Li and Durbin 2010). This allows partial mapping, which is optimally suited for identifying break-spanning reads. The mouse mm10, rat rn5, cow bosTau8, SV40 GCF_000837645.1, rabbit oryCun2, chicken galGal4, and human genome version hg19 were used for mapping. Changes between the genome version hg19 and GRCh38 mainly concern an update on high level chromosome assembly and the addition of 'haplotype alleles.' For the projects we typically perform, we amplify and analyze 'local' DNA sequences, which in general are not affected by this update. Within the aligned data sets, regions were sorted on coverage height to sort out regions with highest coverage, i.e., regions containing sequences that were part of the transgene construct. A visual inspection was performed to discriminate true peaks from background, which yielded fusion fragment(s). Based on the fusion reads present on the outer ends of these fusion fragments, a reconstruction was made of the original transgene construct.

Sequence validation

For all lines where breakpoint-spanning reads were available, TG integration fusions were confirmed by PCR amplification and sequence analysis. The extended reads were analyzed for GC content using ENDMEMO software (http://www.endmemo.com/bio/gc.php), and PCR primers were designed using Primer3 software (Untergasser et al. 2012) to optimize for size and GC content. If the fusion product was larger than 900 bp, the fusion site was confirmed using at least two sets of primers for the long read as well as an internal read to insure adequate coverage of the integration site. PCR amplicons with suitable products were purified and Sanger sequenced. For Vil-cre, because our fusion fragment did not contain sufficient transgene sequence to design a validation primer, we used the primer described in a prior report (Cain-Hom et al. 2017) to confirm the breakpoint of this line.

QPCR analysis

Genomic DNA isolated from tail biopsies was used to analyze loss-of-allele (LOA) or relative concentration qPCR on Applied Biosystem's ViiA 7 (Applied Biosystems).

LOA assay design: The premise behind the LOA assay assumes a one-copy difference between a transgene insertion and the wild-type (WT) sample, while a gain of allele can be used to show a duplication of the genomic target region. Based on transgene integration site sequences and resultant deletions or duplications, target genomic region q-PCR 5′ nuclease assays were designed using PrimerQuest software (Integrated DNA Technologies). The internal reference control *Apob* probe contains a VIC (4,7,2′-trichloro-7′-phenyl-6-carboxyfluorescein) reporter dye (ABI, Applied Biosystems) while all experimental assays use FAM (6-carboxyfluorescein)-labeled probes for detection. An NFQ-MGB (*Apob*) dark or Zen/Iowa Black FQ quencher (IDT) is used for all assays. Primer and probe sequences are provided in Supplemental Table S3.

QPCR samples were then analyzed in triplicate and Cq values for the samples and the internal reference (*Apob*) were calculated using Viia7 software (QuantStudio Software V1.3, ABI, Applied Biosystems). The means of the Cq values were used to calculate ΔCq values, and these were then used to calculate relative copy number of the recombinant region using the $2^{-\Delta\Delta Cq}$ formula (Livak and Schmittgen 2001).

Phenotyping

We employed a modified version of the IMPReSS pipeline (www.mousephenotype.org/impress) for high-throughput clinical phenotyping assessment, which was developed under the IMPC program (de Angelis et al. 2015; Dickinson et al. 2016; Karp et al. 2017; Meehan et al. 2017). The following seven lines (and genotypes) were characterized: B6.Cg-Tg(Alb-cre)21Mgn/J (HOM), B6.Cg-Tg(Ins2-cre)25Mgn/J (HOM), B6.Cg-Tg(Lck-cre)548Jxm/J (HOM), B6.Cg-Tg(Nes-cre)1Kln/J (HEMI), B6.Cg-Tg(Vav1-icre)

A2Kio/J (HEMI), B6.Cg-Tg(Vil1-cre)997Gum/J (HEMI), and B6.Cg-Tg(Wnt1-cre)11Rth Tg(Wnt1-GAL4)11Rth/J (HEMI). Control mice are from a pool consisting of C57BL/6J WT mice and noncarrier (NCAR) controls from the colony for lines maintained in a HEMI × NCAR breeding scheme [Tg(Wnt1-cre)11Rth, Tg(Vav1-icre)A2Kio, Tg(Vil1-cre)997Gum, and Tg(Nes-cre)1Kln]. For each mouse strain, eight male and eight female transgenic animals, NCAR controls, or C57BL/6J mice were processed through the JAX Adult Phenotyping Pipeline. Full details of the JAX Adult Phenotyping Pipeline can be found on the IMPC website (www.mousephenotype.org/impress/procedures/12). Briefly, mice were received into the pipeline at 4 wk of age, body weight was collected weekly, and assays were performed weekly from 8 to 18 wk of age, ordered such that the least invasive, behavioral testing was performed first. The specific assays and age in weeks that the assay was performed in this study are as follows:

Open Field (OFA) (8 wk)
Light-Dark Transition (LD), Holeboard (HB) (9 wk)
Acoustic Startle/Pre-pulse Inhibition (PPI) (10 wk)
Tail Suspension, Electrocardiogram, Rotarod (11 wk)
Body Composition (BC), (14 wk)
Piezoelectric Sleep/Wake (SLEEP) (15 wk)
Auditory Brainstem Response (ABR) (4M + 4F) (16 wk)
Electroconvulsive Seizure Threshold (ECT) (17 wk)
Terminal collection including Hematology (HEM), Clinical Biochemistry (CBC)

JAX-specific sleep test

Sleep and wake states were determined using the PiezoSleep System (Flores et al. 2007; Donohue et al. 2008; Mang et al. 2014). The system is comprised of plexiglass cages lined with piezoelectric films across the cage floor that detect pressure variations. Signal features sensitive to the differences between the sleep and wake states are extracted from 8-sec pressure signal segments, and classification is automatically performed every 2 sec using overlapping windows. From this, the following parameters are calculated: sleep bout lengths (light phase, dark phase, 24-h mean), breathing rate, breathing rate during sleep, percentage daily sleep (light and dark phase), and diurnal wake ratio.

Statistical analysis

Linear mixed models (LMM) were performed to identify phenotypic associations from high-throughput phenotyping experiments. Sex, weight, and batch were a significant source of variation for continuous phenotypes. In the linear mixed model, explanatory factors including sex, weight, and mutant genotype were treated as fixed effects, while batch (date of test) was treated as a random effect adding variation to the data (see Equation 1):

$$\text{Variable} = \text{Genotype} + \text{Sex} + \text{Weight} + (1|\text{Batch}). \quad (1)$$

Parameters from the mixed model are estimated using the method of restricted maximum likelihood (REML). Adjusted *P*-values were calculated from nominal *P*-values in mixed models to control for false discovery rate (FDR). All data analysis was performed using R (R Core Team 2016).

Data access

All sequence data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP156273. Phenotypic data have been submitted to the Mouse Phenome Database (https://phenome.jax.org) under the search term JaxKOMP-Cre.

Competing interest statement

E.S. and M.v.M. are employees at Cergentis B.V.

Acknowledgments

This work was supported in part by the Office of the Director, National Institutes of Health and the National Human Genome Research Institute of the National Institutes of Health under award numbers R24OD011190, U42OD011185, UM1OD023222, U42OD010921, and HG006332. The authors thank Brianna Caddle and Larry Bechtel for their technical assistance in isolating spleen cells and Kevin Peterson for his helpful and thoughtful comments on the manuscript.

References

Bersell K, Choudhury S, Mollova M, Polizzotti BD, Ganapathy B, Walsh S, Wadugu B, Arab S, Kuhn B. 2013. Moderate and high amounts of tamoxifen in *αMHC-MerCreMer* mice induce a DNA damage response, leading to heart failure and death. *Dis Model Mech* **6:** 1459–1469. doi:10.1242/dmm.010447

Cain-Hom C, Splinter E, van Min M, Simonis M, van de Heijning M, Martinez M, Asghari V, Cox JC, Warming S. 2017. Efficient mapping of transgene integration sites and local structural changes in Cre transgenic mice using targeted locus amplification. *Nucleic Acids Res* **45:** e62. doi:10.1093/nar/gkw1329

Chen ZY, He CY, Meuse L, Kay MA. 2004. Silencing of episomal transgene expression by plasmid bacterial DNA elements *in vivo*. *Gene Ther* **11:** 856–864. doi:10.1038/sj.gt.3302231

de Angelis MH, Nicholson G, Selloum M, White J, Morgan H, Ramirez-Solis R, Sorg T, Wells S, Fuchs H, Fray M, et al. 2015. Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. *Nat Genet* **47:** 969–978. doi:10.1038/ng.3360

de Vree PJ, de Wit E, Yilmaz M, van de Heijning M, Klous P, Verstegen MJ, Wan Y, Teunissen H, Krijger PH, Geeven G, et al. 2014. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat Biotechnol* **32:** 1019–1025. doi:10.1038/nbt.2959

Declercq J, Brouwers B, Pruniau VP, Stijnen P, de Faudeur G, Tuand K, Meulemans S, Serneels L, Schraenen A, Schuit F, et al. 2015. Metabolic and behavioural phenotypes in Nestin-Cre mice are caused by hypothalamic expression of human growth hormone. *PLoS One* **10:** e0135502. doi:10.1371/journal.pone.0135502

Dennis JF, Kurosaka H, Iulianella A, Pace J, Thomas N, Beckham S, Williams T, Trainor PA. 2012. Mutations in *Hedgehog Acyltransferase* (Hhat) perturb Hedgehog signaling, resulting in severe acrania-holoprosencephaly-agnathia craniofacial defects. *PLoS Genet* **8:** e1002927. doi:10.1371/journal.pgen.1002927

Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, Meehan TF, Weninger WJ, Westerberg H, Adissu H, et al. 2016. High-throughput discovery of novel developmental phenotypes. *Nature* **537:** 508–514. doi:10.1038/nature19356

Donohue KD, Medonza DC, Crane ER, O'Hara BF. 2008. Assessment of a non-invasive high-throughput classifier for behaviours associated with sleep and wake in mice. *Biomed Eng Online* **7:** 14. doi:10.1186/1475-925X-7-14

Dubose AJ, Lichtenstein ST, Narisu N, Bonnycastle LL, Swift AJ, Chines PS, Collins FS. 2013. Use of microarray hybrid capture and next-generation sequencing to identify the anatomy of a transgene. *Nucleic Acids Res* **41:** e70. doi:10.1093/nar/gks1463

Flores AE, Flores JE, Deshpande H, Picazo JA, Xie XS, Franken P, Heller HC, Grahn DA, O'Hara BF. 2007. Pattern recognition of sleep in rodents using piezoelectric signals generated by gross body movements. *IEEE Trans Biomed Eng* **54:** 225–233. doi:10.1109/TBME.2006.886938

Frendewey D, Chernomorsky R, Esau L, Om J, Xue Y, Murphy AJ, Yancopoulos GD, Valenzuela DM. 2010. The loss-of-allele assay for ES cell screening and mouse genotyping. *Methods Enzymol* **476:** 295–307. doi:10.1016/S0076-6879(10)76017-1

Galichet C, Lovell-Badge R, Rizzoti K. 2010. Nestin-Cre mice are affected by hypopituitarism, which is not due to significant activity of the transgene in the pituitary gland. *PLoS One* **5:** e11443. doi:10.1371/journal.pone.0011443

Giusti SA, Vercelli CA, Vogl AM, Kolarz AW, Pino NS, Deussing JM, Refojo D. 2014. Behavioral phenotyping of Nestin-Cre mice: implications for

genetic mouse models of psychiatric disorders. *J Psychiatr Res* **55:** 87–95. doi:10.1016/j.jpsychires.2014.04.002

Gordon JW, Ruddle FH. 1981. Integration and stable germ line transmission of genes injected into mouse pronuclei. *Science* **214:** 1244–1246. doi:10.1126/science.6272397

Heffner CS, Herbert Pratt C, Babiuk RP, Sharma Y, Rockwood SF, Donahue LR, Eppig JT, Murray SA. 2012. Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource. *Nat Commun* **3:** 1218. doi:10.1038/ncomms2186

Hottentot QP, van Min M, Splinter E, White SJ. 2017. Targeted locus amplification and next-generation sequencing. *Methods Mol Biol* **1492:** 185–196. doi:10.1007/978-1-4939-6442-0_13

Karp NA, Mason J, Beaudet AL, Benjamini Y, Bower L, Braun RE, Brown SDM, Chesler EJ, Dickinson ME, Flenniken AM, et al. 2017. Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat Commun* **8:** 15475. doi:10.1038/ncomms15475

Lee JY, Ristow M, Lin X, White MF, Magnuson MA, Hennighausen L. 2006. RIP-Cre revisited, evidence for impairments of pancreatic β-cell function. *J Biol Chem* **281:** 2649–2653. doi:10.1074/jbc.M512373200

Lewis AE, Vasudevan HN, O'Neill AK, Soriano P, Bush JO. 2013. The widely used *Wnt1-Cre* transgene causes developmental phenotypes by ectopic activation of Wnt signaling. *Dev Biol* **379:** 229–234. doi:10.1016/j.ydbio.2013.04.026

Lexow J, Poggioli T, Sarathchandra P, Santini MP, Rosenthal N. 2013. Cardiac fibrosis in mice expressing an inducible myocardial-specific Cre driver. *Dis Model Mech* **6:** 1470–1476. doi:10.1242/dmm.010470

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26:** 589–595. doi:10.1093/bioinformatics/btp698

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25:** 402–408. doi:10.1006/meth.2001.1262

Loonstra A, Vooijs M, Beverloo HB, Allak BA, van Drunen E, Kanaar R, Berns A, Jonkers J. 2001. Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc Natl Acad Sci* **98:** 9209–9214. doi:10.1073/pnas.161269798

Madison BB, Dunbar L, Qiao XT, Braunstein K, Braunstein E, Gumucio DL. 2002. *cis* elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J Biol Chem* **277:** 33275–33283. doi:10.1074/jbc.M204935200

Mang GM, Nicod J, Emmenegger Y, Donohue KD, O'Hara BF, Franken P. 2014. Evaluation of a piezoelectric system as an alternative to electroencephalogram/electromyogram recordings in mouse sleep studies. *Sleep* **37:** 1383–1392. doi:10.5665/sleep.3936

Meehan TF, Conte N, West DB, Jacobsen JO, Mason J, Warren J, Chen CK, Tudose I, Relac M, Matthews P, et al. 2017. Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat Genet* **49:** 1231–1238. doi:10.1038/ng.3901

Murray SA, Eppig JT, Smedley D, Simpson EM, Rosenthal N. 2012. Beyond knockouts: cre resources for conditional mutagenesis. *Mamm Genome* **23:** 587–599. doi:10.1007/s00335-012-9430-2

Naiche LA, Papaioannou VE. 2007. Cre activity causes widespread apoptosis and lethal anemia during embryonic development. *Genesis* **45:** 768–775. doi:10.1002/dvg.20353

Nakanishi T, Kuroiwa A, Yamada S, Isotani A, Yamashita A, Tairaka A, Hayashi T, Takagi T, Ikawa M, Matsuda Y, et al. 2002. FISH analysis of 142 EGFP transgene integration sites into the mouse genome. *Genomics* **80:** 564–574. doi:10.1006/geno.2002.7008

Orban PC, Chui D, Marth JD. 1992. Tissue- and site-specific DNA recombination in transgenic mice. *Proc Natl Acad Sci* **89:** 6861–6865. doi:10.1073/pnas.89.15.6861

Padmanabhan N, Jia D, Geary-Joo C, Wu X, Ferguson-Smith AC, Fung E, Bieda MC, Snyder FF, Gravel RA, Cross JC, et al. 2013. Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development. *Cell* **155:** 81–93. doi:10.1016/j.cell.2013.09.002

Pease S, Saunders TL, International Society for Transgenic Technologies. 2011. *Advanced protocols for animal transgenesis: an ISTT manual*. Springer, Berlin.

R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rosen B, Schick J, Wurst W. 2015. Beyond knockouts: The International Knockout Mouse Consortium delivers modular and evolving tools for investigating mammalian genes. *Mamm Genome* **26:** 456–466. doi:10.1007/s00335-015-9598-3

Sauer B, Henderson N. 1990. Targeted insertion of exogenous DNA into the eukaryotic genome by the Cre recombinase. *New Biol* **2:** 441–449.

Scrable H, Stambrook PJ. 1997. Activation of the *lac* repressor in the transgenic mouse. *Genetics* **147:** 297–304.

Srivastava A, Philip VM, Greenstein I, Rowe LB, Barter M, Lutz C, Reinholdt LG. 2014. Discovery of transgene insertion sites by high throughput sequencing of mate pair libraries. *BMC Genomics* **15:** 367. doi:10.1186/1471-2164-15-367

Tosh JL, Rickman M, Rhymes E, Norona FE, Clayton E, Mucke L, Isaacs AM, Fisher EMC, Wiseman FK. 2017. The integration site of the *APP* transgene in the J20 mouse model of Alzheimer's disease. *Wellcome Open Res* **2:** 84. doi:10.12688/wellcomeopenres.12237.1

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Res* **40:** e115. doi:10.1093/nar/gks596

Valenzuela DM, Murphy AJ, Frendewey D, Gale NW, Economides AN, Auerbach W, Poueymirou WT, Adams NC, Rojas J, Yasenchak J, et al. 2003. High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat Biotechnol* **21:** 652–659. doi:10.1038/nbt822

White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ, Podrini C, et al. 2013. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154:** 452–464. doi:10.1016/j.cell.2013.06.022

Received December 18, 2017; accepted in revised form January 14, 2019.