bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 **Large-scale discovery of mouse transgenic integration sites reveals frequent structural variation**
2 **and insertional mutagenesis**
3
4
5
6 Leslie O. Goodwin[1], Erik Splinter[2], Tiffany L. Davis[1], Rachel Urban[1], Hao He[3], Robert E. Braun[1], Elissa J.
7 Chesler[1], Vivek Kumar[1], Max van Min[2], Juliet Ndukum[1], Vivek M. Philip[1], Laura G. Reinholdt[1], Karen
8 Svenson[1], Jacqueline K. White[1], Michael Sasner[1], Cathleen Lutz[1], and Stephen A. Murray[1]*
9
10
11 1. The Jackson Laboratory, Bar Harbor, ME, USA
12 2. Cergentis, B.V., Utrecht, The Netherlands
13 3. The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
14
15
16 * Correspondence:
17 steve.murray@jax.org
18
19
20
21
22
23
24
25
26

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ABSTRACT

Transgenesis has been a mainstay of mouse genetics for over 30 years, 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 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.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 **INTRODUCTION**

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

17 Of the 8,012 transgenic alleles published in the Mouse Genome Database, only 416 (5.2%) have
18 an annotated chromosomal location. For transgenic *cre* alleles, the number is even lower, with a known
19 chromosomal location for 36/1,631 (2.3%) lines, highlighting the challenge of identification of integration
20 sites despite widespread acknowledgement that such information is useful and important. Low resolution
21 mapping of transgenes can be achieved through FISH or linkage mapping, but these approaches offer
22 little information about potential mutagenesis at the integration site. Inverse PCR can be used to clone the
23 actual fusion sequence, but has a high failure rate owing to the multi-copy nature of most transgenes.
24 More recently, high-throughput sequencing (HTS) has been employed to identify transgene insertion sites
25 (Dubose et al. 2013), with improvements offered by the use of mate pair libraries(Srivastava et al. 2014).

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 Despite the promise, HTS-based approaches have not seen widespread implementation, possibly due to
2 the cost and/or complexity of the analysis.
3 The identification of transgene insertion sites is useful for a number of reasons. First, it allows the
4 user to avoid experimental designs that attempt to combine linked alleles (e.g. a conditional allele with a
5 cre transgene), obviating a long and possibly fruitless breeding exercise. Second, it enables the design of
6 allele-specific genotyping assays, which assist in colony management and determination of zygosity.
7 Finally, it alerts the investigator to potential confounding effects of insertional mutagenesis through the
8 direct disruption of the coding sequence of endogenous genes, indirect effects on the regulation of nearby
9 genes, or complex structural variations (inversions or duplications) that can accompany the integration
10 event. Cases of insertional mutagenesis with dramatic phenotypic consequences have been reported. For
11 example, the Tg(TFAP2A-cre)1Will allele inserted into the *Hhat* gene, disrupting its function, leading to
12 a variety of severe developmental abnormalities in homozygous embryos including holoprosencephaly
13 with acrania and agnathia, reflecting a disruption of the hedgehog signaling pathway (Dennis et al. 2012).
14 Given the utility of this line in targeting branchial arches of the developing face, this could confound the
15 interpretation of experimental data if the correct breeding scheme and controls are not included. Because
16 so few insertion sites have been mapped, the scale of this issue is unknown. A prior report using FISH
17 found that transgenes tend to insert into G-positve band regions (Nakanishi et al. 2002), which typically
18 have reduced gene density, but the mapped transgenes were not assessed for expression levels, so it is
19 unclear if these data are representative of transgenes used in the wider scientific community. More
20 recently, Targeted Locus Amplification (TLA) (de Vree et al. 2014; Hottentot et al. 2017) has been
21 employed to identify the insertion site for 7 Cre driver lines(Cain-Hom et al. 2017), only one of which
22 was found to insert into an annotated gene. However, because of the small sample size, it is not clear if
23 this rate of mutagenesis is representative of the genome-wide rates in larger collections representing a
24 variety of transgene types.
25 Here, we describe the first large-scale survey of insertion sites in widely used transgenic mouse
26 lines. The TLA capture process is highly efficient, providing high-resolution localization of the transgene

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 insertion site for all 40 lines tested, including 24 with fusion reads discovered on both ends of the
2 insertion. Remarkably, half of the insertion events disrupt one or more genes. Some of these disrupted
3 genes have known knockout phenotypes, including embryonic lethality. We also identify a number of
4 structural variations, including frequent deletions and duplications, and unexpected elements that have co-
5 integrated with some of the transgenes. Finally, we show that transgenic lines can display phenotypes
6 independent of experimental context, consistent with our insertional mutagenesis discoveries. Together,
7 these data demonstrate the clear need to consider the molecular consequences of transgene insertion in
8 any experimental design.
9
10 **Results**
11 We selected a total of 40 transgenic lines from live colonies in the JAX Repository for our study,
12 including 4 lines distributed through the Mouse Mutant Research and Resource Center (MMRRC) at JAX.
13 All lines are broadly utilized and thus represent important research tools that would benefit from insertion
14 site identification. The list comprises 17 genetic tool strains, including 15 cre drivers, many of which have
15 demonstrated off-target or unexpected excision activity (Heffner et al. 2012). In addition, we included 5
16 lines that lack an allele-specific genotyping assay, and 18 critical Alzheimer's or Parkinson's disease
17 models (Figure 1A). We selected lines that were generated through a variety of means (Figure 1B),
18 including standard small plasmid-based transgenes, human and mouse BAC transgenes, a human PAC, a
19 human cosmid, and a transgene generated through lentiviral-mediated transgenesis. The
20 BAC/PAC/cosmid vectors were included both to capture critical lines of interest and to test the feasibility
21 of the TLA process on these larger constructs. A schematic of the TLA process, depicted in Figure 1C,
22 was performed essentially as described ((de Vree et al. 2014; Hottentot et al. 2017); Materials and
23 Methods) using primers specific for known elements of each transgenic line (Supplementary Table 1).
24 Transgene insertion sites result in high sequencing coverage across the transgene and its insertion site(s),
25 and at least one putative fusion read across the transgene-genome breakpoint was identified for all 40
26 lines (Table 1; Supplementary Table 1). In some cases, follow-up TLA analysis using additional primers

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

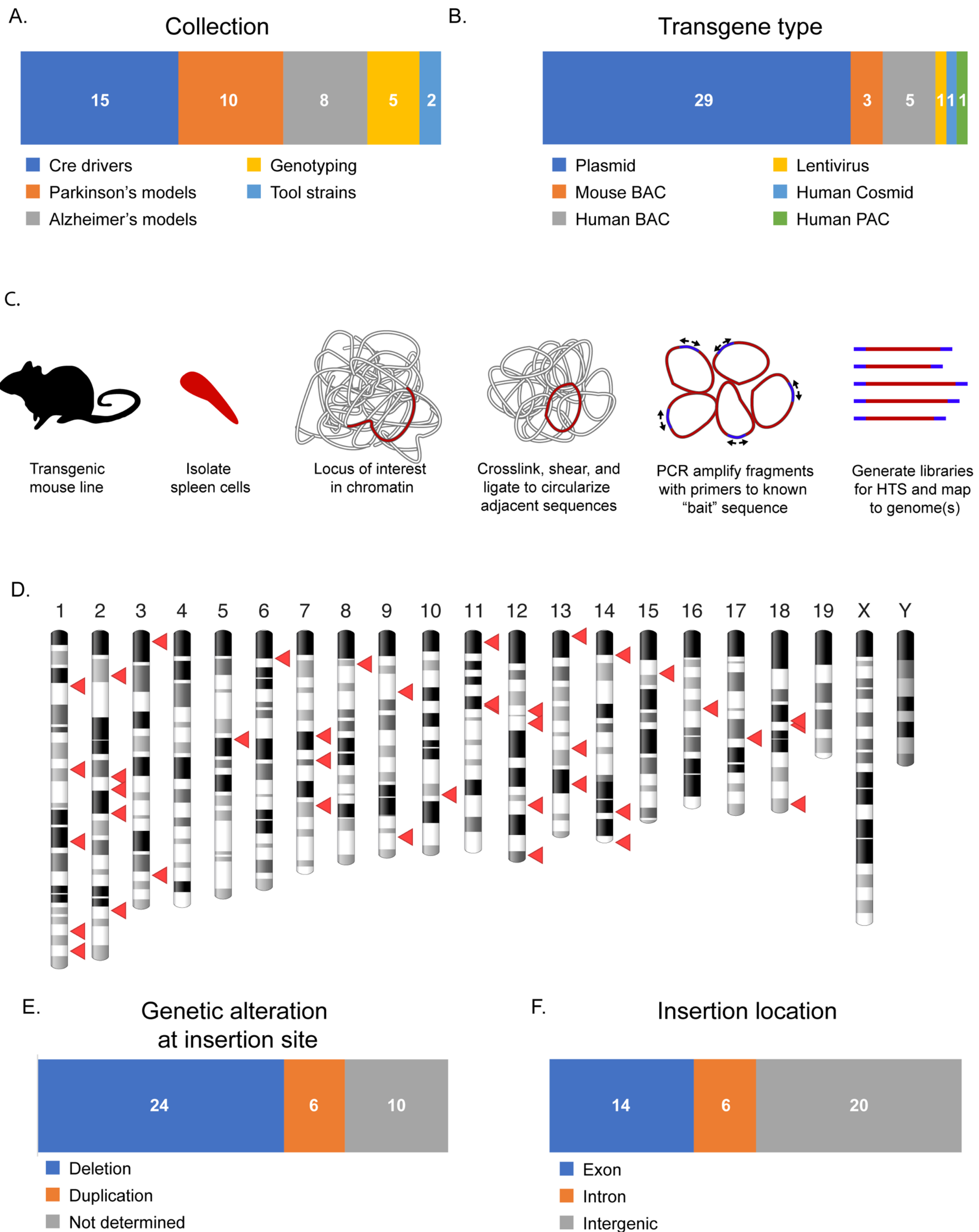

Figure 1. Discovery of the integration loci for 40 transgenic mouse lines. A. Distribution of the categories of transgenes included in this study. B. Distribution of transgenes by molecular type. C. Schematic of the targeted locus amplification process. C. Ideogram showing the physical distribution of 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 non-gene loci (intergenic).

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Allele	JAX STOCK#	Target type	Insertion coordinates	Insertion mutation	Genes affected*
Ins2-cre	3573	Transgene	Chr7:62991157-63008557	Duplication	None
Alb-cre	3574	Transgene	Chr13:3172116-3172120	4bp deletion	*Speer6-ps1*
Nes-cre	3771	Transgene	Chr12:90524592-90524609	17bp deletion	None
Lck-cre	3802	Transgene	Chr11:41490714	Duplication	None
Tek-cre	4128	Transgene	Chr13:68459931-68701276	241 Kb deletion	*Mtrr, Fastkd3, 1700001L19Rik, Adcy2*
Vil1-cre	4586	Transgene	Chr17:55326957-55341510	14.6 Kb deletion	None
Ddx4-cre	6954	Transgene	Chr18:85696612-86794868	1,098 Kb deletion	*Neto1, Cbln2*
UBC-cre/ERT2	8085	Lenti transgenic	Chr2:25249816-25249821	5bp deletion	*Ndor1*
Cspg4-cre	8533	Mouse BAC	Chr1:173692115	ND	*Ifi208* (intron)
Cspg4-cre/Esr1	8538	Mouse BAC	Chr14:106654779-106655407	628bp deletion	None
Th-cre	8601	Transgene	Chr9:33514690-34139124	624 Kb deletion	*7630403G23Rik*
Vav1-icre	8610	Transgene	Chr18:47022629	ND	*Commd10* (intron)
Wnt1-cre	9107	Transgene	Chr11:6425500-6456783	31.2 Kb deletion	*H2afv*
Sox2-cre	14094	Transgene	Chr13:89311726	ND	*Edil3* (intron)
Wnt1-cre(2)	22501	Transgene	Chr2:154561346-154561603	257bp deletion; complex inversion	*E2f1*
Itgax-DTR/EGFP	4509	Transgene	Chr1:80448681-80455738	7,057 bp deletion	*1700016L21Rik*
Camk2a-tTA	7004	Transgene	Chr12:116101154-116609271	508 Kb deletion	*Vipr2, Wdr60, Esyt2, Ncapg2, Ptprn2*
GFAP-APOE_i4	4631	Transgene	Chr15:23364633-23373281	8,648 bp deletion	*Cdh18*
"3XTg-AD"	4807	Transgenes	Chr2:87862466-87862463	3 bp deletion	None
MAPT	5491	Human PAC	Chr7:10447768	ND	None
APPswe,PSEN1dE9	5864	Transgenes	Chr9:113003660	Duplication	None
PDGFB-APPSwInd	6293	Transgene	Chr16:43086322-43127049	40.7 Kb deletion	*Zbtb20* (intron)
Prnp-MAPT*P301S	8169	Transgene	Chr3:140354280-140603283	249 Kb deletion	None
"5XFAD"	8730	Transgenes	Chr3:6297836	ND	None
tetO-MAPT*P301L	15815	Transgene	Chr14:124457842-124702169	244 Kb deletion	*Fgf14*
Prnp-SNCA*A53T (line 83)	4479	Transgene	Chr12:48212716	ND	None
Prnp-SNCA*A53T (line 23)	6823	Transgene	Chr10:95350683-95399000	48.3Kb deletion	*2310039L15Rik*
LRRK2*R1441G	9604	Human BAC	Chr1:32289302-32289738	436 bp deletion	*Khdrbs2* (intron)
Lrrk2*G2019S	12467	Mouse BAC	Chr18:44968085	ND	None
LRRK2	13725	Human BAC	Chr6:16279287-16327995	756 bp deletion	None
Thy1-SNCA (line 12)	16936	Transgene	Chr14:14719103	Duplication	*Slc4a7* (intron)
Thy1-SNCA (line 15)	17682	Transgene	Chr11:40456787-40495044	38.4 Kb deletion	None
SNCA	18442	Human BAC	Chr7:77604164-77605062	898 bp deletion	None
LRRK2*G2019S	18785	Human BAC	Chr1:80896405	ND	None
LRRK2*R1441G(2)	18786	Human BAC	Chr1:121956000-121995855	39.9 Kb deletion	None
HLA-A/H2-D	4191	Transgene	Chr12:41759331-41760601	1,279 bp deletion	*Immp2l*
CAG-FCGRT	4919	Transgene	Chr1:185129377	Duplication	None
FXN*500GAA	8586	Transgene	Chr5:61755638	ND	None
NSG-HLA-A2.1	9617	Transgene	Chr8:18736683-18757058	Duplication	*Mcph1*
hFCGRT	14565	Human Cosmid	Chr2:101081712	ND	None

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

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

designed based on the initial results was required to identify a fusion read at either junction. Insertions were found genome-wide on 17/19 autosomes, with 5 insertions each identified on chromosomes 1 and 2 (Figure 1D). Structural variations accompanying the insertion were identified for a majority of lines (30/40), comprising 24 deletions and 6 duplications (Figure 1E). As some of the fusion contigs were constructed using single reads, we used PCR and Sanger sequencing to verify fusion reads identified by TLA. 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 (Supplementary Table 1 & 2). For deletions where only one fusion read 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 Table 1 & 3). Remarkably, the transgene insertion event in 20 of 40 lines either delete at least one exon of one or more genes (14) or insert into an intron likely affecting its normal transcription (6). 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 (Figure 1E), which varied in size from a few basepairs to 1.1Mb in the case of the Ddx4-cre line (Figure 2A). As noted above, amongst the 24 deletions, we identified a surprisingly high rate of insertional mutagenesis, either deleting or disrupting between 1 and 5 mouse genes (Figure 2B), each with potential phenotypic consequences. In addition, two genes are disrupted through duplication event that accompanied their respective insertion (Supplementary Table 1). Of the total 28 genes disrupted, 15 have a previously reported knockout (KO) phenotype, including 4 with embryonic lethal phenotypes (Figure 2C), and multiple genes are deleted in 3 lines, highlighting the potentially confounding effect of the insertion event. For example, in the Tek-cre line we identified a 241Kb deletion on Chromosome 13 that includes four protein coding genes (*Mtrr, Fastkd3, 1700001L19Rik, Adcy2*) (Figure 2 D,E). We validated both breakpoints and confirmed the

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

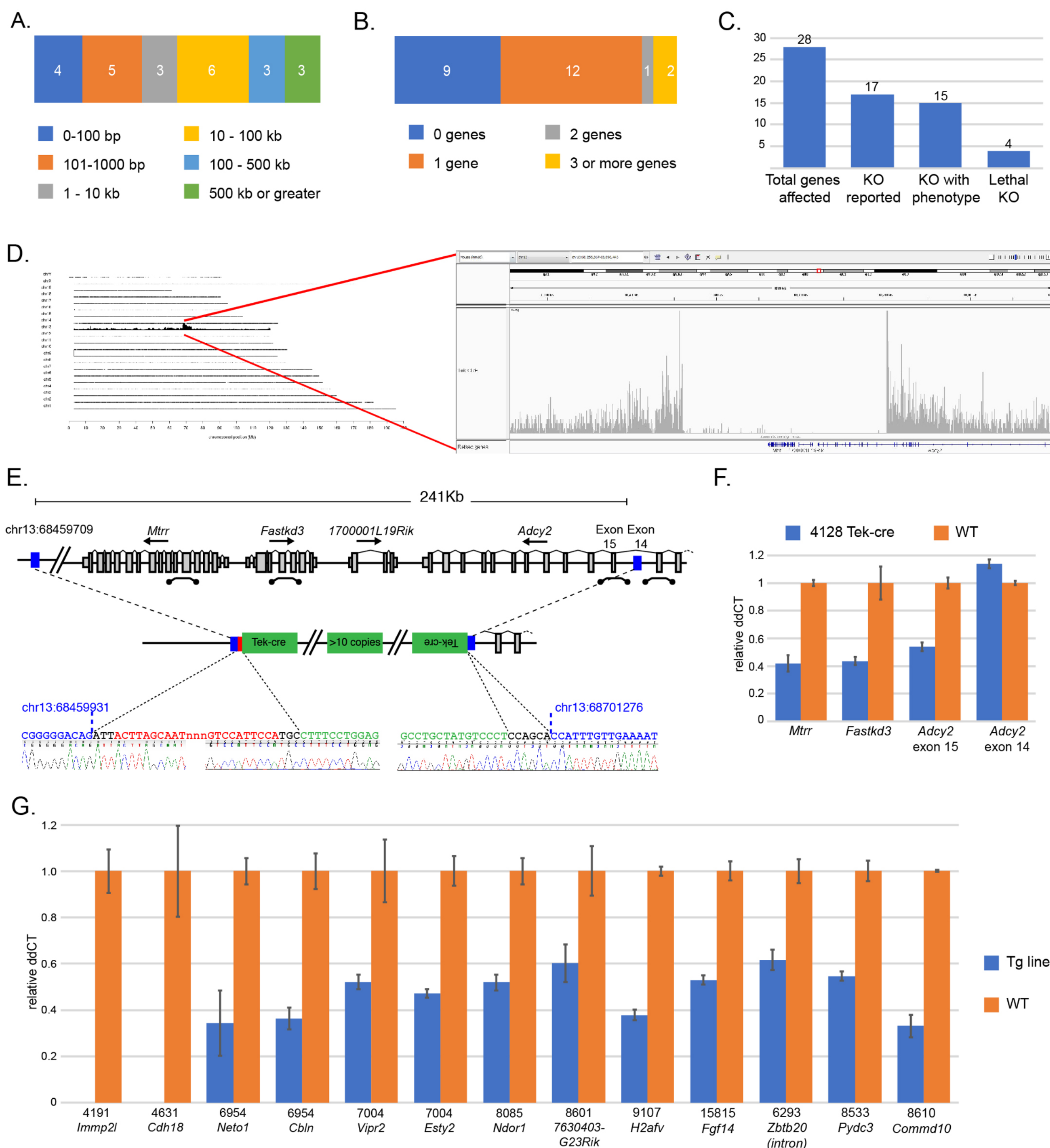

Figure 2. Deletions accompanying transgenic insertions. A. Profile of sizes of deletions identified at integration loci. B. For integrations that occur in genes, the profile of the number of genes affected by the insertion event. C. Illustration of the potential impact of transgene insertions into genes, 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. Schematic of the insertion locus in the Tek-cre 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 reads. Locations of qPCR probes to confirm copy number are shown. E. Results of loss of native allele (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. E. 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.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 deletion of *Mtrr*, *Fastkd3*, and *Adcy2* by loss-of-native-allele qPCR assays, showing clear loss of one
2 copy for all three genes, and in addition confirming the breakpoint between exons 14 and 15 in *Adcy2*
3 (Figure 2F). In addition to the frequent widespread (off-target) activity seen in this line (Heffner et al.
4 2012), prior reports show that an *Mtrr* gene trap allele exhibits trangenerational epigenetic effects leading
5 to severe developmental abnormalities when breeding from a female carrier (Padmanabhan et al. 2013).
6 Given the common use of this line to analyze vascular development, the transgene itself could confound
7 analysis depending on the breeding scheme, highlighting the need for proper controls (i.e. Cre-only) in
8 studies using this transgene. Finally, we confirmed the deletion of an additional 13 genes (Figure 2G,
9 Supplementary Table 1), including two (*Immp2l* and *Cdh18*) that appear null due to the maintenance of
10 the line as a homozygote. Interestingly, one line (PDGFB-APPSwind) showed insertion into the gene
11 *Zbtb20*, but a recent report clearly shows that despite the insertion of more than 10 copies of the transgene,
12 expression of the protein in heterozygous transgenic mice is comparable to WT, suggesting other
13 regulatory mechanisms to maintain a uniform level of expression (Tosh et al. 2017). Therefore, in some
14 cases the consequences of intron insertion require independent validation. Together, these data show that
15 transgene insertions are often associated with large mutagenic deletions, affecting one or more genes,
16 potentially confounding interpretation of results unless the proper control strategies are employed.
17
18 TLA analysis also revealed additional structural variations around the insertion site, including 6
19 instances of duplications and one inversion accompanied by a large deletion. In many duplication cases,
20 fusion reads were only identified on one end of the transgene insertion, so the exact extent of the
21 duplication could only be estimated through read depth. However, we were able to confirm additional
22 copies of parts of the genes *Mcph1* and *Slc4a7* (data not shown), although it is not clear how this might
23 affect gene function. For the Wnt1-Cre(2) line(Lewis et al. 2013), we observed a complex structural
24 variation on chromosome 2, involving a large 45Kb inverted segment inserted into exon 5 of the *E2f1*
25 gene (Figure 3A). The inversion itself contains all of exon 5, but deletes 23Kb including exons 6 and 7 of
26 *E2f1* proximal to the transgene insertion location, all of *Necab3, 1700003F1Rik,* and a portion of the 3'

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

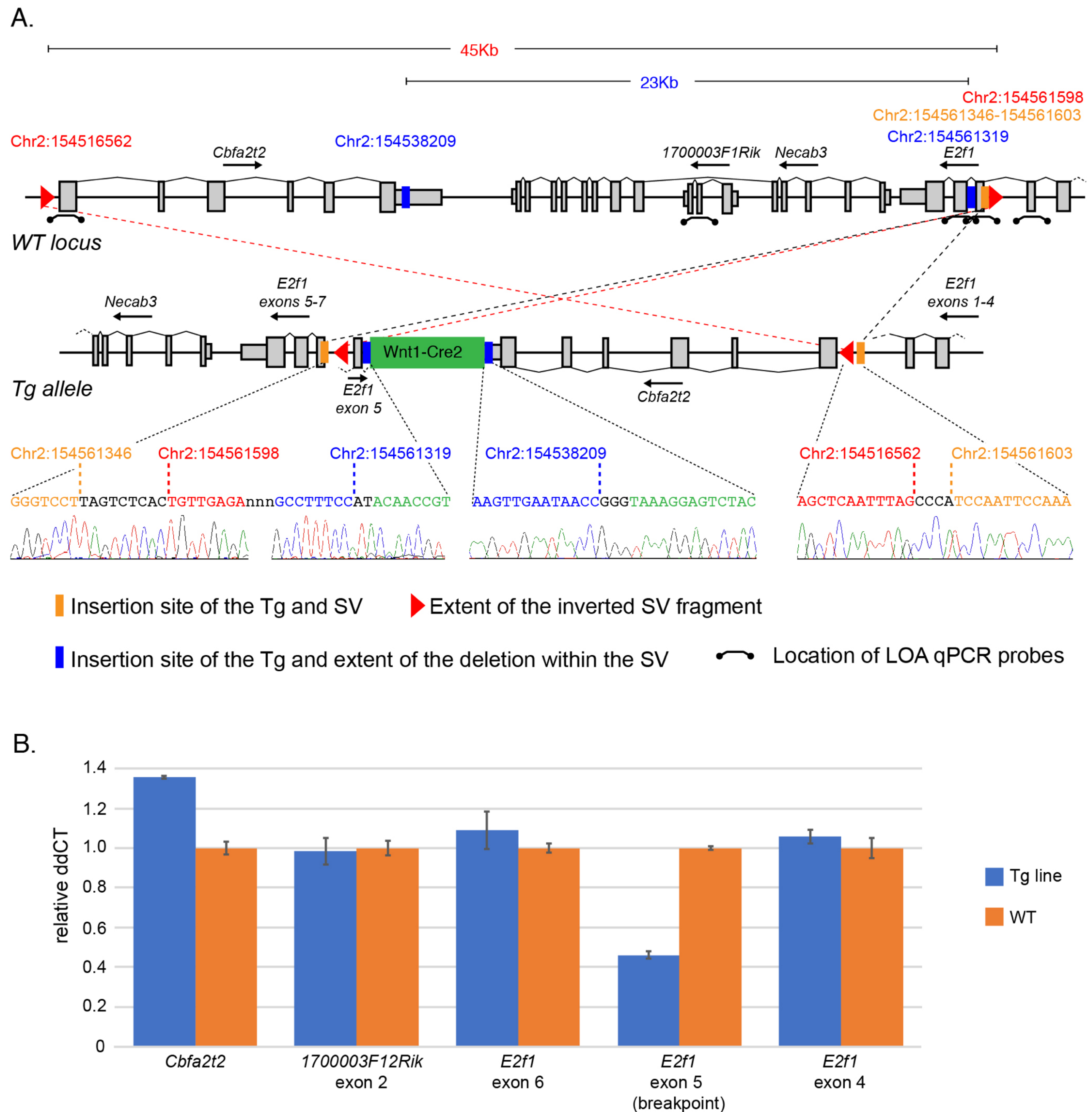

Figure 3. Complex structural variations accompanying transgenic insertions. A. Schematic of the structural variation accompanying the Wnt1-cre2 transgene insertion. The locus includes a large duplication with a partial deletion that accompanies the transgene insertion. The entire duplicated interval 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 read 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.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 UTR of *Cbfa2t2*. As a result, the structural variation disrupts the *E2f1* gene, with the concomitant
2 duplication of exon 5 in the opposite orientation. We used an LOA assay to confirm the disruption of
3 exon 5 (Figure 3B), but does not capture the duplication of the inverted exon, as the inverted fragment is
4 smaller than the amplicon of the qPCR probe. The copy number of *E2f1* exons 4 and 6 are unaffected, as
5 they surround the structural variation. In addition, an LOA assay shows the duplication of exon 6 of
6 *Cbfa2t2*, and an LOA for exon 2 of *1700003F12Rik*, which resides in the deleted portion of the duplicated
7 fragment, shows the normal two copies as expected (Figure 3B). Together, these data illustrate the
8 potential complex structural variations that can occur with transgene integration.
9
10 Because TLA isolates all DNA fragments in close proximity to the transgene integration site, it is
11 possible to identify components of the transgene itself, in addition to the surrounding mouse sequence.
12 The only limitation is the selection of reference genomes for mapping. In this study, we typically mapped
13 to genomes predicted to be part of the transgene, based on the published description of the transgene
14 construction. While for the most part we were able to identify construct elements described in the original
15 publications, unexpected components were seen in several transgenic lines. For example, we found that
16 an entire hGH minigene, described as a polyA sequence, was present in four lines (Ins2-cre, Alb-cre, Nes-
17 cre, and Lck-cre; Figure 4A), as previously reported for the Nes-cre line (Declercq et al. 2015). Of note,
18 publications for these lines reference a vector originally described in (Orban et al. 1992), which clearly
19 describes the minigene structure of the cassette. Similarly, TLA and PCR validation of the Vil-cre line
20 reveals the presence of the entire *Mt1* gene sequence (Figure 4B), despite its description as a
21 "metallothionein poly(A) signal" in the original publication (Madison et al. 2002). The source plasmid
22 does indeed describe it as containing the polyA and several introns (Sauer and Henderson 1990).
23 Although the impact of the presence of the *Mt1* minigene is unclear, there is evidence that the Nes-cre
24 hGH minigene is expressed and that this expression is responsible for some of the metabolic phenotypes
25 observed in the Nes-cre line (Galichet et al. 2010; Giusti et al. 2014). These data indicate that TLA has

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 *E.coli* genome, and not vector sequences, suggesting that fragments of contaminating *E.coli* DNA were co-integrating 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 co-integration in 10/40 strains, with total composition

A. 3573: Ins2-cre

B. 4586: Vil-cre

C.

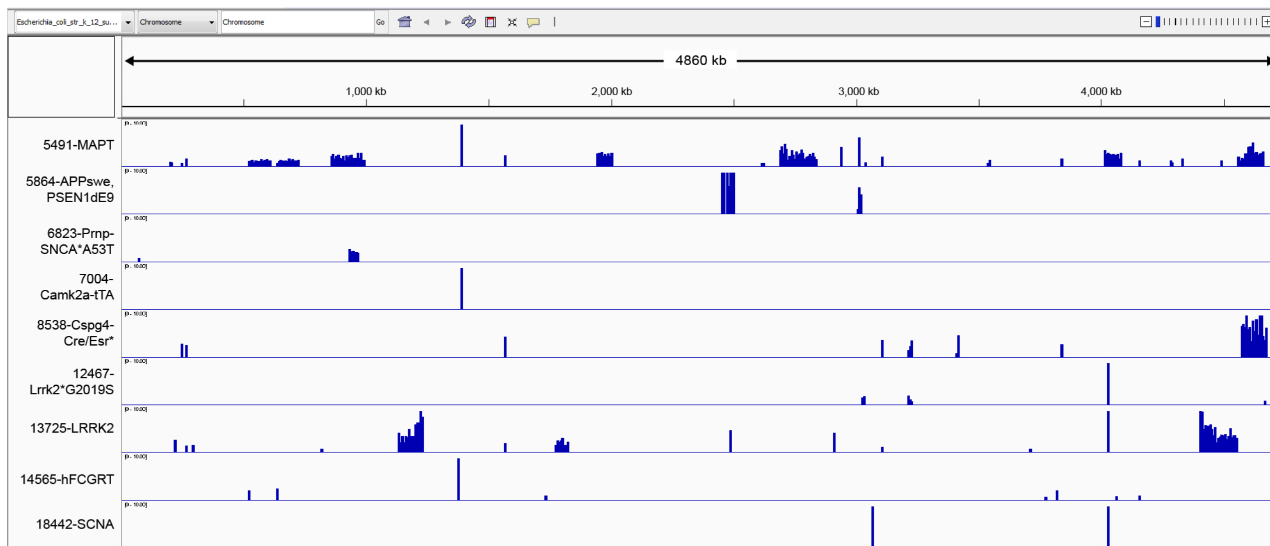

Figure 4

Figure 4. TLA reveals additional passenger cassettes and fragments in transgenes. A, B. View of TLA reads that map to the human growth hormone (hGH) gene for two transgenes (Ins2-cre and Vil-cre), showing the inclusion of the entire gene structure, including coding exons. C. Reads for nine transgenes mapped to the *E.coli* genome indicating a variable level of co-insertion into the transgene integration site. Deep coverage for discrete loci shared between multiple lines suggests that these sequences are part of the transgene vector. The amount of *E.coli* co-integration ranges from a few hundred bp to more than 200Kb.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 ranging from as little as 300bp to more than 200Kb (Figure 4C). Some of the small fragments were
2 identical between samples, suggesting that these were intentional components of the transgene, possibly
3 as part of the vector backbone. However, for lines with significant *E.coli* genome contribution, it is likely
4 that this is the result of contamination in the microinjection preparation of the construct. While the impact
5 of this finding for these specific lines is unclear, prior reports have shown that bacterial sequences can
6 contribute to transgene silencing (Scrable and Stambrook 1997; Chen et al. 2004).
7
8 While some of the small number of transgene insertion sites currently known were discovered
9 following the serendipitous identification of an unexpected transgene-specific phenotype, systematic
10 phenotyping of transgenic lines to assess the impact of transgene insertion has not been reported. Taking
11 advantage of the high-throughput KOMP2 Phenotyping platform at JAX (White et al. 2013; de Angelis et
12 al. 2015; Dickinson et al. 2016; Karp et al. 2017; Meehan et al. 2017), we asked whether we could detect
13 phenotypes in a selection of 7 Cre driver lines from the lines examined above. As colony maintenance
14 strategy differed amongst lines, we pooled control mice and WT C57BL/6J to create a reference
15 population. As shown in Figure 5, we identified 66 significant phenotypes amongst strains, with Nes-cre
16 displaying a the most phenotype hits (21) and Vil-cre displaying the least (2). Physiology phenotypes
17 were most common, with both Nes-cre and Ins2-cre showing 19 and 13 abnormalities, respectively. As
18 noted above, metabolic phenotypes in Nes-cre mice has been reported by others (Galichet et al. 2010;
19 Giusti et al. 2014), and a recent paper suggests these phenotypes are due to presence of the hGH minigene
20 (Declercq et al. 2015). The same hGH minigene is also found in the Ins2-cre allele, possibly explaining
21 the metabolic phenotypes observed. This line is reported to develop age-related impaired glucose
22 tolerance of unknown etiology(Lee et al. 2006), and thus the presence of the hGH minigene should be
23 explored as a possible explanation. Both Alb-cre and Lck-cre also carry this minigene, but do not show
24 the same number of phenotypic hits, suggesting that expression of the minigene varies between
25 transgenes, and thus it cannot be assumed that its presence alone is necessarily confounding. It is
26 interesting that Vav1-icre displayed only two hits despite landing in the *Commd10* gene, for which a KO

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

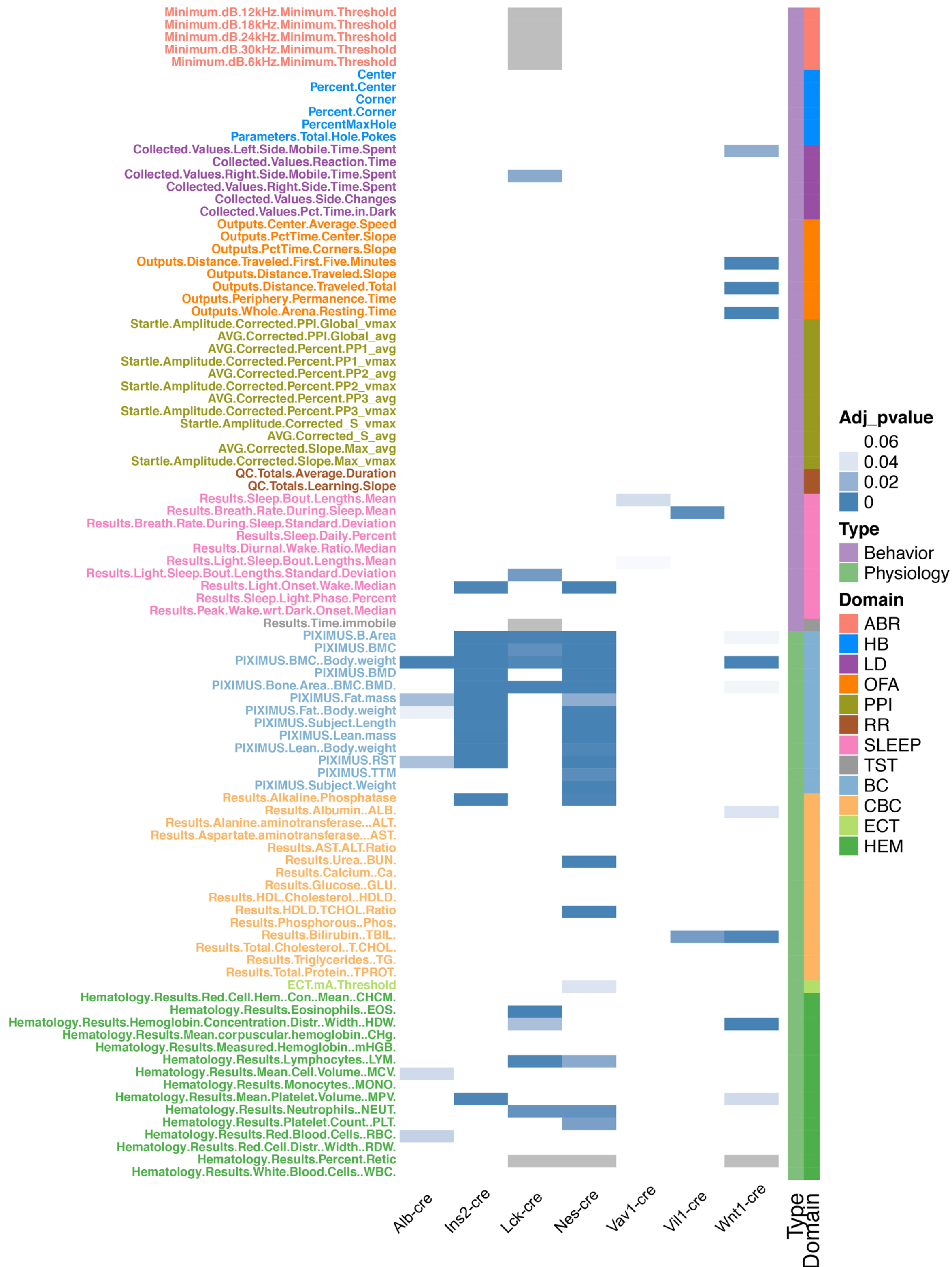

Figure 5. Physiology and behavioral testing of the cre transgenic lines in the KOMP pipeline. Mice were tested in 12 phenotypic domains spanning behavior and physiology (color coded, right bar). Each test is color coded (right bar) and grouped broadly into behavior (purple) and physiology domains (green). Significant differences from controls are shows in the heat map (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.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 allele is homozygous lethal. This is consistent, however, with the IMPC phenotyping data
2 (www.mousephenotype.org), which shows no significant phenotypes in *Commd10*$^{tm1a(EUCOMM)Wtsi/+}$ mice.
3 By contrast, the Wnt1-cre transgene inserts into the histone gene *H2afv* and shows 11 phenotypic hits
4 which span several domains (Figure 5), including four significant behavioral phenotypes. Currently, there
5 are no reports of targeted mutations of phenotypes for this gene.
6
7 **DISCUSSION**

8 Despite widespread use of transgenic lines in the scientific community, the consequence of
9 random transgene insertion in these lines is largely unknown. Here we show, expanding on prior work,
10 that TLA represents a rapid and efficient means to precisely identify the site of insertion, and critically,
11 the corresponding molecular consequences. These events are associated with structural variations,
12 primarily deletions and duplications, including a deletion of greater than 1Mb and a complex structural
13 variation that includes a simultaneous duplication, deletion, and inversion. While TLA simplifies the
14 discovery process, reconstruction of the full transgene structure is difficult due to the high copy
15 concatemerization of the transgene, coupled with complex structural variation that can accompany
16 insertion. Given these contigs are built using relatively short read sequencing technology, including
17 several instances where the fusion was covered by one read, it is critical to validate each putative fusion
18 fragment using PCR-sequencing, and we indeed found small differences in the actual fusion sequence in a
19 few lines. Targeted long-read sequencing approaches (PacBio, Oxford nanopore), might prove to be a
20 useful complement to TLA for full characterization of transgenic alleles. However, such reconstruction is
21 not necessary for typical use, as the key elements of chromosome location, break point, and structural
22 variation are easily obtained with TLA and PCR validation alone.

23 In contrast to a recent similar screen on a small number of Cre driver lines(Cain-Hom et al. 2017),
24 we found a surprisingly high percentage (50%) disrupt annotated genes, the majority of which are protein
25 coding. Many of these genes, when mutated, are known to result in observable phenotypes, including four
26 that are homozygous lethal. At face value, this is unexpected as only 3% of the genome is protein coding,

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 and thus we would expect a similar hit rate with truly random insertion. However, many of these
2 insertions are accompanied by substantial deletions, which would increase the odds of hitting a gene. In
3 addition, our set of transgenic lines in not an unbiased selection of random transgenic animals; they are
4 lines selected for robust transgene expression and activity. Genic regions of the genome are likely to
5 support active transgene expression, as opposed to intergenic stretches and regions of heterochromatin.
6 While we do observe that larger transgenes derived from BAC (or similar) constructs, which contain a
7 larger complement of genetic elements required for proper expression, can insert into and/or delete coding
8 genes (e.g. Cspg4-cre), our sample size is too small to determine if the rate is different than that of smaller
9 transgenic constructs.
10 TLA-based discovery of transgenic insertion sites provides a number of practical benefits that
11 should improve quality control for both public repositories and the end user. For example, allele-specific
12 assays can be developed at the integration site to distinguish all genotype classes, allowing for
13 homozygous mating strategies unless precluded by insertional mutagenesis. End users of Cre lines can
14 use knowledge of the genetic locus before attempting to mate to a floxed target allele that is linked to the
15 Cre line, selecting an alternative unlinked line, or scaling their breeding to assure identification of rare
16 recombinants.
17 Our findings further illustrate the need to use proper controls in all experiments that include
18 transgenic lines. Several studies have shown that expression of Cre itself can have phenotypic or toxic
19 effects (Loonstra et al. 2001; Naiche and Papaioannou 2007; Bersell et al. 2013; Lexow et al. 2013).
20 Given these findings it is clear that animals/embryos expressing Cre alone must be included as a control,
21 and our results provide additional evidence that this control strategy is essential. Therefore, the potentially
22 confounding impact of frequent insertional mutagenesis in Cre driver lines can be managed with the
23 proper use of controls, depending on the research question and phenotype. For transgenic lines that are
24 employed as disease models, a Tg-only control is not possible. In many cases, the original publication
25 included results from multiple founders corroborating the findings of the line that ultimately became the
26 "standard" for subsequent studies. It is interesting, however, that most studies published with an

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 "established" model do not include the same level of independent corroboration, despite significant
2 differences in study design, including analysis of additional phenotypes, inclusion of additional mutant
3 alleles, and/or the use of a distinct genetic background. It is plausible that in those scenarios, effects of
4 insertional mutagenesis not seen in the original publication might manifest, confounding the interpretation
5 of the data. Typically, multiple alleles are not deposited in a public biorepository for distribution or
6 retained at all, and thus reproduction of results with independent transgenic lines is impossible,
7 notwithstanding the practical and financial challenge of reproducing every study with multiple transgenic
8 lines. Thus, it seems prudent, and now feasible, for investigators to determine the insertion site of the
9 transgenic line used in their study if independent corroboration is not possible.

10 One potential use of TLA is to confirm the content of a given transgenic line, providing a level of
11 quality control not available through other means. This includes clarification of the specific details of the
12 constructs components (e.g. hGH or MT1 minigenes) that are either omitted or reported incorrectly. It is
13 worth noting that in the case of both "polyA" signals reported here, a careful review of the literature
14 clearly shows that the original content of the vectors is correctly reported (Sauer and Henderson 1990;
15 Orban et al. 1992), but this information was omitted or incorrectly cited in subsequent descriptions of the
16 construct or mouse strain. Given the number of years and hands involved, this type of "information
17 mutation" is not surprising. Indeed, we have seen that ~20% of all lines submitted to the JAX Repository
18 carry alleles or have been bred to mouse strains not reported by the donating investigator. TLA provides
19 an additional tool for assuring the content and nature of the allele for both investigators sharing their
20 strain and for repositories distributing strains to scientists around the world.

21 Primary phenotyping of a subset of Cre drivers included in this study demonstrate the potential
22 scope of "endogenous" phenotypes in transgenes in common use. While the impact of insertional
23 mutagenesis is clear, for most KO alleles only homozygous mutants are carefully phenotyped, and thus
24 potential confounding heterozygous phenotypes are unclear. Moreover, some transgenes delete multiple
25 genes, and the combinatorial effect on phenotype would require independent evaluation. Finally, the
26 transgene-specific caveats of passengers (minigenes, genes on BACs, *E.coli* genome, etc.) require specific

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 testing. As noted above, the use of proper controls can mitigate most concerns arising from insertional
2 mutagenesis, passenger cassettes, or transgene toxicity, assuming it does not directly impact the
3 phenotype of interest. However, with the emergence of high-throughput phenotyping pipelines
4 (Dickinson et al. 2016; Karp et al. 2017; Meehan et al. 2017), it is now feasible to broadly characterize
5 the phenotypes of a larger collection transgenic tool lines, perhaps in parallel to insertion site discovery.
6 Given our findings, and the potential caveats implied for the use of transgenes, it is tempting to
7 suggest that the community should move away from making, and ultimately using, lines generated by
8 random transgenesis. For Cre lines, knockin alleles targeting the endogenous locus of a desired driver
9 gene has the added potential advantage of providing greater specificity, desirable given the high rate of
10 off-target activity seen in many transgenic lines (Heffner et al. 2012). To this end, the EUCOOMTOOLs
11 program has produced hundreds of new Cre driver lines using this strategy (Murray et al. 2012; Rosen et
12 al. 2015). However, this typically comes at the cost of haploinsufficiency at the driver locus, often a gene
13 that is part of a pathway critical to the development of the cell type or tissue in question. Expression
14 levels in a knockin might be lower than that of a multicopy transgene, thus sacrificing effectiveness for
15 specificity. The use of neutral locus docking sites or targeted transgenesis facilitated by CRISPR can
16 avoid the mutagenic risk associated with random insertion, but the former is typically a single copy event
17 and the latter is relatively untested. Thus, while alternatives to random transgenesis exist, they come with
18 their own caveats and do not necessarily provide a suitable alternative. Rather, given the impact of
19 discoveries enabled by transgenic lines, knowledge of the transgenic insertion site is best viewed as one
20 of many critical pieces of information that should be considered in an experimental design.
21

22 **METHODS**

23 **Mice**

24 All strains used for the TLA analysis were obtained from the Jackson Laboratory Repository, 4 of which
25 are distributed from the JAX Mouse Mutant Research and Resource Center (MMRRC). The specific
26 mouse strains and JAX Stock # (and MMRRC Stock # if applicable) are available in Supplemental Table

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 1. All procedures and protocols (see Phenotyping below) were approved by The Jackson Laboratory
2 Animal Care and Use Committee, and were conducted in compliance with the National Institutes of
3 Health Guideline for Care and Use of Laboratory Animals.
4
5 **Isolation of splenocytes**
6 Splenocytes were isolated from each line as previously described (de Vree et al. 2014).
7 In brief, the spleens were dissected and stored on ice. A single cell suspension was made using a 40
8 micron mesh filter suspending the cells in 10% Fetal Calf Serum (FCS)/Phosphate Buffered Saline (PBS).
9 Following centrifugation at 4◦C at 500 × g for 5 min. The supernatant was discarded, the pellet dissolved
10 in 1ml 1× Pharm Lyse (BD Biosciences) and incubated at room temperature for 3 min to lyse splenic
11 erythrocytes. To terminate the lysis reaction, 0.5 ml phosphate buffered saline (PBS) was added followed
12 by centrifugation at 4◦C, 500 × g for 5 min. The supernatant was discarded and the pellet resuspended in
13 0.5 ml PBS. After one final centrifugation step for 2 min, the supernatant was discarded and cell pellet
14 resuspended in 1 ml freeze medium (PBS with 10% Dimethyl Sulfoxide and 10%fetal calf serum). The
15 samples were stored at minus 80◦C until shipment for TLA processing.
16
17 **TLA procedure**
18 Targeted locus amplification (TLA) was performed as previously described. (de Vree et al. 2014) In brief,
19 Cells were crosslinked using formaldehyde, after which the DNA was digested using the restriction
20 enzyme NlaIII (CATG). Subsequently the sample was ligated, crosslinks were reversed and the DNA was
21 purified. To obtain circular chimeric DNA molecules for PCR amplification, the DNA molecules were
22 trimmed with NspI and ligated at a DNA concentration of 5 ng/μl to promote intramolecular ligation.
23 NspI has a RCATGY recognition sequence that encompasses the CATG recognition sequence of NlaIII,
24 which ensures only a subset of NlaIII (CATG) sites were (re-) digested, generating DNA fragments of
25 approximately 2 kb and allowing the amplification of entire restriction fragments. After ligation the DNA
26 was purified, and eight 25-μl PCR reactions, each containing 100 ng template, were pooled for

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sequencing. Sequences of the inverse primers, which were designed using Primer3 software38, can be found in Supplementary Table 4.

Mapping and sequence alignment

The primer sets were used in individual TLA amplifications. PCR products were purified and library prepped using the Illumina NexteraXT protocol and sequenced on an Illumina Miseq sequencer. Reads were mapped using BWA-SW, which is a Smith-Waterman alignment tool. 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. Identified TG sequences based on these alignments have been confirmed in this paper.

Sequence validation.

Breakpoint-spanning reads-All TG integration fusion sites 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 Primer 3 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 2 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.

QPCR analysis

Genomic DNA isolated from tail biopsies were used to analyze loss of allele (LOA) or relative concentration qPCR on Applied Biosystem's ViiA 7 (Applied Biosystems, Foster City, CA USA).

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

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 duplication of the genomic target region. Based on transgene integration site sequences and resultant
2 deletions or duplications, target genomic region q-PCR 5' nuclease assays were designed using
3 PrimerQuest software (Integrated DNA Technologies). The internal reference control *Apob* probe
4 contains a VIC (4,7,2′-trichloro-7′-phenyl-6-carboxyfluorescein) reporter dye, (ABI, Applied Biosystems,
5 Foster City, CA USA) while all experimental assays use FAM (6-carboxyfluorescein) labeled probes for
6 detection. An NFQ-MGB (*Apob*) dark or Zen/Iowa Black FQ quencher (IDT) is used for all assays.
7 Primer and probe sequences are provided in Supplementary Table S3.
8
9 QPCR samples were then analyzed in triplicate and Cq values for the samples and the internal reference
10 (*Apob*) were calculated using Viia7 Software (QuantStudio™ Software V1.3, ABI, Applied Biosystems,
11 Foster City, CA USA). The means of the Cq values were used to calculate ΔCq values and these were
12 then used to calculate relative copy number of the recombinant region using the $2^{-\Delta\Delta Cq}$ formula (Livak
13 and Schmittgen 2001).
14

15 **Phenotyping**

16 We employed a modified version of the IMPReSS pipeline (www.mousephenotype.org/impress) for high-
17 throughput clinical phenotyping assessment, which was developed under the IMPC program (Dickinson
18 et al. 2016; Karp et al. 2017; Meehan et al. 2017), de Angelis, Nat Genet. 2015 September; 47(9): 969–
19 978. doi:10.1038/ng.33). The following seven lines (and genotypes) were characterized: Alb-cre (HOM),
20 Ins2-cre (HOM), Lck-cre (HOM), Nes-cre (HEMI), Vav1-icre (HEMI), Vil-cre (HEMI) and Wnt1-cre
21 (HEMI). Control mice are from a pool consisting of C57BL/6J WT mice and non-carrier (NCAR)
22 controls from the colony for lines maintained in a HEMI x NCAR breeding scheme (Wnt1-cre, Vav1-icre,
23 Vil-cre, and Nes-cre). For each mouse strain, 8 male and 8 female transgenic animals, NCAR controls, or
24 C57BL/6J mice were processed through the JAX Adult Phenotyping Pipeline. Full details of the JAX
25 Adult Phenotyping Pipeline can be found on the IMPC website
26 (www.mousephenotype.org/impress/procedures/12). Briefly, mice were received into the pipeline at 4

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

weeks of age, body weight was collected weekly, and assays were performed weekly from 8 to 18 weeks 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 wks)
Light-Dark Transition (LD), Holeboard (HB) (9 wks)
Acoustic Startle/Pre-pulse Inhibition (PPI) (10 wks)
Tail Suspension, Electrocardiogram, Rotarod (11 wks)
Body Composition (BC), (14 wks)
Piezoelectric Sleep/Wake (SLEEP) (15 wks)
Auditory Brainstem Response (ABR) (4M + 4F) (16 wks)
Electroconvulsive Seizure Threshold (ECT) (17 wks)
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-second pressure signal segments, and classification is automatically performed every 2 seconds using overlapping windows. From this, the following parameters are calculated: Sleep bout lengths (light phase, dark phase, 24 hour 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

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 phenotypes. In the linear mixed model, explanatory factors including sex, weight and mutant genotype
2 were treated as fixed effects, while batch (date of test) was treated as a random effect adding variation to
3 the data (see Equation 1).

4 $$Variable = Genotype + Sex + Weight + (1|Batch) \qquad \text{Equation 1}$$

5 Parameters from the mixed model are estimated using the method of restricted maximum likelihood
6 (REML). Adjusted p-values were calculated from nominal p-values in mixed models to control for False
7 Discovery Rate (FDR). All data analysis was performed using R.
8

9 **DATA ACCESS**

10 All sequence date will be submitted to the NCBI Sequence Read Archive (SRA;
11 http://www.ncbi.nlm.nih.gov/sra)
12 Phenotypic data have been submitted to the Mouse Phenome Database at phenome.jax.org
13

14 **ACKNOWLEDGEMENTS**

15 This work was supported in part by the Office of the Director, National Institutes of Health and the
16 National Human Genome Research Institute of the National Institutes of Health under award numbers
17 R24OD011190, U42OD011185, UM1OD023222, and HG006332. The authors thank Brianna Caddle and
18 Larry Bechtel for their technical assistance in isolating spleen cells, and Kevin Peterson for his helpful
19 and thoughtful comments on the manuscript.
20

21 **DISCLOSURE DECLARATIONS**

22 ES and MvM are employees at Cergentis b.v.
23

24 **SUPPLEMENTARY INFORMATION**

25 **Supplementary Table 1.** Detailed summary of results for each of the 40 lines included in this study
26 **Supplementary Table 2.** Primers and expected PCR product sizes used for validation of insertion sites.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Table 3. qPCR probes used for LOA validation.

Supplementary Table 4. Primers used for TLA amplifications.

Supplementary Table 5. Table of statistical results for mouse phenotyping.

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 alphaMHC-MerCreMer mice induce a DNA damage response, leading to heart failure and death. *Dis Model Mech* **6**: 1459-1469.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 Heffner CS, Herbert Pratt C, Babiuk RP, Sharma Y, Rockwood SF, Donahue LR, Eppig JT, Murray SA.
2 2012. Supporting conditional mouse mutagenesis with a comprehensive cre characterization
3 resource. *Nat Commun* **3**: 1218.
4 Hottentot QP, van Min M, Splinter E, White SJ. 2017. Targeted Locus Amplification and Next-
5 Generation Sequencing. *Methods Mol Biol* **1492**: 185-196.
6 Karp NA, Mason J, Beaudet AL, Benjamini Y, Bower L, Braun RE, Brown SDM, Chesler EJ, Dickinson
7 ME, Flenniken AM et al. 2017. Prevalence of sexual dimorphism in mammalian phenotypic traits.
8 *Nat Commun* **8**: 15475.
9 Lee JY, Ristow M, Lin X, White MF, Magnuson MA, Hennighausen L. 2006. RIP-Cre revisited,
10 evidence for impairments of pancreatic beta-cell function. *J Biol Chem* **281**: 2649-2653.
11 Lewis AE, Vasudevan HN, O'Neill AK, Soriano P, Bush JO. 2013. The widely used Wnt1-Cre transgene
12 causes developmental phenotypes by ectopic activation of Wnt signaling. *Dev Biol* **379**: 229-234.
13 Lexow J, Poggioli T, Sarathchandra P, Santini MP, Rosenthal N. 2013. Cardiac fibrosis in mice
14 expressing an inducible myocardial-specific Cre driver. *Dis Model Mech* **6**: 1470-1476.
15 Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative
16 PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**: 402-408.
17 Loonstra A, Vooijs M, Beverloo HB, Allak BA, van Drunen E, Kanaar R, Berns A, Jonkers J. 2001.
18 Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc Natl*
19 *Acad Sci U S A* **98**: 9209-9214.
20 Madison BB, Dunbar L, Qiao XT, Braunstein K, Braunstein E, Gumucio DL. 2002. Cis elements of the
21 villin gene control expression in restricted domains of the vertical (crypt) and horizontal
22 (duodenum, cecum) axes of the intestine. *J Biol Chem* **277**: 33275-33283.
23 Mang GM, Nicod J, Emmenegger Y, Donohue KD, O'Hara BF, Franken P. 2014. Evaluation of a
24 piezoelectric system as an alternative to electroencephalogram/ electromyogram recordings in
25 mouse sleep studies. *Sleep* **37**: 1383-1392.
26 Meehan TF, Conte N, West DB, Jacobsen JO, Mason J, Warren J, Chen CK, Tudose I, Relac M,
27 Matthews P et al. 2017. Disease model discovery from 3,328 gene knockouts by The
28 International Mouse Phenotyping Consortium. *Nat Genet* **49**: 1231-1238.
29 Murray SA, Eppig JT, Smedley D, Simpson EM, Rosenthal N. 2012. Beyond knockouts: cre resources for
30 conditional mutagenesis. *Mamm Genome* **23**: 587-599.
31 Naiche LA, Papaioannou VE. 2007. Cre activity causes widespread apoptosis and lethal anemia during
32 embryonic development. *Genesis* **45**: 768-775.
33 Nakanishi T, Kuroiwa A, Yamada S, Isotani A, Yamashita A, Tairaka A, Hayashi T, Takagi T, Ikawa M,
34 Matsuda Y et al. 2002. FISH analysis of 142 EGFP transgene integration sites into the mouse
35 genome. *Genomics* **80**: 564-574.
36 Orban PC, Chui D, Marth JD. 1992. Tissue- and site-specific DNA recombination in transgenic mice.
37 *Proc Natl Acad Sci U S A* **89**: 6861-6865.
38 Padmanabhan N, Jia D, Geary-Joo C, Wu X, Ferguson-Smith AC, Fung E, Bieda MC, Snyder FF, Gravel
39 RA, Cross JC et al. 2013. Mutation in folate metabolism causes epigenetic instability and
40 transgenerational effects on development. *Cell* **155**: 81-93.
41 Pease S, Saunders TL, International Society for Transgenic Technologies. 2011. *Advanced protocols for*
42 *animal transgenesis : an ISTT manual*. Springer, Heidelberg ; New York ;.
43 Rosen B, Schick J, Wurst W. 2015. Beyond knockouts: the International Knockout Mouse Consortium
44 delivers modular and evolving tools for investigating mammalian genes. *Mamm Genome* **26**: 456-
45 466.
46 Sauer B, Henderson N. 1990. Targeted insertion of exogenous DNA into the eukaryotic genome by the
47 Cre recombinase. *New Biol* **2**: 441-449.
48 Scrable H, Stambrook PJ. 1997. Activation of the lac repressor in the transgenic mouse. *Genetics* **147**:
49 297-304.

bioRxiv preprint doi: https://doi.org/10.1101/236307; this version posted December 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 Srivastava A, Philip VM, Greenstein I, Rowe LB, Barter M, Lutz C, Reinholdt LG. 2014. Discovery of
2 transgene insertion sites by high throughput sequencing of mate pair libraries. *BMC Genomics* **15**:
3 367.
4 Tosh JL, Rickman M, Rhymes E, Norona FE, Clayton E, Mucke L, Isaacs AM, Fisher EMC, Wiseman
5 FK. 2017. The integration site of the APP transgene in the J20 mouse model of Alzheimer's
6 disease. *Wellcome Open Res* **2**: 84.
7 Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new
8 capabilities and interfaces. *Nucleic Acids Res* **40**: e115.
9 Valenzuela DM, Murphy AJ, Frendewey D, Gale NW, Economides AN, Auerbach W, Poueymirou WT,
10 Adams NC, Rojas J, Yasenchak J et al. 2003. High-throughput engineering of the mouse genome
11 coupled with high-resolution expression analysis. *Nat Biotechnol* **21**: 652-659.
12 White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ,
13 Podrini C et al. 2013. Genome-wide generation and systematic phenotyping of knockout mice
14 reveals new roles for many genes. *Cell* **154**: 452-464.
15

16