Efficient mapping of transgene integration sites and local structural changes in Cre transgenic mice using targeted locus amplification

Carol Cain-Hom[1], Erik Splinter[2], Max van Min[2], Marieke Simonis[2], Monique van de Heijning[2], Maria Martinez[1], Vida Asghari[1], J. Colin Cox[1,*] and Søren Warming[3,*]

[1]Department of Transgenic Technology, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA, [2]Cergentis BV, Yalelaan 62, 3584 CM Utrecht, the Netherlands and [3]Department of Molecular Biology, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA

ABSTRACT

Cre/LoxP technology is widely used in the field of mouse genetics for spatial and/or temporal regulation of gene function. For Cre lines generated via pronuclear microinjection of a Cre transgene construct, the integration site is random and in most cases not known. Integration of a transgene can disrupt an endogenous gene, potentially interfering with interpretation of the phenotype. In addition, knowledge of where the transgene is integrated is important for planning of crosses between animals carrying a conditional allele and a given Cre allele in case the alleles are on the same chromosome. We have used targeted locus amplification (TLA) to efficiently map the transgene location in seven previously published Cre and CreERT2 transgenic lines. In all lines, transgene insertion was associated with structural changes of variable complexity, illustrating the importance of testing for rearrangements around the integration site. In all seven lines the exact integration site and breakpoint sequences were identified. Our methods, data and genotyping assays can be used as a resource for the mouse community and our results illustrate the power of the TLA method to not only efficiently map the integration site of any transgene, but also provide additional information regarding the transgene integration events.

INTRODUCTION

Cre/LoxP technology is widely used in the field of mouse genetics for spatial and/or temporal regulation of gene function ([1–3]) and hundreds of Cre 'deleter' lines are available to the mouse community. Cell-type specific expression of Cre allows for specific deletion of a gene of interest by the use of a 'conditional knock-out' (CKO) allele of that gene ([2]). Typically, for a conditional allele a critical exon(s) is flanked by two loxP sites ('floxed') and in the cells where Cre is expressed, the floxed exon(s) is removed, resulting in a deletion, or knock-out, allele. Cre deleter lines are generated either by targeted knock-in of the Cre cDNA into an endogenous locus or by pronuclear microinjection of a Cre transgene driven by a cell-type specific promoter. For the latter, the integration site is random and in most cases not known. Knowledge of where the transgene is integrated is important for planning of crosses between animals carrying a conditional allele and a given Cre allele in case the alleles are on the same chromosome. This becomes increasingly important in complex crosses with multiple conditional alleles, as some combinations of alleles might not be possible. Importantly, integration of a transgene can disrupt an endogenous gene, potentially interfering with interpretation of the transgenic phenotype ([4–6]) or preventing the generation of homozygous transgenic animals due to embryonic lethality when the transgene is bred to homozygosity. Transgenes often integrate as a multicopy concatemer ([7]) and in the absence of integration site data, hemi- and homozygous animals have to be distinguished by copy number variation (CNV) analysis which involves quantitative polymerase chain reaction (PCR) and reference DNA with a known copy number. The resolution and reliability of CNV analysis for accurate genotyping decreases as copy number increases. Therefore, another important and practical use of knowing the exact transgene insertion site is that efficient, copy-number independent and locus-specific genotyping assays can be developed. For large transgenic animal facilities, where a large number of genotyping assays are

performed, automated genotyping platforms and robust assays are necessary (8). CNV analysis is more labor-intensive than real-time PCR assays due to the need for an internal copy number standard and the additional requirement for replicates to ensure accurate copy-number calling (9). Alternatively, if the insertion site is known, an assay can be designed for the wild-type DNA and for the insertion, allowing for analysis without having to calculate copy number in order to determine zygosity. Thus the characterization of transgene integration sites is important both to know if any rearrangements have occurred in the endogenous sequence of the integration site and to enable automated genotyping of as many of the Cre transgenic lines as possible.

The most commonly used method for identification of transgene insertion sites is based on inverse PCR (iPCR) (10,11). This method relies on knowledge of appropriate restriction sites located in the transgene and on sequence information to design appropriate primers for the iPCR. Furthermore, iPCR works best if the number of transgene copies is low as selective amplification of the transgene concatemer reduces the chance of identifying a PCR product containing flanking genomic sequence (12). An alternative method for identification of insertion sites, 'Splinkerette', was first developed for the cloning of retroviral integration sites (13). Just as for iPCR, Splinkerette is best suited for single or low copy integrations. While iPCR and Splinkerette can be used to identify the genomic/transgene junction, both methods rely on prior knowledge of the end(s) of integrated transgene sequences and on whether the breakpoint sequence is suited for PCR amplification. In addition, neither method provides any information about structural changes at the integration site. Although both iPCR and Splinkerette methods now take advantage of next-generation sequencing (NGS) (for example, see (14)), NGS technology alternatives to iPCR and Splinkerette for transgene mapping now include whole genome sequencing (15,16) and sequence capture followed by NGS (17). Whole genome sequencing analyses are not suited for the detailed analysis of the transgene sequence since most of the generated sequencing data is uninformative. More importantly, in hemizygous mice, the generated sequencing data is not only derived from the locus of integration but also from the wild type locus, obscuring interpretation of structural changes specific to the integration site. Moreover, many transgenes contain sequences that are homologous to endogenous sequences and the presence of the endogenous homologs in whole genome sequencing data precludes sensitive and specific analyses of the transgene. Capture (or multiplex PCR based) enrichment technologies require detailed knowledge of the transgene sequence and, inherently, only provide information across sequences known to be present in the transgene and for which capture probes/PCR primers were designed. The integration site in the genome is only detected if that specific fragment was captured by the designed probes and successfully sequenced.

As an alternative to whole genome sequencing and capture based targeted re-sequencing, targeted locus amplification (TLA) enables the selective amplification and NGS sequencing of the locus of interest without the need for detailed knowledge of the region (18). The TLA technology is based on the crosslinking, fragmentation, religation and selective amplification of DNA and results in the amplification of >100 kb of sequence information at either end of a primer pair complementary to a short (trans-) gene specific sequence. As such, TLA-based targeted sequencing enables targeted complete sequencing of loci of interest and detection of all single nucleotide variations (SNVs) and structural variants. Here we describe the use of TLA to efficiently map the transgene location in seven previously published Cre and CreERT2 transgenic lines. These seven mouse lines are all widely used by the mouse community. For example, see http://www.informatics.jax.org/home/recombinase for an up-to-date list of references for each transgenic line. With very limited sequence data available for the transgenes used to generate these lines, Cre transgenic lines provided an ideal test-case for the power of the TLA approach to transgene mapping. Because the mouse lines all have the Cre sequence in common, we designed primers on the Cre sequence and these primers could then be used in the analysis for all Cre lines. After mapping the exact integration sites and identifying structural changes, we used the information to design quantitative PCR genotyping assays that distinguish wild-type, hemizygous and homozygous animals for four of the lines. Our data illustrates the power of the TLA method to sequence transgenes and efficiently map their integration sites with very limited requirement for prior knowledge of the transgene sequence. Furthermore, our data illustrates the power of TLA to identify structural changes occurring at the site of transgene integration.

MATERIALS AND METHODS

Mouse strains

The Cre and CreERT2 transgenic strains used in this study were licensed and obtained from different sources and then maintained at Genentech. ACTB-Cre: FVB/N-Tg(ACTB-cre)2Mrt/J (19). Obtained from the Jackson laboratory (stock no. 003376) and backcrossed until congenic on C57BL/6N at Genentech. BEST1-Cre: C57BL/6-Tg(BEST1-cre)1Jdun/J (20). Obtained from the Jackson laboratory (stock no. 017557). Generated on C57BL/6NCrl, then crossed to and maintained on C57BL/6J. Cdh5-CreERT2: Tg(Cdh5-cre/ERT2)CIVE23Mlia (21). Obtained from University of California, LA (UCLA). Generated on C57BL/6J and kept on a C57BL/6J background at Genentech. 'CIVE': C57BL/6J, Inducible, VE-cadherin Cre-recombinase mouse. Pdx1-Cre: Tg(Pdx1-cre)6Tuv (22). Obtained from University of Cincinatti and then back-crossed until congenic on C57BL/6J-Tyr^{c-2J} at Genentech. Syn1-Cre: B6.Cg-Tg(Syn1-cre)671Jxm/J (23). Obtained from The Jackson laboratory (stock no. 003966). This strain originated on a B6:CBAF1 background and then backcrossed to wild-type C57BL6/NHsd mice for at least five generations at the Jackson laboratory. Maintained at Genentech on the mixed C57BL/6N:CBA background. Tyr-CreERT2: Tg(Tyr-cre/ERT2)13Bos (24). Obtained from the University of Vermont, via Charles River Laboratories, then backcrossed until congenic on C57BL/6J at Genentech. Vil1-Cre: B6.Cg-Tg(Vil-cre)997Gum/J (25). Obtained from The Jackson laboratory (stock no. 004586). Congenic on C57BL/6J. All rodent studies were performed

in accordance with Institutional Animal Care and Use Committee–approved guidelines.

Splenocyte preparation

Mice were euthanized and the spleens dissected and stored on ice. Splenocytes were then isolated and purified through a 40 μm mesh filter. The splenocytes were collected by centrifugation at 4°C at $500 \times g$ for 5 min. For each spleen, the supernatant was discarded and the pellet dissolved in 1 ml 1× Pharm Lyse (BD Biosciences) and the samples were incubated at room temperature for 3 min to lyse splenic erythrocytes. To terminate the lysis reaction, 0.5 ml phosphate buffered saline (PBS) was added and the splenocytes were collected by centrifugation at 4°C, $500 \times g$ for 5 min, the supernatant discarded and the pellet resuspended in 0.5 ml PBS. After one final centrifugation step for 2 min, the supernatant was discarded and cell pellet resuspended in 1 ml freeze medium (PBS with 10% Dimethyl Sulfoxide and 10% fetal calf serum). The samples were stored at minus 80°C until TLA processing.

TLA

Preparation of the samples for TLA was performed as described (18). In brief, the splenocytes were crosslinked using formaldehyde and DNA was digested with NlaIII. The samples were ligated, crosslinks reversed, and the DNA 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. Importantly, NspI was chosen for its RCATGY recognition sequence that encompasses the CATG recognition sequence of NlaIII. As a consequence, only a subset of NlaIII (CATG) sites were (re-)digested, generating DNA fragments of approximately 2 kb and allowing the amplification of entire restriction fragments. Sequences of the Cre iPCR primers are (5′ to 3′): 2291_CRE1_F GGAGTTTCAATA CCGGAGAT; 2292_CRE1_R AGGGTGTTATAAGCAA TCCC; 2299_CRE2_F AGTTTCAATACCGGAGATCA; 2300_CRE2_R TTTCGGCTATACGTAACAGG.

After ligation, the DNA was purified, and eight 25-μl PCR reactions, each containing 100 ng template, were pooled for sequencing. Illumina NexteraXT NGS library preparations were performed according to manufacturer's protocols. We performed sequencing of TLA libraries on the Illumina MiSeq platform pooling ∼20 libraries per V2 PE150 sequencing run yielding on average 1 million reads per library.

Bioinformatics/sequence alignment

Because the TLA protocol leads to reshuffling of genomic sequences, reads were mapped using split-read aware alignment with BWA mapping software version 0.6.1-r104, settings: bwasw –b 7 (26). Although we perform paired-end sequencing, we do not use the paired-end information in the mapping owing to reshuffling of the sequence. Paired ends are therefore treated separately in the general analysis. The data were aligned to the human (hg19), mouse (mm9) and

rat (rn5) genome. The resulting BAM files were analyzed using IGV software (27).

Genotyping assays

For all seven lines, the mice were first typed for presence or absense of the Cre transgene using the generic *Cre_Apob* primers and probes listed in Supplementary Table S1. Using the same primers and probes, samples were then analyzed in triplicate or quadruplicate and C_q values for the samples and the internal reference (*Apob*) were calculated using CFX Manager software (Bio-Rad). The means of the C_q values were used to calculate ΔC_q values and these were then used to calculate transgene copy number using the $2^{-\Delta\Delta C_q}$ formula (28). Genomic DNA was isolated from tail biopsies using the Agencourt DNAdvance kit, 96 well (Beckman Coulter) and tail DNA from a homozygous knock-in Cre line (two copies of Cre) was used as copy-number reference. Based on transgene integration site sequences, genotyping assays were designed using PrimerQuest software (Integrated DNA Technologies). 5′ nuclease assays were created such that one primer sits 5′ to the junction site, and one reverse primer sits 3′ to the probe. A wild-type probe is designed to sit entirely on uninterrupted wild-type sequence 3′ to the forward primer and a mutant probe is created to sit 3′ of the junction site, completely in the transgene. Primer and probe sequences are provided in Supplementary Table S2. A total of 5 μl reactions were constructed as follows: 1 μl (∼20–100 ng) gDNA, 2.5 μl of 2× Type-It Fast SNP Probe PCR Master Mix (Qiagen) and 1.5 μl of primers and probes with a final reaction concentration of 500 nM each primer and 250 nM each probe. Assembled reactions were processed on a CFX384 Touch Real-Time PCR System (Bio-Rad) under the following cycling conditions: initial denaturation at 95°C for 8 min followed by 30 cycles of 95°C for 10 s, 60°C for 60 s and 72°C for 15 s. C_q values were calculated using CFX Manager software (Bio-Rad). Relative fluorescence units values were averaged for duplicate samples and plotted using Prism 6 Software (GraphPad) with Cre TG probe signal on the X-axis and WT signal on the Y-axis.

Data access

Fastq files generated using TLA technology have been deposited at the NCBI SRA (Sequence Read Archive) under accession number PRJNA343439 (http://www.ncbi.nlm.nih.gov/sra/?term=PRJNA343439).

RESULTS

Copy number analysis of seven Cre/CreERT2 lines

In the absence of detailed information on the transgene insertion site in our transgenic lines, we routinely genotype and determine zygosity for Cre transgenic mice by CNV analysis. This is done by quantitative PCR (qPCR) using Cre-specific PCR primers and a Cre probe (see Supplementary Table S1). Using a calibrator with a known number of Cre copies, the copy number value is calculated using the standard qPCR formula, $2^{-\Delta\Delta C_q}$ (28) (see 'Materials and
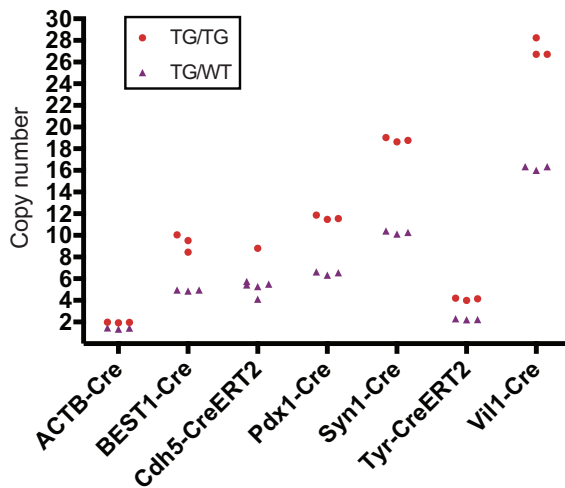
Figure 1. Copy number variation (CNV) analysis for all seven Cre/CreERT2 lines described in this study. Y-axis show calculated copy number by comparison to a homozygous (2 copies) Cre knock-in mouse control. Animals carrying the Cre transgene were first identified and subsequently CNV was performed to determine allelic status. Each data point represents one animal. C_q values used to calculate ΔC_q, $\Delta\Delta C_q$ and copy number were mean of values from 3–4 technical replicates. TG/TG: homozygous transgenic. TG/WT: hemizygous transgenic.

Methods' section for details). An example of CNV genotyping data for all seven Cre/CreERT2 lines used in this study is shown in Figure 1. Although CNV analysis is separate from TLA, and not necessary for accurate TLA analysis, the calculated transgene (hemizygous) copy number for each of the seven lines is listed in Table 1 for easy reference. This method is reliable but time-consuming due to the amount of manual genotype-calling needed after obtaining the qPCR data and thus we were interested in obtaining integration-site specific data for each line to be able to generate genotyping assays that can directly distinguish wild-type, hemizygous and homozygous animals, enabling full automation of the genotyping process. Added benefits to knowing the integration site would include information of proximity to nearby genes that might be affected by the transgene, and information about structural changes at the integration site.

Identification of transgene integration sites

In order to identify the exact transgene integration sites, we performed TLA on splenocytes from an individual animal from each line. An outline of the TLA process is provided in Figure 2 (please see (18) and 'Materials and Methods' section for details). We mapped the sequence data to the mouse, rat and human genomes and analyzed the results to identify integration sites in the mouse genome and to confirm, or identify, the different components of the transgenes (e.g. a rat-derived promoter or the human estrogen receptor sequences found in CreERT2 lines). TLA provides deep sequence coverage of genomic regions in close proximity to the location of the primer set and very low sequence coverage of the rest of the genome (Supplementary Figure S1). This allows for identification of the exact transgene integration site with high confidence and for detailed characterization of any structural changes at the integration site. After
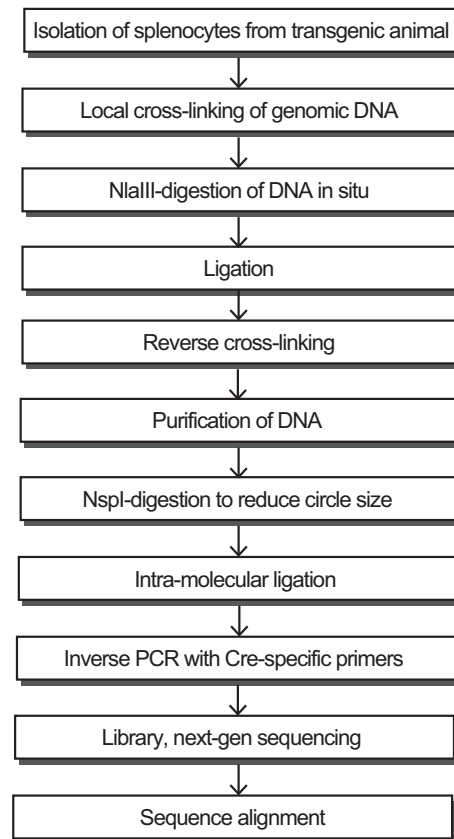

Figure 2. Flow diagram of the TLA process. For details, see the 'Materials and Methods' section.

mapping of the transgene integration sites, the exact transition sequences sites were subsequently confirmed by PCR followed by sequencing using genomic DNA from transgenic mice and primers designed to span the identified junction. The genomic/transgene sequence for one or both of the two junctions for each transgenic line is provided in Supplementary Figure S2, and transgene integration site data for each of the seven transgenic Cre lines is summarized in Figure 3 and Table 1. The ACTB-Cre mouse line contains a transgene consisting of human *ACTB* (beta-actin) promoter, the *Cre* cDNA and human *ACTB* 3′UTR and polyA signal (19). TLA analysis mapped the transgene insertion site to mouse chromosome 1 in intron 4 of *Tmem163*. Our analysis also detected three intronic deletions of 2.2, 7.3 and 3.2 kb, respectively, surrounding the transgene integration site.

The BEST1-Cre transgene contains about 500 bp of the human *BEST1* promoter, *Cre* cDNA, Simian Virus 40 (SV40) exons/introns and Herpes Simplex Virus Thymidine Kinase (HSV TK) polyA (20). TLA analysis mapped the transgene insertion site to mouse chromosome 9, about 58 kb upstream of *Tm108*. Our analysis also detected a 120 bp deletion at the transgene integration site.

The Cdh5-CreERT2 line carries a transgene consisting of a 3 kb mouse *Cdh5* (*VE-cadherin*) promoter, a rabbit beta-globin intron, *CreERT2* and SV40 intron/exon plus polyA (21). TLA analysis maps the integration site to chromosome
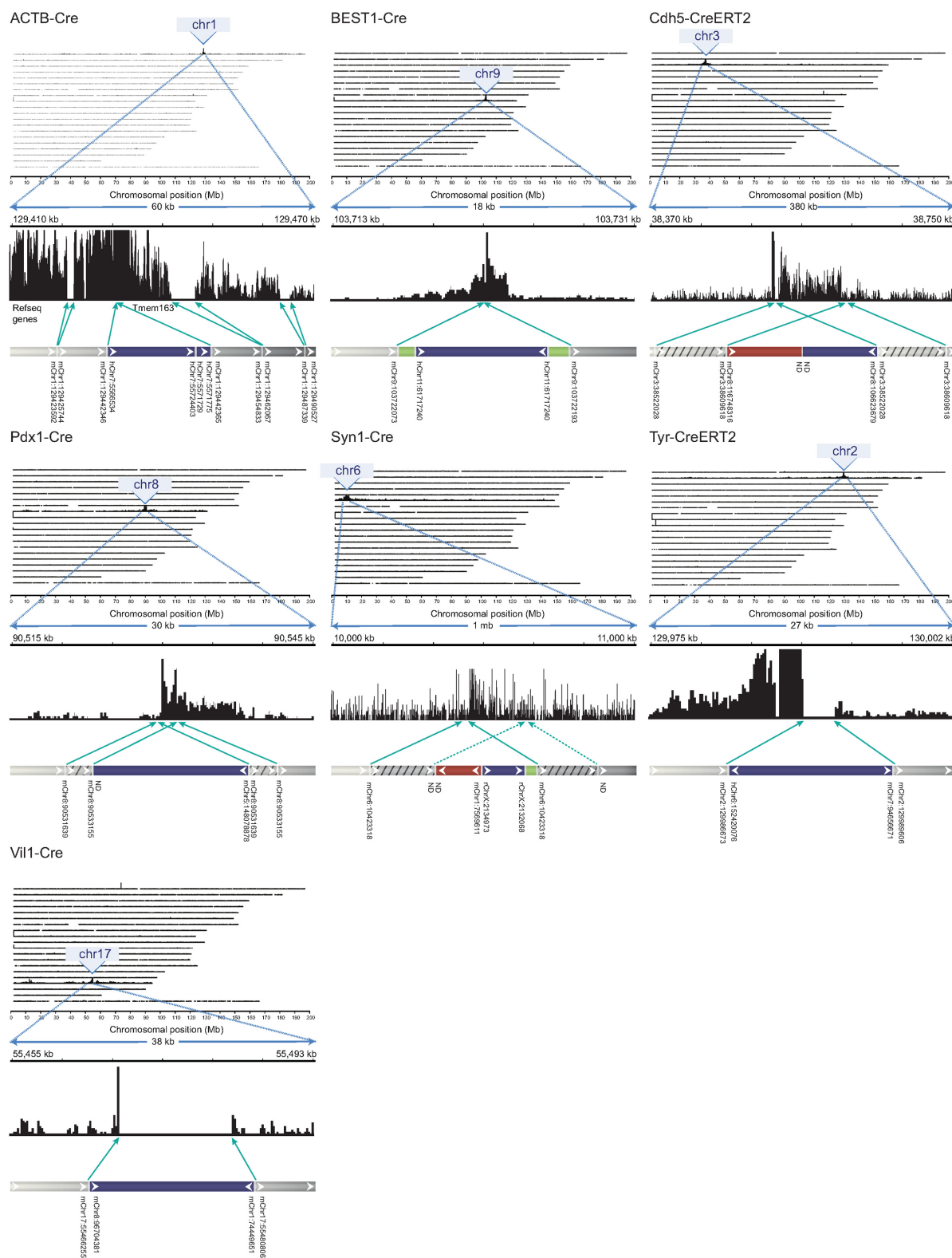
Figure 3. TLA coverage and analysis plots. Upper panels, TLA sequence coverage: mouse chromosomes 1 through X are arranged on the Y-axis. X-axis shows chromosomal position. A detailed view of TLA sequence coverage surrounding the integration site is expanded in the middle panels. Lower panels: graphic representation of transgene integration site and structural changes. Gray: flanking genomic sequence. Blue: transgene and corresponding genomic coordinates of the transgene sequence (mouse, human or rat genome). Only one transgene copy is shown for simplicity. Green: DNA sequences that cannot be mapped to a genome (e.g. vector sequence). Red: a 40 kb mouse *Vat1l* sequence co-integrated with the Cdh5-CreERT2 transgene and a 142 kb mouse Chr1 sequence co-integrated with the Syn1-Cre transgene. Duplicated regions are highlighted by hatching. For Syn1-Cre, one end of the duplicated region is unknown (indicated by a dotted line). ND: no data. Mouse genome assembly: mm9; human genome assembly: hg19. Rat genome assembly: rn5.

Table 1. Summary of Cre transgene integration site data.

Transgenic line	Reference	Position	Nearest gene(s)	Structural variation	Copy number
ACTB-Cre	Lewandoski *et al.* (19)	chr1:129,442,346	*Tmem163*	deletion	1
BEST1-Cre	Iacovelli *et al.* (20)	chr9:103,722,073	*Tmem108*	deletion	5
Cdh5-CreERT2	Monvoisin *et al.* (21)	chr3:38,522,028	*Ankrd50*	duplication, insertion	5
Pdx1-Cre	Hingorani *et al.* (22)	chr8:90,531,639	*Zfp423*	duplication	6
Syn1-Cre	Zhu *et al.* (23)	chr6:10,423,318	n.a[a]	duplication, insertion	10
Tyr-CreERT2	Bosenberg *et al.* (24)	chr2:129,989,606	*Tgm6, Snrpb*	deletion	2
Vil1-Cre	Madison *et al.* (25)	chr17:55,466,255	*St6gal2*	deletion	16

[a]n.a.: Not applicable; nearest gene is >1 Mb away from integration site. Reference genome: mouse genome assembly NCBI37/mm9. Determination of hemizygous transgene copy number was calculated using quantitative PCR data from Figure 1 and was not part of the TLA analysis.

3, about 138 kb upstream of *Ankrd50*. Our analysis also detected a fragment (about 40 kb) from the mouse *Vat1l* gene. As this sequence is not part of the transgene itself, it's possible that the fragment co-integrated with the transgene (see 'Discussion' section). In addition, our analysis indicates that a 87.6 kb genomic duplication occured at the integration site. This duplication unfortunately prevents design of a genotyping assay that can distinguish hemizygous from homozygous mice as the wild-type probe will be able to recognize both the unmodified wild-type allele and the transgene allele. In other words, true homozygous transgenic mice would be falsely typed as hemizygous due to the extra signal from the wild-type probe (see Supplementary Figure S3).

The only description of the Pdx1-Cre transgene in the original publication is that *Cre* is driven by the mouse *Pdx1* promoter (22). By mapping the sequences to both the human and mouse genome, our TLA analysis found that the Pdx1-Cre transgene in addition to the mouse *Pdx1* promoter also contains human *ACTB* 3′UTR and polyA signal. TLA analysis mapped the transgene insertion site to mouse chromosome 8, about 48 kb upstream of *Zfp423*. Our analysis shows that a 1.5 kb genomic duplication occurred at the integration site. As for Cdh5-CreERT2, this duplication prevents design of a genotyping assay that can distinguish hemizygous from homozygous mice.

The Syn1-Cre transgene contains about 4 kb of the rat *Syn1* promoter, 100 bp of chloramphenicol acetyltransferase vector/UTR sequence, *Cre* and the human growth hormone gene (exons, introns, polyA) (23). TLA analysis mapped the transgene insertion site to mouse chromosome 6. For this insertion site, the nearest gene is located more than 1 Mb away. Our analysis identified a duplication of unknown size of the genomic region at the integration site. As for Cdh5-CreERT2 and Pdx-Cre, this duplication prevents design of a genotyping assay that can distinguish hemizygous from homozygous mice. Furthermore, we also found evidence of the co-integration of a 142 Kb region from chromosome 1 along with the transgene. For this transgene we were not able to identify more than one of the two transgene/genomic transitions. A more detailed view of the TLA analysis for Syn1-Cre is provided in Supplementary Figure S1.

The Tyr-CreERT2 line carries a transgene composed of a 3.6 kb mouse *Tyr* enhancer element fused to a 5.5 kb mouse *Tyr* promoter element. The two elements map 12 kb apart on the mouse genome. Between the promoter and *CreERT2* cDNA is a rabbit beta-globin intronic sequence and the

CreERT2 cDNA is followed by the SV40 polyA (24). TLA analysis maps the integration site to chromosome 2, about 7.8 kb downstream of *Snrpb* and 9.6 kb downstream of *Tgm6*. Our analysis also detected a 3 kb genomic deletion at the integration site.

Finally, The Vil1-Cre line carries 12.4 kb of the mouse *Vil1* promoter fused to *Cre*, followed by exons, introns and polyA from the mouse *Mt1* gene (25). TLA analysis maps the integration site to chromosome 17, about 119 kb upstream of *St6gal2*. Our analysis also detected a 14.6 kb genomic deletion at the transgene integration site.

Development of real-time PCR assays based on transgene integration site data

With transgene insertion site sequences available, we next designed genotyping assays for the four lines that did not have genomic duplications surrounding the integration site (ACTB-Cre, BEST1-Cre, Tyr-CreERT2 and Vil1-Cre) to enable high-throughput and fully-automated genotyping using real-time PCR. A genomic duplication flanking the integration site does not prevent the design of a transgene-specific assay. However, the assay for the wild-type allele will also produce a signal from the transgene allele, thus preventing accurate genotyping (Supplementary Figure S3). With the integration site sequences available from this study (Supplementary Figure S2), it is also possible and easy to design three-primer assays for regular PCR. In this type of genotyping assay, a forward genomic primer is shared and reverse primers are designed specific to either the transgene or the wild-type allele, making sure the two kinds of amplicons have different sizes. All three primers are included in the PCR reaction and the three possible genotypes (wild-type, hemizygous and homozygous transgenic) can easily be identified by agarose gel separation. This type of assay, commonly used in smaller mouse facilities, is not practical for high-throughput genotyping, however. As described in the 'Materials and Methods' section and summarized in Supplementary Table S2, we therefore developed a set of primer pairs and matching probes for real-time PCR, specific for each Cre/CreERT2 transgene insertion site and their corresponding wild-type alleles. The assays were then tested using tail-derived genomic DNA. Genotyping results for the four Cre/CreERT2 lines are shown in Figure 4 (A, B, C, D, respectively). For all assays, a clear separation of wild-type, hemizygous and homozygous transgenic animals was obtained, illustrating the robustness of the genotyping as-
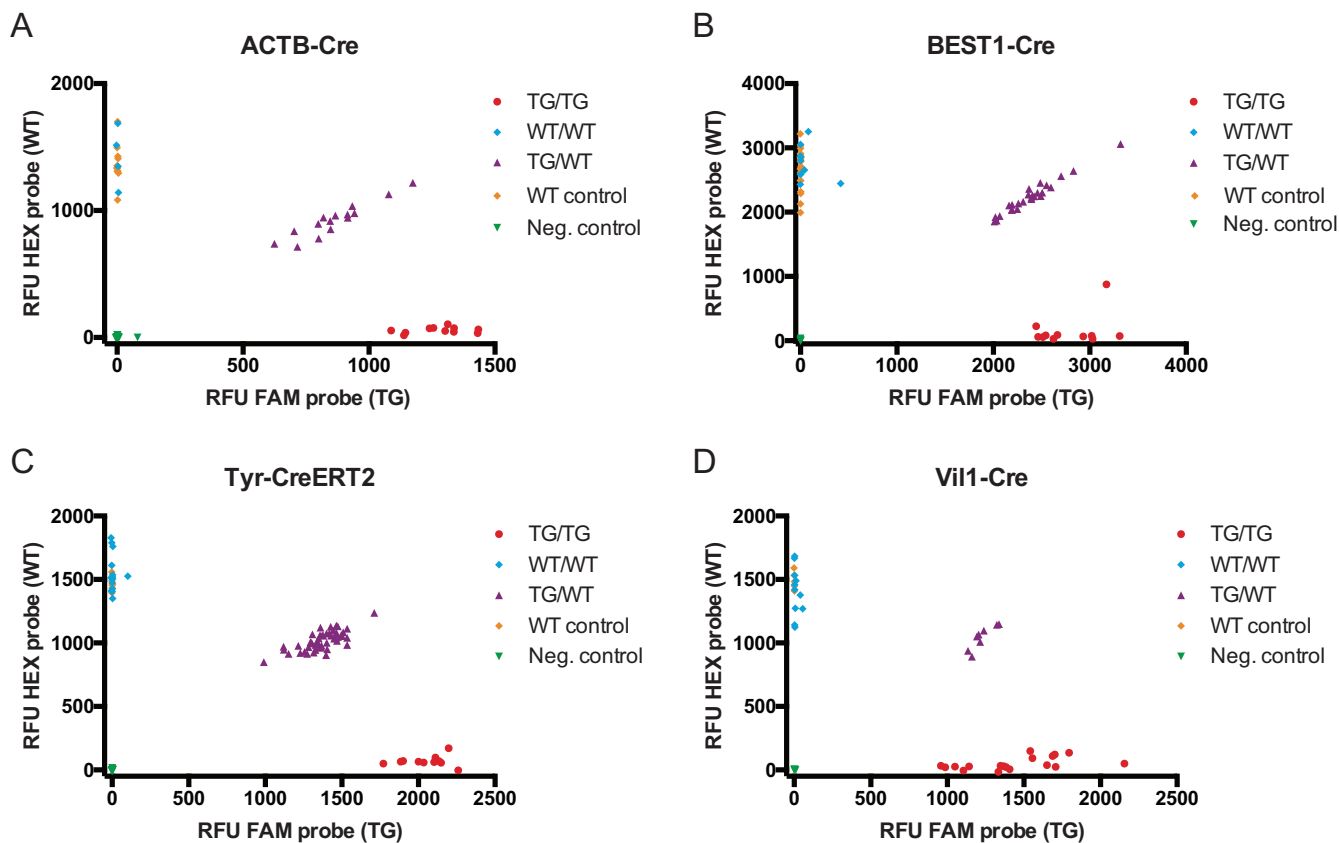
Figure 4. Real-time PCR genotyping plots. X-axis show value of FAM probe signal (transgene specific probe). Y-axis show value of HEX probe signal (wild-type specific probe). RFU: relative fluorescence units. Each data point represents one animal. C_q values are means of values from two technical replicates. For each plot the tested genotypes are indicated in the legend. TG/TG: homozygous transgenic. WT/WT: wild-type. TG/WT: hemizygous transgenic. WT control: C57BL/6N. Neg. control: no template control.

says. The genotype calls were all confirmed by our standard CNV genotyping assay (data not shown).

Effect of transgene insertion on nearby genes

For the ACTB-Cre line, we mapped the insertion site to intron 4 of *Tmem163* (Figure 3 and Table 1). Our analysis also detected three intronic deletions. It is therefore possible that the normal function of *Tmem163* is disrupted in this line. However, we routinely obtain homozygous transgenic mice, so *Tmem163* is either not necessary for normal development or fertility, or the transgene does not significantly alter the normal regulation and expression of *Tmem163*. A *Tmem163* knock-out mouse has to our knowledge not been reported, but a plausible role for *Tmem163* in the regulation of intracellular zinc levels has been suggested (29). It can therefore not be ruled out that ACTB-Cre mice display a Zinc homeostasis phenotype, and this should be taken into account when using the ACTB-Cre line together with alleles of genes relevant for Zinc metabolism. Table 1 lists nearby genes for the other Cre transgenic lines as well. Although none of the other integration sites are located within genes, some transgenes are located in the vicinity of genes and we cannot exclude that transgene integrations influence the expression of these nearby genes. However, since we have been able to generate homozygous mice from all seven lines, it can

be concluded that none of these Cre/CreERT2 transgenes disrupt genes essential for survival and development.

DISCUSSION

Compared to other methods for identifying transgene insertion sites, including other NGS methods, TLA uniquely enables targeted and complete NGS sequencing of transgenes and integration sites and the detection of single nucleotide variants and structural variants within the transgene sequence and in sequences surrounding the integration site. We have demonstrated here that even with limited knowledge of the transgene sequences (limited to the method sections of the papers where the Cre lines were first described), it is possible to generate a complete or near-complete picture of the transgene integration event and provide exact genomic/transgenic borders. The transgenes described in this study all contain the *Cre* cDNA sequence, and we could thus use the same set of Cre-specific primers to perform TLA analysis for all the seven Cre or CreERT2 lines. We are currently using this approach to map the insertion site in additional transgenic Cre lines in our facility but since only minimal information is needed about a given transgene for successful mapping, TLA is a highly efficient method that should be useful for mapping the insertion site of any transgene.

Importantly, our analysis identified structural changes at the site of transgene integration in all seven lines. Three of the lines (Cdh5-CreERT2, Pdx1-Cre and Syn1-Cre) had genomic duplications surrounding the transgene concatemer. These duplications prevent the design of genotyping assays that can distinguish wild-type, hemizygous and homozygous alleles since all the samples will have a positive signal from the wild-type probe/amplicon (Supplementary Figure S3). Four lines had small or large deletions and we identified genomic DNA fragments from another chromosome integrating along with the transgene in two lines. Our observation of structural variations at the integration site is in agreement with a chromothripsis model for transgene insertion (30,31). While the mechanism for transgenesis by random insertion is still largely unknown, it is possible that the linear transgene is simply integrated at a chromosomal site already undergoing active DNA break repair. Alternatively, by an unknown mechanism, the action of injecting linear DNA into the pronucleus of a mouse zygote could itself stimulate chromothripsis followed by DNA repair and insertion of the transgene, accompanied by other structural changes (insertions, deletions, inversions). This model suggests that only when repair by the cell's DNA repair machinery is possible in a manner compatible with survival will a transgenic founder be the result. If DNA repair is not possible, due to extensive chromosomal damage, the early embryo will not survive. To our knowledge there are only a few reports in the literature (30,32) with detailed analysis of transgene integration events. Going forward, TLA should be an excellent tool for this type of analysis.

In conclusion, we have demonstrated the power of TLA not only for precisely mapping the integration site of several Cre transgenes, but also for describing the nature of the structural changes that often accompany transgene insertions. Detailed knowledge of the transgene integration event not only allows for improved genotyping assays, it also helps inform the interpretation of phenotypes obtained when using mice, or any other model organism, carrying transgenic alleles. Finally, for crosses involving conditional alleles, knowing the exact location of the Cre transgene is a clear advantage and since the use of transgenic Cre/CreERT2 lines is widespread, our results and methods should serve as a useful reference for the mouse community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Sauer,B. and Henderson,N. (1988) Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 5166–5170.
2. Gu,H., Marth,J.D., Orban,P.C., Mossmann,H. and Rajewsky,K. (1994) Deletion of a DNA polymerase beta gene segment in T cells using cell type-specific gene targeting. *Science*, **265**, 103–106.
3. Feil,R., Wagner,J., Metzger,D. and Chambon,P. (1997) Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem. Biophys. Res. Commun.*, **237**, 752–757.
4. Mukai,H.Y., Motohashi,H., Ohneda,O., Suzuki,N., Nagano,M. and Yamamoto,M. (2006) Transgene insertion in proximity to the C-myb gene disrupts erythroid-megakaryocytic lineage bifurcation. *Mol. Cell. Biol.*, **26**, 7953–7965.
5. Durkin,M.E., Keck-Waggoner,C.L., Popescu,N.C. and Thorgeirsson,S.S. (2001) Integration of a C-myc transgene results in disruption of the mouse Gtf2ird1 gene, the homologue of the human GTF2IRD1 gene hemizygously deleted in williams–beuren syndrome. *Genomics*, **73**, 20–27.
6. Vogt,T.F., Jackson-Grusby,L., Wynshaw-Boris,A.J., Chan,D.C. and Leder,P. (1992) The same genomic region is disrupted in two transgene-induced limb deformity alleles. *Mamm. Genome*, **3**, 431–437.
7. Bishop,J.O. and Smith,P. (1989) Mechanism of chromosomal integration of microinjected DNA. *Mol. Biol. Med.*, **6**, 283–298.
8. Cain-Hom,C., Pabalate,R., Pham,A., Patel,H.N., Wiler,R. and Cox,J.C. (2016) Mammalian genotyping using acoustic droplet ejection for enhanced data reproducibility, superior throughput, and minimized cross-contamination. *J. Lab. Autom.*, **21**, 37–48.
9. Yuan,J.S., Burris,J., Stewart,N.R., Mentewab,A. and Stewart,C.N. (2007) Statistical tools for transgene copy number estimation based on real-time PCR. *BMC Bioinformatics*, **8**(suppl 7), S6.
10. Liang,Z., Breman,A.M., Grimes,B.R. and Rosen,E.D. (2008) Identifying and genotyping transgene integration loci. *Transgenic Res.*, **17**, 979–983.
11. Uemura,S., Nagaoka,T., Yokoyama,M., Igarashi,M. and Kishi,M. (2014) A simple and highly efficient method to identify the integration site of a transgene in the animal genome. *Neurosci. Res.*, **80**, 91–94.
12. Rosenthal,A. (1992) PCR amplification techniques for chromosome walking. *Trends Biotechnol.*, **10**, 44–48.
13. Uren,A.G., Mikkers,H., Kool,J., van Der Weyden,L., Lund,A.H., Wilson,C.H., Rance,R., Jonkers,J., van Lohuizen,M., Berns,A. *et al.* (2009) A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc.*, **4**, 789–798.
14. Brett,B.T., Berquam-Vrieze,K.E., Nannapaneni,K., Huang,J., Scheetz,T.E. and Dupuy,A.J. (2011) Novel molecular and computational methods improve the accuracy of insertion site analysis in sleeping beauty-induced tumors. *PLoS One*, **6**, e24668.
15. Ji,Y., Abrams,N., Zhu,W., Salinas,E., Yu,Z., Palmer,D.C., Jailwala,P., Franco,Z., Roychoudhuri,R., Stahlberg,E. *et al.* (2014) Identification of the genomic insertion site of Pmel-1 TCR α and β transgenes by next-generation sequencing. *PLoS One*, **9**, e96650.
16. Srivastava,A., Philip,V.M., Greenstein,I., Rowe,L.B., Barter,M., Lutz,C. and Reinholdt,L.G. (2014) Discovery of transgene insertion sites by high throughput sequencing of mate pair libraries. *BMC Genomics*, **15**, 367.
17. DuBose,A.J., Lichtenstein,S.T., Narisu,N., Bonnycastle,L.L., Swift,A.J., Chines,P.S. and Collins,F.S. (2013) Use of microarray hybrid capture and next-generation sequencing to identify the anatomy of a transgene. *Nucleic Acids Res.*, **41**, e70.
18. de Vree,P.J.P., de Wit,E., Yilmaz,M., van de Heijning,M., Klous,P., Verstegen,M.J.A.M., Wan,Y., Teunissen,H., Krijger,P.H.L., Geeven,G. *et al.* (2014) Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.*, **32**, 1019–1025.
19. Lewandoski,M., Meyers,E.N. and Martin,G.R. (1997) Analysis of Fgf8 gene function in vertebrate development. *Cold Spring Harb. Symp. Quant. Biol.*, **62**, 159–168.
20. Iacovelli,J., Zhao,C., Wolkow,N., Veldman,P., Gollomp,K., Ojha,P., Lukinova,N., King,A., Feiner,L., Esumi,N. *et al.* (2011) Generation of Cre transgenic mice with postnatal RPE-specific ocular expression. *Invest. Ophthalmol. Vis. Sci.*, **52**, 1378–1383.

21. Monvoisin,A., Alva,J.A., Hofmann,J.J., Zovein,A.C., Lane,T.F. and Iruela-Arispe,M.L. (2006) VE-cadherin-CreERT2 transgenic mouse: a model for inducible recombination in the endothelium. *Dev. Dyn.*, **235**, 3413–3422.

22. Hingorani,S.R., Petricoin,E.F., Maitra,A., Rajapakse,V., King,C., Jacobetz,M.A., Ross,S., Conrads,T.P., Veenstra,T.D., Hitt,B.A. *et al.* (2003) Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell*, **4**, 437–450.

23. Zhu,Y., Romero,M.I., Ghosh,P., Ye,Z., Charnay,P., Rushing,E.J., Marth,J.D. and Parada,L.F. (2001) Ablation of NF1 function in neurons induces abnormal development of cerebral cortex and reactive gliosis in the brain. *Genes Dev.*, **15**, 859–876.

24. Bosenberg,M., Muthusamy,V., Curley,D.P., Wang,Z., Hobbs,C., Nelson,B., Nogueira,C., Horner,J.W., DePinho,R. and Chin,L. (2006) Characterization of melanocyte-specific inducible Cre recombinase transgenic mice. *Genesis*, **44**, 262–267.

25. Madison,B.B., Dunbar,L., Qiao,X.T., Braunstein,K., Braunstein,E. and Gumucio,D.L. (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.

26. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, **26**, 589–595.

27. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

28. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2-\Delta\Delta CT$ method. *Methods*, **25**, 402–408.

29. Cuajungco,M.P., Basilio,L.C., Silva,J., Hart,T., Tringali,J., Chen,C.-C., Biel,M. and Grimm,C. (2014) Cellular zinc levels are modulated by TRPML1-TMEM163 interaction. *Traffic*, **15**, 1247–1265.

30. Chiang,C., Jacobsen,J.C., Ernst,C., Hanscom,C., Heilbut,A., Blumenthal,I., Mills,R.E., Kirby,A., Lindgren,A.M., Rudiger,S.R. *et al.* (2012) Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.*, **44**, 390–397.

31. Stephens,P.J., Greenman,C.D., Fu,B., Yang,F., Bignell,G.R., Mudie,L.J., Pleasance,E.D., Lau,K.W., Beare,D., Stebbings,L.A. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.

32. Würtele,H., Little,K.C.E. and Chartrand,P. (2003) Illegitimate DNA integration in mammalian cells. *Gene Ther.*, **10**, 1791–1799.