Identification of structural variation in mouse genomes

Thomas M. Keane1*, Kim Wong1, David J. Adams1, Jonathan Flint2, Alexandre Reymond3 and Binnaz Yalcın3,4*

1 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
2 Wellcome Trust Centre for Human Genetics, Oxford, UK
3 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
4 Institute of Genetics and Molecular and Cellular Biology, Illkirch, France

Edited by:
Frederic J. J. Chain, McGill University, Canada

Reviewed by:
Gabriel Da Rocha Fernandes, Universidade Católica de Brasília, Brazil
Omer Gokcumen, University at Buffalo, USA
Hindrik Kerstens, Radboud University Nijmegen, Netherlands

*Correspondence:
Thomas M. Keane, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK
e-mail: thomas.keane@sanger.ac.uk;
Binnaz Yalcın, Center for Integrative Genomics, University of Lausanne, Lausanne, CH-1015, Switzerland
e-mail: binnaz.yalcin@unit.ch

Structural variation is variable in structure of DNA regions affecting DNA sequence length and/or orientation. It generally includes deletions, insertions, copy-number gains, inversions, and transposable elements. Traditionally, the identification of structural variation in genomes has been challenging. However, with the recent advances in high-throughput DNA sequencing and paired-end mapping (PEM) methods, the ability to identify structural variation and their respective association to human diseases has improved considerably. In this review, we describe our current knowledge of structural variation in the mouse, one of the prime model systems for studying human diseases and mammalian biology. We further present the evolutionary implications of structural variation on transposable elements. We conclude with future directions on the study of structural variation in mouse genomes that will increase our understanding of molecular architecture and functional consequences of structural variation.

Keywords: array comparative genome hybridization (aCGH), next-generation sequencing (NGS), structural variation (SV), paired-end mapping (PEM), inbred strains of mice, Heterogeneous Stock (HS), Sanger Mouse Genomes Project

INTRODUCTION

Structural variation (SV) is generally considered as rearrangements of DNA regions affecting DNA sequence length and/or orientation in the genome of one species, and includes deletions, insertions, copy-number gains, inversions, and transposable elements. Structural variation has long been known to be pathogenic, resulting in rare genomic disorders such as those associated with mutations in transposable elements. We conclude with future directions on the study of structural variation in mouse genomes that will increase our understanding of molecular architecture and functional consequences of structural variation.

Table 1

| SV Type                  | Characteristics                      | Examples                  |
|--------------------------|--------------------------------------|---------------------------|
| Deletions                | Loss of DNA sequence                | 5q31 deletion syndrome    |
| Insertions               | Gain of DNA sequence                | 7p11.2 micro-deletion      |
| Copy-number gains        | Duplication of DNA sequence         | 16p11.2 micro-deletion     |
| Inversions               | Rearrangement of DNA sequence       | 17p13.1 micro-deletion     |
| Neuronal mosaicism       | Multiple structural variants         | 18p11.2 micro-deletion     |

SVs have traditionally been observed by array comparative genome hybridization (aCGH), a method for analyzing copy number variations by measuring fluorescence between two differentially labeled DNA samples (DNA of a test sample compared to a reference sample). Using aCGH, the extent of genome-wide SV in the mouse was first demonstrated in 2007 with the detection of 80 high-confidence copy number variants in 20 inbred strains of mice (Graubert et al., 2007), subsequently followed by other studies, summarized in Table 1 (Cutler et al., 2007; Akagi et al., 2008; Cahan et al., 2009; Henrichsen et al., 2009; Agam et al., 2010; Quinlan et al., 2010). These studies, however, have proven to be difficult to interpret due to their poor reproducibility (Agam et al., 2010) and inability to detect certain types of structural variants. For example inversions and insertions of novel sequence are blind to aCGH technology because inversions do not affect copy number, which is what is detected by aCGH technique, and novel sequence insertions have no copy in the reference genome.

With the emergence of next-generation sequencing (NGS) (Mardis, 2011), the Mouse Genomes Project (http://www.sanger.ac.uk/resources/mouse/genomes/) was able to sequence the entire genomes of 18 classical laboratory strains and wild-derived lines of inbred strains of mice, producing detailed maps of SV and retro-transposon elements in each mouse strain, relative to the reference mouse strain C57BL/6J (Keane et al., 2011; Nellaker et al., 2012).
et al., 2012; Wong et al., 2012; Simon et al., 2013). For the first time, this resulted in the detection of an extraordinarily larger number of structural variants than previously observed using aCGH, totaling 710,000 novel structural variants affecting 1% of the mouse genome and encompassing 10 times more total nucleotides than single nucleotide polymorphisms (Yalcin et al., 2011). As a comparison, we had identified 121 deletions in a previous aCGH study of SV in DBA/2J, with SV length ranging between minimum size of 5 kilobases (Kb) and maximum of 260 Kb (median size 48 Kb) (Agam et al., 2010), whereas in a latest NGS study of SV we found far more deletions (a total of 16,318) in that same strain, of much smaller size (minimum size of 100 bp, maximum of 10 Kb, median of 400 bp) (Figure 1).

Such genome-wide abundance in structural variation has led to several important questions: what is the molecular architecture of these variants, what are the mechanisms of SV formation and how do they impact gene function? In this review, we address these questions and redefine what we have learnt so far about the nature, origins, and role of structural variation from current studies in the mouse. Finally, we discuss the promises of novel methods which are likely to facilitate access to repeat-rich regions and assembly of complex genomic regions, in order to assess the origins and functional impact of structural variation in the most challenging regions of the mouse genome.

**DETECTION OF STRUCTURAL VARIANTS USING PAIRED-END MAPPING METHODS**

While most deep-sequencing applications focus on the identification of single-nucleotide polymorphisms (SNPs) or small insertion deletion polymorphisms, structural variation can also be identified from the same data. However, while the basic types of structural variants (deletions, insertions, inversions, and duplications) can be identified using a combination of computational methods, the detection of complex rearrangements remains challenging. We define complex rearrangements as those structural variants consisting of a combination of basic types that directly about each other or that are nested within each other (e.g., an inversion directly flanked by insertions, or a deletion nested within a tandem duplication).

Typically, genomic DNA of a test genome is sheared into fragments of 300–500 bp to generate a sequencing library. Short paired-reads (50–250 bp) from either extremity of the fragment (called paired-end reads) are sequenced and mapped to the reference genome. Structural variants are then called based on orientation, distance, and depth of the mapped paired-reads (also reviewed in Medvedev et al., 2009; Alkan et al., 2011). Depending on the size and type of structural variant, these methods exploit read pairs (Korbel et al., 2007; Chen et al., 2009), split-reads (Ye et al., 2009; Albers et al., 2011), single end clusters and read depth (Simpson et al., 2010).

The most widely used methods are read pair and read depth methods. Read pair based methods analyze distance and orientation of paired reads to infer deletion, insertion, inversion and tandem duplication events as shown in Figure 2. When the paired-end reads are mapping in the correct orientation (“/+” is normal) but to a distance that is significantly larger than the average fragment length, this suggests a deletion, whereas if the distance is smaller than the fragment length, it suggests an insertion. When the two sequenced ends map back to the genome in the wrong orientation (“+/-” and “-/+”), and at a distance that is significantly larger than the size of the fragment itself, this indicates an inversion. Finally, when paired-end reads map with orientation “-/+” to a large distance, it suggest tandem duplication. In the single-end cluster analysis, one of the paired-end reads maps to the reference while its mate map to the inserted sequence (de novo sequence or repeat element insertion). Read depth methods take advantage of the high coverage of next generation sequencing to infer increase or decrease of reads at a locus. When the coverage is higher than the expected genome coverage, duplication is inferred, whereas when it is smaller or null, deletion is inferred. Once the structural variant is detected using

**Table 1 | Summary of mouse studies reporting genome-wide structural variants.**

| Technique | No. of SVs | No. of strains | References |
|-----------|------------|----------------|------------|
| aCGH      | 80         | 20             | Graubert et al., 2007 |
| aCGH      | 2,094      | 42             | Cutler et al., 2007 |
| WGS       | 10,000     | 4              | Akagi et al., 2008 |
| aCGH      | 1,300      | 20             | Cahan et al., 2009 |
| aCGH      | 7,103      | 33*            | Henrichsen et al., 2009 |
| aCGH      | 7,196      | 1              | Quinlan et al., 2010 |
| aCGH      | 1,976      | 7              | Agam et al., 2010 |
| NGS       | 711,920    | 17             | Yalcin et al., 2011 |
| NGS       | 30,048     | 1              | Wong et al., 2012 |
| NGS       | 43         | 1              | Simon et al., 2013 |

Column 1 gives the technique used in the study (aCGH, array comparative genome hybridization; WGS, whole genome sequencing; NGS, next generation sequencing). Column 2 refers to the total number of structural variants (SVs) identified and column 3, to the number of laboratory inbred mouse strains used in the study at the exception of * that includes 21 wild-caught mice. The reference mouse strain (C57BL/6J) is excluded in the count. Column 4 is the reference to the study.
Keane et al. Structural variation in mouse genomes

**FIGURE 2 | Read mapping patterns used by computational methods to detect basic structural variation from NGS data.** This figure shows the principle of SV identification using (i) read-pair analysis, (ii) split-read mapping, (iii) single end cluster analysis, and (iv) read depth analysis. Deletions and insertions are represented using red rectangles, and inversions and duplications using light blue arrows. Reads are represented using solid dark blue arrows. The first step consists in sequencing a test genome. Typically, the genomic test DNA is fragmented into chunks of 300–500 bp. Then, reads of 50–250 bp are sequenced from either side of each fragment (we call these paired-end reads). The second step consists in mapping these paired-end reads to the mouse reference genome. A rightward facing arrow denotes a positive strand alignment, and leftward a negative strand alignment. (i) In the read-pair analysis approach, when the paired-end reads are mapping in the correct orientation (“+/−” is normal) but to a distance that is significantly larger than the average fragment length. If we suppose this distance to be 1100 bp, it suggests a deletion of 600 bp, whereas if the distance is smaller than the fragment length, for example 200 bp, it suggests an insertion of 300 bp. When the two sequenced ends of two fragments map back to the reference genome in the wrong orientation (“+/+” and “−/−”), and at a distance that is significantly larger than the size of the fragment itself, this indicates an inversion. Finally, when paired-end reads map with orientation “−/+” to a large distance, it suggest tandem duplication. (ii) In the split-read approach, one of the paired-end reads map to the reference genome while its mate contains the structural variant, typically a deletion or an insertion of small length. (iii) In the single-end cluster analysis, one of the paired-end reads maps to the reference while its mate map to the inserted sequence that can be either de novo sequence or repeat element such as LINE, SINE, or ERV. (iv) Finally, the read depth approach takes advantage of the high coverage of next generation sequencing that makes it possible to detect copy number changes. Of note, the coverage drops at insertion and inversion breakpoints, which when combined with paired-end reads analysis makes the SV call highly reliable.

these analyses, breakpoint refinement is typically achieved using local sequence assembly.

Remarkably, in the past several years many algorithms have been developed to discover basic structural variation in paired-end next generation sequencing data. There are over 50 programs to date (Table 2), however none is as yet considered to reach a community standard and only a handful combine multiple methods for the detection of structural variation (Medvedev et al., 2010; Wong et al., 2010; Rausch et al., 2012b; Sindi et al., 2012; Hart et al., 2013). Accurate structural variant calling depends on many factors such as sequencing library biases, read length, uniform sequencing coverage, and proximity of SVs to repeat sequences. Some of the most frequent sequencing library biases that can detrimentally affect SV detection are high PCR duplicates, non-normal fragment size distributions, and uneven representation of the genome at varying levels of GC content. Therefore, false negative rates of most studies remain high (20–30%) compared to SNP calling (<5%). False positive rates are also high and are often caused by misalignment of the short reads and sometimes by reference genome assembly errors.

There is a growing awareness of complex structural variants (Berger et al., 2011; Stephens et al., 2011; Quinlan and Hall, 2012; Yalcin et al., 2012a; Malhotra et al., 2013), however, their genome-wide detection is much more challenging and less intuitive as they often generate ambiguous paired-end mapping patterns. Complex structural variants are very often completely or partially missed, or incorrectly classified because a single method on its own might not be sufficient to capture the whole complexity of the structural variant (e.g., an apparent deletion and inversion may be simultaneously part of a tandem duplication region). Thus, it is important to combine multiple methods, something that the community has begun to do. Sindi and colleagues, for example, used an algorithm combining both read pairs and read depth signals into a probabilistic model implemented in a software GASV-PRO that significantly improves detection specificity (Sindi et al., 2012). Rausch and colleagues have developed DELLY that integrates short insert paired-ends,
Table 2 | Algorithms for the detection of structural variation.

| Algorithm | Description | Download | References |
|-----------|-------------|----------|------------|
| BreakDancer | Predicts del, ins, inv, and translocations using PEM. Performance examined in an ind. with acute myeloid leukemia and samples from the 1000 Genomes trio. Compared with VariationHunter and MoDIL | [http://gmt.genome.wustl.edu/breakdancer/current/](http://gmt.genome.wustl.edu/breakdancer/current/) | Chen et al., 2009 |
| CNAnseg | Identifies CNVs from NGS data. Uses depth of coverage to estimate copy number states in cancer and normal samples | [http://www.compbio.group.cam.ac.uk/software.html](http://www.compbio.group.cam.ac.uk/software.html) | Ivakhno et al., 2010 |
| cnD | HMM that uses read coverage to determine genomic copy number. Tested on short read sequence data generated from re-sequencing chr. 17 of the mouse strains A/J and CAST/EiJ with the Illumina platform | [http://www.sanger.ac.uk/resources/software/cnd.html](http://www.sanger.ac.uk/resources/software/cnd.html) | Simpson et al., 2010 |
| cn.MOPS | Mixture Of PoissonS Bayesian approach to detect CNVs. Compared with mrFast, EWT, JointSLM, CNV-Seq, and FREEC using data from a male HapMap individual and high coverage data from the 1000 Genomes Project | [http://www.bioinf.jku.at/software/cnmops](http://www.bioinf.jku.at/software/cnmops) | Klambauer et al., 2012 |
| CNVer | Method that supplements the depth-of-coverage with PEM information, where mate pairs mapping discordantly to the reference serve to indicate the presence of variation | [http://compbio.cs.toronto.edu/cnver](http://compbio.cs.toronto.edu/cnver) | Medvedev et al., 2010 |
| CNVnator | Method for CNV discovery and genotyping from read-depth analysis of personal genome sequencing | [http://sv.gersteinlab.org/cnvnator](http://sv.gersteinlab.org/cnvnator) | Abyzov et al., 2011 |
| CNV-Seq | Method to detect CNV using shotgun sequencing | [http://tiger.dbs.nus.edu.sg/CNV-seq](http://tiger.dbs.nus.edu.sg/CNV-seq) | Xie and Tammi, 2009 |
| CREST | Clipping Reveals Structure, uses NGS reads with partial alignments to a ref. to map SVs at nucleotide level resolution. Used for 5 pediatric acute lymphoblastic leukemias and a human melanoma cell line | [http://www.stjuderesearch.org/site/lab/zhang](http://www.stjuderesearch.org/site/lab/zhang) | Wang et al., 2011 |
| DELLY | Integrates paired-end and split-read analysis | [www.korbel.embl.de/software.html](http://www.korbel.embl.de/software.html) | Rausch et al., 2012b |
| Dindel | Bayesian method to call small indels by realigning reads to candidate haplotypes that represent alternative sequence to the reference, using a split-read approach. Used in the 1000 Genomes Project call sets | [http://www.sanger.ac.uk/resources/software/dindel](http://www.sanger.ac.uk/resources/software/dindel) | Albers et al., 2011 |
| EWT | Event-wise testing, method based on significance testing. Error rate tested using the analysis of chromosome 1 from paired-end shotgun sequence data (30×) on 5 individuals | [http://dxplorer.sourceforge.net](http://dxplorer.sourceforge.net) | Yoon et al., 2009 |
| FREEC | Control-FREE Copy number caller that automatically normalizes and segments copy number profiles | [http://bioinfo-out.curie.fr/projects/freec](http://bioinfo-out.curie.fr/projects/freec) | Boeva et al., 2011 |
| GASV-PRO | Combines both paired read and read depth signals into a probabilistic model for greater specificity | [http://compbio.cs.brown.edu/software](http://compbio.cs.brown.edu/software) | Sindi et al., 2012 |
| GenomeSTRiP | Genome STRucture In Populations, toolkit for discovering and genotyping structural variations using sequencing data. Twenty to thirty genomes required to get good results | [http://www.broadinstitute.org/software/genomestrip/download-genomestrip](http://www.broadinstitute.org/software/genomestrip/download-genomestrip) | Handsaker et al., 2011 |
| HYDRA | Localizes SV breakpoints by PEM. Uses a similar clustering strategy to VariationHunter. Accuracy evaluated using VWS split-read mappings. Maps repetitive elements such as transposons and SD | [http://code.google.com/p/hydra-sv](http://code.google.com/p/hydra-sv) | Quinlan et al., 2010 |
| inGAP-sv | Scheme that uses abnormally mapped read pairs. Possible to distinguish HOM and HET variants. Compared with VariationHunter, Breakdancer, PEMer, Spanner, Cortex, and Pindel | [http://ingap.sourceforge.net](http://ingap.sourceforge.net) | Qi and Zhao, 2011 |
| JointSLM | Allows to detect common CNVs among individuals using depth of coverage | [http://www.mybiosoftware.com/population-genetics/11185](http://www.mybiosoftware.com/population-genetics/11185) | Magi et al., 2011 |
| MoDIL | Detection of small indels from clone-end sequencing with mixtures of distributions | [http://compbio.cs.toronto.edu/modil](http://compbio.cs.toronto.edu/modil) | Lee et al., 2009 |
| mrFast | Allows for the prediction of absolute copy-number variation of duplicated segments and genes | [http://mrfast.sourceforge.net](http://mrfast.sourceforge.net) | Alkan et al., 2009 |
| PEMer | Compatible with several NGS platforms. Simulation-based error models, yielding confidence-values for each SV | [http://sv.gersteinlab.org/pemer](http://sv.gersteinlab.org/pemer) | Korbel et al., 2009 |
long-range mate-pairs and split-read alignments to accurately delineate genomic rearrangements at single-nucleotide resolution (Rausch et al., 2012b). In our studies, we used SVMerge (Wong et al., 2010), a pipeline that integrates structural variation calls from five existing software, and validates breakpoints using local de novo assembly.

Unbiased exploration of next-generation sequencing data is laborious, however it is essential for deciphering the true complex nature of structural variants. Toward this goal, we visualized read mappings to the whole of mouse chromosome 19 as well as a random set of regions on other chromosomes using the short-read visualization tool LookSeq (Manske and Kwiatkowski, 2009) in 17 inbred strains of laboratory mice (Yalcin et al., 2012a) as well as in C57BL/6J mice (Simon et al., 2013). We were able to recognize classical paired-end mapping (PEM) patterns, but unexpectedly we were also able to detect a number of other patterns, of greater diversity and complexity that would have been missed or miscalled by existing computational SV detection methods. When two (or more) structural variants co-locate at a locus in the genome (right next to each other), or when one or more structural variants are embedded within another one of larger size (nested), it creates confusing paired-end mapping patterns and incoherent read depth. Figure 3 highlights some complex rearrangements that cause conflicting signals during automatic detection. For example, a deletion directly flanked by a large insertion is characterized by null read depth as expected, however paired reads supporting the deletion are missing because of the insertion. However, we showed that it is possible to train genome-wide computational analysis to detect most of these atypical patterns using integration of multiple detection methods (Wong et al., 2010).

In conclusion, to study the whole diversity and complexity of structural variants, future algorithms need to integrate multiple signals and sequence analyses features based on what we have learnt so far about the architecture of structural variants, while visual approaches will continue to increase our understanding of complex forms of structural variants such as inversions and translocations that remain to be fully resolved. It is important to gain better sensitivity and specificity in the identification of structural variants especially those that have complex architecture to study accurately their impact on diseases such as tumor heterogeneity (Russnes et al., 2011), and on the evolution of genomes.
FIGURE 3 | Complex rearrangements in mouse genomes. We highlight three examples of complex rearrangements that cause ambiguous signals during their detection (for a full list of complex rearrangements see Yalcin et al., 2012a): (A), a deletion directly flanked by an insertion; (B), an inversion directly flanked by two deletions; and (C), an inversion directly flanked by an insertion. For each complex rearrangements, we provide: (1) a drawing of the paired-end mapping (PEM) pattern, (2) an illustration using the short read visualization tool LookSeq (Manske and Kwiatkowski, 2009), and (3) PCR validation. We draw paired-end reads (black arrows) and how they map to the reference genome (dashed gray lines). Green arrows represent primer pairs used for PCR validation. PCR amplification was carried out across eight inbred strains of mice (A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, and LP/J), which are the parental strains of the Heterogeneous Stock population (Valdar et al., 2006). Hyperladder II is the size marker. Genomic coordinates refer to the mm9 mouse assembly. (A) Deletion of 836 bp directly flanked by an insertion of 1200 bp on mouse chromosome 19 (chr19: 48,061,057–48,061,892 bp) in mouse strains A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, and LP/J. In LookSeq, the two back arrows show singleton reads suggesting an insertion (their mates are within the inserted sequence). Read depth is null but paired-end reads in support of the deletion are missing because of the insertion. PCR in four strains (A/J, BALB/cJ, DBA/2J, and LP/J) does not show directly the presence of the 836-bp deletion but instead reveals the presence of an insertion of about 400bp that is in fact the size difference between the deletion and the insertion. (B) Inversion of 325 bp on mouse chromosome 5 (chr5: 148,925,249–148,925,573 bp), directly flanked on the left by a deletion of 71 bp (chr5: 148,925,178–148,925,248 bp) and on the right by another deletion of 645 bp (chr5: 148,925,574–148,926,218 bp). In LookSeq, the top arrow shows the PEM pattern of the deletion. Normally, the underlying read depth should be null, however, it is only null at the regions shown by the two bottom arrows. This is caused by an intervening inversion. PCR in four strains (A/J, AKR/J, BALB/cJ, and C3H/HeJ) confirms the presence of the two deletions. (C) An inversion of 548 bp on mouse chromosome 8 (chr8: 77,137,213–77,137,760 bp) directly flanked by an insertion of 400 bp in mouse strain BALB/cJ, C3H/HeJ, CBA/J, and DBA/2J. In LookSeq, the bottom arrows show a dip in the coverage; on the right, it is caused by an insertion and on the left by an inversion. The presence of the insertion results in missing reads (“−”/“−”), supporting the inversion. PCR shows an amplification band of about 1400 bp in BALB/cJ, C3H/HeJ, CBA/J, and DBA/2J, whereas, in the remaining strains, the band is at about 1000 bp. This confirms the insertion of 400 bp in BALB/cJ, C3H/HeJ, CBA/J, and DBA/2J.
FUNCTIONAL IMPACT OF STRUCTURAL VARIANTS

The functional impact of structural variants is still controversial in the literature. On one hand, some studies showed that SNPs are more likely to contribute to individual phenotypic differences than structural variants (Conrad et al., 2010; Keane et al., 2011); on the other hand, several studies have estimated the impact of structural variation using its effect on gene expression, and these estimates ranged from 10 to 74% (Stranger et al., 2007; Cahan et al., 2009; Henrichsen et al., 2009; Yalcin et al., 2011). It has also been reported that structural variation can influence gene expression both spatially and temporally (Chagnat et al., 2011), including genes outside of SV margins (Henrichsen et al., 2009), and can do so through chromatin conformation changes (Gheldof et al., 2013). The influence of structural variation on gene expression is specifically reviewed in Harewood et al. (2012).

Interpreting the phenotypic consequences of structural variation can be done using different methods. In this review, we describe three methods with specific emphasis on genome wide association studies. Genome wide association studies (GWASs) identify genomic loci associated with individual differences (these regions are called Quantitative Trait Loci, QTLs) using large populations of outbred mice, while taking advantage of recombinants that have naturally accumulated during breeding (Valdar et al., 2006; Yalcin et al., 2010). When combined with the availability of full genome sequences, GWASs in outbred mice are providing significant advances into the understanding of the genotype-phenotype relationship (reviewed in Yalcin and Flint, 2012), especially the impact of structural variants on phenotypic differences.

To test causality of a structural variant within a QTL region, Richard Mott and colleagues have developed a statistical test (called merge) to identify genomic variants likely to be functional from those less likely to be functional (Yalcin et al., 2011), including genes outside of SV margins (Henrichsen et al., 2009), and can do so through chromatin conformation changes (Gheldof et al., 2013). The influence of structural variation on gene expression is specifically reviewed in Harewood et al. (2012).

TABLE 3 | Structural variants associated with quantitative traits in outbred mice.

| Chr | Start | Stop | Type | Gene | Region | Quantitative trait |
|-----|-------|------|------|------|--------|-------------------|
| 1   | 175158884 | 175158885 | Ins  | Fcer1a | Upstream | Mean platelet volume |
| 2   | 144402760 | 144402971 | SINE Ins | Sec23b | Intron | OFT total activity |
| 4   | 49690362 | 49690363 | Del | Grin3a | Intron | HP cellular proliferation marker |
| 4   | 108951263 | 108951264 | IAP Ins | Eps15 | Upstream | Home cage activity |
| 4   | 130038388 | 130038389 | SINE Ins | Snpnp40 | Intron | T-cells: %CD3 |
| 7   | 90731819 | 90731820 | IAP Ins | Tmc3 | Upstream | Wound healing |
| 7   | 111397607 | 111479433 | Ins | Trim5 | Exon | Mean cellular hemoglobin |
| 7   | 111504989 | 111505193 | Del | Trim30b | UTR | Mean cellular hemoglobin |
| 8   | 87957244 | 87957245 | LINE Ins | 4921524J17Rik | Upstream | Mean cellular volume |
| 11  | 11506127 | 11506250 | Del | Tnem104 | UTR | Serum urea concentration |
| 13  | 113783196 | 113783359 | Del | Gmr6320 | Upstream | HP cellular proliferation marker |
| 17  | 344838681 | 344838682 | Del | H2-Ea | Upstream | T-cells: CD4/CD8 ratio |

Columns 1, 2, and 3 give positional information about the structural variant (coordinates refer to the mm9 mouse assembly). Column 4 is the type of the variant. Column 5 and 6 give information about the underlying gene. Column 7 is the quantitative trait associated with the structural variant. Ins, insertion; Del, deletion; UTR, untranslated region; SINE, short interspersed nuclear element; LINE, long interspersed nuclear element; IAP, intracisternal A-particle; HP, Hippocampus; OFT, open field test.
Phenotype) terms that coincide with the phenotype differentiating C57BL/6J and C57BL/6N. The first is an intronic LINE insertion found in the intron of Chl1 (Cell adhesion molecule with homology to L1CAM). C57BL/6N mice displayed abnormal spatial memory in the Morris water maze test compared to C57BL/6J mice. Interestingly, knockout mice of Chl1 also show abnormal spatial working memory. The second is an intronic Rptor (Regulatory associated protein of MTOR, complex 1) in C57BL/6J mice. These mice were characterized with decreased fat mass and blood glucose. Knockout mice of Rptor interestingly also showed decreased fat mass and blood glucose amongst other metabolic phenotypes. The third is the ERV insertion in the intron of Soat1 (Sterol O-acyltransferase 1) results in hair interior defects (Wu et al., 2010; Tareen et al., 2009; Wu et al., 2010) and, in some cases, are giving rise to specific phenotype coding exon of Fv1 (Friend-virus-susceptibility-1) is associated with retrovirus replication (Best et al., 1996; Yalcin et al., 2011), and a deletion of 6817 bp on the first exon of Soat1 (Sterol O-acyltransferase 1) results in hair interior defects (Wu et al., 2010; Yalcin et al., 2011).

![Image](https://example.com/image1.png)

**Table 4** | Structural variants differentiating C57BL/6J and C57BL/6N.

| Chr | Start | Stop | Type | Gene | Region |
|-----|-------|------|------|------|--------|
| 2   | 70619835 | 70620080 | SINE Ins | Tik1 | Intron |
| 3   | 60336036 | 60336037 | Del (large) | Mbn1 | Intron |
| 4   | 101954274 | 101954395 | Del | Pde4b | Intron |
| 5   | 116051393 | 116051799 | MalR Ins | Mast2 | Intron |
| 6   | 103669536 | 103676487 | LINE Ins | Chl1 | Intron |
| 7   | 92096990 | 92096149 | Del | Vmn2r65 | Exon |
| 8   | 27636128 | 27748456 | Ins | Cyp2a22 | Entire |
| 9   | 139306094 | 139307981 | MalR Ins | Cpnm | Intron |
| 10  | 16716381 | 16716382 | Del (large) | Csmd1 | Intron |
| 11  | 58544415 | 58546304 | MalR Ins | 2410076121Rik | Intron |
| 12  | 32536420 | 32543464 | LINE Ins | Nkain2 | Intron |
| 13  | 119560391 | 119568827 | MTA Ins | Rptor | Intron |
| 14  | 42023964 | 4202747 | Del | Immpl2 | Intron |
| 15  | 12016268 | 12016426 | Del (large) | Nnt | Intron |
| 16  | 12863187 | 12863188 | Del (1800 bp) Zfp91 | Intron |

Columns 1, 2, and 3 give positional information about the structural variant (coordinates refer to the mm9 mouse assembly). Column 4 is the type of the variant. Column 5 and 6 give information about the underlying gene. Ins, insertion; Del, deletion; LINE, Long Interspersed Nuclear Element; IAP, Intracisternal A-particle; SINE, Short Interspersed Nuclear Element; MalR, Mammalian-Apparent Long-Term Repeat Retrotransposon; MTA, Mammalian Transposable Element; VNTR, Variable Number Tandem Repeat.
Human GWAS have shown that common SNPs (minor allele frequency >5%) explain only some fraction of the heritability, suggesting that SVs might also be contributing to individual phenotypic variation (Manolio et al., 2009). Results presented in this review suggest that, given the abundance of structural variants in mouse genomes, SVs make less of a contribution to individual phenotypic variation than SNPs. However, when they do, structural variants have a large effect size on the phenotype, providing a unique opportunity to investigate the relationship between structural variants and phenotypic differences, at a molecular as well as mechanistic level.

**EVOLUTIONARY IMPLICATIONS AND TRANPOSABLE ELEMENTS**

Transposable elements (TEs) have been highly influential in shaping the structure and evolution of mammalian genomes, as exemplified by TE-derived sequence contributing between 38 and 69% of genomic sequence (Ruzdin, 2004; Cordaux and Batzer, 2009; Shapiro, 2010; de Koning et al., 2011). TE insertions also can influence the transcription, translation or function of genes. Functional effects of TE insertions include their regulation of transcription by acting as alternative promoters or as enhancer elements and via the generation of antisense transcripts, or of transcriptional silencers. TEs are classified on the basis of their transposition mechanism (Goodier and Kazazian, 2008). Class I retrotransposons propagate in the host genome through an intermediate RNA step, requiring a reverse transcriptase to revert it to DNA before insertion into the genome. Class II DNA transposons do not have an RNA intermediate, and translocate with the aid of transposases and DNA polymerase. The overwhelming majority, over 96%, of TEs in the mouse genome, are of the retrotransposon type. These are further classified into three distinct classes: short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and the endogenous retrovirus (ERV) superfamily (Kvikstad and Makova, 2010).

Different classes of TEs are subsequently cleared from the genome: LINE elements are depleted from gene-rich sequence, whilst SINE TEVs tend to reside in G+C-rich sequence (Korenberg and Rykowski, 1988; Boyle et al., 1990). It was also observed that ERV TEVs are more heterogeneous than SINEs or LINEs in their G+C bias, with MuLV TEVs being enriched in high G+C sequence as SINEs and LINEs. Interestingly, by contrast to monomorphic TEs, polymorphic TEVs are more unevenly distributed among the chromosomes (having accounted for G+C content) with, for example, chromosome 19 exhibiting a significant enrichment of SINEs and the X chromosome showing a strong deficiency of all three TEV classes (Nellaker et al., 2012). The depletion of polymorphic LINEs on the X chromosome was previously seen in a study of four mouse strains (A/J, DBA/2J, 129S1/SvImJ, and 129X1/SvJ) (Akagi et al., 2008). TEVs from all three classes show strong and significant depletions in protein-coding gene exons, implying that such insertions are strongly deleterious (assuming that most TEVs across the noncoding genome are neutral or deleterious). The significant deficits of ERV or LINE TEVs in introns indicate that many were deleterious and thus were selectively purged over these strains’ evolutionary history. These observations agree with previous findings that LINE TE insertions are less tolerated within gene-rich sequence (Kvikstad and Makova, 2010).

A strong orientation bias is evident for each of the three TE classes (32.6, 41.7, and 41.6% for ERV, LINE, and SINE TEVs, respectively) (Nellaker et al., 2012). The orientation bias for IAP TEVs was recently reported to be 25.9% for a redundant set of 3317 intronic IAPs (Li et al., 2012). The strong biases for ERV and LINE TEVs in introns indicate that many were deleterious and thus were selectively purged over these strains’ evolutionary history. These observations agree with previous findings that LINE TE insertions are less tolerated within gene-rich sequence (Kvikstad and Makova, 2010).

Indeed, using a stringent statistical re-sampling approach to take into account confounding influences of strain and expression divergence, TEVs were found to be twice as likely to reside in a differentially expressed gene as expected by chance (Nellaker et al., 2012). However, when TEVs are considered with other forms of potential co-segregating mutations (SNPs, indels, and other structural variations), only 34 TEVs passed a stringent
genome-wide test, and these TEVs contain significantly fewer LINEs than the null expectation that all TEV classes have equal effects (Nellaker et al., 2012). While it has been extensively documented in the literature that de novo LINE insertions can cause changes in gene expression, it appears that, in Mus musculus, purifying selection has preferentially purged such variants. However, given that the proportion of expression heritability attributable to TEVs generally is no more than 10% (Yalcin et al., 2011).

To summarize, transposable elements make up almost half of the mouse genome (Gogvadze and Buzdin, 2009) and importantly their activity is the most prevalent mechanism for generating large structural variations in laboratory inbred mouse strains (Yalcin et al., 2011). However, as we demonstrated in this review, transposable elements appear to be under strong purifying selection for deleterious insertions with the majority of insertions observable in present day mouse strains having little phenotypic effects (Nellaker et al., 2012).

**DATA ACCESS AND VISUALIZATION**

The entire set of structural variation calls across 18 mouse genomes (129P2/OlaHsd, 129S1/SvImJ, 129S5SvEvBrd, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CAST/EiJ, CBA/J, C57BL/6N, C57BL/10ScCr, C57L/J, C57BL/6J, 129S6/SvEv, C3H/HeNCr, 129S6/SvEvCrl, C57BL/10MstCr, 129S1/SvImJ) is available online. Access to the data is straightforward and can be achieved through the Mouse Genomes Project website. The data can be queried automatically and manually through a web-based tool called LookSeq (Manske and Kwiatkowski, 2009). The tool allows users to search for structural variants by selecting a genome build, entering a search item or location, choosing the type of structural variant (deletion, insertion, gain, inversion or complex), selecting a strain, and more. The choice of the insert size depends on the size of the underlying structural variant, so that when the variant is large the insert size should also be large. Types of structural variants can be recognized using our comprehensive catalog of paired end mapping (PEM) patterns described in Yalcin et al. (2012a).

![Figure 6](image.png)
FUTURE WORK AND CONCLUDING REMARKS

The current approaches for cataloging mutations are primarily based on aligning sequencing reads to the appropriate reference genome to identify SNPs, indels, and structural variations. The majority of SV discovery methods to date have been based on observing patterns of clusters of aberrant read mappings to the reference genome. However, for many groups of strains or individuals there are many haplotypes that are not present on the reference genome and therefore are excluded from the catalog of mutations. This is especially true for the wild-derived mouse strains such as SPRET/EiJ, CAST/EiJ, and PWK/PhJ. So while the current approaches can often detect the presence of a non-reference haplotype in the form of a large insertion, they are blind to sequence variation occurring on the haplotype.

One solution to this problem is to create data structures capable of representing all of the haplotypes present in a group of related samples. In a recent study, Iqbal et al. developed de Bruijn graph methods for detecting and genotyping simple and complex genetic variants in an individual or population without a reference genome and were able to discover more than 3 Mb of sequence absent from the human reference genome (Iqbal et al., 2012).

The String Graph Assembler (SGA) was the first sequence assembly pipeline for next-generation data based on sequence overlaps (Simpson and Durbin, 2012). At the heart of SGA is the use of a compressed data structure called the FM-index, which is used to model the read sequence overlap graph of all the samples. Recently, work has been carried out to investigate building these structures using reads from multiple samples to represent all of the haplotypes present in the samples (Simpson, 2012).

An alternative approach is to first create individual whole-genome de novo assemblies for each sample and then subsequently carry out whole-genome alignments of the pre-assembled sequences. Several algorithms have been proposed for creating whole-genome alignments taking into account substitutions, insertions, deletions, and larger structural rearrangements. One such implementation of this approach is the combined Progressive Cactus and Hierarchical Alignment (HAL) graph pipeline (Paten et al., 2011). HAL is a graph-based hierarchical alignment format for storing multiple genome alignments arranged phylogenetically with the corresponding ancestral sequence reconstructions as internal nodes (Hickey et al., 2013).

The Mouse Genomes Project (http://www.sanger.ac.uk/resources/mouse/genomes/) has made a substantial contribution toward our understanding of structural variation diversity in mouse genomes and in their correlation to phenotypic variation. However, as explained in this review, there are ongoing challenges in computational detection of SVs with complex molecular architecture. Improved sequencing technologies with longer read lengths, along with the completion of de novo assemblies of mouse genomes, will be crucial in the identification of the remaining structural variants. De novo assembly also avoids reference bias in ascertainment of SVs (Sousa and Hey, 2013). Using longer fragments in sequencing library construction also aids in de novo assembly and SV detection in genomic regions that are “inaccessible” to short-read mapping due to their repetitive nature.

AUTHOR CONTRIBUTIONS

All authors read and approved the final manuscript. Thomas Meane and Binnaz Yalcin wrote the paper.

ACKNOWLEDGMENTS

This work was supported by the Medical Research Council and the Wellcome Trust. Binnaz Yalcin is supported by the French National Research Agency and the Swiss National Science Foundation.

REFERENCES

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., et al. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061–1068. doi: 10.1038/nature09534

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, 974–984. doi: 10.1101/gr.114876.110

Agam, A., Yalcin, B., Bhomra, A., Cubin, M., Webber, C., Holmes, C., et al. (2010). Elusive copy number variation in the mouse genome. PLoS ONE 5:e12839. doi: 10.1371/journal.pone.0012839

Akagi, K., Li, J., Stephens, R. M., Volfovsky, N., and Symer, D. E. (2008). Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res. 18, 869–880. doi: 10.1101/gr.75770.107

Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. Genome Res. 21, 961–973. doi: 10.1101/gr.112326.110

Alkan, C., Coo, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376. doi: 10.1038/nrg2958

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. 41, 1061–1067. doi: 10.1038/ng.437

Bauer, E., Schweimer, K., Kluver, E., Conejo-Garcia, J. R., Forssmann, W. G., Rosch, P., et al. (2005). Structure determination of human and murine beta-defensins reveals structural conservation in the absence of significant sequence similarity. Protein Sci. 10, 2470–2479. doi: 10.1100/sps.24401

Beck, J. A., Lloyd, S., Hafizparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M. E., et al. (2000). Genealogies of mouse inbred strains. Nat. Genet. 24, 23–25. doi: 10.1038/71641
Elia, J., Glessner, J. T., Wang, K., Takahashi, N., Shtir, C. J., Hadley, D., et al. (2012). Gogvadze, E., and Buzdin, A. (2009). Retroelements and their impact on genome evolution and functioning. Cell. Mol. Life Sci. 66, 3722–3742. doi: 10.1007/s00018-009-0107-2

Goodier, J. L., and Kazazian, H. H. Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135, 23–35. doi: 10.1016/j.cell.2008.09.022

Graubert, T. A., Cahan, P., Edwin, D., Selzer, R. R., Richmond, T. A., Eis, P. S., et al. (2007). A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 3:e3. doi: 10.1371/journal.pgen.0030003

Handsaker, R. E., Korn, J. M., Nemesh, J., and McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. 43, 269–276. doi: 10.1038/ng.678

Harewood, L., Chaing, A., and Reymond, A. (2012). Structural variation and its effect on expression. Methods Mol. Biol. 838, 173–186. doi: 10.1007/978-1-61779-357-7_8

Hart, S. N., Sarangi, V., Moore, R., Baheti, S., Bhavsar, J. D., Couch, F. J., et al. (2013). SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. PLoS ONE 8:e83356. doi: 10.1371/journal.pone.0083356

Helbig, I., Swinkels, M. E., Aten, E., Caliebe, A., van’t Slot, R., Boor, R., et al. (2014). Structural genomic variation in childhood epilepsies with complex phenotypes. Eur. J. Hum. Genet. 22, 896–901. doi: 10.1038/ejhg.2013.262

Henrichsen, C. N., Vinckenbosch, N., Zollner, S., Chaing, A., Pradervand, S., et al. (2009). Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. 41, 424–429. doi: 10.1038/ng.345

Hickey, G., Paten, B., Eeal, D., Zerbinio, D., and Haussler, D. (2013). HAL: a hierarchical framework for storing and analyzing multiple genome alignments. Bioinformatics 29, 1341–1342. doi: 10.1093/bioinformatics/btt128

Hollox, E. J., Huffmanie, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olhthuis, D., et al. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. Nat. Genet. 40, 23–25. doi: 10.1038/ng.2007.48

Hormozdizari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. 19, 1270–1278. doi: 10.1101/gr.086833.108

Iqbal, Z., Caccamo, M., Turner, I., Flice, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Brujin graphs. Nat. Genet. 44, 226–232. doi: 10.1038/ng.2028

Ibarra, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., et al. (2010). Skint1, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects epidermal gammadelta T cells. Cell. Mol. Life Sci. 67, 163–174. doi: 10.1007/s00018-010-0104-9

Iqbal, Z., Caccamo, M., Turner, I., Flice, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet. 44, 226–232. doi: 10.1038/ng.2028

Iturbe, A., Hart, S. N., Sarangi, V., Moore, R., Baheti, S., Bhavsar, J. D., Couch, F. J., et al. (2013). SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. PLoS ONE 8:e83356. doi: 10.1371/journal.pone.0083356

Keane et al. Structural variation in mouse genomes

Keane, T. M., Goodstadt, L., Dankecz, P., White, M. A., Wong, K., Yalcin, B., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289–294. doi: 10.1038/nature10413

Keane, T. M., Wong, K., and Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29, 389–390. doi: 10.1093/bioinformatics/bts697

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterrecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 40, e69. doi: 10.1093/nar/gkt603

Koolsen, D. A., Kramer, J. M., Neveling, K., Nilesen, W. M., Moore-Barton, H. L., Elmslie, E. V., et al. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat. Genet. 44, 639–641. doi: 10.1038/ng.2262

Korbel, J. O., Abyzov, A., Wu, X. J., Carrierio, N., Cayting, P., Zhang, Z., et al. (2009). PMEm: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 10:R23. doi: 10.1186/gb-2009-10-2-r23
Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420–426. doi: 10.1126/science.1149504

Korenberg, J. R., and Rykowski, M. C. (1988). Human genome organization: alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 53, 391–400. doi: 10.1016/0092-8674(88)90195-6

Kvikstad, E. M., and Makova, K. D. (2010). The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. *Genome Res.* 20, 600–613. doi: 10.1101/gr.099044.109

Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. (2009). MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* 6, 473–474. doi: 10.1038/nmeth.f.236

Li, J., Akagi, K., Hu, Y., Trivett, A. L., Hlynialuk, C. J., Swing, D. A., et al. (2012). Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. *Genome Res.* 22, 870–884. doi: 10.1101/gr.130470.111

Luptsi, J. R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14, 417–422. doi: 10.1016/S0168-9525(98)01555-8

Luptsi, J. R. (2009). Genomic disorders ten years on. *Genome Med.* 1, 42. doi: 10.1186/gm42

Luptsi, J. R., de Oca-Luna, R. M., Slagnauenhart, S., Pentao, L., Guzzetta, V., Trask, B. J., et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* 66, 219–232. doi: 10.1016/0092-8674(91)90613-4

Macaya, G., Thiery, J. P., and Bernardi, G. (1976). An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* 108, 237–254. doi: 10.1016/0022-2836(76)90105-2

Magi, A., Benelli, M., Yoon, S., Roviello, F., and Torricelli, F. (2011). Detecting common copy number variants in high-throughput sequencing data by using JointSILM algorithm. *Nucleic Acids Res.* 39, e65. doi: 10.1093/nar/gkr068

Malhotra, A., Lindberg, M., Faust, G. G., Leibowitz, M. L., Clark, R. A., Layer, R. M., et al. (2013). Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res.* 23, 762–776. doi: 10.1101/gr.136777.112

Malhotra, D., McCarthy, S., Michaelson, J. J., Vacic, V., Burdick, K. E., Yoon, J., et al. (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 72, 951–963. doi: 10.1016/j.neuron.2011.11.007

Manoli, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Manske, H. M., and Kwiatkowski, D. P. (2009). LookSeq: a browser-based viewer for deep sequencing data. *Genome Res.* 19, 2125–2132. doi: 10.1101/gr.095443.109

Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature* 470, 198–203. doi: 10.1038/nature09796

McConnell, M. J., Lindberg, M. R., Brennand, K. J., Pipher, J. C., Voet, T., Cowing-Zitron, C., et al. (2013). Mosaic copy number variation in human neurons. *Science* 342, 632–637. doi: 10.1126/science.1234372

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome Res.* 20, 1613–1622. doi: 10.1101/gr.106344.110

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6, 513–520. doi: 10.1038/nmeth.1374

Nellaker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. M., et al. (2012). The genomic landscape shaped by selection on transposable elements during cancer development. *Cell* 148, 59–71. doi: 10.1016/j.cell.2012.11.013

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., and Korbel, J. O. (2012b). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, 1333–1339. doi: 10.1093/bioinformatics/bts378

Robertson, J. T., Thorvaldsdottir, H., Winckler, W., Gutmann, M., Lande, S. R., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Russnes, H. G., Navin, N., Hicks, J., and Borresen-Dale, A. L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest.* 121, 3810–3818. doi: 10.1172/JCI57088

Sebat, J., Libkhardi, M., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445–449. doi: 10.1126/science.1138659

Shapiro, J. A. (2010). Mobile DNA and evolution in the 21st century. *Mol. DNA 14*: 10.1186/1759-8753-1-4

Simons, M. M., Greenaway, S., White, J. K., Fuchs, H., Gallus-Durner, V., Wells, S., et al. (2013). A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol.* 14:R82. doi: 10.1186/1475-2198-14-7-r82

Simpson, J. T. (2012). Efficient Sequence Assembly and Variant Calling Using Compressed Data Structures. Ph.D thesis, University of Cambridge. Available online at: ftp://ftp.sanger.ac.uk/pub/resources/theses/js18/

Simpson, J. T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 22, 549–556. doi: 10.1101/gr.126953.111

Simpson, J. T., Mcintyre, R. E., Adams, D. J., and Durbin, R. (2010). Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* 26, 565–567. doi: 10.1093/bioinformatics/btp693

Sindi, S. S., Onal, S., Peng, L. C., Wu, H. T., and Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.* 13:R22. doi: 10.1186/1475-2198-13-7-r22

Souza, V., and Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* 14, 404–414. doi: 10.1038/nrg3464

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O. P., Ingason, A., Steinberg, S., et al. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature* 455, 232–236. doi: 10.1038/nature07229

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. I., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27–40. doi: 10.1016/j.cell.2011.10.055

Stocking, C., and Kozak, C. A. (2008). Murine endogenous retroviruses. *Cell. Mol. Life Sci.* 65, 3383–3398. doi: 10.1007/s00018-008-8497-0
Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–853. doi: 10.1126/science.1136678

Sudmant, P. H., Huddleston, J., Catanach, C. R., Malig, M., Hillier, L. W., Baker, C., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* 23, 1373–1382. doi: 10.1101/gr.158543.113

Talbot, C. J., Nicod, A., Cherny, S. S., Fuller, D. W., Collins, A. C., and Flint, J. (1999). High-resolution mapping of quantitative trait loci in outbred mice. *Nat. Genet.* 21, 305–308. doi: 10.1038/6825

Tareen, S. U., Sawyer, S. L., Malik, H. S., and Emerman, M. (2009). An expanded clade of rodent Trim5 genes. *Virology* 385, 473–483. doi: 10.1016/j.virol.2008.12.018

Thiery, J. P., Macaya, G., and Bernardi, G. (1976). An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* 108, 219–235. doi: 10.1016/S0022-2836(76)80104-0

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Valdar, W., Solberg, L. C., Gauguer, D., Burnett, S., Klernerman, P., Cookson, W. O., et al. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* 38, 879–887. doi: 10.1038/ng1840

van de Lagemaat, L. N., Medstrand, P., and Mager, D. L. (2006). Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol.* 7:R86. doi: 10.1186/gb-2006-7-9-r86

Walters, R. G., Jacquemont, S., Valsesia, A., de Smith, A. J., Martinet, D., Andersson, J., et al. (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* 463, 671–675. doi: 10.1038/nature08727

Wong, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* 8, 652–654. doi: 10.1038/nmeth.1628

Wong, K., Bumpstead, S., Van Der Weyden, L., Reinholdt, L. G., Wilming, L. G., Adams, D. J., et al. (2012). Sequencing and characterization of the FVB/NJ mouse genome. *Genome Biol.* 13:R72. doi: 10.1186/gb-2012-13-8-r72

Wong, K., Keane, T. M., Stalker, J., and Adams, D. J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* 11:R128. doi: 10.1186/gb-2010-11-12-r128

Wu, B., Potter, C. S., Silva, K. A., Liang, Y., Reinholdt, L. G., Alley, L. M., et al. (2010). Mutations in sterol O-acyltransferase 1 (Soat1) result in hair interior defects in AKR/J mice. *J. Invest. Dermatol.* 130, 2666–2668. doi: 10.1038/jid.2010.168

Xie, C., and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80. doi: 10.1186/1471-2105-10-80

Yalcın, B., Adams, D. J., Flint, J., and Keane, T. M. (2012b). Next-generation sequencing of experimental mouse strains. *Mamm. Genome* 23, 490–498. doi: 10.1007/s00335-012-9402-6

Yalcın, B., and Flint, J. (2012). Association studies in outbred mice in a new era of full-genome sequencing. *Mamm. Genome* 23, 719–726. doi: 10.1007/s00335-012-9409-z

Yalcın, B., Flint, J., and Mott, R. (2005). Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* 171, 673–681. doi: 10.1373/gene/1002890

Yalcın, B., Nicod, J., Bhomra, A., Davidson, S., Cleak, J., Farinelli, L., et al. (2010). Commercially available outbred mice for genome-wide association studies. *PLoS Genet.* e1001085. doi: 10.1371/journal.pgen.1001085

Yalcın, B., Wong, K., Agam, A., Goodson, M., Keane, T. M., Gan, X., et al. (2011). Sequence-based characterization of structural variation in the mouse genome. *Nature* 477, 326–329. doi: 10.1038/nature10432

Yalcın, B., Wong, K., Bhomra, A., Goodson, M., Keane, T. M., Adams, D. J., et al. (2012a). The fine-scale architecture of structural variants in 17 mouse genomes. *Genome Biol.* 13:R18. doi: 10.1186/gb-2012-13-3-r18

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871. doi: 10.1093/bioinformatics/btp394

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-ne, P., Nicolas, A., et al. (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 26, 1895–1896. doi: 10.1093/bioinformatics/btq293

Zhang, J., Wang, J., and Wu, Y. (2012). An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics* 13(Suppl. 6):S6. doi: 10.1186/1471-2105-13-S6-S6

Zhang, Y., Romanish, M. T., and Mager, D. L. (2011a). Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput. Biol.* 7:e1001846. doi: 10.1371/journal.pcbi.1002046

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., et al. (2011b). Identification of genomic indels and structural variations using split reads. *BMC Genomics* 12:375. doi: 10.1186/1471-2164-12-375

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.