

# Sequence-based characterization of structural variation in the mouse genome

Binnaz Yalcin<sup>1\*</sup>, Kim Wong<sup>2\*</sup>, Avigail Agam<sup>1,3\*</sup>, Martin Goodson<sup>1\*</sup>, Thomas M. Keane<sup>2</sup>, Xiangchao Gan<sup>1</sup>, Christoffer Nellåker<sup>3</sup>, Leo Goodstadt<sup>1</sup>, Jérôme Nicod<sup>1</sup>, Amarjit Bhomra<sup>1</sup>, Polinka Hernandez-Pliego<sup>1</sup>, Helen Whitley<sup>1</sup>, James Cleak<sup>1</sup>, Rebekah Dutton<sup>1</sup>, Deborah Janowitz<sup>1,4</sup>, Richard Mott<sup>1</sup>, David J. Adams<sup>2</sup> & Jonathan Flint<sup>1</sup>

**Structural variation is widespread in mammalian genomes<sup>1,2</sup> and is an important cause of disease<sup>3</sup>, but just how abundant and important structural variants (SVs) are in shaping phenotypic variation remains unclear<sup>4,5</sup>. Without knowing how many SVs there are, and how they arise, it is difficult to discover what they do. Combining experimental with automated analyses, we identified 711,920 SVs at 281,243 sites in the genomes of thirteen classical and four wild-derived inbred mouse strains. The majority of SVs are less than 1 kilobase in size and 98% are deletions or insertions. The breakpoints of 160,000 SVs were mapped to base pair resolution, allowing us to infer that insertion of retrotransposons causes more than half of SVs. Yet, despite their prevalence, SVs are less likely than other sequence variants to cause gene expression or quantitative phenotypic variation. We identified 24 SVs that disrupt coding exons, acting as rare variants of large effect on gene function. One-third of the genes so affected have immunological functions.**

The pre-eminent organism for modelling the relationship between phenotype and genotype, including SVs, is the mouse, but our catalogue of SVs in this animal is incomplete<sup>6</sup> and most of what we know about the impact of SVs on phenotypes comes from analyses of gene expression<sup>7,8</sup>. Up to 28% of the between-strain variation in gene expression in haematopoietic stem and progenitor cells has been attributed to SVs<sup>7</sup>; SVs may account for between 66% and 74% of between-strain expression variation in kidney, liver, lung and testis<sup>8</sup>. Because gene expression variation is believed to contribute to variation in phenotypes in the whole organism<sup>9</sup>, SVs may turn out to have a major role in the genetic determination of many aspects of mouse biology.

Combining short-read paired-end mapping with experimental analyses (Supplementary Methods), we found SVs greater than 100 base pairs (bp) at 281,243 sites in the mouse genome, amounting to 711,920 SVs in thirteen classical and four wild-derived inbred strains of mice (Supplementary Table 1a), affecting 1.2% (33.0 Mb) and 3.7% (98.2 Mb) of the genome respectively (Supplementary Table 1b). Deletions, a category we can measure accurately, have a median size of 349 bp with modes at 100 bp and 6,400 bp (Supplementary Fig. 1a).

Our catalogue contains far more SVs than previously identified: 99.4% of SVs are simple and 0.6% are complex (Supplementary Table 1a), where simple SVs include insertions, deletions, inversions and copy number gains, and complex SVs consist of a mixture of events that abut each other. From experimental analyses of simple deletion SVs, we estimated an average false-negative rate of 17% in the classical inbred strains (Supplementary Tables 2a, b and 3a) and 24% in the wild-derived strains (Supplementary Table 2b); false-positive rates were below 5% for all strains (Supplementary Table 2c). False-negative rates for non-deletion simple SVs as well as complex SVs were higher than for simple deletions, ranging from 24% to 31% and 35% to 54% per strain, respectively (Supplementary Table 3b).

It proved difficult to obtain robust estimates of SVs smaller than 100 bp. Our best estimate of the rate of SVs between 30 and 100 bp is based on combining manual and automated methods over a region of 7.2 Mb (Supplementary Methods). Assuming that this region is typical, the rest of the genome (in classical laboratory strains) should contain approximately 49,000 SVs in this size range.

Microhomology at SV breakpoints, as well as the sequence content within SVs and the SV's ancestral state, were used to infer the likely mechanism of formation for simple SVs. To obtain breakpoint sequence, we performed *de novo* local assembly for 80.3% of deletions. Comparison of 1,314 predicted deletion breakpoints to the breakpoint delineated by PCR and sequencing (Supplementary Table 4) revealed that 57.7% of breakpoint predictions are exact and 86.5% are within 20 bp (Supplementary Table 5a). In cases where the local assembly strategy failed, we relied on the original breakpoint estimates obtained from the mapping of reads to the reference genome: 83.3% of these estimates are within 100 bp of the actual breakpoint (Supplementary Table 5b). Breakpoint accuracy for insertions, inversions and copy number gains is presented in Supplementary Table 5c, d and e, respectively.

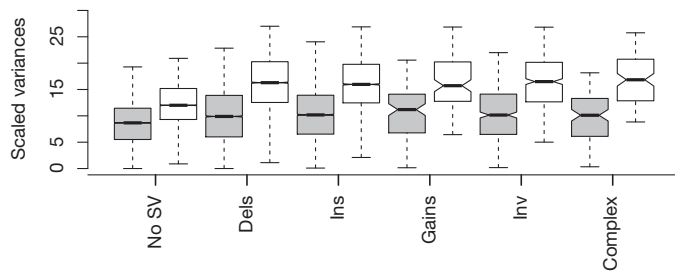
Genome-wide estimates of the contribution of each mechanism to SV formation were derived from analysis of breakpoint sequence of deletions relative to C57BL/6J. We have highly accurate breakpoint sequence for this SV category, which should be unbiased with respect to ancestry. Using rat as an outgroup, we classified 19% of relative deletion SVs as ancestral deletions, 57% as ancestral insertions and the remainder (24%) were indeterminate (Supplementary Fig. 2).

SVs are most often due to retrotransposons (long interspersed nuclear elements (LINEs; 25%), long terminal repeats (LTRs; 14%) and short interspersed nuclear elements (SINEs; 15%)), followed by variable number tandem repeats (VNTRs) (15%) and pseudogenes (2%). Other mechanisms, not involving retrotransposons, account for 29% of SVs. Outgroup analysis showed that the transposon-associated SVs arose almost exclusively from ancestral insertions events (98.8%). Target site duplications (12–16 bp) surround the breakpoints of LINE and SINE derived SVs; shorter (6–8 bp) sequences are associated with LTR SVs (Supplementary Fig. 1b). Non-repeat-mediated SVs are mainly a result of ancestral deletion events (79%), and are associated with microhomologies up to 7 bp in length (Supplementary Fig. 1b), consistent with either microhomology-mediated break-induced replication<sup>10</sup> or microhomology-mediated end joining<sup>11</sup>.

Given their potential role in human disease<sup>12</sup>, we were interested to document the occurrence of SVs that arise at the same genomic locus independently in unrelated strains (recurrent SVs). Non-allelic homologous recombination (NAHR) is the major mechanism for recurrent SVs<sup>13</sup>, whereas fork stalling and template switching and/or microhomology-mediated break-induced replication mechanisms may be important for non-recurrent SVs<sup>14</sup>.

<sup>1</sup>The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. <sup>2</sup>The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. <sup>3</sup>MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK. <sup>4</sup>Department of Psychiatry and Psychotherapy, Ernst-Moritz-Arndt-Universität Greifswald Klinikum der Hansestadt Stralsund, Rostocker Chaussee 70, 18437 Stralsund, Germany.

\*These authors contributed equally to this work.



**Figure 1 | Impact of SVs on gene expression.** Within-strain (grey boxes) and between-strain (white boxes) gene expression variances for transcripts which are not overlapped by any structural variant (No SV) and for those which are overlapped. Within-strain variance is due to environmental effects; between-strain variance is due to environmental and genetic effects. The difference between the two variances is a measure of heritability. Six categories are shown: No SV, deletions (Dels), insertions (Ins), copy number gains (Gains), inversions (Inv), and complex rearrangements (Complex).

Using the SV breakpoints obtained from PCR sequencing (249 SV sites in eight strains, accounting for over 4,000 breakpoints, Supplementary Table 4), we identified SVs occurring at the same locus in different strains, but with different breakpoints, indicating independent origins. In the classical strains, only 2.5% of deletions at the same locus had different breakpoint sequences. However within all 17 strains we found multiple alleles at 12% of SVs, due almost entirely to the presence of different alleles originating from the wild-derived inbred strains. Consistent with the low frequency of recurrent SVs, breakpoint features associated with NAHR are rare. We estimated that 0.13% of deletions are due to NAHR, when we required a signature of  $\geq 200$  bp of  $\geq 90\%$  sequence identity.

We assessed the impact of SVs on phenotypes by first estimating the proportion of heritability attributable to SVs<sup>8</sup> from brain RNA-seq and found that no category accounts for more than 10% (Fig. 1). To determine if these results were specific to brain tissue, we analysed gene expression data for the eight founder strains of the heterogeneous stock (HS) population ( $n = 5$  for each) from liver, measured on Illumina gene-expression arrays<sup>15</sup>. Mean heritability attributable to an SV, for transcripts overlapping one or more SVs, was 9.5%. Because many transcripts overlap multiple small SVs (median of 3, maximum of 216), we proposed that SV heritability might be related to the amount of gene overlapped. For each transcript we summed the amount of DNA overlapping a gene and expressed this as a proportion of the total length of the gene. SVs that overlap 50% or more of a gene make a large contribution to heritability: in brain tissue, such SVs contribute to 25% of the variance, compared to 7.8% for transcripts where SVs overlap less than 50% of the gene. However, large overlaps (50% or more) are rare, affecting less than 3% of transcripts. Thus, whereas SVs make a modest contribution to the overall heritability of expression variance, at individual transcripts they may be the main cause of between-strain differences in expression.

As another method to assess the impact of SVs on phenotype, we applied a test of functionality<sup>16</sup> to 281,246 SVs in association with 100 phenotypes measured in over 2,000 HS mice<sup>17</sup>. We identified 290 quantitative trait loci (QTLs) where SVs were among the variants most likely to be functional, but in all these cases the SVs were only a subset of the total number of functional variants. We found a small but highly significant deficit in SVs among the functional variants (0.36% compared to 0.54% among the non-functional,  $P < 10^{-16}$ ,  $\chi^2 = 72.1$ ).

Whereas SVs make a relatively small contribution to the total amount of quantitative phenotypic variation, at a small number of QTLs they are the cause of variation. As shown in an accompanying paper<sup>18</sup>, larger effect QTLs are more likely to arise from SVs. We identified 12 QTLs where the SV overlapped a gene or flanking region (2 kb up and downstream), and where the QTL effect size is in the top 5% of the distribution. Table 1 lists these SVs, the genes they affect and the putative phenotype with which they are associated. Two associations have been directly tested: complementation of the deletion of the *H2-Ea* promoter has confirmed the effect of this SV on the T-cell phenotype<sup>19</sup>; analysis of a knock out of *Eps15* showed the predicted lower locomotor activity (Fig. 2a).

There are relatively few examples where an SV can be said unequivocally to delete one, or more, coding exons. Without nucleotide resolution accuracy we cannot be certain whether the breakpoint of an SV lies within an exon. Therefore to find SVs overlapping exons we used our most accurate and complete category of SV calls: deletions relative to C57BL/6J. We identified 210 that overlap exons (Ensembl Build 58); after removing pseudogenes, and genes not annotated as 'protein coding', we were left with 24 SVs that affect coding exons, including six that encompass a gene in its entirety (Table 2).

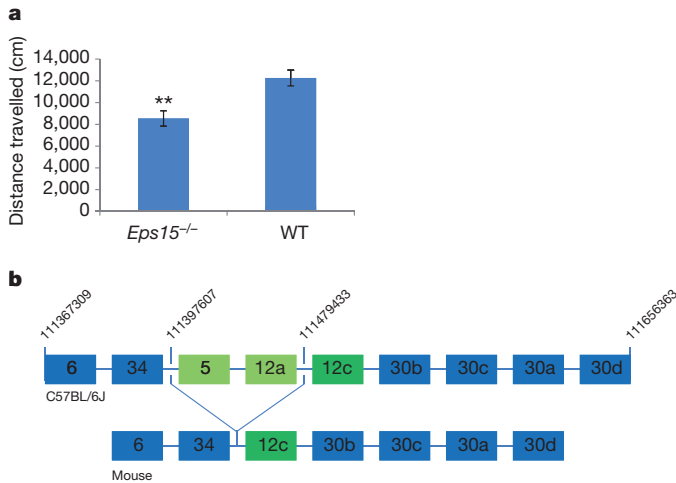
Five of the 24 SVs are already known<sup>20–24</sup>, the remaining 19 are novel. A third of the genes affected are involved in immunity and infection. Our data expand current knowledge of the molecular architecture of these SVs. Figure 2b shows that antiviral genes *Trim5* and *Trim12a* are unique to C57BL/6J, due to segmental duplication<sup>25</sup>. All the other strains contain only the *Trim12c* gene. Therefore the mouse contains a unique homologue of the human *TRIM5* gene. A similar analysis revealed that documented exonic changes in the defensin beta 8 gene (*Defb8*)<sup>26</sup> are linked to a previously undetected 3,192-bp ancestral insertion plus a 54-bp deletion (Table 2).

Our results are important in three respects. First, we find an unexpectedly large number of SVs with diverse molecular architecture, thus providing a catalogue of the most dynamic and variable regions of the mouse genome. Second, we were able to map almost 60% of deletions to base-pair resolution, allowing us to classify SVs by the mechanism that created them. In contrast to human SV studies, the great majority of SVs that we have discovered are non-recurrent rearrangements, based on two observations: among the classical strains, only 2.5% of deletions at the same locus had different breakpoint sequences and less than 1% of deletions are due to NAHR<sup>12</sup>. Third, SVs have relatively little impact on gene function, a conclusion based on the following observations. We found that SVs overlapping a gene account for less than 10% of variation in gene

**Table 1 | QTLs associated with SVs**

Phenotype	Chromosome	SV start	SV stop	Ancestral event	Gene	SV overlap
Mean platelet volume	1	175158884*	175158885*	Ins (large)	<i>Fcer1a</i>	Upstream
OFT total activity	2	144402760	144402971	SINE Ins	<i>Sec23b</i>	Intron
Hippocampus cellular proliferation marker	4	49690362	49690363	Del (137 bp)	<i>Grin3a</i>	Intron
Home cage activity	4	108951263	108951264	IAP Ins (~6,400 bp)	<i>Eps15</i>	Upstream
T-cells: %CD3	4	130038388	130038389	SINE Ins (202 bp)	<i>Snrnp40</i>	Intron
Wound healing	7	90731819	90731820	IAP Ins (~6,400 bp)	<i>Tmc3</i>	Upstream
Red cells: mean cellular haemoglobin	7	111397607	111479433	Ins	<i>Trim5</i>	Exon
Red cells: mean cellular haemoglobin	7	111504989	111505193	Del	<i>Trim30b</i>	UTR
Red cells: mean cellular volume	8	87957244	87957245	LINE Ins (~500 bp)	<i>4921524J17Rik</i>	Upstream
Serum urea concentration	11	115106127	115106250	Del	<i>Tmem104</i>	UTR
Hippocampus cellular proliferation marker	13	113783196	113783359	Del	<i>Gm6320</i>	Upstream
T-cells: CD4/CD8 ratio	17	34483681	34483682	Del (629 bp)	<i>H2-Ea</i>	Upstream

Start and stop coordinates are given for MGSCv37 of the mouse reference genome. Unless there is an asterisk, coordinates refer to the exact coordinates as delineated by Sanger sequencing. IAP, intracisternal A particle; Ins, insertion; Del, deletion; LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements. OFT, open-field test. UTR, untranslated region.



**Figure 2 | Experimental analysis of SVs.** **a**, Locomotor activity in *Eps15*<sup>-/-</sup> mice. Activity was recorded during a period of 10 min in an open field arena.  $n = 7$  for *Eps15*<sup>-/-</sup> male mice and  $n = 16$  for matched control wild type. **\*\*** $P$  value  $< 0.05$ . **b**, Schematic representation of the *Trim6-Trim30* genes cluster on chromosome 7. Boxes represent the sequential positions of the *Trim6*, *Trim34*, *Trim5/12* and *Trim30* genes. *Trim5* and *Trim12a* genes, which are only present in the C57BL/6J genome, occurred by segmental duplication of the *Trim12c* gene present in all 17 strains. The flanking *Trim34* and *Trim30* genes do not vary between strains. Coordinates are given for MGSCv37 assembly of the mouse reference genome.

expression, three to four times less than that found by studies using expression arrays<sup>7,8</sup>. SVs overlapping exons are rare: because the frequency of insertions is equal to that of deletions, and because these two categories make up 98% of all SVs, extrapolating from the 24 SVs that delete exons, we predict that there are only about 50 SVs that directly overlap exons, or about 0.2% of the total burden of SVs in the genome. Finally, our analysis of the phenotypic consequences of SVs on QTLs for multiple phenotypes points to a relative deficit of SVs as the molecular basis of complex phenotypes. For the classical laboratory strains, single nucleotide polymorphisms (SNPs) and indels affect 0.5% of the genome, whereas on average 33 Mb (1.2%) of each classical laboratory strain falls into structurally variant regions of the genome. This implies that SVs are at least twice as likely to have phenotypic consequences than the combined effect of SNPs and indels. Yet we find that SVs contribute only 10%

**Table 2 | SVs affecting coding regions**

Gene	Chromosome	SV start	SV stop	Ancestral event	Known function
<i>Soat1</i>	1	158394620	158401436	Del	Hair morphogenesis
<i>Olfir1055</i>	2	86179898	86186982	IAP Ins	Olfactory
<i>Fcrl5</i>	3	87245084	87245947	Del	Infection and immunity
<i>Nes</i>	3	87780530	87780662	VNTR	Brain development
<i>Pglyrp3</i>	3	91831862	91835385	Del	Infection and immunity
<i>Skint4,3,9</i> <sup>+</sup>	4	111731004*	112272814*	Ins	Infection and immunity
<i>Fv1</i>	4	147244398	147245739	Del	Infection and immunity
<i>Ugt2b38</i>	5	87850554	87854999	Del	Metabolism
<i>Klrb1a</i>	6	128559593	128559740	Del	Infection and immunity
<i>Klri2</i>	6	129689526	129691211	Del	Infection and immunity
<i>Tas2r120</i> <sup>+</sup>	6	132580541	132613777	Del+Ins 326-bp	Taste
<i>Tas2r103</i>	6	132985563	132986696	Del	Taste
<i>Zfp607</i> <sup>+</sup>	7	28646761	28671650	Del	DNA-binding
<i>Krtap5-5</i>	7	149415121	149415210	VNTR	Hair formation
<i>Trim5,12a</i> <sup>+</sup>	7	111397607	111479433	Ins	Infection and immunity
<i>Defb8</i>	8	19447465	19450575	Ins+54-bp Del	Infection and immunity
<i>Zfp872</i>	9	22004856	22005023	VNTR	DNA-binding
<i>Olfir913</i>	9	38402589	38403498	Del	Olfactory
<i>Rtp3</i>	9	110889280	110889465	VNTR	Bone density
<i>Nlrp1c</i> <sup>+</sup>	11	71046193*	71101410*	Ins	Embryonic development
<i>Fam110c</i>	12	31759321	31759461	VNTR	Cell migration
<i>Olfir234</i>	15	98328544	98328861	Del	Olfactory
<i>Krtap16-1</i>	16	88874294	88874392	VNTR	Hair formation
<i>Amd2</i> <sup>+</sup>	18	64607747	64609669	Ins	Biosynthesis of polyamines

Start and stop coordinates are given for MGSCv37 assembly of the mouse reference genome. Plus signs (+) indicate that the structural variant overlaps the entire gene. Unless there is an asterisk, coordinates refer to the exact coordinates as delineated by Sanger sequencing. VNTR, variable number tandem repeat.

to the heritability of gene expression, not the 50% implied by the genomic size argument.

It is important to note that conclusions based on our analysis of the HS outbred population may not apply to other outbred populations. The mouse population we tested is derived from inbred progenitors whose homozygosity will have purged their genomes of variants that could otherwise be maintained in heterozygous freely mating populations. Nevertheless, despite their relative rarity in the mouse genome, SVs that cause phenotype change are likely to provide biological insights out of proportion to their relative small contribution to phenotypic variance. We expect that the alleles we have described will provide a starting point for investigating the relationship between phenotype and genotype in mice.

## METHODS SUMMARY

**SV discovery.** We used a combination of four computational methods: split-read mapping<sup>27</sup>, mate-pair analysis<sup>28</sup>, single-end cluster analysis (SECluster and RetroSeq), and read-depth<sup>29</sup>. These methods identify deletions, insertions, inversions and copy number gains. We also derived methods to recognize other types of rearrangements, such as inversion plus insertion or inversion plus deletion, newly revealed from our experimental analysis.

**Experimental analysis.** We visually inspected short-read sequencing data using LookSeq<sup>30</sup> and manually detected SVs across mouse chromosome 19 in its entirety and a random set of other chromosomal regions. We analysed molecular structures of these SVs at nucleotide-level resolution using PCR and Sanger-based sequencing.

**Outgroup analysis.** The rat was used as an outgroup species to classify each mouse SV as either an ancestral deletion or an ancestral insertion. We predicted the ancestral state in the rat by estimating the size of the region in the rat genome that was homologous to the region that encompassed the mouse SV.

**SV classification.** We developed a machine learning method to classify SVs. The method used a random forest classifier, trained using sequence features within the SVs. Microhomology between breakpoints was determined by recording the longest sequence of bases that was identical between each breakpoint of each SV.

**Functional impact of SVs.** We tested whether an SV is likely to be functional using merge analysis<sup>16</sup>. The variances of expression data were calculated using ANOVA in the statistical software R using formulae described in ref. 8 and also by comparing a model where the expression value is explained by the strain, to a model in which the expression is explained by strain and whether or not the animal has an SV.

Received 5 July; accepted 4 August 2011.

1. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).

2. Quinlan, A. R. *et al.* Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* **20**, 623–635 (2010).
3. Zhang, F., Gu, W., Hurler, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
4. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
5. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature Genet.* **39**, 1217–1224 (2007).
6. Agam, A. *et al.* Elusive copy number variation in the mouse genome. *PLoS ONE* **5**, e12839 (2010).
7. Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nature Genet.* **41**, 430–437 (2009).
8. Henriksen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. *Nature Genet.* **41**, 424–429 (2009).
9. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37**, 710–717 (2005).
10. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genet.* **41**, 849–853 (2009).
11. Ma, J. L., Kim, E. M., Haber, J. E. & Lee, S. E. Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. *Mol. Cell. Biol.* **23**, 8820–8828 (2003).
12. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
13. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
14. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
15. Huang, G. J. *et al.* High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res.* **19**, 1133–1140 (2009).
16. Yalcin, B., Flint, J. & Mott, R. Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* **171**, 673–681 (2005).
17. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet.* **38**, 879–887 (2006).
18. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* doi:10.1038/nature10413 (this issue).
19. Yalcin, B. *et al.* Commercially available outbred mice for genome-wide association studies. *PLoS Genet.* **6**, e1001085 (2010).
20. Best, S., Le Tissier, P., Towers, G. & Stoye, J. P. Positional cloning of the mouse retrovirus restriction gene *Fv1*. *Nature* **382**, 826–829 (1996).
21. Boyden, L. M. *et al.* *Skint1*, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects epidermal  $\gamma\delta$  T cells. *Nature Genet.* **40**, 656–662 (2008).
22. Nelson, T. M., Munger, S. D. & Boughter, J. D. Jr. Haplotypes at the *Tas2r* locus on distal chromosome 6 vary with quinine taste sensitivity in inbred mice. *BMC Genet.* **6**, 32 (2005).
23. Persson, K., Heby, O. & Berger, F. G. The functional intronless S-adenosylmethionine decarboxylase gene of the mouse (*Amd-2*) is linked to the ornithine decarboxylase gene (*Odc*) on chromosome 12 and is present in distantly related species of the genus *Mus*. *Mamm. Genome* **10**, 784–788 (1999).
24. Wu, B. *et al.* Mutations in sterol O-acyltransferase 1 (*Soat1*) result in hair interior defects in AKR/J mice. *J. Invest. Dermatol.* **130**, 2666–2668 (2010).
25. Tareen, S. U., Sawyer, S. L., Malik, H. S. & Emerman, M. An expanded clade of rodent *Trim5* genes. *Virology* **385**, 473–483 (2009).
26. Taylor, K. *et al.* Defensin-related peptide 1 (*Defr1*) is allelic to *Defb8* and chemoattracts immature DC and CD4+ T cells independently of CCR6. *Eur. J. Immunol.* **39**, 1353–1360 (2009).
27. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. 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 (2009).
28. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009).
29. Simpson, J. T., McIntyre, R. E., Adams, D. J. & Durbin, R. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* **26**, 565–567 (2010).
30. Manske, H. M. & Kwiatkowski, D. P. LookSeq: a browser-based viewer for deep sequencing data. *Genome Res.* **19**, 2125–2132 (2009).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank A. Whitley, G. Durrant, A. M. Hammond, D. J. Fabrigar, L. Chen, M. Johannesson, E. Cong and G. Blázquez for helping B.Y. with various laboratory-based work. We also thank C. P. Ponting for comments on the manuscript. This project was supported by The Medical Research Council, UK, and the Wellcome Trust. D.J.A. is supported by Cancer Research UK.

**Author Contributions** D.J.A. and J.F. conceived the study and directed the research. J.F. wrote the core of the paper. K.W. and T.K. performed the genome-wide SV discovery and local assembly for SV breakpoint resolution. K.W. carried out the sensitivity and specificity analyses. K.W. and B.Y. liaised regularly to integrate experimental work into genome-wide SV discovery pipeline. This resulted in a highly accurate map of SV across the mouse genome, essential to downstream analyses. A.B., P.H.P., H.W., J.C., R.D. and D.J. carried out experimental work, led by B.Y. A.B. and B.Y. analysed Sanger-based sequencing data, resolved SV breakpoints at nucleotide-level resolution and inferred mechanism of SV formation. M.G. performed the genome-wide SV mechanism of formation and outgroup analysis, with contributions from A.A. and B.Y.; J.F. and A.A. analysed functional impact of SVs on expression and phenotypes. C.N., L.G., J.N., A.A. and R.M. carried out additional analyses. B.Y. characterized function of individual SV examples.

**Author Information** Data sets described here will be available under study accession number estd118 from the Database of Genomic Variants archive (DGVA) at <http://www.ebi.ac.uk/dgva/page.php>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at [www.nature.com/nature](http://www.nature.com/nature). Correspondence and requests for materials should be addressed to J.F. (jf@well.ox.ac.uk) or D.J.A. (da1@sanger.ac.uk).