A genome-wide association study for regulators of micronucleus formation in mice

Rebecca E. McIntyre¹, Jérôme Nico², Carla Daniela Robles-Espinoza¹, John Maciejowski³, Na Cai², Jennifer Hill⁴, Ruth Verstraten¹, Vivek Iyer¹, Alistair G. Rust¹⁵, Gabriel Balmus¹⁶, Richard Mott²⁷, Jonathan Flint², David J. Adams¹⁸


Correspondence to:

Dr. David Adams
Experimental Cancer Genetics
Wellcome Trust Sanger Institute
Hinxton, Cambridge, CB10 1HH
Ph: +44 1223 83496862
Email: da1@sanger.ac.uk

© The Author(s) 2013. Published by the Genetics Society of America.
Abstract

In mammals the regulation of genomic instability plays a key role in tumor suppression and also controls genome plasticity, which is important for recombination during the processes of immunity and meiosis. Most studies to identify regulators of genomic instability have been performed in cells in culture or in systems that report on gross rearrangements of the genome, yet subtle differences in the level of genomic instability can contribute to whole organism phenotypes such as tumour predisposition. Here we performed a genome-wide association study in a population of 1379 outbred Crl:CFW(SW)-US_P08 mice to dissect the genetic landscape of micronucleus formation, a biomarker of chromosomal breaks, whole chromosome loss, and extra-nuclear DNA. Variation in micronucleus levels is a complex trait with a genome-wide heritability of 53.1%. We identify seven loci influencing micronucleus formation (false discovery rate<5%), and define candidate genes at each locus. Intriguingly at several loci we find evidence for sex-specific genetic effects in micronucleus formation, with a locus on chromosome 11 being specific to males.
Introduction

Genomic instability is a key hallmark of nearly all cancer cells (Hanahan and Weinberg 2011), and patients carrying loss-of-function mutations in components of the DNA damage response (DDR) machinery, including BRCA1, BRCA2, ATM, BUB1 and RAD51, are highly predisposed to tumorigenesis due to a significant increase in their basal somatic mutation rate (Watson et al. 2013). First-degree relatives of mutation carriers have also been shown to have a higher cancer incidence, suggesting that the link between mutations in DDR genes and cancer is heritable. In the general population, genomic instability has been linked to tumor incidence, higher levels of genomic instability being associated with malignancies such as lung and prostate cancer, and tumours of the skin and brain (Bonassi et al. 2011; Forsberg et al. 2014). Furthermore, sequencing and genome-wide association studies have linked variants in DNA repair genes such as CHEK2, ATM, RAD50, BRIP1 and PALB2 with breast cancer disease risk (Rahman 2014), highlighting the importance of DNA repair in tumor predisposition. In the same way, genes that regulate cell division and chromosome segregation also prevent genomic instability, and their mutation has been linked to tumorigenesis (Godinho and Pellman 2014; Lee 2014). Micronuclei have also been linked to chromothripsis (Zhang et al. 2015), the phenomenon by which up to thousands of clustered chromosomal rearrangements occur in a single event in localised and confined genomic regions.

One test to assess genomic instability in vivo is the micronucleus assay (Evans et al. 1959), which enumerates the number of peripheral erythrocytes that carry DNA. Since the nucleus is expelled during erythropoiesis, the presence of micronuclei in these cells is easily quantified and indicates that DNA damage occurred prior to enucleation. The frequency of micronucleated cells indicates the basal level of somatic genomic instability. However, it is unknown whether the presence and extent of micronuclei are heritable traits, and therefore under genetic control. Understanding their genetic architecture may therefore indicate pathways by which genetic instability and cancer arise.
To identify genetic mediators of micronucleus formation *in vivo* we used an outbred mouse population. Performing these studies in the mouse has several distinct advantages: Firstly, in mice, micronucleated erythrocytes are not removed by the spleen meaning that micronucleus formation can be measured with high sensitivity and accuracy (Balmus *et al.* 2015). In humans these cells are rapidly cleared from the circulation. Secondly, functional studies can be followed up in the mouse by use of knockout or clustered regularly-interspaced short palindromic repeats (CRISPR) technology (Wang *et al.* 2013), and thus candidate causal variants/genes can be assessed *in vivo*. Lastly, since multiple aspects of whole animal physiology, such as metabolism and endocrine function, may influence micronucleus formation we can also look for the contribution of these factors to micronucleus levels, which is not possible using *in vivo* culture systems.

Genetic mapping in the mouse using conventional crosses suffers from poor resolution, which presents a challenge when trying to identify the causal genes (Yalcin *et al.* 2010). To circumvent this issue we used a Swiss Webster outbred stock Crl:CFW(SW)-US_P08 (hereafter CFW) more suited to high-resolution mapping. This stock carries a limited number of segregating alleles at each locus, and a large number of recombination events leading to rapid linkage disequilibrium decay (Yalcin *et al.* 2010). We measured micronucleus levels and genotypes in 1379 CFW mice that were part of a larger multi-phenotype study (Nicod *et al.*, 2016), and for which genotypes were already available, to show that this is a heritable polygenic trait. We map numerous associated quantitative trait loci (QTL) and identify potential causal genes, as a prelude to further functional investigation.

**Methods**

**Study animals and phenotyping**

All mice (Crl:CFW(SW)-US_P08) were purchased from Charles River, Portage, USA at 4-7 weeks of age and shipped to MRC Harwell, Oxfordshire, UK. Mice were housed in IVC cages (three per cage) on an ad-lib diet and at 16 weeks of age started a 4-weeks phenotyping pipeline where behavioural and physiological measures were collected. This phenotyping
pipeline and the data collected is described elsewhere (Nicod et al., 2016). At 20 weeks of age mice were weighed and sacrificed between 8 am and 12 pm, after overnight food restriction, and blood was collected by cardiac puncture into EDTA-coated vials. Full blood count analysis was performed with a Siemens Advia 2120 hematology analyzer using 200 µl of whole blood. In parallel a 50 µl aliquot of whole blood was collected into heparin solution, fixed in ice-cold methanol and stored at -80°C until micronucleus levels were measured by flow cytometry as described previously (Balmus et al. 2015).

Analysis of micronucleus levels was performed using the R statistical analysis software package using purpose-written software available on request from the authors. Outliers, defined as data points that were more than three standard deviations from the mean, were excluded. The effects of covariates such as sex, body weight and batch on micronucleus level were assessed by ANOVA, with each explaining more than 1% of the variance at a significant level ($P<0.05$) being included in a linear regression model from which residual measures were obtained. The linear model used for the micronucleus measure was the following: Micronucleus_level ~ 1 + Sex + (1|Batch) + Year_of_measure. The residuals were quantile-normalised and used for genome-wide association testing. For the sex-specific genome-wide association testing, sex was omitted from the covariates in the modelling of the micronucleus measure.

**Studies in knockout lines**

We analysed knockouts for the genes Trex1, Nfkb1l1, Trub2 and H2-Eb1 (Trex1<sup-tm1(KOMP)Wtsi</sup>, Nfkb1l1<sup-tm1a(KOMP)Wtsi</sup>, Trub2<sup-tm2a(EUCOMM)Wtsi</sup>, and H2-Eb1<sup-tm1a(KOMP)Wtsi</sup>) coming from the Sanger Mouse Genetics Project (White et al., 2013). Blood from these animals was collected at 16 weeks and analysed as previously described (Balmus et al. 2015).
Sequencing, variant calling and genotype imputation

Nicod et al., 2016 provides full details of the sequencing and genotyping protocol deployed in this study. In summary, DNA was extracted from tissues collected at the time of sacrifice. Sequencing libraries of 95 barcoded DNA samples were pooled and 100bp paired-end sequencing reads were generated, using one lane of a HiSeq (Illumina), per pool yielding 30 Gb of sequence data. Reads were mapped to the mouse mm10 reference genome and variants called using all chromosomes pooled together. Imputation of genotype probabilities was performed using STITCH, (Davies et al 2016) which models the chromosomes in the CFW mice as mosaics of a limited number of founder haplotypes. Optimization showed that the most probable number of founder haplotypes was 4. The catalogue of segregating variation in the CFW was derived from a ~370X pileup of all mice, combined with the positions of known SNPs in the Mouse Genomes Project (Keane et al., 2011). We identified 7,073,398 SNPs in this way, at which we imputed genotype dosages using STITCH (Davies et al 2016). After stringent post-imputation quality control we retained 5,766,828 high-quality imputed SNPs for subsequent analysis. We used two quality control measures on selecting well imputed versus poorly imputed SNPs: the $P$-value for violation of Hardy-Weinberg equilibrium $P$-value and IMPUTE2-style INFO scores, as described previously (Howie et al. 2009). The mean SNP-wise correlation ($r^2$) with sites that were also polymorphic on a genotyping microarray (Yang et al., 2009) using 44 samples was 0.974 before QC and 0.981 after QC.

Genome-wide association study

Details of QTL mapping are fully described in (Nicod et al., 2016). In brief, we identified a subset of 359,559 SNPs tagging the entire genome and used the imputed allele dosages at these loci to compute the genome-wide additive genetic relationship matrix (GRM). We tested for the association between each tagging SNP (represented by its imputed dosage) and the quantile-normalized residuals of the micronucleus and hematological measures as fixed effect in a mixed-model, controlling for relatedness and population structure using a GRM as
random effect. We used a leave-one chromosome-out strategy in which the GRM used to test association of SNPs on a given chromosome was computed from all other autosomes. Statistical significance at each locus was measured by ANOVA, comparing the fit of the allele dosage model to the null model. We estimated the FDR by permutation and called QTLs when at least one SNP had an FDR<5%. After the discovery phase using the 360k tagging SNPs, the genetic analysis was repeated in a 20Mb window around the mapped QTLs using the complete set of SNPs, to determine confidence intervals at each QTL using a logP-drop method. The same analysis was repeated testing males and females separately. We next tested for gene-by-sex interaction at all QTLs detected with all mice (males and females). Significance of the interaction effect at the QTL was determined by ANOVA, comparing the fit of the interaction model between sex and the dosage at the most strongly associated SNP (pheno~geno*sex) to the direct additive effect model (pheno~geno+sex).

Variant Functional Annotations

Putative SNPs were annotated using ANNOVAR (Wang et al. 2010) with gene annotations/proteins from the UCSC mouse genome annotation database (mouse assembly GRCm38/mm10). Unless otherwise stated, version 73 of the Ensembl mouse genome annotation database (assembly GRCm38/mm10) and software were used.

All SNPs

For each SNP position, its GERP sequence conservation score (Pollard et al. 2010) was obtained from the Ensembl Compara database based on the alignments of 36 Eutherian mammals (from the EPO whole-genome multiple alignment pipeline [ref http://sep2013.archive.ensembl.org/info/genome/compara/epo_anchors_info.html]). A score for sequence constraint was also reported derived from stretches of the local, multiple alignment around each SNP [ref http://sep2013.archive.ensembl.org/info/genome/compara/analyses.html#conservation].
Both coding and non-coding sequence nucleotide variants (SNVs) were analyzed using the Ensembl Variant Effect Predictor (VEP) (Yourshaw et al. 2015). For coding SNPs, SIFT (Ng and Henikoff 2003) was used to predict whether an amino acid substitution affects protein function, based on sequence homology and the physical properties of amino acids. Using transcription factor binding site data [ref http://sep2013.archive.ensembl.org/Mus_musculus/Experiment/Sources?db=funcgen;ex=all;f db=funcgen;r=17:46617590-46621119#ExperimentalMetaData], non-coding SNPs were assessed for potential disruption of binding sites.

**Coding SNPs**

A more detailed analysis to that performed for all the SNPs was performed for non-synonymous SNPs predicted by ANNOVAR. GERP (Pollard et al. 2010) and sequence conservation scores were obtained as described previously (Yang and Wang 2015). In addition, text-based alignments were obtained from Ensembl EPO alignments of 13 eutherian mammals (Yates et al. 2016). A region of ten nucleotides up- and down-stream for each SNP was specified, and the alignments across all potential 13 mammals was extracted for this 21 nucleotides region. Not all local regions had an alignment as some were unique to mouse.

With respect to protein-based analyses, VEP was configured to report whether a coding SNP lay within a protein domain. As before SIFT consequences (Ng and Henikoff 2003) were reported for each SNP, if available. In the case of multiple SNP predictions for a single SNP, the most deleterious prediction (lowest score) was reported.

In addition the effects of the amino acid change were assessed using the Grantham matrix (Grantham 1974). SNPs causing amino acid changes were classified by their Grantham scores as follows. A SNP was classified as 'conservative' for a Grantham score of less than 60. A score greater than 60 but less than 100 was classified as 'non-conservative'. A SNP was classified as 'radical' for scores equal to or greater than 100.

To obtain text-based protein alignments around a SNP, the following strategy was used. The ANNOVAR SNPs were predicted using UCSC protein data, identified using UCSC gene
identifiers. These UCSC gene identifiers were first used to link to Ensembl mouse transcript identifiers and then to Ensembl gene identifiers. The Ensembl gene identifiers were used to query the Ensembl Compara database for protein alignments across seven species; human, rat, dog, chicken, pig, cow and platypus.

For computational efficiency the Ensembl Compara software pipeline selects a single (usually canonical) transcript for each gene on which to compute homologies across species. Similarity is determined by aligning protein alignments of genes. Therefore, for genes that have multiple translations, features of interest can occur in transcripts from which it is not possible to determine their context in terms of homology via Ensembl. This is the case here for coding SNPs reported by ANNOVAR. The set of UCSC proteins used included proteins that were not considered the canonical transcript by Ensembl and upon which homologies were computed. Hence for ANNOVAR, predicted coding SNPs reported on a transcript not pre-computed in Ensembl, these were reported as 'UCSC_protein_not_used_for_homology_in_Ensembl'. Also there can be examples where the gene in which the SNP lies has no sequence homology to another species; these were tagged as 'no_mouse_sequence_in_Ensembl_homology'.

For UCSC-based SNPs present in the Ensembl Compara database (Yates et al. 2016) the site of the amino acid was used to create an expanded peptide sequence of four amino acids up- and down-stream of the mutated site. This gave a nine amino acid sequence to match against alignments from other species. This sequence was shortened appropriately if the SNP site was close to the start or end of the protein in which it lay. The localized protein alignments were reported in text format for the seven species as listed previously.

*Analysis of micronuclei in TREX1 wildtype and mutant cells.*

RPE-1 derived cell lines were constructed as described previously (Maciejowski et al. 2015). Cells were plated onto 35 mm glass bottom dishes (MatTek) 48 h before imaging. One hour before imaging cell culture media was replaced with phenol red-free DMEM/F12 medium. Live cell imaging was performed using a CellVoyager CV1000 spinning disk confocal system
(Yokogawa, Olympus) equipped with 445, 488, and 561 nm lasers, a Hamamatsu 512 x 512 EMCCD camera. Pinhole size was 50 µm. Images were acquired at the indicated intervals using a UPlanSApo 60x/1.3 silicone oil objective with the correction collar set to 0.17. The pixel size in the image was 0.27 µm. The 617/73 emission filter was used for image acquisition of mCherry-tagged proteins. Sixteen µm z-stacks were collected at 2.0 µm steps. Temperature was maintained at 37°C in a temperature-controlled enclosure with CO2 support. Maximum intensity projection of z-stacks and adjustment of brightness and contrast were performed using Fiji software. Image stitching was done with the Fiji plugin Grid/Collection stitching (Preibisch et al. 2009) with 20% tile overlap, linear blending, a 0.30 regression threshold, a 2.50 max/avg. displacement threshold, and a 3.50 absolute displacement threshold. Images were cropped and assembled into figures using Photoshop CS5.1 (Adobe). Evaluated for statistical significance using ANOVA and Kruskal-Wallis post hoc test.

Ethics Statement

The work described here was approved by the Oxford local ethics committee and was performed in accordance with Home Office Regulations, UK. A detailed description of the procedures is provided in Nicod et al., 2016.

Results

Strain specific and sex differences in micronucleus levels

To determine whether genetic diversity drives differences in micronucleus levels in mice, we first analyzed males from four inbred strains (Fig. 1A) finding significant differences in micronucleus levels between strains (BALB/cJ>C57BL/6NTAC ≠ CBA/2J>129S5SvEv^Brd, \( P<0.05 \); one-way analysis of variance (ANOVA)). Micronucleus levels were sexually dimorphic (Fig. 1B), with females of two C57BL/6 substrains (C57BL/6J and C57BL/6^C-c[Tyr^C-brd]) and 129S5SvEv^Brd showing significantly lower micronucleus levels than male littermates (\( P<0.0001 \); Student’s two-tailed t-test).
Genetic mapping in outbred mice

We provide a detailed description and characterization of the CFW mouse population used in this study in (Nicod et al., 2016). To map loci linked to micronucleus formation in the CFW we used a highly sensitive and reproducible high-throughput flow-cytometric micronucleus assay to score the frequency of micronucleated erythrocytes in blood obtained from 1,485 unrelated CFW animals (733 males and 752 females) aged 20 weeks of age (Balmus et al. 2015). Initial analysis revealed micronucleus levels to be approximately normally distributed (Figure S1). Body weight and a full haematological profile were also measured at the time of blood collection. DNA was obtained from 1,379 of these mice (677 males and 702 females) and, following sparse whole-genome sequencing at an average coverage of 0.15X (range 0.06X to 0.51X), genotype probabilities imputed at 7,073,398 single-nucleotide polymorphisms (SNPs) segregating within this population (Nicod et al., 2016). For the genetic analysis we retained 5,766,828 SNPs that passed a stringent post-imputation quality control threshold (Impute2-like INFO score > 0.4 and (in autosomes only) $P$-value for Hardy-Weinberg equilibrium $R^2 > 1 \times 10^{-6}$).

To map quantitative trait loci (QTL) we used a subset of 359,559 SNPs tagging all other SNPs with a minor allele frequency (MAF) > 0.1% at LD $R^2 > 0.98$, thereby capturing the common genetic variability present in this population. The heritability of micronucleus levels in the CFW, as estimated from the additive genetic relationship matrix (GRM) based on these tagging SNPs, was 53.1% (se 7.7%) (Nicod et al., 2016). Genetic mapping was performed using a mixed-model by testing the association of each tagging SNP as a fixed effect with the level of micronuclei, using the same GRM as a random effect, and controlling for relevant covariates (see Methods and Nicod et al., 2016). This genome-wide analysis revealed seven QTLs at a false discovery rate (FDR) < 5%. (Fig. 1C & 1D). We then repeated the genetic analysis around each QTL using all nearby SNPs from the total catalogue of 5,766,828, to fine-map and determine the confidence interval (CI) at each locus (see Methods).
Table 1 shows a summary of the seven loci identified and their 95% CI, with sizes ranging from 476Kb to 1.46Mb (mean 905Kb). In total, the 95% CI include 197 coding genes, or an average of 28 genes identified at each QTL. The most significant locus was on chromosome 8 with a –log10 P-value of 13.05 (Table 1). At this locus the most significant SNP fell next to the Werner syndrome Wrn gene. WRN is a RecQ helicase having intrinsic 3’ to 5’ DNA helicase activity (Gray et al. 1997; Shen and Loeb 2000). It interacts with Ku70/80, and participates in DNA end processing. Defects in WRN are associated with premature ageing, chromosomal instability, and tumorigenesis (Shen and Loeb 2000). Candidates at other loci include Trp53, Rassf3, and Trub2, all of which have established roles in the regulation of DNA repair/genomic stability (Smith and Fornace 1995; Zucchini et al. 2003; van der Weyden and Adams 2007), Fig. 2 & Fig. 3).

Interestingly, after correcting for the effect of potential confounding variables (see Methods), we found small but highly significant positive correlations between micronucleus level and several haematological measures; the red blood cells distribution width (RDW, $P = 2.55 \times 10^{-26}$, Spearman $R^2 = 0.1$), the hemoglobin concentration distribution width (HDW, $P = 1.63 \times 10^{-12}$, $R^2 = 0.04$, Spearman), the mean cellular hemoglobin concentration (CHCM, $P = 3.14 \times 10^{-11}$, $R^2 = 0.04$, Spearman), and the measure of blood hemoglobin (measHGB, $P = 1.88 \times 10^{-6}$, $R^2 = 0.02$, Spearman). Elevation of RDW, which is a measure of the variance of red blood cell width, often occurs together with elevated HDW, which measures the variation in hemoglobin content of red blood cells. For example, iron deficiency may cause a reduction in hemoglobin production (elevated HDW), which causes smaller red blood cells (elevated RDW). Elevation of RDW and/or HDW is a characteristic of anemia, a condition that may be caused by genomic instability (O’Driscoll 2012), and mouse models of genome instability disorders often display hematological abnormalities (Gallagher et al., 2011; with Crossan et al., 2011; with Nijnik et al., 2012). With this in mind we performed a genetic analysis with these hematological measures and discovered that measHGB and CHCM are both associated with the same locus on chromosome 5 ($P = 3.67 \times 10^{-7}$ and $5.05 \times 10^{-7}$, respectively [Figure S2]). We also mapped a micronucleus QTL at this locus, that contains only three genes within
its 95% CI, including Slc7a1. Mice carrying a homozygous deleterious mutation of this gene die of anemia at birth with 50% fewer red blood cells and reduced hemoglobin levels (Perkins et al. 1997). Hence Slc7a1 is a potential causative gene for the hematological traits and with a potential indirect effect on micronucleus levels. However, another gene within the QTL is the Microtubule Associated Tumor Suppressor Candidate 2 (Mtus2) (Jiang et al. 2009), which has been implicated in the function of the microtubule cytoskeleton (Ward et al. 2013).

Erythropoietic micronucleus formation and sex

Consistent with our observation of differences between sexes in micronucleus levels, genetic mapping performed on each sex independently revealed that, when applying the same FDR<5% threshold for QTL discovery, two loci (on Chrs 10 and 17) were detected only in females and one, on Chr 11, only in males (Fig. 1C). We tested for gene-by-sex interactions at all QTLs using the entire dataset and found that the locus on chromosome 11 shows a sex-specific effect on micronucleus formation in male mice (ANOVA $P=3.13 \times 10^{-4}$). The QTL on Chr 8 is present in both sexes while the QTLs on chromosomes 2, 5 and 9 do not reach genome-wide significance (FDR<5%) when males and females are tested separately, presumably due to a lack of power (Fig. 1D).

Identification of candidate genes at QTL

Gene ontology (GO) analysis revealed four genes under the micronucleus QTLs (Table S1) with annotations associated with DNA repair; Wrn, Hspa1b, Hspa1a and Hspa1l (GO 0006281). Other genes such as Aurora kinase B (Aurkb) have established roles in the regulation of processes such as the cell cycle (Adams et al. 2001), and links to processes such as mitosis, which might be expected to also result in elevated levels of micronuclei if dysregulated.

Our genotyping by low-coverage sequencing methodology means we have identified most high-frequency SNPs segregating in the CFW, and can test them for association with the trait, opening up the possibility of identifying functional variant(s) at each QTL. From our set of
5,766,828 SNPs we annotated all variants in the 95% CI at each QTL provided they had a P-value of association with micronucleus levels < 10^{-3}, to ascertain if they could disrupt gene function, and hence contribute to elevated micronucleus levels. We used a combination of approaches, annotating variants with ANNOVAR, applying GERP scores, Grantham and SIFT analysis, and also by using the Ensembl variant effector predictor (VEP) from Ensembl (McCarthy et al. 2014). **Table 2** shows a list of the top five scoring SNPs at each locus ranked by P-value; at some loci there were fewer than five or no coding SNPs that could be scored in this way.

We found missense variants with strong predicted deleterious effects in the *Wrn, Atrip, Trex1* and *Aurkb* genes, which, as noted above, have annotated roles in the regulation of genomic stability making these variants high value candidates. For *Aurkb* the SNP identified (rs29417126) results in a F45S change, and falls into a highly conserved region of the protein ([Fig. 4](#)). We projected this position onto the human AURKB protein revealing that it falls into a residue previously found to be post-translationally modified/phosphorylated (Daub et al. 2008). Indeed, this position has been found to be mitotically phosphorylated, and thus is likely to play a regulatory role in AURKB function. Thus *Aurkb* is a likely gene responsible for the association at the QTL on chromosome 11. Importantly this QTL also includes the gene *Trp53*, which is involved in the cellular response to DNA repair, and thus is another potential candidate at this locus. However, we did not identify any missense coding variant on the *Trp53* gene in the CFW, an observation in line with the absence of variants in the classical inbred strains (Keane et al. 2011). However, it is possible that a causal regulatory variant that controls expression of *Trp53* may yet be identified.

We next took advantage of both published and novel mouse knockout data to study genes within our QTL intervals. Over 1800 mammalian phenotype (MP) terms have been assigned to the 197 genes within the QTLs, including some associated with genomic instability, spindle abnormalities and defects in replication (**Table S1**). For many of the genes, phenotypes/functions associated with genomic instability have been assigned (**Table S2**). We
then attempted to predict genes associated with elevated levels of micronuclei. To do this we searched for live mice generated by the Wellcome Trust Sanger Mouse Genetics Project (White et al., 2013). This analysis revealed four strains (Trex1, Nfkbi1, Trub2 and H2-Eb1) available for testing using the micronucleus assay. Our analysis of potentially pathogenic variants within QTLs (Table 2) revealed a non-synonymous change within Trex1 (R269G; rs386972414) as a candidate variant. TREX1 is a major 3’-5’ DNA exonuclease linked to systemic lupus erythematosus and also Aicardi-Goutières syndrome (Fauré et al. 1999; Crow et al. 2000). Analysis of blood from Trex1+/− and Trex1−/− mice revealed significantly elevated levels of DNA positive red blood cells, making this gene a strong candidate at this locus (Fig. 2 and Fig. 4). The micronucleated erythrocyte frequency for Nfkbi1, Trub2 and H2-Eb1 mutants was not significantly different to that of wildtype control mice (Figure S3).

The identification of elevated levels of micronuclei in Trex1 mice was unexpected given the role of this gene in single-stranded DNA processing but not directly in DNA repair or chromosome segregation. We next used time lapse microscopy of micronuclei in isogenic human cell lines (RPE-1) in which chromatin (H2B) was labelled with mCherry in the context of TREX1 disruption, and in a matched control (TREX1 wildtype) line (Fig. 4). Analysis of 808 cell divisions from TREX1 mutant cells and 405 from wildtype cells revealed no difference in the frequency of large micronuclei, a result in keeping with the observation that Trex1 mutant mice do not show elevated levels of spontaneous mutation in the Big Blue assay (Morita et al. 2004). We conjecture that the elevated frequency of RBCs staining with propidium iodide may be the result of the accumulation of single-stranded DNA fragments previously reported to accumulate in tissues from Trex1 mutants (Morita et al. 2004), rather than bulky micronuclei resulting from chromosomal breaks or whole chromosome loss. This suggests that the micronucleus assay used here is capable of identifying genes with a range of DNA processes functions beyond those involved in DNA repair/chromosome segregation. It is important to note that Trex1 abuts Atrip, a known DNA repair gene, and we cannot exclude an indirect effect of the targeting event on Atrip gene function.
Discussion

We report an in vivo genetic screen for mediators of micronucleus formation and identify seven loci that reach genome-wide significance. At these loci we mapped genes that have established roles in DNA repair or the regulation of the cell cycle, and we annotated variants at these loci to identify possible candidates associated with elevated micronucleus levels. One locus also affects two haematological measures and contains a gene (Slc7a1) causing severe anaemia in the mouse. Further, functional evidence in vivo in a mouse mutant supports the role of Trex1 in the formation of extra-nuclear DNA.

We found evidence for a role of sex in the formation of micronuclei, with a locus on chromosome 11 containing the Aurkb gene being male-specific. In the case of erythrocyte micronucleus levels it seems unlikely that this phenotype is mediated by sex hormones or anatomical differences between male and female mice, although these factors could possibly contribute. Interestingly, men have a higher incidence of and mortality rate from sex-unspecific cancers, a fact that is unexplained by known risk factors (Cook et al. 2011; Edgren et al. 2012). Recently, mosaic loss of the chromosome Y in peripheral blood cells, which indicates loss of Y in hematopoietic progenitor cells, was associated with reduced survival and a higher risk of cancer in men (Forsberg et al. 2014). In CFW mice we found elevated micronucleus levels in males and find a sex-specific locus controlling this trait, further highlighting the role of sex in predisposition to tumorigenesis.

What might explain the differences in micronucleus levels in male and female mice that we observed? It has previously been shown that the regulatory genome is sexually dimorphic (Yang et al. 2006) with as many as 70% of transcripts showing sex-specific differences in expression (Yang et al. 2006), and recent work has shown that the expression of several DNA repair proteins is influenced by sex. Sexually dimorphic regulation of DNA repair genes could have evolved to support processes such as meiosis, which is regulated differently
between the sexes (Morelli and Cohen 2005). In addition to the locus on chromosome 11 six other loci containing both known and novel candidate DNA repair genes were identified. It is important to note that while micronuclei are a marker of genomic instability they also measure the genotoxic effects of disease processes. For example, micronuclei are elevated in autoimmune disorders such as systemic lupus erythematosus and thus may be associated with other traits/phenotypes defined in this outbred population (Al-Rawi et al. 2014). Future studies will involve validating candidate variants at these loci, for example using CRISPR-mediated gene editing.

Collectively, this work reveals the landscape of micronucleus formation in the mouse and the value of studying traits in an outbred mouse population, and at the whole organism level.
Data availability

The data and the results of the analysis described in this paper are available in an open access database at http://outbredmice.org. Mapping results can be visualised at http://mus.well.ox.ac.uk/gscandb/ (GScanViewer).
Acknowledgements

This work was supported by Cancer Research UK and the Wellcome Trust.

Author contributions

REM, JN, JM, DS, NC, VI, AR, RM and CDRE performed analysis. JF, JN, RM, DJA and REM wrote the paper.

Funder Statement

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
References


Evans, H. J., G. J. Neary, and F. S. Williamson, 1959 The relative biological efficiency of single doses of fast neutrons and gamma-rays on Vicia faba


Legends to Figure

Figure 1: Micronucleus levels are genetically controlled. A). Graph shows the percentage of micronucleated (MN) normochromatic erythrocytes (NCE) in males from four different inbred strains of mice. Boxplot shows the median and the min-max (whiskers) of at least six measurements for each strain. Data were analysed by one-way analysis of variance ($P<0.0001$) followed by Tukey’s post hoc test ($P<0.05$ for all combinations except C57BL/6NTAC and CBA/2J). B). Micronucleus levels are influenced by sex in mice. Shown is the difference in micronucleus levels in males (blue) and females (pink) for three strains of inbred laboratory mice. Data were analysed by unpaired Student’s two-tailed t-test, $P<0.00001$. C). Manhattan plots of genome-wide analyses for micronucleus levels performed in all mice (top) or in female (middle) or male (bottom) mice only using the subset of tagging SNPs (n=359,559). Horizontal red lines represent $-\log_{10} P$-values at which QTLs reach FDR $< 5%$. D). QTLs detected in all mice, males and females. Loci with FDR$<5\%$ are shaded in grey. These QTL positions were defined using the abovementioned 359,559 tagging SNPs.

Figure 2: Genome-wide significant loci for micronuclei on Chromosomes 2, 5, 8 and 9. The $-\log_{10} P$-values of imputed single-nucleotide polymorphisms (SNPs) associated with micronucleus levels are shown on the Y axis. The X axis gives chromosome and position in megabases (Mb). Genes within the regions are shown in the bottom panels (for clarity, as indicated on the figure, some gene names have been omitted). Linkage disequilibrium of each SNP with top SNP, shown in large purple diamond, is indicated by its colour. The plots were drawn using LocusZoom (Pruim et al. 2010).

Figure 3: Genome-wide significant loci for micronuclei on Chromosomes 10, 11 and 17. The $-\log_{10} P$-values of imputed SNPs associated with micronucleus levels are shown on the left y axis. The horizontal axis gives chromosome and position in megabases (Mb). Genes within the regions are shown in the bottom panels (for clarity, as indicated on the figure, some
gene names have been omitted). Linkage disequilibrium of each SNP with top SNP, shown in large purple diamond, is indicated by its colour. The plots were drawn using LocusZoom (Pruim et al. 2010).

**Figure 4: Candidate genes from genome-wide significant loci. A.** Frequency of propidium iodide positive, micronucleated (MN) normochromatic erythrocytes (NCE) in wildtype (+/+), heterozygote (tm1/+) and homozygous (tm/tm) Trex1 knockout male mice. Each circle, square or triangle indicates an individual mouse. Mutant mice had significantly elevated MN-NCE when compared to wild-type control mice (Student’s two-tailed t-test; \( P < 0.0001 \)), but heterozygote and homozygote mice showed comparable levels of MN-NCE. **B & C.** Micronucleus formation in human TREX1 null and wildtype control cells. Chromatin was labelled with H2B- mCherry. The data shown is the result of three independent experiments where >100 mitoses were counted. Cell lines 2.2 and 2.5 are TREX1 null RPE-1 cells. For a full description of the lines used in these experiments see Maciejowski et al. 2015. Wildtype (WT) refers to an isogenic control. **D.** Schematic to show alignment of mouse and human AURKB proteins. A candidate sequence nucleotide polymorphism (SNP) in Aurkb (rs29417126) falls into a highly conserved residue of AURKB that is known to be mitotically phosphorylated and to regulate the cell cycle.

**Table 1: Genome-wide significant loci.** Shown are the seven genome-wide significant loci for micronucleus levels and the genes within the 95% confidence intervals. The start and end positions of each QTL are provided, and the minor allele frequency (MAF), beta and effect size (variance explained) of the top scoring SNP. These QTL positions were defined using the entire collection of SNPs for higher mapping resolution.

**Table 2: Pathogenicity analysis of variants at loci associated with elevated micronucleus levels.** Shown are the top five scoring variants at each of the seven genome-wide significant
loci ranked by –log10 $P$-value. See Methods for a description of the approach used for variant annotation.
Table 1: Genome-wide significant loci.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Top SNP position</th>
<th>Top SNP logP</th>
<th>FDR</th>
<th>Size 95% Confidence Interval (bp)</th>
<th>Start position 95% CI</th>
<th>End position 95% CI</th>
<th>Number of genes in 95% CI</th>
<th>Gene names</th>
<th>MAF (%)</th>
<th>beta (%)</th>
<th>Variance explained (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>23690374</td>
<td>5.21</td>
<td>0.017</td>
<td>681130</td>
<td>23686966</td>
<td>30876935</td>
<td>20</td>
<td>Rapgef1, G6, Sf13-1, Tnfr1, Ccn2, Ccne1, Ccne2, G6, Sf13-1, Tnfr1, Ccn2</td>
<td>0.49</td>
<td>0.33</td>
<td>2.19</td>
</tr>
<tr>
<td>5</td>
<td>14678119</td>
<td>4.80</td>
<td>0.045</td>
<td>749781</td>
<td>14605075</td>
<td>14684175</td>
<td>3</td>
<td>Mrch1, Sm2, Ix6, 1x3</td>
<td>0.46</td>
<td>0.27</td>
<td>1.13</td>
</tr>
<tr>
<td>8</td>
<td>53158259</td>
<td>13.05</td>
<td>0</td>
<td>1312071</td>
<td>32889029</td>
<td>34289139</td>
<td>13</td>
<td>Wnt, Fgr, Tnfr1, Ccn2, Ccne1, Ccne2, Sm2, Ix6, 1x3</td>
<td>0.38</td>
<td>0.59</td>
<td>4.46</td>
</tr>
<tr>
<td>9</td>
<td>19927889</td>
<td>5.43</td>
<td>0.015</td>
<td>624562</td>
<td>10358568</td>
<td>10458101</td>
<td>20</td>
<td>Ccne1, Ccne2, Ccne1, Ccne2, Sm2, Ix6, 1x3, Tnfr1, Ccn2</td>
<td>0.20</td>
<td>0.46</td>
<td>1.77</td>
</tr>
<tr>
<td>10</td>
<td>121477912</td>
<td>0.55</td>
<td>0</td>
<td>4790032</td>
<td>121220924</td>
<td>121703505</td>
<td>6</td>
<td>Tnfr1, Ccn2, Sm2, Ix6, 1x3, Ccne1, Ccne2</td>
<td>0.50</td>
<td>0.46</td>
<td>2.98</td>
</tr>
<tr>
<td>11</td>
<td>86573996</td>
<td>0.21</td>
<td>0</td>
<td>1052568</td>
<td>88747919</td>
<td>69663195</td>
<td>49</td>
<td>Myc, Iklk3, Iklk2, Iklk1, Ccmk, Ccmx, Ccmx, Ccmx, Ccmx, Ccmx</td>
<td>0.45</td>
<td>0.45</td>
<td>2.05</td>
</tr>
<tr>
<td>17</td>
<td>34159695</td>
<td>9.25</td>
<td>0</td>
<td>14053751</td>
<td>34145626</td>
<td>35062386</td>
<td>86</td>
<td>H2-DMb2, H2-DMb1, Psmb9, Tap1, Tap1, Tap1, Tap1, Tap1, Tap1, Tap1</td>
<td>0.36</td>
<td>0.74</td>
<td>3.16</td>
</tr>
<tr>
<td>Mouse_genomic_location</td>
<td>Mouse_gene</td>
<td>Consequence</td>
<td>log2_ratio</td>
<td>GENEenan</td>
<td>Grumman_classification</td>
<td>Grumman_variant</td>
<td>GOE_class</td>
<td>Lowest_GOE_score</td>
<td>ResNISV_TSV_class</td>
<td>IndelMvn_TSV_class</td>
<td>Rare impact</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------</td>
<td>-------------</td>
<td>------------</td>
<td>----------</td>
<td>------------------------</td>
<td>----------------</td>
<td>----------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>chr11:69107564-69107564</td>
<td>Atrp</td>
<td>C&gt;T</td>
<td>0.74</td>
<td>-9.38</td>
<td>radical</td>
<td>205</td>
<td>tolerated</td>
<td>0.16</td>
<td>downstream_gene_variant,missense_variant</td>
<td>2.18</td>
<td>2.185</td>
</tr>
<tr>
<td>chr1:29788325-29788325</td>
<td>Atrp</td>
<td>G&gt;A</td>
<td>0.74</td>
<td>-9.38</td>
<td>radical</td>
<td>205</td>
<td>tolerated</td>
<td>0.16</td>
<td>downstream_gene_variant,missense_variant</td>
<td>2.18</td>
<td>2.185</td>
</tr>
<tr>
<td>chr1:29702527-29702527</td>
<td>Atrp</td>
<td>G&gt;A</td>
<td>0.74</td>
<td>-9.38</td>
<td>radical</td>
<td>205</td>
<td>tolerated</td>
<td>0.16</td>
<td>downstream_gene_variant,missense_variant</td>
<td>2.18</td>
<td>2.185</td>
</tr>
<tr>
<td>chr2:30174134-30174134</td>
<td>Atrp</td>
<td>G&gt;A</td>
<td>0.74</td>
<td>-9.38</td>
<td>radical</td>
<td>205</td>
<td>tolerated</td>
<td>0.16</td>
<td>downstream_gene_variant,missense_variant</td>
<td>2.18</td>
<td>2.185</td>
</tr>
<tr>
<td>chr2:30174134-30174134</td>
<td>Atrp</td>
<td>G&gt;A</td>
<td>0.74</td>
<td>-9.38</td>
<td>radical</td>
<td>205</td>
<td>tolerated</td>
<td>0.16</td>
<td>downstream_gene_variant,missense_variant</td>
<td>2.18</td>
<td>2.185</td>
</tr>
<tr>
<td>chr5:148288935-148288935</td>
<td>D2Wsu81e</td>
<td>A&gt;C</td>
<td>0.625</td>
<td>3.14</td>
<td>radical</td>
<td>113</td>
<td>0.0</td>
<td>3.14</td>
<td>downstream_gene_variant,missense_variant</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>chr17:34711321-34711321</td>
<td>Tex15</td>
<td>-3.07</td>
<td>-3.07</td>
<td>60</td>
<td>deleterious</td>
<td>56</td>
<td>102</td>
<td>0.02</td>
<td>downstream_gene_variant,splice_region_variant,missense_variant</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>
A

B

C

D

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>All mice (n=1379)</th>
<th>Males (n=677)</th>
<th>Females (n=702)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>29991588</td>
<td>5.14</td>
<td>1.85</td>
<td>3.70</td>
</tr>
<tr>
<td>5</td>
<td>148775593</td>
<td>4.58</td>
<td>2.12</td>
<td>2.64</td>
</tr>
<tr>
<td>8</td>
<td>33138260</td>
<td>13.01</td>
<td>6.74</td>
<td>7.35</td>
</tr>
<tr>
<td>9</td>
<td>109064030</td>
<td>5.32</td>
<td>1.88</td>
<td>3.53</td>
</tr>
<tr>
<td>10</td>
<td>121477912</td>
<td>8.55</td>
<td>2.84</td>
<td>7.76</td>
</tr>
<tr>
<td>11</td>
<td>69550069</td>
<td>8.09</td>
<td>10.40</td>
<td>0.58</td>
</tr>
<tr>
<td>17</td>
<td>34732496</td>
<td>8.42</td>
<td>2.75</td>
<td>7.39</td>
</tr>
</tbody>
</table>
**A**

![Graph](image)

**B**

![Image](image)

**C**

![Bar Graph](image)

**D**

![Diagram](image)