RE McIntyre, J Nicod, et al (2016), A genome-wide association study for regulators of micronucleus formation in mice, G3- Genes Genomes Genetics, Vol. 6(8), 2343-2354
A Genome-Wide Association Study for Regulators of Micronucleus Formation in Mice


*Experimental Cancer Genetics and §‡‡Microbial Pathogenesis, The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK; †Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK; ‡Laboratory for Cell Biology and Genetics, The Rockefeller University, New York, New York 10065; **Tumour Profiling Unit, The Institute of Cancer Research, London SW3 6JB, UK; ††The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, CB2 1QH, UK; and §§Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Boulevard Juriquilla 3001, Santiago de Querétaro 76230, Mexico

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 tumor 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 extranuclear 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 sexual dimorphism in micronucleus formation, with a locus on chromosome 11 being specific to males.

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 cancer is heritable. In the general population, genomic instability has been linked to tumor incidence, with higher levels being associated with malignancies such as lung and prostate cancer, and tumors 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, BRI1, 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 localized 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: First, 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. Second, 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 multiphenotype 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.

**MATERIALS AND METHODS**

**Study animals and phenotyping**

All mice [Crl:CFW(SW)-US_P08] were purchased from Charles River, Portage, at 4–7 wk of age and shipped to MRC Harwell, Oxfordshire, UK. Mice were housed in IVC cages (three per cage) on an ad-libitum diet and at 16 wk of age started a 4-wk phenotyping pipeline where behavioral and physiological measures were collected. This phenotyping pipeline and the data collected are described elsewhere (Nicod et al. 2016). At 20 wk of age mice were weighed and killed 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° until micronucleus levels were measured by flow cytometry as described previously (Balms 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 analysis of variance (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-normalized and used for genome-wide association testing.

**Studies in knockout lines**

We analyzed knockouts for the genes Trex1, Nfkbi1, Trub2, and H2-Eb1 (Trex1<sup>fl<sup>–</sup>l(ROMPF)<sup>W60</sup>, Nfkbi1<sup>fl<sup>–</sup>l(ROMPF)<sup>W60</sup>, Trub2<sup>fl<sup>–</sup>l(EUCOMM)<sup>W60</sup>, and H2-Eb1<sup>fl<sup>–</sup>l(ROMPF)<sup>W60</sup>) coming from the Sanger Mouse Genetics Project (White et al. 2013). Blood from these animals was collected at 16 wk and analyzed as previously described (Balms 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 death. Sequencing libraries of 95 barcoded DNA samples were pooled and 100 bp 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 catalog of segregating variation in the CFW population 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 vs. 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 (r²) 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**

Details of QTL mapping are fully described in Nicod et al. (2016). In brief, we identified a subset of 359,539 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 effects 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 false discovery rate (FDR) by permutation and called QTL when at least one SNP had an FDR < 5%. After the discovery phase using the 360 K tagging SNPs, the genetic analysis was repeated in a 20 Mb window around the mapped QTL 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 QTL detected with all mice (males and females). Significance of the interaction effect at the QTL was determined by ANOVA, comparing the sex-specific genome-wide association testing, sex was omitted from the covariates in the modeling of the micronucleus measure.
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 Annotate Variation (ANNOVAR) (Wang et al. 2010) with gene annotations/proteins from the University of California, Santa Cruz (UCSC) mouse genome annotation database (mouse assembly GRCh38/mm10). Unless otherwise stated, version 73 of the Ensembl mouse genome annotation database (assembly GRCh38/mm10) and software were used.

**All SNPs**

For each SNP position, a 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 noncoding sequence nucleotide variants (SNVs) were analyzed using the Ensembl Variant Effect Predictor (VEP) (Yourschaw et al. 2015). For coding SNPs, SIFT (Ng and Henikoff 2003) was used to predict whether an amino acid substitution may influence 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;fdb=funcgen;r=17:46617590-46621119#ExperimentalMetaData), noncoding 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 nonsynonymous 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 10 nucleotides up- and down-stream for each SNP was specified, and the alignments across all potential 13 mammals were extracted for this 21-nucleotide 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 effect of amino acid changes were assessed using the Grantham matrix (Grantham 1974). A SNP was classified as ‘conservative’ when it had a Grantham score of < 60. A score > 60 but < 100 was classified as ‘nonconservative.’ A SNP was classified as ‘radical’ when it had a score > 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 precomputed 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 wild-type and mutant cells**

RPE-1 derived cell lines were constructed as described previously (Maciejowska et al. 2015). Cells were plated onto 35 mm glass bottom dishes (MatTek) 48 hr 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 and a Hamamatsu 512 × 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-micrometer 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). Evaluation for statistical significance was carried out using ANOVA and the 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).

**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.
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 (Figure 1A) finding significant differences in micronucleus levels between strains (BALB/cJ > C57BL/6NTAC ≥ CBA/2J > 129S5SvEvBrd; P < 0.05; one-way ANOVA). Micronucleus levels were sexually dimorphic (Figure 1B), with females of two C57BL/6 substrains [C57BL/6J and C57BL/6c-/c-(C57BL/6-Tyr C-Brd)] and 129S5SvEvBrd showing significantly lower micronucleus levels than male littermates (P < 0.00001; 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 1485 unrelated CFW animals (733 males and 752 females) culled at 20 wk of age (Balmus et al. 2015). Initial analysis revealed micronucleus levels to be approximately normally distributed (Supplemental Material, Figure S1). Body weight and a full hematological profile were also measured at the time of blood collection. DNA was obtained from 1379 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 were imputed at 7,073,398 SNPs segregating within this population (Nicod et al. 2016).

For the genetic analysis we retained 5,766,828 SNPs that passed a stringent postimputation quality control threshold [Impute2-like INFO score >0.4 and (in autosomes only) P-value for Hardy-Weinberg equilibrium R² > 1 · 10⁻⁶].

To map QTL we used a subset of 359,559 SNPs tagging all other SNPs with a minor allele frequency (MAF) < 0.1% at LD R² > 0.98, thereby capturing the common genetic variability present in this population. The heritability of micronucleus levels in the CFW population, as estimated from the additive 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 Materials and Methods). This genome-wide analysis revealed seven QTL at an FDR < 5% (Figure 1, C and D). We then repeated the genetic analysis around each QTL using all nearby SNPs from the total catalog of 5,766,828, to refine-map and determine the confidence interval (CI) at each locus (see Materials and Methods).

Table 1 shows a summary of the seven loci identified and their 95% CI, with sizes ranging from 476 kb to 1.46 Mb (mean 905 kb). In total, the 95% CI include 197 coding genes, or an average of 28 genes
## Table 1 Genome-wide significant loci

| Chromosome | Start Position | Top SNP Position | Size 95% CI (bp) | Start Position 95% CI | End Position 95% CI | Number of Genes in 95% CI | Gene Names | MAF | β | Variance Explained (%) |
|------------|----------------|-----------------|------------------|------------------------|---------------------|--------------------------|------------|-----|------------------------|
| 2          | 29990674       | 2                | 5.21             | 0.017                  | 661130              | 30267695                 | Rapgef1, Gm13547, Trub2, Coq4, Slc27a4, Umt1, Ceram, Odl2, Gle1, Sptan1, Wdx34, Set, Pik3, Zdhrc12, Zer1, Tbc1d13, Endog, D2Wsu81e, Ccb1, Lrcc8a | 0.49 | 0.33 | 1.29 |
| 5          | 148761610      | 4.66             | 0.045            | 747871                 | 148093705           | 148841575                | Mu6s2, Scl7a1, Ubi3 | 0.46 | 0.27 | 1.13 |
| 8          | 33158129       | 13.05            | 0                | 1312071                | 32889539            | 34201609                 | Wn, Purg, Tex15, Ppp2cb, Ubxn8, Gsr, Gft2e2, Smm18, Rbpsms, Dctn6, Mboat4, Leprorf1, Sraf | 0.38 | 0.59 | 4.46 |
| 9          | 109076890      | 5.43             | 0.015            | 624542                 | 108855860           | 109480401                | Scl26a6, Tmem89, Uqrrc1, Col7a1, Ucn2, Pklfb4, Shisa5, Trex1, Atrip, Tma7, Ccsc51, Plxb1, Fbxw21, Fbxw13, Fbxw20, Fbxw14, Fbxw14, Fbxw22, Fbxw16, Fbxw19 | 0.20 | 0.46 | 1.77 |
| 10         | 121477912      | 8.55             | 0                | 476002                 | 121229924           | 121705925                | Tbc1d30, Gns, Rass3, Tbk1, Xpot, D930020B18Rik | 0.50 | 0.46 | 2.98 |
| 11         | 69570999       | 8.21             | 0                | 1055268                | 68747199            | 69803186                 | Myh10, Nlde1, Rnf222, Rpl26, Odf14, Arhgef15, Slc25a35, Rangfrf, Fias, Ctc1, Aurb6, 2310047M10Rik, Tmem107, Vamp2, Per1, Hez7, Alox3, Alox12b, Alox8, Gucy2e, Ctnrob, Trappc1, Kncab3, Chd3, Cyb5d1, Naa38, Tmem88, Kdm6b, Dnah2, Efnb3, Wrap53, Trp53, Aph1b2, Sibg, Sat2, Fxr2, Sox15, Mpu1, Cad68, Eif4a1, Senp3, Trnsf13, Trnsf13, Trnsf13, Trnsf13, Fbxw20, Fbxw14, Fbxw14, Fbxw22, Fbxw16, Fbxw19 | 0.45 | 0.45 | 2.55 |
| 17         | 34159865       | 9.25             | 0                | 1456751                | 34145616            | 35602366                 | H2-DMb2, H2-DMb1, Pmb9, Tap1, Pmb8, Tap2, H2-Ob, H2-Ab1, H2-AA, H2-Eb1, H2-Eb2, Bln12, Bn1f1, BCDS1142, Bn1f14, Bn16, Notch4, Gpn2m3, Fbx2, Ager, Rnf5, Agpat1, Egfl8, Ppl2, Pmt1, Fkbp1, Atf6b, Tnxb, C4B, Cyp21a1, S1k19, Dxo, Skiv2l, Nelf6, Ceb, C2, Zbb12, Ehmt2, Slc44a4, Neu1, Hepa1b, Hepa1a, Hepa1f, Lsm2, Vars, Wva7, Sapcd1, Msh5, Clic1, Ddah2, G6b, Ly6g6c, Ly6g6d, Ly6g6e, Ly6g6f, Abhd16a, Ly6g5c, Ly6g5b, Csnk2b, Gpank1, D17H6553E, Apom, Bag6, Prcc2a, Aif1, Lst1, Ltb, Tnf, Lta, Nkfb1, Atp6v1g2, Ddx39b, H2-D1, H2-Q1, H2-Q2, H2-Q4, H2-Q7, H2-Q10, Pousf1, Tcf19, Chcc1, Pors1c2, Cdsn, 2300002M23Rik, Sf1a2 | 0.38 | 0.74 | 3.16 |

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), β, 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. Note that the -logP value given in the table is the maximum among all imputed SNPs under the QTL, which is generally higher than that shown on Fig 1D.
The most significant locus was on chromosome 8 with a $-\log_{10} P$-value of 13.05 (Table 1). At this locus the most significant SNP fell next to the Werner syndrome \textit{Wrn} gene. \textit{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 \textit{WRN} are associated with premature aging, chromosomal instability, and tumorigenesis (Shen and Loeb 2000). Candidates at other loci include \textit{Trp53}, \textit{Rassf3}, and \textit{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) (Figure 2 and Figure 3).

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 sometimes display hematological abnormalities (Crossan et al. 2011; 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].
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 at this locus with a possible 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 (Figure 1C). We tested for gene-by-sex interactions at all QTL 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 × 10^{-4}). The QTL on Chr 8 is present in both sexes while the QTL 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 (Figure 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

Figure 3: Genome-wide significant loci for micronucleus formation on chromosomes 10, 11 and 17. The −log10 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 color. The plots were drawn using LocusZoom (Pruim et al. 2010).
<table>
<thead>
<tr>
<th>Mouse genomic location</th>
<th>Nearest gene</th>
<th>Consequence</th>
<th>Log P value</th>
<th>GERP Score</th>
<th>Grantham classification</th>
<th>Grantham score</th>
<th>SIFT class</th>
<th>Lowest SIFT score</th>
<th>EnsEMBL VEP class</th>
<th>EnsEMBL aa change</th>
<th>AA Change</th>
<th>Nuc Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr8:33576360-33576360</td>
<td>Tex15</td>
<td>A &gt; T</td>
<td>12.48026273</td>
<td>-1.05</td>
<td>Radical</td>
<td>110</td>
<td>Tolerated</td>
<td>0.4</td>
<td>Downstream gene variant; missense variant</td>
<td>R/S</td>
<td>R1939S</td>
<td>A5817T</td>
</tr>
<tr>
<td>chr8:33268792-33268792</td>
<td>Wm</td>
<td>G &gt; A</td>
<td>12.45662832</td>
<td>-3.07</td>
<td>Radical</td>
<td>145</td>
<td>Tolerated</td>
<td>0.3</td>
<td>Missense variant</td>
<td>S/L</td>
<td>S1021L</td>
<td>C3062T</td>
</tr>
<tr>
<td>chr8:33557389-33557389</td>
<td>Tex15</td>
<td>A &gt; C</td>
<td>12.43606688</td>
<td>3.19</td>
<td>Radical</td>
<td>110</td>
<td>Tolerated</td>
<td>0.44</td>
<td>NMD transcript variant; downstream gene variant; missense variant</td>
<td>S/R</td>
<td>S160R</td>
<td>A478C</td>
</tr>
<tr>
<td>chr8:33557366-33557366</td>
<td>Tex15</td>
<td>G &gt; T</td>
<td>12.4319775</td>
<td>3.14</td>
<td>Radical</td>
<td>205</td>
<td>Tolerated</td>
<td>0.13</td>
<td>NMD transcript variant; downstream gene variant; missense variant</td>
<td>C/F</td>
<td>C152F</td>
<td>G455T</td>
</tr>
<tr>
<td>chr8:33546343-33546343</td>
<td>Tex15</td>
<td>C &gt; G</td>
<td>12.42045548</td>
<td>3.56</td>
<td>Nonconservative</td>
<td>60</td>
<td>Deleterious; tolerated</td>
<td>0.02</td>
<td>NMD transcript variant; missense variant</td>
<td>A/G</td>
<td>A94G</td>
<td>C281G</td>
</tr>
<tr>
<td>chr17:34804124-34804124</td>
<td>Cyp21a1</td>
<td>A &gt; G</td>
<td>8.239597012</td>
<td>4.11</td>
<td>Conservative</td>
<td>5</td>
<td>Deleterious</td>
<td>0.01</td>
<td>Downstream gene variant; missense variant; upstream gene variant</td>
<td>V/A</td>
<td>V71A</td>
<td>T212C</td>
</tr>
<tr>
<td>chr17:34711526-34711526</td>
<td>Tnxb</td>
<td>G &gt; A</td>
<td>8.19865514</td>
<td>4.53</td>
<td>Radical</td>
<td>125</td>
<td>Deleterious; tolerated</td>
<td>0.05</td>
<td>Missense variant</td>
<td>G/R</td>
<td>G397R</td>
<td>G1189A</td>
</tr>
<tr>
<td>chr17:34711722-34711722</td>
<td>Tnxb</td>
<td>G &gt; A</td>
<td>8.188940007</td>
<td>-9.05</td>
<td>Conservative</td>
<td>29</td>
<td>Deleterious</td>
<td>0.02</td>
<td>Missense variant</td>
<td>R/H</td>
<td>R462H</td>
<td>G1385A</td>
</tr>
<tr>
<td>chr17:34713138-34713138</td>
<td>Tnxb</td>
<td>T &gt; A</td>
<td>8.185003889</td>
<td>-2.37</td>
<td>Radical</td>
<td>113</td>
<td>Tolerated</td>
<td>0.21</td>
<td>Missense variant</td>
<td>L/Q</td>
<td>L550Q</td>
<td>T1649A</td>
</tr>
<tr>
<td>chr17:35161042-35161042</td>
<td>Prcc2a</td>
<td>T &gt; G</td>
<td>8.088956801</td>
<td>4.11</td>
<td>Conservative</td>
<td>15</td>
<td>Deleterious</td>
<td>0.02</td>
<td>Downstream gene variant; missense variant; upstream gene variant</td>
<td>M/L</td>
<td>M249L</td>
<td>A745C</td>
</tr>
<tr>
<td>chr11:69236603-69236603</td>
<td>Gucy2e</td>
<td>C &gt; G</td>
<td>7.758017279</td>
<td>-9.38</td>
<td>Radical</td>
<td>125</td>
<td>Tolerated</td>
<td>0.46</td>
<td>Missense variant; upstream gene variant</td>
<td>G/R</td>
<td>G15R</td>
<td>G43C</td>
</tr>
<tr>
<td>chr11:69047870-69047870</td>
<td>Aurkb</td>
<td>T &gt; C</td>
<td>7.543972828</td>
<td>0.726</td>
<td>Radical</td>
<td>155</td>
<td>Tolerated</td>
<td>0.85</td>
<td>Missense variant</td>
<td>F/S</td>
<td>F75S</td>
<td>T224C</td>
</tr>
<tr>
<td>chr11:69134001-69134001</td>
<td>Aloxe3</td>
<td>G &gt; A</td>
<td>6.687629704</td>
<td>3.92</td>
<td>Conservative</td>
<td>56</td>
<td>Deleterious</td>
<td>0.01</td>
<td>NMD transcript variant; missense variant; upstream gene variant</td>
<td>G/S</td>
<td>G341S</td>
<td>G1021A</td>
</tr>
<tr>
<td>chr11:69107564-69107564</td>
<td>Perf</td>
<td>C &gt; T</td>
<td>6.65511366</td>
<td>2.92</td>
<td>Nonconservative</td>
<td>64</td>
<td>Deleterious</td>
<td>0.01</td>
<td>Downstream gene variant; missense variant</td>
<td>A/V</td>
<td>A1014V</td>
<td>C3041T</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Mouse genomic location</th>
<th>Nearest gene</th>
<th>Consequence</th>
<th>Log P value</th>
<th>GERP Score</th>
<th>Grantham classification</th>
<th>Grantham score</th>
<th>SIFT class</th>
<th>Lowest SIFT score</th>
<th>Ensembl VEP class</th>
<th>AA change</th>
<th>Nuc change</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr9:109061596-109061596</td>
<td>Atrip</td>
<td>A &gt; T</td>
<td>5.330403284</td>
<td>−0.969</td>
<td>Radical</td>
<td>113</td>
<td></td>
<td>NMD transcript variant; downstream gene variant; missense variant; synonymous variant; upstream gene variant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr9:109073661-109073661</td>
<td>Atrip</td>
<td>T &gt; A</td>
<td>5.271500634</td>
<td>0.625</td>
<td>Radical</td>
<td>152</td>
<td>Deleterious</td>
<td>0.03</td>
<td>NMD transcript variant; downstream gene variant; missense variant; splice region variant; upstream gene variant</td>
<td>D/V</td>
<td>NA</td>
</tr>
<tr>
<td>chr9:109432436-109432436</td>
<td>Fbxw16</td>
<td>A &gt; G</td>
<td>5.023943697</td>
<td>0</td>
<td>Radical</td>
<td>155</td>
<td>Tolerated</td>
<td>0.56</td>
<td>Missense variant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2:30174134-30174134</td>
<td>D2Wsu81e</td>
<td>C &gt; T</td>
<td>4.890540776</td>
<td>4.1</td>
<td>Radical</td>
<td>125</td>
<td>Tolerated</td>
<td>0.34</td>
<td>Downstream gene variant; missense variant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2:30174134-30174134</td>
<td>D2Wsu81e</td>
<td>C &gt; T</td>
<td>4.890540776</td>
<td>4.1</td>
<td>Radical</td>
<td>125</td>
<td>Tolerated</td>
<td>0.34</td>
<td>Downstream gene variant; missense variant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2:29702527-29702527</td>
<td>Rapgef1</td>
<td>G &gt; T</td>
<td>4.806085521</td>
<td>1.45</td>
<td>Conservative</td>
<td>24</td>
<td>Deleterious; tolerated</td>
<td>0.02</td>
<td>NMD transcript variant; missense variant;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2:29702527-29702527</td>
<td>Rapgef1</td>
<td>G &gt; T</td>
<td>4.806085521</td>
<td>1.45</td>
<td>Conservative</td>
<td>24</td>
<td>Deleterious; tolerated</td>
<td>0.02</td>
<td>NMD transcript variant; missense variant;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2:29788325-29788325</td>
<td>Coq4</td>
<td>G &gt; T</td>
<td>4.57110541</td>
<td>0</td>
<td>Radical</td>
<td>102</td>
<td>Tolerated</td>
<td>0.32</td>
<td>NMD transcript variant; missense variant; upstream gene variant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr9:109060491-109060491</td>
<td>Atrip</td>
<td>T &gt; G</td>
<td>3.712727076</td>
<td>0.625</td>
<td>Conservative</td>
<td>15</td>
<td>Deleterious; tolerated</td>
<td>0.03</td>
<td>NMD transcript variant; downstream gene variant; missense variant; upstream gene variant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr5:148288927-148288927</td>
<td>Mtus2</td>
<td>C &gt; A</td>
<td>3.699186538</td>
<td>0</td>
<td>Radical</td>
<td>110</td>
<td></td>
<td>NMD transcript variant; missense variant; synonymous variant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr5:148288935-148288935</td>
<td>Mtus2</td>
<td>T &gt; C</td>
<td>3.692862836</td>
<td>0</td>
<td>Radical</td>
<td>145</td>
<td></td>
<td>NMD transcript variant; missense variant; synonymous variant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr9:109058117-109058117</td>
<td>Trex1</td>
<td>G &gt; C</td>
<td>3.215586866</td>
<td>−1.25</td>
<td>Radical</td>
<td>125</td>
<td>Tolerated</td>
<td>0.18</td>
<td>Downstream gene variant; missense variant</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shown are the top five scoring variants at each of the seven genome-wide significant loci ranked by $-\log_{10}$ P-value. See Materials and Methods for a description of the approach used for variant annotation.
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 population, 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 thus potentially contribute to elevated micronucleus levels. We used a combination of approaches to assess the pathogenicity of variants by using ANNOVAR, applying GERP scores, and by performing a Grantham analysis. We also used 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 (Figure 4). We projected this position onto the human AURKB protein revealing that it falls into a highly conserved residue of AURKB that is polymorphism (SNP) in Aurkb (rs29417126) falls into a residue previously found to be mitotically phosphorylated. 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 Tp53, 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 Tp53 gene in the CFW population, 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 Tp53 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 QTL, 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 find 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, Nfkbil1, Trub2, and H2-Eb1) available for testing using the micronucleus assay. Our analysis of potentially pathogenic variants within QTL (Table 2) revealed a nonsynonymous 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 (Figure 2 and Figure 4). The micronucleated erythrocyte frequency for Nfkbil1, Trub2, and H2-Eb1 mutants was not significantly different to that of wild-type 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 inbred mouse cell lines (RPE-1) in which chromatin (H2B) was labeled with mCherry in the context of TREX1 disruption, and in a matched control (TREX1 wild-type) line (Figure 4). Analysis of 808 cell divisions from TREX1 mutant cells and 405 from wild-type 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 process 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 hematological measures and contains a gene (Slc7a1) causing severe anemia in the mouse. Further, functional evidence in vivo in a mouse mutant supports the role of Trex1 in the formation of extranuclear 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.

ACKNOWLEDGMENTS

This work was supported by Cancer Research UK and the Wellcome Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Author contributions: R.E.M., J.N., J.M., N.C., V.I., A.G.R., R.M., and C.D.R.E. performed micronucleus assays and statistical analysis. J.F., J.N., R.M., D.J.A., and R.E.M. wrote the paper.

LITERATURE CITED
