Strand asymmetry influences mismatch repair during single-strand annealing

Victoria O. Pokusaeva[1,2], Aránzazu Rosado Diez[1,3], Lorena Espinar[1], Guillaume J. Filion[1,4,5,*]

[1] Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
[2] Present address: Institute of Science and Technology Austria, Am Campus 1, Klosterneuburg, Austria
[3] Present address: H12O-CNIO Lung Cancer Clinical Research Unit, i + 12 Research Institute, Spanish National Cancer Research Center (CNIO), Madrid, Spain
[4] University Pompeu Fabra (UPF), Barcelona, Spain
[5] Present address: Dept. Biological Sciences, University of Toronto Scarborough
* to whom correspondence should be addressed: guillaume.filion@gmail.com

ABSTRACT

Biases of DNA repair can shape the nucleotide landscape of genomes at evolutionary timescales. However, such biases have not yet been measured in chromatin for lack of technologies. Here we develop a genome-wide assay whereby the same DNA lesion is repaired in different chromatin contexts. We insert thousands of barcoded transposons carrying a reporter of DNA mismatch repair in the genome of mouse embryonic stem cells. Upon inducing a double-strand break between tandem repeats, a mismatch is generated when the single strand annealing repair pathway is used. Surprisingly, the mismatch repair machinery favors the same strand 60-80% of the time. The location of the lesion in the genome and the type of mismatch have little influence on the repair bias in this context. Using machine learning, we further show that both the repair bias and the efficiency of the repair are independent of known chromatin features. These results suggest that some intrinsic property of the lesion can have a large influence on the outcome of DNA repair, irrespective of the surrounding chromatin context.

INTRODUCTION

The genome of every organism is the result of a mutation-selection process that unfolds since the origins of life. Mutations have a dual role in this process: on the one hand they generate the diversity for selection to act upon, on the other hand they drive evolution through non-selective forces (Filipski 1990). Non-selective forces are changes that drive a genome away from its current state without affecting the fitness of the organism. For instance, small asymmetries in mutations can accumulate over evolutionary timescales so as to form sequence patterns in a genome (Freese 1962; Sueoka 1962).

At least two features of mammalian genomes are shaped by non-selective forces: the depletion of CpG dinucleotides (Sinsheimer 1955) and the 10 bp periodicity of ApA dinucleotides (Gale, Nissen, and Smerdon 1987). The first is due to increased C to T mutations when C is methylated, which takes place only within CpG dinucleotides (Rideout *et al*. 1990; Hwang and Green 2004). The second is due to increased damage on nucleotides facing outward the nucleosome, where ApA is the least exposed dinucleotide (Ramstein and Lavery 1988; Filipski 1990; A. J. Brown *et al*. 2018).

The mechanisms that underlie mutational biases in mammals are otherwise poorly understood. For instance, one of the most enigmatic features of mammalian, and more generally, vertebrate genomes is that they are organized in megabase-scale domains called isochores, where the GC-content is relatively uniform (Bernardi *et al*. 1985). However, the average GC-content varies from 30% to 70% between isochores. Several lines of evidence suggest that the pattern may emerge from an asymmetry in meiotic recombination known as biased gene conversion (Galtier *et al*. 2001; Duret and Galtier 2009). The theory postulates that mismatched heteroduplexes are repaired in favor of G/C alleles over A/T alleles, with the consequence that the GC-content increases at recombination hotspots (Duret, Eyre-Walker, and Galtier 2006).

Initial estimates in mammalian genomes suggested that mismatch repair is indeed biased toward G and C nucleotides (T. C. Brown and Jiricny 1989), but those estimates were obtained on circular unintegrated plasmids. In more realistic conditions where heteroduplexes are integrated and repaired in the genome, the bias disappeared, except for the G:T mismatch, handled by a specific repair pathway (Bill *et al*. 1998). It is thus doubtful that any nucleotide is intrinsically favored by the mismatch repair system. Instead, it appears that there exists a hierarchy of factors influencing repair biases, but which take precedence over the others is largely unknown.

Recent insight into this question came from cancer genome sequencing (Pleasance *et al*. 2010; Hoadley *et al*. 2014). In particular, this made it possible to show that the mismatch repair system in healthy cells is more accurate at some loci than others. For instance, mismatches in late-replicating regions are repaired less efficiently (Supek and Lehner 2015), a feature that seems to be shared among eukaryotes (Weber, Pink, and Hurst 2012). It is presently unknown whether the chromatin context can bias the mismatch repair toward one nucleotide or another, mostly because it is difficult to tease apart the contributions of DNA damage and DNA repair to mutation patterns.

In sum, the fact that DNA repair is context-dependent suggests that it may have a large influence on the local nucleotide composition of a genome. However, the importance of chromatin compared to the molecular features of the lesion is an open question. More generally, it has been so far impossible to measure the biases of mismatch repair in chromatin separately from DNA damage for lack of technologies to engineer and track mismatches genome-wide.

Here we set out to measure such biases in the chromatin of wild-type mouse embryonic stem cells (ES cells). We develop an assay where a mismatch is produced in the genome as a byproduct of the single-strand annealing pathway (SSA). Using reporters integrated at tens of thousands of locations, we pit nucleotides against each other and measure the effect of the chromatin context on the outcome of mismatch repair. This allows us to directly test the hypothesis that the mismatch repair pathway favors G/C alleles over A/T alleles in different genomic contexts.

RESULTS

A TRIP assay to measure mismatch repair biases in SSA

TRIP (Thousands of Reporters Integrated in Parallel) is a shotgun technique to assay the influence of the chromatin context on a phenomenon of interest (Akhtar *et al*. 2013; Corrales *et al*. 2017; Gisler *et al*. 2019). The principle is to insert reporters at different locations of the genome, and to measure a readout in bulk using DNA barcodes. The experiment typically consists of two phases: in the first, transposons are inserted in a cell pool and the barcodes are mapped to generate a lookup table; in the second, the phenomenon of interest is measured for all the barcodes simultaneously and the outcome is demultiplexed using the lookup table.

In this study we developed a TRIP assay to measure biases of DNA mismatch repair (**Figure 1**). The reporter construct consists of two nearly identical 152 bp sequences separated by a restriction site for the meganuclease I-SceI. As the 152 bp sequences were originally a central segment of the GFP gene, we refer to them as F segments throughout. The F segments differ by one nucleotide located at the center, so annealing two strands from different F segments creates 152 bp heteroduplex with a central mismatch. The assay is initiated by expressing I-SceI, which cleaves the integrated reporters. The double-strand breaks are repaired by either non-homologous end joining (NHEJ) or single-strand annealing (SSA). In the first case, DNA ends are blunted and ligated so the final product consists of two F segments with distinct alleles flanking the scarred I-SceI site. In the second case, 5' DNA ends are resected and the two strands anneal to each other, forming a mismatched duplex that is eventually repaired (Spies and Fishel 2015). The final product contains only one F segment with only one of the two original alleles.

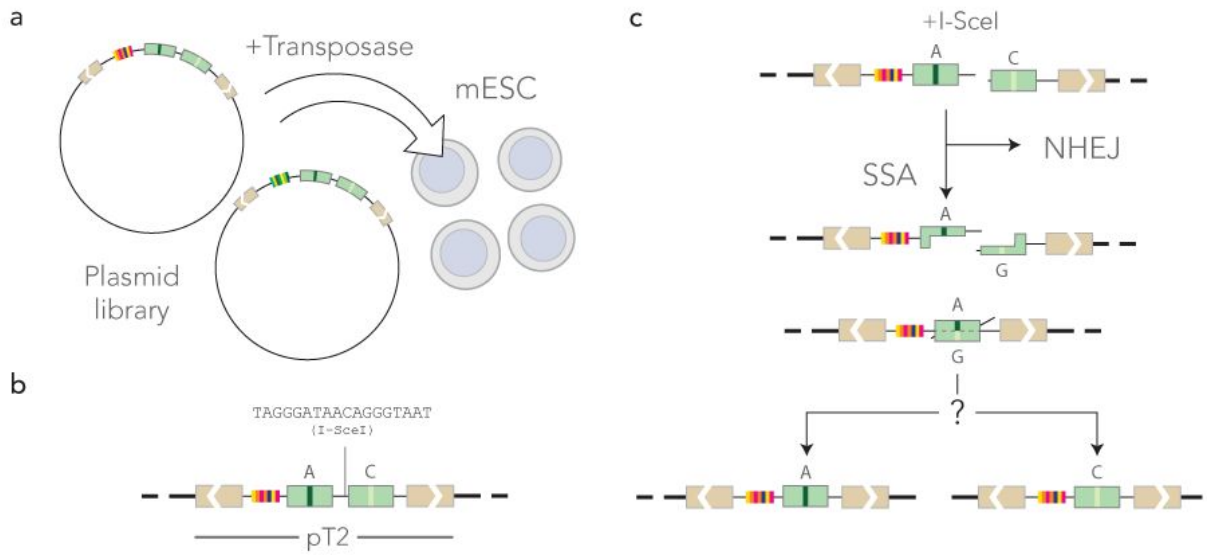

Figure 1. Experimental approach. **a)** mouse ES cells in culture are co-transfected with a barcoded reporter library and the Sleeping Beauty 100X transposase. **b)** The reporters in the pT2 transposon backbone are integrated at random in the mouse genome. **c)** Mismatches occur during the repair of a double-strand break induced by the transient expression of I-SceI. If the double-strand break is repaired through Non Homologous End Joining (NHEJ), no mismatch is formed. If it is repaired by Single Strand Annealing (SSA), a mismatch is formed and repaired in favor of one of the two nucleotides. Sequencing the construct reveals the outcome of DNA repair at different locations identified with the barcode.

The outcome of mismatch repair is revealed by sequencing the reporters that have only one F segment. The barcode flanking the sequence allows us to know the location of the reporter in the genome of mouse ES cells thanks to the lookup table.

The integrated reporters cover the mouse genome

We designed four similar constructs where SSA produces heteroduplexes with A:G, T:G, A:C and T:C mismatches. In those constructs, the A and T alleles are on the left F segment in the orientation of **Figure 1b,** while G and C alleles are on the right one. This means that during heteroduplex formation, A/T nucleotides are always of the top strand in the orientation of **Figure 1c**, and G/C nucleotides on the bottom one. We therefore included a strand-swap control of the A:C mismatch, referred to as C:A, where C is on the top strand and A is on the bottom one. For concreteness, we will refer to the strands as top and bottom in what follows, always assuming that the reference orientation is that of **Figure 1b**.

The five constructs were barcoded using random 20-mers so that each integrated reporter contains a different barcode (see Methods). The barcoded TRIP reporter libraries were inserted by Sleeping Beauty transposition in the genome of E14 mouse ES cells (Mátés et al. 2009). After two weeks of growth, the reporters were mapped by inverse PCR (see Methods). The number of unambiguously mapped reporters varied from 9 to 50 thousand, for a total around 120 thousand (see **Table 1**).

Construct	*Mapped*	*In repeats*	*Repair events*	*Mapped events*
A:G	20,240	6,761	5,522	795
T:G	9,570	2,123	29,499	1,388
A:C	13,071	4,188	4,754	614
T:C	26,843	9,943	3,120	692
C:A	50,769	10,466	51,5848	26,832
Total	120,493	33,481	94,749	30,321

Table 1. Mapping and repair statistics. *Construct*: code of the mismatch produced during SSA, with the convention that the left nucleotide is proximal to the barcode (for instance, the construct in **Figure 1b** is A:G). *Mapped*: number of barcodes unambiguously mapped in the mouse genome. *In repeats*: number of barcodes with ambiguous location in repeated sequences. *Repair events*: number of repair events observed on all the barcodes. *Mapped events*: number of repair events observed on unambiguously mapped barcodes.

At chromosomal scale, the transposons were found everywhere except on the Y chromosome (**Figure 2a**). E14 ES cells are male, but the repetitive nature of the Y chromosome makes it impossible to map the reporters unambiguously. The insertion rate on the X chromosome was approximately half the value observed on the autosomes, in line with the expectation for a male cell line (**Figure 2b**). This shows that the reporters were distributed evenly throughout the mappable genome.

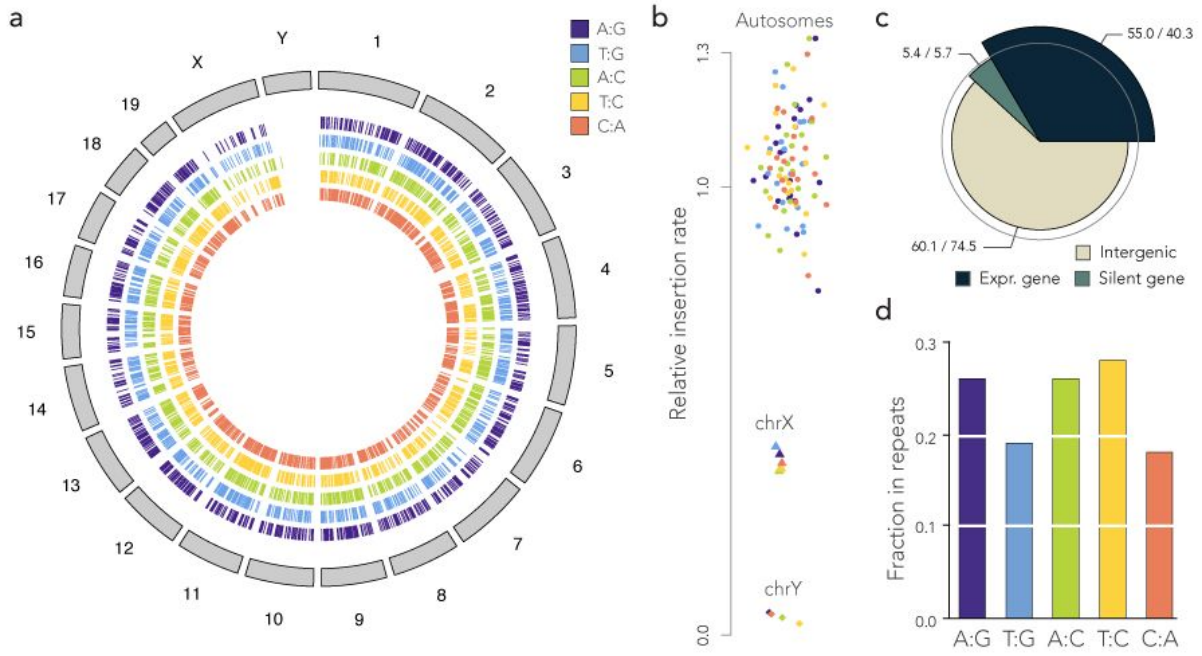

Figure 2. Mapping of the reporters. **a)** Overview of the insertions. For each construct, 500 insertions were drawn at random and plotted on a circular representation of the mouse genome. Each tick mark represents a mapped insertion. **b)** Dot plot of the relative insertion rate per chromosome. For each chromosome and each construct, the insertion rate was computed as the number of insertions per bp, normalized by the expected number of insertions under the uniform model. **c)** Insertions relative to genes. The spie chart represents the global number of observed vs expected insertions inside and

outside genes. The area of a wedge is proportional to the observed value and its angle is proportional to the expected value (so depleted categories are within the grey circle and enriched categories protrude outside). The numbers represent observed over expected insertions, expressed in thousand. **d)** Insertion sites in repeats. The bar plot shows the proportion of barcodes mapping to ambiguous locations in repeated sequences (see **Table 1**).

Overall, the mapped reporters were enriched in transcribed genes, with a 35% excess over random (**Figure 2c**). Genes are typically more mappable than the rest of the genome, but mapped reporters were not enriched in silent genes, suggesting that ongoing transcription facilitates the insertion of the transposon. Meanwhile, reporters were depleted from intergenic regions with a 20% deficit over random. In conclusion, insertion biases toward transcribed chromatin are present but minor.

Some parts of the mouse genome are unmappable. In particular, the reporters inserted in repeated sequences such as retrotransposons cannot be assigned to a unique location. The proportions of barcodes ambiguously mapped to repeated sequences were similar between experiments (**Figure 2d**). More generally, not all the barcodes could be mapped, so most of the repair events occurred at unknown locations (**Table 1**). Yet, the subset that was mapped shows that the integrated reporters cover the mouse genome with sufficient uniformity to study regional repair biases.

Mismatch repair on the reporters is strand-biased

To quantify the outcome of DNA repair on the integrated reporter, we designed two sequencing assays based on Unique Molecular Identifiers (UMIs). The DNA extracted from the cell pools was first digested with I-SceI in order to eliminate the reporters that were not cleaved in ES cells—repairing the double-strand break destroys the I-SceI restriction site. Using primers decorated with UMIs, we then performed either 50 cycles of linear amplification or 6 cycles of PCR (**Figure 3a**). The reasons for using UMIs are twofold: First, they make the quantification more robust (random fluctuations in the first PCR cycles can have large effects on the read numbers). Second and more importantly, they were used to mitigate template switching (Meyerhans, Vartanian, and Wain-Hobson 1990), a common artefact in PCR that can potentially shuffle the barcodes between templates and make quantifications inaccurate. The UMIs were inserted opposite to the barcode, so that we could discard UMIs associated with multiple barcodes (those can occur only through template switching). Counting UMIs thus provides more accurate measurements than counting sequencing reads.

Figure 3b shows the global repair bias toward A/T for each construct 24 and 48 hours post I-SceI induction (the barcodes detected before I-SceI induction are discarded, see below). This represents a total of almost 95 thousand repair events from mapped and unmapped barcodes (**Table 1**). Mismatch repair was reproducibly biased in all the tested conditions, with a bias in the range 60-80% toward A/T for the first four constructs, and 17% for the last. For each construct, the biases were similar between replicates, between measurement methods and between time points, showing that the assays are reproducible in the given experimental conditions.

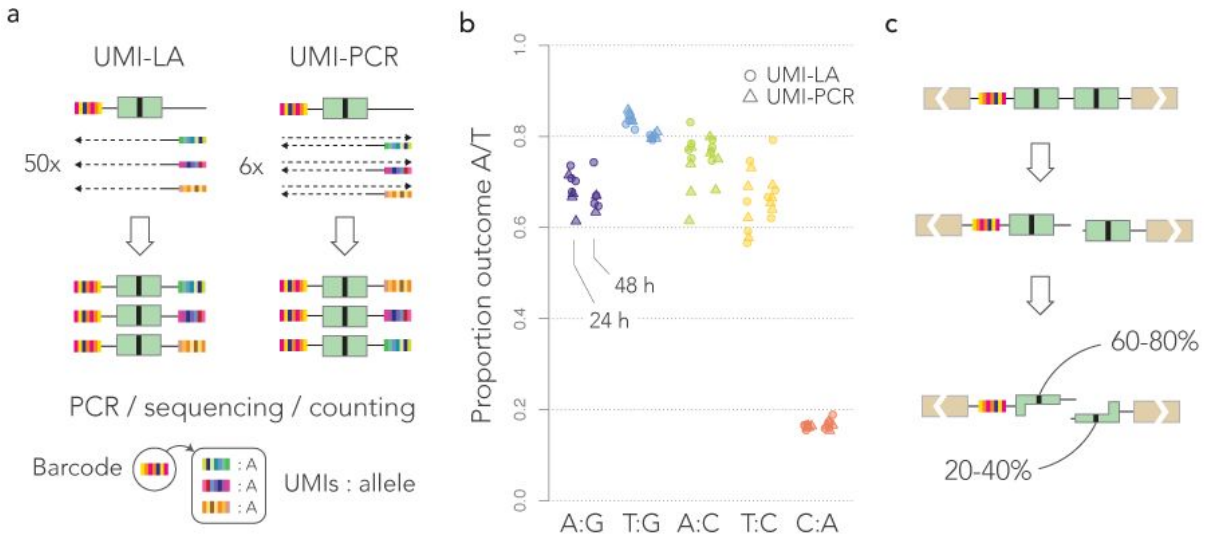

Figure 3. Measure of repair biases. **a)** Quantification methods. In UMI-LA (left), barcoded reporters are amplified by 50 cycles of linear amplification using a primer decorated by UMIs. In UMI-PCR, reporters are amplified by 6 PCR cycles where one primer is decorated by UMIs. Either way, the products are further amplified by regular PCR. After sequencing, each barcode is associated to several UMIs, themselves associated to alleles. Repair biases are quantified by giving one vote per UMI. **b)** Repair outcome. The dot plot shows the measured bias toward A or T (whichever applies) in all the replicates. For each construct, data points obtained 24 and 48 hours post I-SceI induction are shown on the left and on the right, respectively. **c)** Graphical summary of the results. The nucleotide of the top strand is most frequently kept by the mismatch repair system.

Strikingly, the dominant outcome did not correspond to a nucleotide but to a strand. Indeed, the A:C mismatch was resolved in favor of A in the A:C construct (green), but in favor of C in the C:A construct (red) where the nucleotides were swapped. The measured repair biases are roughly symmetric between the two constructs (75% vs 17%). We therefore conclude that in this assay, the top strand is more likely to be used as a template during mismatch repair.

The magnitudes of the repair bias in favor of A or T when they are on the top strand (purple, cyan, green and yellow) are comparable to each other. This means that in the present context, the nature of the mismatch has less influence than the nature of the allele on the top strand. Taken together, these results suggest that mismatch repair during SSA can be strongly biased toward a strand, regardless which nucleotides are mismatched (**Figure 3c**).

All the reporters have similar strand biases

Is the 60-80% bias above a typical value for most reporters or a conflated average? We addressed this question in several steps by leveraging the properties of the UMI-amplicons.

First we made sure that the reporters are cleaved by I-SceI. Here we took advantage of the fact that the same barcode is sometimes found in different UMI-LA or UMI-PCR replicates. The repair events observed without I-SceI induction tend to be identical among replicates (**Figure 4a**). Those represent barcodes from reporters where the F segments have recombined earlier than I-SceI induction (e.g., during the preparation of the barcoded library), they are replicated through cell division and therefore are all idendentical. In

contrast, the repair events observed 24 and 48 hours post I-SceI induction can differ between replicates, even if they take place on identical integrated reporters. This shows that the repair events observed upon I-SceI induction are distinct from the spurious repair events that took place earlier.

Second, we ruled out that the mismatches are unrepaired at the time they are assayed. If a mismatch is not repaired, the two strands are not complementary so UMI-LA and UMI-PCR give different results (**Figure 4b**). UMI-LA uses only one strand as a template (the top one), so all the associated UMIs must report the same outcome for a given barcode, even when the mismatch is not repaired. In contrast, UMI-PCR uses both strands as a template, so the UMIs are associated with both nucleotides of the unrepaired mismatch. Therefore, UMI-LA and UMI-PCR should be strongly discordant if they are used to amplify unrepaired mismatches.

As noted in **Figure 3b**, UMI-LA and UMI-PCR are concordant, suggesting that the mismatches are repaired. To confirm this conclusion, we measured the proportion of barcodes with conflicting UMIs (**Figure 4c**). UMI-PCR did not produce more conflicts than UMI-LA, confirming that the repair biases observed in **Figure 3b** apply to fully repaired mismatches.

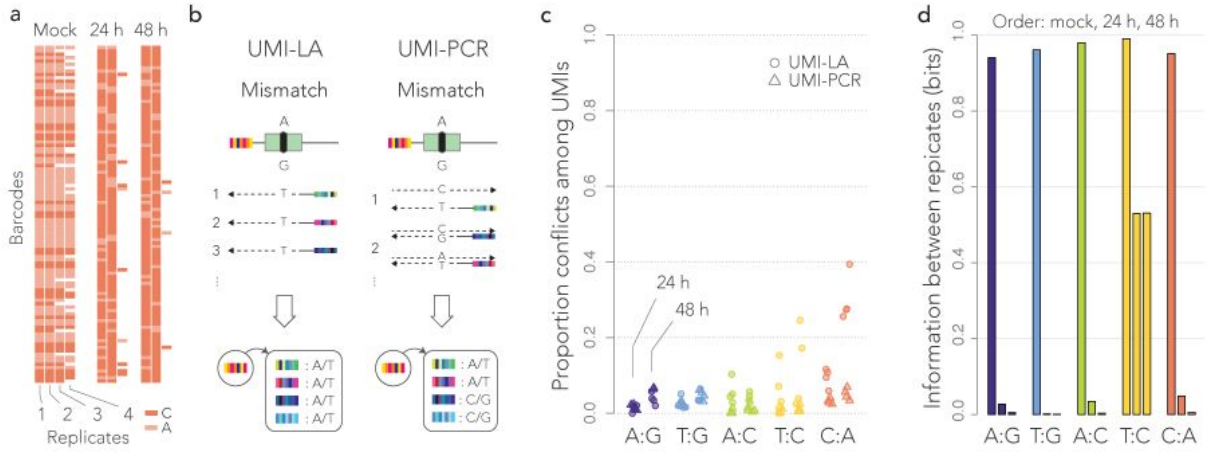

Figure 4. Global differences between reporters. **a)** Repair across replicates construct. Each row shows a random barcode from the C:A construct and each column shows a replicate where it appears. The color of the rectangle indicates the repair outcome. Without I-SceI induction (Mock), barcodes are typically present in more than two replicates, always with the same outcome. 24 or 48 h post I-SceI induction, barcodes are typically present in two replicates with different outcomes. The results are similar for all the constructs. **b)** Amplification of unrepaired mismatches. If the mismatch is unrepaired, UMI-LA (left) produces UMIs with the allele of the top strand only, whereas UMI-PCR produces UMIs with both alleles. **c)** Conflicts among UMIs. The dot plot shows the proportion of barcodes such that at least one UMI reports a different allele than the majority. Colors and symbols are the same as in **Figure 3b**. **d)** Mutual information between replicates. The bar graph shows the average mutual information per barcode between pairs of replicates (experiments with the T:C construct had only 98 pairs, compared to > 1000 pairs for the other constructs and the no I-sceI controls).

Note that in **Figure 4c** the great majority of barcodes do not have a single conflicting UMI. This suggests that the input DNA in UMI-based assays consists of just one molecule per

barcode. The alternative would be that copies of the same reporter are consistently repaired the same way, i.e. that reporters have distinct predefined biases and that the 60-80% bias is a conflated average.

To distinguish between these two hypotheses, we computed the mutual information between repair outcomes on the same barcodes (see Materials and Methods). Mutual information is more adequate than the Pearson and Spearman coefficients for categorical variables, and the interpretation is similar in the sense that a value of 0 indicates that the variables are independent. We can thus use this metric to test whether reporters have individual biases: if they do, the repair outcome is partly determined by this individual bias, so knowing how a reporter is repaired in one replicate gives some information about how it is repaired in other replicates (the mutual information is nonzero).

We collected the barcodes appearing in at least two replicates and assigned them to a single dominant repair outcome. For each construct and each time point, we filled a 2x2 contingency table with all the replicate pairs, from which we computed the mutual information shown in **Figure 4d**. Without I-SceI induction, the mutual information between replicates is close to 1, meaning that the reported outcome is always the same for a given barcode, consistently with **Figure 4a** (recall that those repaired reporters were replicated through cell division). In contrast, the mutual information drops to 0 for barcodes that are amplified after I-SceI induction. This means that the repair of a reporter in one replicate has no predictive value for the repair in another replicate. In other words, if a reporter is repaired toward A, say, it is not more likely to be repaired toward A in other experiments (as apparent in **Figure 4d**). Therefore, there exists no group of reporters with a much higher (or lower) bias than average.

Taken together, these results show that in the conditions of our assay, the mismatches occurring through SSA are repaired with a ubiquitous bias toward the top strand. In other words, regardless of their location, the reporters all have the same propensity to repair the mismatch toward this outcome.

Mismatch repair on the reporters is independent of chromatin features

Our results so far indicate that mismatch repair on the reporters shows a universal bias toward the top strand. If the bias is an intrinsic property of the repair process on this construct, it should not depend on any local property of the chromatin at or near the site of the integrated reporter. One way to test this hypothesis directly is to use a model-loose machine learning approach such as deep learning to evaluate how well the information about chromatin predicts the local repair bias.

To this end, we designed an artificial neural network with a residual network architecture (He *et al*. 2016). The output of hidden layers is batch-normalized (Ioffe and Szegedy 2015) and passed through a standard ReLU activation function (Nair and Hinton 2010). We also included two Dropout steps to mitigate overfitting (Srivastava *et al*. 2014). Finally, the output layer projects onto the parameters of a Beta distribution, so that the network can be used to predict proportions between 0 and 1 (Sadowski and Baldi 2018). The architecture of the network is sketched in **Figure 5a**.

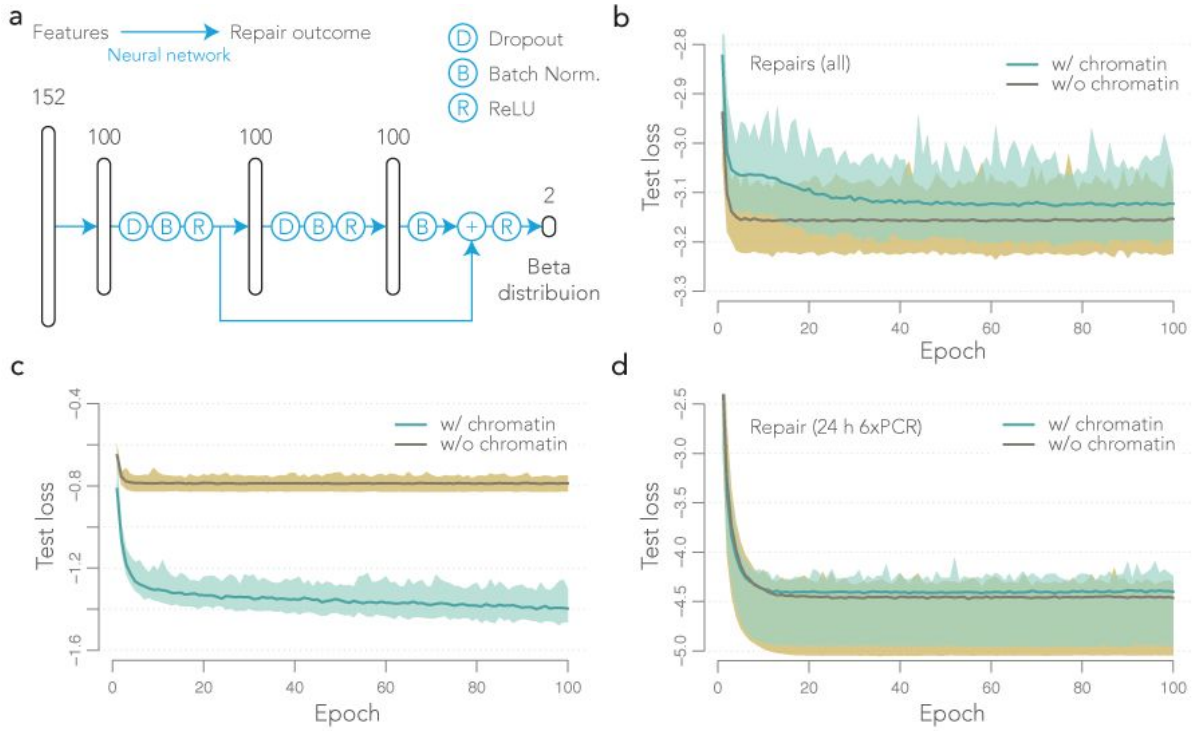

Figure 5. Machine learning approach to fit repair outcome from chromatin features. **a)** Sketch of the architecture. Vertical cartridges represent fully connected neuron layers with indicated dimensions. The blue arrows indicate the forward information flow. **b)** Learning curves for the full data set. The median loss on the test set is shown in green for 100 learnings with random restart. The interval between the 1st and 99th percentile is shown with the green shaded area. The median loss for the null model without chromatin features is shown in brown, and percentiles are shown with the brown shaded area. **c)** Learning curves for the local GC-content. The repair outcome was replaced by the GC-content in a 2 kb window around the reporter. **d)** Learning curves for the occurrence of repair. The only observations used for learning are UMI-PCR 24 h post I-SceI induction.

We trained this network to predict the outcome of mismatch repair on the mapped reporters from a compendium of chromatin proteins and histone marks mapped in mouse ES cells (Juan et al. 2016). The dataset combines ChIP-seq, meDIP, and GLIB assays for 3 cytosine modifications, 13 histone marks, and 61 chromatin proteins. We also added the local GC-content of the genome, and the local level of gene expression, for a total of 152 variables. The data set consisted of 30,321 observations (**Table 1**), of which 10% were kept for testing (see Materials and Methods). We performed 100 independent trainings with random restart. The median performance of the network on the test set at different epochs of training is shown in **Figure 5b**. For comparison, we included 100 trainings of a null model where the chromatin features were removed (this model can only learn the genome-wide averages per experiment). Surprisingly, the full model performs less well than the null model, meaning that chromatin features have no predictive value for the outcome of mismatch repair.

To confirm that the network has an architecture that is suitable for learning, we predicted the local GC-content from the 151 remaining chromatin features. In this case the full model clearly outperforms the null model and still shows some evidence of learning after 100 epochs (**Figure 5c**), indicating that slow learning or overfitting are not a concern on this data

set. Instead, we conclude that the network fails to predict the outcome of mismatch repair because the chromatin context has little influence on the process.

Finally, we asked whether the efficiency of mismatch repair itself depends on the chromatin context. To address this question, we took advantage of two facts established previously: The first is that the input DNA usually consists of one molecule per barcode; the second is that UMI-PCR is expected to produce mixed populations of UMIs if the mismatch is not repaired, as mentioned above. We thus used the network to predict the ratios of UMIs assayed by UMI-PCR 24 hours post I-SceI induction (**Figure 5d**). In this case, the performance of the full model was similar to that of the null model, indicating that the efficiency of mismatch repair also does not depend on the chromatin context.

Barcodes mapped to ambiguous locations (**Figure 2d**) did not contribute to this analysis, so it is possible that repair proceeds differently in repeated regions. However, this is unlikely because **Figure 4d** suggests that reporters have no individual biases. In summary, our results show that the repair of the reporters is unconditionally biased. Neither the efficiency of mismatch repair nor the magnitude of the repair bias depends on known elements of the chromatin context.

DISCUSSION

Here we used a TRIP assay to study the process of DNA repair in the chromatin of mouse ES cells. Our construct is designed to produce a mismatch if the reporter is repaired through the SSA pathway, allowing us to study how the same mismatch is repaired at different loci. With TRIP reporters inserted throughout the genome (**Figure 1** and **Table 1**), we obtained a global view of the DNA mismatch repair landscape. We found no evidence that repair is intrinsically biased toward G and C nucleotides. Instead, we found a persistent 60-80% bias toward one of the strands, regardless the mismatch that was induced. We also showed that the repair bias is uniform throughout the genome, suggesting that the chromatin surrounding the lesion has little or no influence in this context. Overall, these results have important implications regarding the factors influencing mismatch repair.

Teasing apart the individual contributions of DNA damage and repair to mutational processes is challenging. This has been possible on plasmids (T. C. Brown and Jiricny 1989), but the known interactions of mismatch repair with chromatin suggest that mutational biases should be studied directly in the genome (Goellner 2019). In this regard, the TRIP assay developed here is a technical step forward. A similar principle was already used by Gisler *et al*. (2019), with the difference that they focused on the repair of double-strand breaks induced by the CRISPR-Cas9 system. Interestingly, they found that the local state of chromatin is somewhat predictive of indel rates. The contrast with our findings suggests that the importance of the chromatin context varies between repair pathways.

It is important to highlight that the mismatches produced in our repair assay are coupled to the repair of a double-strand break. In that sense they differ from mismatches occurring during DNA replication. We nevertheless gained general insight into the mechanism of

mismatch repair. By far the most striking is that the process can be substantially biased. Such bias must *in fine* come from an asymmetry in or around the construct because the bias was constant throughout the genome. Here it is important to highlight that this asymmetry was never intended in the experimental design (the mismatch is at the center of the heteroduplex), so we can only comment on this in hindsight.

The state of the lesion after the two F segments have annealed is sketched in **Figure 6a**. The asymmetry that is responsible for the repair bias may lie in the sequences flanking the F segments (in our case this is mostly the pT2 transposon), the sequence of the heteroduplex, the flaps forming 3' extensions, or the gaps generated upon resection of the 5' ends early in SSA. Among those, only the flaps have an obvious asymmetry in our construct: the 3' extension of the dominant strand is 34 nucleotides, while the 3' extension of the other strand is only 13 nucleotides. This is so because the I-SceI site is not exactly halfway between the two F segments (the reason for this design was to perform a preliminary control in the early stages of this work).

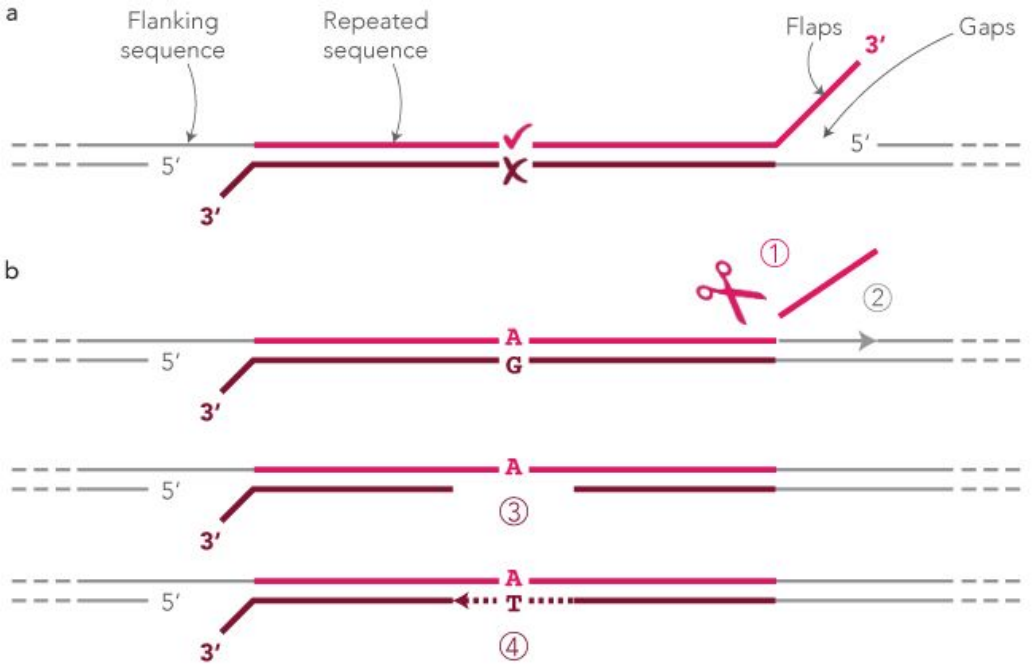

Figure 6. Summary and model. **a)** Sketch of the SSA intermediate. The length of the sequence features is drawn to scale. The 3' flap of the top strand is 34 nucleotides, and that of the bottom strand is 13 nucleotides. The repair bias may be due to some asymmetry in the sequence flanking the F segments, in the sequence of the F segments surrounding the mismatch, in the flaps or in the gaps generated by 5' resection early in SSA. **b)** A possible influence of the flaps. If the time required to clip the flaps depends on their length (1), the gap will be sealed earlier on one side (2), giving a structure with only one damaged strand, thought to be important for discrimination by the mismatch repair system (Pavlov, Mian, and Kunkel 2003; Kunkel and Erie 2005). The allele on the damaged strand would be removed (3) and replaced by that on the intact strand (4).

Could it be that the asymmetry of the flaps influences the repair of the mismatch? It was recently discovered in yeast that the flaps influence the efficiency of mismatch repair in break-induced replication (Anand *et al*. 2017), so there exists a cross-talk between flap excision and mismatch repair. In addition, repair in a common yeast assay to study SSA also

shows an approximately 70% bias toward the strand with the longer flap (Chakraborty *et al*. 2016, Table 3). Importantly, the work showing that there is no repair bias except on G:T mismatches in rodent cells was based on the formation of heteroduplexes without flaps (Bill *et al*. 1998). So overall, the view that the flaps contribute to the bias is compatible with the present state of knowledge.

One can only speculate as to how the flaps may contribute to mismatch repair biases; one plausible scenario is shown in **Figure 6b**. Another possibility is that the nucleotides close to the mismatch have a strong influence. Since the only difference between the heteroduplexes is the mismatch, it is presently impossible to evaluate the contribution of the sequence.

Using the properties of UMI-PCR we could establish that the assayed mismatches were repaired. The DNA cannot be amplified before the flaps are removed and the gaps are sealed (the PCR primers do not anneal), so the mismatch must be repaired shortly thereafter. In any event, the location of the reporter did not seem to influence whether the mismatch would be repaired, which seems to contradict the well established fact that mismatches in late-replicating regions are repaired less efficiently (Weber, Pink, and Hurst 2012; Lujan et al. 2014; Supek and Lehner 2015). It may be that mismatches occurring through SSA are more accessible because of the double-strand break. Alternatively, mismatches in late-replicating regions could be repaired as often as in early-replicating regions, but not as faithfully: toward the end of the S phase, replication forks merge and DNA loses the asymmetry that may be required for the mismatch repair to discriminate the strands .

The potential sources of asymmetry can be tested and studied experimentally. For instance, the length of the flaps can be adjusted by changing the location of the I-SceI site in the construct. Also, the F segments can be replaced by any sequence of interest, so future experiments will shed light on the nature of this repair bias.

In conclusion, the work presented here shows that factors other than the chromatin context have a strong influence on the bias of mismatch repair. We found no evidence that any nucleotide was intrinsically favored in our assay. Instead, we found that the mismatch repair system can be persistently biased. Elucidating the molecular mechanisms of this bias will require future investigations.

METHODS

Plasmid construction and library preparation

Plasmid pCBASceI for I-SceI expression and plasmid pcDNA3.1-mCherry were obtained from Addgene (#26477 and #128744 respectively). Plasmid pCMV(CAT)T7-SB100X for Sleeping Beauty 100X expression was kindly provided by Zsuzsanna Izsvák, plasmid pcDNA3.1-mCherry. FF fragments (each with a precursor for one of heteromismatches) were synthesised by Life Technologies, and cloned into plasmid pT2 using Gibson Assembly Cloning Kit (NEB, E5510S). Obtained pT2_FF plasmids were used as templates for PCR-based barcoded library preparation (Corrales *et al*. 2017).

For barcoding PCR, 100 pg of each of pT2_FF plasmid was used as template in 50 µL Phusion DNA polymerase reaction mix (Thermo Fisher Scientific, F530S) with GC buffer, using PCR primers L1-6 from **Table 2** in the following cycling conditions: 98°C for 1min; 98°C for 30s, 60°C for 30s and 72°C for 3 min (25 cycles); and 72°C for 5 min. The template was destroyed by adding 1 µL 20,000 U/mL DpnI (NEB, R0176S) to the mix and incubation at 37 °C for 1 hour. The products were purified with a QIAquick Gel Extraction Kit (Qiagen). For T:G-library preparation PCR product was self-ligated with T4 DNA ligase (Thermo Fisher Scientific, EL0013) with 5% PEG 4000 at 4°C overnight. For other three libraries PCR products were digested using NheI restriction enzyme (NEB, R0131S) at 37 °C for 3 hours, and self-ligated with T4 DNA ligase (Thermo Fisher Scientific, EL0013) with 5% PEG 4000 at 4°C overnight. Ligated products (100-400 ng/µL) were desalted by drop dialysis using 13 mm diameter, Type-VS Millipore membrane (Merck Millipore, #VSWP01300). 20 µL ElectroMAX DH10B competent cells (Invitrogen, 18290015) was electroporated with 3 µL ligated products. 0.01% of the electroporated bacteria were plated on ampicillin-containing medium in order to estimate the complexity of the libraries; the remaining cultures were grown overnight in 100ml of liquid medium, and the plasmids were extracted the next day. Barcoded plasmid libraries with complexity of 0.8-2 million independent clones were used for further experiments.

Cell culture.

mESCs were grown at 37 °C under a 95% air and 5% CO2 atmosphere on gelatin in serum/LIF medium composed of GMEM (Sigma, G5154) supplemented with 15% FBS (HyClone™), 1X MEM Non-Essential Amino Acids (Gibco, 11140 - 050), 1X GlutaMax (Gibco, 35050 - 061), 1mM Sodium Pyruvate (Gibco, 11360-070), 0.1 mM 2-Mercaptoethanol (Thermo Fisher, 31350-010) and 1,000 U/ml LIF ESGRO® (Merck Millipore, ESG1106). mESCs were passaged every 2 d with a 1:8 dilution. Cells were tested yearly for mycoplasma contamination.

Transfection and transposon integration.

To integrate the construct into the genome of mouse ES cells, 1 million cells in 6-well plates were transfected with 2 µg of plasmid pT2_FF together with 2 µg of plasmid pCMV(CAT)T7-SB100x and 2 µg of plasmid pcDNA3.1-mCherry using Lipofectamine 2000 (Thermo Fisher, #11668027). After 24 h, mCherry-positive cells were FACS sorted. Pools of 20,000 cells were plated on 24-well plates and grown for two weeks, transferring to 100 mm dishes when the cultures reached a density of $5x10^6$ cells/mL. Two independent cell pools of 20,000 cells were prepared for each construct.

To generate the mismatches, pools of mouse ES cells harboring integrated transposons were transfected with 5 µg of plasmid pCBAScel using Lipofectamine 2000 (Thermo Fisher, #11668027). The growth medium was changed after 16 h later. 24 h and 48 h post transfection, cells were collected using Trypsin-EDTA (Gibco, #25200056), washed with PBS, and used for genomic DNA isolation. Genomic DNA from transfected cells was extracted using DNeasy Blood and Tissue Kit (Qiagen, #69504).

Inverse PCR.
10 µg genomic DNA from transfected mESCs were digested using 10 µl 10 U/µL NlaIII (NEB, #R0125S) in a 50 µL final volume for 3 hours at 37 °C. The reaction was heat-inactivated at 65 °C for 20 minutes. The reaction was diluted to a final volume of 1.8 mL in 1X T4 ligase buffer (Thermo Fisher Scientific, #EL0013) to favor self-ligation events, and ligation was carried out with 600 U of T4 ligase (Thermo Fisher Scientific, #EL0013) at 16°C overnight. After ligation, samples were precipitated by ethanol, pellets were resuspended in water and column-purified (QIAGEN, QIAquick PCR purification kit #28104) eluting with 100 uL EB. Non circularized templates were eliminated by 2 h digestion at 37°C with Plasmid-safe DNAse (Epicentre, #E3101K), the enzyme was inactivated by heating for 30 minutes at 70°C. The product was column-purified (QIAGEN, QIAquick PCR purification kit #28104). The backbone of the TRIP reporters contains a I-CeuI site outside the transposable cassette, taking advantage of this, non integrated plasmids were cut by 2 h digestion at 37 °C with I-CeuI restriction enzyme (NEB, R0699S) in a total volume of 70 ul followed by 20 minutes heat inactivation at 65 °C. All enzymatic reactions were carried out in the recommended manufacturer's buffer.

For the first round of nested PCR, 10µL of I-CeuI-digestion products was mixed in 50 µL standard Phusion polymerase reaction mix (Thermo Fisher Scientific, F530S) in GC buffer, with 0.1 µM primers IR1,2 (annealing to IR/DR sequence). The cycling conditions were as follows: 98 °C for 5 min; 98°C for 20 s, 60°C for 1 min and 72°C for 5 min (1 cycle); 98°C for 20 s, 60°C for 1 min and 72°C for 2 min (20 cycles); and 72°C for 5 min. Products of the reaction were purified using Agencourt AMPure XP beads (Beckman Coulter, A63880), and eluted in 40 µL of water . For the second round of nested PCR, 37 µL of the products was diluted to 50 µL of standard Phusion polymerase reaction mix in GC buffer with 0.1 µM primer M1 (annealing to the Illumina PE1.0 primer) and one of indexing primers M2,3. The cycling conditions were as follows: 98°C for 2 min; 98°C for 20 s, 60°C for 1 min and 72°C for 1 min (10 cycles); and 72°C for 5 min. Primers M2,3 added the Illumina PE2.0 primer and one of indices to the amplicons. Products of the reaction were purified using Agencourt AMPure XP beads (Beckman Coulter, A63880), and eluted in 40 µL of water .
PCR products ran as a smear on agarose gel. The smears were specific, because they failed to appear when the mESCs were not transfected.

Genomic DNA preparation for Linear Amplification and UMI-PCR.
To eliminate I-SceI sites that were not cut during DSB induction and limit the size of PCR extension, 2 µg genomic DNA from mESCs (both, with DSB induction and control without the induction) were digested using 1 µl 100 U/µL XbaI (NEB, #R0145T) and 4 µl 5 U/µL I-SceI (NEB, #R0694L) in a 50 µL final volume for 3 hours at 37 °C. Digested DNA was column-purified (QIAGEN, QIAquick PCR purification kit #28104).

Linear amplification.
500 ng of genomic DNA obtained after I-SceI/XbaI digestion was used as a template in a 50 µL of Q5 DNA polymerase reaction mix (NEB, M0491S), using 50 nM of a UMI-containing primer U1 in the following cycling conditions: 98°C for 4 min; 98°C for 30s, 60°C for 1 min and 72°C for 1 min (50 cycles); and 72°C for 5 min. Products of linear amplification were purified using Agencourt AM Pure XP beads (Beckman Coulter, A63880), and eluted in 40 µL of water. Linear amplification was repeated 4 times for every sample to account for the technical variability.

UMI-PCR.
500 ng of genomic DNA obtained after I-SceI/XbaI digestion was used as a template in a 50 µL of Q5 DNA polymerase reaction mix (NEB, M0491S), using 50 nM of primers U1 and U2 in the following cycling conditions: 98°C for 1 min; 98°C for 20s, 60°C for 1 min and 72°C for 4 min (6 cycles); and 72°C for 10 min. Products of UMI-PCR were purified using Agencourt AM Pure XP beads (Beckman Coulter, A63880), and eluted in 40 µL of water. UMI-PCR was repeated 4 times for every sample to account for the technical variability.

Sequencing sample preparation of UMI-amplicons.
Products of Linear Amplification and UMI-PCR were used as a template for indexing PCR. 50 µL of Q5 DNA polymerase reaction mix (NEB, M0491S) was used. Every sample was amplified using 100nM of primer U2 (annealing to the Illumina PE1.0 primer) and one of indexing primers with Illumina PE2.0 sequence (UIND1-12). The cycling conditions were as follows: 98°C for 1 min; 98°C for 20s, 60°C for 30s and 72°C for 4 min (30 cycles); and 72°C for 10 min. Products of the indexing PCR were pooled into 3 final samples: 1) control (without DSB induction), 2) 24 hours after I-SceI transfection, and 3) 48 hours after I-SceI transfection. The samples were purified using 2% E-Gel EX precast agarose gels (Thermo Fisher Scientific, G401002).
Each sample was visualized on a Bioanalyzer (Agilent Technologies) and quantified by qPCR using a Kapa Library Quantification Kit (Kapa Biosystems, KK4835).

High throughput sequencing.
Final samples for both inverse PCR and UMI-amplicons (concentration 4 nM) were sequenced as paired-end reads on HiSeq2500 and NextSeq500 sequencers (Illumina).

Processing inverse PCR reads
The paired-end reads were pre-processed by custom Python scripts. The forward read consists of the barcode, a fixed 20 nucleotide watermark sequence used for identification, the NlaIII restriction site and some mouse genome sequence ligated to the NlaIII site. The reverse read consists of the last 25 nucleotides of pT2 and the mouse genome sequence at the insertion site of the transposon. We used seeq version 1.1 (https://github.com/ezorita/seeq) allowing up to three errors (mismatches and indels) to identify the watermark and isolate the barcode sequence. Reads were discarded if the watermark was not found or if the barcode was not between 14 and 24 nucleotides long. We removed the first 25 nucleotides of the reverse read and cut the sequence at the first NlaIII site, if any. The reads for which this sequence was shorter than 20 nucleotides were

discarded. The genomic sequences thus obtained were mapped in the mouse genome release mm9 using the GEM mapper build 1.376 (Marco-Sola *et al*. 2012) with options `-q ignore --unique-mapping`. The reason for not using the more recent mouse release mm10 was that the chromatin from Juan *et al*. (2016) were only available in release mm9.

After mapping, the barcodes were clustered using Starcode (Zorita, Cuscó, and Filion 2015) allowing up to two errors (options `-d2 --print-clusters`). This removes potential sequencing errors and consolidates the barcode sequences.

The barcodes were assigned to a genomic location using custom Python scripts. Unmapped genomic sequences were discarded and sequences mapping to multiple locations were flagged as such. For each barcode, we collected all the insertion sites that totalled at least 10% of the reads and we computed their diameter, equal to the maximum of their pairwise genomic distances (infinite for two sites on different chromosomes or if one of them maps to multiple locations). If the diameter was greater than 30 nucleotides, the barcode was discarded for being used in reporters mapping to different locations. Otherwise, the barcode was kept and its location was attributed to the most frequent insertion site (they are usually within 1-2 nucleotides of each other because of small mapping artefacts).

Processing UMI-amplicon reads

Paired-end reads were preprocessed using custom Python scripts. The forward read consists of the barcode, the watermark sequence and the right half of the second F segment in the orientation of **Figure 1b**. The reverse read consists of the UMI and the left half of the first F segment in the orientation of **Figure 1b**. Both reads extend the mid point of the F segment by three nucleotides. If the reporter is uncut or repaired by NHEJ, forward and reverse reads do not overlap. If the reporter is repaired by SSA, forward and reverse reads overlap because there is only one F segment. We can thus isolate the reads from reporters that went through SSA by ensuring that the nucleotides in the mismatch position are reverse-complement of each other in the forward and reverse reads.

Thus, we used seeq with up to 10 errors to identify the half F segments and isolate the nucleotides in mismatch position on the forward and reverse reads, together with the barcode and the UMI. Barcodes and UMIs were clustered with Starcode allowing up to 2 errors (options `-d2 --print-clusters`) and the repair events were quantified for each barcode. The barcode–UMI pairs with only one read per run were discarded. After this operation, UMIs that were associated with more than one barcode were discarded. The remaining UMIs were classified as NHEJ or SSA as explained above, and those classified as SSA were further split into A/T or G/C. This provided for each barcode the full list of events reported by UMIs.

Barcodes with more UMIs reporting NHEJ than SSA and those with only one UMI were removed. Barcodes that passed all these criteria in the control experiments without I-SceI induction were removed. The global proportion of remaining UMIs reporting A/T versus G/C was used a measure of repair bias.

Mutual information

The operative definition of mutual information is the Kullback-Leibler between the joint distribution of two variables and their product distribution (whereby we assume independence). Joint and product distributions are particularly easy to compute for categorical variables, which makes mutual information more adapted than the Pearson and Spearman coefficients of correlation in this context.

We collected the barcodes with at least 5 UMIs that appeared in at least two replicates and we assigned them to a single repair outcome by majority vote (i.e. each barcode was called either A/T or G/C, even in case of conflicts between UMIs). For every pair of replicates where the barcode appears, there are thus 4 possible outcomes. We used the number of occurrences of the 4 outcomes as estimate of their joint distribution, and the product of their margins as estimate of the product distribution.

We computed the mutual information using the log2 function instead of the natural logarithm so that the result is expressed in bits. There is in general no upper bound on mutual information, but for two categorical variables with two outcomes each, the maximum is 1 bit

Neural network training

Neural networks were trained using custom scripts written in Python with Pytorch version 1.0.1 and Numpy version 1.15.4 on a desktop running Ubuntu 16.04 and equipped with a graphics card GeForce GTI 1660Ti (Nvidia), CUDA version 10.1. Networks were optimized with the Adam optimizer (Kingma and Ba 2014) with a learning rate empirically set to 0.001 for 100 epochs.

Code and data access

The data will be released on GEO after the manuscript is peer-reviewed. The scripts used in this study are available on Github at https://github.com/gui11aume/REPLAY_FF. A Docker container to reproduce the results is available on Dockehub at https://hub.docker.com/r/gui11aume/ff.

Table 2. List of primers.

Primer name	Sequence
L1	NNNNNNNNNNNNNNNNNNNagatcggaagagcgtcgtg
L2	CGCTAATTAATGGAATCATGAACA
L3	catagGCTAGC NNNNNNNNNNNNNNNNNNNagatcggaagagcgtcgtg
L4	catagGCTAGC TCCGCAGAATCATGAACA
L5	catagGCTAGC AGTCAGGAATCATGaaca
L6	catagGCTAGC TCGTTGGAATCATGaaca
IR1	TGTATTTGGCTAAGGTGTATGTAA
IR2	ATTCCCAGTGGGTCAGAAGT
M1	AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGACGCTCTTCCGATCT
M2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACC GCTCTTCCGATCT ACT AAACTTCCGACTTCAACTGT
M3	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACC GCTCTTCCGATCT TGT AAACTTCCGACTTCAACTGT
U1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNNNNN NNNNTGCAACGAATTCATTAGTGCG
U2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG
UIND1	CAAGCAGAAGACGGCATACGAGAT CAAGCT GTGACTGGAGTTC
UIND2	CAAGCAGAAGACGGCATACGAGAT GTGAAC GTGACTGGAGTTC
UIND3	CAAGCAGAAGACGGCATACGAGAT ACTGCA GTGACTGGAGTTC
UIND4	CAAGCAGAAGACGGCATACGAGAT CTAAGA GTGACTGGAGTTC
UIND5	CAAGCAGAAGACGGCATACGAGAT CTTGGC GTGACTGGAGTTC
UIND6	CAAGCAGAAGACGGCATACGAGAT AGACAT GTGACTGGAGTTC
UIND7	CAAGCAGAAGACGGCATACGAGAT TGTAGA GTGACTGGAGTTC
UIND8	CAAGCAGAAGACGGCATACGAGAT TTCAGC GTGACTGGAGTTC
UIND9	CAAGCAGAAGACGGCATACGAGAT GTCCTA GTGACTGGAGTTC
UIND10	CAAGCAGAAGACGGCATACGAGAT ATCCAG GTGACTGGAGTTC
UIND11	CAAGCAGAAGACGGCATACGAGAT ACATCG GTGACTGGAGTTC
UIND12	CAAGCAGAAGACGGCATACGAGAT GCCTAA GTGACTGGAGTTC

ACKNOWLEDGEMENTS

We would like to thank William R. Engels, Carlos Flores and Laurent Duret for early discussions about this project; and Zsuzsanna Izsvák for kindly providing the Sleeping Beauty constructs. We acknowledge the financial support of the Spanish Ministry of Economy, Industry and Competitiveness ('Centro de Excelencia Severo Ochoa 2013-2017', Plan Estatal PGC2018-099807-B-I00), of the CERCA Programme / Generalitat de Catalunya, and of the European Research Council (Synergy Grant 609989). V. P. was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie programme (665385). We also acknowledge the support of the Spanish Ministry of Economy and Competitiveness (MEIC) to the EMBL partnership.

REFERENCES

Akhtar, Waseem, Johann de Jong, Alexey V. Pindyurin, Ludo Pagie, Wouter Meuleman, Jeroen de Ridder, Anton Berns, Lodewyk F. A. Wessels, Maarten van Lohuizen, and Bas van Steensel. 2013. "Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel." *Cell* 154 (4): 914–27.

Anand, Ranjith, Annette Beach, Kevin Li, and James Haber. 2017. "Rad51-Mediated Double-Strand Break Repair and Mismatch Correction of Divergent Substrates." *Nature* 544 (7650): 377–80.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. "The Mosaic Genome of Warm-Blooded Vertebrates." *Science* 228 (4702): 953–58.

Bill, C. A., W. A. Duran, N. R. Miselis, and J. A. Nickoloff. 1998. "Efficient Repair of All Types of Single-Base Mismatches in Recombination Intermediates in Chinese Hamster Ovary Cells. Competition between Long-Patch and G-T Glycosylase-Mediated Repair of G-T Mismatches." *Genetics* 149 (4): 1935–43.

Brown, Alexander J., Peng Mao, Michael J. Smerdon, John J. Wyrick, and Steven A. Roberts. 2018. "Nucleosome Positions Establish an Extended Mutation Signature in Melanoma." *PLoS Genetics* 14 (11): e1007823.

Brown, T. C., and J. Jiricny. 1989. "Repair of Base-Base Mismatches in Simian and Human Cells." *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada* 31 (2): 578–83.

Chakraborty, Ujani, Carolyn M. George, Amy M. Lyndaker, and Eric Alani. 2016. "A Delicate Balance Between Repair and Replication Factors Regulates Recombination Between Divergent DNA Sequences in Saccharomyces Cerevisiae." *Genetics* 202 (2): 525–40.

Corrales, Marc, Aránzazu Rosado, Ruggero Cortini, Joris van Arensbergen, Bas van Steensel, and Guillaume J. Filion. 2017. "Clustering of Drosophila Housekeeping Promoters Facilitates Their Expression." *Genome Research* 27 (7): 1153–61.

Duret, Laurent, Adam Eyre-Walker, and Nicolas Galtier. 2006. "A New Perspective on Isochore Evolution." *Gene* 385 (December): 71–74.

Duret, Laurent, and Nicolas Galtier. 2009. "Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes." *Annual Review of Genomics and Human Genetics* 10: 285–311.

Filipski, J. 1990. "Evolution of DNA Sequence Contributions of Mutational Bias and Selection to the Origin of Chromosomal Compartments." In *Advances in Mutagenesis Research*, edited by Günter Obe, 1–54. Berlin, Heidelberg: Springer Berlin Heidelberg.

Freese, Ernst. 1962. “On the Evolution of the Base Composition of DNA.” *Journal of Theoretical Biology* 3 (1): 82–101.

Gale, J. M., K. A. Nissen, and M. J. Smerdon. 1987. “UV-Induced Formation of Pyrimidine Dimers in Nucleosome Core DNA Is Strongly Modulated with a Period of 10.3 Bases.” *Proceedings of the National Academy of Sciences of the United States of America* 84 (19): 6644–48.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. “GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis.” *Genetics* 159 (2): 907–11.

Gisler, Santiago, Joana P. Gonçalves, Waseem Akhtar, Johann de Jong, Alexey V. Pindyurin, Lodewyk F. A. Wessels, and Maarten van Lohuizen. 2019. “Multiplexed Cas9 Targeting Reveals Genomic Location Effects and gRNA-Based Staggered Breaks Influencing Mutation Efficiency.” *Nature Communications* 10 (1): 1598.

Goellner, Eva M. 2019. “Chromatin Remodeling and Mismatch Repair: Access and Excision.” *DNA Repair* 85 (October): 102733.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–78.

Hoadley, Katherine A., Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D. M. Leiserson, *et al*. 2014. “Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin.” *Cell* 158 (4): 929–44.

Hwang, Dick G., and Phil Green. 2004. “Bayesian Markov Chain Monte Carlo Sequence Analysis Reveals Varying Neutral Substitution Patterns in Mammalian Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (39): 13994–1.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1502.03167.

Juan, David, Juliane Perner, Enrique Carrillo de Santa Pau, Simone Marsili, David Ochoa, Ho-Ryun Chung, Martin Vingron, Daniel Rico, and Alfonso Valencia. 2016. “Epigenomic Co-Localization and Co-Evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs.” *Cell Reports*. https://doi.org/10.1016/j.celrep.2016.01.008.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.

Kunkel, Thomas A., and Dorothy A. Erie. 2005. “DNA Mismatch Repair.” *Annual Review of Biochemistry* 74: 681–710.

Lujan, Scott A., Anders R. Clausen, Alan B. Clark, Heather K. MacAlpine, David M. MacAlpine, Ewa P. Malc, Piotr A. Mieczkowski, *et al*. 2014. “Heterogeneous Polymerase Fidelity and Mismatch Repair Bias Genome Variation and Composition.” *Genome Research* 24 (11): 1751–64.

Marco-Sola, Santiago, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. 2012. “The GEM Mapper: Fast, Accurate and Versatile Alignment by Filtration.” *Nature Methods* 9 (12): 1185–88.

Mátés, Lajos, Marinee K. L. Chuah, Eyayu Belay, Boris Jerchow, Namitha Manoj, Abel Acosta-Sanchez, Dawid P. Grzela, *et al*. 2009. “Molecular Evolution of a Novel Hyperactive Sleeping Beauty Transposase Enables Robust Stable Gene Transfer in Vertebrates.” *Nature Genetics* 41 (6): 753–61.

Meyerhans, Andreas, Jean-Pierre Vartanian, and Simon Wain-Hobson. 1990. “DNA Recombination during PCR.” *Nucleic Acids Research*.

https://doi.org/10.1093/nar/18.7.1687.

Nair, V., and G. E. Hinton. 2010. “Rectified Linear Units Improve Restricted Boltzmann Machines.” *Proceedings of the 27th International Conference*. https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf.

Pavlov, Youri I., Ibrahim M. Mian, and Thomas A. Kunkel. 2003. “Evidence for Preferential Mismatch Repair of Lagging Strand DNA Replication Errors in Yeast.” *Current Biology: CB* 13 (9): 744–48.

Pleasance, Erin D., R. Keira Cheetham, Philip J. Stephens, David J. McBride, Sean J. Humphray, Chris D. Greenman, Ignacio Varela, et al. 2010. “A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome.” *Nature* 463 (7278): 191–96.

Ramstein, J., and R. Lavery. 1988. “Energetic Coupling between DNA Bending and Base Pair Opening.” *Proceedings of the National Academy of Sciences of the United States of America* 85 (19): 7231–35.

Rideout, W. M., 3rd, G. A. Coetzee, A. F. Olumi, and P. A. Jones. 1990. “5-Methylcytosine as an Endogenous Mutagen in the Human LDL Receptor and p53 Genes.” *Science* 249 (4974): 1288–90.

Sadowski, Peter, and Pierre Baldi. 2018. “Neural Network Regression with Beta, Dirichlet, and Dirichlet-Multinomial Outputs.” https://openreview.net/pdf?id=BJeRg205Fm.

Sinsheimer, R. L. 1955. “The Action of Pancreatic Deoxyribonuclease. II. Isomeric Dinucleotides.” *The Journal of Biological Chemistry* 215 (2): 579–83.

Spies, Maria, and Richard Fishel. 2015. “Mismatch Repair during Homologous and Homeologous Recombination.” *Cold Spring Harbor Perspectives in Biology* 7 (3): a022657.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” *Journal of Machine Learning Research: JMLR* 15 (1): 1929–58.

Sueoka, N. 1962. “On the Genetic Basis of Variation and Heterogeneity of DNA Base Composition.” *Proceedings of the National Academy of Sciences of the United States of America* 48 (April): 582–92.

Supek, Fran, and Ben Lehner. 2015. “Differential DNA Mismatch Repair Underlies Mutation Rate Variation across the Human Genome.” *Nature* 521 (7550): 81–84.

Weber, Claudia C., Catherine J. Pink, and Laurence D. Hurst. 2012. “Late-Replicating Domains Have Higher Divergence and Diversity in Drosophila Melanogaster.” *Molecular Biology and Evolution* 29 (2): 873–82.

Zorita, Eduard, Pol Cuscó, and Guillaume J. Filion. 2015. “Starcode: Sequence Clustering Based on All-Pairs Search.” *Bioinformatics* 31 (12): 1913–19.