Strand asymmetry influences mismatch resolution 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

Background: 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.

Results: 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. Regardless of the mismatch, we observed a 60-80% bias in the resolution in favor of one strand. The location of the lesion in the genome and the type of mismatch had little influence on the bias in this context. Instead, changing the position of the double-strand break in the reporter gave a complete reversion of the bias.

Conclusion: These results suggest that the processing of the double-strand break has a major influence on the repair of mismatches during single-strand annealing, irrespective of the surrounding chromatin context.

Keywords: mismatch repair; single-strand annealing; chromatin; genome-wide technologies; mouse embryonic stem cells

BACKGROUND

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 (1). 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 (2,3).

At least two features of mammalian genomes are shaped by non-selective forces: the depletion of CpG dinucleotides (4) and the 10 bp periodicity of ApA dinucleotides (5). The first is due to increased C to T mutations when C is methylated, which takes place only within CpG dinucleotides (6,7). The second is due to increased damage on nucleotides facing outward the nucleosome, where ApA is the least exposed dinucleotide (1,8,9).

The mechanisms that underlie mutational biases in mammals are otherwise poorly understood. For instance, one of the most enigmatic features of vertebrate genomes is that they are organized in megabase-scale domains called isochores, where the GC-content is relatively uniform (10). 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 (11,12). 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 (13).

Initial estimates in mammalian genomes suggested that mismatch repair is indeed biased toward G and C nucleotides (14), 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 (15). Also, it was later shown that in mice GC-biased gene conversion is restricted to non-crossover events with a single mismatch (16). 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 (17,18). 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 (19), a feature that seems to be shared among eukaryotes (20). 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 separate the biases due to damage from the

biases due to repair in the context of chromatin, mostly for lack of a technology 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 (21–23). 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 phase, 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 mismatch resolution in the DNA (**Figure 1**). The reporter construct consists of two nearly identical 152 bp sequences separated by a restriction site for the meganuclease I-SceI. The 152 bp sequences were originally a central segment of the GFP gene, so 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 a 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 (24). The final product contains only one F segment with only one of the two original alleles.

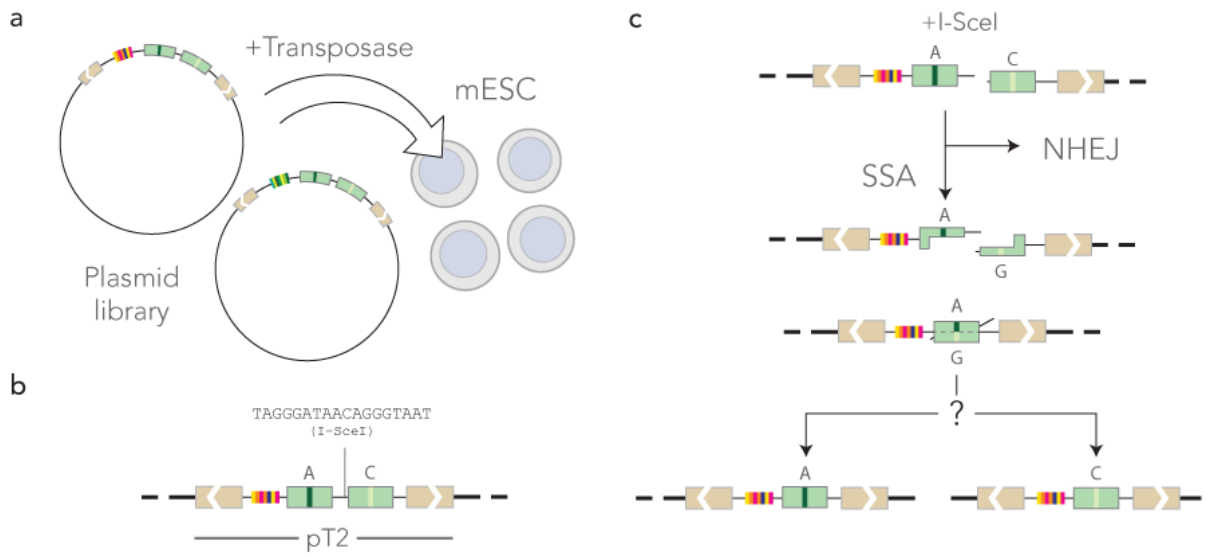

Figure 1. Experimental approach. **a)** Mouse ES cells in culture are co-transfected with a barcoded reporter library and with the Sleeping Beauty 100X transposase. **b)** The reporters in the pT2 transposon backbone are integrated at random in the mouse genome. **c)** Mismatches occur during the repair 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 strand. We therefore included a strand-swap control of the T:G mismatch, referred to as G:T, where G is on the top strand and T is on the bottom strand. 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 (25). After two weeks of growth, the reporters were mapped by inverse PCR (see Methods). For each construct, we mapped between 9 and 48 thousand reporters, with 107 thousand known reporter locations in total (see **Table 1**). The substantial variations in the number of recorded events are due to batch effects, rather than differences in the induction of the double-strand break or in the efficiency of the DNA repair system. For the T:C construct in particular, many sequencing reads were lost due to contaminations of the PCR products.

Construct	*Mapped*	*In repeats*	*Repair events*	*Mapped events*
A:G	15,601	6,346	28,115	8,738
T:G	9,451	3,619	188,575	24,856
A:C	12,445	4,963	32,872	9,138
T:C	21,448	8,363	4,189	965
G:T	48,627	19,575	75,775	39,227
Total	107,572	42,866	329,526	82,924

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 (sum of two TRIP pools). *In repeats*: number of barcodes mapped in repeated sequences. *Repair events*: number of barcodes for which the outcome of mismatch resolution was measured (technical replicates are counted separately). *Mapped events*: subset of the repair events where the barcode is mapped unambiguously.

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.

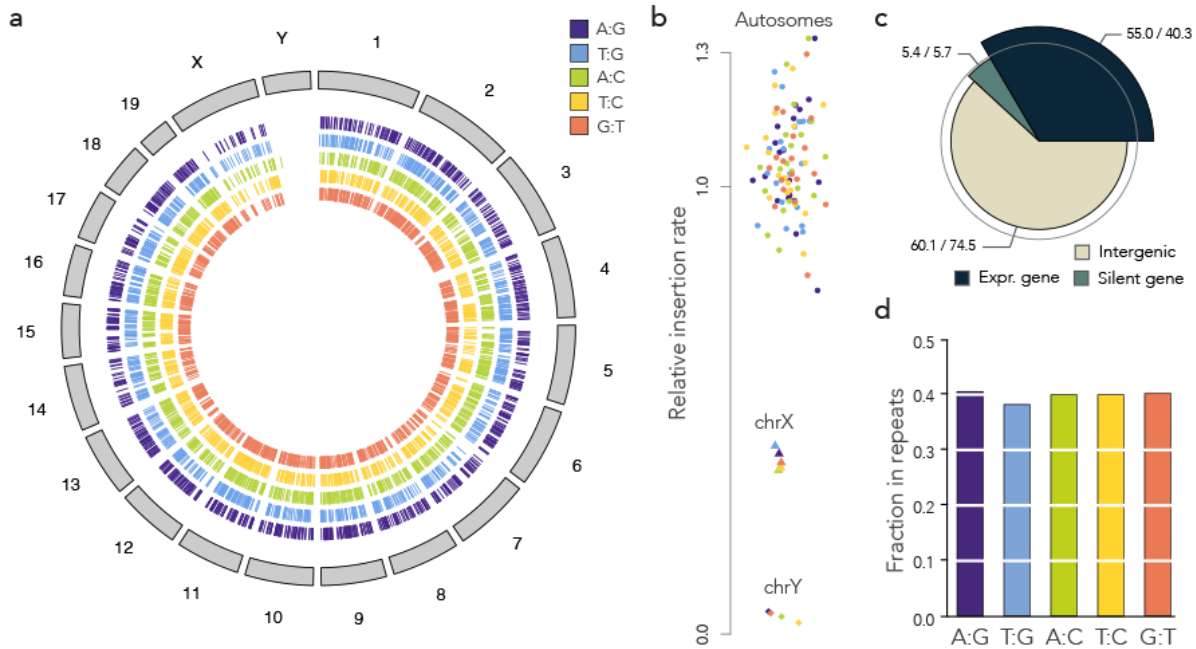

Figure 2. Mapping of the reporters. **a)** Overview of the insertions. For each construct, 1,000 insertions were drawn at random and plotted on a circular representation of the mouse genome. Each tick mark represents a mapped insertion. **b)** Dot plot of the relative insertion rate per chromosome. For 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 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.

The inserted reporters also allowed us to interrogate repeated sequences (**Figure 2c**) but not all the barcodes could be mapped, so most of the repair events occurred at unknown or ambiguous 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 through either NHEJ or SSA 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 (26), a common PCR artefact that can potentially shuffle the barcodes between templates and make quantification 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.

For each construct, we performed 4 technical replicates per UMI-PCR and per UMI-LA, at 24 and 48 hours post I-SceI induction, plus 4 technical replicates with UMI-PCR and 4 with UMI-LA for each construct without I-SceI induction. Two UMI-PCRs for the A:G construct at 48 hours failed to yield any sequencing read. **Figure 3b** shows the global repair bias toward A/T for each construct (the barcodes detected without I-SceI induction are discarded, see below). This represents a total of almost 330,000 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 around 20% for

the last. For each construct, the biases were similar between replicates, between amplification methods and between time points, showing that the assays are reproducible in the given experimental conditions.

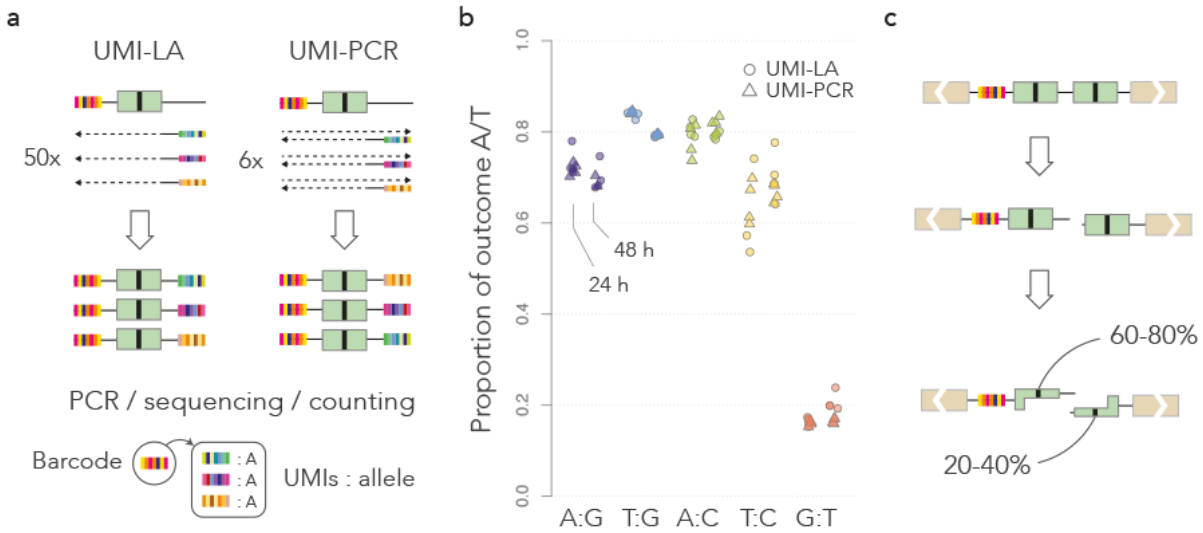

Figure 3. Measure of repair biases. **a)** Quantification methods. In UMI-LA (left), barcoded reporters are amplified by 50 cycles of linear amplification using a primer decorated by UMIs. In UMI-PCR (right), reporters are amplified by 6 PCR cycles where one primer is decorated by UMIs. Either way, the products are further amplified by regular PCR. After sequencing, each barcode is associated with several UMIs, themselves associated with 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 each technical replicate. 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 during the resolution of the mismatch.

Strikingly, the dominant outcome did not correspond to a nucleotide but to a strand. Indeed, the T:G mismatch was resolved in favor of T in the T:G construct (cyan, **Figure 3b**), but in favor of G in the G:T construct (red, **Figure 3b**) where the nucleotides were swapped. The measured repair biases are roughly symmetric between the two constructs (80% vs 20%). We therefore conclude that in this assay, the top strand is more likely to be used as a template during the repair of the mismatch.

The magnitudes of the repair bias in favor of A or T when they are on the top strand (purple, cyan, green and yellow, **Figure 3b**) are comparable to each other, except for the T:C construct, which showed greater variations (we obtained fewer events for this construct, see **Table 1**). 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 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 were 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 the I-SceI induction (*e.g.*, during the preparation of the barcoded library), they are replicated through cell division and therefore are all identical. 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.

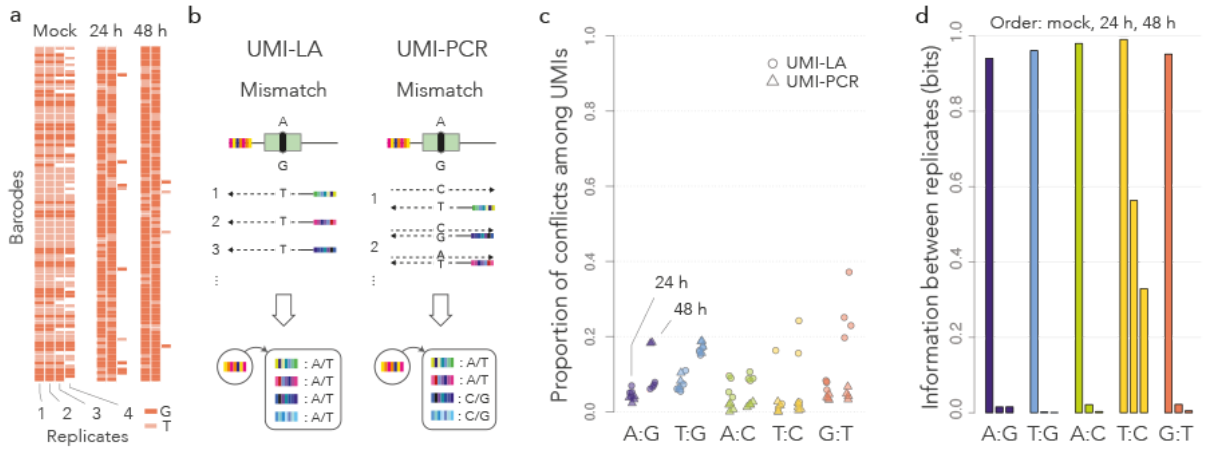

Figure 4. Global differences between reporters. **a)** Repair across replicates. Each row shows a random barcode from the G:T construct and each column shows a replicate where it appears. The color of the rectangle indicates the repair outcome. Without I-SceI induction (Mock), barcodes are typically 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 (not shown). **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 100 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 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 (and 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 resolution on the reporters has a global bias toward the top strand. If the bias is an intrinsic property of the repair process, it should not depend on any local property of the chromatin at or near the site of the integrated reporter.

Figure 5a shows the repair bias of the G:T construct (that with the highest number of mapped events) when the reporters are inserted in expressed genes, silent genes or intergenic regions. In all cases, the bias is toward G and it has the same magnitude as the

genome-wide average (**Figure 3b**). **Figure 5b** shows a similar analysis for insertions in sites enriched in H3K9me3 or H3K36me3. Both epigenomic marks were previously associated with DNA repair: somatic mutations occur mostly in domains enriched in H3K9me3 (27), and H3K36 interacts with MutSα (28). The enrichment domains were defined by Juan *et al*. (29), and when both marks were present, the domain was attributed to the less common mark H3K9me3. Once again, the results show a similar bias toward G in all the cases, suggesting that the chromatin context has little influence over the bias.

It is possible that the influence of the chromatin context is too complex to be captured by a single epigenomic mark. One way to test this hypothesis is to use a flexible machine learning approach such as deep learning to evaluate how well the information about chromatin predicts the local repair bias.

We designed an artificial neural network with a residual network architecture (30). The output of hidden layers is batch-normalized (31) and passed through a standard ReLU activation function (32). We also included two Dropout steps to mitigate overfitting (33). 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 (34). The architecture of the network is sketched in **Figure 5a** (see Methods for details).

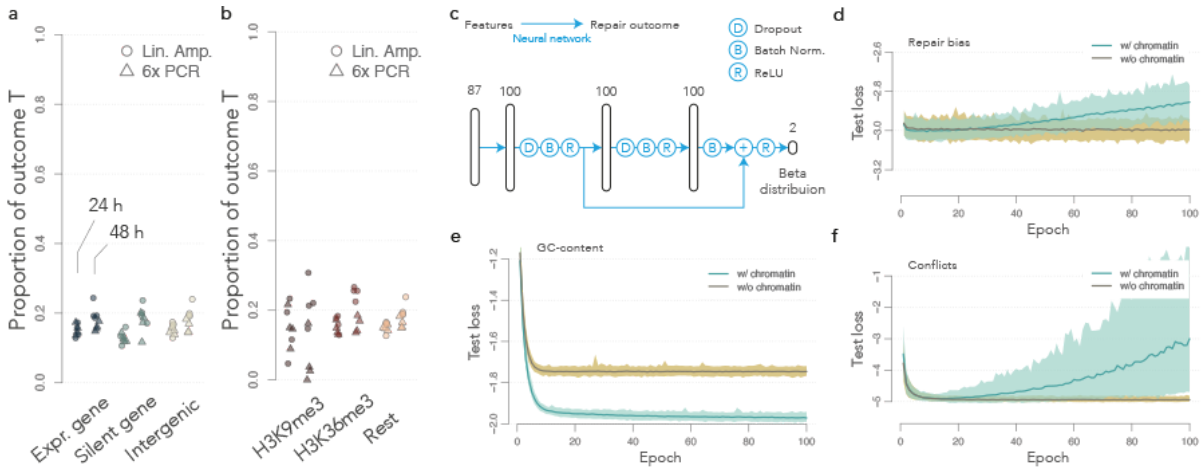

Figure 5. Mismatch resolution and chromatin context. **a)** Biases in genomic contexts. The dot plot shows the measured bias toward T with the G:T construct. Expressed genes, silent genes and intergenic regions are the same as in **Figure 2c**. **b)** Biases in chromatin contexts. Same representation as in a), where reporters are split based on the local enrichment of H3K9me3, or H3K36me3. The spread for H3K9me3 is higher because there are fewer mapped events in this context. **c)** Architecture of the artificial neural network. Vertical cartridges represent fully connected neuron layers with indicated dimensions. The blue arrows indicate the forward information flow. **d)** Learning curves for the full data set. The median loss on the test set is shown in green for 100 trainings 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. **e)** Learning curves for the local GC-content. The repair outcome was replaced by the GC-content in a 10 kb window around the reporter. **f)** Learning curves for conflicts among UMIs. The only observations used for learning are UMI-PCR 24 h post I-SceI induction.

The dataset consists of 43,864 mapped repair events for which the chromatin features at the insertion site are available. Each record consists of the repair outcome for one barcode together with the local GC-content and the chromatin features compiled by Juan *et al*. (29). If a barcode was present in several replicates, the outcomes were considered to be independent events within the same chromatin context. The chromatin features include 3 cytosine modifications, 13 histone marks, and 62 chromatin proteins. Chromatin features are mapped at a resolution of 200 bp and each record was constructed by inheriting the full chromatin profile of the window where the reporter is inserted. The records also include the construct type, the time point, the amplification technology and the TRIP pool. Those bookkeeping variables were introduced to buffer non-biological variations such as batch effects. A random set of 10% of the records was held out for testing (see Methods).

We performed 100 independent trainings with random restarts and tracked the performance with a loss function measuring the discrepancy between predicted and observed outcomes of repair. The learning curves represent the mean value of the loss function on the test set as learning unfolds: As long as the score goes down, the capacity of the model to predict the outcome on new data improves; when the loss stabilizes, the model reaches its maximum performance; and if the loss increases, the model is overfitted and further training damages the performance.

The median performance of the network on the test set is shown in **Figure 5d**. For comparison, we included 100 trainings of a null model where the chromatin features were removed from the predictors, leaving only the construct, the time point, the amplification technology and the TRIP pool. In other words, the null model cannot fit any context-dependence for repair events. Surprisingly, the learning curve of the null model is below that of the full model, indicating a better performance. In addition, the full model shows clear signs of overfitting from 30 epochs.

Is the full model mis-specified or is there no relationship to be discovered between the repair bias and the chromatin context? To answer this question, we used the neural network for the comparable task of predicting the local GC-content within a 10 kb window. In this case the full model clearly outperforms the null model and still shows some evidence of learning after 100 epochs (**Figure 5e**), demonstrating that the model can discover complex associations between the local chromatin context and other variables. Thus, the previous results suggest that the chromatin context has little influence over the repair outcome in this assay. If such influence existed, it would have been discovered by the neural network.

Finally, we asked whether the occurrence of repair itself depends on the chromatin context. To address this question, we took advantage of a fact established previously: UMI-PCR is expected to produce mixed populations of UMIs if the mismatch is not repaired (**Figure 4b**). We used the neural network to predict whether a barcode will have conflicting UMIs, when assayed by UMI-PCR 24 hours post I-SceI induction (total of 8,734 records, **Figure 5f**). As for the repair bias, the learning curve of the null model is below that of the full model. This indicates that the chromatin context had no detectable effect on the probability of repairing the mismatch.

The unmapped reporters were excluded from this analysis, so it is possible that repair proceeds differently in some unmapped regions, such as the Y chromosome for instance. Otherwise, our results show that neither the occurrence of mismatch repair nor the magnitude of the repair bias depends on known elements of the chromatin context.

Asymmetries between flaps influence the repair bias

The results of our repair assay show that one strand is favored during mismatch resolution. What could create such an asymmetry between the strands? The position of the I-SceI site is skewed toward one of the F segments, meaning that after the resection of 5' ends, one 3' flap is longer than the other. Could this cause the repair bias observed throughout the genome?

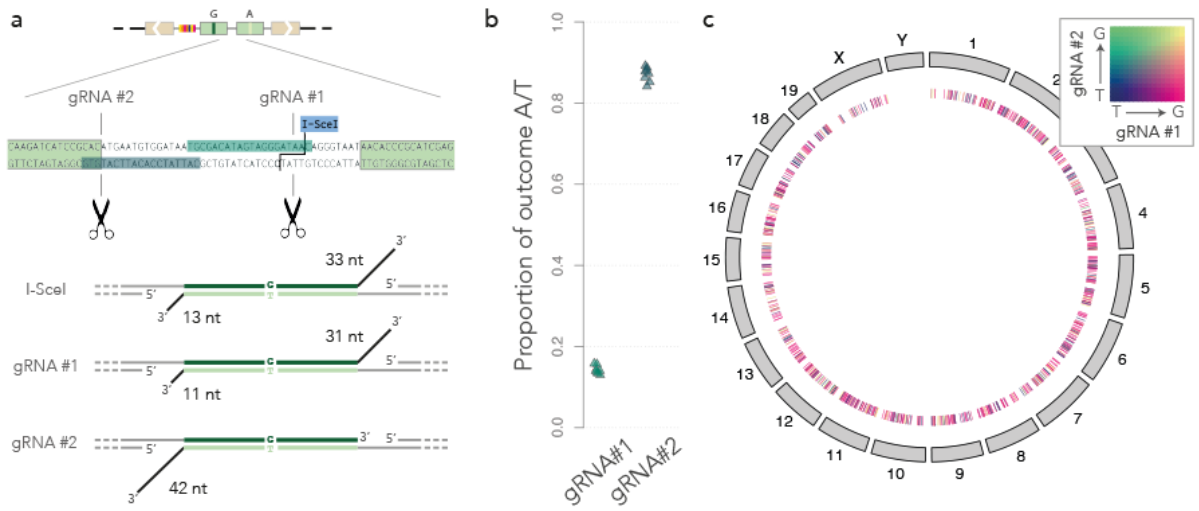

Figure 6. Mismatch resolution and flap asymmetry. **a)** Sketch of the construct and of the guide RNAs used to induce double-strand breaks at alternative locations. **b)** Global bias reversal. The dot plot shows the measured bias toward T with the G:T construct, measured by UMI-PCR 24 hours after inducing a double-strand break with the guide RNAs shown in a). **c)** Local bias reversal. The circular map of the mouse genomes shows a random sample of 1,000 inserted G:T reporters where the mismatch resolution was known with both guide RNAs shown in a). Each tick mark represents an integrated reporter and its color represents the outcome of both experiments, as per the heat map in the top right corner. Most tick marks are magenta, indicating that the reporter was repaired toward G when using guide RNA #1, and toward T when using guide RNA #2.

To test this hypothesis, we used the CRISPR-Cas9 system to change the position of the double-strand break. I-SceI leaves a 3' flap of 33 nucleotides on the top strande, and 13 nucleotides on the bottom strand. We designed a control guide RNA that induces a double-strand break near the I-SceI site, leaving 3' flaps of 31 and 11 nucleotides, respectively. We also designed a second guide RNA that induces a double-strand break at a symmetrically opposite location, leaving 3' flaps of 0 and 42 nucleotides respectively (**Figure 6a**). If the asymmetry between the flaps induces the repair bias, one expects that the first guide RNA should induce the same bias as I-SceI, whereas the second guide RNA should induce a bias in the opposite direction.

Figure 6b suggests that this is indeed the case. The G:T mismatch is repaired in favor of G when initiating the double-strand break with I-SceI or guide RNA #1, and in favor of T when

initiating it with guide RNA #2. To test whether the bias can be reverted on the exact same reporter, we collected all the mapped barcodes for which the repair was measured for both guide RNAs. We then used a color code to depict both outcomes simultaneously (**Figure 6c**). The results show that the bias is reversed on individual reporters when replacing guide RNA #1 with guide RNA #2. The switch is uniform throughout the genome, consistent with the previous finding that the insertion site of the reporter has little effect on the resolution of the mismatch.

Taken together, our results show that the position of the double-strand break can bias the mismatch resolution during SSA. More specifically, in our repair assay, the asymmetry between the 3' flaps results in a repair bias where the strand with the longest flap is more likely to be used as a template.

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 single-strand annealing (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 landscape of the biases in mismatch resolution. We found no evidence that repair is intrinsically biased toward G and C nucleotides. Instead, we found a persistent 60-80% bias toward the strand with the longest 3' flap (**Figure 6**), regardless of the mismatch that was induced. We also observed that the repair bias is uniform throughout the genome, suggesting that the chromatin surrounding the lesion has little influence in this context. Overall, these results have important implications regarding the factors influencing the repair of mismatches.

Teasing apart the individual contributions of DNA damage and repair to mutational processes is challenging. This has been possible on plasmids (14), but the known interactions of mismatch repair with chromatin suggest that mutational biases should be studied directly in the genome (35). In this regard, the TRIP assay developed here is a technical step forward. A similar principle was already used by Gisler *et al*. (23), and Schep *et al*. (36). Interestingly, the authors found that the local state of chromatin influences the choice of repair pathways, which we did not investigate in this study.

We did not find any evidence that the chromatin context influenced the mismatch resolution on our inserted reporters. A limitation of this study is that we have not established that the reporters blend in with the genomic context and acquire all the associated chromatin features. The Sleeping Beauty backbone was previously used by our and other teams to build sensors of the chromatin context (22,37), and the repeated F segment is taken from the green fluorescent protein (GFP), which is commonly used in reporter assays. It is unlikely that the reporter construct contains a promoter, enhancer or insulator, but the generality of our finding will have to be established in future studies. In particular, it will be important to test whether the global bias reversal observed in **Figure 6c** can be generalized to other sequences.

The mismatches produced in our repair assay are coupled to the repair of a double-strand break. In that sense they differ from the typical mismatches that occur during DNA replication. We nevertheless gained general insight into the mechanism of mismatch repair, the most striking of which being that the process can be substantially strand-biased. Here it is worth mentioning that the asymmetry between the strands was not intended in our experimental design; the serendipitous discovery that it was due to the skewed location of the double-strand break was as described in the results. In hindsight, the bias reversal is a strong internal control that the mismatches are repaired within 24 hours and that our experimental system can detect a skew in either direction.

It was recently discovered in yeast that the flaps influence the efficiency of mismatch repair in break-induced replication (38), 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 (39). 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 (15). So overall, our results are in line with previous experiments in the field.

We have not established experimentally that the flaps are exactly as shown in **Figure 6a**, but the repair mechanism of SSA is well enough established that we can accept the model as a starting point for further discussion. In particular, it is interesting to speculate as to how the flaps may influence the choice of the nucleotide to be excised. The repair of the mismatch must be concomitant with that of the double-strand break, and the fact that one strand is favored suggests that there is a race condition between the processing of flaps. In the experiment with guide RNA #2, the disfavored strand has no flap at all, which suggests that the asymmetry is in the timing of flap removal. We thus propose the following “first-nick” model (**Figure 7**): The flaps prevent strand discrimination by the MSH complex; when a flap is removed, a nick or a breach is exposed on the associated strand, providing the signal needed by MSH2-MSH6 to backtrack and resolve the mismatch (40). At that stage, the involvement of the mismatch repair system has not been established experimentally and the central hypothesis of our model is that it takes longer to remove long flaps than short flaps. Whether this is the case will require further investigations.

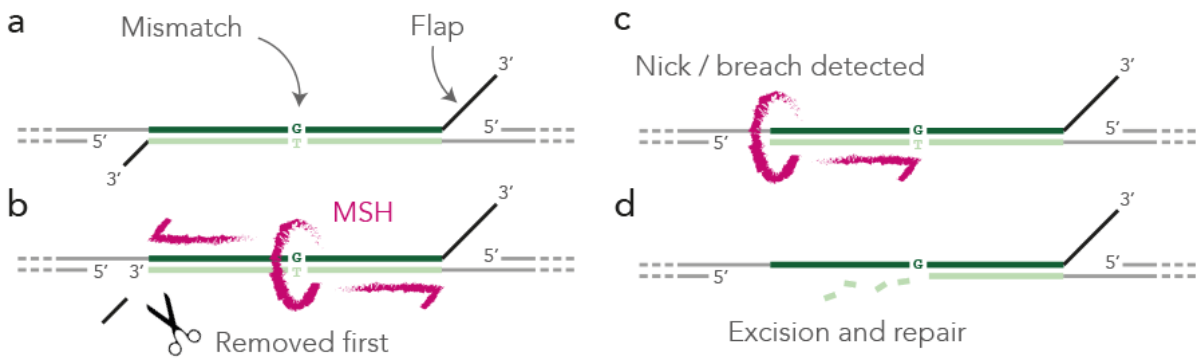

Figure 7. First-nick model. **a)** After resection of the 5’ ends and annealing of the complementary strands, unannealed flaps extend in 3’. **b)** The MSH complex recognizes the mismatches and slides to initiate strand discrimination but the flaps jam the process. Shorter flaps tend to be removed earlier,

allowing for strand discrimination. **c)** The nick of breach exposed by the removal of the flap is recognized by the MSH complex that backtracks to initiate the repair of the mismatch. **d)** The mismatch is repaired as usual by excising and re-synthesizing the strand.

Using the properties of UMI-PCR we could establish that the mismatches were repaired at the time of the assay. The DNA cannot be amplified before the flaps are removed and the gaps are sealed because the PCR primers would not anneal to the template. This suggests that the mismatch is repaired before the double-strand break is fully sealed or shortly thereafter, in line with the first-nick model. In any event, the location of the reporter did not seem to influence whether the mismatch is repaired, which seems to contradict the well established fact that mismatches in late-replicating regions are repaired less efficiently (19,20,41). It may be that mismatches occurring through SSA are more accessible because of the double-strand break. Alternatively, the unique chromatin features of ES cells may favor widespread mismatch repair throughout the genome (42). But this study raises the possibility that a different mechanism is at work: mismatches in late-replicating regions may be repaired with the same probability, but with less accurate strand discrimination. Our model suggests that unannealed regions such as those arising at the convergence of replication forks (43) can mask downstream nicks and obfuscate strand discrimination.

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 molecular details of the damage such as the position of the cut and the length of the 3’ flaps had a large influence on the outcome mismatch resolution.

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). Guide RNAs TGCGACATAGTAGGGATAAC (gRNA1) and GCATTATCCACATTCATGTG (gRNA2) were cloned into plasmid pSpCas9(BB)-2A-GFP (Addgene #48138) by the company VectorBuilder and shipped as pRP[CRISPR]-EGFP-hCas9-U6> {VP_gRNA1} and pRP[CRISPR]-EGFP-hCas9-U6> {VP_gRNA2}, with internal identifiers VB200402-2406dex and VB200402-3189qpa, respectrively. 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 (22).

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 incubating 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) were electroporated with 3 µL ligated products. 0.01% of the electroporated bacteria were plated on an 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 a 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 pCBASceI 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).

Mismatch generation via CRISPR-Cas9 was done similarly. 5 µg of either plasmid pRP[CRISPR]-EGFP-hCas9-U6> {VP_gRNA1} or plasmid pRP[CRISPR]-EGFP-hCas9-U6> {VP_gRNA2} were used for the transfection. Transfected cells were collected 24h post transfection.

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 100 nM 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 (44) 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*. (29) were only available in release mm9.

After mapping, the barcodes were clustered using Starcode (45) 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 midpoint 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-complements 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 mismatched 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 as 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 an estimate of their joint distribution, and the product of their margins as an 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

The networks have three hidden layers with 100 neurons each, and one output layer with 2 neurons. The vectors in output of each hidden layer are batch-normalized and the ReLU activation function is applied (*i.e.*, negative values are set to 0). Batch normalization was shown to improve training speed (31) and the rectified linear unit or ReLU is a common activation function that was empirically shown to mitigate the problem of vanishing gradients (32). The fist two hidden layers include a Dropout step where a random set of input values are set to 0 with probability 0.3. This was shown to mitigate overfitting by forcing some redundancy in the encoding of the information (33). Including a Dropout step at the output of the last hidden layer showed no benefit. Finally, the networks have a residual connection bypassing the last two hidden layers. Such residual connections were shown to improve training speed and model performance (30).

The input layer of the null model has 7 neurons (4 for the construct, 1 for the time point, one for the amplification technology and 1 for the TRIP pool). The input layer of the full model has 87 neurons (the 7 neurons of the null model plus 78 for the chromatin features from Juan *et al*. (29), plus 1 for the GC-content within 10 kb, plus one for the GC-content within 1 Mb). When predicting the GC-content within 10 kb, the input layer of the full model has 85 neurons (the two variables for the GC-content are removed).

The 2 neurons in the output layer encode the two parameters of a Beta distribution. The networks thus transform their input into a distribution over the interval (0,1) that reflects the confidence of the model for all the possible values of the bias at the given genomic location. For instance, when the parameters are both close to 1, the distribution is near uniform and the model is “clueless” about the bias because all the values are equally likely. When one parameter is substantially larger than the other, the model expresses confidence that the bias is strong in the given direction. In line with this interpretation, the loss function that the optimizer sets to minimize is defined as the (negative) log-likelihood of the observations under the local Beta distribution associated with a genomic location.

The datasets were split randomly so as to keep 10% of the records for testing. The remaining 90% were used for training for 100 epochs. At the end of every epoch, the performance of the model was evaluated on the test set.

The neural networks were implemented using custom scripts written in Python 3.7.6 with Pytorch version 1.9.0 for CUDA version 10.2 and Numpy version 1.19.2. Networks were optimized in mini batches of size 256 with the Adam optimizer (46) with a learning rate equal to 0.001.

Code and data access

The data has been deposited in the Gene Expression Omnibus database (GEO accession GSE141211). The scripts used in this study are available on Github at https://github.com/cellcomplexitylab/strand_asymmetry. A Docker container to reproduce the results is available on Dockehub at https://hub.docker.com/r/gui11aume/strand_asymmetry.

Table 2. List of primers.

Primer name	Sequence
L1	NNNNNNNNNNNNNNNNNNNNagatcggaagagcgtcgtg
L2	CGCTAATTAATGGAATCATGAACA
L3	catagGCTAGC NNNNNNNNNNNNNNNNNNNNagatcggaagagcgtcgtg
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

DECLARATIONS

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The raw data for this study is available from the Gene Expression Omnibus under GEO accession GSE141211. Mouse E14 ES cells are available from the American Type Culture Collection under ATCC accession CRL-1821.

Competing interests
The authors declare that they have no competing interests

Funding
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). VOP 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.

Authors' contributions
VOP, ARD, LE and GJF carried out the experiments. GJF performed the analyses. VOP and GJF wrote the manuscript.

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 thank the flow cytometry and the sequencing facility of the CRG for their support. Our most sincere gratitude goes to Fyodor Kondrashov, Maximilian Jösch and Adam Mott for providing material support during the COVID-19 epidemic.

REFERENCES

1. Filipski J. Evolution of DNA Sequence Contributions of Mutational Bias and Selection to the Origin of Chromosomal Compartments. In: Obe G, editor. Advances in Mutagenesis Research. Berlin, Heidelberg: Springer Berlin Heidelberg; 1990. p. 1–54.

2. Freese E. On the evolution of the base composition of DNA. J Theor Biol. 1962;3(1):82–101.

3. Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A. 1962 Apr 15;48:582–92.

4. Sinsheimer RL. The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. J Biol Chem. 1955 Aug;215(2):579–83.

5. Gale JM, Nissen KA, Smerdon MJ. UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. Proc Natl Acad Sci U S A. 1987 Oct;84(19):6644–8.

6. Rideout WM 3rd, Coetzee GA, Olumi AF, Jones PA. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science. 1990 Sep 14;249(4974):1288–90.

7. Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A. 2004 Sep 28;101(39):13994–4001.

8. Ramstein J, Lavery R. Energetic coupling between DNA bending and base pair opening. Proc Natl Acad Sci U S A. 1988 Oct;85(19):7231–5.

9. Brown AJ, Mao P, Smerdon MJ, Wyrick JJ, Roberts SA. Nucleosome positions establish an extended mutation signature in melanoma. PLoS Genet. 2018 Nov;14(11):e1007823.

10. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, et al. The mosaic genome of warm-blooded vertebrates. Science. 1985 May 24;228(4702):953–8.

11. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 2001 Oct;159(2):907–11.

12. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.

13. Duret L, Eyre-Walker A, Galtier N. A new perspective on isochore evolution. Gene. 2006 Dec 30;385:71–4.

14. Brown TC, Jiricny J. Repair of base-base mismatches in simian and human cells. Genome. 1989;31(2):578–83.

15. Bill CA, Duran WA, Miselis NR, Nickoloff JA. 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. 1998 Aug;149(4):1935–43.

16. Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nat Commun. 2019 Aug 29;10(1):3900.

17. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010 Jan 14;463(7278):191–6.

18. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014 Aug 14;158(4):929–44.

19. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation

across the human genome. Nature. 2015 May 7;521(7550):81–4.

20. Weber CC, Pink CJ, Hurst LD. Late-replicating domains have higher divergence and diversity in Drosophila melanogaster. Mol Biol Evol. 2012 Feb;29(2):873–82.

21. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013 Aug 15;154(4):914–27.

22. Corrales M, Rosado A, Cortini R, van Arensbergen J, van Steensel B, Filion GJ. Clustering of Drosophila housekeeping promoters facilitates their expression. Genome Res. 2017 Jul;27(7):1153–61.

23. Gisler S, Gonçalves JP, Akhtar W, de Jong J, Pindyurin AV, Wessels LFA, et al. Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. Nat Commun. 2019 Apr 8;10(1):1598.

24. Spies M, Fishel R. Mismatch repair during homologous and homeologous recombination. Cold Spring Harb Perspect Biol. 2015 Mar 2;7(3):a022657.

25. Mátés L, Chuah MKL, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat Genet. 2009 Jun;41(6):753–61.

26. Meyerhans A, Vartanian J-P, Wain-Hobson S. DNA recombination during PCR [Internet]. Vol. 18, Nucleic Acids Research. 1990. p. 1687–91. Available from: http://dx.doi.org/10.1093/nar/18.7.1687

27. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012 Aug 23;488(7412):504–7.

28. Li F, Mao G, Tong D, Huang J, Gu L, Yang W, et al. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. Cell. 2013 Apr 25;153(3):590–600.

29. Juan D, Perner J, de Santa Pau EC, Marsili S, Ochoa D, Chung H-R, et al. Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs [Internet]. Vol. 14, Cell Reports. 2016. p. 1246–57. Available from: http://dx.doi.org/10.1016/j.celrep.2016.01.008

30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.

31. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Internet]. arXiv [cs.LG]. 2015. Available from: http://arxiv.org/abs/1502.03167

32. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference [Internet]. 2010; Available from: https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf

33. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res.

2014;15(1):1929–58.

34. Sadowski P, Baldi P. Neural Network Regression with Beta, Dirichlet, and Dirichlet-Multinomial Outputs [Internet]. 2018 [cited 2019 Nov 4]. Available from: https://openreview.net/pdf?id=BJeRg205Fm

35. Goellner EM. Chromatin remodeling and mismatch repair: Access and excision. DNA Repair . 2019 Oct 17;85:102733.

36. Schep R, Brinkman EK, Leemans C, Vergara X, van der Weide RH, Morris B, et al. Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. Mol Cell. 2021 May 20;81(10):2216–30.e10.

37. Chen C-K, Symmons O, Uslu VV, Tsujimura T, Ruf S, Smedley D, et al. TRACER: a resource to study the regulatory architecture of the mouse genome [Internet]. Vol. 14, BMC Genomics. 2013. p. 215. Available from: http://dx.doi.org/10.1186/1471-2164-14-215

38. Anand R, Beach A, Li K, Haber J. Rad51-mediated double-strand break repair and mismatch correction of divergent substrates. Nature. 2017 Apr 20;544(7650):377–80.

39. Chakraborty U, George CM, Lyndaker AM, Alani E. A Delicate Balance Between Repair and Replication Factors Regulates Recombination Between Divergent DNA Sequences in Saccharomyces cerevisiae. Genetics. 2016 Feb;202(2):525–40.

40. Putnam CD. Strand discrimination in DNA mismatch repair. DNA Repair . 2021 Jun 19;105:103161.

41. Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. Genome Res. 2014 Nov;24(11):1751–64.

42. Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Res. 2015 Aug;25(8):1091–103.

43. Yao NY, O'Donnell ME. Replication fork convergence at termination: A multistep process. Proc Natl Acad Sci U S A. 2018 Jan 9;115(2):237–9.

44. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012 Dec;9(12):1185–8.

45. Zorita E, Cuscó P, Filion GJ. Starcode: sequence clustering based on all-pairs search. Bioinformatics. 2015 Jun 15;31(12):1913–9.

46. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization [Internet]. arXiv [cs.LG]. 2014. Available from: http://arxiv.org/abs/1412.6980