Chromatin structure influences rate and spectrum of spontaneous mutations in *Neurospora crassa*

Mariana Villalba de la Peña[1], Pauliina A. M. Summanen[1],

Martta Liukkonen[1], Ilkka Kronholm[1]

[1]Department of Biological and Environmental Science, University of Jyväskylä, FI-40014 Jyväskylä, Finland

Running Head: Influence of chromatin on mutation rate

Keywords: fungi, mutation accumulation, genomics, epigenetics, H3K9 methylation, H3K27 trimethylation

Corresponding author:

Ilkka Kronholm

Department of Biological and Environmental Science,

University of Jyväskylä,

P.O. Box 35, FI-40014

Jyväskylä,

Finland

Fax +358 14 617 239

Email: ilkka.kronholm@jyu.fi

Abstract

Mutation rates in different species have been extensively studied in eukaryotes. However, much less is known about variation in mutation rate across the genome. Chromatin modifications may be an important factor in determining mutation rate variation in eukaryotic genomes. We performed a mutation accumulation experiment in the filamentous fungus *Neurospora crassa* and detected mutations in the MA-lines by genome sequencing. We detected 1322 mutations, which happened during asexual propagation. Our results show a markedly different mutation rate and spectrum than previously reported during sexual reproduction. We observed that mutation rate was higher in regions of low GC, in domains of H3K9 methylation, in centromeric regions, and in domains of H3K27 methylation. Rate of single nucleotide mutations in euchromatin was $2.46\ [2.19, 2.77] \times 10^{-10}$. In contrast, this rate in H3K9me domains was tenfold higher: $2.43\ [2.25, 2.62] \times 10^{-9}$. We also observed that the spectrum of single nucleotide mutations was different in H3K9me and euchromatic domains. We validated our statistical model of mutation rate variation by comparing its predictions to natural genetic variation and observed that a moderate amount of extant genetic variation can be predicted by our model of mutation rate variation. Furthermore, we characterize mutation rates of structural variants and complex mutations in *N. crassa*. Our study highlights that chromatin modifications influence mutation rate and accurate evolutionary inferences should take variation in mutation rate across the genome into account.

Introduction

New mutations are the source of all genetic diversity. While considerable amount of contemporary phenotypic changes in populations, such as that observed in plant and animal breeding, may be due to frequency changes of alleles already segregating in the population, all evolutionary change is ultimately dependent on the input of new mutations into the population. Yet, organisms also pay a terrible cost for their ability to evolve, as majority of mutations are deleterious and can lead to adverse outcomes such as hereditary diseases and cancer. Therefore, rates of spontaneous mutation are fundamental to our understanding of evolution and certain aspects of medicine.

Spontaneous mutations are rare events, and used to be difficult to study due to their rarity. However, with the advent of new sequencing technologies it has become possible to capture large numbers of spontaneous mutations for analysis (Katju and Bergthorsson, 2018). Several studies have now estimated mutation rates by directly observing mutations by sequencing mutation accumulation lines or parent-offspring trios (Ossowski et al., 2010; Ness et al., 2012; Keightley et al., 2014; Zhu et al., 2014; Sung et al., 2015; Ness et al., 2015; Keightley et al., 2015; Keith et al., 2016; Wang et al., 2020). These studies have allowed us to estimate mutation rates with high precision and have revealed the spectrum of spontaneous mutations.

We now understand mutation rates rather well, but detailed studies have also revealed unexpected patterns. The process of mutation itself is stochastic, but not all mutations are equally likely. While this has been appreciated for a long time for certain classes of mutations, such as transitions and transversions, there is also variation in mutation rate that seems to depend on the structural features of the genome, such as the organization of chromatin (Makova and Hardison, 2015). In particular, nucleosome positioning (Tolstorukov et al., 2011; Chen et al., 2012; Li and Luscombe, 2020) and chromatin structure due to epigenetic modifications have a strong influence on mutation rates (Schuster-Böckler and Lehner, 2012; Polak et al., 2015; Weng et al., 2019; Monroe et al., 2022). Areas of closed chromatin seem to have higher mutation rates than areas of open chromatin (Makova and Hardison, 2015). In addition, local sequence context, such as GC-content, also has a strong effect on mutation rate (Makova and Hardison, 2015; Ness et al., 2015; Sung et al., 2015).

While we have some idea how chromatin structure influences mutation rates, the picture is far from clear as data comes mainly from only a few species. We don't know if the spectrum of mutations is different in different regions of the genome. Furthermore, to what extent variation in mutation rate determines patterns of observed genetic diversity along with other evolutionary mechanisms is understood mostly from population genetic data rather than direct measurements of mutation, with few exceptions e.g. Monroe et al. (2022). Understanding mutation rate variation across the genome will help us parameterize population genetic models better and infer how mutation, natural selection, and demographic processes jointly shape genetic diversity (Johri et al., 2020).

To examine patterns of mutation rate variation we performed a mutation accumulation experiment (Halligan and Keightley, 2009) in the filamentous fungus *Neurospora crassa*, and we sequenced the genomes of these mutation accumulation lines using short read sequencing. *Neurospora crassa* is a filamentous fungus that has a facultative sexual cycle. A previous study in *N. crassa* investigated mutation rate during sexual reproduction and revealed that mutation rate is especially high in regions of the genome targeted by unique genome defense mechanism (Wang et al., 2020). However, in Wang et al. (2020) only a small number of mutations was collected during asexual reproduction, which was not enough to infer variation in mutation rate across the genome. Our study complements this previous study by allowing us to characterize the determinants of mutation rate and spectrum during asexual reproduction. We used information about the chromatin structure of *N. crassa* to model variation in mutation rate across the genome. We also resequenced strains of *N. crassa* obtained from natural populations and compared our mutation model to patterns of natural genetic variation to asses whether natural genetic variation reflects observed mutation rate variation.

Materials and methods

Mutation accumulation experiment

To be able to obtain a large number of mutations we used a mutation accumulation (MA) experiment. In MA-experiments an ancestor is line is split into multiple lines and these lines are grown in a manner that minimizes the effective population size and thus the efficacy of natural selection (Figure 1A). This way even mildly deleterious mutations can accumulate in these lines (Halligan and Keightley, 2009). We started the MA experiment with two different strains: 2489 *mat A* and 2489 *mat a*. These strains share an identical genetic background, but have different mating types. We have previously generated these strains by backcrossing mating type *a* from strain 4200 into 2489 (Kronholm et al., 2020). Their Fungal Genetics Stock Center ID's are: B 26708 and B 26709. We used 20 lines for both these strains, giving 40 MA-lines in total. Common protocols for culturing *N. crassa* were followed, and sorbose plates were used to induce colonial morphology on plates (Davis and de Serres, 1970). The experiment was started by picking a single colony from a sorbose plate for both ancestors and transferring that colony into a test tube with 1 ml of flat VM with 1.5% agar and 1.5% sucrose (Metzenberg, 2003). Tubes were incubated at 25 $^{\circ}\mathrm{C}$ for 3 days to allow conidia (asexual spores) to develop. Then we picked small amount of conidia with a loop into a tube with 1.4 ml of 0.01% Tween-80, we then pipetted 1 µl of this conidial suspension into a 50 µl water droplet on a sorbose plate and spread it. We incubated the plates at room temperature for 2 days and picked single colonies to establish the MA lines. The MA lines were transferred the same way, so that a single colony was always picked randomly from a sorbose plate to propagate the MA line (Figure 1B). We tested that 2 days of incubation was enough time for all colonies to appear on plates. Combining the time of 2 days on plates and 3 days in a tube a single transfer took 5 days. We propagated the MA-lines for 40 transfers, the ancestors and periodic samples were stored frozen in suspended animation until sequencing.

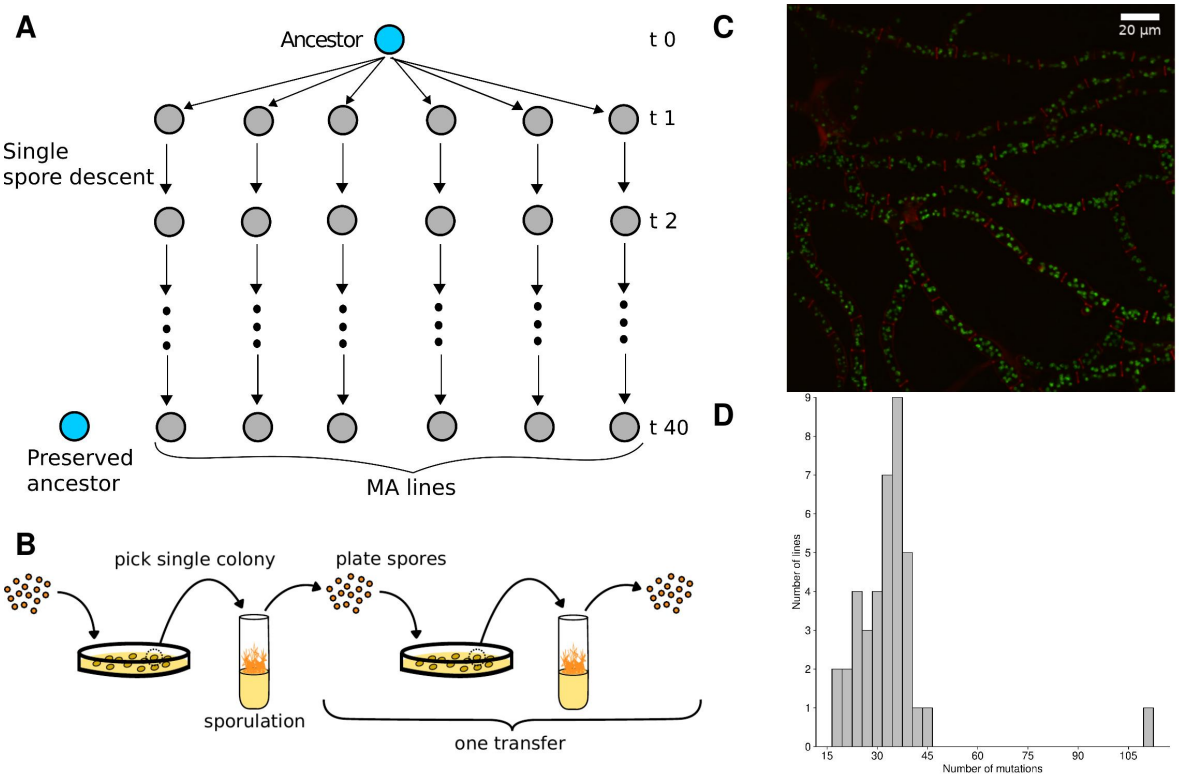

Figure 1: A) Overview of the mutation accumulation experiment. Ancestor is split into multiple lines, which are propagated via single spore descent. This minimizes the efficacy of natural selection and allows even deleterious mutations to accumulate in the MA-lines. B) The MA-experiment was transferred by always picking a colony originating from a single spore from a plate, moving this to a VM-tube to allow sporulation, then picking spores, diluting them and spreading to a sorbose plate. C) Micrograph of *N. crassa* mycelium, showing nuclei fluorescing green and cell walls in red. D) Distribution of mutations in the MA-lines.

Estimating number of mitoses in the MA-experiment

Neurospora crassa is a filamentous organism, and it does not have a defined germline. All parts of the mycelium are able to produce structure that make asexual spores. Thus the number of transfers during the MA-experiment does not correspond to a generation in a natural way. Therefore the reasonable unit for measuring mutation rate is the number of mutations per mitosis.

To estimate the number of mitoses that happened in the MA-experiment during one transfer, we needed to obtain data about the number of nuclei present in each phase of a transfer: in a colony on a sorbose plate, in the mycelium on VM in a test tube, and the conidia produced in the test tube. To estimate the density of nuclei per μm^2 of hyphae we used the strain *mat A his-3*$^+$*::Pccg-1-hH1*$^+$*-sgfp*$^+$ (FGSC# 9518) which expressed a green fluorescent protein that had been fused into histone H1 (Freitag et al., 2004). We grew the strain on plates with either normal VM medium or sorbose medium, cut out a piece of the agar, and mounted it on a glass coverslip using the inverted agar block method (Lichius and Zeilinger, 2019). We used Congo Red to stain cell walls: a 20 µl droplet with 2 µM Congo Red was pipetted to a glass coverslip and an agar block with the side carrying the mycelium was placed face down in the droplet.

Samples were imaged with a Nikon A1R confocal microscope, GFP was excited with a 488 nm laser and detected with a 515/30 emission filter, Congo Red was excited with a 561 nm laser and detected with a 595/50 emission filter. Plan apochromat air objectives 20x (numerical aperture 0.75) and 40x (numerical aperture 0.95) were used. Laser power was set as low as possible to avoid saturated pixels. We imaged vertical stacks of the mycelium, and used imageJ2 (Rueden et al., 2017) to count the area covered by hyphae in sections of the image, and counted the number of nuclei in these areas (Figure 1C). We then estimated the number of nuclei in the different phases of a transfer, and calculated the number of mitoses the MA-lines went through during each transfer (see Supplementary methods for details).

Strains from natural populations

We obtained 33 strains from the Fungal Genetics Stock Center (McCluskey et al., 2010), and re-sequenced these strains. In addition we obtained genome sequencing data for additional 23 strains from Zhao et al. (2015) by downloading this data from the short read archive. In total, including the laboratory strain 2489, we had 57 strains with sequencing data (Table S2).

DNA extraction

To get high quality DNA for sequencing, the natural strains, MA-lines, and the ancestors were grown in 5 ml of liquid VM for two days at 25 $^{\circ}\mathrm{C}$ with shaking. We harvested the mycelium and freeze dried it over night in a lyophilizer. Dried mycelium was then ground with a glass bead in Qiagen Tissue Lyzer for two times 20 s with frequency of 25 s^{-1}. Then 500 µl of extraction buffer was added (10 mM Tris pH 8, 0.1 M EDTA, 150 mM NaCl, and 2% SDS), and the powdered tissue dissolved by shaking. Then samples were extracted with 750 µl of 25:24:1 Phenol:Chloroform:Isoamylalcohol and keeping the aqueous phase. We added 2 µl of RNAse A (10 $\mathrm{mg/ml}$) and 50 U of RNAse I to each sample and incubated them for 1 h at 37 $^{\circ}\mathrm{C}$. Samples were then extracted with 750 µl of chloroform, 1 ml of 100% ethanol was added to samples and DNA precipitated for 1 h at -20 $^{\circ}\mathrm{C}$. Then DNA was pelleted with centrifugation at 4 $^{\circ}\mathrm{C}$, ethanol aspirated, pellet washed with 70% ethanol, and air dried. We then added 77.5 µl of TE-buffer to elute the samples and incubated at 37 $^{\circ}\mathrm{C}$ to help dissolve the pellets. We observed that occasional small DNA fragments would remain in the samples and to remove these we did a polyethyleneglycol precipitation: we added 12.5 µl of 4 M NaCl, mixed and added 12 µl 50% PEG (P3350), mixed and precipiated DNA over night at 4 $^{\circ}\mathrm{C}$. DNA was then pelleted with centrifugation and supernatant aspirated, pellet was washed twice with 70% ethanol, and aspirated. Pellets were eluted to 55 µl of TE-buffer as above. DNA concentrations were measured with Qubit Broad Range Kit, and DNA quality was checked by running 2 µl of sample on an 0.8% agarose gel.

Genome sequencing

Sample DNA was sent for sequencing at Novogene (Cambridge, UK) using Illumina platform with paired-end 150 bp libraries. Libraries were prepared by fragmenting the DNA by sonication, adapter ligation and PCR-amplification. Libraries were sequenced to 30x target coverage. Reads with adapter sequences, > 10% N's, or > 50% bases of low quality Q score ≤ 5 were removed.

Read mapping and genotyping

For reference genome in read mapping we used the *N. crassa* reference genome (NC12), where we had included the mitochondrial genome and mating type *a* region as additional contigs, to be able to map reads to this region in *mat a* strains. Reads were mapped using BWA-MEM version 0.7.12-r1039 with default parameters (Li, 2013). Alignment files were sorted and indexed with samtools and read groups were added with picardtools.

We used the GATK version 4.2.0.0 (McKenna et al., 2010) pipeline to call single nucleotide mutations (SNMs) and small indels. First we ran Haplotypecaller for each sample individually to make a g.vcf file. Haplotypecaller was run with otherwise default parameters, emitting all sites, and in diploid mode. *N. crassa* is haploid but in our experience mapping errors manifest as heterozygous sites in haploid organisms (Kronholm et al., 2017), see also Li (2014) and Ness et al. (2012). This makes it easier to filter out mapping errors afterwards. We then consolidated all of the samples together into a database using the GenomicsBDImport function in GATK. Samples were then jointly genotyped with the GenotypeGVCFs function to produce a vcf file with all samples.

We used wormtable version 0.1.5 (Kelleher et al., 2013) to convert the vcf file into an indexed database and then a custom python script to filter for high quality sites. For a site to be included as a candidate mutation, we required the genotypes of the ancestor and the MA-line to differ for that site, the site had to have five or more reads from both the ancestor and the sample, the site had to have genotype quality greater or equal to 30 for both the ancestor and the sample, and sites that were called heterozygous in either the ancestor or the sample were excluded. We considered invariant sites to have been called if their reference genotype quality was greater or equal to 30. Fur-

thermore, we checked all candidate mutations manually by inspecting the alignments from BWA and or Haplotypecaller in IGV (Thorvaldsdóttir et al., 2013) to produce the final dataset of curated mutations. Based on our manual inspection our filtering criteria were stringent enough for our high coverage haploid genomes to remove mapping errors and leave only real mutations, as only very few candidate mutations had to be rejected based on manual inspection and most mutations were unambiguous.

For genotyping SNPs in the strains sampled from natural populations, the above pipeline was used to call genotypes. Other variants than SNPs were excluded. For a site to be included, it had to be polymorphic in the sample, with mean read depth five or greater, genotype quality 30 or greater, and mapping quality 40 or greater across all samples. Then these same criteria were applied for each individual sample, and if a sample failed to meet the quality filters, its genotype was recorded as missing data. Heterozygous sites were excluded. Sites were also excluded if > 90% of samples had missing data. Sites were called as monomorphic if mean reference genotype quality was 30 or greater and read depth 5 of greater across all samples. Then these same criteria were applied to individual samples, genotypes were recorded as missing data if a sample did not pass the filters.

Genotyping structural variants

There are several algorithms available to detect structural variants (SVs) from short-read sequencing data. However, because this kind of data is prone to base calling and alignment errors, none of the available computational algorithms can accurately and sensitively detect all types and sizes of SVs (Kosugi et al., 2019). To overcome this limitation it is common to use several algorithms and merge their outputs to increase sensitivity and precision. We assessed the performance of four different SVs algorithms (DELLY, Lumpy, PINDEL and SVaba, see supplementary methods) using simulated data, and selected the best performing ones (DELLY and Lumpy) to call SVs on the MA-lines (Table S3).

For calling SVs in the MA-lines we first aligned the reads to the reference genome using BWA-MEM, excluded duplicated reads with SAMBLASTER version 0.1.26 (Faust and Hall, 2014)

and extracted the discordant paired-end and split-read alignments using SAMTOOLS version 1.9 (Danecek et al., 2021). DELLY was used as indicated in the recommended workflow (Rausch et al., 2012). For LUMPY the read and insert length were extracted from alignment files using SAMTOOLS and the SVs were genotyped using SVTyper version 0.7.1 (Chiang et al., 2015). To filter out SVs that were not exclusive of the MA-lines we used SnpSift version 5.0e (Cingolani et al., 2012). We removed those calls with a genotype quality score lower than 30 and read depth below 10. We also removed SVs where the MA-lines had the same genotype as the ancestor. The analysis with both callers were carried out in somatic-germline mode, considering MA-line as somatic and the ancestor as germline. All of the SVs detected by each caller were manually verified by inspecting the alignment files in IGV. We also scanned for CNV variants using two detection programs, CNVnator version 0.4.1 (Abyzov et al., 2011) and CNV-seq version 0.2-7 (Xie and Tammi, 2009), see also supplementary methods. However, we did not find any evidence of CNV events in the MA-lines.

Validation of mutations with Sanger sequencing

To verify a sample of the observed mutations, we performed Sanger sequencing using standard methods. We selected mutations that passed our threshold but had the lowest quality scores. We attempted to confirm 83 annotated low quality score mutations, 30 of them were point mutations, 37 small indels and 16 larger structural variants. List of primers and confirmed mutations are given in the supplementary file S1. Additionally, complex mutations were selected for verification and we sequenced time points from the middle of the MA-experiment to see if the multiple changed sites appeared together in the MA-line.

Chromatin modifications

ChIP-seq reads for H3K9 methylation and H3K27 trimethylation were obtained from Jamieson et al. (2013), accession numbers SRX248101 and SRX248097. Data for H3K36 methylation was obtained from Bicocca et al. (2018), accession number SRX4549854. Reads were aligned to the

reference genome using BWA, and duplicate reads were removed by Picard tools. Domains of chromatin modifications were identified using RSEG 0.4.9 (Song and Smith, 2011). Data for centromeric regions were obtained from Smith et al. (2011) and Wang et al. (2020). Furthermore, we used the data of duplicated regions that were defined by Wang et al. (2020).

Statistical analysis

Mutation rates were estimated using Bayesian Poisson models implemented with the Stan language (Carpenter et al., 2017) interfaced from R 3.6.0 (R Core Team, 2019) with the "brms" package (Bürkner, 2017). The basic model for estimating mutation rate was

$$\begin{aligned} y_i &\sim \mathrm{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \alpha \\ \alpha &\sim \mathrm{N}(0, 10) \end{aligned} \tag{1}$$

where y_i is the number of mutations in ith MA-line, λ is the poisson rate parameter, and α is the intercept. The linear model part was modified accordingly if other predictors were used. We can then calculate mutation rate, μ, from posterior distributions as

$$\mu = \frac{\exp(\alpha)}{Ntm} \tag{2}$$

where N is the number of called nucleotides, t is the number of transfers the MA-lines went through, and m is the number of mitoses per transfer. To get mutation rate per genome N is removed from the denominator. We used a weakly regularizing prior for α. Settings for MCMC estimation were: 1000 iterations of warm-up followed by 3000 iterations of sampling with four independent chains. MCMC convergence was monitored by traceplots and $\hat{R}$ values. No con-

vergence problems were observed. Differences in mutation rates were tested by computing ratios using posterior distributions, generally estimates and 95% highest posterior density intervals were reported.

For cases where relative mutation rates were computed for different classes of bases or trinucleotides the linear model part in equation 1 was modified so that model was

$$\log(\lambda_i) = \log \tau_j + \alpha_{[j]} \tag{3}$$

where τ_j is an offset term for class j that allows taking into account differences in the abundance of certain classes (McElreath, 2015), such as higher frequency of A's and T's than G's and C's in the genome. Furthermore, if we calculate the expected frequencies of mutations across different classes under the assumption that all mutations are equally likely from the total number of observed mutations and use this expected frequency as the offset parameter, then $\exp(\alpha)$ yields the relative mutation rate of class j.

To perform model comparisons we used the leave-one-out information criterion (LOOIC) (Vehtari et al., 2017) where possible, and widely applicable information criterion (WAIC) otherwise (McElreath, 2015; Vehtari et al., 2017). When we assessed how well did the mutation model predict the natural genetic variation we used Bayesian version of R^2 (Gelman et al., 2019) to assess the model fit.

To analyze effects of local base composition on mutation rate, we extracted the adjacent base-pairs for every point mutation. There are 64 different trinucleotides, but as we cannot know in which strand the mutation originally occurred we grouped the trinucleotides into 20 different classes based on complementary and symmetry. For example, trinucleotides ATA and TAT are complementary and were grouped, as were complementary trinucleotides that are mirror images of each other, such as: GCC, CGG, CCG, and GGC. Relative mutation rate was analyzed using a model following equation 3.

We further analyzed the relative mutation rates of different trinucleotides based on how many pyrimidine dimers they contain. We used a model where we predicted the relative mutation rate of a

trinucleotide with the number of different classes of pyrimidine dimers. We included the estimated error of relative mutation rate in the model. The model used for the whole genome was

$$
\begin{aligned}
y_{est,i} &\sim \mathrm{N}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{TT} x_i + \beta_{CCT} c_i \\
y_{obs,i} &\sim \mathrm{N}(y_{est,i}, y_{sd,i}) \\
\alpha, \beta_{TT}, \beta_{CCT} &\sim \mathrm{N}(0, 10)
\end{aligned} \tag{4}
$$

where $y_{obs,i}$ is the median of ith observed relative mutation rate, $y_{sd,i}$ is the standard deviation of ith relative mutation rate, $y_{est,i}$ is the ith estimated relative mutation rate, α is the intercept, β_{TT} is the slope effect for TT dimers, and β_{CCT} is the slope effects for CC or CT dimers. We tested different models with different classes of pyrimidine dimers, either across the whole genome or in different domains (Table S5).

For analysis of homopolymers, repeated sequences in the genome were detected using MISA (Beier et al., 2017), we used 5 bp as the minimum homopolymer length and extracted repeat counts for all homopolymer loci in the genome. Counts of homopolymers of different lengths were used as an offset term in a model used to estimate mutation rates in homopolymers.

Population genetics

To obtain measures of genetic variation across the genome we used the sample of genomes from strains sampled from natural populations, and calculated θ_W, which is a measure of DNA sequence variation, across the genome in 200 bp windows. We estimated θ_W following Ferretti et al. (2012), which allows us to deal with missing data.

Results

Number of mitoses in the experiment

We estimated the number of mitoses the MA-lines went through based on counting nuclei in different phases of one transfer (Figure 1B), based on our estimate the MA-lines went through 25 [25, 26] mitoses in a single transfer. For the whole experiment of 40 transfers, this means that the MA-lines went through 1015 [1003, 1026] mitoses.

Mutations in the MA-lines

After sequencing the MA-lines, it became apparent that one of the lines had many of the same mutations as another line, likely due to a mislabeling or contamination at some part of the MA-experiment. This line was excluded from the analysis, leaving 39 MA-lines in the data. Of the low quality mutations selected for Sanger verification: PCR or sequencing failed in 6 out 30 point mutations, and the remaining 24 point mutations were all confirmed. For the 37 small indels, PCR or sequencing failed from 10, while 20 were confirmed, and 7 were false positives. Of the 16 tested SVs, PCR or sequencing failed for 9, 5 were confirmed and 2 were false positives. Since these mutations were a sample of the mutations with the worst genotype qualities, we consider that our genotyping for point mutations is accurate, while there is some uncertainty for small indels and SVs. All mutations that were false positives were excluded from the data.

In total we observed 1322 mutations, with a median of 33 mutations per MA-line. One of the MA-lines had an excess amount of mutations (Figure 1D), and it is possible that a mutation happened in this line that increased mutation rate. The breakdown of different mutation types was 1077 single nucleotide mutations, 134 insertions, 97 deletions, 9 complex mutations where a single mutational event created multiple nucleotide changes, and 5 translocations. Total mutation rate during asexual propagation was 0.03 [0.03, 0.04] mutations / genome / mitosis.

Mutation rate variation across the genome

Next we examined the distribution of mutations across the seven *N. crassa* chromosomes. We observed that mutations were not uniformly distributed along the chromosomes, but were concentrated in centromeric regions and regions of the genome marked by H3K9 methylation (Figure 2A). The duplicated regions *N. crassa* genome defined by Wang et al. (2020) were almost perfectly correlated with H3K9me domains (Figure 2A). Furthermore, H3K9me and H3K36me domains were almost perfect mirror images of each other (Figure 2A), so we did not use H3K36me domains in the analysis as they contain the same information as absence of H3K9me. Examining relative mutation rates confirmed that mutation rate was over six fold higher in centromeric and H3K9me domains (Figure 2B), while the effect of H3K27me domains was much smaller: mutation rate in H3K27me domains relative to euchromatin was only 1.4 [1.1, 1.78].

We also observed that GC-content displayed a bimodal distribution and was lower in H3K9me and centromeric domains, with distribution of GC-content overlapping very little with euchromatic domains (Figure 2C). This raises the question does the higher mutation rate in H3K9me domains arise from the lower GC-content itself, some other factor related to the chromatin modifications, or a combination of GC-content and chromatin modifications. To rule out a Simpson's paradox type of effect we examined the effects of GC-content on mutation rate within the different domains, and observed that lower GC-content leads to higher mutation rates within each domain, at different ranges of GC-content (Figure S1). Although the pattern was not that clear in H3K27me domains, as few mutations were observed in H3K27me regions that had low GC-content. Thus, we could not ignore GC-content as an explanation for this pattern.

To investigate the joint effects of GC-content and chromatin modifications we fitted models with different predictors, including GC-content, H3K9 domain, H3K27 domain, and centromeric domains. Based on model comparisons the model with best predictions included the effect of GC-content, the effect of H3K9 domain, effect of H3K27 domain, the effect of centromeric region, and the interaction between GC-content and H3K9-domain (Table S7). Based on model weights, the next best model that included additional interaction between H3K27 domain and GC-content was

also plausible (Table S7). However, the overall predictions from these two models were similar. There were only a few mutations in low GC areas of H3K27me domains, so there is some uncertainty in estimating a different slope for H3K27me domains, therefore we prefer the first model with the highest weight. Based on model estimates GC-content has a strong effect on mutation rate, with areas of low GC having higher mutation rates (Figure 2D). Within H3K9me domains, GC-content has a smaller effect on mutation rate, and centromeric regions have a statistically detectable increase in mutation rate on top of H3K9me domain effect (Figure 2D, Table S8), even if centromeric regions always have H3K9me as well. H3K27me also increased mutation rate on top the GC-effect (Table S8).

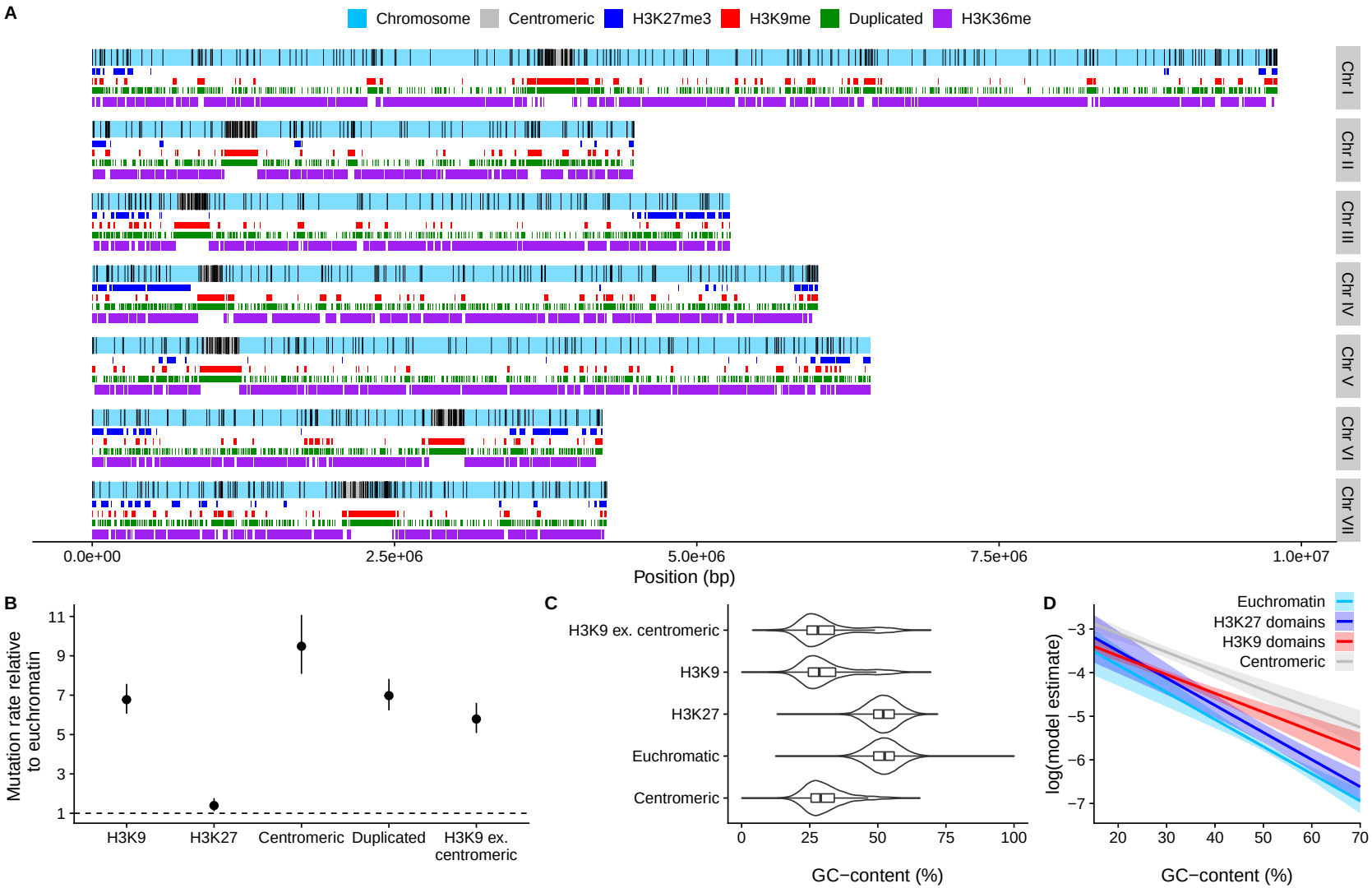

Figure 2: A) Distribution of mutations along the seven *N. crassa* chromosomes. Black vertical lines indicate mutations. Centromeric regions, H3K27 trimethylation, H3K36 methylation, H3K9 methylation domains, and duplicated regions are shown. B) Relative mutation rates for different genomic domains. H3K9 ex. centromeric are H3K9 domains where those domains that overlap with centromeric domains have been excluded. C) GC-content for the different genomic domains. D) Model estimates (on a log-scale) for mutation rate from a model with GC-content, H3K9, and centromeric domains as predictors.

Genetic variation in natural populations and mutation rate

To further validate our mutation model results, we examined if natural genetic variation can be explained by our mutation model. First we calculated θ_W, a measure of genetic variation, across the genome in 200 bp windows. We observed that θ_W was different for the different genomic domains (Figure 3A). θ_W was 0.012 [0.002, 0.043] in euchromatin, and elevated in centromeric and H3K9me domains like mutation rate. In addition, θ_W was elevated in H3K27me domains, 0.024 [0.004, 0.064], perhaps partly reflecting elevated mutation rate in H3K27me domains.

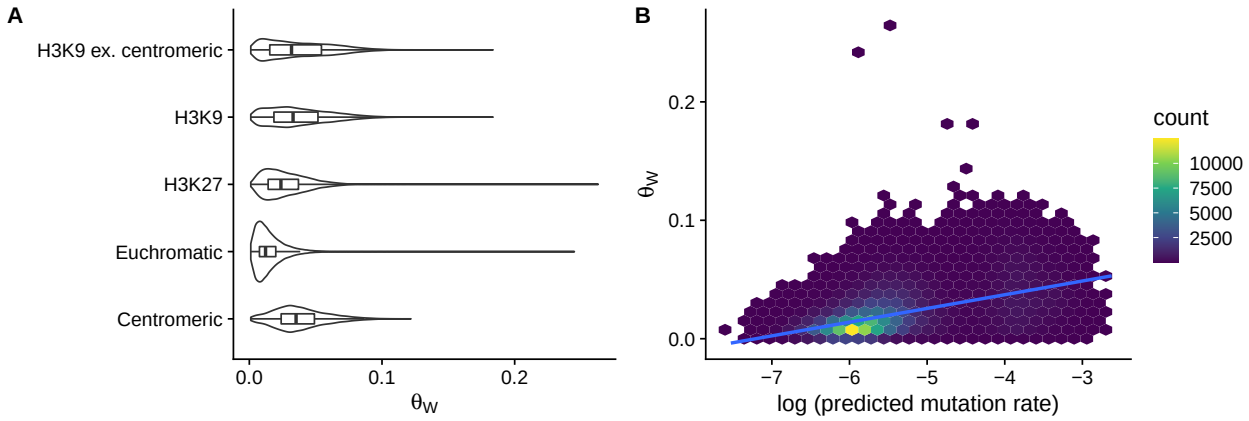

Figure 3: A summary statistic of genetic variation, θ_W, was calculated across the genome in 200 bp windows. Only windows where $\theta_W > 0$ were included in the analysis, $n = 190669$. A) Distribution of θ_W among different genomic domains. B) Relationship with θ_W and predicted mutation rate. For plotting data was been binned into hexes, because of the high number of overlapping points, and number of windows falling into each hex is shown by the legend.

To examine how genetic variation segregating in natural populations is dependent on mutation rate variation across the genome, we used our mutation model (Table S8) to predict variation in θ_W across the genome. We calculated predicted mutation rate for each 200 bp window across the genome and observed that our model did predict moderate amount of variation in θ_W (Figure 3B). A measure of model fit, the Bayesian R^2 value was 0.3 [0.29, 0.3]. While this may seem a rather low R^2 one should take into account that this is after our mutation model has been challenged with completely new data, and not only mutation, but other evolutionary mechanisms also influence θ_W. Therefore mutation rate has a substantial influence on the amount of genetic variation that is present across the genome of *N. crassa*.

Rate and spectrum of single nucleotide mutations

Next we examined the rate and spectrum of different types of mutations more carefully. The rate of single nucleotide mutations (SNM) was 6.7 [6.32, 7.11] $\times 10^{-10}$ mutations / bp / mitosis across the whole genome. SNM rate in euchromatic regions was 2.46 [2.19, 2.77] $\times 10^{-10}$ mutations / bp / mitosis, and SNM rate in H3K9me domains was 2.43 [2.25, 2.62] $\times 10^{-9}$ mutations / bp / mitosis. The ratio of transition to transversion rates over the whole genome was 1.08 [0.96, 1.21], which is on the low end of reported Ts/Tv ratios. The Ts/Tv ratio of euchromatic regions was 1.49 [1.17, 1.91], which was higher then the Ts/Tv ratio in H3K9me domains 0.93 [0.8, 1.08].

As can be seen from the different transition over transversion ratios, we observed that spectra of single nucleotide mutations was different for H3K9me domains versus rest of the genome (Figure 4A). A:T $\rightarrow$ C:G and C:G $\rightarrow$ G:C transversions were more common in H3K9 domains, while the A:T $\rightarrow$ G:C transition was less common (Figure 4B). Over the whole genome the four different transversions occur at similar rates (Figure 4A), and the 2 transitions at higher rates. The A:T $\rightarrow$ G:C transition was the most common SNM, and it occurred at a higher rate than the other transition. The ratio of A:T $\rightarrow$ G:C transitions to C:G $\rightarrow$ T:A transitions was 1.23 [1.04, 1.44].

Mutational mechanisms of SNMs

In order to understand what factors are influencing SNM rates we looked at the effect of local base pair context on SNM rate. For each SNM we extracted the two adjacent base pairs to form a trinucleotide. We combined trinucleotides with respect to both sequence complementary and directional symmetry, which leaves 20 trinucleotides in total. After taking into account the frequencies at which different types of trinucleotides occur, mutations were not equally distributed across trinucleotides. For trinucleotides where the mutating base was A:T we observed that mutations occurred more likely in those trinucleotides that were flanked by either two A's or T's or an A and a T. Whenever one of the flanking base pairs was C or G mutations were less common (Figure 5). When the mutating base was C:G, we observed that when C was flanked by a T and no purines were in the trinucleotide, mutation rate was increased. For instance, TCT:AGA trinucleotides had the highest

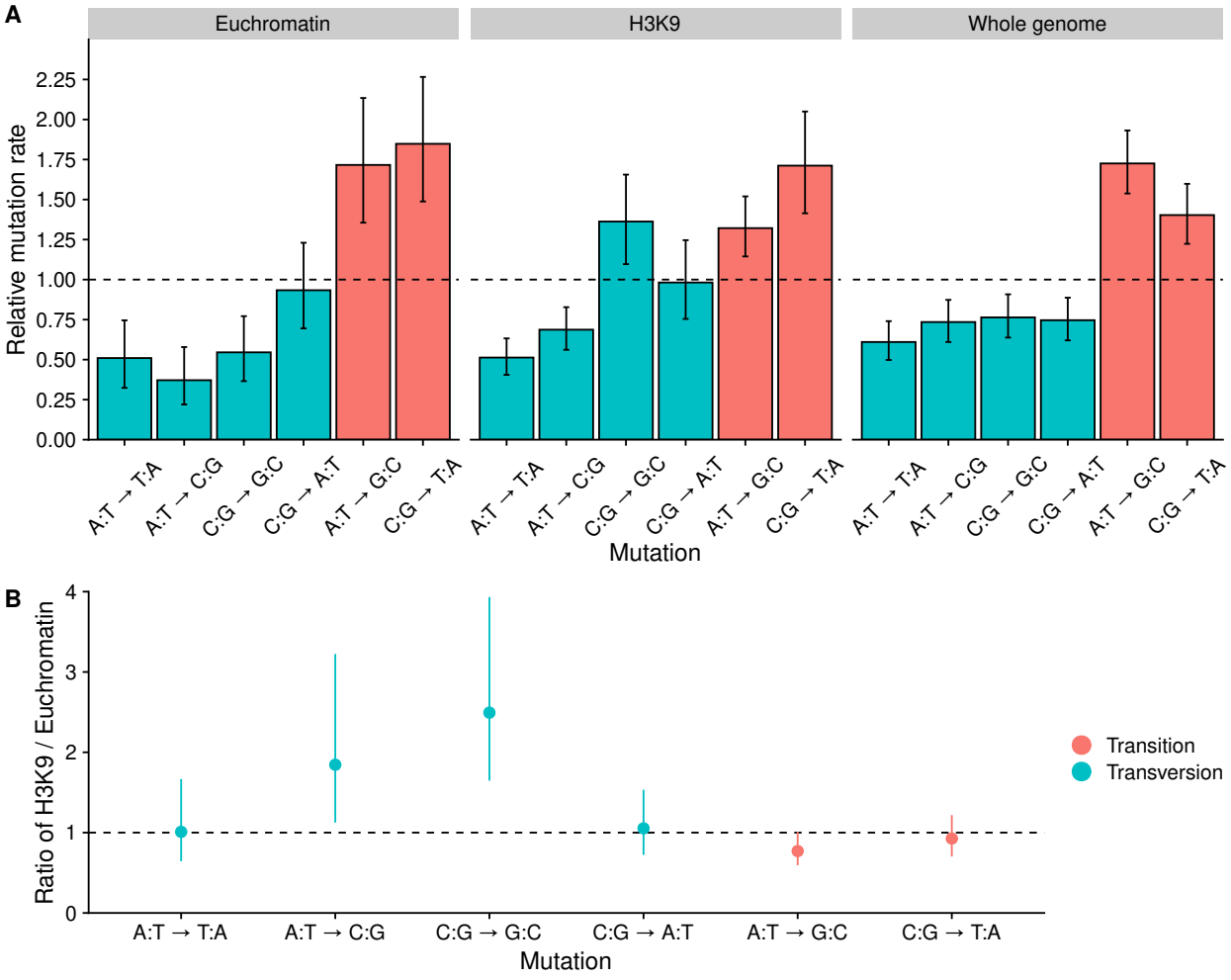

Figure 4: Spectrum of relative SNM rates, dashed line shows the expected rate if all mutations occurred at equal frequencies. Nucleotide frequencies were taken into account in calculating the relative rates. Error bars are 95 % HPD intervals. Transversions are colored blue, transitions red.

mutation rates. However, whenever C was flanked by an A (or G by a T) these triplets had lower relative mutation rates (Figure 5).

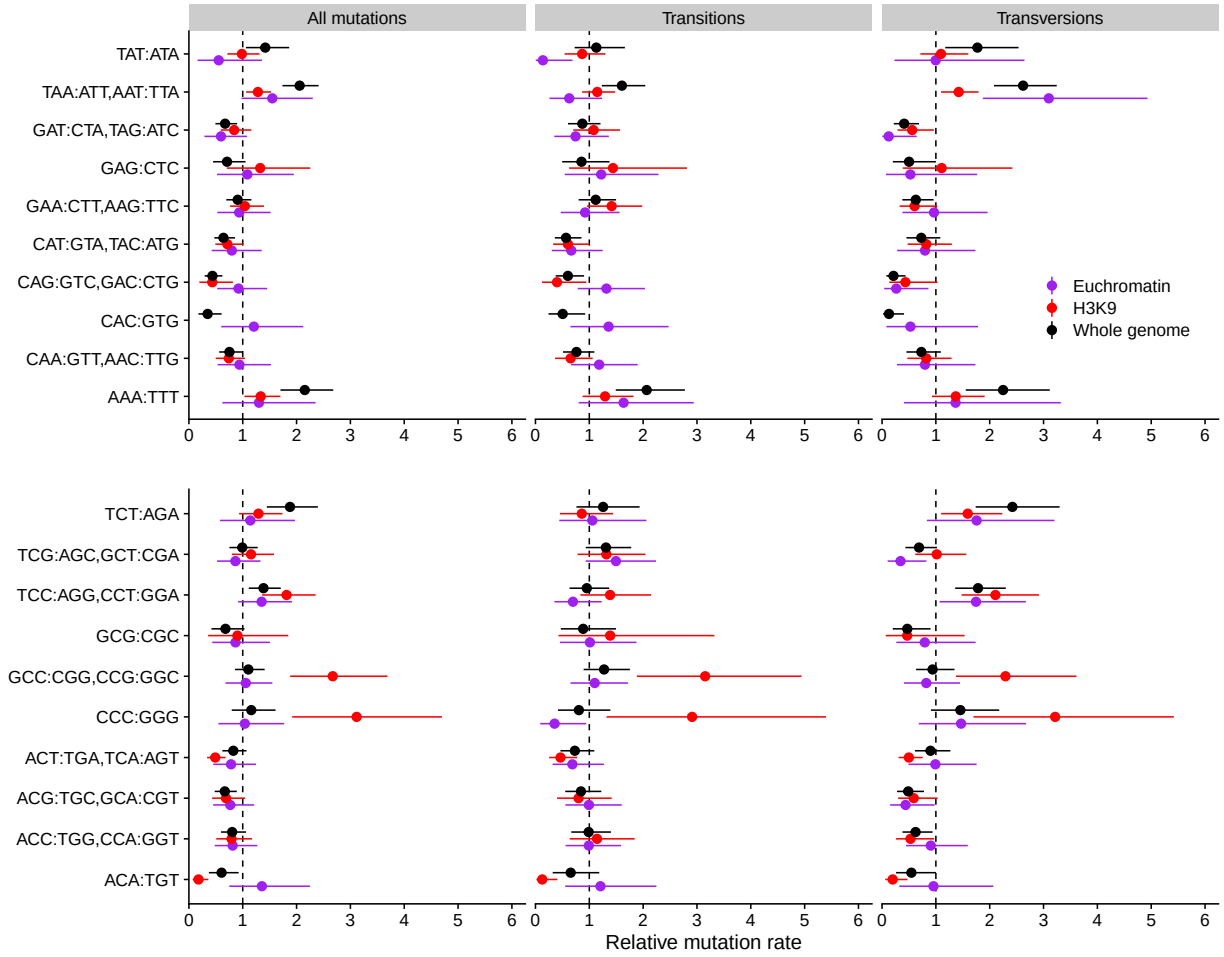

Figure 5: Relative mutation rates for different classes of trinucleotides. Trinucleotides where the middle base is A:T are on the top row and C:G trinucleotides on the bottom row. Range shows 95% HPD interval of the relative mutation rate. Colours show genomic domains.

We also observed domain specific differences in relative trinucleotide mutation rates. In particular trinucleotides containing two adjacent C's had much higher relative mutation rates in H3K9me domains (Figure 5). TAT:ATA trinucleotides also had a higher relative rate in H3K9me domains than in euchromatic regions. In ACA:TGT trinucleotides relative mutation rate in euchromatic regions was higher than in H3K9 domains, and this was also seen in transversions for TAA:ATT trinucleotides (Figure 5).

To examine what could be driving differences in trinucleotide mutation rates, we tested the idea

that trinucleotides with one or more pyrimidine dimers have increased mutation rate, as pyrimidine photodimerization has been suggested to be an important mutational mechanism (Rochette et al., 2003). We used a model with the observed trinucleotide mutation rates as a response and the number of pyrimidine dimers the trinucleotides contained as predictors. We tested different models with separate classes for different dimers, some, or all pyrimidine dimers combined for the different genomic domains. For the whole genome, none of the models were clearly preferred, except that the models with pyrimidine effects gave superior predictions compared to a model with intercept only (Table S5). We show results for a model with separate predictors for TT dimers and CT and CC dimers combined (Table S6). For H3K9 domains, a model with all classes as separate was strongly preferred (Table S5), and for euchromatic regions none of the models were strongly preferred. Accordingly, in euchromatic regions the pyrimidine dimers had no effect on mutation rate (Table S6).

We observed that pyrimidine dimers did increase relative mutation rate of trinucleotides over the whole genome (Figure 6, Table S6), with the strongest effect for TT dimers. In H3K9me domains CC dimers had a strong effect, while the other pyrimidine dimers had no effect (Figure 6). In euchromatic domains there was no effect of pyrimidine dimers (Figure 6). These observations suggest that mutation rate is generally higher for adjacent thymines. This accounts for some of the increased mutation rate in H3K9 domains, due to their low GC-content. However, there are some additional mechanisms in operation in H3K9 domains, as adjacent cytosines have increased mutation rate.

Deletions, insertions, and translocations

In addition to point mutations, we also examined structural variants that happened in the MA-lines. We observed 97 deletions and 134 insertions in the MA-lines. Most common length for deletions and insertions were changes of one base pair, out of 96 1 bp indels, 88 occurred in homopolymer stretches. The second most common length was 3 bp, which were predominantly changes in microsatellite repeats. Some large deletions were observed: the largest deletion was

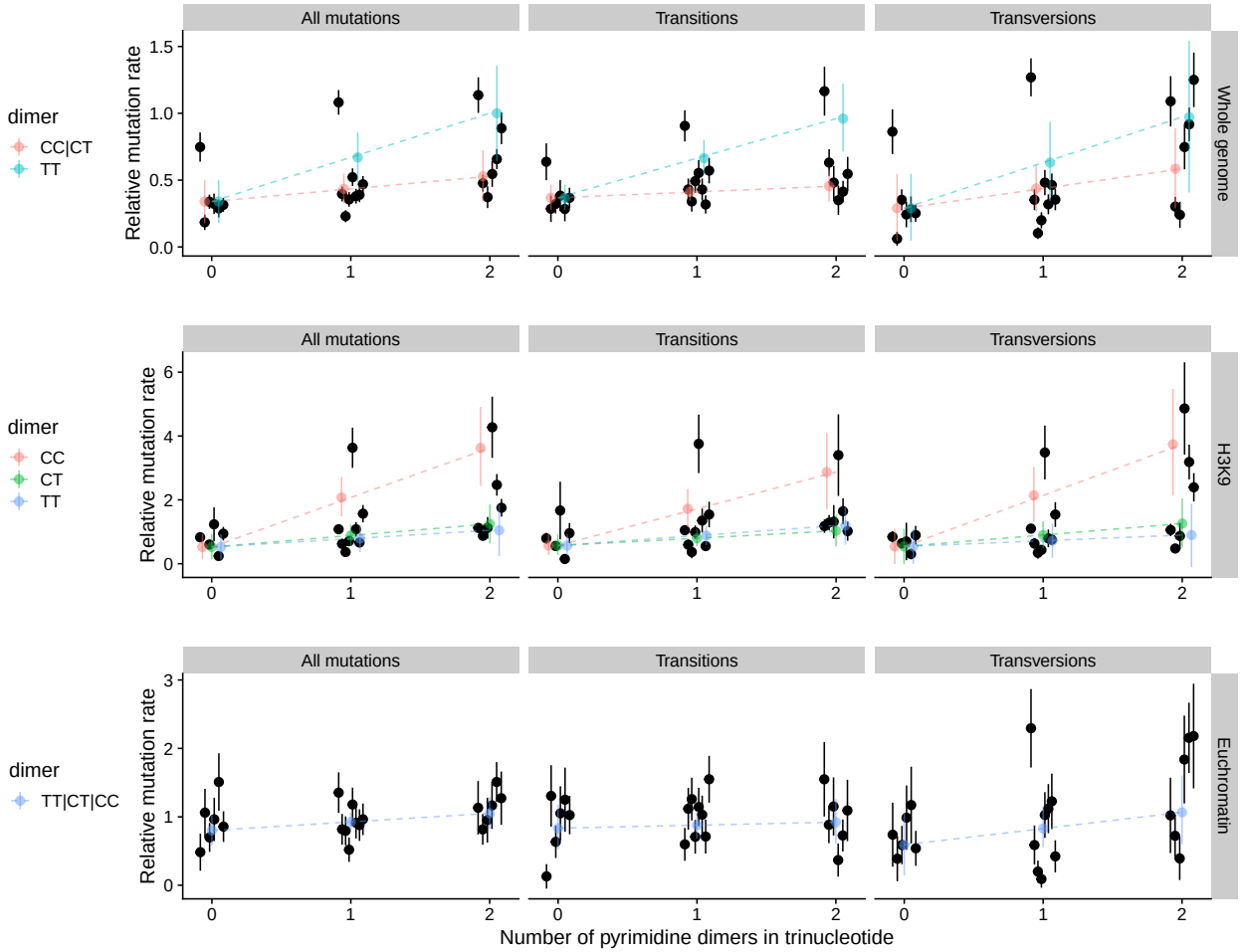

Figure 6: Relative mutation rate and number of pyrimidine dimers. Black points are the observed relative mutation rates for different triplets, blue, red, and green points are model estimates for TT and CC or CT dimers respectively.

17.8 kb, there were 3 deletions around 8.8 kb, and one around 2.5 kb. Otherwise most deletions were under 100 bp (Figure 7A). The largest observed insertion was 130 bp, with most insertions under 20 bp (Figure 7B).

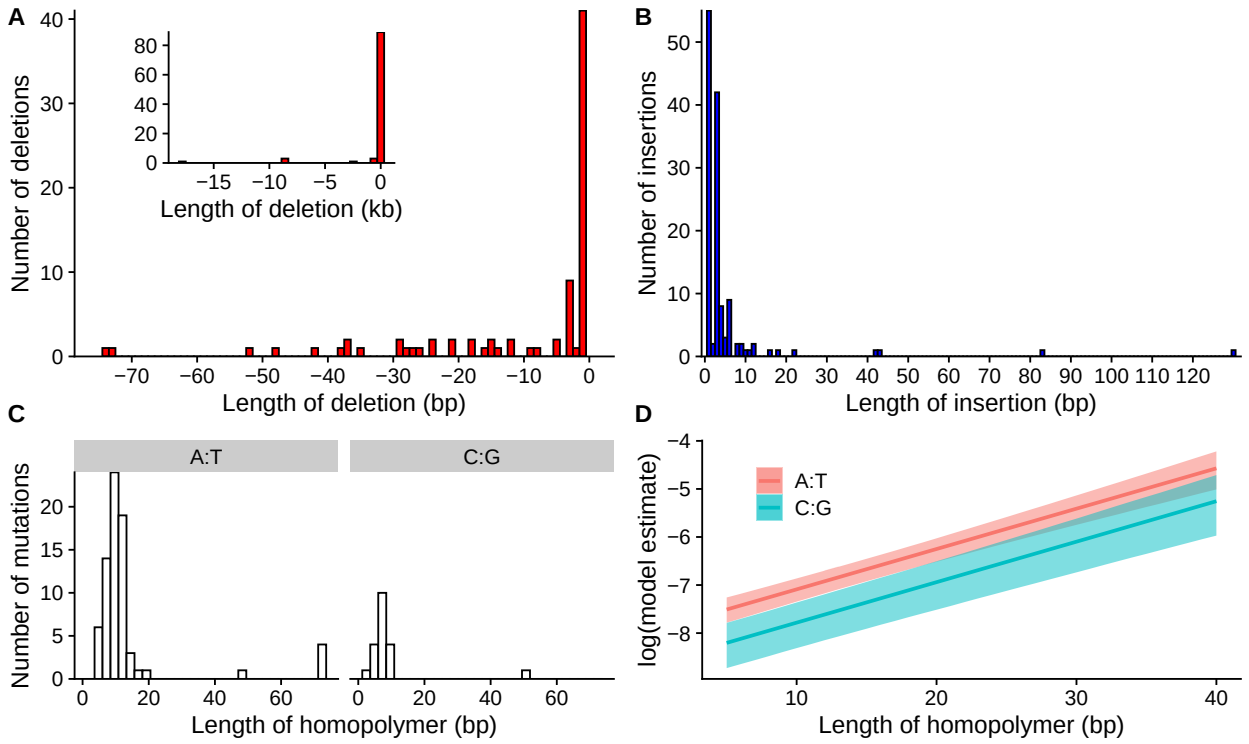

Figure 7: A) Distribution of deletion lengths in the range of 0 to 75 bp, inset shows the full distribution that includes few large deletions of several kb. B) Distribution of insertion lengths. C) Distribution of homopolymer lengths for those mutations that occurred in either A:T or C:G homopolymer stretches. D) Model estimates for mutation rate in homopolymers, A:T homopolymers had an overall higher mutation rate, and longer homopolymers had higher mutation rates.

Mutation rates for insertions and deletions are shown in table 1. When all deletions and insertions were included in the analysis, insertions had slightly higher rates than deletions (Table 1). However, when we excluded homopolymers, microsatellites, and other repeats from the analysis we observed that deletions had a three times higher mutation rate than insertions (Table 1). Given that mean length of deletion for mutations that do not occur in repeats was 1160 bp, which was much longer than the 27 bp mean length for insertions, there is mutation pressure to lose DNA. Some of the large deletions observed in our data are outliers, but deletions tended to be longer in the smaller scale mutations as well (Figure 7). When we only considered mutations occurring in

repeats, mutation rate of insertions was over two fold higher than deletions.

Table 1: Mutation rates (mutations / genome / mitosis) for deletions and insertions and their ratio.

	Deletions	Insertions	Insertion / deletion ratio
All indels	0.0024 [0.0020, 0.0029]	0.0034 [0.0028, 0.0039]	1.38 [1.06, 1.77]
Repeats excluded	0.0011 [0.0008, 0.0014]	0.0003 [0.0002, 0.0005]	0.3 [0.14, 0.51]
Repeats only	0.0014 [0.0010, 0.0017]	0.0030 [0.0025, 0.0036]	2.25 [1.6, 3.05]

Homopolymer stretches had particularly high rates indel mutations, we observed 92 mutations in homopolymers, most mutations in homopolymers were indels of 1 bp. More mutations occurred in A:T than in C:G homopolymers (Figure 7C). A:T homopolymers are approximately 1.7 times more common in the genome, but even when taking numbers of homopolymers into account, mutation rate in A:T homopolymers was 1.79 [1.4, 2.22] $\times 10^{-8}$ mutations / locus / mitosis compared to rate in C:G homopolymers of 8.15 [4.89, 11.86] $\times 10^{-9}$ mutations / locus / mitosis. Thus, mutations in A:T homopolymers were 2.2 [1.27, 3.49] times more common than in C:G polymers. We also observed that longer homopolymers had higher mutation rates. In a model with polymer length and polymer type, length had the same effect in for both A:T and C:G homopolymers (Figure 7D). This suggest that replication slippage, the mechanism suggested to be involved in indel mutations in repeats, tends to occur more often in longer repeats as expected. Detailed analysis of microsatellite mutations and population variation will be presented elsewhere.

We observed 5 translocations in the MA-lines, of which two were among the SVs confirmed by PCR and Sanger sequencing. Of these mutations, 3 were from one chromosome to another and two occurred among unmapped contigs. One translocation was 2813 bp and the others where length could be determined were around 300 bp. The translocation rate was 1.19 [0.33, 2.41] $\times 10^{-4}$ translocations / genome / mitosis. Translocations were much rarer than other insertions or deletions, thus we don't have enough mutations to do further analyses of their properties.

Complex mutations

We observed 9 cases were two SNMs or 1 bp indels occurred within few base pairs of each other in the same MA-line. While it is possible that two independent mutations happened next to each other,

this seems unlikely. Rather, these changes were more likely caused by a single mutational event. The observed complex mutational events are listed in table 2. To confirm whether these changes were caused by a single mutational event or two independent events, we used Sanger sequencing to check the genotypes of MA-lines from intermediate transfers stored during the experiment. In all cases that could be examined, we always observed that the two changes appeared together (Table 2). Therefore, we consider that the most parsimonious explanation is that these changes appeared as a result of single mutational events, likely caused by DNA repair error by an error prone DNA-polymerase.

Table 2: Complex mutations. Chr. is the chromosome or contig, position is the coordinate of the most 5' end change, dist. is the distance between nucleotide changes. Sequence column shows the ancestor sequence on first row and the MA-line sequence on second row, mutated bases are in bold. Final column contains the MA-experiment transfer numbers (mutations absent, mutations present) of those transfers when was the last transfer where none of the changes could be observed in the MA-line, and the transfer when all changes were observed in the MA-line. We always observed all changes in complex mutations to be present together. tv. = transversion, ts. = transition

MA-line	Chr.	Position	Dist. (bp)	Sequence	Changes	Mutation occurred between transfers
Line 27	1	1194971	2	GG-TGC GG**T**T**T**C	Insertion (T) and G → T tv.	(35, 38)
Line 31	1	4086493	1	GTAA G**AG**A	T → A tv. and A → G ts.	(10, 40)
Line 18	1	5852116	NA	GATACTATCGTAATATATCACTCTTGAACTAC G**TAGTTCAAGAGTAT----------**C	Deletion and insertion	(20, 25)
Line 21	1	9717311	3	CGAACT C**-**AA**A**T	Deletion (G) and C → A tv.	(15, 20)
Line 24	3	742257	1	AAAA A**TT**A	Two A → T tv.	(10, 15)
Line 35	3	1051825	2	ATTTG A**C**T**G**G	T → C ts. and T → G tv.	(15, 20)
Line 12	6	2848187	2	CTATA C**-**A**A**A	Deletion (T) and T → A tv.	not checked
Line 23	contig 8	176939	4	ATTTTTT A**C**TTT**C**T	Two T → C ts.	(7, 8)
Line 20	contig 9	121389	3	AATTTA A**T**TT**A**A	A → T tv. and T → A tv.	(30, 35)

We treated complex mutations as single events in calculations were all mutations were used to calculate overall mutation rates. The rate of complex mutations was 5.42 [2.22, 9.3] $\times 10^{-12}$ mutations / bp / mitosis. The rate of SNMs over rate of complex mutations was 123.7 [57.48, 231.49], making point mutations over 100-fold more common than complex events.

Discussion

We have estimated spontaneous mutation rate during asexual growth to high precision in *Neurospora crassa*. Our estimate of point mutation rate across the whole genome of 6.7 [6.32, 7.11] $\times 10^{-10}$ mutations / bp / mitosis is slightly higher, although quite close to an estimate of asexual mutation rate of 6.03×10^{-10} obtained by Wang et al. (2020). For mutations during asexual growth, the dataset of Wang et al. (2020) is considerably smaller than ours, with only 64 mutations, and they estimated the number of mitoses that happened only indirectly, in contrast to our study. Nevertheless, their estimate is the same order of magnitude as ours. Previous combined estimate of mutation rate from marker gene studies suggested that mutation rate is 4.10–4.66×10^{-9} (Lynch et al., 2016) but neither our results nor the results of Wang et al. (2020) agree with this. When considering only point mutations that happen in euchromatic regions, we obtained a result of 0.007 [0.006, 0.008] mutations / genome / mitosis, which is in line with results already obtained by Drake (1991) who observed that mutation rate per genome for microbes seems to be around 0.003 mutations per genome per generation with approximately two fold variation around this mean. Estimate of Drake (1991) was based on marker genes, that were in euchromatin, so comparing mutations in euchromatin seems reasonable. Thus asexual mutation rate in *N. crassa* in euchromatic regions seems to be rather typical for a microbe.

What makes *N. crassa* peculiar in terms of mutation rate, is that there is a striking difference in rate and spectrum of mutation during sexual and asexual reproduction (Wang et al., 2020). During sexual reproduction a genome defence mechanism called RIP is activated which recognizes duplicated regions during the pairing of homologous chromosomes and C:G $\rightarrow$ T:A transitions

are induced in those regions (Gladyshev et al., 2017). Mutations induced during sexual reproduction happen mainly in these duplicated regions (Wang et al., 2020), and these duplicated regions nearly completely overlap with H3K9 methylated domains. In contrast, during asexual reproduction, while H3K9me domains have a higher mutation rate, the bias is much smaller and the spectrum of mutations very different.

While many studies have reported effects on chromatin structure on mutation rate, these have often been based on indirect inference from species divergence and polymorphism (Sasaki et al., 2009; Ying et al., 2010; Washietl et al., 2008). Our study adds to the growing body of direct observations that chromatin structure is an important determinant of spontaneous mutation rate. We observed extensive variation across the *N. crassa* genome in mutation rate and mutation spectra due to chromatin modifications. Mutation rate for SNMs in domains marked by H3K9 methylation that determines heterochromatic regions in *N. crassa* was tenfold higher than in euchromatic regions. In *N. crassa* centromeric regions determined by the presence of centromeric histone variant CenH3 always overlap with H3K9 methylation (Smith et al., 2011). In centromeric regions there was an additional effect of increased mutation rate on top of the H3K9 effect. Increased mutation rate in heterochromatic and centromeric regions has also been found by Weng et al. (2019) in the plant *Arabidopsis thaliana*. These results are also in line with observations from human cancer cells that have higher rates of mutation in heterochromatic regions (Schuster-Böckler and Lehner, 2012; Polak et al., 2015). In *N. crassa* facultative heterochromatin is marked by H3K27 trimethylation (Jamieson et al., 2013), and we observed that mutation rate was slightly elevated in H3K27me3 domains, although this effect was much smaller than for H3K9 methylation.

In *N. crassa* methylation in H3K9 and H3K36 are almost completely mutually exclusive, and other mark occurs where the other does not. However, H3K36 is not a straightforward mark of euchromatin as it can be deposited by two enzymes: SET-2 and ASH1. Genes marked with H3K36 methylation by SET-2 are actively transcribed, while genes marked with H3K36 methylation by ASH1 are silenced and can be further marked by H3K27 methylation (Bicocca et al., 2018). SET-2 is responsible for most of H3K36 methylation, and in our study we considered all regions that

lacked H3K9 and H3K27 methylation to be euchromatin. Monroe et al. (2022) observed that in *Arabidopsis* mutation rate was affected by several different epigenetic marks, and they suggested that mutation rate was lower in genes that were actively transcribed and perhaps even fine tuned for highly expressed genes, possibly due to the presence of H3K36 methylation or other epigenetic marks of active transcription. While we also observed that H3K36 domains (absence of H3K9) had a lower mutation rate compared to heterochromatin, so in that sense genes also had lower mutation rate. However, our dataset does not contain enough mutations to address if gene expression levels quantitatively influence mutation rate in *N. crassa*.

Given that we have observed variation in mutation rate across the genome, what are the mechanism behind this phenomenon? In *N. crassa* domains with H3K9 methylation and centromeric regions have much lower GC-content than euchromatic domains, and GC-content has a strong effect on mutation rate. While the effect of GC-content is seen in both in euchromatin and heterochromatic domains and these effects can be statistically separated, effect of CG-content in the total increase in mutation rate in heterochromatic regions seems substantial. However, we cannot determine whether increased mutation rate in H3K9 regions, especially during the sexual cycle, has led to lower GC-content over evolutionary time, or whether these regions are heterochromatic because of their low GC-content.

Chromatin structure is involved in DNA lesion and repair, and this may explain why heterochromatic regions have a higher mutation rate. In yeast, actively transcribed regions contain acetylation at H3K56 which suppresses spontaneous mutations (Kadyrova et al., 2013). Moreover, analysis based on human tumours suggest that DNA mismatch repair works more efficiently in euchromatin than in heterochromatin and it is DNA repair that is variable, not the supply of mutations themselves (Supek and Lehner, 2015). Further evidence supporting this explanation was gathered by a study performed by Yazdi et al. (2015) in which the authors observed that, when using DNA repair machinery knockout yeast strains, mutation rate did not increase in high nucleosome occupancy regions like in wild type strains. Differential exposure of heterochromatin to natural mutagens, such as oxidative damage, does not seem to explain our results. Mutations typically associated with

oxidative damage: G → C transversions, C → T transitions, and G → T transversions (McBride et al., 1991; Cheng et al., 1992), were not systematically overrepresented in regions with H3K9 methylation, only the relative amount of G → C transversions differed between euchromatic and H3K9 methylated domains.

Among the mutations that were observed, sequence context did have large effect on mutation rate. It is known that local sequence context can influence the probability of mutation and this phenomenon has been frequently observed (Ness et al., 2015). However, the mechanisms of why some sequence contexts are prone to mutation are less clear. For example, the thymine-thymine cyclobutane dimer, a form of DNA damage caused by sunlight, formation rates are dependent on flanking sequence (Law et al., 2013), but the biophysical basis is unclear (Law et al., 2013). We did observe that TT-dimers had high relative mutation rate across the whole genome, but since this effect was not observed within heterochromatin or euchromatin this seems an effect of the high mutation rate in heterochromatin and not a signature of mutations that were caused by UV-induced DNA damage in our experiment. AT-rich trinucleotides in general had higher mutation rates, and like Law et al. (2013) in the context of TT-dimers, we also observed that flanking G's supressed mutation rates in trinucleotides otherwise TT.

In regions of H3K9 methylation we observed a striking effect of increased relative mutation rate in trinucleotides composed of entirely C's and G's, but not in CGC:GCG trinucleotides. One hypothesis for this would be that increased mutation rate in these sequences could be due to the presence of DNA methylation. In *N. crassa* DNA methylation occurs only at H3K9 methylated domains (Tamaru and Selker, 2001), and deamination of 5-methylcytosine is a known cause of mutations (Cooper et al., 2010). However, deamination should cause mainly C → T transitions while we observed that both transitions and transversions are elevated in these trinucleotides and no excess of C → T transitions were observed in H3K9 methylated domains. Therefore DNA methylation seems to contribute quite little to spontaneous genetic mutations in *N. crassa*, possibly because DNA methylation is so rare in this species (Hosseini et al., 2020). Alternatively, it could be that these trinucleotides are susceptible to oxidation as GG dimers have been found to be more

susceptible to oxidation (Hanrahan et al., 1997). Nevertheless, we cannot conclusively explain this pattern.

For spontaneous insertions and deletions, we observed that when repeated sequences, which were mainly homopolymers and microsatellite sequences, were excluded, deletions were more common than insertions. Since deletions also tended to be longer, mutations have the tendency to reduce the size of the genome. Similar patterns in mutational bias have been observed for *Drosophila melanogaster* (Leushkin et al., 2013). For repeated sequences, we analyzed homopolymer sequences in more detail. We found that while A:T homopolymers are more common in *N. crassa* they also have a higher mutation rate. This is in contrast to observations in nematodes, where C:G homopolymers had much higher mutation rates than A:T homopolymers (Denver et al., 2004) suggesting that there are species specific differences in mutation rates of these sequences.

While the mutation rate during asexual reproduction is comparable to other eukaryotes, the spectrum of single nucleotide mutations in *N. crassa* is different from yeast, in *N. crassa* in euchromatic regions both transition occur at nearly equal frequencies, where as in *Saccharomyces cerevisiae* C:G $\rightarrow$ T:A transtions are much more common (Zhu et al., 2014), as is also the case in the algae *Chlamydomonas reinhardtii* (Ness et al., 2015) and crustacean *Daphnia* (Keith et al., 2016). It appears that DNA repair machinery of *N. crassa* has somewhat different properties than other microbial eukaryotes, in particular that A:T $\rightarrow$ C:G transversions are common than in other species. Alternatively, it may be that C:G $\rightarrow$ T:A transitions are more rare in *N. crassa*.

We identified mutations with multiple base changes due to a single mutational event. Such complex mutations are thought to arise from the action of error-prone translesion DNA polymerases, such as Pol ζ (Stone et al., 2012). A recent large scale survey of human trios identified many multinucleotide mutations, and observed that mutations that were 2–10 bp apart showed an overrepresentation of A:T $\rightarrow$ T:A, and A:T $\rightarrow$ G:C mutations (Besenbacher et al., 2016). While we did not observe enough complex mutations to conduct a statistical analysis, our results are compatible with this pattern as most nucleotide changes in complex mutations were either A:T $\rightarrow$ T:A or A:T $\rightarrow$ G:C. However, higher mutation rate in AT-rich regions could also be important. We

did not observe GC $\rightarrow$ AA or GA $\rightarrow$ TT tandem mutations, that are the most common tandem mutation in humans (Harris and Nielsen, 2014; Besenbacher et al., 2016). In *N. crassa* complex mutations represented only 0.7% of observed mutations, which is smaller than approximately 3% observed in humans, thus it may be that the impact of multinucleotide mutations on population genetic inference may be smaller in than in humans.

Conclusions

We have observed variation in mutation rate and mutation spectrum across the genome of *N. crassa*. It appears that variation in mutation rate in eukaryotes due to chromatin modifications will be the norm rather than the exception. This raises an issue that there is no single mutation rate for a given species, as chromatin state has such a large effect. In the future it will be more informative to speak about mutation rates specific to different chromatin domains. The implication for evolutionary genetics is that variation in mutation rates must be taken into account in order to quantify which evolutionary forces act on the genome. Genes residing in different regions of the genome can differ substantially in their mutation loads, and it seems that variation in mutation rate does also have a large part in determining how much genetic variation is segregating in a given regions of the genome. To accurately estimate imprints of selection on natural genetic variation, we must account for variation in mutation rates in evolutionary models. Therefore, detailed models of mutation rates are needed for different species. This study is the first step in order to do this in *N. crassa*, a model microbial eukaryote.

Acknowledgments

This study was supported by a grant from the Academy of Finland (no. 321584) to IK. We thank Dr. Visa Ruokolainen for help with confocal microscopy, and the Finnish CSC-IT Center for Science Ltd. for providing computational resources.

Data availability

Sequencing data generated in this study has been deposited to the short read archive, project number PRJNA839531, with sequence accession numbers SRX15345434–SRX15345507. Data and scripts will be available in Dryad XXXXX.

References

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein, 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21:974–984. URL http://genome.cshlp.org/content/21/6/974.abstract.

Beier, S., T. Thiel, T. Münch, U. Scholz, and M. Mascher, 2017. MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583–2585. URL https://doi.org/10.1093/bioinformatics/btx198.

Besenbacher, S., P. Sulem, A. Helgason, H. Helgason, H. Kristjansson, A. Jonasdottir, A. Jonasdottir, O. T. Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, D. F. Gudbjartsson, and K. Stefansson, 2016. Multi-nucleotide *de novo* mutations in humans. PLOS Genetics 12:1–15. URL https://doi.org/10.1371/journal.pgen.1006315.

Bicocca, V. T., T. Ormsby, K. K. Adhvaryu, S. Honda, and E. U. Selker, 2018. ASH1-catalyzed H3K36 methylation drives gene repression and marks H3K27me2/3-competent chromatin. eLife 7:e41497. URL https://doi.org/10.7554/eLife.41497.

Bürkner, P.-C., 2017. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80:1–28. URL https://www.jstatsoft.org/v080/i01.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo,

P. Li, and A. Riddell, 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76:1–32. URL https://www.jstatsoft.org/v076/i01.

Chen, X., Z. Chen, H. Chen, Z. Su, J. Yang, F. Lin, S. Shi, and X. He, 2012. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science 335:1235–1238. URL https://www.science.org/doi/abs/10.1126/science.1217580.

Cheng, K. C., D. S. Cahill, H. Kasai, S. Nishimura, and L. A. Loeb, 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G-T and A-C substitutions. Journal of Biological Chemistry 267:166–172. URL https://www.sciencedirect.com/science/article/pii/S0021925818484748.

Chiang, C., R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan, and I. M. Hall, 2015. Speedseq: ultra-fast personal genome analysis and interpretation. Nature Methods 12:966–968. URL https://doi.org/10.1038/nmeth.3505.

Cingolani, P., A. Platts, M. Coon, T. Nguyen, L. Wang, S. Land, X. Lu, and D. Ruden, 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6:80–92.

Cooper, D. N., M. Mort, P. D. Stenson, E. V. Ball, and N. A. Chuzhanova, 2010. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Human Genomics 4:406. URL https://doi.org/10.1186/1479-7364-4-6-406.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li, 2021. Twelve years of SAMtools and BCFtools. GigaScience 10. URL https://doi.org/10.1093/gigascience/giab008. Giab008.

Davis, R. H. and F. J. de Serres, 1970. Genetic and microbiological research techniques for *Neurospora crassa*. Methods in Enzymology 17:79–143.

Denver, D. R., K. Morris, A. Kewalramani, K. E. Harris, A. Chow, S. Estes, M. Lynch, and W. K. Thomas, 2004. Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of *Caenorhabditis elegans*. Journal of Molecular Evolution 58:584–595. URL https://doi.org/10.1007/s00239-004-2580-4.

Drake, J. W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proceedings of the National Academy of Sciences 88:7160–7164. URL https://www.pnas.org/doi/abs/10.1073/pnas.88.16.7160.

Faust, G. G. and I. M. Hall, 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30:2503–2505. URL https://doi.org/10.1093/bioinformatics/btu314.

Ferretti, L., E. Raineri, and S. Ramos-Onsins, 2012. Neutrality tests for sequences with missing data. Genetics 191:1397–1401. URL https://doi.org/10.1534/genetics.112.139949.

Freitag, M., P. C. Hickey, N. B. Raju, E. U. Selker, and N. D. Read, 2004. GFP as a tool to analyze the organization, dynamics and function of nuclei and microtubules in *Neurospora crassa*. Fungal Genetics and Biology 41:897–910. URL http://www.sciencedirect.com/science/article/pii/S1087184504000957.

Gelman, A., B. Goodrich, J. Gabry, and A. Vehtari, 2019. R-squared for Bayesian regression models. The American Statistician 73:307–309. URL https://doi.org/10.1080/00031305.2018.1549100.

Gladyshev, E., J. Heitman, and E. H. Stukenbrock, 2017. Repeat-induced point mutation and other genome defense mechanisms in fungi. Microbiology Spectrum 5:5.4.02. URL https://journals.asm.org/doi/abs/10.1128/microbiolspec.FUNK-0042-2017.

Halligan, D. L. and P. D. Keightley, 2009. Spontaneous mutation accumulation studies in evolu-

tionary genetics. Annual Review of Ecology, Evolution, and Systematics 40:151–172. URL http://dx.doi.org/10.1146/annurev.ecolsys.39.110707.173437.

Hanrahan, C. J., M. D. Bacolod, R. R. Vyas, T. Liu, N. E. Geacintov, E. L. Loechler, and A. K. Basu, 1997. Sequence specific mutagenesis of the major (+)-anti-benzo[a]pyrene diol epoxide–DNA adduct at a mutational hot spot *in vitro* and in *Escherichia coli* cells. Chem. Res. Toxicol. 10:369–377. URL https://doi.org/10.1021/tx9601925.

Harris, K. and R. Nielsen, 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Research 24:1445–1454. URL http://genome.cshlp.org/content/24/9/1445.abstract.

Hosseini, S., C. Meunier, D. Nguyen, J. Reimegård, and H. Johannesson, 2020. Comparative analysis of genome-wide dna methylation in *Neurospora*. Epigenetics 15:972–987. URL https://doi.org/10.1080/15592294.2020.1741758. PMID: 32228351.

Jamieson, K., M. R. Rountree, Z. A. Lewis, J. E. Stajich, and E. U. Selker, 2013. Regional control of histone H3 lysine 27 methylation in *Neurospora*. Proceedings of the National Academy of Sciences 110:6027–6032. URL http://www.pnas.org/content/110/15/6027.abstract.

Johri, P., B. Charlesworth, and J. D. Jensen, 2020. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics 215:173–192. URL https://doi.org/10.1534/genetics.119.303002.

Kadyrova, L. Y., T. M. Mertz, Y. Zhang, M. R. Northam, Z. Sheng, K. S. Lobachev, P. V. Shcherbakova, and F. A. Kadyrov, 2013. A reversible histone h3 acetylation cooperates with mismatch repair and replicative polymerases in maintaining genome stability. PLOS Genetics 9:1–16. URL https://doi.org/10.1371/journal.pgen.1003899.

Katju, V. and U. Bergthorsson, 2018. Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with

high-throughput genomic approaches. Genome Biology and Evolution 11:136–165. URL https://doi.org/10.1093/gbe/evy252.

Keightley, P. D., R. W. Ness, D. L. Halligan, and P. R. Haddrill, 2014. Estimation of the spontaneous mutation rate per nucleotide site in a drosophila melanogaster full-sib family. Genetics 196:313–320. URL http://www.genetics.org/content/196/1/313.abstract.

Keightley, P. D., A. Pinharanda, R. W. Ness, F. Simpson, K. K. Dasmahapatra, J. Mallet, J. W. Davey, and C. D. Jiggins, 2015. Estimation of the spontaneous mutation rate in heliconius melpomene. Molecular Biology and Evolution 32:239–243. URL http://mbe.oxfordjournals.org/content/32/1/239.abstract.

Keith, N., A. E. Tucker, C. E. Jackson, W. Sung, J. I. Lucas Lledó, D. R. Schrider, S. Schaack, J. L. Dudycha, M. Ackerman, A. J. Younge, J. R. Shaw, and M. Lynch, 2016. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome Research 26:60–69. URL http://genome.cshlp.org/content/26/1/60.abstract.

Kelleher, J., R. W. Ness, and D. L. Halligan, 2013. Processing genome scale tabular data with wormtable. BMC Bioinformatics 14:1–5. URL http://dx.doi.org/10.1186/1471-2105-14-356.

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani, 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology 20:117. URL https://doi.org/10.1186/s13059-019-1720-5.

Kronholm, I., A. Bassett, D. Baulcombe, and S. Collins, 2017. Epigenetic and genetic contributions to adaptation in *Chlamydomonas*. Molecular Biology and Evolution 34:2285–2306. URL + http://dx.doi.org/10.1093/molbev/msx166.

Kronholm, I., T. Ormsby, K. J. McNaught, E. U. Selker, and T. Ketola, 2020. Marked *Neurospora crassa* strains for competition experiments and Bayesian methods for fitness estimates. G3: Genes|Genomes|Genetics 10:1261–1270.

Law, Y. K., R. A. Forties, X. Liu, M. G. Poirier, and B. Kohler, 2013. Sequence-dependent thymine dimer formation and photoreversal rates in double-stranded DNA. Photochem. Photobiol. Sci. 12:1431–1439. URL http://dx.doi.org/10.1039/C3PP50078K.

Leushkin, E. V., G. A. Bazykin, and A. S. Kondrashov, 2013. Strong mutational bias toward deletions in the *Drosophila melanogaster* genome is compensated by selection. Genome Biology and Evolution 5:514–524. URL https://doi.org/10.1093/gbe/evt021.

Li, C. and N. M. Luscombe, 2020. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. Nature Communications 11:1363–. URL https://doi.org/10.1038/s41467-020-15185-0.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] .

———, 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30:2843–2851. URL https://doi.org/10.1093/bioinformatics/btu356.

Lichius, A. and S. Zeilinger, 2019. Application of membrane and cell wall selective fluorescent dyes for live-cell imaging of filamentous fungi. JoVE P. e60613. URL https://www.jove.com/t/60613.

Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster, 2016. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17:704–714. URL http://dx.doi.org/10.1038/nrg.2016.104.

Makova, K. D. and R. C. Hardison, 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nature Reviews Genetics 16:213–223. URL https://doi.org/10.1038/nrg3890.

McBride, T. J., B. D. Preston, and L. A. Loeb, 1991. Mutagenic spectrum resulting from DNA damage by oxygen radicals. Biochemistry 30:207–213. URL https://doi.org/10.1021/bi00215a030.

McCluskey, K., A. Wiest, and M. Plamann, 2010. The fungal genetics stock center: a repository for 50 years of fungal genetics research. J Biosci 35:119–126.

McElreath, R., 2015. Statistical Rethinking - A Bayesian course with examples in R and Stan. CRC Press, New York.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303. URL http://dx.doi.org/10.1101/gr.107524.110.

Metzenberg, R. L., 2003. Vogel's medium N salts: Avoiding the need for ammonium nitrate. Fungal Genetics Newsletter 50:14.

Monroe, J. G., T. Srikant, P. Carbonell-Bejerano, C. Becker, M. Lensink, M. Exposito-Alonso, M. Klein, J. Hildebrandt, M. Neumann, D. Kliebenstein, M.-L. Weng, E. Imbert, J. Ågren, M. T. Rutter, C. B. Fenster, and D. Weigel, 2022. Mutation bias reflects natural selection in *Arabidopsis thaliana*. Nature Pp. –. URL https://doi.org/10.1038/s41586-021-04269-6.

Ness, R. W., A. D. Morgan, N. Colegrave, and P. D. Keightley, 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. Genetics 192:1447–1454.

Ness, R. W., A. D. Morgan, R. B. Vasanthakrishnan, N. Colegrave, and P. D. Keightley, 2015. Extensive de novo mutation rate variation between individuals and across the genome of chlamydomonas reinhardtii. Genome Research URL http://genome.cshlp.org/content/early/2015/08/10/gr.191494.115.abstract.

Ossowski, S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, and M. Lynch, 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92–94. URL http://dx.doi.org/10.1126/science.1180677.

Polak, P., R. Karlić, A. Koren, R. Thurman, R. Sandstrom, M. S. Lawrence, A. Reynolds, E. Rynes, K. Vlahoviček, J. A. Stamatoyannopoulos, and S. R. Sunyaev, 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518:360–364. URL https://doi.org/10.1038/nature14221.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339. URL https://doi.org/10.1093/bioinformatics/bts378.

Rochette, P. J., J. Therrien, R. Drouin, D. Perdiz, N. Bastien, E. A. Drobetsky, and E. Sage, 2003. UVA-induced cyclobutane pyrimidine dimers form predominantly at thymine–thymine dipyrimidines and correlate with the mutation spectrum in rodent cells. Nucleic Acids Research 31:2786–2794. URL https://doi.org/10.1093/nar/gkg402.

Rueden, C. T., J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri, 2017. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics 18:529. URL https://doi.org/10.1186/s12859-017-1934-z.

Sasaki, S., C. C. Mello, A. Shimada, Y. Nakatani, S.-i. Hashimoto, M. Ogawa, K. Matsushima, S. G. Gu, M. Kasahara, B. Ahsan, et al., 2009. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 323:401–404.

Schuster-Böckler, B. and B. Lehner, 2012. Chromatin organization is a major influ-

ence on regional mutation rates in human cancer cells. Nature 488:504–507. URL https://doi.org/10.1038/nature11273.

Smith, K. M., P. A. Phatale, C. M. Sullivan, K. R. Pomraning, and M. Freitag, 2011. Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. Molecular and Cellular Biology 31:2528–2542. URL https://journals.asm.org/doi/abs/10.1128/MCB.01285-10.

Song, Q. and A. D. Smith, 2011. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics 27:870–871. URL https://doi.org/10.1093/bioinformatics/btr030.

Stone, J. E., S. A. Lujan, and T. A. Kunkel, 2012. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in *Saccharomyces cerevisiae*. Environmental and Molecular Mutagenesis 53:777–786. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/em.21728.

Sung, W., M. S. Ackerman, J.-F. Gout, S. F. Miller, E. Williams, P. L. Foster, and M. Lynch, 2015. Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. Molecular Biology and Evolution 32:1672–1683. URL https://doi.org/10.1093/molbev/msv055.

Supek, F. and B. Lehner, 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature 521:81–84. URL https://doi.org/10.1038/nature14173.

Tamaru, H. and E. U. Selker, 2001. A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. Nature 414:277–283. URL http://dx.doi.org/10.1038/35104508.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualiza-

tion and exploration. Briefings in Bioinformatics 14:178–192. URL http://bib.oxfordjournals.org/content/14/2/178.abstract.

Tolstorukov, M. Y., N. Volfovsky, R. M. Stephens, and P. J. Park, 2011. Impact of chromatin structure on sequence variability in the human genome. Nature Structural & Molecular Biology 18:510–515. URL https://doi.org/10.1038/nsmb.2012.

Vehtari, A., A. Gelman, and J. Gabry, 2017. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27:1413–1432. URL https://doi.org/10.1007/s11222-016-9696-4.

Wang, L., Y. Sun, X. Sun, L. Yu, L. Xue, Z. He, J. Huang, D. Tian, L. D. Hurst, and S. Yang, 2020. Repeat-induced point mutation in *Neurospora crassa* causes the highest known mutation rate and mutational burden of any cellular life. Genome Biology 21:142. URL https://doi.org/10.1186/s13059-020-02060-w.

Washietl, S., R. Machné, and N. Goldman, 2008. Evolutionary footprints of nucleosome positions in yeast. Trends in Genetics 24:583–587.

Weng, M.-L., C. Becker, J. Hildebrandt, M. Neumann, M. T. Rutter, R. G. Shaw, D. Weigel, and C. B. Fenster, 2019. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. Genetics 211:703–714. URL https://www.genetics.org/content/211/2/703.

Xie, C. and M. T. Tammi, 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10:80. URL https://doi.org/10.1186/1471-2105-10-80.

Yazdi, P. G., B. A. Pedersen, J. F. Taylor, O. S. Khattab, Y.-H. Chen, Y. Chen, S. E. Jacobsen, and P. H. Wang, 2015. Increasing nucleosome occupancy is correlated with an increasing mutation rate so long as DNA repair machinery is intact. PLoS One 10:e0136574.

Ying, H., J. Epps, R. Williams, and G. Huttley, 2010. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. Molecular biology and evolution 27:637–649.

Zhao, J., P. Gladieux, E. Hutchison, J. Bueche, C. Hall, F. Perraudeau, and N. L. Glass, 2015. Identification of allorecognition loci in *Neurospora crassa* by genomics and evolutionary approaches. Molecular Biology and Evolution 32:2417–2432. URL http://mbe.oxfordjournals.org/content/32/9/2417.abstract.

Zhu, Y. O., M. L. Siegal, D. W. Hall, and D. A. Petrov, 2014. Precise estimates of mutation rate and spectrum in yeast. Proceedings of the National Academy of Sciences 111:E2310–E2318. URL https://www.pnas.org/content/111/22/E2310.

Supplementary Information

Supplemental information for article: Chromatin structure influences rate and spectrum of spontaneous mutations in *Neurospora crassa*. By Mariana Villalba de la Peña, Pauliina A. M. Summanen, Martta Liukkonen, Ilkka Kronholm.

Supplementary methods

Estimating the number of mitoses

To count the number of mitoses that happened in the MA-lines during the experiment we needed to know how many nuclei were present in the colonies on sorbose plates that were picked and transferred to slants, and how many nuclei were in the mycelium that formed in the test tube, and finally how many nuclei were in the conidia that formed in the test tube (Figure 1B). This allows us to calculate how many mitoses happened during one transfer of the MA-experiment, from one spore to a spore. There are multiple sources of error in these calculations, so we used a Bayesian framework to do the calculations using posterior distributions of the estimates to incorporate all sources of error in the final estimate.

Nuclei were counted from the microscope images using Fiji2 version 2.0.0-rc-54/1.51g (Schindelin et al., 2012). Short sections of mycelium were surrounded with the rectangular selection tool and the area inside was measured. All nuclei with more than 50% of their diameter inside the selection were counted manually. Multiple sections were counted from each image, with no overlap. In some fainter images, the contrast was enhanced with the enhance contrast tool, with the default value 0.3% saturated pixels and no histogram equalization. To estimate the number of nuclei in a given area of hyphae, we used the counts of nuclei and hyphal areas calculated from the microscope images to obtain counts of number of nuclei per μm^2. We had images for both VM and sorbose plates, in total we collected 519 measurements. To estimate average density of nuclei for VM and sorbose we used a model where we allowed standard deviations to differ for VM and

sorbose media:

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \sigma_i) \\
\mu_i &= \rho + \beta_s x_i \\
\log(\sigma_i) &= \alpha_\sigma + \beta_\sigma x_i \\
\rho, \beta_s &\sim \mathrm{N}(0, 0.1) \\
\alpha_\sigma, \beta_\sigma &\sim \mathrm{hT}(3, 0, 10)
\end{aligned}
\tag{S1}
$$

where y_i is the ith density measurement, ρ is the intercept, β_s is the effect of sorbose medium, x_i an indicator variable for sorbose, α_σ is the intercept for standard deviation, and β_σ is the effect of sorbose medium on standard deviation. Average density of nuclei in VM medium is ρ and density of nuclei in sorbose is obtained as $\rho_s = \rho + \beta_s$.

To estimate the average size of colonies on sorbose plates, we plated conidia on sorbose plates as in the MA-experiment and photographed the plates. Millimeter paper was used as a scale. Colony area was measured from these images with ImageJ2 version 2.0.0-rc-43/1.50e. The pixels per millimeter calibration value was set by measuring the number of pixels per $1\ \mathrm{mm}$ of millimeter paper. The images were enhanced with the sharpen tool to make the colony outlines more distinct. The colony area was measured using the elliptical selection tool. We used 10 different genotypes from different MA-lines and timepoints in this experiment, including the 2 ancestors. We collected a dataset with 482 area measurements. To estimate average colony size, we fitted a multilevel model

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \sigma) \\
\mu_i &= \alpha_{g[i]} \\
\alpha_g &\sim \mathrm{N}(\bar{\alpha}, \sigma_g) \\
\bar{\alpha} &\sim \mathrm{N}(0, 3) \\
\sigma, \sigma_g &\sim \mathrm{hT}(3, 0, 10)
\end{aligned}
\tag{S2}
$$

where y_i is the ith area measurement, $\bar{\alpha}$ is the overall mean, α_j is the mean for jth genotype, σ_g is the genotype standard deviation, and σ is the error standard deviation. Standard deviations had a weakly informative prior, which was the half-location scale version of Student's t-distribution, where 3 is the degrees of freedom, 0 is the location, and 10 is the scale parameter. We estimated the number of nuclei in a sorbose plate colony, n_s, as

$$
n_s = \bar{\alpha} \times 10^6 \times \rho_s \tag{S3}
$$

the average colony size is multiplied by 10^6 to transform the unit from mm^2 to $\mu\mathrm{m}^2$.

Once the sorbose colony is transferred to the test tube, the mycelium will cover the surface of the growth media. We estimated the number of nuclei present in the mycelium, n_v by multiplying the surface of the media in the test tube with the density of nuclei in the hyphae in VM medium:

$$
n_v = \pi(d/2)^2 \times 10^6 \times \rho \tag{S4}
$$

where d is the diameter of the test tubes used in the experiment, area is multiplied by 10^6 to transform the unit to $\mu\mathrm{m}^2$.

To estimate the number of conidia produced by the mycelium in the test tube, we counted conidia by suspending them in 1 ml of 0.01% Tween-80, making a 10000-fold dilution of the

suspension, and plating 10 µl of the dilution on sorbose plates. We counted the colonies that were formed, and estimated the original number of conidia produced. We used 10 different genotypes, including the ancestors from the MA-experiment to estimate produced conidia. We collected 71 measurements, the model was

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \sigma) \\
\mu_i &= \nu_{g[i]} \\
\nu_g &\sim \mathrm{N}(\bar{\nu}, \sigma_g) \\
\bar{\nu} &\sim \mathrm{hT}(3, 40, 21) \\
\sigma, \sigma_g &\sim \mathrm{hT}(3, 0, 21)
\end{aligned}
\tag{S5}
$$

where y_i is the ith conidial number measurement, $\bar{\nu}$ is the overall mean, α_j is the mean for jth genotype, σ_g is the genotype standard deviation, and σ is the error standard deviation. Priors followed Student's t-distribution. The number of nuclei contained by the conidia, n_c was estimated as

$$n_c = 2\bar{\nu} \tag{S6}$$

since the mode of nuclei in conidia of *N. crassa* is two.

Number of mitotic divisions separating two time points is the base 2 logarithm of final number of nuclei over initial number of nuclei. Thus, using the posterior distributions of numbers of nuclei in the different phases of the transfer, we can calculate the number of mitoses that happen during a transfer, m, as:

$$m = \log_2\left(\frac{n_s}{2}\right) + \log_2\left(\frac{n_s + n_v}{n_s}\right) + \log_2\left(\frac{n_c + n_v + n_s}{n_v + n_s}\right) \tag{S7}$$

this estimate incorporates all sources of measurement error since posterior distributions are used in every step of the calculations.

Performance of SV callers on simulated data

To evaluate the performance of different SV callers, we simulated 22, 39, 62 and 160 SVs with mutation rate of 0.001 on the fasta file of the reference genome (NC12) using SURVIVOR version 1.0.7 (Jeffares et al., 2017). Based on the modified fasta we created 150 bp paired end reads with an error rate of 0.03% and a mean coverage of 30X using DWGSIM version 0.1.11 (Homer, 2021). Simulated reads were then aligned to the reference genome using BWA-MEM (Li, 2013) with default parameters, and SVs were called using DELLY version 0.8.7, LUMPY version 0.2.13, PINDEL version 0.2.5b9, SVaba version 1.1.0 (Rausch et al., 2012; Layer et al., 2014; Ye et al., 2009; Wala et al., 2018). We used the eval function of SURVIVOR to evaluate the performance of each SV caller. The SV calls were considered correct if the simulated and detected SV were 1) of the same type 2) on same chromosome and 3) both start and stop locations were within 50 bp. The callers that performed the best were DELLY and LUMPY as they showed high sensitivity score and low false discovery rate (FDR) score (Table S3).

Performance of CNV callers on simulated data

We scanned for CNV variants using two detection programs, CNVnator version 0.4.1 (Abyzov et al., 2011) and CNV-seq version 0.2-7 (Xie and Tammi, 2009). CNV-seq was used with default parameters while CNVnator was used with two different bin sizes, 75 and 1670. Bins of 75 bp allowed the detection of small events, while bins of 1670 bp, which is the average gene length of *N. crassa* (Galagan et al., 2003), allowed the detection larger-scale events. We excluded MA-line sites if the start or stop location of these where within 500 bp of any site detected in the ancestor. Also, we only retained the sites that were detected by both callers CNVnator and CNVseq (if 1000 bp or less overlapped at the start or end location). The remaining sites were manually verified by inspecting the alignment file in IGV. To evaluate the performance of this workflow we simulated 32 CNVs using SECNVs version 2.7.1 (Xing et al., 2020), then simulated 150 bp paired end reads with an error rate of 0.03% and a mean coverage of 30X using DWGSIM. The merged data set performed better that any of the callers individually by showing the lowest FDR rate score of 0.482, and good

sensitivity score of 0.906 (Table S4).

Supplementary results

Re-analysis of data from Wang et al. 2020

Wang et al. (2020) estimated rate of spontaneous mutation during meiosis in *N. crassa*. During meiosis a genome defence mechanism called RIP induces mainly C $\rightarrow$ T transitions in duplicated regions of the genome resulting in a very high overall mutation rate (Wang et al., 2020). While not made explicit by Wang et al. (2020), the duplicated regions correspond almost completely to H3K9 methylated domains. In order to better compare our results for asexual mutation rate in different domains to the sexual mutation rate estimated in their study, we re-analyzed the data from Wang et al. (2020) provided in their supplementary material, and included the information about chromatin domains. Their data are comprised of mutations in sequenced tetrads, which correspond to the products of a single meiosis. We included only those tetrads originating from crosses between non mutant strains. This leaves 67 tetrads in the data that originate from five different crosses.

First we split the mutations in those that occurred in euchromatin and those that occurred in H3K9 methylated domains. We observed that the numbers of mutations occurring in euchromatin and H3K9 domains for a given tetrad had very different distributions (Figure S2A), number of mutations occurring per tetrad in H3K9 domains had a very long tail. When we examined mutations per tetrad by cross, we observed that for mutations that occurred in euchromatic regions there were a median of 22 mutations per tetrad with some differences among the five crosses, but variation among tetrads from different crosses was similar (Figure S2B). However, for mutations that occurred in H3K9 domains, there were a median of 38 mutations per tetrad but huge variation among tetrads, even within tetrads from a single cross (Figure S2B). For example, some tetrads from the same cross had 20 to 40 mutations, while others could have hundreds. In cross E the range of mutations was from 27 in one tetrad to 1187 in another. Variation among mutations in H3K9 domains per tetrad suggest that while there probably were some genetic influences on mutation rate in the different crosses, there was substantial heterogeneity in the activation of RIP that was independent

of genetic effects.

We calculated mutation rate per meiosis for the euchromatic regions of the genome using a multilevel model with cross as a random factor. The model was

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i) \qquad \text{(S8)} \\
\log(\lambda_i) &= \bar{\alpha} + \alpha_{c[i]} \\
\bar{\alpha} &\sim \text{N}(0, 10) \\
\alpha_c &\sim \text{N}(0, \sigma_c) \\
\sigma_c &\sim \text{hT}(3, 0, 10)
\end{aligned}
$$

where y_i is the number of mutations in euchromatic regions in the ith tetrad, $\bar{\alpha}$ is the average intercept, α_c is deviation from average intercept for each cross, and σ_c is the cross standard deviation. Prior for σ_c was the half-location scale version of Student's t-distribution, with 3 degrees of freedom, location 0, and scale 10. Based on posterior predictive checks, this model fitted the data. Mutation rate was calculated from posterior distribution of $\bar{\alpha}$ as

$$
\mu = \frac{\exp(\bar{\alpha})}{N n_t} \qquad \text{(S9)}
$$

where N is the number of called nucleotides, and n_t is the number of tetrads. The mutation rate in euchromatic regions during sexual reproduction was 1.07 [0.6, 1.67] $\times 10^{-8}$ mutations / meiosis / bp.

The data for mutations that occurred in H3K9 domains are clearly overdispersed. To calculate mutation rate per meiosis for H3K9 methylated domains we also modelled the heterogeneity among tetrads. We fitted a gamma-poisson model, also called a negative binomial model, to the data. A gamma-poisson model allows each observation, a tetrad in our case, to have a different poisson rate

allowing us to model this heterogeneity in observed rates (McElreath, 2015). We fitted a model

$$
\begin{aligned}
y_i &\sim \text{Gamma-Poisson}(\lambda_i, \phi) \\
\log(\lambda_i) &= \bar{\alpha} + \alpha_{c[i]} \\
\bar{\alpha} &\sim \text{N}(0, 10) \\
\alpha_c &\sim \text{N}(0, \sigma_c) \\
\sigma_c &\sim \text{hT}(3, 0, 10) \\
\phi &\sim \Gamma(0.01, 0.01)
\end{aligned}
\tag{S10}
$$

where y_i is the number of mutations in H3K9 methylated domains in the ith tetrad, ϕ is the dispersion parameter, and other parameters were same as above. The prior for ϕ was a gamma distribution with shape of 0.01 and scale 0.01. Posterior predictive check indicated that the model fit the data reasonably well. Then mutation rate was calculated from average intercept as above. The mutation rate in H3K9 methylated regions during sexual reproduction was 2.54 [0.11, 7.55] $\times 10^{-7}$ mutations / meiosis / bp. As a result of rate heterogeneity there is quite a bit of uncertainty in the estimate. The ratio of mutation rates in H3K9 regions over euchromatic regions was 23.7 [0.99, 76.38]. While the 95% interval of the ratio slightly overlaps one due to large uncertainly in mutation rate in H3K9 regions, mutation rate those regions seems higher.

We examined the spectrum of mutations that occurred for euchromatic and H3K9 methylated regions separately in the same way we did for asexual mutations. We observed that in H3K9 regions there was a substantial overrepresentation of C:G $\rightarrow$ T:A transitions due to the action of RIP (Figure S2C). However, the mutation spectra that occurred in euchromatic regions was much more similar to the one we observed during asexual reproduction in euchromatic regions. There was no difference in the relative mutation rate of C:G $\rightarrow$ T:A transitions during sexual and asexual reproduction in euchromatic regions. Some of the transversion did have different relative rates: C:G $\rightarrow$ G:C transversions had higher rate during sexual reproduction, while C:G $\rightarrow$ A:T transversions

had a lower relative rate.

Our analysis gives somewhat different results compared to those of Wang et al. (2020), who only calculated mutation rates across the whole genome, and did not take variation among tetrads or crosses into account. We do find higher mutation rates during sexual reproduction than during asexual reproduction, suggesting that meiosis is mutagenic in *N. crassa* is addition to the RIP effect in H3K9 domains. However, the mutation rate per meiosis was much smaller than that estimated by Wang et al. (2020). H3K9 methylated regions contain mainly transposable elements, and are quite gene poor. If we compare non-synonymous mutations in euchromatic and H3K9 regions, of those mutations that occurred in euchromatic regions 22.16% were non-synonymous, while only 0.17% were non-synonymous in H3K9 methylated regions. Thus, the very high mutation rate observed in H3K9 regions due to action of RIP, does not necessarily translate into a high genetic load. We suggest that mutation load during sexual reproduction may not be as high as it has been suggested for *N. crassa*.

Supplementary References

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein, 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21:974–984. URL `http://genome.cshlp.org/content/21/6/974.abstract`.

Galagan, J. E., S. E. Calvo, K. A. Borkovich, E. U. Selker, N. D. Read, D. Jaffe, W. FitzHugh, L.-J. Ma, S. Smirnov, S. Purcell, B. Rehman, T. Elkins, R. Engels, S. Wang, C. B. Nielsen, J. Butler, M. Endrizzi, D. Qui, P. Ianakiev, D. Bell-Pedersen, M. A. Nelson, M. Werner-Washburne, C. P. Selitrennikoff, J. A. Kinsey, E. L. Braun, A. Zelter, U. Schulte, G. O. Kothe, G. Jedd, W. Mewes, C. Staben, E. Marcotte, D. Greenberg, A. Roy, K. Foley, J. Naylor, N. Stange-Thomann, R. Barrett, S. Gnerre, M. Kamal, M. Kamvysselis, E. Mauceli, C. Bielke, S. Rudd, D. Frishman, S. Krystofova, C. Rasmussen, R. L. Metzenberg, D. D. Perkins, S. Kro-

ken, C. Cogoni, G. Macino, D. Catcheside, W. Li, R. J. Pratt, S. A. Osmani, C. P. C. DeSouza, L. Glass, M. J. Orbach, J. A. Berglund, R. Voelker, O. Yarden, M. Plamann, S. Seiler, J. Dunlap, A. Radford, R. Aramayo, D. O. Natvig, L. A. Alex, G. Mannhaupt, D. J. Ebbole, M. Freitag, I. Paulsen, M. S. Sachs, E. S. Lander, C. Nusbaum, and B. Birren, 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859–868. URL https://doi.org/10.1038/nature01554.

Homer, N., 2021. DWGSIM: Whole genome simulator for next-generation sequencing. github repository. URL https://github.com/nh13/DWGSIM.

Jeffares, D. C., C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, F. Balloux, C. Dessimoz, J. Bähler, and F. J. Sedlazeck, 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nature Communications 8:14061. URL http://dx.doi.org/10.1038/ncomms14061.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall, 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 15:R84. URL https://doi.org/10.1186/gb-2014-15-6-r84.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] .

McElreath, R., 2015. Statistical Rethinking - A Bayesian course with examples in R and Stan. CRC Press, New York.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339. URL https://doi.org/10.1093/bioinformatics/bts378.

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri,

P. Tomancak, and A. Cardona, 2012. Fiji: an open-source platform for biological-image analysis. Nature Methods 9:676–682. URL https://doi.org/10.1038/nmeth.2019.

Wala, J. A., P. Bandopadhayay, N. F. Greenwald, R. O'Rourke, T. Sharpe, C. Stewart, S. Schumacher, Y. Li, J. Weischenfeldt, X. Yao, C. Nusbaum, P. Campbell, G. Getz, M. Meyerson, C.-Z. Zhang, M. Imielinski, and R. Beroukhim, 2018. SvABa: genome-wide detection of structural variants and indels by local assembly. Genome Research 28:581–591. URL http://genome.cshlp.org/content/28/4/581.abstract.

Wang, L., Y. Sun, X. Sun, L. Yu, L. Xue, Z. He, J. Huang, D. Tian, L. D. Hurst, and S. Yang, 2020. Repeat-induced point mutation in *Neurospora crassa* causes the highest known mutation rate and mutational burden of any cellular life. Genome Biology 21:142. URL https://doi.org/10.1186/s13059-020-02060-w.

Xie, C. and M. T. Tammi, 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10:80. URL https://doi.org/10.1186/1471-2105-10-80.

Xing, Y., A. R. Dabney, X. Li, G. Wang, C. A. Gill, and C. Casola, 2020. SECNVs: a simulator of copy number variants and whole-exome sequences from reference genomes. Frontiers in Genetics 11:82.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871. URL https://doi.org/10.1093/bioinformatics/btp394.

Supplementary Figures

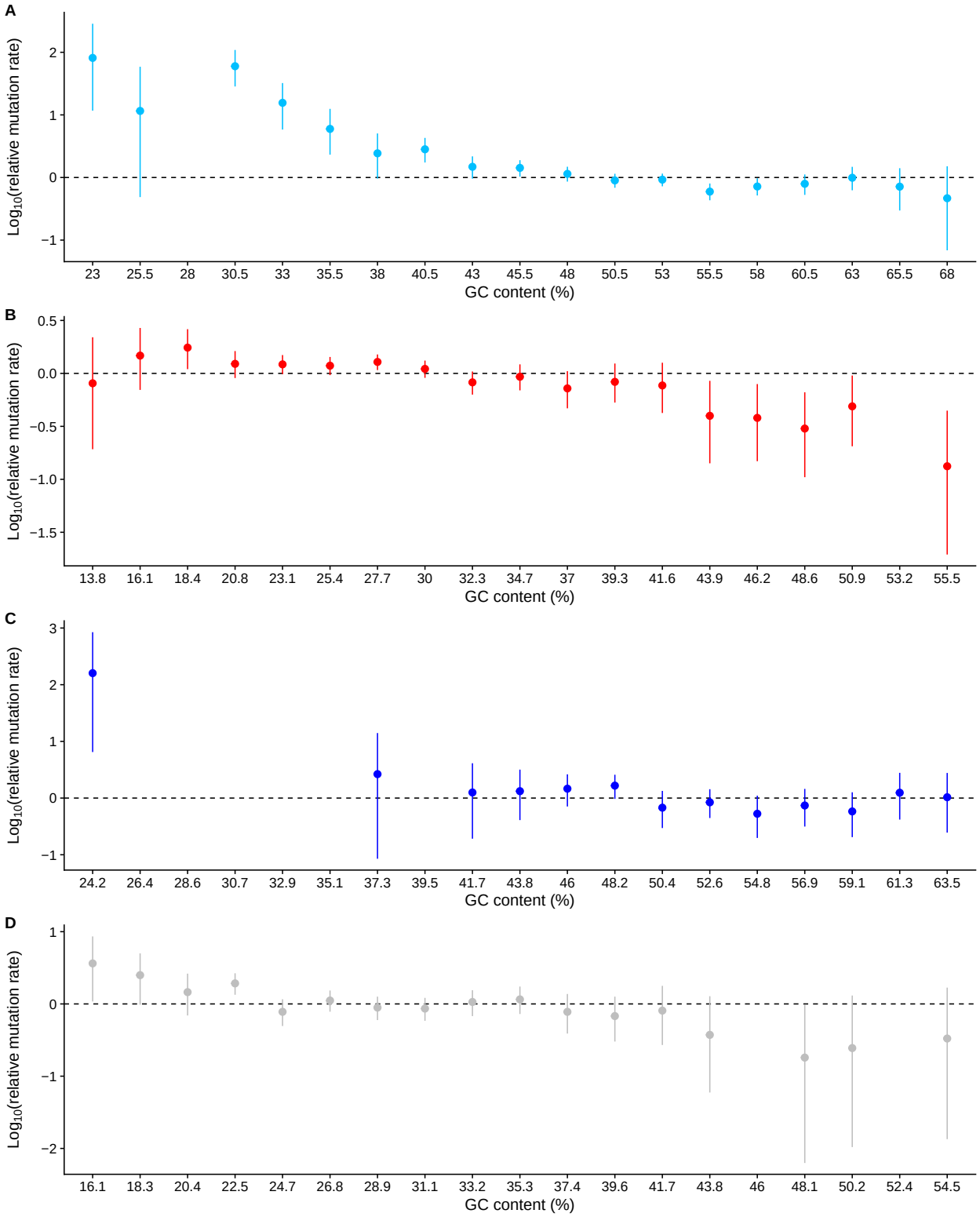

Figure S1: Relative mutation rates for windows of 200 bp binned for GC-content at 2.5 percentage point intervals. Note that y-axis is on $\log_{10}$ scale, dashed line indicates relative mutation rate of one. A) Euchromatic regions B) H3K9 domains C) H3K27 domains D) Centromeric regions.

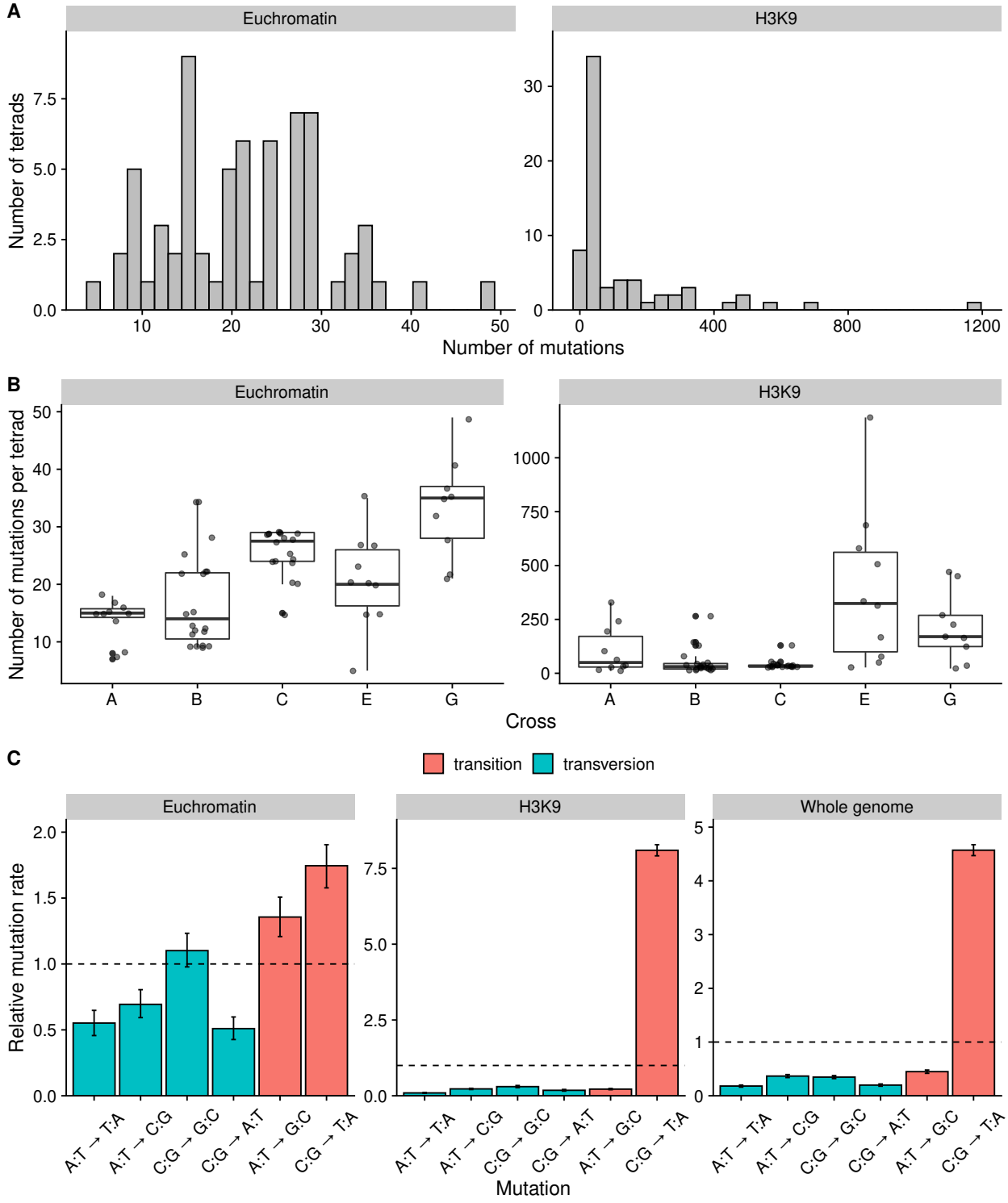

Figure S2: Mutations that occurred during sexual reproduction. Data is from Wang et al. (2020). Note that y-axis scales are different in different panels. A) Distribution of number of mutations in the tetrads in euchromatin and H3K9 methylated domains. B) Number of mutations per tetrad for the different crosses. C) Spectrum of mutations for different regions of the genome.

Supplementary Tables

Table S1: Summary of alignment metrics for genomes used in this study. The ancestors used to start the MA-experiment were: B 26708, which is 2489 *mat A*, and B 26709, which is 2489 *mat a*. Lines L1–L20 are *mat A* and L21–L40 are *mat a*.

Natural strains				MA-lines			
Line	Number of reads	Depth	Mapped reads (%)	Line	Number of reads	Depth	Mapped reads (%)
10882	30683221	112	87.2	B 26708	15183018	55	98.5
10883	20920000	76	69.7	B 26709	18679696	68	98.3
10884	32294020	118	72.8	L1	13215012	48	98.2
10886	15265173	56	92.7	L2	15658340	57	98.2
10892	32124218	117	62.3	L3	14699884	54	98.3
10904	14718008	54	91.1	L4	16424902	60	98.2
10906	14935778	55	89.7	L5	16328002	60	97.6
10907	32213115	118	68.1	L6	15189850	55	98.3
10908	13924689	51	91.7	L7	14006386	51	98.4
10912	14829204	54	92	L8	13888092	51	98
10914	30158241	110	61.1	L9	14386986	53	98.3
10915	24410441	89	61.1	L10	15458492	56	98.6
10918	40163515	147	65.4	L11	13884490	51	98.1
10923	16231892	59	92	L12	15913482	58	98.2
10925	37031022	135	85.2	L13	20192750	74	97.8
10926	32682749	119	66	L14	17445964	64	98.1
10927	17741573	65	73.1	L15	14776588	54	98.5
10928	15048163	55	92	L16	14562124	53	98
10932	13402421	49	91.1	L17	16043040	59	97.9
10935	22304835	81	81.1	L18	14755826	54	97.7
10937	29757556	109	85.4	L19	13712542	50	98.1
10943	32838278	120	72.8	L20	18685746	68	96.9
10946	17261401	63	90.1	L21	14814794	54	98.5
10948	13835447	50	91	L22	15849832	58	98
10950	14364201	52	98	L23	17528602	64	98.3
10951	13354998	49	89.5	L24	14415798	53	97.8
10983	38349194	140	72.8	L25	14773870	54	98.1
1131	16974127	62	88.7	L26	17436754	64	98.2
1133	14540257	53	88.3	L27	14047860	51	98.2
1165	15095250	55	90	L28	16295790	59	98.3
3210	13681263	50	90.4	L29	16244750	59	98.1
3211	13908534	51	89.4	L30	19847786	72	98.1
3223	14095397	51	92.1	L31	19613222	72	97.9
3943	143993668	525	82.6	L33	33430658	122	97.9
3975	13671543	50	88.8	L34	15456254	56	97.9
4708	14310459	52	87.7	L35	14405808	53	98
4712	12602174	46	86	L36	15567642	57	97.8
4716	18336337	67	86.7	L37	15232240	56	98
4730	15588817	57	87.4	L38	14265830	52	98
4824	12921275	47	88.3	L39	16540462	60	98.1
5910	12494976	46	83.4	L40	18856392	69	98.3
6203	13266522	48	89.3				
851	12956372	47	89.7				
8783	15381794	56	88				
8790	13205097	48	90.1				
8816	15452652	56	89.6				
8819	13736103	50	87.5				
8845	15868843	58	86.9				
8850	12695261	46	89.3				
P4452	172179164	628	86.9				
P4463	21708426	79	69.7				
P4468	21129155	77	68.4				
P4471	24828717	91	72.4				
P4476	92452134	337	86.8				
P4479	22220857	81	72.5				
P4489	34624044	126	72.5				

Table S2: Natural strains with sequencing data included in this study. Strains were obtained from FGSC. 33 strains were sequenced in this study and data 23 strains were obtained from Zhao et al. (2015).

Strain	Source	Strain	Source
10948	This study	P4452	(Zhao et al., 2015)
10886	This study	P4463	(Zhao et al., 2015)
10932	This study	P4468	(Zhao et al., 2015)
1165	This study	P4471	(Zhao et al., 2015)
8816	This study	P4476	(Zhao et al., 2015)
3223	This study	P4479	(Zhao et al., 2015)
8845	This study	10882	(Zhao et al., 2015)
10908	This study	10883	(Zhao et al., 2015)
10904	This study	10884	(Zhao et al., 2015)
851	This study	10892	(Zhao et al., 2015)
1131	This study	10907	(Zhao et al., 2015)
8850	This study	10914	(Zhao et al., 2015)
8819	This study	10915	(Zhao et al., 2015)
4708	This study	10918	(Zhao et al., 2015)
4712	This study	10925	(Zhao et al., 2015)
6203	This study	10926	(Zhao et al., 2015)
4824	This study	10927	(Zhao et al., 2015)
8783	This study	10935	(Zhao et al., 2015)
8790	This study	10937	(Zhao et al., 2015)
3975	This study	10943	(Zhao et al., 2015)
10928	This study	10983	(Zhao et al., 2015)
10912	This study	3943	(Zhao et al., 2015)
3210	This study	P4489	(Zhao et al., 2015)
10923	This study		
10950	This study		
10951	This study		
10946	This study		
3211	This study		
10906	This study		
5910	This study		
4730	This study		
1133	This study		
4716	This study		

Table S3: Detecting structural variants with different callers from simulated data.

							Delly		Lumpy		SVaba		Pindel	
set	Deletion	Duplication	Inversion	Translocation	Insertion	Total number of SV	sensitivity	FDR	sensitivity	FDR	sensitivity	FDR	sensitivity	FDR
1	1	5	4	8	2	2	0.90	0	0.90	0.55	0.60	0.60	0.55	0.65
2	4	10	4	16	5	39	0.84	0.02	0.84	0.54	0.51	0.55	0.48	0.60
3	5	10	10	30	7	62	0.88	0	0.88	0.54	0.64	0.52	0.38	0.61
4	49	20	40	40	11	160	0.54	0.33	0.54	0.60	0.36	0.77	0.28	0.94

Table S4: Calling CNVs on simulated data using different approaches.

	Sensitivity score	FDR score
CNVnator (1670 bin size)	0.375	0.556
CNVnator (75 bin size)	0.968	0.797
CNV-seq	0.937	0.999
Merged data	0.906	0.482

Table S5: Model comparison among different models of relative mutation rate of different triplets and the number of pyrimidine dimers they contain. Model terms are the different linear model parts in equation XXX, β_{TT} is the effect of TT dimers, β_{CCT} is the effects of CC or CT dimers, β_p is the effects of all pyrimidine dimers, β_{CC} is the effect of CC dimers, and β_{CT} is the effect of CT dimers. LOOIC = Leave-one-out information criterion, SE = standard error.

Whole genome			
Model terms	LOOIC	diff ($\pm$ SE)	weight
$\alpha + \beta_{TT}x_i + \beta_{CCT}c_i$	-2.31	0 (0)	0.4
$\alpha + \beta_{TT}x_i$	-2.13	0.18 (2.56)	0.36
$\alpha + \beta_{TT}x_i + \beta_{CC}d_i + \beta_{CT}z_i$	-0.65	1.66 (1.01)	0.17
$\alpha + \beta_p p_i$	1.53	3.84 (4.62)	0.06
α	4.75	7.06 (5.76)	0.01
H3K9me domains			
Model terms	LOOIC	diff ($\pm$ SE)	weight
$\alpha + \beta_{TT}x_i + \beta_{CC}d_i + \beta_{CT}z_i$	36.47	0 (0)	0.98
$\alpha + \beta_{TT}x_i + \beta_{CCT}c_i$	45.25	8.78 (5.99)	0.01
$\alpha + \beta_p p_i$	45.61	9.13 (6.48)	0.01
α	51.79	15.32 (6.84)	0
$\alpha + \beta_{TT}x_i$	53.93	17.46 (6.45)	0
Euchromatin			
Model terms	LOOIC	diff ($\pm$ SE)	weight
$\alpha + \beta_p p_i$	10.94	0 (0)	0.26
$\alpha + \beta_{TT}x_i + \beta_{CC}d_i + \beta_{CT}z_i$	11.17	0.23 (1.02)	0.24
α	11.43	0.49 (1.19)	0.21
$\alpha + \beta_{TT}x_i + \beta_{CCT}c_i$	12.06	1.12 (0.27)	0.15
$\alpha + \beta_{TT}x_i$	12.17	1.23 (1.16)	0.14

Table S6: Results of for a model estimating the effects of pyrimidine dimers on relative mutation rate of different triplets. The full model is given in equation 4. For parameters: α is the intercept, β_{TT} is the effect of TT dimers, β_{CCT} is the effect of CC or CT dimers, and σ is the standard deviation.

Whole genome	Estimate [95% HPDI]		
Parameter	All mutations	Transitions	Transversions
α	0.34 [0.18, 0.5]	0.37 [0.28, 0.47]	0.29 [0.05, 0.55]
β_{TT}	0.33 [0.13, 0.53]	0.29 [0.16, 0.44]	0.34 [0.02, 0.66]
β_{CTT}	0.1 [-0.04, 0.23]	0.04 [-0.04, 0.13]	0.15 [-0.07, 0.37]
σ	0.2 [0.13, 0.31]	0.08 [0.01, 0.17]	0.32 [0.21, 0.49]
H3K9me domains	Estimate [95% HPDI]		
Parameter	All mutations	Transitions	Transversions
α	0.52 [0.14, 0.92]	0.56 [0.28, 0.92]	0.54 [0, 1.1]
β_{TT}	0.26 [-0.22, 0.72]	0.32 [-0.05, 0.66]	0.18 [-0.43, 0.76]
β_{CT}	0.36 [-0.03, 0.75]	0.23 [-0.1, 0.55]	0.35 [-0.13, 0.87]
β_{CC}	1.54 [0.91, 2.25]	1.15 [0.53, 1.79]	1.58 [0.73, 2.56]
σ	0.47 [0.29, 0.78]	0.28 [0.11, 0.57]	0.58 [0.33, 1]
Euchromatin	Estimate [95% HPDI]		
Parameter	All mutations	Transitions	Transversions
α	0.8 [0.59, 1.02]	0.83 [0.53, 1.18]	0.58 [0.14, 1.05]
β_p	0.12 [-0.06, 0.31]	0.05 [-0.22, 0.29]	0.23 [-0.13, 0.63]
σ	0.11 [0.01, 0.3]	0.29 [0.11, 0.52]	0.47 [0.25, 0.81]

Table S7: Model comparisons among different models of mutation rate predicted by GC-content and chromatin modifications. Model terms are different linear model parts, α is the intercept, β_{GC} is the slope effect of GC-content, β_{K9} is the effect of H3K9 domain, β_{K27} is the effect of H3K27 domain, β_C is the effect of centromeric domain, β_I is the interaction effect between GC-content and H3K9 domain, β_{I2} is the interaction effect between GC-content and centromeric domain, and β_{I3} is the interaction effect between GC-content and H3K27 domain. WAIC = widely applicable information criterion, SE = standard error.

Model terms	WAIC	diff ($\pm$ SE)	weight
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i + \beta_{K27}g_i + \beta_C c_i + \beta_I x_i d_i$	454.47	0 (0)	0.63
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i + \beta_{K27}g_i + \beta_C c_i + \beta_I x_i d_i + \beta_{I3}x_i g_i$	456.86	2.39 (2.19)	0.19
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i + \beta_C c_i + \beta_I x_i d_i$	458.67	4.2 (6.94)	0
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i$	458.84	4.37 (10.59)	0
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i + \beta_C c_i + \beta_I x_i d_i + \beta_{I2}x_i c_i$	460.35	5.88 (7.17)	0
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i + \beta_I x_i d_i$	495.86	41.39 (19.86)	0
$\alpha + \beta_{GC}x_i + \beta_{K9}d_i$	496.65	42.18 (22.43)	0
$\alpha + \beta_{GC}x_i$	546.62	92.15 (33.84)	0
$\alpha + \beta_{K9}d_i + \beta_C c_i$	614.83	160.36 (42.53)	0
$\alpha + \beta_{K9}d_i$	645.82	191.35 (49.19)	0
$\alpha + \beta_C c_i$	1290.02	835.55 (255.52)	0
α	1689.56	1235.09 (264.15)	0

Table S8: Model estimates for a model predicting mutation rate by GC-content, centromeric, H3K9, and H3K27 domains, α is the intercept, β_{CG} is the slope effect of GC-content, β_{K9} is the effect of H3K9me domain, β_{K27} is the effect of the H3K27me3 domain, β_C is the effect of centromeric domain, and β_I is the interaction effect between GC-content and H3K9me domain.

Parameter	Estimate [95% HPDI]
α	-2.61 [-3.34, -1.88]
β_C	0.51 [0.35, 0.67]
β_{K9}	-0.14 [-0.93, 0.63]
β_{K27}	0.32 [0.08, 0.55]
β_{GC}	-0.06 [-0.08, -0.05]
β_I	0.02 [0.00, 0.04]