Variation in spontaneous mutation rate and spectrum across the genome of *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: Variation in mutation rate

Keywords: fungi, mutation accumulation, genomics, chromatin, 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 rate have been extensively studies in eukaryotes. However, much less is known about the variation of 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 over 1300 mutations, which happened during asexual propagation. This leads to markedly different mutation rate and spectrum than during sexual reproduction as previously investigated. We observed that GC-content, H3K9 methylation, and centromeric regions have large influence on mutation rate, with higher mutation rate in H3K9 and centromeric regions. H3K27 trimethylation had no effect on mutation rate. We also observe that the spectrum of single nucleotide mutations is different in H3K9 and euchromatic domains. We validate our mutation model by comparing the predictions to natural genetic variation and observed that a moderate amount of extant genetic variation can be predicted by our model. Furthermore, we characterize structural variants and complex mutations that happen in *Neurospora crassa*. Our study highlights that chromatin modifications influence mutation rate and accurate evolutionary inferences must take this variation in mutation rates into account.

Introduction

New mutations are the ultimate source of all genetic diversity, and evolutionary change is dependent on the input of new mutations into the population. As such, mutations generate the genetic diversity used for plant and animal breeding. 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.

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 mutation 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; Ness et al., 2015; Keightley et al., 2015; Keith et al., 2016). These studies have allowed us to estimate mutation rates with high precision and have revealed the spectrum of mutations that happen.

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 certain classes of mutations, such as for transitions and transversions, there is also variation across the genome in mutation rates that was not easy to explain initially. Certain features of the genome including nucleosome positioning (Chen et al., 2012; Li and Luscombe, 2020) and chromatin structure due to epigenetic modifications has a strong influence on mutation rates (Schuster-Böckler and Lehner, 2012; Weng et al., 2019; Monroe et al., 2022).

While we have now some idea how chromatin structure influences mutation rates, at the moment this data comes mainly only from a few species. 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. Understanding mutation will help us parameterize population genetic models better and infer how natural selection and demographic forces 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 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 defence mechanism of *N. crassa* (Wang et al., 2020). However, in this study only a small number of mutations that happened during asexual reproduction was investigated 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 acress 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 does it reflect 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 °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 1.5 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. 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.

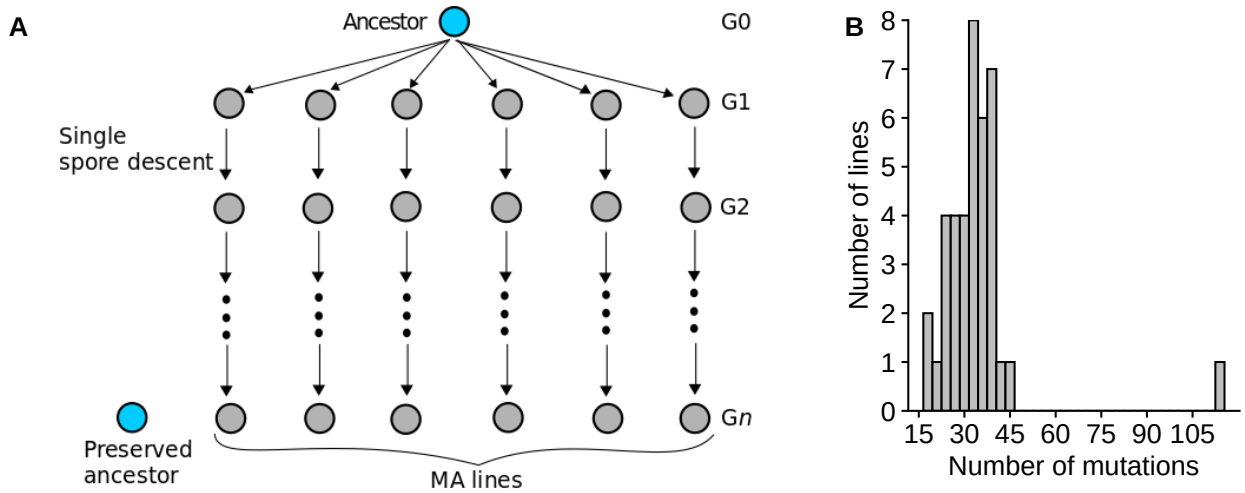

Figure 1: A) Overview of the mutation accumulation experiment. Ancestor is split into multiple lines, which are propagated via single spore descent. This minimizes the efficacy of natural selection and allows even deleterious mutations to accumulate in the MA-lines. B) Distribution of mutations in the MA-lines.

Strains from natural populations

We obtained 33 strains from the Fungal Genetics Stock Center (McCluskey et al., 2010), in addition we obtained genome sequencing data for additional 23 strains from Zhao et al. (2015) by

downloading this data from short read archive. In total, including the laboratory strain 2489, we had 57 strains with sequencing data (Table S1).

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 (concentration?) 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 precipiated for 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. Furthermore, 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

We aligned each MA-line to the reference genome using BWA v0.7 (Li, 2013), and exclude duplicated reads using SAMBLASTER v0.1.26 (Faust and Hall, 2014). From each alignment we extracted the discordant paired-end and the split-read alignments using SAMTOOLS v1.9 (Danecek et al., 2021). To detect structural variation we ran the samples as somatic-germline mode, considering the mutation accumulation line as somatic and the ancestor as germline. For LUMPY v0.2.13 the read and insert lengths were extracted from mapping files (bam files) using samtools and the SVs were genotyped using SVTyper v0.7.1 (Chiang et al., 2015). To filter SV that were exclusive of the mutation accumulation lines (somatic SV) we used SnpSift v5.0e(Cingolani et al., 2012) to filter out the calls in which the genotype of the sample and of the ancestor were the same, those in which the genotype quality was lower than 30 and read depth below 10. DELLY v0.8.7 was used as the recommended workflow (Rausch et al., 2012). Each of the SV independently identified by each caller were verified manually using the genome browser IGV.

To asses the performance of four different SV (structural variants) callers we simulated four data sets with different types and amount of SV's. Using SURVIVOR v1.0.7 (Jeffares et al., 2017),

we modified the fasta file of the reference genome to simulate 20, 39, 62 and 160 SV, we set the mutation rate to be 0.001. Based on the modified fasta files we simulated 150 bp pair end reads with an error rate of 0.03% and a mean coverage of 30X using DWGSIM v0.1.11. The simulated reads were aligned to the reference genome using BWA using the default parameters and analyzed with: DELLY v0.8.7, LUMPY v0.2.13, PINDEL v0.2.5b9, SVaba v1.1.0 (Rausch et al., 2012; Layer et al., 2014; Ye et al., 2009; Wala et al., 2018). To evaluate the performance of each SV caller, we used the eval function of SURVIVOR v1.0.7 to evaluate the reads, in which a prediction is considered correct if the simulated and simulated SV were of the same type, were predicted to be on same chromosome and their start and stop locations were with 50 bp. The callers that performed best were DELLY v0.8.7 and LUMPY v0.2.13 as they had high sensitivity score and low false discovery rate (FDR) score (Table S2). Therefore we used these two callers to detect SV in the mutation accumulation lines.

Chromatin modifications

ChIP-seq reads for H3K9 methylation and H3K27 trimethylation were obtained from Jamieson et al. (2013), accession numbers SRX248101 and SRX248097. 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, duplicated regions that were defined by Wang et al. (2020) were obtained.

Validation of mutations with Sanger sequencing

To confirm that the mutations we called from the Illumina sequencing data were real, we selected a sample of mutations that passed our threshold but had the lowest quality scores as these were the mutations we were most uncertain of. Furthermore, 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. Sanger sequencing was performed using standard

methods.

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

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

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

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

where N is the number of called nucleotides and t is the number of transfers the MA-lines went through. 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 convergence 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 trinu-

cleotides 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.

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 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

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.

In total we observed 1329 mutations, with a median of 33 mutations per MA-line. One of the MA-lines had an excess amount of mutations (Figure 1B), 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, 101 deletions, 9 complex mutations where a

single mutational event created multiple nucleotide changes, 7 translocations, and 1 inversion. Total mutation rate during asexual propagation was 0.85 [0.81, 0.9] mutations / genome / TRANSFER.

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 domains (Figure 2A). The duplicated regions *N. crassa* genome defined by Wang et al. (2020) were almost perfectly correlted with H3K9me domains (Figure 2A). Examining relative mutation rates confirmed that mutation rate was higher in centromeric and H3K9me domains, while mutation rate H3K27me domains was the same as in euchromatin (Figure 2B). We also observed that GC-content displayed a bimodal distribution and was lower in H3K9me and centromeric domains, with distribution of CG content overlapping very little euchromatic domains (Figure 2C). This raises the question does the lower mutation rate in H3K9me regions arise from the lower CG-content itself, some other factor related to the chromatin modifications, or a combination of GC-content and chromatin modifications. First to rule out Simpsons paradox type of effects we examined the effects of CG-content on mutation rate and observed that there was an effect of CG-content within each domain (Figure S2). Thus, we could not ignore CG-content as an explanation for this pattern.

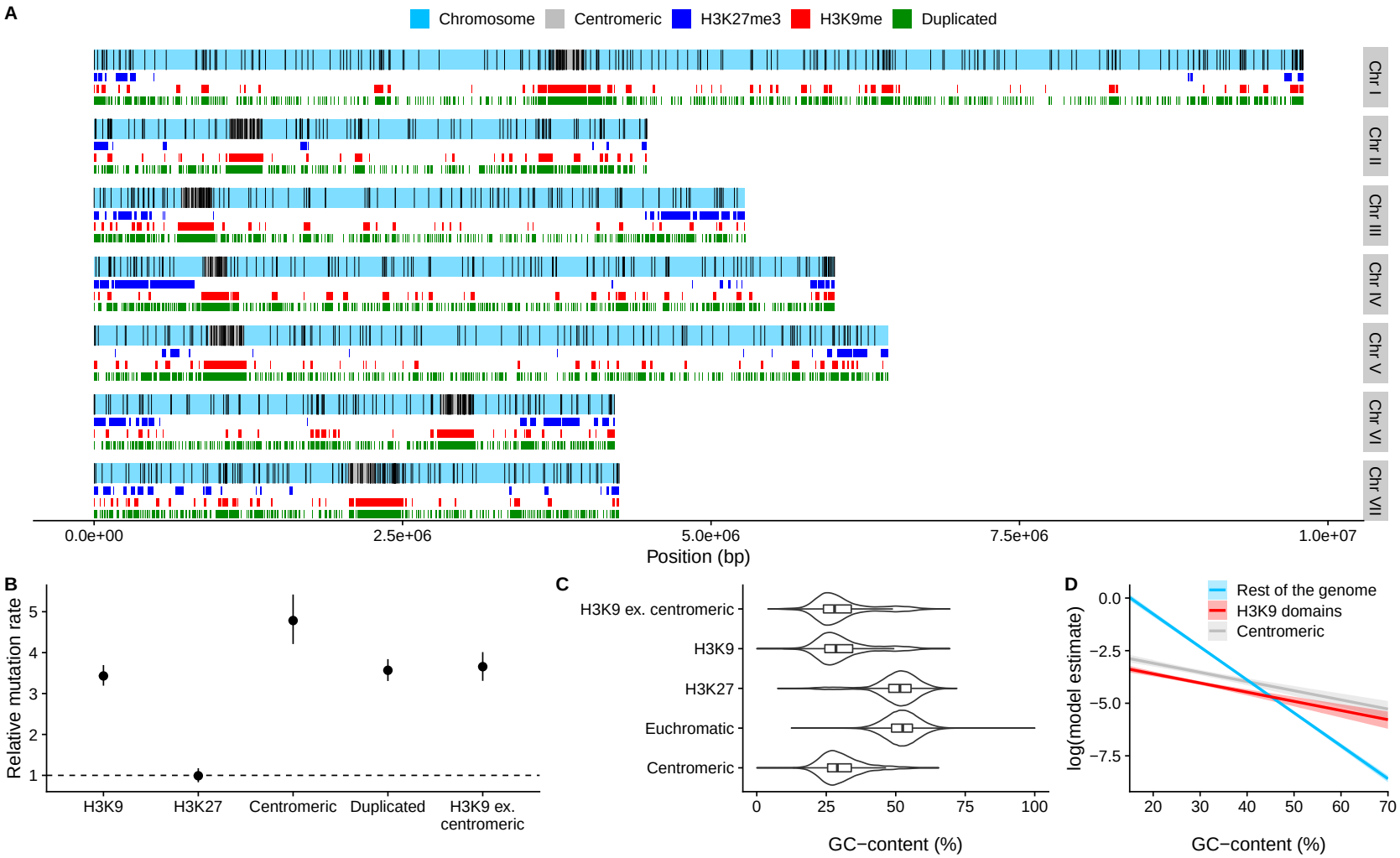

Figure 2: A) Distribution of mutations along the seven *N. crassa* chromosomes. Black vertical lines indicate mutations. Centromeric regions, H3K27 trimethylation and H3K9 methylation domains, and duplicated regions are shown. B) Relative mutation rates for different genomic domains. H3K9 ex. centromeric are H3K9 domains where those domains that overlap with centrometic 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.

To investigate the joint effects of GC-content and chromatin modifications we fitted a models with different predictors, including GC-content, H3K9 domain, and centromeric domains. Based on model comparisons the best model included the effect of GC-content, the effect of H3K9 domain, the effect of centromeric region, and the interaction between CG-content and H3K9-domain (Table S5). Based on model estimates GC-content has a strong effect on mutation rate euchromatic domains. However, as there are very few regions with low GC-content the very high mutation rate estimates for low GC-content in euchromatic regions must be viewed with certain scepticism (Figure 2D). In H3K9 regions, CG-content has a smaller effect on mutation rate, and centromeric regions have a statistically detectable increase in mutation rate on top of H3K9 domain (Figure 2D), even if centromeric regions always have H3K9me as well.

Genetic variation in natural populations and mutation rate

To further validate our mutation model we were interested to examine if natural genetic variation can be explained by our mutation model. First we calculated θ_W 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 H3K9 domains like mutation rate. In addition, θ_W was elevated in H3K27 domains, 0.024 [0.004, 0.067] even though mutation rate in H3K27 was the same as in euchromatin.

To examine how genetic variation segregating in natural populations is dependent on mutation rate variation across the genome, we used our mutation model (Table S6) 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 correctly predict moderate amount of variation in θ_W (Figure 3B). A measure of model fit, the Bayesian R^2 value was 0.22 [0.22, 0.23]. 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*.

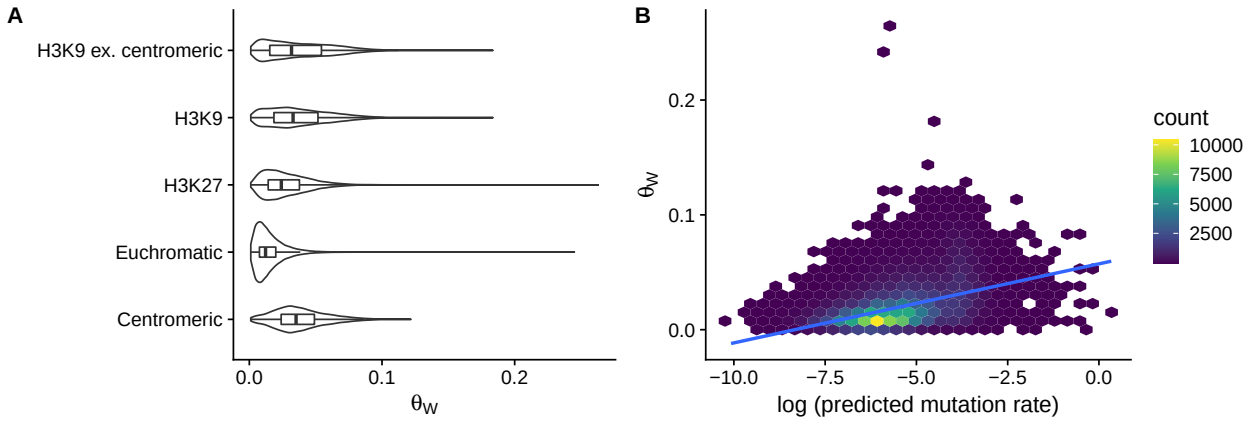

Figure 3: A summary statistic of genetic variation, θ_W, was calculated across the genome in 200 bp windows. Only windows where $\theta_W > 0$ were included in the analysis, $n = 190669$. A) Distribution of θ_W among different genomic domains. B) Relationship with θ_W and predicted mutation rate. For 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.

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 1.7×10^{-8} [1.6×10^{-8}, 1.8×10^{-8}] mutations / bp / TRANSFER. The ratio of transition to transversion rates was 1.08 [0.95, 1.21], which is on the low end of reported Ts/Tv ratios. We observed that spectra of single nucleotide mutations was different for H3K9 domains versus rest of the genome (Figure 4A). A:T $\rightarrow$ C:G and C:G $\rightarrow$ G:C transversion were more common in H3K9 domains, while A:T $\rightarrow$ G:C transition were 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. A:T $\rightarrow$ G:C transitions was the most common SNM, and they 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 [1.03, 1.44].

SNM mutational mechanisms

In ordet 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

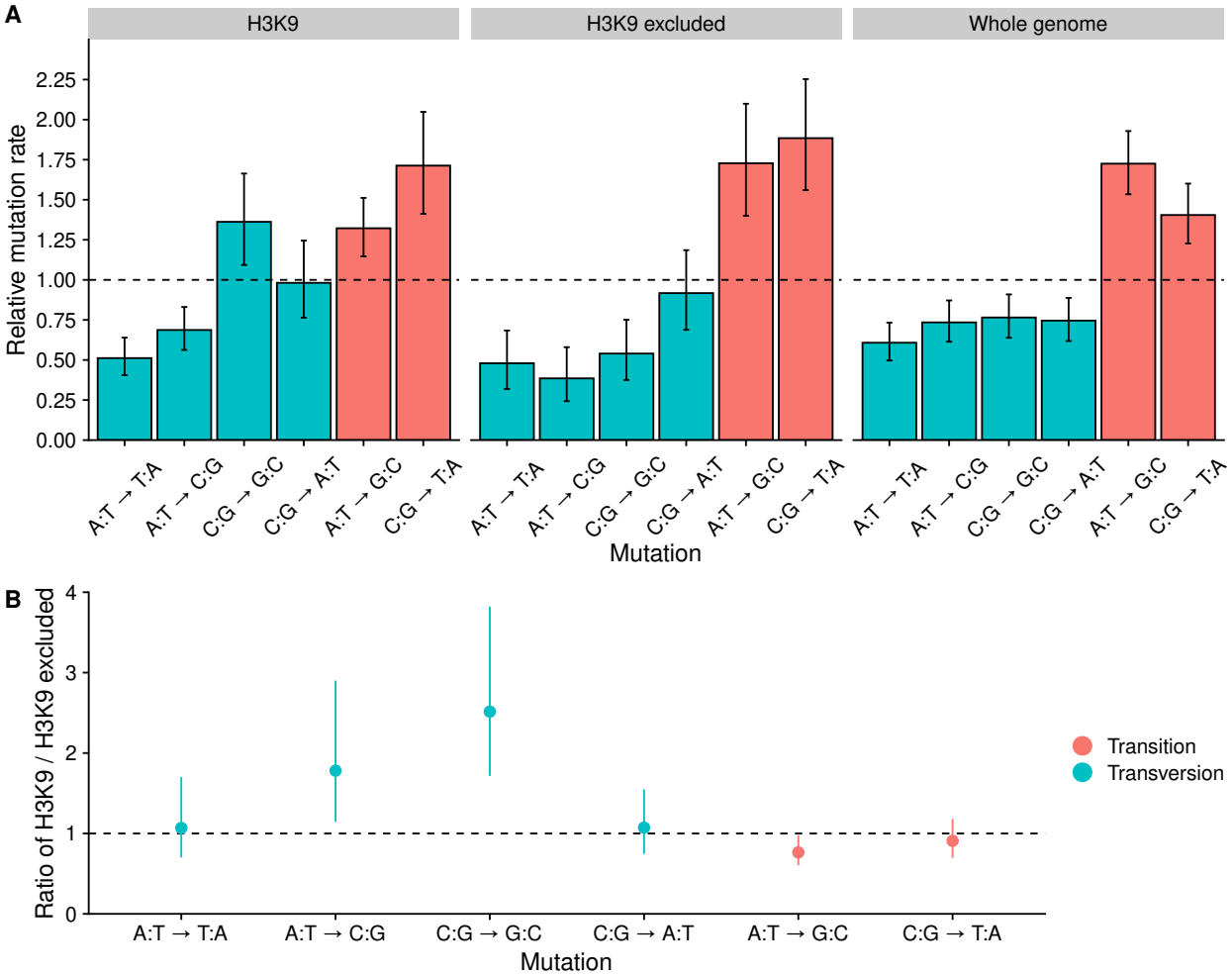

Figure 4: Spectrum of relative SNM rates, dashed line shows the expected rate if all mutations occurred at equal frequencies. Nucleotide frequencies were taken into account in calculating the relative rates. Error bars are 95 % HPD intervals. Transversions are colored blue, transitions red.

trinucleotide. We combined trinucleotides with respect to sequence complementary and directional symmetry, which leaves 20 trinucleotides in total. After taking into account the frequencies at which different types of triplets occur, mutations were not equally distributed across triplets. For triplets where the mutating base was A:T we observed that mutations occurred more likely in those triplets 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 TCT:AGA triplets had the highest mutation rates. However, whenever C was flanked by an A (or G by a T) these triplets had lower relative mutation rates (Figure 5).

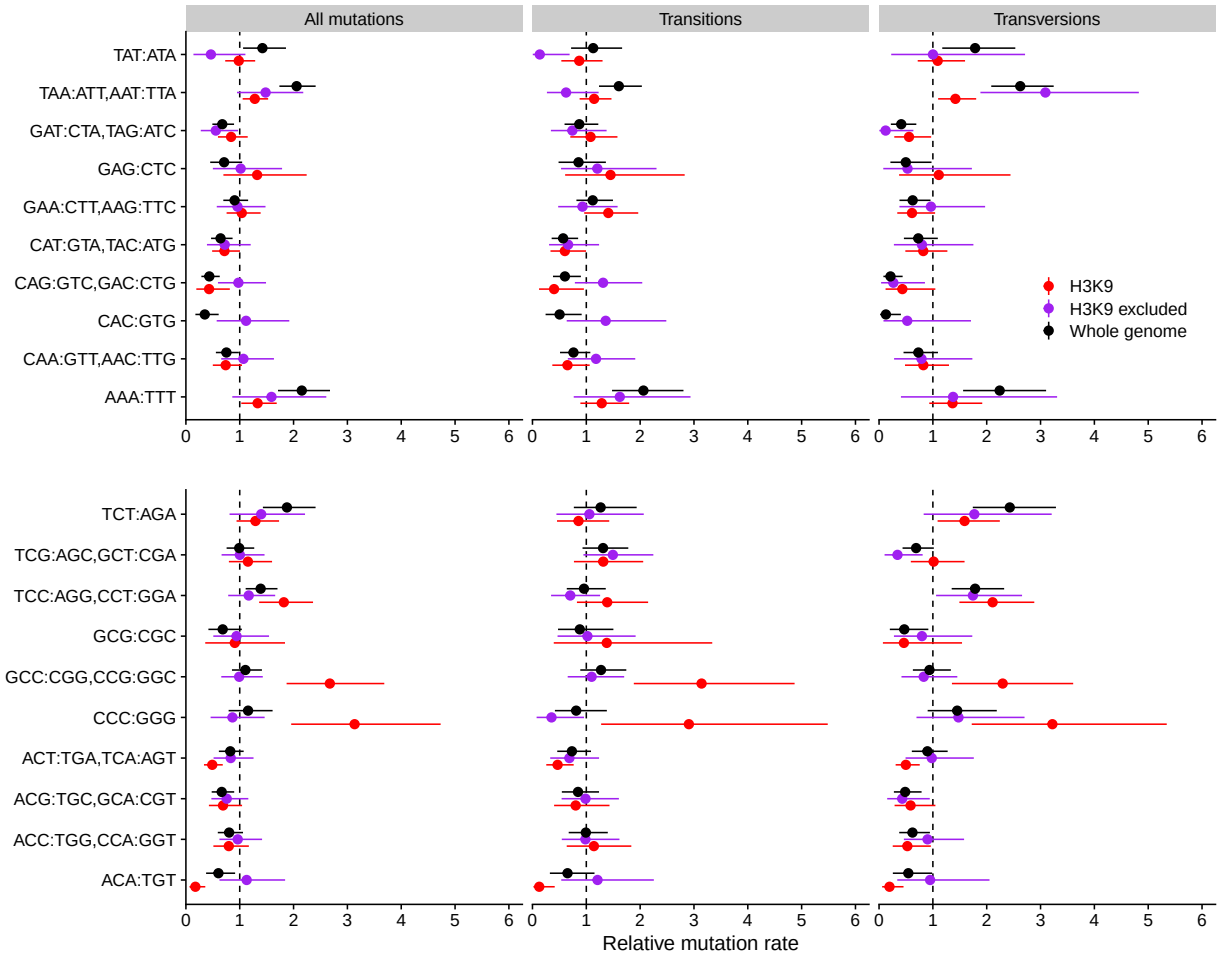

Figure 5: Relative mutation rates for different classes of triplets. Triplets where the middle base is A:T are on the top row and C:G triplets on the bottom row. Range shows 95% HPD interval of the relative mutation rate. Colours show genomic domains.

We also observed domain specific differences in relative trinucleotide mutation rates. In partic-

ular 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. 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 rate we looked at pairs of adjacent bases in the trinuceotides as thymine photodimerization has been suggested to be an important mutational mechanism. We tested the idea that having one or more pyrimidine dimers in a trinucleotide increases mutation rate, by using 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 was clearly preferred, except that the models with pyrimidine effects gave superior predictions compared to a model with intercept only (Table S3). Here we show results for a model with separate predictors for TT dimers and CT and CC dimers combined. For H3K9 domains, a model with all classes as separate was preferred, whoe euchromatic regions a model with all classes combined was preferred.

We observed that pyrimidine dimers did increase relative mutation of trinucleotide over the whole genome (Figure 6). In H3K9 domains CC dimers had a strong effect, while the other pyrimidine dimers had no effect. In euchromatic domains the effect of pyrimidine dimers was weak at best (Figure 6). These observations suggest that mutation rate is generally higher for adjacent thymines, likely due to photodimerization. This accounts for some of the increased mutation rate in H3K9 domains, due to their low GC-content. However, there are some additional mechanism in operation in H3K9 domains, as adjacent cytosines increases the muration rate. This is most likely due to deamination of methylated cytosines, as methylation is present only in H3K9 domains in *N. crassa*.

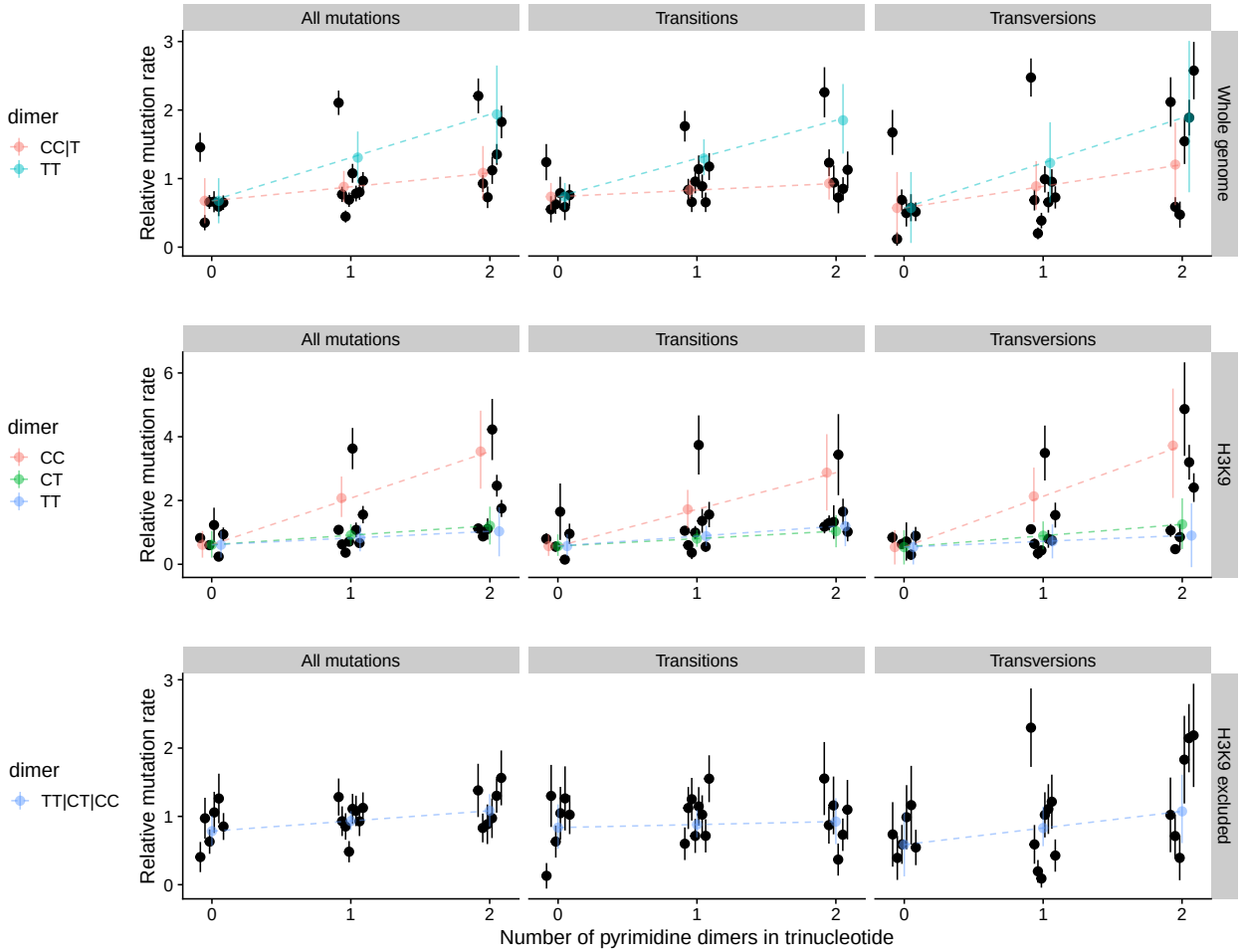

Figure 6: Relative mutation rate and number of pyrimidine dimers. Black points are the observed relative mutation rates for different triplets, blue, red, and green points are model estimates for TT and CC or CT dimers respectively.

Deletions, insertions, and translocations

In addition to point mutations, we also examined structural variants that happened in the MA-lines. We observed 101 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 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).

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 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 1061 bp, which was much longer than the 27 bp mean lenght 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 for deletions and insertions and their ratio.

	Mutation rate (mutations / genome / TRANSFER)		
	Deletions	Insertions	Insertion / deletion ratio
All indels	0.06 [0.05, 0.08]	0.09 [0.07, 0.1]	1.33 [1.01, 1.69]
Repeats excluded	0.03 [0.02, 0.04]	0.01 [0, 0.01]	0.27 [0.13, 0.46]
Repeats only	0.03 [0.03, 0.04]	0.08 [0.06, 0.09]	2.25 [1.58, 3.03]

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 we found that even when taking numbers of homopolymers into

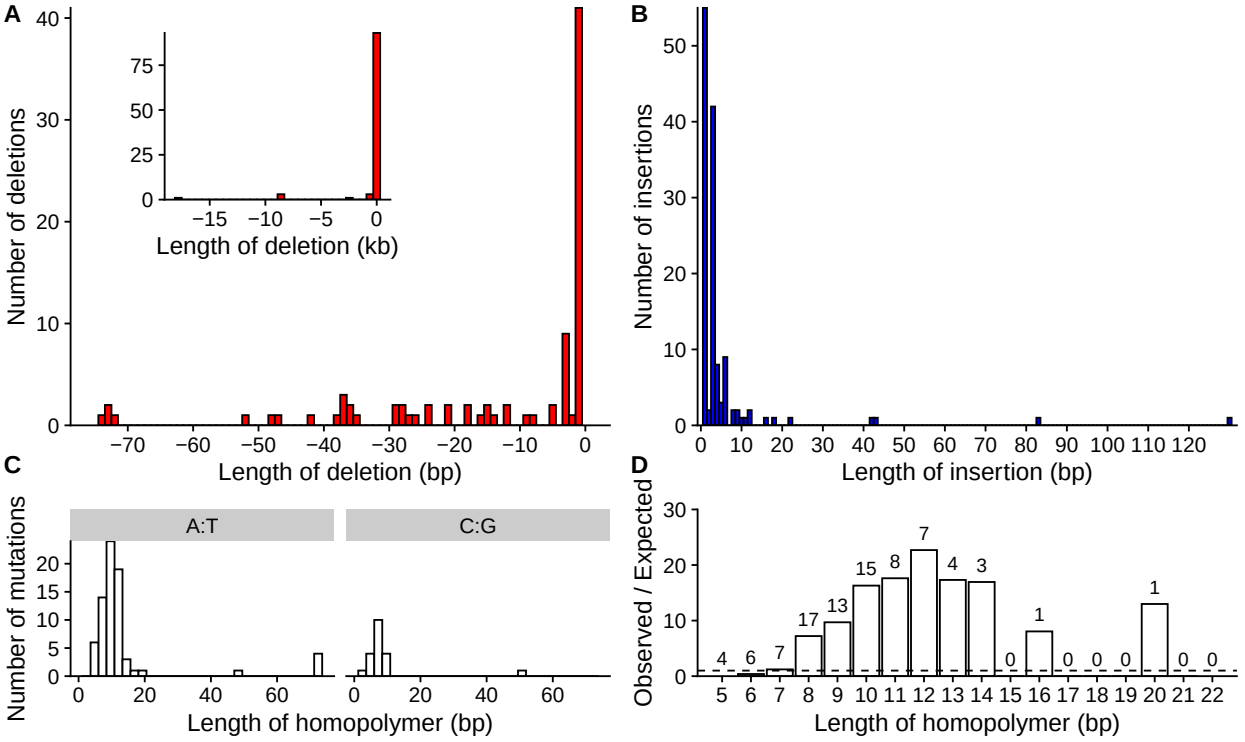

Figure 7: A) Distribution of deletion lenghts in the range of 0 to 75 bp, inset shows that the full distribution that includes few large deletions of several kb. B) Distribution of insertion lenghts. C) Distribution of homopolymer lengths for those mutations that occurred in either A:T or C:G homopolymer streches. D) Relative numbers of mutations in homopolymers of different lengths calculated by taking into account the frequency of homopolymers of different lengths in the genome. Dashed line shows where numbers observed mutations are equal to expected numbers if mutations in homopolymers of different sizes were equally common. Numbers on top of bars are counts of observed mutations. Some rare occurrances in repeats longer than 22 bp are excluded from the figure.

account, mutation rate in A:T homopolymers was 4.55×10^{-7} [3.6×10^{-7}, 5.69×10^{-7}] mutations / locus / TRANSFER compared to rate in C:G homopolymers of 2.06×10^{-7} [1.19×10^{-7}, 2.96×10^{-7}] mutations / locus / TRANSFER. Thus, mutations in A:T homopolymers were 2.21 [1.25, 3.49] times more common than in C:G polymers. We also observed that longer homopolymers had higher mutation rates. The number of mutations that occurred in homopolymers 8 bp or longer greatly exceeded the expected number of mutations based on the occurrance of such repeats in the genome (Figure 7D) under the hypothesis that mutation rate was equal for all homopolymer lenghts. Mutations in homopolymers of 7 bp or shorter were underrepresented relative to their occurrance in the genome. In a model that included both homopolymer sequence and lenght, both of these affected homopolymer mutation rate (Figure S1). This suggest that replication slippage, the mechanism suggested to be involved in indel mutations in repeats, tends to occur more often in longer repeats and threshold of mutation rate increase starts from 7 to 8 bp. Detailed analysis of microsatellite mutations and population variation will be presented elsewhere.

We observed 7 translocations in the MA-lines that we could confirm by PCR and Sanger sequencing. Of these mutations 5 were from one chromosome to another and two occurred among unmapped contigs. One translocation was 2813 bp and the others were length could be determined were around 300 bp. The translocation rate was 0.004 [0.002, 0.008] translocations / genome / TRANSFER. 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 also observed a curious phenomenon were two We observed 9 cases where two SNMs or 1 bp indels where 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, it is unlikely. These changes were more likely caused by a single mutational event. The observed complex mutational events are listed in table 2. To confirm where the changes caused by a single mutational event or two independent events we used Sanger sequencing to check the genotypes of MA-lines from intermediate transfers

during the experiment. We had stored intermediate time points periodically 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 1.37×10^{-10} [5.35×10^{-11}, 2.35×10^{-10}] mutations / bp / TRANSFER. The rate of SNMs over rate of complex mutations was 123.88 [60.69, 238.04], making point mutations over 100-fold more common than complex events.

Discussion

We observed extensive variation across the *Neurospora crassa* genome in mutation rate and mutation spectra. Monroe et al. (2022) observed that mutation rate was lower in genes that were actively transcribed, possibly due to the presence of H3K36 methylation or other epigenetic marks of active transcription. Our dataset does not contain enough mutation to conclusively address if the same phenomenon can be observed in *N. crassa*. However, mutation rate was higher in H3K9 domains and centromeric regions as has been observed for human cancer cells and the plant *Arabidopsis thaliana* (Schuster-Böckler and Lehner, 2012; Weng et al., 2019). Showing that chromatin structure does influence mutation in *N. crassa* and that transcriptionally silent heterochromatin marked by H3K9me has a higher mutation also in this organism.

In *N. crassa* the H3K9me and centromeric regions have much lower GC-content than euchromatic domains, and we cannot unambiguously determine whether increased mutation rate is due to elevated GC-content or because H3K9me regions have increased mutation rate that has lead to low CG-content over evolutionary time. During sexual cycle there is a genome defence mechanism operating in H3K9 regions that causes high rate of G:C $\rightarrow$ A:T transitions (Wang et al., 2020), so GC-content is not necessarely causal.

We did observe additional mechanisms at work in H3K9me regions, in particular adjacent cytosines have high relative mutation rate in these regions. The mechanistic explanation for this is most likely methylation mediated cytosine deamination (Cooper et al., 2010). However, increased mutation rate due to presence of DNA methylation in H3K9 regions cannot account overall the increased mutation rate observed in these regions. Moreover, we observed increased mutation rate

at centromeric regions on top of H3K9me effects, that cannot be accounted with only GC-content.

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 are common then in other species. Alternatively, it may be that C:G $\rightarrow$ T:A are more rare in *N. crassa*.

What makes *N. crassa* peculiar in terms of mutation rate, is that there is a striking different in rate and spectrum of mutation during sexual and asexual reproduction (Wang et al., 2020). During sexual reproduction a genome defence mechanism 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). In sexual reproduction the most common mutation is overwhelmingly the C:G $\rightarrow$ T:A transition, which are induced in H3K9me domains (Wang et al., 2020). In contrast, during asexual reproduction, while H3K9me domains have higher mutation rate, the bias is much smaller.

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. In order quantify which evolutionary forces act on the genome and leave imprints on natural genetic variation, we must account for variation in mutation rates in evolutionary models. Detailed knowledge of mutation rate is required as organisms can have substantial variation in different parts of the genome. This study is the first step in order to do this in *Neurospora crassa*, a model microbial eukaryote.

Acknowledgments

This study was supported by a grant from the Academy of Finland (no. 321584) to IK. We thank the Finnish CSC-IT Center for Science Ltd. for providing computational resources.

Data availability

Sequencing data generated in this study will be deposited to the short read archive, accession numbers XXX-XXX. Data and scripts will be available in Dryad XXXXX.

References

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`.

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.

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.

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

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.

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.

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.

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.

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.

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, 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.

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`.

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`.

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

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14:178–192. URL http://bib.oxfordjournals.org/content/14/2/178.abstract.

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.

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.

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.

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

Supplementary Figures

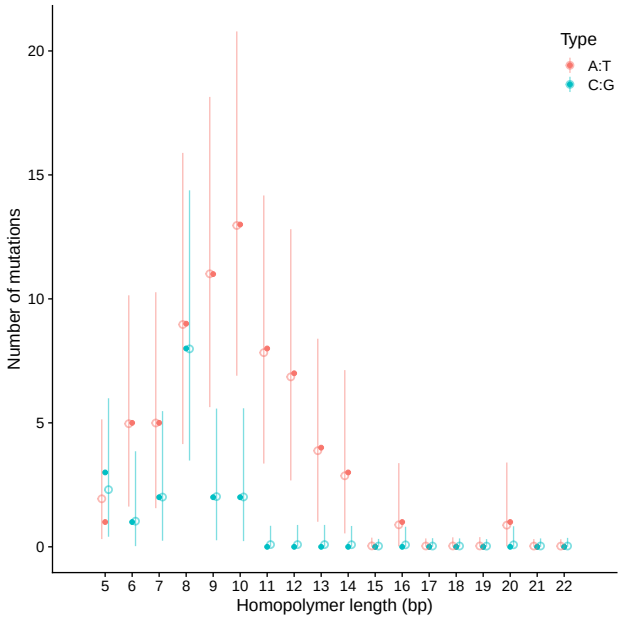

Figure S1: Model estimates of expected numbers of mutations for homopolymers. Estimates were calculated based on estimated mutation rates. Filled points are observed data and open circles are model estimates, with range showing 95% HPD interval.

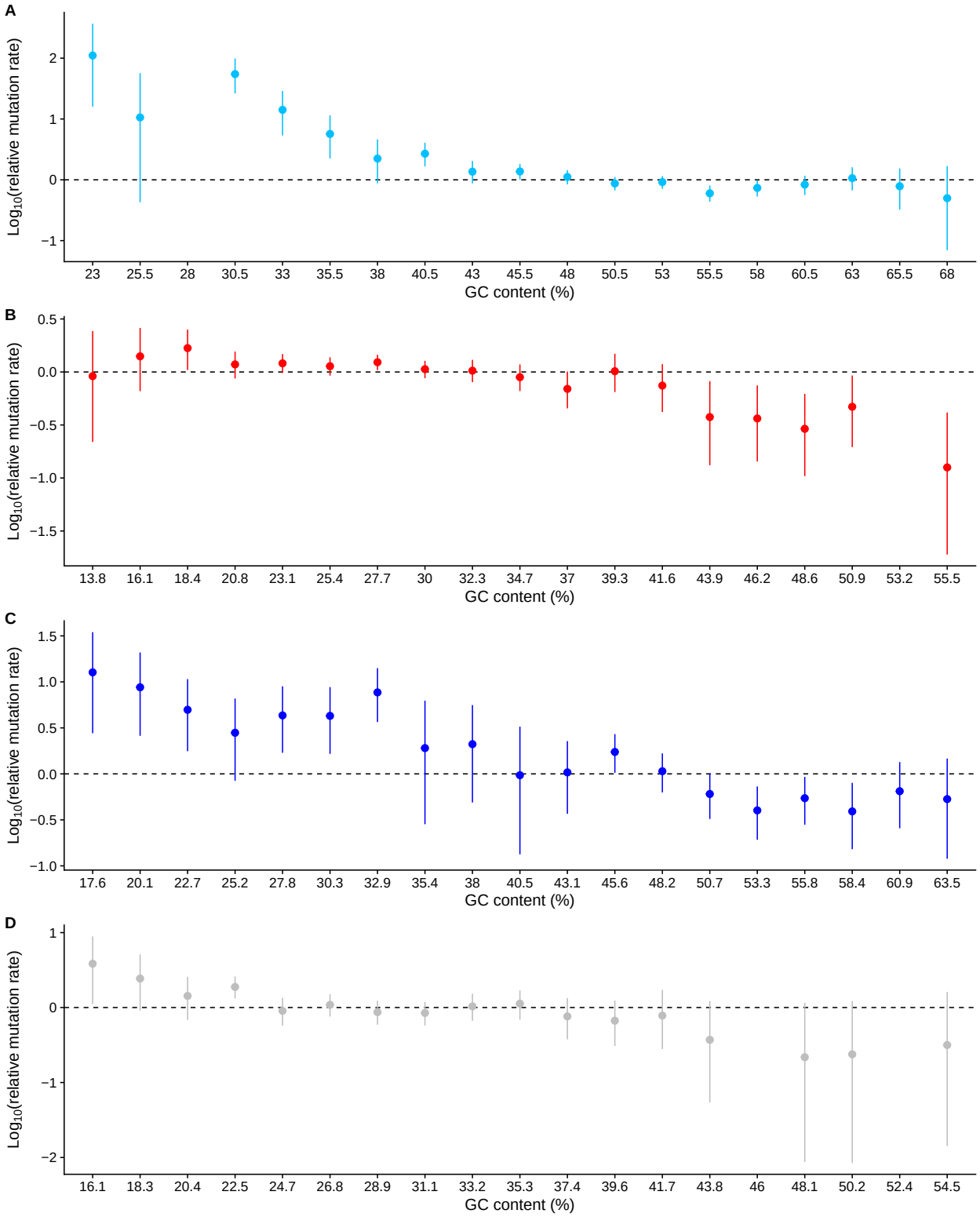

Figure S2: Relative mutation rates for windows of 200 bp binned for GC-content at 2.5 percentage point intervals. Note that y-axis is on $\log_{10}$ scale, dashed line indicates relative mutation rate of one. A) Euchromatic regions B) H3K9 domains C) H3K27 domains D) Centromeric regions.

Supplementary Tables

Table S1: 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 S2: 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 S3: 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.

Model terms	LOOIC	diff ($\pm$ SE)	weight
$\alpha + \beta_{TT}x_i + \beta_{CCT}c_i$	27.69	0 (0)	0.37
$\alpha + \beta_p p_i$	28.62	0.93 (2.57)	0.24
$\alpha + \beta_{TT}x_i$	29.12	1.43 (3.86)	0.18
$\alpha + \beta_{TT}x_i + \beta_{CC}d_i + \beta_{CT}z_i$	29.16	1.47 (1.8)	0.18
α	32.88	5.19 (4.6)	0.03

Table S4: 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 XXX. 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.

	Estimate [95% HPDI]		
Parameter	All mutations	Transitions	Transversions
α	0.65 [0.31, 1]	0.71 [0.5, 0.93]	0.55 [0.05, 1.06]
β_{TT}	0.67 [0.24, 1.1]	0.62 [0.32, 0.93]	0.67 [0.03, 1.31]
β_{CTT}	0.23 [-0.07, 0.52]	0.12 [-0.08, 0.32]	0.33 [-0.11, 0.79]
σ	0.43 [0.3, 0.67]	0.22 [0.09, 0.4]	0.65 [0.45, 1.01]

Table S5: Model comparisons among different models of mutation rate predicted by GC-content and H3K9 domain. Model terms are different linear model parts, α is the intercept, β_{CG} is the slope effect of GC-content, β_H is the effect of H3K9 domain, β_C is the effect of centromeric domain, β_I is the interaction effect between GC-content and H3K9 domain, and β_X is the interaction effect between GC-content and centromeric domain. WAIC = widely applicable information criterion, SE = standard error.

Model terms	WAIC	diff ($\pm$ SE)	weight
$\alpha + \beta_{GC}x_i + \beta_H d_i + \beta_C c_i + \beta_I x_i d_i$	839.73	0 (0)	0.78
$\alpha + \beta_{GC}x_i + \beta_H d_i + \beta_C c_i + \beta_I x_i d_i + \beta_X x_i c_i$	842.23	2.51 (1.72)	0.22
$\alpha + \beta_{GC}x_i + \beta_H d_i + \beta_I x_i d_i$	876.41	36.68 (18.08)	0
$\alpha + \beta_{GC}x_i + \beta_H d_i + \beta_C c_i$	1205.68	365.95 (99.64)	0
$\alpha + \beta_{GC}x_i + \beta_H d_i$	1258.92	419.19 (116.8)	0
$\alpha + \beta_{GC}x_i$	1685.18	845.46 (397.09)	0
$\alpha + \beta_D d_i + \beta_C c_i$	3207	2367.27 (931.36)	0
$\alpha + \beta_D d_i$	3238.54	2398.82 (930.3)	0
$\alpha + \beta_C c_i$	3573.64	2733.92 (858.76)	0
α	3855.25	3015.52 (828.82)	0

Table S6: Model estimates for a model predicting mutation rate by GC-content, centromeric, and H3K9 domains, α is the intercept, β_{CG} is the slope effect of GC-content, β_H is the effect of H3K9 domain, β_C is the effect of centromeric domain, β_I is the interaction effect between GC-content and H3K9 domain.

Parameter	Estimate [95% HPDI]
α	2.37 [2.16, 2.59]
β_C	0.50 [0.35, 0.66]
β_H	-5.10 [-5.45, -4.75]
β_{CG}	-0.16 [-0.16, -0.15]
β_I	0.11 [0.10, 0.12]