Rate volatility and asymmetric segregation diversify mutation burden in mutator cells.

Dowsett, I.T.[a], Sneeden, J. [a], Olson, B.J. [b, c], McKay-Fleisch, J. [a], McAuley, E. [a], Kennedy, S.R. [a], Herr, A.J.* [a]

Author Information

[a]Department of Pathology, University of Washington, Seattle, WA 98195-7705 [b]Department of Statistics, University of Washington, Seattle, WA 98195-7705 [c]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Mutations that compromise mismatch repair (MMR) or DNA polymerase exonuclease domains produce mutator phenotypes capable of fueling cancer evolution. Tandem defects in these pathways dramatically increase mutation rate. Here, we model how mutator phenotypes expand genetic heterogeneity in budding yeast cells using a single-cell resolution approach that tallies all replication errors arising from individual divisions. The distribution of count data from cells lacking MMR and polymerase proofreading was broader than expected for a single rate, consistent with volatility of the mutator phenotype. The number of mismatches that segregated to the mother and daughter cells after the initial round of replication co-varied, suggesting that mutagenesis in each division is governed by a different underlying rate. The distribution of "fixed" mutation counts that cells inherit is further broadened by an unequal sharing of mutations due to semiconservative replication and Mendelian segregation. Modeling suggests that this asymmetric segregation may diversify mutation burden in mutator-driven tumors.

Introduction

All tumors contain genetically divergent cells spawned by the evolutionary processes of mutation and selection. In some tumors, genetic heterogeneity arises from a "mutator phenotype"[1] due to mismatch repair (MMR) defects[2] or heterozygous exonuclease domain mutations (EDM) affecting the leading or lagging strand DNA polymerases (pol), Polε or Polδ [3-9]. Since MMR corrects polymerase errors, when MMR and EDM mutations occur together they produce a dramatic increase in the number of unrepaired polymerase errors. The resulting tumors rapidly evolve and possess "ultra-hypermutated" genomes. Yet a full understanding of the relative contributions of mutagenesis and selection to the rise of heterogeneity within these tumors remains elusive, since cells with more mutations tend to adapt more readily.

A key unanswered question is whether the mutation rate is constant within mutator cell populations. The two most common ways of measuring mutation rates are fluctuation analysis [10] and mutation accumulation lines[11]. Both assume a uniform mutation rate and report the average of hundreds or thousands of cell divisions. However, in recent years, evidence has emerged that mutagenic processes may vary from one division to the next. Kataegis and chromothripsis, for instance, sharply increase mutation burden in a single cell division[12-14]. Indirect evidence for highly mutagenic sub-populations of cells also comes from studies of yeast exposed to 6-hydroxylaminopurine or AID/APOBEC cytosine deaminase. Drug-resistant mutants in these

studies had substantially higher mutation burdens than non-selected isolates from the same population[15]. More recently, limited single-cell propagation of human cancer cell lines coupled to whole genome sequencing revealed broader than expected variation in mutation rate in closely related subclones[16]. Observations such as these challenge the assumption that mutation rate is constant and beg higher resolution studies of mutator cells.

The asymmetrically dividing budding yeast, *Saccharomyces cerevisiae*, is ideal for studying mutator phenotypes with high resolution. It encodes many of the same DNA replication and mismatch repair genes found in humans. Yeast "daughter" cells can be separated from their larger "mother" cell at each division by micromanipulation and then moved to defined locations on an agar plate, forming a "single cell lineage". Whole genome sequencing (WGS) of cultures derived from these cells permits the number of new mutations that arose in the mother cell at each division to be counted. Moreover, the small size of the genome (12 megabases) makes it cost effective to score enough cell divisions to see whether the distribution of mutation counts conforms to that expected from a single underlying mutation rate.

We previously pioneered this approach with haploid mutator mother cells deficient in Polε proofreading and MMR (*pol2-4 msh6Δ*)[17]. A single underlying mutation rate could not explain the distribution of mutation counts from 87 divisions. However, the distribution did fit a model with two underlying mutation rates that differed by 10-fold (0.4 and 4 mutations/genome/division). This led to a hypothesis of "mutator volatility" in which cells assumed one of two mutator states as they passed through the cell cycle[17]. But since we only scored mutations retained by the mother, we could not exclude an alternative hypothesis: that polymerase errors sporadically segregated asymmetrically between mother and daughter cells, either as mismatches at the initial division or as permanent, "fixed" mutations following the next round of synthesis. Here, to distinguish between these two hypotheses, we sought to score all replication errors that arose in individual cell divisions using more extensive single cell lineages. Examination of the distribution of the full replication error counts from individual divisions provided a way to test the mutator volatility hypothesis apart from the confounding influence of segregation. At the same time, sequencing complete lineages gave us the means to determine whether replication errors segregate equally on their way to fixation.

Results

To confidently score replication errors arising on all nascent DNA strands from each division, we devised a scheme that ensured that all mutations were observed in at least two members of a single cell lineage. After moving each daughter by micromanipulation from the founding mother cell, we isolated a sublineage of three additional cells to help score the number of errors segregated to that daughter. These cells included the first and second granddaughter (born to the daughter cell) as well as the first great-granddaughter cell derived from the first granddaughter (Fig.1a). Errors segregated to the daughter as mismatches in the first division segregate as fixed mutations in the next division when the daughter produces the first granddaughter. Mutations retained by the daughter after that segregation event will be inherited by the second granddaughter, forming what we call the "Da" segregant group. Mutations segregated to the first granddaughter will be inherited by the great-granddaughter, forming the "Db" segregant group. In theory, the Da and Db segregant groups represent half of the errors made by the mother cell during a given division. The remaining errors, retained initially by the mother as mismatches, segregate between the mother and her next daughter as fixed mutations in the next division. Fixed mutations segregated to that daughter will be uniquely present in the next sublineage, forming the "Ma" segregant group. Mutations retained by the mother will be found in all later sublineages, defining the "Mb" segregant group. After colony

formation and WGS, a full error count for a given division can be determined by simply summing the number of fixed mutations in the Da, Db, Ma, and Mb segregant groups. With a complete set of sublineages from the same mother cell, the full replication error counts from several sequential cell divisions can be determined from the nested data (Extended Data Fig.1). By requiring that all errors be observed in at least two members of the lineage, this approach eliminates false positives due to sequencing errors or clonal sweeps within the cultures.

We initially began our experiments with the *pol2-4 msh6Δ* haploid strain used in the previous study[17]. We found evidence for a more limited mutator volatility but were concerned that lethality within some sublineages may have introduced a bias (see Supplementary Information and Extended Data Fig. 2). To improve viability and the mutational signal, we switched to using diploid yeast with a 10-fold higher mutation rate due to homozygous mutations affecting Polδ proofreading and base-base mismatch repair (*pol3-01/pol3-01 msh6Δ/msh6Δ*)[18,19]. To obtain *pol3-01/pol3-01 msh6Δ/msh6Δ* cells, we mated *pol3-01 msh6Δ* haploids, freshly dissected from sporulated *POL3/pol3-01 MSH6/msh6Δ* diploids. We isolated the newly formed zygotes and then used the first or second diploid daughters as founding mother cells for the isolation of single-cell lineages, noting the time and placement of each cell. Following colony formation, and WGS, we scored 13,801 mutations from 50 divisions obtained from 7 different lineages (Fig.1b, Table 1, Extended Data Fig. 3). The mutations were distributed across the genome and displayed a spectrum consistent with combined proofreading and MMR deficiency (Extended Data Fig. 4). We only scored mutations at genomic sites confidently called in all members of a lineage and carefully vetted the resulting variant lists. Having complete lineage information allowed us to assign when the mutations arose using the logic described above. In addition, we visually inspected the variant sites in all genomes from a given lineage using the Integrative Genomics Viewer, which allowed us to detect discrepancies in the lineage order or whether mutations had been incorrectly assigned (see Methods). We tallied the full replication error counts from each division and determined whether the distribution could be explained by a single underlying mutation rate.

Mutagenesis has been modeled for more than 70 years[18-20] with the Poisson distribution, which is a discrete probability distribution of the number of expected independent events occurring within a defined interval, assuming a constant rate (λ). A simple test of whether a distribution matches a single Poisson is to calculate the index of dispersion ($\hat{D}$), which is equal to the variance of the distribution divided by the mean (σ^2/μ). The variance of Poisson distributions always equals the mean, which results in a $\hat{D}$ of 1. The *pol3-01/pol3-01 msh6Δ/msh6Δ* mother cells committed an average of 276 (±37.7, standard deviation (σ)) replication errors per division. This corresponds to a $\hat{D}$ of 5.15 ($37.7^2/276$), which suggests that the distribution does not conform to a single Poisson (Fig.1b). Two alternative explanations failed to account for the overdispersion. For instance, we did not observe any relationship between the mother's replicative age and the number of errors made by Polδ (Spearman's rank correlation coefficient: 0.007209, p = 0.9604)(Extended Data Fig.5), nor did the number of mutations correlate with the size of the scored genome, which differed between lineages due to variation in sequencing depth and the number of members in each lineage (Spearman's rank correlation coefficient: -0.0416, p = 0.7743)(Extended Data Fig.5). Instead, the broad distribution of full replication error counts, free from the confounder of segregation, is consistent with mutator volatility.

To better understand the nature of mutator volatility in *pol3-01/pol3-01 msh6Δ/msh6Δ* cells, we used finite mixture modeling, which employs a maximum likelihood framework to identify mixtures of two or more Poisson distributions that better fit the data. We also modeled the data as a negative binomial (nb), which is a discrete distribution with separate rate (μ) and shape parameters (θ) commonly used to interpret over dispersed count data. The rate

parameters λ and μ, for the Poisson and nb distributions, both define the mean number of events. Since these models derive from different distributions, they cannot be directly compared using standard statistical tests. Non-nested models such as these, however, can be evaluated with Akaike Information Criteria (AIC), which uses maximum likelihood to estimate the loss of information of each model relative to the observed distribution. To prevent overfitting, AIC penalizes models with more parameters. Lower AIC values correspond to a more parsimonious fit; however, interpreting differences in raw AIC values can be enigmatic. Thus, we transformed the raw AIC values to "Akaike weighted values", which conveys their relative likelihood (Fig.1b)[21,22]. We found that the negative binomial model was the most likely (relative likelihood of 0.9999), followed by the two-Poisson-mixture model (2.2×10^{-6}), and the single Poisson (4.3×10^{-28}) (Fig.1b). Similar results were obtained using Bayesian Information Criteria (BIC), which imposes stronger penalties for overfitting. Thus, mutator volatility in *pol3-01/pol3-01 msh6Δ/msh6Δ* cells is more complex than just two distinct mutator states.

The superiority of the negative binomial model suggests that the mutator phenotype may vary continuously. This rationale derives from the ability to describe a negative binomial as a gamma-Poisson distribution (Fig.2a). The gamma function is a continuous, rather than discrete, distribution. Here, it takes the same shape parameter (θ) as the negative bionomial and serves as a conjugate-prior to define variation in the rate parameter λ of a mixture of Poisson distributions. The variation in λ that creates a negative binomial occurs between replication events at the same site, or a collection of sites such as a chromosome or genome. Having complete lineage information provided an opportunity to test whether λ varies at a chromosomal or genome-wide level. The distributions of mismatches segregated to mother (Mm) or daughter cells (Dm) across all divisions were the same and fit a negative binomial (Fig. 2b). If λ varied widely during the replication of individual replicons (the units of DNA replication on a chromosome), this could introduce asymmetry in the number of errors on sister chromatids, which would then propagate to the daughter and mother cells (Fig. 2d). Consequently, Dm and Mm from the same division would be free to vary within the observed negative binomial distribution. Alternatively, if the genome-wide value for λ varies between cell divisions, a single mutation rate would govern mismatch formation for both the mother and daughter genomes (Fig. 2e). Dm and Mm would co-vary within the constraints of the corresponding Poisson distribution. To distinguish between these two hypotheses, we first compared the correlation of mismatches segregated to mother and daughter cells to simulated data generated under the constraints of the two models. While no correlation was seen between Dm and Mm in the simulated data from the first model (R^2=0.001), similar correlations were observed for both the simulated data from the second model (R^2=0.47) and the actual data (R^2=0.37). This correspondence in the number of mismatches segregated to mother and daughter cells extended down to the level of chromosomes (Fig. 2f). The R^2 values are lower than typically seen with strong correlations, but as our modeling shows, this is expected since both X and Y values are randomly drawn from a Poisson distribution. As a second test of the hypotheses, we also performed 10,000 simulations of how each model would affect the distribution of full replication error counts from 50 divisions (Fig. 2g). With the first model, the simulated index of dispersion (3.28 ± 0.66, σ) was substantially less than observed with the actual data ($\hat{D}$=5.15), while the second model produced a good match (5.54 ± 1.12, σ). Together, these analyses strongly suggest that the source of mutator volatility is variation in the genome-wide mutation rate from one division to the next.

With this support for the mutator volatility hypothesis, we turned our attention to the question of asymmetric inheritance. Individual cells averaged 69 (±18, σ) fixed mutations/diploid

genome/division (*n* = 200) (Fig. 3a) with an index of dispersion of 4.8. A negative binomial fit the distribution most closely (relative likelihood = 0.82), followed by a four-Poisson mixture model (relative likelihood = 0.18). A close examination of mutations arising from the same division revealed a striking asymmetric pattern of inheritance. When pairs of segregant groups were compared (e.g. Da vs Db or Ma vs Mb), half of the time one segregant group inherited all of the mutations for a given chromosome while the other received none (Fig. 3b,c). This pattern is explained by the sequential actions of semiconservative DNA replication and Mendelian segregation (Extended Data Fig.6). At the end of the first S-phase, due to semiconservative replication, all errors arising due to the Poisson process of polymerase error formation reside on one of the two strands of each sister chromatid. These strands segregate equally between mother and daughter cells. The next round of replication produces two new duplexes per cell, only one of which contains fixed mutations. At metaphase, cells receive either all or none of the fixed mutations for that chromosome from the previous division. This binomial process occurs twice for every chromosome number in diploid cells. Consequently, for each chromosome number, cells receive 0%, ~50%, or 100% of the mutations in a given division with a Mendelian ratio of 1:2:1 (Fig. 3c) (actual ratio, 876:1490:834). Thus, we can describe how polymerase errors arise in an individual division and later become fixed as a compound Poisson-binomial process.

To determine the contribution of the Poisson-binomial process to the overdispersion of mutation counts, we simulated mutagenesis in *pol3-01/pol3-01 msh6Δ/msh6Δ* cells assuming a constant error rate. Given that we observed an average of 138 mismatches per diploid mother or daughter cell (Fig. 2c), the average rate of error formation was 69 errors/haploid genome/division. Since cells only inherit, on average, half of the polymerase errors, the observed mutation rate in *pol3-01/pol3-01 msh6Δ/msh6Δ* cells was 34.5 fixed mutations/haploid genome/division. To model the Poisson-binomial process we simulated mutagenesis on each chromosome by setting λ equal to 69 errors/haploid genome and then, to mimic segregation, multiplied the number of mutations apportioned to each chromosome by a randomly chosen 1 or 0, before summing the total fixed mutations (Fig.3d). For comparison, we simulated mutation accumulation assuming a simple Poisson process in which mutations accumulated with a rate of 34.5 mutations per haploid genome (Fig.3d). With 1000 simulations of 200 cell cohorts, the Poisson-binomial model produced a broader index of dispersion ($\hat{D}$ = 3.58±0.49, σ) than the Poisson model ($\hat{D}$ = 1.0±0.1, σ) (Fig.3e), but narrower than the observed data ($\hat{D}$ = 4.8). However, substituting the constant mutation rate with the gamma-distributed set of λ values from Fig.2c yielded simulated data with an equivalent dispersion ($\hat{D}$ = 4.80±0.49, σ) (Fig. 3e). Thus, the combination of mutator volatility and asymmetric segregation of mutations — a gamma-Poisson-binomial process —accounts for the observed distribution of fixed mutations in individual *pol3-01/pol3-01 msh6Δ/msh6Δ* cells.

To understand the potential implications of our findings for mutator-driven cancers, we first focused on how the Poisson-binomial process would influence the heterogeneity of mutation burden within a dividing population of tumor cells. Assuming a constant mutation rate comparable to *pol3-01/pol3-01 msh6Δ/msh6Δ* yeast, the expected distribution of simulated mutation counts in human cells after one division ($\hat{D}$ = 50) was far broader than in yeast (Fig.3f) and persisted through 30 simulated divisions (Fig. 3g,h). Adding a comparable level of volatility to the mutator phenotype further increased the simulated dispersion ($\hat{D}$ = 82) (Fig.3f). Using the Poisson-binomial model, we simulated a range of mutator phenotypes observed in cancer cells and found a linear relationship between mutation rate and predicted dispersion. For instance,

mutation accumulation in HCT116, the well-known MLH1 mutant colon cancer cell line, increases from 48 to 190 mutations/haploid genome/division upon introduction of a heterozygous *POLE* proofreading-deficient allele[9]. In these cells, the predicted index of dispersion expanded from 3.4 to 10.8 (Fig.3i). Even greater heterogeneity may arise in human cancers when more potent *POLE* mutator alleles occur in combination with MMR deficiency[5,7,23,24]. Thus, the fundamental Poisson-binomial process of asymmetric segregation has the potential to dramatically expand the diversity of mutation burdens present among a population of human mutator cells.

Discussion

Genetic heterogeneity progressively increases in a dividing population of cells as an unavoidable consequence of errors made during DNA synthesis. Here, for the first time, we describe the fate of polymerase errors made on all nascent DNA strands synthesized in individual cell divisions. We developed this single cell resolution approach in order to understand previous observations that the distribution of new fixed mutations in individual mutator cells was broader than expected. To explain the phenomenon, we proposed two hypotheses: (1) that mutator phenotypes are volatile and (2) that polymerase errors arise with a constant rate but segregate asymmetrically on the way to fixation. The design of our single cell pedigrees ensured at least two independent biological observations for each mutation, which allowed us to confidently assign more than 13,000 mutations to fifty divisions. From the resulting mutation count data, we found strong evidence that both mutator volatility and asymmetric segregation significantly expand genetic heterogeneity in *pol3-01/pol3-01 msh6Δ/msh6Δ* yeast.

Historically, mutagenesis has been modeled with the Poisson distribution, which describes the probability of the number of independent events per unit time given a constant rate. The observed distribution of full replication error counts of mutator cells, free from the influence of segregation, best a fit a negative binomial and not a single Poisson (Fig.1b). Negative binomials are equivalent to a continuous mixture of Poisson distributions whose rates vary according to a gamma distribution (Fig.2a). This suggests that mutator volatility may create a continuum of mutation rates rather than discrete mutator states. We explored the idea that mutation rate varies from one division to the next by simulating the number of mismatches segregated to mother and daughter cells (Fig 2d,e) and the dispersion of full replication error counts expected from small cohorts of cells (Fig.2f). Both simulations closely matched the observed data, consistent with the hypothesis that mutator volatility derives from continuous variation in mutation rate between divisions. Mutator polymerases do not operate as a closed system. They interface with a myriad of other replication components and metabolites, such as dNTPs, that influence their fidelity[25,26]. Variation in the timing and duration of perturbations to these interactions may produce a continuum of rates. The observed overall mutation rate that cells exhibit represents a composite of mutation rates at all sites within the genome. Conceivably, the change in replication fidelity could be localized to certain parts of the genome in a given division. But if so, our data suggests, that the nascent strands from each pair of sister chromatids in the affected region must be equally influenced by the change in rate (Fig.2c,f).

The asymmetric inheritance of mutations observed in mutator cells results from the fundamental processes of semi-conservative replication and Mendelian inheritance acting in concert. Current models of mutation accumulation generally ignore the potential for this synergy to expand genetic heterogeneity, although there are exceptions. John Cairns proposed a far

more extreme asymmetric inheritance of mutations in the "Immortal Strand Hypothesis" in which stem cells always segregated away newer DNA duplexes with fixed mutations[27]. In keeping with this hypothesis, a recent computational analysis of human somatic variants argued that the high variance of mutation burden in adult stem cells with age supports a preferential inheritance of ancestral strands[28]. A second study from the field of evolutionary biology examined the potential influence of disparate mutagenesis of leading and lagging strand synthesis to promote variable evolutionary trajectories from the same cell population[29]. Our findings here demonstrate that, in the context of a mutator phenotype, the *normal* process of semi-conservative replication and Mendelian inheritance has the potential to create unequal sharing of mutations. For every cell that inherits disproportionately more mutations there will be another cell with fewer mutations. The predicted impact of this process on the variation in mutation burden is larger in human cells than in yeast due to the vast differences in chromosome length, and the correspondingly larger number of fixed mutations per chromosome. However, with longer chromosomes comes an increased likelihood that sister chromatid exchanges (SCEs) may mitigate the asymmetry. SCEs clearly to do not homogenize mutation burden in diploid mutator yeast cells as half of cells either received all or none of the new fixed mutations for a given chromosome (Fig. 3c). This finding is in keeping with recent evidence from a sensitive Next Generation Sequencing methodology (Strand-seq) that SCE occurs with a rate of 0.26 events/division in yeast[30]. Strand-seq experiments of normal human fibroblasts and lymphoblasts indicate the SCEs occur with a rate of 5 events/cell division[31]. At this rate, most chromatid pairs in mutator cells would be free of SCEs even after the two divisions it takes for errors to become fixed mutations. Of course, the frequency of SCEs may increase in some cancer cells, especially those with intrinsic DNA repair defects [31]. However, the mutator yeast strains studied here do not show obvious signs of elevated SCEs. Performing single cell lineage analysis of human mutator cells in future studies should address both the prevalence of SCEs and the asymmetric inheritance of mutations.

Our simulation of a mutator-driven tumor rapidly generated substantial intra-tumoral genetic heterogeneity during expansion (colored lines, Fig.3h) compared to a population in which mutations accumulated by a simple Poisson process (black line, Fig.3h). The associated variability in mutation load may be relevant to cancer evolution. Early during tumorigenesis the subpopulation of cells that inherit disproportionately more mutations may adapt more readily. With elevated mutation rates, polyclonal adaptation is almost certain. The unifying feature of these adapted cells is a high mutation burden. As mutation burden mounts and mutator cells contend with increasingly strong negative selection pressure due to immune surveillance and negative epistatic interactions[32,33], adapted cells that inherit fewer new mutations due to asymmetric inheritance may be at a relative fitness advantage. Selectively increasing mutation rate in mutator cancer cells could represent a novel therapy[25]. If, as a means of treatment, the mutation rate of cancer cells is only transiently elevated to induce extinction, this subpopulation may persist. Sustained elevation of mutation rate over many divisions of mutator cells may be required to drive their extinction.

Methods

Yeast strains and culture conditions.

The diploid strains AH2801 (*POL2/URA3::pol2-4 MSH6/msh6Δ::LEU2*) and AH2601 (*POL3/URA3::pol3-01 MSH6/msh6Δ::LEU2*) were previously described[17,34]. They are derived from AH0401, a BY4743 derivative engineered to be heterozygous at the *CAN1* locus (*CAN1::natMX/can1Δ::HIS3*) to facilitate forward mutation rate assays[32]. We followed standard

procedures for yeast propagation and tetrad dissection[35]. For general propagation, we grew liquid YPD cultures (1% wt/vol yeast extract, 2% wt/vol peptone, 2% wt/vol dextrose) at 30°C. For sporulation, we diluted overnight YPD cultures 1:100 in 3 mls of YPD and grew until the culture was 1-2 x 10^7 cells/ml. We recovered the cells by centrifugation, resuspended and pelleted the cells once in 1 ml H_2O, and then resumed growth at 22-25°C in sporulation media (1% potassium acetate, 0.1% yeast extract, 0.05% dextrose) for five days. For rich solid media, we used synthetic complete (SC) [6.7 g Difco yeast nitrogen base without amino acids, 2% wt/vol dextrose, 2 g/L SC amino acid Mix (SCM) (Bufferad)] supplemented with 2% wt/vol agar. For plates lacking leucine and uracil (SC-Leu-Ura), SCM was substituted for SCM-Leu-Ura (Bufferad). Archival frozen stocks were stored in 23% glycerol at -80°C.

Single cell lineage isolation

To isolate *pol2-4 msh6Δ* lineages we dissected AH2801 tetrads on SC-Leu-Ura selective media and chose one germinating spore per plate to serve as the founding mother cell. To obtain *pol3-01::URA3/pol3-01::URA3 msh6Δ::LEU2/msh6Δ::LEU2* cells for pedigree analysis we first dissected *POL3/ pol3-01::URA3 MSH6/ msh6Δ::LEU2* tetrads on SC-Leu-Ura plates. After two divisions, double mutant haploid cells from different tetrads were placed next to each other to allow mating. Upon isolation of a zygote, the first or second daughter was used as the founding mother (M) for the lineage. Mothers were placed at an isolated location and we separated daughter cells (designated Dn, Dn+1, etc.) from the mother as they were generated and moved them to select areas 5 mm apart on the plate. We repeated the procedure to obtain each daughter's first daughter (GD.1, Fig.1b), second daughter (GD.2), and first granddaughter (GGD, born to GD.1). This strategy was repeated for each daughter up to either the 20th division or the end of the mother's replicative lifespan, whichever occurred first. In a typical experiment, we pre-punched the agar with the dissecting needle at each drop-off location so that we would always put the cell in a defined place, making it easy to later find the cell for inspection and manipulations. We isolated lineages over the span of a week by performing rounds of dissections every 90-120 minutes. Only a few cells on a plate were moved in any one round, and then, only one cell at a time. We noted the timing of each round of bud dissections. We incubated plates at 30°C between dissections. At the end of the day, plates were wrapped in parafilm and stored overnight at 4°C. When plate dissections were concluded, we incubated each plate an additional 48 hours at 30°C to allow colonies to fully develop. Prior to sequencing, the *pol3-01/pol3-01 msh6Δ/msh6Δ* and *pol2-4 msh6Δ* genotypes were confirmed by previously described allele-specific PCR assays[34].

Genome sequencing

We performed whole genome sequencing of yeast cultures as described[34]. Briefly, each colony in a pedigree was used to inoculate overnight 5 ml liquid YPD cultures. Glycerol stocks were made and genomic DNA extraction extracted with the ZR Fungal/bacterial purification kit (Zymo Research). DNA was sheered into 500 to 1000 bp fragments by sonication. After end-repair, Illumina sequencing libraries were made by ligating on dsDNA adapters and indexing by quantitative PCR. The samples were then sequenced on the HiSeq 2500 or Nextseq platforms. We performed sequencing alignments and variant calling using a custom pipeline (eex_yeast_pileline.sh) that runs in the Unix command-line (see Github link below). Reads were aligned to a repeat-masked S288C yeast genome[17] using the Burrows-Wheeler Aligner (0.7.17)[36]. Discordant and split read groups were removed using Samblaster (0.1.24) [37]. We used Picard tools (2.21.9) AddorReplaceReadGroups to add information to the header used for later steps. We indexed the BAM files with Samtools (1.8)[38] and then sequentially processed them with functions from the Genome Analysis Toolkit (GATK3)[39] to minimize false variant calls: RealignerTargetCreator, IndelRealigner, LeftAlignIndels, BaseRecalibrator, and PrintReads. We made a pileup file with Samtools and used VarScan (v2.3.9) mpileup2snp to call single

nucleotide variants[40]. We limited our analysis to single nucleotide variants, which are by far the most abundant polymerase error type in these cells. We used the Varscan2 tool to identify variants present in our colonies with the following parameters. For *pol2-4 msh6Δ* haploid lineages we used a variant frequency cut-off of 0.8 with a minimum read-depth of 18 (daughter and GD.1 positions) or 10 (for GD.2 and GGD positions). Since these are haploid cells, new variants should be present in 100% of reads. Setting the cut-off at 0.8 accommodates sites with low read depth and one sequencing error. For *pol3-01/pol3-01 msh6Δ/msh6Δ* diploids, we used a minimum read-depth of 18 for all strains and a variant frequency cut-off of 0.22. With a read depth of 18, clonal heterozygous variants in diploid cells have a false negative rate of 6.1×10^{-5}. With 1000 mutations we have a 6% chance of having 1 false negative in a genome. We filtered the above results to remove variants present in the parental strains as well as recurrent sequencing artifacts. A small number of variants (<0.1%) could be reliably scored with the above parameters but fell below a quality threshold for a subset of genomes. These were manually curated for inclusion. We detected these by visually inspecting the BAM files for all strains in a single cell lineage at the same time using the Integrated Genome Viewer (IGV).

Scoring of mutations and detection of assignment errors

We used a custom Python script (JLSLineageCaller) to determine the number of shared variants within each lineage. The program first determines all genomic positions with 18-fold read-depth in all members of the lineage and then filters the called variant lists for mutations at positions within the shared genome. Pairwise comparisons are done between certain strains to identify shared mutations at different branch points in the lineage, resulting in a data-frame of comparisons that allows all mutations arising in a lineage to be sorted and examined in Microsoft Excel. The mutation counts for division *n* were determined by summing the number of new mutations identified at branchpoints Da (GDn.1 vs GGDn.1), Db (Dn vs GDn.2), Ma (Dn+1 vs GDn+1.1), and Mb (Dn+2 vs Dn+3). Da mutations are only found in the daughter (Dn) and her second daughter (GDn.2). Likewise, Db mutations are only found in GDn.1 and her first daughter GGDn.1. Mismatches retained by the mother after the first division become fixed in the next division and are either passed on to her next daughter (Dn+1) or are retained by the mother and passed on to all future offspring. The fixed mutations inherited by Dn+1 that form the Ma segregant group are only found in this branch of the lineage. Finally, the fixed mutations retained by the mother, the Mb segregant group, first appear in Dn+2 and her offspring, but also show up in all subsequent daughters (Dn+3, Dn+4, etc) and their offspring. Any deviation from this pattern of inheritance indicates an “assignment error” has occurred, and that a cell was inadvertently placed in the wrong position in the lineage. In the Supplementary Information we describe two such cases. The divisions encompassing these strains were censored from the analysis. Below we describe how these errors arise and are detected to illustrate the reliability of the method.

One possible assignment error could occur at dissection when the daughter and mother cells both divide before the next round of dissection. On the basis of size, the first daughter (Dn) can be easily distinguished from the mother, the second daughter (Dn+1), and her own daughter (GDn.1). Usually Dn+1 and GDn.1 can also be distinguished because Dn+1 buds before GDn.1. However, in rare cases Dn+1 and GDn.1 are adjacent and similarly sized. If Dn+1 is moved in place of GDn.1, we will have a sublineage consisting of Dn, Dn+1, GDn.2, and GDn+1.1 (instead of Dn, GDn.1, GDn.2, and GGDn.1). Every sublineage should normally contain subsets of mutations from different divisions (Da and Db mutations from the “n” division; Ma mutations from the “n-1” division; and Mb mutations from the “n-2” division). In this sublineage, the Ma segregant group mutation count will be 0, since there are no *new* mutations that will be shared by these four colonies. However, a substantial subset of the mutations assigned to the Db segregant group will also be found in later sublineages indicating that they

are *not* Db mutations but Mb mutations from a later division. The other half of what appear to be Db mutations will in fact be Ma mutations from a different division. Added confirmation of the dissection error comes from the analysis of the next sublineage, which will consist of GDn.1 (not Dn+1 as it should be), GGDn.1, GGDn.2, GGGDn.1 (great-great-great granddaughter 1). There will be 0 Mb mutations in this sublineage since all of these cells are directly descended from Dn. These problematic cell divisions would be censored because we lack key lineage members necessary to obtain a full replication error count. Another type of assignment errors could occur during dissections to isolate the sublineages. For instance, if Dn divides twice in the interval before the next round of dissection we would have to distinguish between GDn.1 and GDn.2. This is usually easy to do because, as above, GDn.1 would be forming a bud while GDn.2 would be unbudded. If we inadvertently reversed those two cells, we would have a sublineage consisting of Dn, GDn.2, GDn.1, and a great granddaughter born to the second granddaughter. When calling the Da segregant group we would be calling shared mutations between Dn and GDn.1 (and not between Gn and GDn.2). We would quickly see that these are, in fact, Ma segregant group mutations because they would also be present as a subset of Db mutations obtained in the comparison between GDn.2 and her offspring.

The most difficult potential assignment errors to detect would occur in the Da and Db segregant groups. For example, if GDn.1 divided twice, producing GGDn.1 and GGDn.2, and we selected GGDn.2 instead of GGDn.1, the mutation count for the Db segregant group would be derived from two divisions instead of one. Again, this is unlikely, because GGDn.1 would begin budding long before GGDn.2. But we lack an obvious distortion to the pattern of mutation inheritance to flag this as an error. We don't think this is a common problem given the correspondence between mismatches segregated to the mother (Mm) and daughter (Dm) cells illustrated in Fig.2d,e. As described above, we regard the Ma and Mb segregant groups as highly reliable because dissection errors lead to obvious perturbations in the pattern of mutation inheritance. In favor of the reliability of the Da/Db data, an XY scatter plot of mutation counts observed in pairs of Ma/Mb segregant groups corresponds very well to that observed with pairs of Da/Db segregant groups (Extended Dataset Fig. 7). Both sets also correspond with what would be expected based on simulated data. (The simulation assumed a gamma-Poisson distribution as in Fig. 2). Interestingly, there are two Ma/Mb (47,36) and Da/Db (52, 20) segregant pairs in the lower left-hand quadrant that appear as outliers. Both pairs are derived from Division 15 (see Supplementary Dataset 1), leading to the conclusion that the mutation rate in that division was inherently low. The highest Da/Db outlier (51,120), derived from Division 8, is also associated with a Ma/Mb pair with high mutation counts (120,65), leading to the conclusion that this division had a high mutation rate.

Statistical modeling

We grouped the fixed mutation counts from the above branch points into Da, Db, Ma, and Mb segregant groups to determine their distributions. We also joined all segregant groups into one larger group to examine the distribution of fixed mutation counts across all cell divisions. To determine the distributions of mismatches segregated to the daughter (Dm) and mother (Mm) cells, we first summed the Da and Db or Ma and Mb fixed mutation counts from each division. We also combined these two sets into one group to view the distribution of mismatches across all cell divisions. To determine the distribution of total polymerase errors per division, we summed all fixed mutations from individual divisions (Da+Db+Ma+Mb). We considered two common approaches for modeling over dispersed count data: the Poisson mixture distribution and the negative binomial distribution.

A *K*-component Poisson mixture distribution, which we denote PM(*K*), has a probability mass function (pmf) given by

(1)

$$f_{\mathrm{PM}}(\mathrm{x}; K, \mathbf{p}_K, \boldsymbol{\lambda}_K) := \sum_{k=1}^{K} p_k f_{Poisson}(x; \lambda_k)$$

where $\mathbf{p}_K$ = (p_1,…,p_K) is a vector of mixture proportions, $\boldsymbol{\lambda}_K$= (λ_1,…, λ_K) is a vector of Poisson means, and f is the pmf of a Poisson(λ_k):

(2)

$$f_{Poisson}(x; \lambda_k) := \frac{\lambda_k^x e^{-\lambda_k}}{x!}.$$

From this formulation, we see that the full density of the distribution is decomposed as a sum of the scaled Poisson densities. In (1), p_k represents the prior probability that a given count measurement will be generated from the *k*th Poisson component distribution, parameterized by λ_k. Since a given count measurement could have been generated from any of these *K* components, we average over their densities based on their prior probabilities to get the full density of that count.

The negative binomial distribution can be specified by the following probability mass function:

(3)

$$f_{NegBinom}(x; \mu, \theta) := \frac{\Gamma(x+\theta)}{x!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\theta}{\theta+\mu}\right)^{x}$$

where μ is the rate parameter and θ is the shape or dispersion parameter. As θ tends towards zero, the variance increases. As $\theta \to \infty$, the negative binomial reduces to a Poisson distribution.

We implemented these principles using a single R script (FMM.R, see Github link below).To fit Poisson mixture models we used the flexmix R package in R v3.5.3[41]. To fit negative binomial models we used the glm.nb function of the MASS R package[42]. Goodness of fit testing of the models was performed using both Akaike information criterion (AIC) and Bayesian information Criterion (BIC) in R. Although these two approaches score fit in slightly different ways, BIC returned results consistent with AIC and we thus report only the more commonly used AIC scores. We scored each tested distribution against up to 4 parameters. We reported only up to the number of parameters that improved model fit. Lower raw AIC values indicate better fit; however, the relative differences are not immediately intuitive and so we calculated Akaike weighted values as described[21,22]. To illustrate this approach, the AIC values in Fig.1b were 637, 537, and 511. The first step in getting weighted AIC values is to determine Δ_iAIC: the difference between each AIC value and the AIC with the lowest value (so for these numbers: 126, 26, 0). The likelihood of each is then calculated by exp(-1/2 x Δ_iAIC). The weighted AIC value for a given model is its likelihood divided by the sum of all competing likelihoods. From these calculations the weighted AIC values are 4.3e-28 (P, k=1), 2.2e-6 (P, k=2), and 0.9999978 (nb), respectively. Thus, the negative binomial model is far more likely than the other two models to account for the observed data. Mixture model graphs were constructed using the ggplot2 package R[43]. Spearman rank correlation coefficients were calculated using the Scipy Stats package in Python and graphs generated with Seaborn 0.9.

Simulation of negative binomial models

We wrote a Python script (Fig2a-d.py, see Github link below) to simulate the expected correlation between Dm and Mm under two distinct models of mutagenesis (Fig.2). The script uses the θ (60.42) and μ (138) parameters estimated by glm.nb for the negative binomial model of mismatches segregated to mother (Mm) or daughter (Dm) cells (see FMM.R). (Note that glm.nb actually returns the natural log value for μ (in this case 4.927), which must be exponentiated ($e^{4.927}$) to get 138). In the first model, we assumed that the negative binomial

distribution was created by variation in mutation rate along chromatid pairs, so that upon segregation, Dm and Mm from the same division were free to vary within the predicted negative binomial distribution. To simulate this process with Scipy.stats.nbinom.rvs, we converted the θ and μ shape parameters to the n and p inputs (see script for details) for nbinom.rvs and then, for each division, we selected two random values from the distribution to represent the Dm and Mm counts. In the second model, we assumed that the negative binomial was created by a gamma distribution of λ values for a series of Poisson processes acting in different cell divisions. We used Scipy.stats.gamma.rvs to simulate λ values from a gamma distribution with shape and scale parameters derived from those of the negative binomial. The shape parameter for the gamma distribution is simply equal to θ. With variance (v) equal to μ^2/θ, the scale parameter is equal to v/μ. With a random λ from the gamma distribution as an input for Scipy.stats.poisson.rvs, we selected two values from the associated Poisson distribution to serve as Dm and Mm counts for each division. To examine the relationship between Dm and Mm in these different models and the actual data, we performed linear regression with Scipy.stats.linregress and visualized the data and regression line using Seaborn 0.9 regplot.

Simulation of gamma-Poisson-binomial process

We created Python scripts to create a Poisson-binomial model of the contributions of semi-conservative DNA replication and Mendelian segregation to the over-dispersion of fixed mutations in individual yeast (Fig3de.py) and human cells (Fig3f-i.py). For yeast simulations, we determined the amount of unmasked DNA on each chromosome in the repeat-masked genome and then divided these values by the total length of unmasked DNA in the haploid genome. The rate of mismatches per haploid genome (69 mismatches/haploid genome/division for *pol3-01/pol3-01 msh6Δ/msh6Δ* cells) was then multiplied in each case by these fractions to obtain the per chromosome rate of mismatch formation. These values were used as input for scipy.stats.poisson.rvs to simulate the number of errors per chromosome in a single division. We created two independent entries per chromosome to model the diploid genome. To mimic the binomial process of Mendelian segregation, we then multiplied the number of simulated errors on each chromosome by a randomly chosen 1 or 0. Finally, we summed the mutation counts from all chromosomes to obtain the total number of fixed mutations per cell. To create a gamma-Poisson-binomial model, we selected a value for lambda at each division from the gamma distribution described in Fig.2 rather than using a constant rate for mismatch formation. As a control we performed the above simulation without the binomial process, using the rate of fixed mutations per haploid genome (34.5 fixed mutations/haploid genome/division). We used the same approach for the human simulations except that we multiplied the fraction of each human chromosome of the total genome (GRCh38) by a mismatch rate comparable to that observed with *pol3-01/pol3-01 msh6Δ/msh6Δ* yeast: 69 mismatches/ haploid yeast genome/division x (3.03×10^9 bp/human haploid genome / 11×10^7 bp/yeast haploid genome) = 1900 mismatches/human haploid genome/division. We compared the resulting distribution to that from a Poisson distribution with λ equal to 950 fixed mutations/haploid genome. To simulate the diversity in mutation burdens that this process generates, we summed the simulated mutation counts for individual lines from 30 divisions.

Data Availability

Sequence data used to generate the findings of this study have been deposited in the NCBI Sequence Read Archive (SRA), BioProject accession: PRJNA586886.

Scripts used to generate figures and perform statistical tests have been deposited to github: https://github.com/idowsett/Asymmetric-segregation-of-polymerase-errors-and-rate-volatility-diversify-mutation-burden

References

1 Beckman, R. A. & Loeb, L. A. Evolutionary dynamics and significance of multiple subclonal mutations in cancer. *DNA repair* **56**, 7-15, doi:https://doi.org/10.1016/j.dnarep.2017.06.002 (2017).

2 Lynch, H. T. *et al.* Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clinical Genetics* **76**, 1-18, doi:10.1111/j.1399-0004.2009.01230.x (2009).

3 Network, C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

4 Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics* **45**, 136-144, doi:10.1038/ng.2503 (2012).

5 Church, D. N. *et al.* DNA polymerase ε and δ exonuclease domain mutations in endometrial cancer. *Human Molecular Genetics* **22**, 2820-2828, doi:10.1093/hmg/ddt131 (2013).

6 Yoshida, R. *et al.* Concurrent genetic alterations in DNA polymerase proofreading and mismatch repair in human colorectal cancer. *European Journal of Human Genetics* **19**, 320-325, doi:10.1038/ejhg.2010.216 (2011).

7 Kandoth, C. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

8 Barbari, S. R. & Shcherbakova, P. V. Replicative DNA polymerase defects in human cancers: Consequences, mechanisms, and implications for therapy. *DNA repair* **56**, 16-25, doi:https://doi.org/10.1016/j.dnarep.2017.06.003 (2017).

9 Hodel, K. P. *et al.* Explosive mutation accumulation triggered by heterozygous human Pol ε proofreading-deficiency is driven by suppression of mismatch repair. *eLife* **7**, e32692, doi:10.7554/eLife.32692 (2018).

10 Foster, P. L. Methods for determining spontaneous mutation rates. *Methods Enzymol* **409**, 195-213 (2006).

11 Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**, 9272-9277, doi:10.1073/pnas.0803466105 (2008).

12 Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979-993 (2012).

13 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

14 Zhang, C. Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179-184, doi:10.1038/nature14493 (2015).

15 Lada, A. G. *et al.* Genome-wide mutation avalanches induced in diploid yeast cells by a base analog or an APOBEC deaminase. *PLoS Genet* **9**, e1003736, doi:10.1371/journal.pgen.1003736 (2013).

16 Brody, Y. *et al.* Quantification of somatic mutation flow across individual cell division events by lineage sequencing. *Genome Res* **28**, 1901-1918, doi:10.1101/gr.238543.118 (2018).

17 Kennedy, S. R. *et al.* Volatility of Mutator Phenotypes at Single Cell Resolution. *PLoS Genet* **11**, e1005151, doi:10.1371/journal.pgen.1005151 (2015).

18 Luria, S. E. & Delbruck, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491-511 (1943).

19 Youn, A. & Simon, R. Using passenger mutations to estimate the timing of driver mutations and identify mutator alterations. *BMC Bioinformatics* **14**, 363, doi:10.1186/1471-2105-14-363 (2013).

20 Jackson, A. L. & Loeb, L. A. The mutation rate and cancer. *Genetics* **148**, 1483-1490 (1998).

21 Wagenmakers, E.-J. & Farrell, S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* **11**, 192-196, doi:10.3758/bf03206482 (2004).

22 Burnham, K. P. & Anderson, D. R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **33**, 261-304, doi:10.1177/0049124104268644 (2004).

23 Li, H. D. *et al.* Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load. *Journal of Clinical Investigation* **128**, 4179-4191, doi:10.1172/JCI122095 (2018).

24 Kane, D. P. & Shcherbakova, P. V. A common cancer-associated DNA polymerase ε mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from loss of proofreading. *Cancer Research*, doi:10.1158/0008-5472.CAN-13-2892 (2014).

25 Williams, L. N. *et al.* dNTP pool levels modulate mutator phenotypes of error-prone DNA polymerase ε variants. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1422948112 (2015).

26 Mertz, T. M., Sharma, S., Chabes, A. & Shcherbakova, P. V. Colon cancer-associated mutator DNA polymerase δ variant causes expansion of dNTP pools increasing its own infidelity. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1422934112 (2015).

27 Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197-200, doi:10.1038/255197a0 (1975).

28 Werner, B. & Sottoriva, A. Variation of mutational burden in healthy human tissues suggests non-random strand segregation and allows measuring somatic mutation rates. *PLOS Computational Biology* **14**, e1006233, doi:10.1371/journal.pcbi.1006233 (2018).

29 Furusawa, M. The disparity mutagenesis model predicts rescue of living things from catastrophic errors. *Frontiers in Genetics* **5**, doi:10.3389/fgene.2014.00421 (2014).

30 Claussin, C. *et al.* Genome-wide mapping of sister chromatid exchange events in single yeast cells using Strand-seq. *eLife* **6**, e30560, doi:10.7554/eLife.30560 (2017).

31 van Wietmarschen, N. & Lansdorp, P. M. Bromodeoxyuridine does not contribute to sister chromatid exchange events in normal or Bloom syndrome cells. *Nucleic Acids Research* **44**, 6787-6793, doi:10.1093/nar/gkw422 (2016).

32 Herr, A. J., Kennedy, S. R., Knowels, G. M., Schultz, E. M. & Preston, B. D. DNA replication error-induced extinction of diploid yeast. *Genetics* **196**, 677-691, doi:10.1534/genetics.113.160960 (2014).

33 Shlien, A. *et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nature Genetics* **47**, 257-262, doi:10.1038/ng.3202 (2015).

34 Lee, M. B. *et al.* Defining the impact of mutation accumulation on replicative lifespan in yeast using cancer-associated mutator phenotypes. *Proc Natl Acad Sci U S A* **116**, 3062-3071, doi:10.1073/pnas.1815966116 (2019).

35 Sherman, F. in *Part B: Guide to Yeast Genetics and Molecular and Cell Biology* Vol. 350 *Methods in Enzymology* (eds Christine Guthrie & Gerald R. Fink) 3-41 (Academic Press, 2002).

36 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

37 Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics (Oxford, England)* **30**, 2503-2505, doi:10.1093/bioinformatics/btu314 (2014).

38 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
39 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).
40 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).
41 Leisch, F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *2004* **11**, 18, doi:10.18637/jss.v011.i08 (2004).
42 Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S.* Fourth edn, (Springer, 2002).
43 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. 213 (Springer_Verlag, 2009).
44 Shinbrot, E. *et al.* Exonuclease mutations In DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Research*, doi:10.1101/gr.174789.114 (2014).

Acknowledgments

We would like to thank Larry Loeb, Sage Malingen and Marina Watowich for their critical review of our manuscript. This study was supported by National Institute for General Medical Sciences (NIH/NIGMS R01GM118854). ITD was supported by the UW Molecular Medicine Training Grant (NIH/NIGMS T32GM095421) and Genetic Approaches to Aging Training Grant (NIH/NIA T32AG000057). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NIA, or NIGMS.

Author Contributions

I.T.D., S.R.K., and A.J.H. designed research; I.T.D., J.S., J.M., E.M., and A.J.H. performed research; I.T.D., J.S., B.J.O., and A.J.H. analyzed data; I.T.D., and A.J.H. wrote the paper.

Competing Interests

The authors declare no competing interests.

Materials & Correspondence

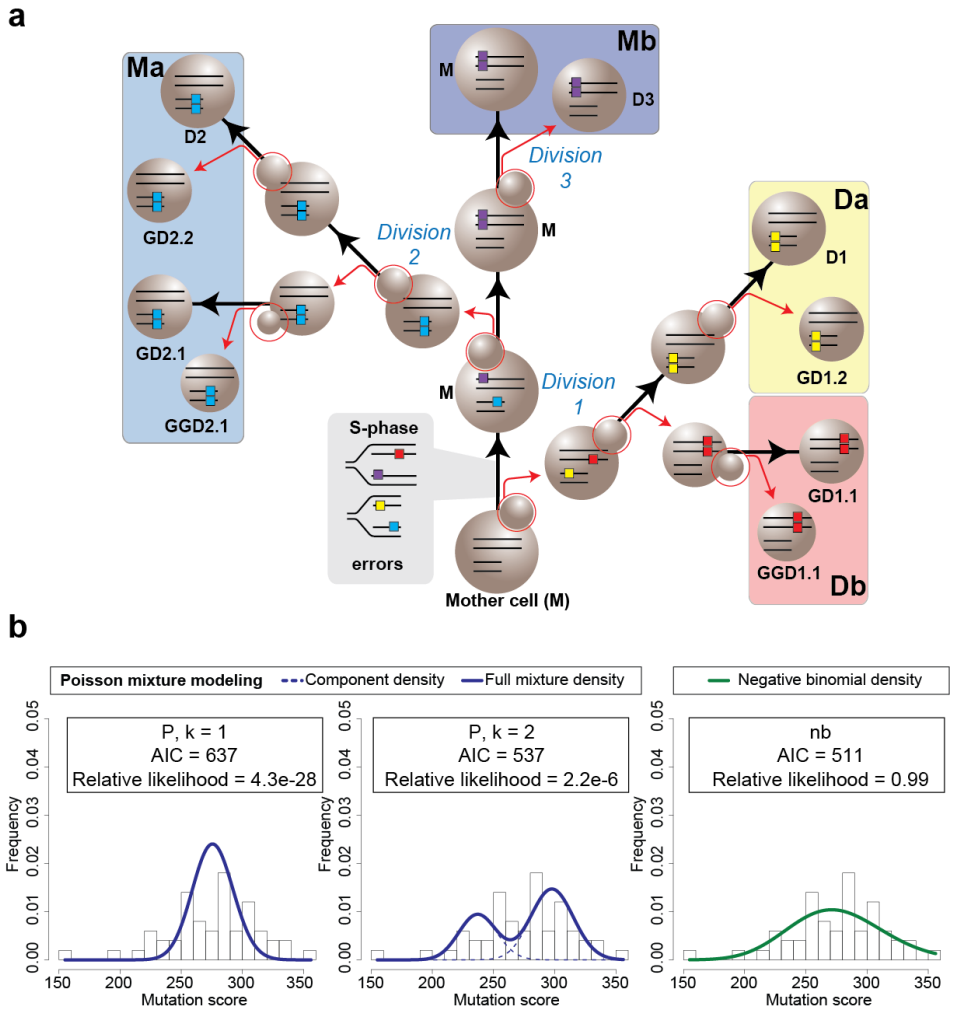
Additional Information

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Alan Herr (alanherr@uw.edu).

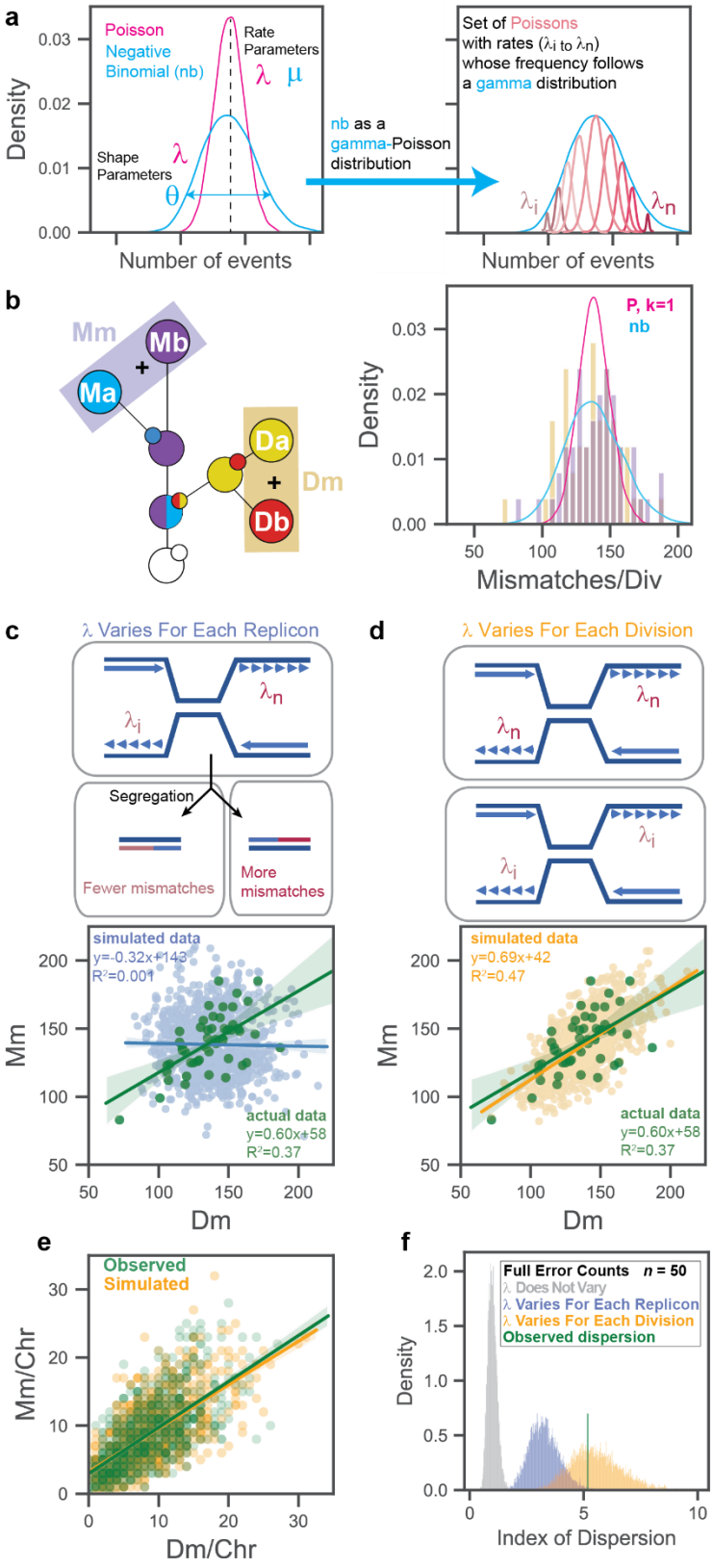
Figures

Fig.1: Mutator DNA polymerase errors at single cell resolution



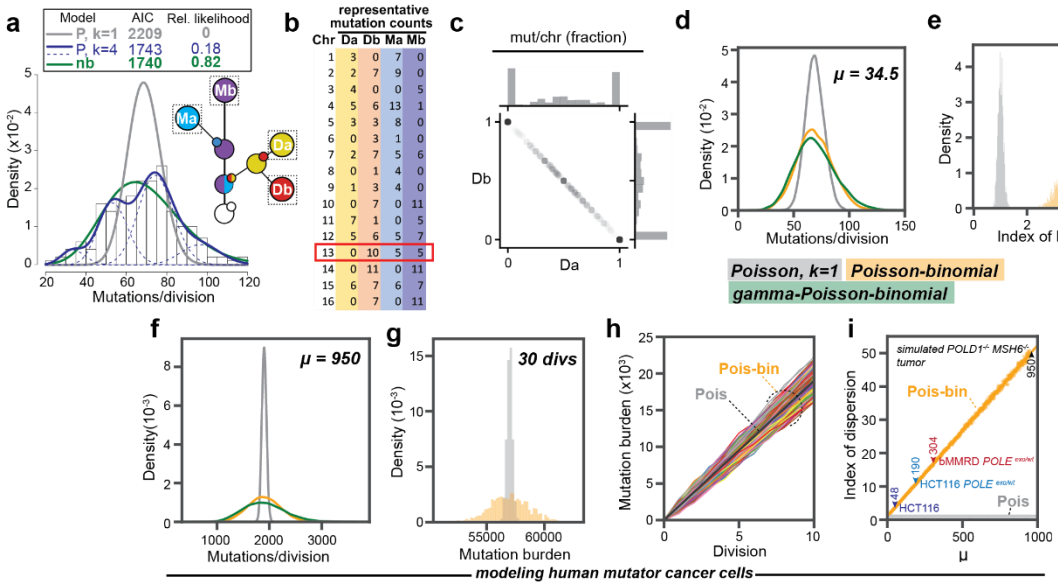
a, Isolation of single cell pedigrees. Using microdissection, the founding mother (M), daughter (e.g. D1), granddaughter (e.g. GD1.1, GD1.2) and great-granddaughter (e.g. GGD1.1) cells from each maternal division (*n* = 50) are separated (red arrows) and moved to isolated regions on the plate to form colonies, which are then sequenced. Polymerase errors arising during the initial S-phase are passed on to four segregant groups, highlighted by large colored boxes (Da, Db, Ma, Mb), the sum of which represents the full error count for that division. Large spheres connected by black arrows represent the same cell through multiple divisions. Small spheres circled in red represent budding daughter cells; parallel lines in cells, double-stranded DNA; colored boxes on lines, polymerase errors. **b,** Fitting the distributions of full error counts from diploid *pol3-01/pol3-01 msh6Δ/msh6Δ* divisions to different models. k = 1, single Poisson; k = 2, two-Poisson; nb, negative binomial; AIC, Akaike information criterion.

Fig. 2: Evidence that mutation rate varies between divisions



a, The negative binomial as a gamma-Poisson distribution. The gamma distribution takes the same shape parameter (θ) as the negative binomial and describes the variation in the rate parameter (λ) of a continuous mixture of Poisson distributions. **b,** Schematic of single cell lineage showing summing of segregant groups to determine the number of mismatches segregated to the mother (Mm) or daughter (Dm) in a single division. Actual distributions are represented by gold (Dm) and purple (Mm) bars. Lines depict models of data: pink, single Poisson (P, k = 1); aqua, negative binomial (nb). **c, d,** Correlations between Mm and Dm counts from actual data (green, *n* = 50) and simulations (*n* = 1000) under two different models. In **c**, top panel depicts a cell with converging replication forks from two replicons with different mutation rates. Bottom panel shows correlation of simulated Mm and Dm values (blue) drawn from the full negative binomial and their linear regression. In **d,** top panel depicts two cells replicating DNA with different mutation rates. Bottom shows correlation of simulated Mm and Dm values (orange) and their linear regression. **e,** Correlation between the number of mismatches per chromosome segregated to Mother (Mm) or Daughter cells (Dm). green, observed counts; orange, simulated counts from model in (**d**). **f,** Simulated index of dispersion of full replication error counts from small cohorts (n = 50) assuming the models from (**c** and **d**).

Fig. 3: Asymmetric segregation broadens the distribution of mutation burden in mutator cell populations.



a, Combined distribution of mutations fixed in the Da, Db, Ma, and Mb segregant groups (see inset) from *pol3-01/pol3-01 msh6Δ/msh6Δ* lineages (*n = 200*). Key of models (top): gray line, single Poisson (P, k = 1); blue lines, four-Poisson (P, k = 4), green line, negative binomial (nb). AIC, Akaike information criterion. **b**, Table of representative mutation counts from one division of a diploid mutator cell. Columns represent different segregant groups; rows, the chromosome (chr) number; values, the total number of new mutations found on homologous chromosome pairs. Red box indicates a chromosome with both asymmetric and equal sharing of mutations. **c,** Segregation of fixed mutations between Da and Db. For each division, the fraction of mutations observed in Da or Db on each chromosome was determined and then plotted against each other. **d,** Simulated distributions of mutations/division at a rate of μ = 34.5 (*n* = 10000) assuming a single Poisson process (grey), a Poisson-binomial process (orange), or a gamma-Poisson-binomial process (green). **e,** Variation in the index of dispersion of simulated data from

748 the 3 models (n = 200) over 1000 iterations. **f**, Simulated distribution of mutations/division in
749 human ultra-mutator cells assuming a fixed mutation rate (μ = 950, n = 10000) comparable to
750 *pol3-01/pol3-01 msh6Δ*/*msh6Δ* yeast and a single Poisson process (grey), Poisson-binomial
751 process (orange), or gamma-Poisson-binomial process (green). **g,** The cumulative mutation
752 burden of a human ultra-mutator cell after 30 simulated divisions with (orange) and without
753 (grey) asymmetric segregation. **h,** Simulated trajectory of mutation burden of human mutator
754 tumor cells (Colored lines, n = 1000) undergoing a Poisson-binomial process compared to a
755 Poisson process (black line). **i,** Change in the index of dispersion under a Poisson-Binomial
756 process (orange line) compared to the static index of dispersion under a Poisson process (grey
757 line at bottom) with an increasing mutation rate. Colored markers represent estimated mutation
758 rates for clinically relevant mutator-driven HCT116-derived mammalian cancer cell lines[9] and a
759 tumor from a patient with biallelic MMR deficiency (bMMRD)[33].

Supplementary Information

Corresponding author: Alan J. Herr, alanherr@uw.edu

Haploid *pol2-4 msh6Δ* Mutators

Prior to switching to stronger diploid mutators, we first obtained full replication error counts for 44 *pol2-4 msh6Δ* divisions from 7 independent lineages, encompassing 308 mutations (Extended Data Table 1, Supplementary Dataset 2). We sequenced only those clones that would contribute to a full error count (Extended Data Fig.2). Since our previous study suggested that mutations were fixed in *pol2-4 msh6Δ* mother cells at a rate of 0.4 or 4 mutations/genome/division, with full replication error counts, the volatility model predicts two well-separated Poisson distributions centered around 1.6 and 16 replication errors per division. Instead, we observed a single distribution centered around 6.5 (±3.9) replication errors per division. The distribution of full replication error counts in *pol2-4 msh6Δ* cells had a $\hat{D}$ of 2.2, which is consistent with a less pronounced volatility of the *pol2-4 msh6Δ* mutator phenotype. In keeping with this interpretation, fitting these data to different probability distributions revealed they matched a negative binomial better than a single or two-Poisson mixture as judged by AIC. Parsing this data into the number of mutations fixed per individual cell produces a distribution (N=176) that fits a single Poisson with a rate of 1.75 mutations/division. This finding does not negate the hypothesis of a mild mutator volatility based on the full replication error counts. The expected dispersions of fixed mutations in *pol2-4 msh6Δ* haploid cells (n=176, λ=1.75) are comparable for the Poisson-binomial ($\hat{D}$ = 1.13±0.12) and Poisson ($\hat{D}$ = 1.0±0.1) models (Extended Data Fig.8). The rate of 1.75 mutations/division lies almost directly between the predicted underlying rates from our published two-Poisson Model[17]. Thus, our previous distribution likely contained a preponderance of cell divisions with this intermediate mutation rate. The high number of divisions in that earlier dataset with no mutations could have partly been the result of a biological “zero-inflation” due to the unequal sharing of mutations described in Fig. 3 for *pol3-01/pol3-01 msh6Δ/msh6Δ* cells. If so, why are there fewer cells with 0 mutations in the current distribution? We suspect that the stringent requirement of eight viable clones to obtain a full replication error count may have introduced an ascertainment bias. Due to unequal sharing of mutations, members of the lineage with the highest number of mutations may fail to form a colony. The reciprocal clones with no errors from that same division would also not be scored. This potential ascertainment bias would affect our estimates of mutator volatility, since divisions with a higher mutation rate are more likely to have at least one progeny fail to form a colony.

Colonies Not Included In Analysis

Many complete sub-lineages (comprised of d, gd1, gd2, and ggd) were not sequenced because inviability later in the lineage prevented us from gaining a full replication error count. For instance, full replication error counts for AH120 divisions that yielded d3 and d4 are not possible because the d5 sublineage was completely inviable (Extended Fig.2). Likewise, sometimes colonies within informative sub-lineages (e.g. AH119 gd8-2, AH120 gd10-2/ggd10, AH121 gd9-1) were not sequenced because they were not required for a full replication error count. In some cases (AH156 d14, AH160 d5, AH157 d8), colonies that provided identical information on a division segregant subgroup were sequenced when the preferred colony failed to form a viable clone (Extended Fig.3). In other cases (AH158 gd4-2/ggd4, AH158 gd7-2/ggd7, AH162 gd6-2/ggd6), colonies had poor sequencing coverage and were censored from the analysis. In rare cases, sequence analysis and review of dissection notes suggested an

assignment error occurred; however, our careful analysis of the patterns of shared mutations enabled post-hoc deconvolution of the events. In one example, in the 8th division of AH156, we separated 3 cells from the mother. The two larger, similarly sized cells were both clearly daughter cells and the smaller cell was a granddaughter cell, although it was not clear which was its parent. We moved the daughter cells to the d8 (AH15629) and d9 positions (AH15633) and placed the smaller cell below in the gd8-1 position (AH15630). We then proceeded to isolate the remaining members of the lineage. Sequencing analysis of the resulting colonies revealed that AH15633 was in fact d8, AH15629 was d9, and AH15630, the daughter of AH15633, not AH15729. This error meant we had to reorder the lineage. AH15631 was gd9-1 not gd8-2. AH15632, daughter of AH15630, was ggd8 and AH15634 was gd8-2. Since we didn't realize AH15631 was gd9-1, we failed to dissect her first daughter to serve as ggd9 and consequently were unable to obtain a full replication error count for the 9th division. In another example, AH160 division 2 was censored entirely from analysis after finding from the pattern of mutations that AH16005 was a granddaughter derived from d1 of this lineage (AH16001).

Distribution of Mutations and Spectrum

Plotting the mutations scored from all divisions of *pol3-01/pol3-01 msh6Δ/msh6Δ* mutator mother cells reveals mutations were generated across much of the unmasked portions of the sequenced genome (Extended Data Fig.4a). Upon investigation, the few tracts of unmasked chromosomes lacking mutations are likely artifacts of regions of low sequence coverage which consistently fell below our target thresholds for quality and depth in at least one or more members of a lineage.

As expected, C→T mutations are the most abundant single nucleotide substitutions, followed by T→C and C→A. The trinucleotide context reveals a prominent peak of C→A mutations at a TCT context, a hallmark of proofreading deficiency[13,44], as well as a peak at CCT (Extended Data Fig.4b).

Extended Data Table 1

Lineage[a]	Scored Sites[b]	Mutations[c]	Divisions[d]	Mutation Rate[e]
pol2-4 msh6				
119	11,080,506	33	5	0.006
120	11,120,599	51	7	0.0066
121	11,203,731	38	6	0.0057
122	10,909,400	59	7	0.0077
123	11,099,548	24	4	0.0054
124	10,094,023	58	10	0.0057
125	10,897,295	45	5	0.0083
Total		*308*	*44*	
Mean (stdev ±)	10,915,014 (3.51 x 10^5)	44 (12)	6.3 (1.8)	0.0065 (0.001)
pol3-01/pol3-01 msh6Δ/msh6Δ				
151	10,541,044	1,597	6	2.53
153	10,636,360	1,582	6	2.48
156	9,521,963	2,976	10	3.13
157	10,594,546	1,599	6	2.52
158	9,818,623	1,347	5	2.74
160	9,679,326	1,543	5	3.19
162	8,578,576	3,157	12	3.07
Total		*13,801*	*50*	
Mean (stdev ±)	9,910,063 (6.95 x 10^5)	1,972 (699)	7 (2.5)	2.81 (0.29)

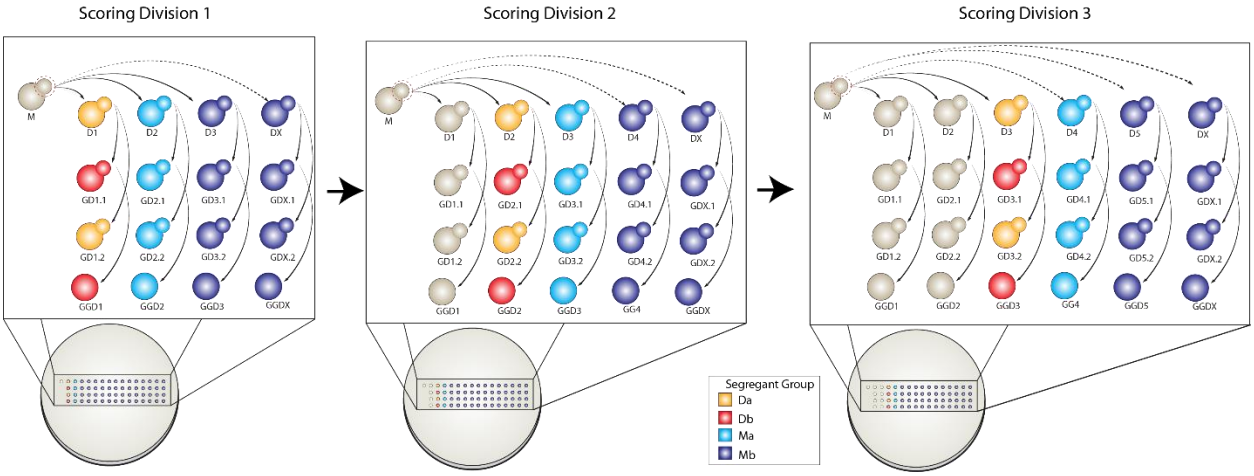
[a] Lineage refers to descendants of the same mother cell. See Extended Data Figs.2 and 3 for images of colonies and Supplementary Datasets for mutations.

836 [b] Scored sites refers to the number of genomic nucleotide positions confidently scored in all members of the
837 lineage.
838 [c] The total number of independent mutations identified within each lineage.
839 [d] The number of divisions with full replication error counts (see Fig.1).
840 [e] Mutation rate (x 10^{-5} mutations/bp/division): the number of mutations divided by the total number scored sites
841 divided by the number of divisions.

842

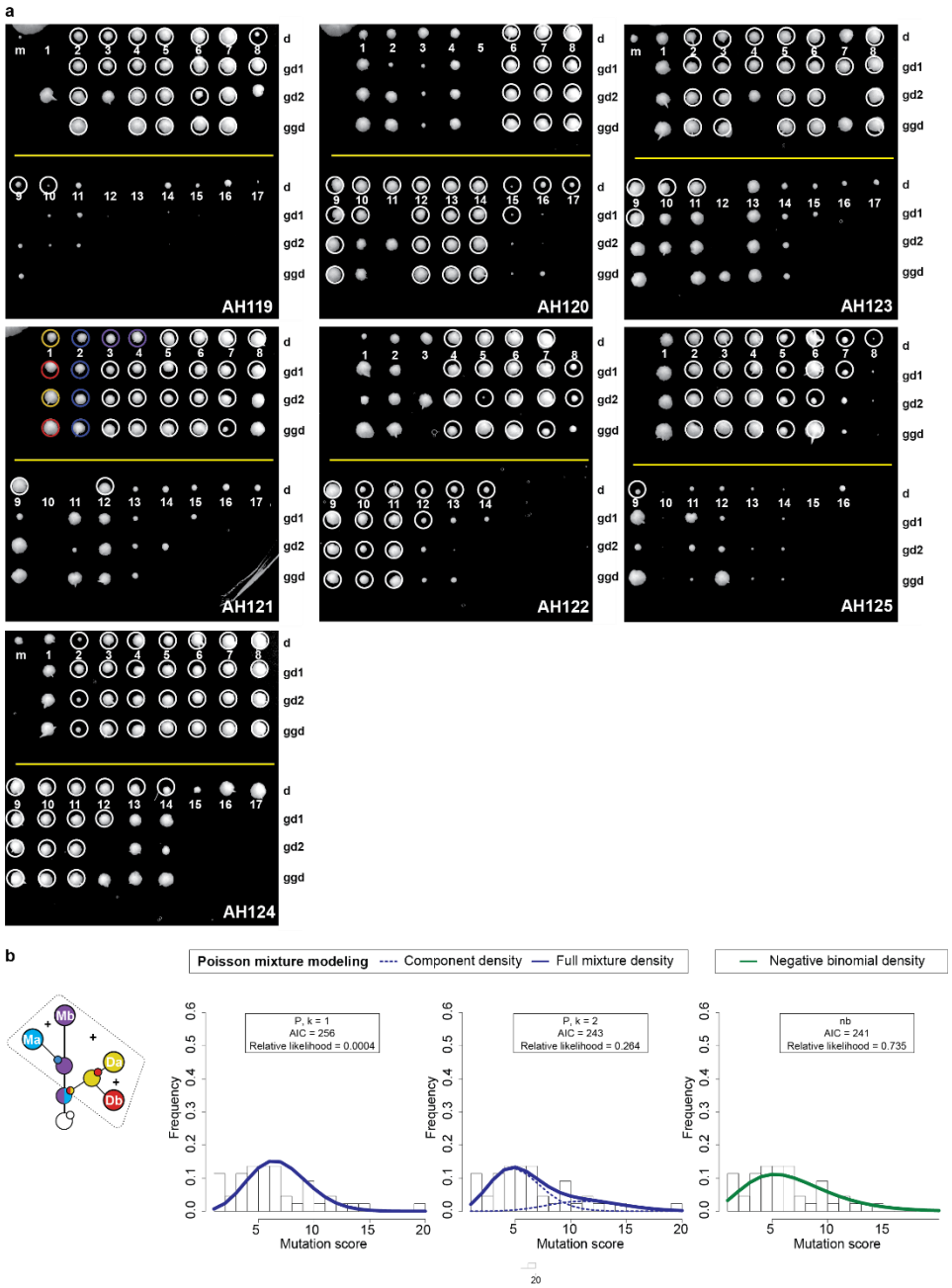
Extended Data Figures

Extended Data Fig. 1: Scoring mutations from multiple divisions from the same lineage.



Arrows depict movement of dissected daughter cells and their descendants to unique positions on the plate, where they form colonies that are sequenced. For each division, mutations unique to the yellow (Da) or red (Db) colonies are counted as errors made during the division. Mutations uniquely shared by all blue cells (from the next division) represent segregant group Ma, and new mutations shared by all subsequent offspring of the mother (Purple) represent group Mb. A full replication error count is the sum of Da, Db, Ma, and Mb. Obtaining full error counts in subsequent divisions shifts the identities of the red, yellow, blue and purple cells one column to the right and repeats the procedure (See AH121 in Extended Data Fig. 2 for an example). While the genomes of colonies used to score mutations from prior divisions are used again, the new mutations scored in the current division were not previously assigned to one of the prior segregant groups.

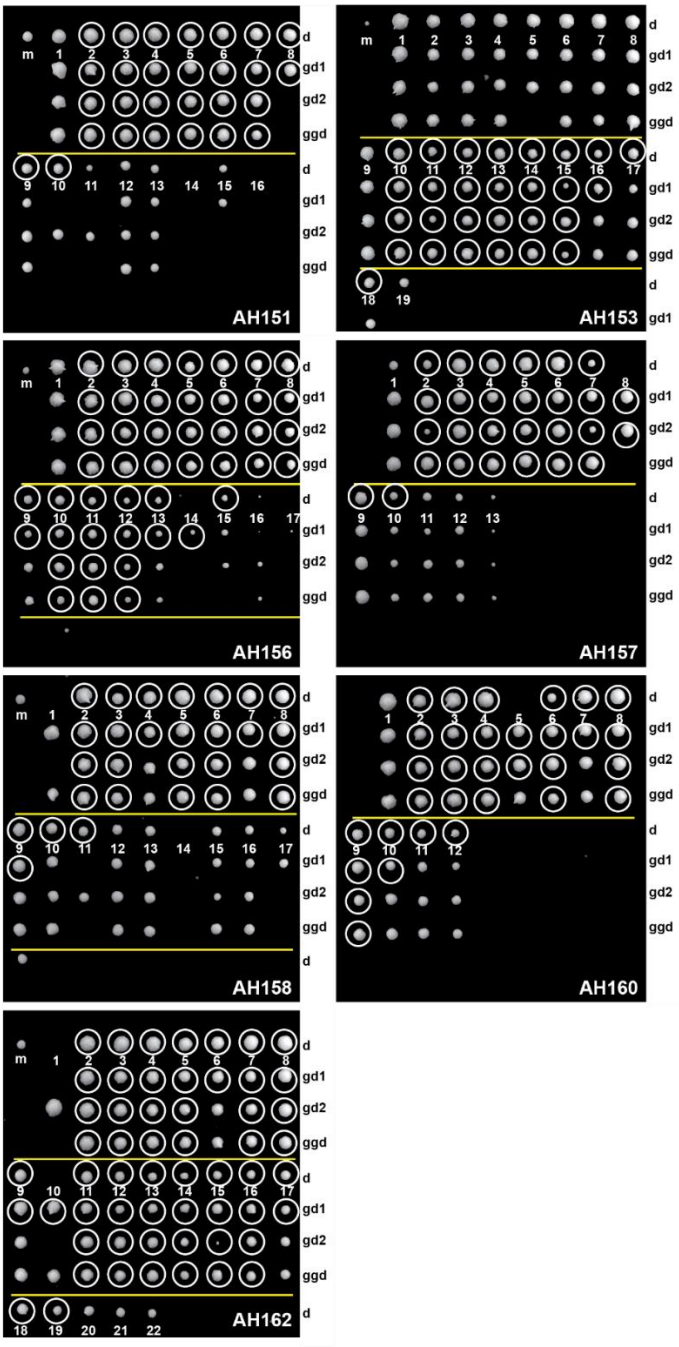
858 **Extended Data Fig. 2: *pol2-4 msh6Δ* lineages.**



859

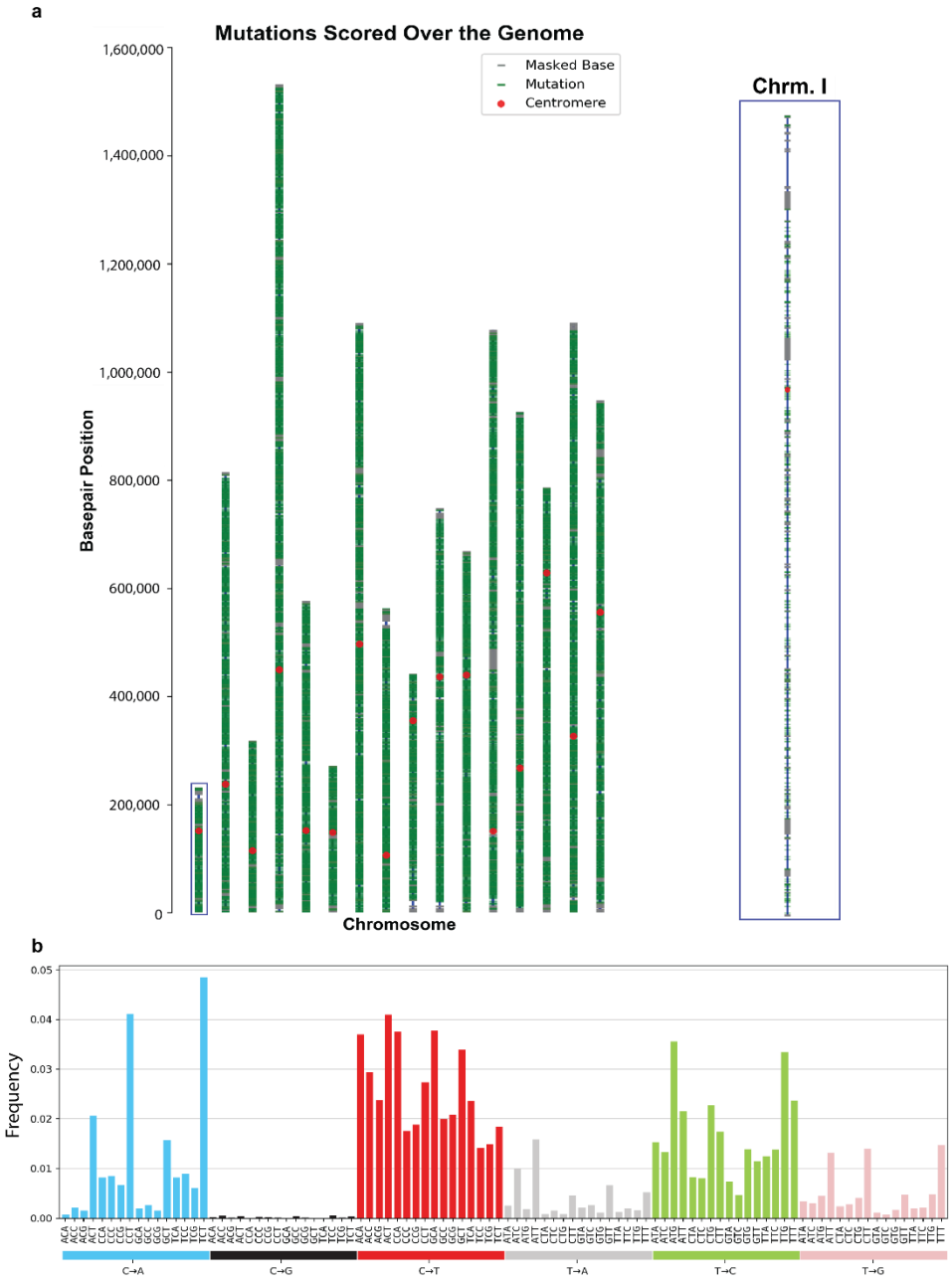
860 **a**, Photographs of agar plates with colonies formed from single cell lineages. The lineage
861 number is given in the lower right-hand corner. Locations of rows of daughter (d), first
862 granddaughter (gd1), second granddaughter (gd2), and great-granddaughter (ggd) colonies are
863 given on the right-hand side of the images. Sublineage number is indicated below each
864 daughter colony. Circles indicate sequenced colonies. Colored circles in Lineage AH121
865 illustrate segregant groupings for the first division (see Extended Data Fig. 1). Yellow line
866 divides earlier sublineages from later sublineages. Gaps in colony growth reflect lethality. **b**,
867 Fitting the distributions of full error counts from haploid *pol2-4 msh6Δ* to alternative models. k=1,
868 single Poisson; k=2, two-Poisson; nb, negative binomial; AIC, Akaike information criterion.

Extended Data Fig. 3: *pol3-01/pol3-01 msh6Δ/msh6Δ* lineages.



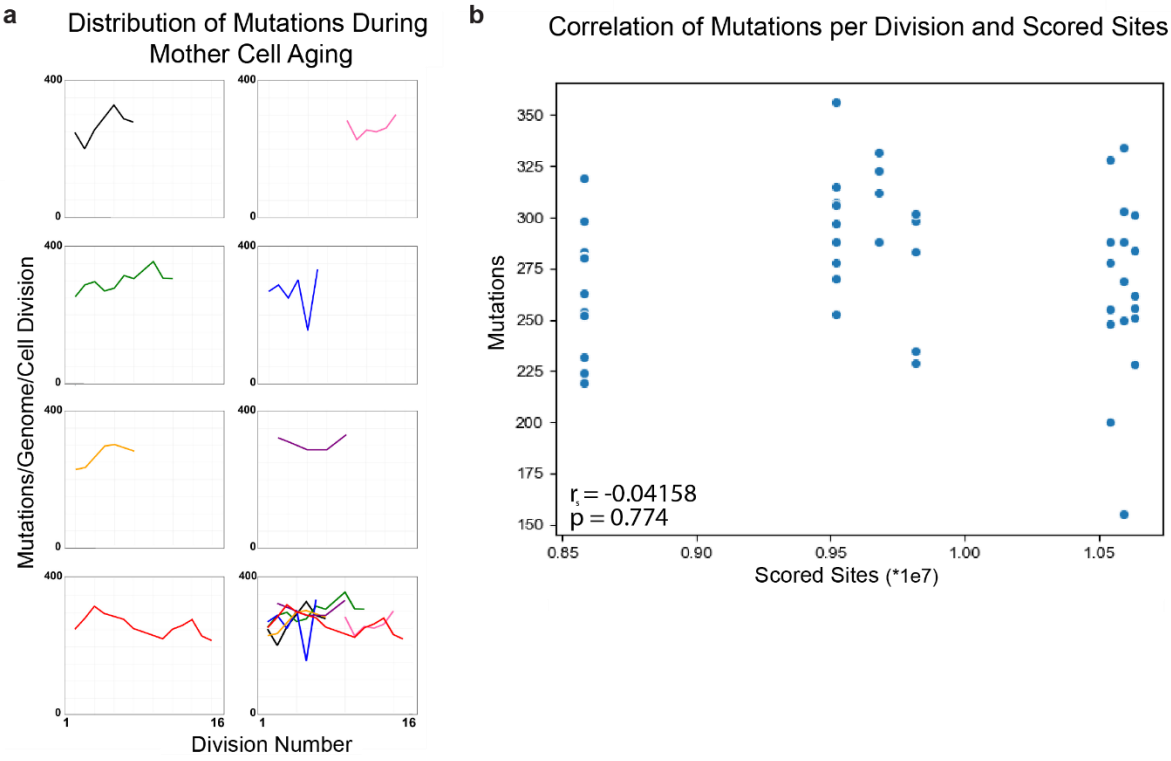
Photographs of agar plates with colonies formed from single cell lineages. The lineage number is given in the lower right-hand corner. Locations of rows of daughter (d), first granddaughter (gd1), second granddaughter (gd2), and great-granddaughter (ggd) colonies are given on the right-hand side of the images. Sublineage number is indicated below each daughter colony. Yellow line divides earlier sublineages from later sublineages. Gaps in colony growth reflect lethality.

Extended Data Fig. 4: Distribution of mutations and spectra in *pol3-01/pol3-01 msh6Δ/msh6Δ* lineages.



a, Genome level distribution of mutations (green) over yeast chromosomes (blue lines). 13,801 mutations pooled from 50 scored divisions of *pol3-01/pol3-01 msh6Δ/msh6Δ* diploid mother cells, representing approximately 1 mutation per 1000 bases of the yeast genome. Close up view of representative chromosome I (right). Masked bases are represented by grey ticks. **b**, 96-trinucleoide mutation spectra context of all mutations (spectrum) by frequency that arose over 50 divisions of *pol3-01/pol3-01 msh6Δ/msh6Δ* diploid mother cells, generated using the snv-spectrum program (https://github.com/aroth85/snv-spectrum).

888 **Extended Data Fig. 5: Excluding simple explanations for *pol3-01/pol3-01 msh6Δ/msh6Δ***
889 **mutator volatility.**



890

891 **a**, Mutation counts and maternal age. The total mutation counts from individual divisions is
892 plotted relative to maternal age (Division number). **b**, Mutation counts and size of scored
893 genome. The proportion of the genome scored in all members of a lineage varies between
894 lineages due to sequencing depth and number of lineage members, but is not correlated with
895 mutation counts (Spearman Correlation).
896

Extended Data Fig. 6: Semi-conservative DNA replication and segregation cause asymmetrical inheritance of mutations.

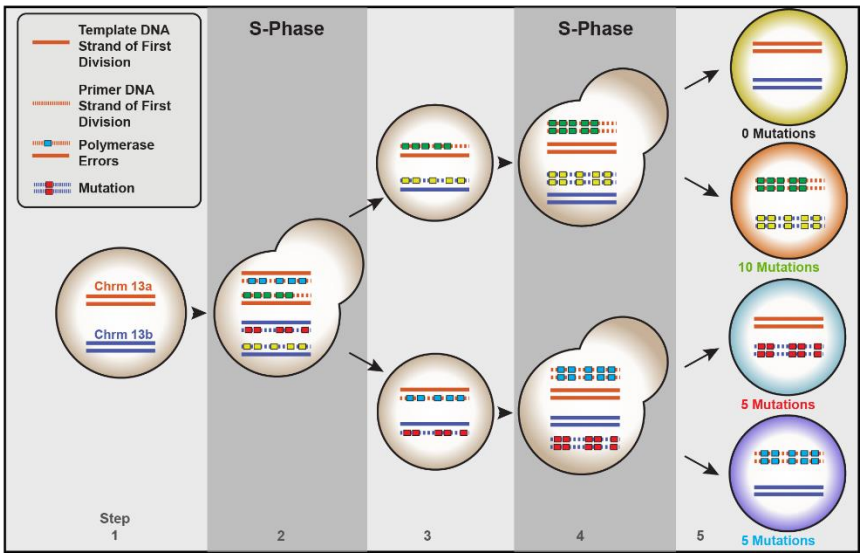
a

Segregant group

Yeast Chromosomes

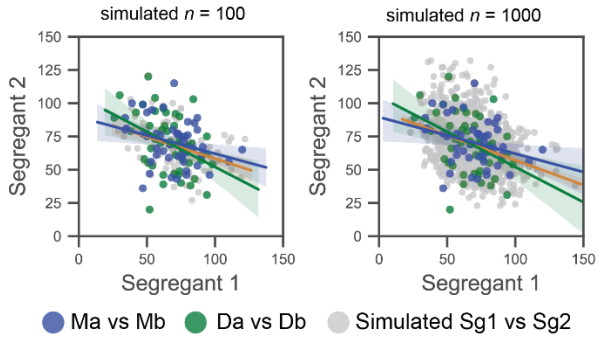
Chr	Da	Db	Ma	Mb
1	3	0	7	0
2	2	7	9	0
3	4	0	0	5
4	5	6	13	1
5	3	3	8	0
6	0	3	1	0
7	2	7	5	6
8	0	1	4	0
9	1	3	4	0
10	0	7	0	11
11	7	1	0	5
12	5	6	5	7
13	0	10	5	5
14	0	11	0	11
15	6	7	6	7
16	0	7	0	11

b



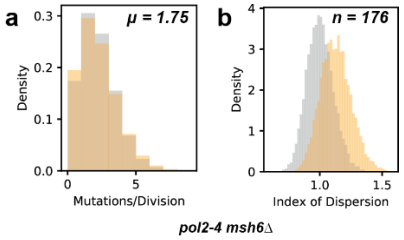
a, Table of representative mutation data from one division of a diploid mutator cell. Columns represent different segregant groups (see Fig.1); rows, the chromosome number; values, the total number of new mutations found on homologous chromosome pairs. The red box indicates an example with both asymmetric and equal sharing of mutations. **b**, Five-step model of unequal segregation: 1) Two homologous chromosomes (orange and blue) prior to scored division of mother cell. 2) Mother cell duplicates chromosomes and mutator Pol δ generates errors (colored boxes) on the nascent strands. 3) Progeny each inherit two chromosomes with mismatches. 4) Each unresolved mispair is 'fixed' as a point mutation in the next S-phase. Error-free strands are free of newly fixed mutations. 5) segregation results in cells with 0, 1, or 2 mutagenized chromosomes.

Extended Data Fig. 7: Correlations between segregation groups.



X/Y Scatter plots of segregant group pairs (Da/Db, Ma/Mb, *n* = 50 for each) from *pol3-01/pol3-01 msh6Δ/msh6Δ* divisions are plotted alongside simulated data (*n* = 100, left; *n* = 1000, right). Segregant 1 corresponds to Da or Ma. Segregant 2 corresponds to Db or Mb. The highest outlier point for Da/Db in the upper left- hand quadrant (51, 120) comes from Division 8 (Supplementary Dataset 1), which produced a Ma/Mb point located at (120, 65) that yielded similar mismatch totals (Dm, 171; Mm, 185). This suggests the division had a high mutation rate. Ma/Mb (47,36) and Da/Db (52, 20) segregant pairs in the lower left-hand quadrant also appear as outliers. Both pairs are derived from Division 15 (see Supplementary Dataset 1), leading to the conclusion that the mutation rate in that division was inherently low.

Extended Data Fig. 8: Simulation of *pol2-4 msh6Δ* haploid mutagenesis.



a, The simulated distribution of mutations from haploid *pol2-4* m*sh6Δ* cells at a rate of μ = 1.75 mutations/division (n=10000) assuming a single Poisson process (grey) or a Poisson-binomial process (orange). **b**, Variation in the index of dispersion of simulated data from Poisson and Poisson-binomial models (n=176) over 10,000 iterations.

Supplementary Dataset 1: *pol3-01/pol3-01 msh6Δ/msh6Δ* sequencing data

Page	*Description*
SRA submission	List of sequence files submitted to the Short Read Archive (SRA)
Mutation Summary	Summary of mutation counts from different single cell lineages
MutInSegGroups	Table of mutation counts broken down into segregant groups
MutInSegGroupsChr	Segregant group table further broken down by chromosome
Full Mutation List	All mutations observed, organized by segregant groups.
Lineage 151	Sorted spreadsheet from Lineage 151 showing variants in segregant groups

Lineage 153	Sorted spreadsheet from Lineage 153 showing variants in segregant groups
Lineage 156	Sorted spreadsheet from Lineage 156 showing variants in segregant groups
Lineage 157	Sorted spreadsheet from Lineage 157 showing variants in segregant groups
Lineage 158	Sorted spreadsheet from Lineage 158 showing variants in segregant groups
Lineage 160	Sorted spreadsheet from Lineage 160 showing variants in segregant groups
Lineage 162	Sorted spreadsheet from Lineage 162 showing variants in segregant groups

Supplementary Dataset 2: *pol2-4 msh6Δ sequencing data*

Page	*Description*
SRA submission	List of sequence files submitted to the Short Read Archive (SRA)
Mutation Summary	Summary of mutation counts from different single cell lineages
Full Mutation List	All mutations observed, organized by segregant groups.
R1_Lineage	Sorted spreadsheet from Lineage R1 showing variants in segregant groups
R2_Lineage	Sorted spreadsheet from Lineage R2 showing variants in segregant groups
R4_Lineage	Sorted spreadsheet from Lineage R4 showing variants in segregant groups
R5_Lineage	Sorted spreadsheet from Lineage R5 showing variants in segregant groups
R6_Lineage	Sorted spreadsheet from Lineage R6 showing variants in segregant groups
R9_Lineage	Sorted spreadsheet from Lineage R9 showing variants in segregant groups
R10_Lineage	Sorted spreadsheet from Lineage R10 showing variants in segregant groups