The origin of human mutation in light of genomic data

Vladimir B. Seplyarskiy[1,2] *and Shamil Sunyaev*[1,2] ✉

Abstract | Despite years of active research into the role of DNA repair and replication in mutagenesis, surprisingly little is known about the origin of spontaneous human mutation in the germ line. With the advent of high-throughput sequencing, genome-scale data have revealed statistical properties of mutagenesis in humans. These properties include variation of the mutation rate and spectrum along the genome at different scales in relation to epigenomic features and dependency on parental age. Moreover, mutations originated in mothers are less frequent than mutations originated in fathers and have a distinct genomic distribution. Statistical analyses that interpret these patterns in the context of known biochemistry can provide mechanistic models of mutagenesis in humans.

Germline mutation
Nucleotide substitutions accumulated in the germ line that, thus, could be transmitted to the offspring.

Somatic mutations
Nucleotide substitutions accumulated during cell divisions of somatic cells in human tissues.

Interest in the origin of human mutation extends well beyond the fields of DNA replication and DNA repair into numerous areas of the life sciences (BOX 1). The role of mutation as the source of genetic variation and, eventually, differences between species brings mutagenesis to the focus of evolutionary biology and population genetics. As a cause of cancer, Mendelian disorders and complex disease, mutations attract the attention of oncologists and medical geneticists. More specifically, models of the mutation rate, the rate at which nucleotide substitutions occur spontaneously during transmission of genetic information, are at the core of many statistical methods of evolutionary genetics, cancer genomics and human disease genetics[1–3]. Mutational footprints guide the search for mutagenic agents posing environmental health risks[4–6]. Patterns of mutations are informative about the underlying mutational processes — for example, DNA repair deficiency in cancer — and can suggest avenues for therapeutic intervention[7,8]. Considering the importance of mutagenesis for numerous fields of basic and medical research, the paucity of knowledge on the origin of human germline mutation is highly surprising.

Large-scale DNA sequencing data sets are a new bottomless source of mutation data. Changes in the DNA sequence that arise during parent to offspring transmission represent de novo germline mutations. These de novo mutations are directly detectable in parent–child trio sequencing studies as nucleotide variants that are present in children but absent in both parents. In the course of population history, some mutations are inherited by many individuals and appear as segregating single-nucleotide variants (SNVs) in population sequencing data sets. As a source of information on mutational characteristics, SNVs do not represent direct observations, because they have passed through the sieve of natural selection and biased gene conversion and are influenced by population demographic history. Still, the amount of data on SNVs is orders of magnitude greater than the number of de novo mutations detected by trio sequencing, making them an attractive proxy for mutations in statistical analyses of mutagenesis. Divergence between genomes of different species, available from comparative genomics data sets, also represents mutations accumulated and fixed at longer evolutionary timescales.

For decades preceding the advent of sequencing, direct experimentation was the only effective way to investigate the role of DNA repair and replication in generating mutations. For example, in vitro reconstitution of the highly complex process of DNA replication propelled studies of fine details of initiation and extension of DNA synthesis[9,10]. Advances in experimental biochemistry enabled the creation of genome-wide maps of the efficiency of some DNA repair processes[11,12]. Experiments on genetically manipulated model organisms or cell lines have been informative about mutational footprints associated with DNA repair deficiency or hyperactivity of mutagenic proteins[13–15]. Mutation accumulation lines emerged as a distinct powerful approach to assess the effects of exogenous mutagens in wild-type or genetically modified organisms[15–18]. Now, the accumulation of next-generation sequencing data opens a new avenue into mutation research. Indeed, recent progress in understanding mechanisms of cancer somatic mutations has been driven by the statistical analysis of cancer genomes. Somatic mutations can be found by deep sequencing[19–23], single-cell sequencing[24] or clonal passage of individual cells[25,26]. In cancer, somatic mutations accumulated in cells become clonal and are revealed in sequencing studies of bulk cancer specimens[27,28].

[1]*Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.*

[2]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.*

✉*e-mail: ssunyaev@rics.bwh.harvard.edu*

https://doi.org/10.1038/s41576-021-00376-2

Box 1 | **Impact of statistical and mechanistic models of mutagenesis on other fields**

Statistical models of mutagenesis have found multiple applications in medical genetics and evolutionary biology. De novo mutations are a frequent cause of rare monogenic diseases[145,146]. They also contribute to polygenic neuropsychiatric diseases such as autism spectrum disorder[2,135,147]. Recurrent de novo mutations in multiple patients point to genes causally involved in disease, which has motivated parent–child trio sequencing studies aimed at identifying disease genes[148]. In sufficiently large data sets, however, recurrent non-pathogenic mutations are also detected in genes unrelated to disease. Studies of both monogenic and polygenic diseases have adopted statistical mutation models that take into account nucleotide context dependency. To map disease genes, the number of recurrent de novo mutations in patients has to be compared with the prediction of the background mutation model[2]. For somatic mutations, conceptually similar approaches are at the core of gene discovery approaches in cancer genomics that aim to identify 'drivers' of cancer development and progression[149–151].

In evolutionary biology, adaptation is commonly inferred from an increased level of genetic divergence between species or population polymorphisms compared with the expected action of mutation alone. This inference is only possible in the presence of a statistical model of mutation in the locus to assess the background level of divergence or variation. The mutation models also help to infer selective constraint of genes or functional elements, which manifests as a lack of variation or divergence[152,153]. Estimates of selective constraint aid the identification of non-coding functional elements by comparative genomics and the prediction of genes involved in human autosomal dominant disorders[144].

Mechanistic models of mutagenesis also have applications outside the immediate field. The effect of epigenetic features on both germline and somatic mutations allows the prediction of epigenetic features of inaccessible cells on the basis of mutations. For example, replication timing and direction significantly influence mutagenesis and leave footprints in mutation data. These patterns of mutations enable the reconstruction of replication timing and replication fork orientation in germline cells[67]. Correlation of epigenomic profiles of different cell types with mutation patterns is informative about cells in which mutations originated. For somatic cancer mutations, this approach can predict the cell of origin of a tumour[154], which is especially important for metastatic tumours whose primary location is unknown. Signatures of somatic mutations associated with mismatch repair deficiency[155] and deficiency of homologous repair identify underlying molecular mechanisms that serve as targets for cancer treatment[8,156,157].

Mutations that occur early in development are shared by many cells in various tissues. Some of them are present in the germ line and may contribute to de novo human mutations[29,30], bringing them into the focus of this Review. Differential exposure of individual tumours to exogenous or endogenous mutagens provides the main instrument for the statistical analysis of cancer genomes. For example, genomes of lung tumours of cigarette smokers are strongly enriched in G>T mutations induced by benzo[*a*]pyrene (a component of tobacco smoke), whereas lung tumours of non-smokers do not exhibit this mutational signature[31]. The same logic has been applied to the analysis of DNA repair system deficiencies or the activity of endogenous mutagens, such as AID, APOBEC and reactive oxygen species[32,33]. A strong association between exposure to a chemical, genetic or epigenetic dysregulation of DNA maintenance and a specific mutational footprint is considered evidence of a causative role of the corresponding process. This epidemiological approach to the analysis of mutagenesis has had a strong impact in oncology[33,34], guiding strategies for prevention and therapeutic intervention.

Although these elegant experimental studies and further cancer research have identified multiple mechanisms that may lead to mutation, they are fundamentally unable to estimate the relative contributions of each mechanism to spontaneous human germline mutations. Some of the findings on sources of extreme mutagenic events cannot be transferred to germ cells at all, where such events should not play a substantial role. Germline mutations accumulate over long periods of time in cells with proficient DNA repair, posing a challenge for studies in a laboratory or in the context of cancer. Now, computational studies of numerous naturally occurring human mutations have provided a new perspective on mutagenesis research.

Here, we review statistical patterns in de novo germline mutations from parent–offspring trio sequencing and in human genetic variation data from large-scale genome sequencing projects. As statistical studies can only be informative if conducted in light of accumulated experimental knowledge about mutational mechanisms, the main focus of this Review is on the interpretation of data from a biochemical point of view. We discuss ways to incorporate existing knowledge of DNA replication and repair pathways into statistical data analysis to identify key mechanisms underlying the incessant influx of mutations in the human population.

The sources of mutations

Cellular processes that lead to mutations leave footprints in the DNA sequence. The statistical analysis of genomic data can detect these footprints and quantify the contribution of specific mechanisms to the overall stream of mutations. In this section, we provide an overview of various mutagenic processes, starting with the basic division between replication errors and DNA damage, and the detectable traces they leave. As with all simplified organizational principles, it is necessary to add a note of caution that this division is not absolute, and some processes described at the end of this section involve interactions between DNA replication, damage and repair.

Mutational properties of DNA replication and unrepaired DNA damage. Mutations arise via two major mechanisms: base misincorporation during replication of non-damaged DNA; and accumulation of DNA damage that has not been properly repaired, leading to mutation (FIG. 1). Even at this most basic level, the relative contributions of DNA replication infidelity and DNA damage to human mutation are unknown.

At first glance, replication infidelity is expected to leave a major statistical footprint in data — the number of replication-induced mutations should scale with the number of cell divisions. This would most obviously manifest as a dependency of the mutation rate on paternal age, as sperm cells continually divide through adulthood. However, to complicate matters, unrepaired DNA lesions are primarily converted to mutations during DNA replication[35,36] (FIG. 2). Thus, not only the number of mutations caused by replication errors but also the number of damage-induced mutations are likely to scale with the number of cell divisions[37], making projection of age dependency to the origin of mutations non-trivial.

There are other ways in which replication infidelity may leave a detectable footprint in DNA sequencing data. For example, replication of leading and lagging DNA strands requires two different machineries[38], potentially leading to asymmetric mutation accumulation

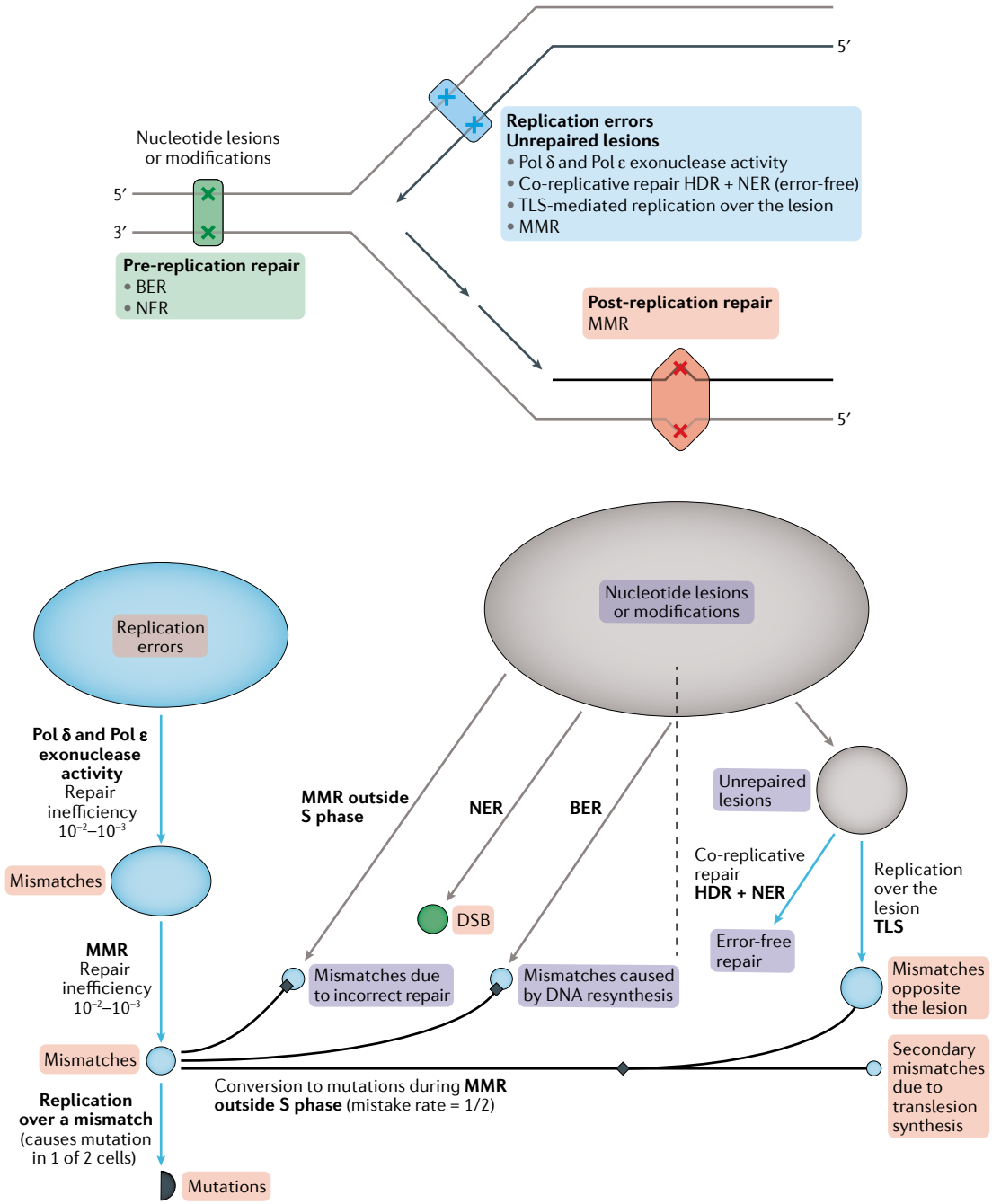


Fig. 1 | **Sources of point mutations.** Mutations occur due to replication errors (~10^{-4} per nucleotide[39,136]) or as a consequence of DNA damage (~70,000 nucleotide lesions or modifications per day[137]). Replication involves a few systems that secure high fidelity (left). Highly selective active sites of major DNA polymerases misincorporate 1 nucleotide per 1,000 (REF.[136]); the resulting mismatches are effectively purged by the exonuclease activity of polymerases Pol δ and Pol ε and by the mismatch repair (MMR) system[39]. Unrepaired mismatches later become substitutions that affect both DNA strands. Consequences of frequent DNA lesions are shown on the right. The majority of lesions are processed by DNA repair processes, such as nucleotide excision repair (NER) and base excision repair (BER), before replication, with a fraction of them repaired incorrectly. Resulting mismatches can be repaired by MMR outside replication, with 50% of them being converted to mutations due to loss of information on the original nucleotide state. Alternatively, DNA damage is tolerated and damaged nucleotides proceed to replication without repair, which frequently generates mismatches or triggers recruitment of low-fidelity translesion polymerases (TLSs)[138]. Blue arrows represent repair pathways active in S phase, and grey arrows represent DNA repair pathways active outside S phase. The size of the oval qualitatively reflects the amount of lesions or mutations. In the lower figure part, red boxes indicate co-replicative processes, and purple boxes indicate replication-independent mutagenesis. DSB, DNA double-strand break; HDR, homology-directed repair.

R-asymmetry
The direction of the replication fork creates a natural asymmetry between the two DNA strands, which provides a very informative statistic, similar to T-asymmetry.

between the two strands, known as R-asymmetry. Also, DNA replication is accompanied by high activity of mismatch repair (MMR)[39] and homology-directed repair[40]. The variable activity of these repair systems along the genome, between leading and lagging DNA strands, and throughout the cell cycle leads to characteristic and potentially interpretable mutational patterns.

Mutations caused by DNA damage similarly leave statistically interpretable traces well beyond a simple correlation of the mutation rate with the exposure to a mutagen. There are three general pathways for DNA lesions to be converted to mutations (FIG. 2). First, smaller lesions do not prevent replication by high-fidelity polymerases[41], but DNA replication over the damaged nucleotide has much lower precision of nucleotide incorporation[39] due to the imperfect template. Second, the attempted repair may be erroneous, resulting in the DNA repair system placing a wrong nucleotide in place of the lesion[42]. Last, bulky lesions block the progression of DNA replication, requiring the recruitment

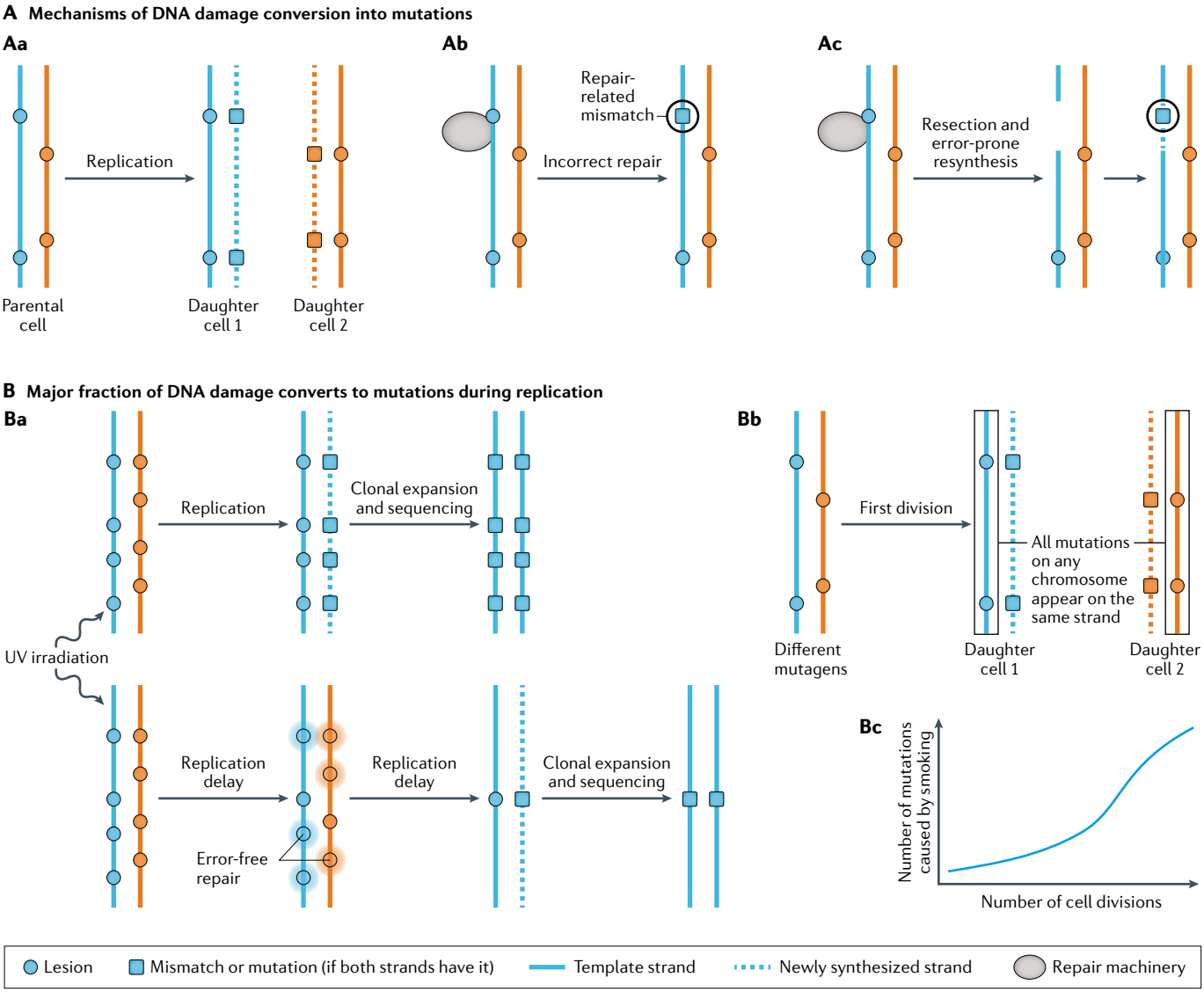
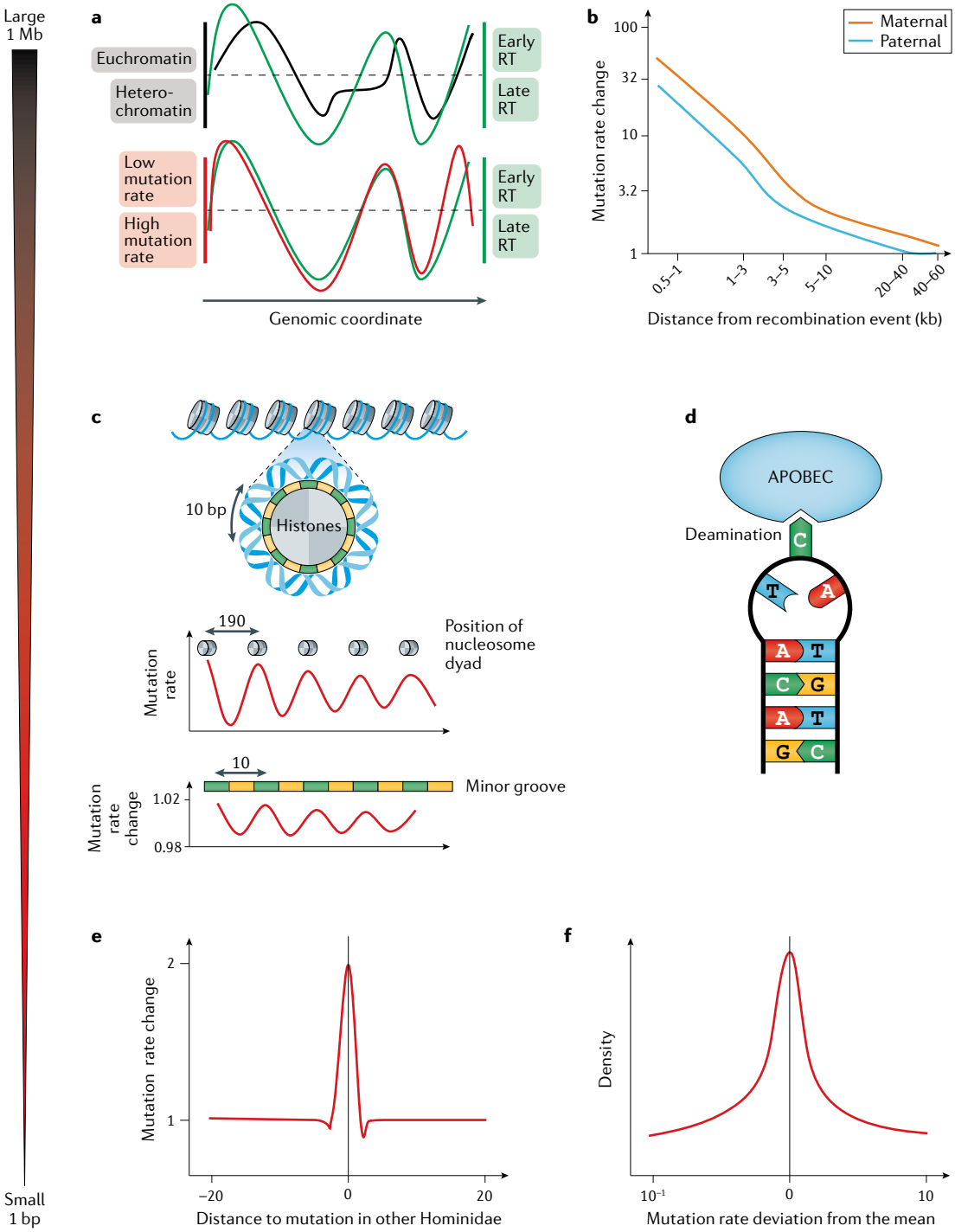


Fig. 2 | **Replication converts DNA damage into mutations. A** | Mechanisms of conversion of DNA damage into mutations are complex. They include inaccurate replication over DNA lesions[139–141] (part **Aa**), inaccurate repair inserting a wrong nucleotide instead of the lesion or opposing the lesion[42] (part **Ab**) and mutagenic repair leading to mutations in the vicinity of the lesion[42,142] (part **Ac**). **B** | Although all of these mechanisms contribute to the mutagenic effect of DNA damage, recent findings suggest that replication is by far the most prevalent mechanism converting lesions to mutations. The number of UV-induced mutations in clones created from an irradiated cell decreased by a factor of 30 when replication was delayed for 48 h post-UV pulse[96] (part **Ba**). This observation suggests that DNA repair before replication is essentially error-free, and that mutations appear during replication over unrepaired lesions. A recent cancer genomics study provides additional support for the role of replication in damage-induced mutagenesis[35]. After pulses of various mutagens, the vast majority of mutations in descendant tumour cells were shown to be caused by lesions in only one out of two DNA strands[35] (part **Bb**). This pattern suggests that all of the lesions on a chromosome are turned into mutations by a single event, most likely represented by replication. Finally, the cancer mutation signature driven by cigarette smoking is much more prevalent in actively dividing bronchi cells[36] (part **Bc**). Collectively, these findings support a model of DNA damage being primarily converted to mutations during replication and suggest that the rate of damage-induced mutations is tightly linked to the number of rounds of DNA replication (schematically summarized from REF.[36]).



of low-fidelity 'translesion' polymerases or leading to replication fork restart downstream of the damage with post-replicative resolution of the lesion featuring a translesion polymerase[38]. These polymerases generally have a larger active site, allowing passage across the bulky lesion while necessarily trading the replication accuracy of non-damaged DNA[43]. This process frequently results in mutations placed opposite the damaged nucleotide. It can also lead to additional mutations in adjacent sites, depending on how far the translesion polymerase travels before the high-fidelity polymerase is swapped back in[44]. There is a growing amount of evidence that replication through the lesion is the dominant mechanism of damage-induced mutagenesis for a wide variety of mutagenic agents[35] (FIG. 2).

If most lesions are converted to mutations during replication, the damage-induced mutation rate should scale with the amount of damage left unrepaired before being replicated during S phase. Thus, the relationship between the mutation rate and the activity of DNA repair systems along the genome carries information about the role of DNA damage in mutagenesis. Base excision repair (BER) and nucleotide excision repair (NER) are the pathways that specifically repair DNA lesions. Both

Bulky lesions
Characterized by DNA helix distortion, DNA damage that is detectable for global genomic nucleotide excision repair. Usually, bulky lesions are a serious obstacle for replication forks and transcription.

Fig. 3 | **Determinants of mutation rate variation on different scales. a** | Replication timing (RT) co-varies with large-scale (200–1,000 kb) epigenomic features, such as the euchromatin and heterochromatin state determined by Hi-C (high-throughput chromosome conformation capture). Replication timing strongly correlates with specific mutation types or processes[64,66,67,143]. **b** | De novo mutations co-occur with recombination events; the effect of recombination on the mutation rate is localized and becomes undetectable at distances of ~50 kb from recombination breakpoints (based on data from REF. [70]). **c** | Average distance between nucleosomes is 190 nucleotides (top panel). The mutation rate is slightly elevated under the nucleosome dyad, resulting in a periodicity 190 nucleotides long (middle panel). Additionally, DNA is wrapped around the nucleosome with a ten-nucleotide period. This periodicity also slightly influences susceptibility to mutations (bottom panel)[55]. **d** | Inverted repeats may allow formation of DNA hairpins. Loops of suspected hairpins mutate more frequently, possibly due to activity of APOBEC or as a consequence of exposure of single-stranded DNA to water[76,78]. **e** | Probability of mutation in the human genome is greatly increased at sites that are mutated in other Hominidae. This observation has been attributed to cryptic single-nucleotide mutational hotspots that are conserved between species[85–87]. **f** | Alternatively, hotspots in panel **e** may be driven by unaccounted effects of extended nucleotide context. Indeed, the mutation rate of some mutation types varies up to 100-fold due to extended contexts[89,90], as shown for the mutation rate distribution across heptanucleotide contexts of A>T mutation.

of these systems exhibit regional and sequence-specific variation in activity, as reviewed in subsequent sections.

Not all non-canonical nucleotides result from DNA damage. Cytosine methylation in CpG sites and cytosine hydroxymethylation are nucleotide modifications actively maintained by human cells. These common forms of non-canonical nucleotides also seem to be highly mutagenic[45].

For the sake of completeness, it is important to note that, in some cases, DNA damage introduced by mutagenic agents and replication fidelity and integrity are not independent. They can interact, breaking the simple dichotomy between damage-induced mutations and replication errors. In addition to directly attacking DNA strands, mutagenic agents may modify unincorporated nucleotides, producing pools of non-canonical bases that subsequently increase the rate of replication errors[46]. In addition to mutations introduced by translesion polymerases in sites opposing lesions (as discussed above), other mutations downstream of the damage may arise during translesion synthesis. Translesion polymerases synthesize short stretches of DNA following the lesion and introduce mutations due to their generally lower precision[17,43,44]. Additionally, replication fork stalling triggered by lesions occasionally drives massive perturbations in the replication process and even rearrangements[38]. Moreover, replication may expose single-stranded DNA to mutagens. For example, in cancer, APOBEC attacks the lagging strand during replication[47–49].

Mutational effect of deficiencies of co-replicative repair. The rate of nucleotide misincorporation during replication is kept in check by the exonuclease (that is, 'proofreading') activity of major DNA polymerases (Pol ε and Pol δ) and by co-replicative MMR (FIG. 1). Most of the information about the role of co-replicative repair in vivo comes from experimental systems with mutated Pol ε and Pol δ, and from MMR-deficient systems[14,33,50]. Loss of co-replicative repair is carcinogenic, and analyses of cancer genomes identified genome-wide patterns associated with loss of proofreading activity and loss of MMR[51,52].

Reduction of replication fidelity in these systems increases the mutation rate by one to three orders of magnitude[51,53]. Failure of co-replicative repair is characterized by unique mutational spectra that may serve as statistical signatures of specific deficiencies. Mutations are asymmetric with respect to the direction of the replication fork. Error-prone mutants of Pol ε and Pol δ introduce mutations on the leading strand and the lagging strand, respectively[50–52]. MMR is more proficient in removing replication errors on the lagging strand than on the leading strand[51,52]. MMR activity was also shown to correlate with the chromatin structure[53]. All of these features are potentially detectable in sequencing data.

Mutational properties of other DNA repair mechanisms. Most DNA lesions are repaired outside replication. NER is the major pathway to correct bulky DNA damage. NER employs a single mechanism to remove and replace the damaged nucleotide but consists of two separate pathways that locate the lesion on DNA. One branch of NER, global genomic NER (GG-NER), detects distortions in the DNA structure. This system, which scans genomic DNA, is more efficient in regions of active chromatin. GG-NER has reduced efficiency at DNA covered by the nucleosome dyad and shows 10-nucleotide periodicity within the dyad[54,55]. The second branch, transcription-coupled NER (TC-NER), is a mechanism of recruiting NER by the RNA polymerase. During transcription, bulky DNA lesions on the transcribed strand block the progression of RNA polymerase. The stalled RNA polymerase recruits the NER system to the damage. As a result, bulky DNA damage in actively transcribed genes is preferentially repaired on the transcribed strand, causing depletion of damage-induced mutations on this strand compared with the non-transcribed strand.

Deficiency of NER has been studied in the context of associated monogenic diseases. Xeroderma pigmentosum is a disease caused by the defect of the GG-NER branch leading to a higher level of somatic mutagenesis, especially on non-transcribed strands of genes[56] effectively invisible to TC-NER[57]. Loss of the TC-NER function causes Cockayne syndrome. In cells with TC-NER deficiency, mutations occur less asymmetrically between the transcribed and non-transcribed DNA strands[24].

BER is another system to actively repair damaged nucleotide bases. Lesion-specific glycosylases recognize DNA damage and excise the affected base, leaving the sugar DNA backbone intact. This process creates one or more abasic sites that are subsequently repaired. Recently, mutational spectra for human cells or cancers lacking the glycosylases OGG, MUTYH, UNG, NTHL1 and MBD4 have become available[58,59]. Regional variation of BER activity has been studied in the context of cancer genomics[60]. Some tumours are deficient in *MUTYH*, which encodes a glycosylase that recognizes 8-oxoguanine lesions. Somatic mutations in tumours with *MUTYH* deficiency suggest that BER efficiency has a very weak association with the megabase-scale chromatin structure, but decreases at the nucleosome dyad[55,60]. A similar effect of nucleosome positioning has been demonstrated for other types of DNA damage repaired by BER in yeast[61]. Considering the availability

of new data sets and unique mutational spectra observed under conditions of glycosylase deficiency, it may be possible to extract statistical signals of spatially variable BER efficiency in repair-proficient cells.

Many other repair systems deal with DNA double-strand breaks (DSBs) and other DNA damage[40,62,63], but their discussion is beyond the scope of this Review.

Mutation rate variation along the genome

Variation in the mutation rate along different genomic scales provides the most obvious statistical instrument for inferring mutational mechanisms from DNA sequencing data (FIG. 3). However, the observation of regional differences by itself does not point to a specific mutagenic force. If mutation rate variation across the genome is mimicking DNA susceptibility to a specific damage or inefficiency of a specific repair system, this damage or repair machinery plays a dominant role. In this section, we discuss factors that influence the mutation rate in the germ line and to what extent statistical observations may be linked to specific mutational pathways.

Large-scale chromatin structures and replication timing. A weak association between the mutation rate and timing of DNA replication has been known for over a decade[64]. The effect could be due to a reduction of replication accuracy caused by a depletion of the free nucleotide pool and frequent replication fork stalling late in S phase[65]. However, this interpretation is complicated by the strong correlation of replication timing with the large-scale chromatin structure (FIG. 3a), which is known to affect the activity of DNA repair systems. The efficiency of both NER and MMR systems has qualitatively similar associations with chromatin[53,56,60]. Mutations of the dinucleotides TC into TA have the strongest enrichment in late replicating regions, which correspond to heterochromatin, as was shown recently[66,67]. This observation may provide a link to one of the repair systems, once supplemented with knowledge of the sequence context-specific efficiency of NER and MMR. More generally, whereas qualitative patterns of mutation rate variation on a megabase scale could not clearly be linked to molecular mechanisms[68,69], models considering the sequence context and additional layers of information may help map mutational patterns to mechanisms.

Small-scale effect of recombination. In contrast to mutation heterogeneity on the megabase scale, factors shaping the mutation rate on the scale of hundreds and even dozens of nucleotides frequently imply a clear biological interpretation. Both direct[70,71] and indirect[72,73] evidence concordantly suggests that recombination is mutagenic. There is less agreement on the quantitative contribution of its role in mutagenesis, with early literature on the subject attributing from 1% to 10% of all mutations to this process[73–75]. Studies that directly map de novo recombination events and point mutations helped resolve the controversy and demonstrated that the density of nucleotide changes within a 1-kb window around a recombination event is 5–50-fold higher than the genome average[70,71] (FIG. 3b). Overall, these results suggest that a measurable effect of recombination induces less than 1% of all mutations genome-wide.

Nucleosome positioning and non-canonical DNA structures. Numerous studies point to the mutagenic effect of non-canonical DNA structures. For example, inverted repeats in proximity to each other show an increase in the mutation rate[76–78]. Such repeats can form hairpins, and DNA in the hairpin loop adopts a single-stranded conformation, making it susceptible to damage. Although the mutation rate of non-canonical DNA structures can be high under experimental conditions[78], detecting this signal in genomic mutation data is hampered by the rarity of highly structure-prone DNA in the genome, as well as by technical challenges owing to the difficulty of sequencing these regions.

At short scale with respect to sequence length, DNA structure is mostly affected by nucleosome positioning[55,79]. Data from cancer genomes and model organisms suggest that the mutation rate has a 10-nucleotide periodicity, corresponding to sites with the minor groove facing towards ('in') or away from ('out') the nucleosome. Interestingly, in these systems, damage-induced mutagenesis is slightly accelerated in positions with the minor groove facing out, echoing the activity of GG-NER[80], whereas replication mismatches are more frequent in positions with the minor groove facing in[54,55,60]. Germline mutations demonstrate a slight tendency to preferentially occur in the 'in' orientation (FIG. 3c). The overall magnitude of the effect is smaller than 5%[55], and the observation is not conclusive about the preferential role of replication errors in nucleosome-related periodicity. Although possibly playing a role, nucleosome positioning is not a major factor shaping the human germline mutation rate.

Transcription factor binding. Transcription factor binding was also shown to be mutagenic in cancer genomes, with the effect attributed to blocking access to the NER system[81,82] (FIG. 4) or, alternatively, to local changes in the DNA conformation induced by the transcription factor, which makes DNA more sensitive to UV light[12]. It was shown that the germline mutation rate is 20% higher in transcription factor-binding sites, but this effect can be explained by the biased nucleotide composition of these sites[83]. Further studies implementing more sophisticated methods are required to investigate the effect of transcription factor binding on the mutation rate beyond the nucleotide composition.

Single-nucleotide mutational hotspots. Strong single-nucleotide mutational hotspots were first discovered in human Mendelian genetics (see REF.[84] for a review). Most common hotspot examples are mutations giving proliferative advantage to spermatocytes and, as a consequence, frequently transmitted to offspring. Cryptic variation in the mutation rate at the single-nucleotide level is not limited to selfish selection in sperm. More than a decade ago, a twofold to threefold increase in the rate of human SNVs was observed in positions divergent between human and chimpanzee genomes[85–87] (FIG. 3e). It was estimated that some individual nucleotide sites have up to a 100-fold increase in the mutation rate[88].

Replication timing
Human DNA replication is a complex process involving hundreds of thousands of origins. Despite firing of individual replication origins being stochastic, genomic loci have a tendency to be replicated early or late in S phase.

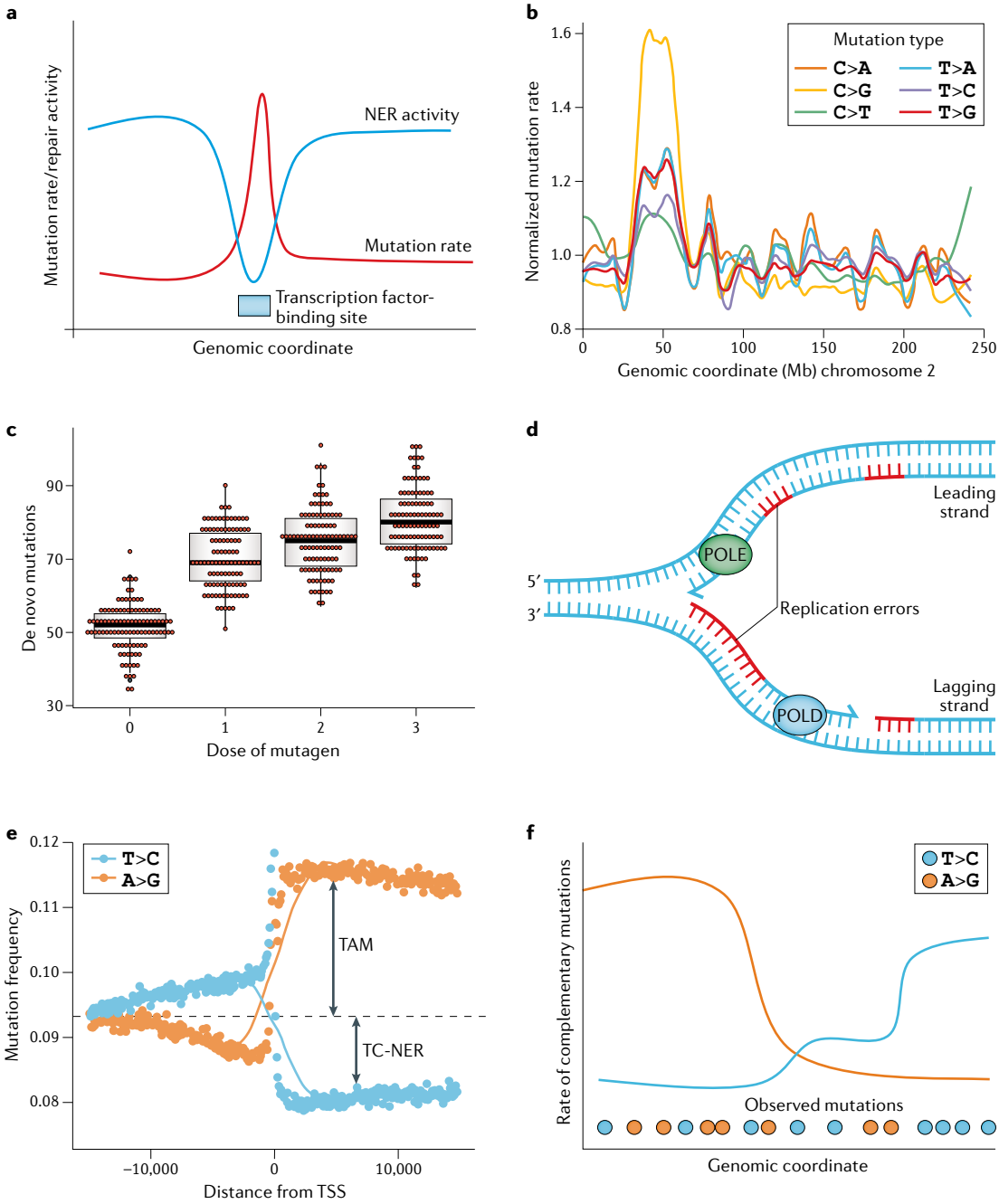


Fig. 4 | **Statistical footprints of mutational processes. a** | Localized spike of the mutation rate may point to a specific mechanism; for example, the rate of UV-induced mutations dramatically increases and the efficiency of nucleotide excision repair (NER) is decreased at transcription factor-binding sites. **b** | Regional increase of a certain mutation type can suggest an underlying mechanism. For example, loci with an elevated rate of C>G substitutions have a disproportionate contribution of mutations of maternal origin (calculated based on data from REF.[144]). **c** | Hypothetical example of a dose-dependent response of the mutation rate to an environmental agent. **d** | R-asymmetry reflects co-replicative mutation accumulation. **e** | T-asymmetry is the footprint of transcription-coupled NER (TC-NER) activity and transcription-associated mutagenesis (TAM) (calculated based on data from REF.[144]). **f** | Hypothetical example of localized imbalance of reverse complementary mutations suggesting the action of an asymmetric mutational process. POLD, polymerase δ; POLE, polymerase ε; TSS, transcriptional start site. Part **a** adapted from REF.[81], Springer Nature Limited.

More recently, efforts in statistical modelling attributed some hotspots to extended (heptameric) nucleotide contexts (FIG. 3f). Some non-CpG extended contexts result in an up to 13-fold increase in the mutation rate[89,90]. Mutagenicity of some of these contexts may be potentially linked to known biological processes. For example, A>T substitutions in the TTTAAAA context (mutated nucleotide underlined) are probably caused by polymerase slippage during replication[59,91]. These findings explain some of the 'cryptic effects' reported by

earlier studies, which only accounted for trinucleotide contexts[85–87].

Mechanism behind CpG to TpG substitutions. Cytosines followed by guanines, usually denoted CpG, are commonly methylated in humans. It is well known that the rate of CpG>TpG mutations is elevated ~15-fold compared with C>T transitions in other contexts[92–94]. The strong correlation of the CpG mutation rate with methylation levels suggests a key role for methylation in the hypermutability of CpG contexts. There are four possible mechanisms that may drive prominent mutability of methylated cytosines (mCs). First, deamination of a mC creates a T:G mismatch (instead of the canonical mC:G). T:G mismatches are usually repaired by BER to C:G. Erroneous repair of the mismatch to T:A will lead to mutation[42]. This mechanism causes CpG>TpG mutations independently of replication. Second, a potential mechanism generates mutations from mismatches that are formed immediately before cell division and unrepaired until replication. One of the two daughter cells inherits a T resulting from the unrepaired deamination of a mC. Third, deamination can happen co-replicatively. The deamination rate is probably increased for single-stranded DNA, and the lagging strand is exposed to the single-strand conformation during replication stress[95]. Fourth, mC may be an inferior template for major polymerases, and the high rate of CpG>TpG mutations can be a consequence of nucleotide misincorporation opposite mC.

CpG>TpG mutations are strongly asymmetric with respect to the direction of the replication fork. All CpG-bearing trinucleotide contexts, except ACG, show significantly higher (by 33%) mutability on the lagging strand[96–98] (FIG. 5). Neither mis-repair (mechanism 1) nor lack of repair (mechanism 2) can generate R-asymmetry. In the case of mechanism 2, T:G mismatch entering replication presents canonical nucleotides (T and G) as templates to both replicating DNA strands. By contrast, co-replicative mechanisms 3 and 4 are potentially R-asymmetric.

The replication-mediated mechanism 4 finds additional support in data from cancers with deficient MMR or with Pol ε lacking proofreading activity.

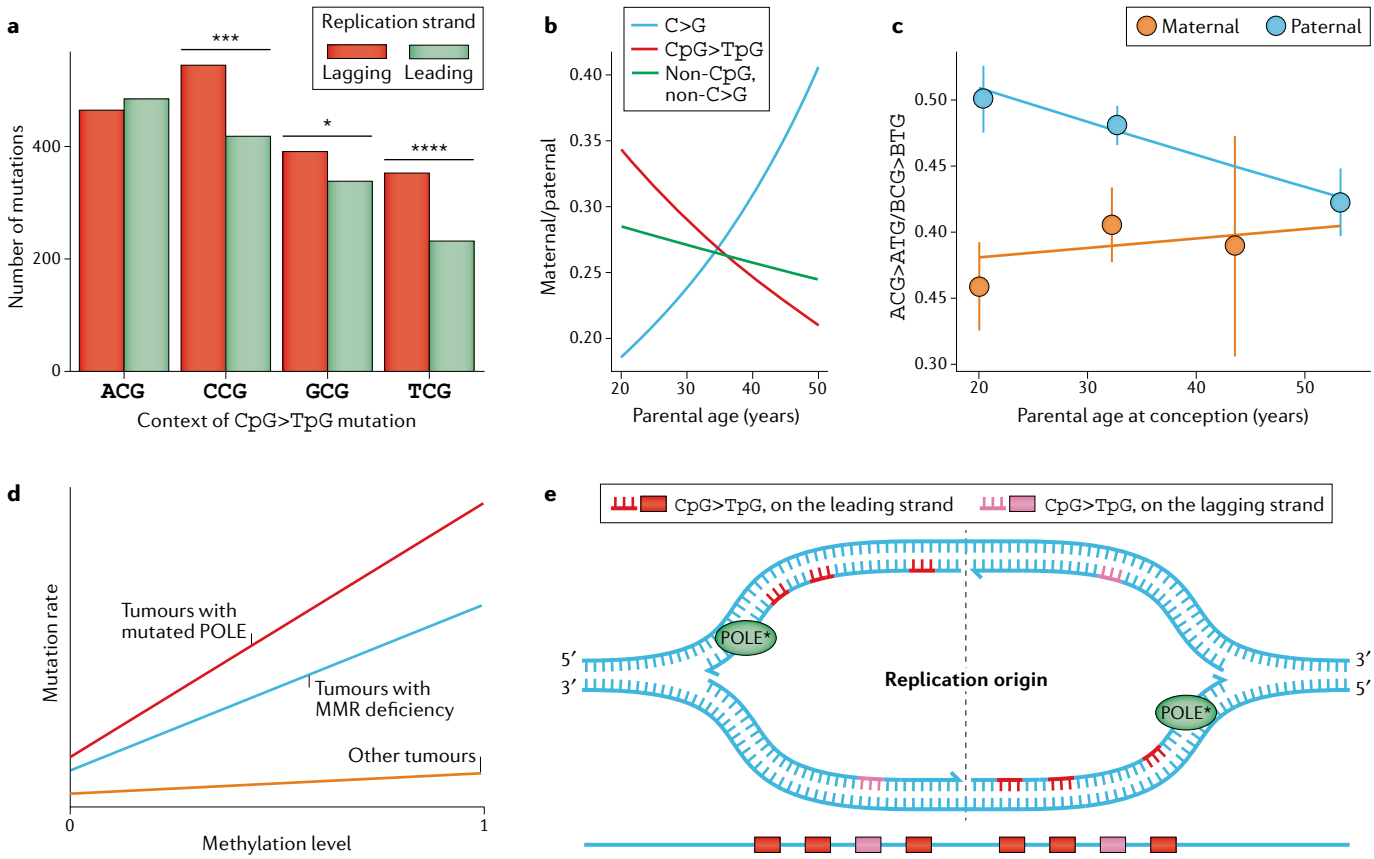


Fig. 5 | **Some CpG>TpG mutations are replicative origin. a** | CpG>TpG mutations other than ACG>ATG exhibit replication asymmetry in the human germ line (data merged from REFS[112,135]). *$P < 0.05$, ***$P < 0.001$ and ****$P < 0.0001$. **b** | Fractions of CpG>TpG and non-CpG mutations (excluding C>G) of maternal origin decrease with the average age of the mother and father. The decrease is strongest for CpG mutations (data from REF.[12]). **c** | Ratio between ACG>ATG and BCG>BTG mutations (where B stands for non-A nucleotides) decreases with age for paternal but not maternal mutations. Lines show estimates obtained with binomial regression. Points show the ratio between ACG>ATG and BCG>BTG for parents stratified by age into three bins (younger than 25 years, 25–40 years and older than 40 years). The *x* axis shows the mean age of parents within the bin; error bars are 95% binomial confidence intervals. **d** | Mutation rate in cancers with compromised mismatch repair (MMR) or disrupted exonuclease activity of polymerase ε (POLE) is elevated in a CpG context in a methylation-dependent manner[97]. **e** | CpG>TpG mutations in cancers with compromised MMR or disrupted exonuclease activity (denoted by *) of POLE demonstrate a high level of replication asymmetry[97].

T-asymmetry
The direction of transcription creates a natural asymmetry between the two DNA strands. Although it is not always possible to estimate the mutation rate separately on the transcribed and the non-transcribed strands, because fixed mutations affect both strands, it is easy to calculate the imbalance between reverse complementary mutations located on one of the strands. For example, if lesions leading to A>G mutations are depleted on the transcribed strand, then the A>G rate would be higher than the T>C rate on this strand. If the aetiology of the damaging agent and resulting types of mutations are known, one may estimate differences in the mutation rate between the two strands.

R-loops
DNA–RNA hybrids that are usually formed co-transcriptionally on the transcribed strand. R-loops are important in the context of mutagenesis, because they stabilize non-transcribed strands in the single-stranded state.

The CpG>TpG mutation rate in these cancers is elevated, sensitive to the methylation level and shows remarkable R-asymmetry[97,99]. It is not always correct to extend insights from cancer data to germline mutagenesis. However, individuals with inherited Pol ε mutation show noticeable similarities of mutational spectra between somatic cells and germ cells[100].

Assuming a constant deamination rate and efficiency of repair, the rate of mutations generated by mis-repair (mechanism 1) is expected to scale with time rather than the number of cell divisions. Many publications, especially in the field of molecular evolution, use the number of CpG mutations as a proxy for physical time[101,102]. They argue that the rate of CpG>TpG substitutions accumulated along phylogenies varies less across species than other mutation types. By contrast, mutations caused by unrepaired T:G mismatches (mechanism 2) are expected to scale with the number of cell divisions because BER is efficient at short timescales[103–105], and only mismatches arising a few minutes before replication are expected to be left unrepaired. Indeed, dysfunction of the MBD4 glycosylase — the protein responsible for repair of mC deamination — delays BER and leads to a substantially higher rate of transitions in the CpG context[27,58]. Co-replicative mechanisms (mechanisms 3 and 4) obviously scale with the number of cell divisions rather than time. It was shown that the CpG>TpG mutation rate strongly correlates with the rate of cell divisions in somatic cells and cancer precursors[26,106,107]; the CpG mutation rate is negligible in non-dividing neurons[24].

Oocytes are arrested non-dividing cells throughout most of the mother's life, so the only mutation process occurring in non-dividing cells (mechanism 1) can increase with maternal age. CpG>TpG mutations have the weakest, albeit still significant, association with maternal age[108]. As a result, the fraction of de novo CpG mutations occurring in oocytes decreases with maternal age[108,109]. By contrast, the number of CpG mutations strongly increases with paternal age in constantly dividing sperm cells[108–110]. Some of the differences between maternal and paternal CpG mutations may be attributed to differences in methylation levels; the fraction of methylated CpG sites in germinal vesicle oocytes is slightly lower than in spermatocytes (~0.5 versus ~0.9)[111]. However, this difference should not be sufficient to explain the fourfold difference in the number of CpG mutations inherited from the mother and the father[93,94,110,112]. It is also unlikely to fully explain the magnitude of the difference in age dependencies between parents.

ACG>ATG mutations lack R-asymmetry and, therefore, may less frequently originate by mechanisms 3 and 4. Interestingly, the fraction of ACG>ATG mutations of paternal, but not of maternal, origin significantly decreases with age at conception ($P = 0.025$ and $P = 0.73$, binomial regression for paternal and maternal mutations, respectively; FIG. 5c). This observation suggests that R-asymmetry may be informative about scaling with the replication rate.

These observations suggest that there is more to CpG mutagenesis than mis-repair of deaminated methylcytosines (mechanism 1); co-replicative deamination and reduced fidelity of replication over methylcytosines (mechanisms 3 and 4) probably play a non-negligible role.

Mutational asymmetry between DNA strands

The key feature of the structure of DNA molecules is the symmetry of the two strands. Given this symmetry, it would be natural to expect that reverse complementary mutations would exhibit identical statistical properties. However, this is not always the case. Processes such as transcription and replication break symmetry between DNA strands (that is, into transcribed and non-transcribed strands of a gene, or into leading and lagging strands). In turn, the mutagenic forces coupled to transcription and replication act in a strand-dependent manner. These differences are imprinted in sequence data as imbalances between reverse complementary mutations and are called transcriptional asymmetry (T-asymmetry) and R-asymmetry[48]. In addition, strand asymmetries with respect to yet unknown DNA features may be studied in an unbiased way by calculating imbalances of complementary mutations within loci[67] (FIG. 4). Mutational asymmetries provide a very clear statistical signal that in many cases can easily be mapped to a mutational mechanism.

Transcriptional asymmetry. One of the major footprints observed in mutational data is left by TC-NER. Recruitment of the NER system to bulky DNA lesions on the transcribed strand of genes by stalled RNA polymerase causes T-asymmetry (reviewed in REF.[113]). Therefore, depletion of some mutation types in a strand-specific manner within gene bodies can be interpreted as a footprint of TC-NER activity. The analysis of asymmetry with respect to the direction of transcription reveals mutation types caused by bulky DNA damage without any previous knowledge of the source of mutagenic exposure[96,114]. The depletion of damage-induced mutations on the transcribed strand compared with adjacent non-genic regions provides an estimate for the lower bound of damage-induced mutations throughout the genome. In the human germ line, at least 10% of all mutations are caused by bulky damage, judging from the degree of T-asymmetry[96].

Interestingly, the mutation rate may also increase on the non-transcribed strand of genes rather than in surrounding intergenic regions[114,115] (FIG. 4e). The increase in mutation rate associated with transcription is called 'transcription-associated mutagenesis'. The exact mechanism leading to transcription-associated mutagenesis remains unknown[114,115], but plausible hypotheses include a mutagenic effect of DNA cleavage by topoisomerases or the formation of RNA–DNA hybrids, so-called R-loops, which may leave the non-transcribed strand unpaired and exposed to damage (reviewed in REF.[116]).

Replication asymmetry. R-asymmetry has been reported for human germline mutations and for various cancers[48,96,98,117]. As discussed above, the literature offers two different, although not mutually exclusive, explanations for this asymmetry. The most obvious explanation attributes the effect to replication-induced mutations. The asymmetry may arise from differential fidelity of the

two major polymerases or from differential efficiency of co-replicative repair. Indeed, cancers with mutations of major polymerases exhibit the highest levels of R-asymmetry[48,51,52,118].

Alternatively, R-asymmetry may be attributed to damage-induced mutations. Bulky damage blocking the progression of major polymerases may be resolved asymmetrically between leading and lagging strands. Bulky damage may trigger fork collapse followed by error-free repair on the leading strand and error-prone bypass on the lagging strand[38]. For example, mutation signatures associated with bulky mutagens show R-asymmetry in blood cancers[119]. Also, a large fraction of mutation types in the germ line have correlated levels of T-asymmetry and R-asymmetry, suggesting a role for differential damage resolution underlying both types of asymmetry[96].

With the help of more subtle statistical approaches, it is possible to extract mutational processes in the germ line that have a much stronger R-asymmetry than T-asymmetry, separating a class of mutations primarily originating as replication errors[67].

Influence of parental age and sex

Sequencing parent–child trios is the most direct way to detect de novo mutations and study mutagenesis in the germ line by statistical means. In contrast to other data types, trio sequencing studies help identify sex-specific patterns and the effect of parental age at conception. Trio sequencing can also discriminate between mutations private to sperm or oocytes, or mutations occurring early in development that affect somatic and germline cells. Contrasting rates and patterns of mutations in children of old versus young parents and mutations of maternal versus paternal origin is another tool to infer mechanisms of mutations.

Sex-specific differences in the rate and spectra of mutations. Overall, the number of mutations accumulated during spermatogenesis exceeds the number of mutations accumulated in oogenesis by a factor of four[93,94,109,110,120]. The most likely explanation for the increased mutation rate in the male germ line is that spermatocytes go through multiple rounds of cell divisions, whereas oocytes remain arrested in prophase I of meiosis. Despite the dissimilarity between female and male germ lines and the larger male contribution to the overall mutation rate, mutations of maternal and paternal origin have similar distributions across all six types of nucleotide substitutions. Only four out of six single-nucleotide mutation types are significantly different between sexes, and these differences do not exceed 40%[109]. However, mutation spectra appear more different if the extended sequence context is taken into account. For example, both transversions in AGCCT context (mutated nucleotide underlined) are 12-fold more common in the male germ line than in the female germ line, which is threefold more than randomly expected (FIG. 6a). Using data from Jónsson et al.[109], we identified 10 pentanucleotides most significantly biased towards either maternal or paternal mutations. These pentanucleotides are twice as likely to mutate in mothers than in fathers (or the other way around), as evident from an independent data set[110]. Numerous additional contexts deviate from the ratio expected for the corresponding single-nucleotide and CpG>TpG mutation types (FIG. 6b). These context-specific differences probably point to mutational mechanisms specific to spermatogenesis or oogenesis. However, extreme outliers are uncommon, and generally deviations are observed in opposite directions, making these observations consistent with quantitatively modest differences between paternal and maternal mononucleotide mutation spectra and the mostly stable ratio of maternal and paternal age effects on the total number of mutations[108].

Age-dependent mutation accumulation in oocytes. Oocytes do not replicate postnatally, but the number of maternal mutations increases substantially with the mother's age at conception[109,120,121]. This implies that the effect of maternal age on the mutation rate cannot be driven by a co-replicative process. As expected for mutations caused by DNA damage, maternal mutations show strong T-asymmetry compared with paternal mutations ($P = 8.37 \times 10^{-6}$, chi-squared test of homogeneity; FIG. 6c). Mutations accumulated in oocytes are highly enriched in several regions of the genome; just 10% of the genome is responsible for 35% of the maternal age effect and harbours 22% of all mutations of maternal origin[67]. Noticeably, these regions have characteristic mutational spectra marked by a high fraction of C>G mutations, probably reflecting a mechanism of mutagenesis specific to oocytes. Maternal mutations in these loci tend to occur in clusters[67,109,121]; two or more mutations of maternal origin in an offspring co-localize within a few kilobases with a frequency dramatically exceeding random expectation, suggesting that they arose from a single complex event. Clusters are approximately sixfold more frequent in these regions[109,121]. Mutation clusters are also highly enriched in C>G substitutions.

Although we lack a mechanistic understanding of localized mutagenesis in oocytes, two non-mutually exclusive hypotheses exist. The first hypothesis assumes that point mutations are a by-product of the repair of DSBs that accumulate with age[70,109,121]. The second hypothesis considers DNA lesions as a major contributor to maternal mutations[67,108]. If the lesions are left unrepaired or the repair is error-prone, the resulting mutations would scale with the mother's age.

The DSB hypothesis relies on the solid experimental evidence that the homologous repair system is progressively less efficient in ageing oocytes[122,123]. It is also supported by the spatial correlation between the maternal mutation rate (primarily for C>G mutations) and the non-crossover recombination rate, the latter serving as a proxy for the DSB rate[109]. One of the caveats of the DSB explanation is the difference between the known mutation spectra of recombination-induced mutations[70] and the mutation spectra in regions with a strong maternal age effect[67,109,121].

The main observation that supports DNA damage as an explanation of the maternal age effect is a very strong T-asymmetry in regions with an accelerated maternal age effect[67]. It is also known that DNA lesions mediate

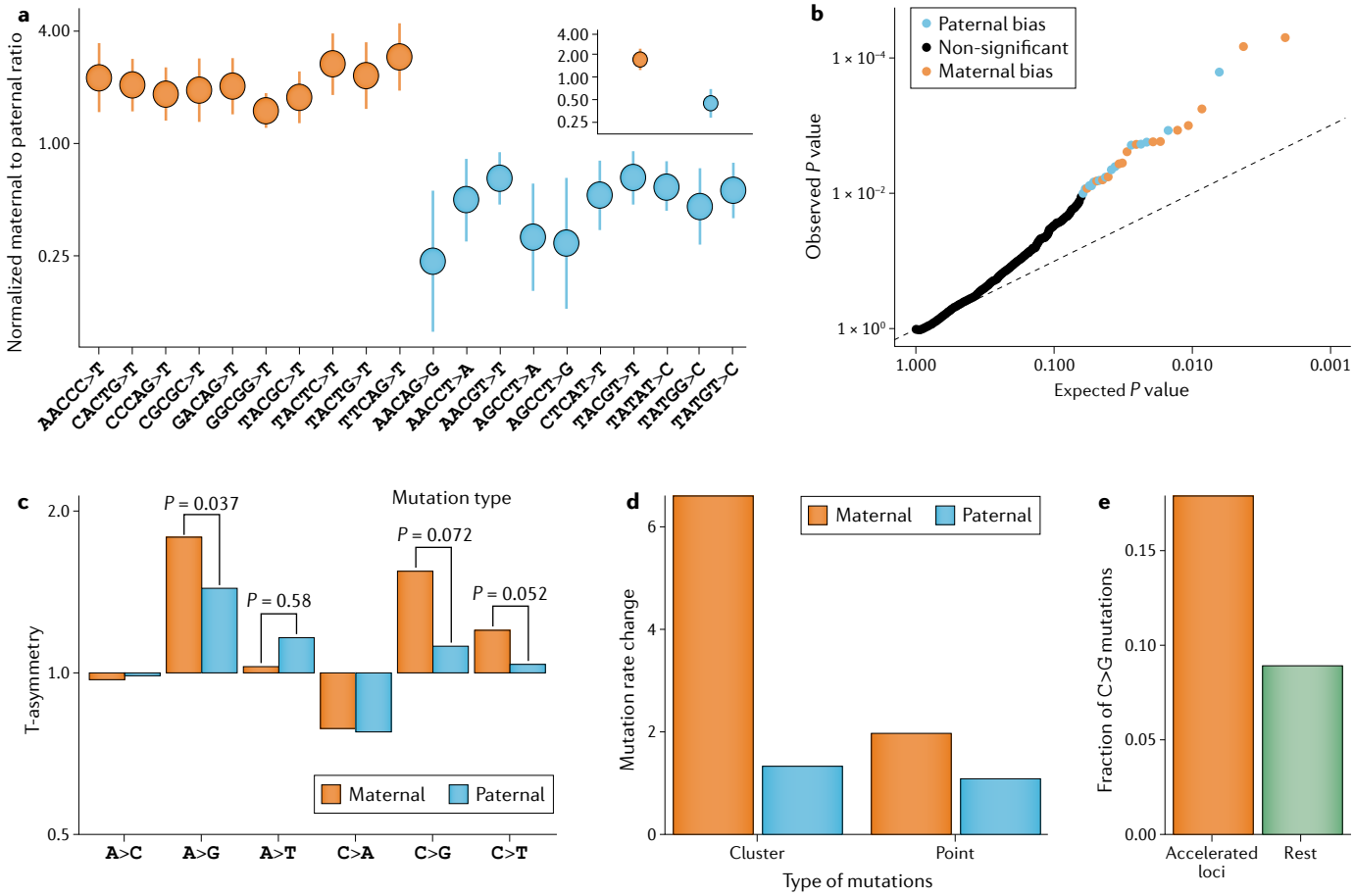


Fig. 6 | **Sex-specific mutational patterns. a** | Ratios of maternal to paternal mutations in specific pentanucleotide contexts. The ratios are normalized by the maternal to paternal ratio for all mutations. Ten contexts most over-represented among maternal mutations (ratios above one) shown in pink, and ten contexts most over-represented among paternal mutations (ratios below one) shown in cyan (data from Jónsson et al.[112]). The average maternal and paternal bias for the same ten maternal and ten paternal contexts estimated in an independent data set is shown in the inset (data from Goldmann et al.[110]). **b** | Quantile–quantile (Q–Q) plot of chi-squared test *P* values of the maternal to paternal ratio deviation from the mean for all pentanucleotide contexts with more than 50 mutations (data from Jónsson et al.[112]). **c** | Level of T-asymmetry for different mutation types for maternal and paternal mutations. Overall, T-asymmetry of maternal mutations is significantly higher ($P = 8.37 \times 10^{-6}$, chi-squared test of homogeneity; data from Jónsson et al.[112]). **d** | Regions with an accelerated maternal mutation rate show a sixfold increase of clustered mutations and a twofold increase of other point mutations (data from Jónsson et al.[112]; annotations of maternal regions from Seplyarskiy et al.[67]). **e** | Fraction of C>G mutations among all maternal mutations in genomic regions with higher maternal mutation rates (accelerated regions) and in the remaining genome.

DSB formation[124], so the correlation between the DSB rate and the mutation rate may be driven by an uneven distribution of DNA lesions across the genome.

Age-dependent mutation accumulation in sperm. The discovery of the effect of paternal age on mutation rate predates the discovery of DNA structure. Recent data recapitulate a linear dependence between paternal age and the number of mutations of paternal origin[93,94,125]. This observation was considered evidence for the major role of replicative errors leading to mutations arising in spermatocytes. However, as discussed above, theoretical considerations suggest that DNA damage may leave the same statistical footprint, making it harder to establish the origin of paternal mutations from trio sequencing data. Notably, the age-dependent accumulation of mutations in both parents results in a relatively stable ratio between maternal and paternal mutations across ages of conception[108].

Variability in mutation rate between families. Early trio sequencing studies suggested that the mutation rate is not substantially variable in the human population. Recent studies are inconclusive about mutation rate variation between families, with two papers reporting statistically significant variability in both the number and the patterns of mutations[30,126]. Another re-analysis study claims that this effect is very minor[127]. According to another recent report, the mutation rate in the Amish population is reduced by ~10% compared with other populations[73]. None of these studies was able to attribute the variability in mutation rate to genetic background and suggested that mutation rate differences, if present, are primarily environmental. Still, genetic changes in some individuals may have a profound effect on the mutation rate. Preliminary results of an ongoing study suggested that rare genetic deficiency in repair systems increases the number of de novo nucleotide changes by a factor of four[128]. Rare genetic repair deficiencies are

unlikely to generate substantial mutation rate variation at the population scale.

Early developmental mutations

Mutations that occur early in development may be broadly represented across tissues, including blood and germ cells. If present in germ cells, these mutations can be transmitted to children. Early developmental mutations have different properties from transmittable mosaic mutations that originated during gametogenesis, which are present in a fraction of germ cells but absent from other tissues[29,30,129]. A growing body of evidence suggests that the mutation rate during the first zygotic divisions is threefold to tenfold higher than the average rate during gametogenesis[25,130]. Mutations believed to arise early in development are characterized by a specific spectrum marked by elevated rates of TCT>TAT and GCA>GAA substitutions[29,131,132].

It is still unclear whether this mutation rate increase is unique to early zygotic divisions or simply reflects the higher mutation rate per division in any tissue compared with the differentiated germ line[130]. If it is a specific feature of early cell divisions, two biological explanations could be readily offered. First, DNA repair might be inefficient before zygotic genome activation, which takes place at the stage of the third cell division[108,133]. Alternatively, a temporary increase of the mutation rate during the first divisions could be caused by unrepaired DNA damage accumulated in gametes[108,134].

Overall, a fraction (possibly as high as 5%) of de novo mutations arise in parents as early developmental mutations[30,112]. Owing to a smaller number of mutations of maternal origin, a higher proportion of maternal mutations are not private to the germ line[112].

Conclusions

The field of germline mutagenesis is entering a new exciting phase. Extensive sequencing data offer a clear picture of the mutation process as it occurs in humans, as opposed to an artificial experimental system. Genomic data clearly point to a role of DNA damage in generating some de novo human mutations. It is impossible to explain asymmetry of mutation spectra with respect to transcription direction (that is, gene orientation) by replication errors. This asymmetry can easily be explained by the action of TC-NER repairing bulky DNA lesions.

By contrast, large-scale sequencing data reveal features of replicative origin for one class of mutations that was widely believed to only result from DNA damage[101,102,106]. From experimental systems, it is known that transitions within the CpG context (CpG>TpG mutations) can arise from spontaneous deamination of methylated cytosines. Genome-scale analysis demonstrated that the rate of CpG mutations depends on the direction of replication, implying that some CpG transitions result from base misincorporation during replication or from co-replicative deamination.

The analysis of sequencing data has also identified numerous surprising patterns that await mechanistic interpretation. Clustered mutations that accumulate in ageing oocytes provide an arguably most interesting example. The mutation clusters arise in non-dividing cells and cannot be attributed to inaccurate replication. They point to a specific mutagenic process that is highly localized in the genome and produces multiple mutations at a scale of tens of kilobases.

In this Review, we argue that the computational analysis of genomic data sets should not be devoid of the body of existing knowledge on DNA replication and repair. Combined approaches that simultaneously aim for statistical detection of footprints of known mutagenic processes in the data and reconstruction of observed patterns in controlled experimental systems seem most promising. We believe that extensive collaborative efforts involving groups across fields ranging from DNA replication and repair to statistical and population genetics could be a highly productive way to translate analyses of mutational patterns into biochemical mechanisms. Indeed, we are optimistic that new data combined with innovative analytic approaches will explain statistical patterns in data with biochemical models.

Data availability

De novo mutations from trio sequencing studies were obtained from REFS[110,112,135]. The authors used rare polymorphisms from gnomAD v2 (https://gnomad.broadinstitute.org/downloads).

Published online 23 June 2021

1. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
2. Kosmicki, J. A. et al. Refining the role of de novo protein truncating variants in neurodevelopmental disorders using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
3. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
4. Hoang, M. L. et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* **5**, 197ra102 (2013).
5. Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
6. Huang, M. N. et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* **27**, 1475–1486 (2017).
7. Liu, J. F., Konstantinopoulos, P. A. & Matulonis, U. A. PARP inhibitors in ovarian cancer: current status and future promise. *Gynecol. Oncol.* **133**, 362–369 (2014).
8. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).
9. Kurat, C. F., Yeeles, J. T. P., Patel, H., Early, A. & Diffley, J. F. X. Chromatin controls DNA replication origin selection, lagging-strand synthesis, and replication fork rates. *Mol. Cell* **65**, 117–130 (2017). **This study reports a clever experimental system that recreates replication in vivo**.
10. Devbhandari, S., Jiang, J., Kumar, C., Whitehouse, I. & Remus, D. Chromatin constrains the initiation and elongation of DNA replication. *Mol. Cell* **65**, 131–141 (2017).
11. Adar, S., Hu, J., Lieb, J. D. & Sancar, A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl Acad. Sci. USA* **113**, E2124–E2133 (2016). **This study creates a single-nucleotide resolution map of NER activity in UV-irradiated cells**.
12. Mao, P. et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* **9**, 2626 (2018).
13. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
14. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
15. Volkova, N. V. et al. Mutational signatures are jointly shaped by DNA damage and repair. *Nat. Commun.* **11**, 2169 (2020). **This article is a comprehensive study of the mutational footprints of DNA mutagens and repair deficiencies in *Caenorhabditis elegans***.
16. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019). **This study creates an encyclopaedia of mutational signatures caused by mutagenic agents in human cells**.
17. Segovia, R., Shen, Y., Lujan, S. A., Jones, S. J. M. & Stirling, P. C. Hypermutation signature reveals a slippage and realignment model of translesion synthesis by Rev3 polymerase in cisplatin-treated yeast. *Proc. Natl Acad. Sci. USA* **114**, 2663–2668 (2017).

18. Törmä, L., Burny, C., Nolte, V., Senti, K.-A. & Schlötterer, C. Transcription-coupled repair in *Drosophila melanogaster* is independent of the mismatch repair pathway. Preprint at *bioRxiv* https://doi.org/10.1093/bib/bbaa295 (2020).
19. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
20. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
21. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
22. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell* **177**, 608–621.e12 (2019).
23. Franco, I. et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biol.* **20**, 285 (2019).
24. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
This study describes mutational processes that operate in non-dividing neurons.
25. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
26. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
27. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
28. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
29. Harland, C. et al. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. Preprint at *bioRxiv* https://doi.org/10.1101/079863 (2017).
30. Sasani, T. A. et al. Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Preprint at *bioRxiv* https://doi.org/10.1101/552117 (2019).
31. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
32. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
33. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
34. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
This paper is a milestone in the statistical analysis of mutational signatures extracted from cancer genomic data.
35. Aitken, S. J. et al. Pervasive lesion segregation shapes cancer genome evolution. *Nature* **583**, 265–270 (2020).
36. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
37. Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the dependence of mutation rates on age and time. *PLoS Biol.* **14**, e1002355 (2016).
This paper develops a theory to investigate the relationship between damage-induced mutations and the replication rate.
38. Yeeles, J. T. P., Poli, J., Marians, K. J. & Pasero, P. Rescuing stalled or damaged replication forks. *Cold Spring Harb. Perspect. Biol.* **5**, a012815 (2013).
39. Kunkel, T. A. & Erie, D. A. Eukaryotic mismatch repair in relation to DNA replication. *Annu. Rev. Genet.* **49**, 291–313 (2015).
40. Chen, C.-C., Feng, W., Lim, P. X., Kass, E. M. & Jasin, M. Homology-directed repair and the role of BRCA1, BRCA2, and related proteins in genome integrity and cancer. *Annu. Rev. Cancer Biol.* **2**, 313–336 (2018).
41. Hsu, G. W., Ober, M., Carell, T. & Beese, L. S. Error-prone replication of oxidatively damaged DNA by a high-fidelity DNA polymerase. *Nature* **431**, 217–221 (2004).
42. Chen, J., Miller, B. F. & Furano, A. V. Repair of naturally occurring mismatches can induce mutations in flanking DNA. *eLife* **3**, e02001 (2014).
43. Goodman, M. F. & Woodgate, R. Translesion DNA polymerases. *Cold Spring Harb. Perspect. Biol.* **5**, a010363 (2013).
44. Kochenova, O. V., Daee, D. L., Mertz, T. M. & Shcherbakova, P. V. DNA polymerase ζ-dependent lesion bypass in *Saccharomyces cerevisiae* is accompanied by error-prone copying of long stretches of adjacent DNA. *PLoS Genet.* **11**, e1005110 (2015).
45. Supek, F., Lehner, B., Hajkova, P. & Warnecke, T. Hydroxymethylated cytosines are associated with elevated C to G transversion rates. *PLoS Genet.* **10**, e1004585 (2014).
46. Satou, K., Kawai, K., Kasai, H., Harashima, H. & Kamiya, H. Mutagenic effects of 8-hydroxy-dGTP in live mammalian cells. *Free Radic. Biol. Med.* **42**, 1552–1560 (2007).
47. Seplyarskiy, V. B. et al. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **26**, 174–182 (2016).
48. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
This article is the first systematic study of T-asymmetry and R-asymmetry in cancer.
49. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
This study investigates differences in mutation rate distributions between mutational processes.
50. Lujan, S. A. et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.* **24**, 1751–1764 (2014).
51. Andrianova, M. A., Bazykin, G. A., Nikolaev, S. I. & Seplyarskiy, V. B. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. *Genome Res.* **27**, 1336–1343 (2017).
This paper provides statistical evidence that MMR is more active on the lagging strand in human cells.
52. Haradhvala, N. J. et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018).
53. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
54. Mao, P., Smerdon, M. J., Roberts, S. A. & Wyrick, J. J. Asymmetric repair of UV damage in nucleosomes imposes a DNA strand polarity on somatic mutations in skin cancer. *Genome Res.* **30**, 12–21 (2020).
55. Pich, O. et al. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell* **175**, 1074–1087.e18 (2018).
56. Zheng, C. L. et al. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Rep.* **9**, 1228–1234 (2014).
57. Hu, J., Adebali, O., Adar, S. & Sancar, A. Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl Acad. Sci. USA* **114**, 6758–6763 (2017).
58. Sanders, M. A. et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood* **132**, 1526–1534 (2018).
59. Zou, X. et al. Dissecting mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.04.234245 (2020).
60. Vöhringer, H. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Preprint at *bioRxiv* https://doi.org/10.1101/850453 (2019).
This article presents a fascinating tool that uses differences in the spatial distribution of mutational processes to extract mutational signatures from cancer genomes.
61. Mao, P. et al. Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Res.* **27**, 1674–1684 (2017).
62. Malkova, A. & Ira, G. Break-induced replication: functions and molecular mechanism. *Curr. Opin. Genet. Dev.* **23**, 271–279 (2013).
63. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* **79**, 181–211 (2010).
64. Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
This study is the first to find the association between replication timing and the mutation rate.
65. Rhind, N. & Gilbert, D. M. DNA replication timing. *Cold Spring Harb. Perspect. Biol.* **5**, a010132 (2013).
66. Agarwal, I. & Przeworski, M. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc. Natl Acad. Sci. USA* **116**, 17916–17924 (2019).
67. Seplyarskiy, V. B. et al. Population sequencing data reveal a compendium of mutational processes in human germline. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.10.893024 (2020).
This study is the first to use variation in mutational spectra across the genome to extract mutational processes in the human germ line.
68. Terekhanova, N. V., Seplyarskiy, V. B., Soldatov, R. A. & Bazykin, G. A. Evolution of local mutation rate and its determinants. *Mol. Biol. Evol.* **34**, 1100–1109 (2017).
69. Smith, T. C. A., Arndt, P. F. & Eyre-Walker, A. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet.* **14**, e1007254 (2018).
70. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
This study provides direct genome-wide data on the relation between crossovers, complex crossovers and the mutation rate.
71. Arbeithuber, B., Betancourt, A. J., Ebner, T. & Tiemann-Boege, I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl Acad. Sci. USA* **112**, 2109–2114 (2015).
72. Huang, S.-W., Friedman, R., Yu, N., Yu, A. & Li, W.-H. How strong is the mutagenicity of recombination in mammals? *Mol. Biol. Evol.* **22**, 426–431 (2005).
73. Kessler, M. D. et al. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc. Natl Acad. Sci. USA* **117**, 2560–2569 (2020).
74. Duret, L. & Arndt, P. F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4**, e1000071 (2008).
75. Spencer, C. C. A. et al. The influence of recombination on human genetic diversity. *PLoS Genet* **2**, e148 (2006).
76. Zou, X. et al. Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res.* **45**, 11213–11221 (2017).
77. Löytynoja, A. & Goldman, N. Short template switch events explain mutation clusters in the human genome. *Genome Res.* **27**, 1039–1049 (2017).
78. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).
79. Li, C. & Luscombe, N. M. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nat. Commun.* **11**, 1363 (2020).
80. Brown, A. J., Mao, P., Smerdon, M. J., Wyrick, J. J. & Roberts, S. A. Nucleosome positions establish an extended mutation signature in melanoma. *PLoS Genet* **14**, e1007823 (2018).
81. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lõpez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
82. Perera, D. et al. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
Together with Sabarinathan et al. (2016), this article shows that the damage-induced mutation rate is increased at transcription factor-binding sites, probably due to the interference between NER and transcription factor binding.
83. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
84. Goriely, A. & Wilkie, A. O. M. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
85. Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A. Cryptic variation in the human mutation rate. *PLoS Biol.* **7**, e1000027 (2009).
86. Seplyarskiy, V. B., Kharchenko, P., Kondrashov, A. S. & Bazykin, G. A. Heterogeneity of the transition/transversion ratio in *Drosophila* and Hominidae genomes. *Mol. Biol. Evol.* **29**, 1943–1955 (2012).
87. Johnson, P. L. F. & Hellmann, I. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol. Evol.* **3**, 842–850 (2011).

88. Smith, T. et al. Extensive variation in the mutation rate between and within human genes associated with Mendelian disease. *Hum. Mutat.* **37**, 488–494 (2016).
89. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48**, 349–355 (2016).
90. Carlson, J. et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* **9**, 3753 (2018).
91. Montgomery, S. B. et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761 (2013).
92. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
93. Francioli, L. C. et al. Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
94. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
95. Klein, H. L. Stressed DNA replication generates stressed DNA. *Proc. Natl Acad. Sci. USA* **117**, 10108–10110 (2020).
96. Seplyarskiy, V. B. et al. Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nat. Genet.* **51**, 36–41 (2019).
97. Poulos, R. C., Olivier, J. & Wong, J. W. H. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.* **45**, 7786–7795 (2017).
This study shows that deficiency of co-replicative repair increases the rate of CpG mutations in cancer genomes.
98. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
99. Fang, H. et al. Mutational processes of distinct POLE exonuclease domain mutants drive an enrichment of a specific TP53 mutation in colorectal cancer. *PLoS Genet.* **16**, e1008572 (2020).
100. Robinson, P. S. et al. Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.23.167668 (2020).
101. Kim, S.-H., Elango, N., Warden, C., Vigoda, E. & Yi, S. V. Heterogeneous genomic molecular clocks in primates. *PLoS Genet.* **2**, e163 (2006).
102. Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl Acad. Sci. USA* **113**, 10607–10612 (2016).
103. Petronzelli, F. et al. Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA *N*-glycosylase. *J. Biol. Chem.* **275**, 32422–32429 (2000).
104. Schermerhorn, K. M. & Delaney, S. A chemical and kinetic perspective on base excision repair of DNA. *Acc. Chem. Res.* **47**, 1238–1246 (2014).
105. Sassa, A., Çağlayan, M., Dyrkheeva, N. S., Beard, W. A. & Wilson, S. H. Base excision repair of tandem modifications in a methylated CpG dinucleotide. *J. Biol. Chem.* **289**, 13996–14008 (2014).
106. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
This paper classifies mutational processes in cancers into processes that scale with time and processes that do not scale with time.
107. Williams, N. et al. Phylogenetic reconstruction of myeloproliferative neoplasm reveals very early origins and lifelong evolution. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.09.374710 (2020).
108. Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl Acad. Sci. USA* **116**, 9491–9500 (2019).
109. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
110. Goldmann, J. M. et al. Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).
111. Yu, B. et al. Genome-wide, single-cell DNA methylomics reveals increased non-CpG methylation during human oocyte maturation. *Stem Cell Rep.* **9**, 397–407 (2017).
112. Jónsson, H. et al. Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018).
113. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014).
114. Polak, P. & Arndt, P. F. Transcription induces strand-specific mutations at the 5′ end of human genes. *Genome Res.* **18**, 1216–1223 (2008).
115. Xia, B. et al. Widespread transcriptional scanning in the testis modulates gene evolution rates. *Cell* **180**, 248–262.e21 (2020).
116. Jinks-Robertson, S. & Bhagwat, A. S. Transcription-associated mutagenesis. *Annu. Rev. Genet.* **48**, 341–359 (2014).
117. Chen, C.-L. et al. Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* **28**, 2327–2337 (2011).
118. Shinbrot, E. et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* **24**, 1740–1750 (2014).
119. Yurchenko, A. A. et al. XPC deficiency increases risk of hematologic malignancies through mutator phenotype and characteristic mutational signature. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.13.200667 (2020).
120. Wong, W. S. W. et al. New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
121. Goldmann, J. M. et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
Together with Jónsson et al. (2017), this paper describes a localized increase of clustered mutations in human oocytes.
122. Oktay, K., Turan, V., Titus, S., Stobezki, R. & Liu, L. BRCA mutations, DNA repair deficiency, and ovarian aging. *Biol. Reprod.* **93**, 67 (2015).
123. Titus, S. et al. Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Sci. Transl. Med.* **5**, 172ra21 (2013).
124. Ma, W., Westmoreland, J. W., Gordenin, D. A. & Resnick, M. A. Alkylation base damage is converted into repairable double-strand breaks and complex intermediates in G2 cells lacking AP endonuclease. *PLoS Genet.* **7**, e1002059 (2011).
125. Goldmann, J. M., Veltman, J. A. & Gilissen, C. De novo mutations reflect development and aging of the human germline. *Trends Genet.* **35**, 828–839 (2019).
126. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
127. Goldmann, J. M. et al. Stochasticity explains differences in the number of de novo mutations between families. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.18.303727 (2020).
128. Kaplanis, J. et al. Identifying and characterising germline hypermutators [abstract]. *Eur. J. Hum. Genet.* **28** (Suppl. 1), 712 https://doi.org/10.1038/s41431-020-00739-z (2020).
129. Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T. & Hurles, M. E. Similarities and differences in patterns of germline mutation between mice and humans. *Nat. Commun.* **10**, 4053 (2019).
130. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
131. Rodin, R. E. et al. The landscape of mutational mosaicism in autistic and normal human cerebral cortex. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.11.944413 (2020).
132. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
133. Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* **332**, 459–461 (1988).
134. Smith, T. B. et al. The presence of a truncated base excision repair pathway in human spermatozoa that is mediated by OGG1. *J. Cell. Sci.* **126**, 1488–1497 (2013).
135. An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
136. Kunkel, T. A. & Bebenek, K. DNA replication fidelity. *Annu. Rev. Biochem.* **69**, 497–529 (2000).
137. Tubbs, A. & Nussenzweig, A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **168**, 644–656 (2017).
138. Hedglin, M. & Benkovic, S. J. Eukaryotic translesion DNA synthesis on the leading and lagging strands: unique detours around the same obstacle. *Chem. Rev.* **117**, 7857–7877 (2017).
139. Pagès, V. & Fuchs, R. P. How DNA lesions are turned into mutations within cells? *Oncogene* **21**, 8957–8966 (2002).
140. Varga, Á., Marcus, A. P., Himoto, M., Iwai, S. & Szüts, D. Analysis of CPD ultraviolet lesion bypass in chicken DT40 cells: polymerase η and PCNA ubiquitylation play identical roles. *PLoS ONE* **7**, e52472 (2012).
141. Chan, K., Resnick, M. A. & Gordenin, D. A. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair* **12**, 878–889 (2013).
142. Simonelli, V., Narciso, L., Dogliotti, E. & Fortini, P. Base excision repair intermediates are mutagenic in mammalian cells. *Nucleic Acids Res.* **33**, 4404–4411 (2005).
143. Ryba, T. et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
144. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
145. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
146. Ng, S. B. et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
147. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
148. Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P. & Eichler, E. E. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Med.* **9**, 101 (2017).
149. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).
150. Dietlein, F. et al. Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
151. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041. e21 (2017).
152. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
153. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
154. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
This study shows that the cancer cell of origin can be inferred from mutational patterns.
155. Touat, M. et al. Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature* **580**, 517–523 (2020).
156. Bryant, H. E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913–917 (2005).
157. Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).

Acknowledgements
The authors thank C. Gilissen, F. Supek and the other, anonymous, reviewers for valuable discussions. This work was supported by US National Institutes of Health grants R35GM127131, R01MH101244 and U01HG009088.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Genetics thanks C. Gilissen, F. Supek and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.