Research Article

Mutational selection: fragile sites, replicative stress, and genome evolution

David Haig[1]

1. Department of Organismic and Evolutionary Biology, Harvard University, United States

New mutations compete in the germline with their unmutated progenitors. Successful genes accumulate functions that are easily broken by small mutational changes, thus placing their own mutants at a selective disadvantage. Their germline phenotypes evolve to be fragile rather than robust to the effects of mutation. Genetic recombination, by contrast, favors robust interactions among parts that must maintain function in many different combinations. Fragile epistatic dependencies are predicted among closely-linked sites with more robust interactions predicted among sites that recombine freely on the time-scale of mutation. Genes can compete for dominance with alternative alleles subjecting the 'dominant' allele to mutational selection but protecting the 'recessive' allele from such selection. Genes that function in DNA replication and repair are predicted to evolve features that challenge their own abilities because these challenges place their own loss-of-competence mutations at a selective disadvantage. This process may help to explain the evolution of fragile sites that test the competence of the machinery of replication and repair.

Every mutation originates in a single cell, each body develops from a single cell, but these are rarely the same cell. Before a new mutation reaches a zygote, it must occur in a cell of the germline and then make its way from this first cell into a gamete. Two evolutionary phases can be distinguished: differential survival and proliferation of germline cells before mutations are inherited by zygotes followed by differential survival and fertility of multicellular bodies. I will refer to the first phase as 'mutational selection' and the second as 'individual selection.' After the generation in which a variant originates by mutation, its long-term fate will be determined by individual selection 'choosing' among bodies with and without its copies and recurrent bouts of mutational selection 'choosing' between cells with and without the variant because it has mutated to something else. From a genetic

 doi.org/10.32388/9ERYW5

perspective, mutational selection involves competition between a mutation and its unmutated progenitor whereas individual selection involves competition among all alleles at a locus.

The definition of the germline used in this paper will be Weismann's (1892, p. 242f) definition of the *Keimbahn* as a cellular lineage that connects an egg cell to a reproductive cell. This definition *includes* the zygote and early embryonic cells. It contrasts with a common definition of the germline that *excludes* totipotent and pluripotent cells from the germline because these have somatic as well as germ-cell descendants (see Haig 2022). The inclusive, or Weismannian, germline is the appropriate definition for this paper because mutations that occur in any one of its cells can potentially be transmitted to offspring whereas mutations in somatic cells cannot. All multicellular organisms, including plants, possess continuous Weismannian germlines.

The phrase 'germline selection' has been used to encompass all selective processes acting on intercellular genetic variation within individual germlines (Hastings 1989, 1991). I will distinguish germline segregation from germline mutation as different causes of intercellular variation. Germline segregation involves mitotic crossing-over and gene conversion that generate mosaics of homozygous and heterozygous cells in germlines derived from heterozygous zygotes. Differential proliferation of these cell-lineages can result in segregation distortion among the gametes produced by an initially heterozygous germline. This is analogous to meiotic drive (Hastings 1989, 1991). I will restrict 'mutational selection' to selection between cells that differ because of a *de novo* mutation not present in the zygote. The focus of this paper will be on mutational selection and will not consider germline segregation in detail.

At each locus, mutational selection occurs only in generations in which a mutation occurs because all cells of subsequent germlines either possess or do not possess the mutation until a mutation to the mutation creates new germline mosaicism. This is a pure form of selection because it involves a choice between cells that differ for the presence versus absence of a mutation on a shared genetic background. It is also a cheap form of selection because mutations can be eliminated by the death of a single cell or small clone of cells (Otto and Orive 1995). Individual selection, by contrast, involves differential survival and reproduction of organisms that vary at many loci simultaneously with mutations eliminated by deaths of multicellular bodies. Germline segregation resembles mutational selection in that changes in gene frequency occur within germlines but resembles individual selection in that selection occurs among alleles segregating in the zygotic gene pool.

In an idealized model of mutational selection, a mutation occurs in one daughter cell of a stem cell. In the absence of selection, half of the descendants of the stem cell would carry a copy of the mutation and half would carry a copy of the unmutated allele. However, if the mutated daughter is eliminated by mutational selection, and its place taken by an extra division of its unmutated sister, then all the surviving descendants of the stem cell will carry the unmutated allele. The form of selection assumed here, and throughout this paper, is one in which the number of cells in the germline is subject to regulatory control, with some degree of reproductive compensation. Differential fertility of germlines will be considered an aspect of individual selection.

Genes favored by mutational selection exclude their own mutants from subsequent germlines. Such mutants must have dominant-negative effects. The function of both alleles in a diploid germline could be maintained by mutational selection if their combined expression were sufficiently low that loss-of-function of one allele resulted in haploinsufficiency, or if gene expression was monoallelic with periodic switching between alleles. Such processes would render mutations subject to selection on their haploid effects (Maley and Tapscott 2003).

Co-dominance of the effects of loss-of-function mutations can be considered a cooperative outcome in which both alleles at a focal locus are subject to mutational selection. This outcome benefits genes at all other loci because these genes are necessarily associated with a functional allele at the focal locus in the next generation of gametes. Therefore, *trans* modifiers favor increased dominance in the germline. However, each allele at the focal locus necessarily segregates away from mutations to its other allele at meiosis and therefore bears no cost from an absence of purifying selection on the other allele. Therefore, selection on *cis* modifiers of germline dominance can favor a non-cooperative outcome in which alleles compete for dominance. In this scenario, successful alleles are ‘hard’ on their own mutants, subjecting them to mutational selection, but ‘soft’ on mutations to the other allele, exempting them from mutational selection.

Germline segregation and mutational selection have different consequences for recessive germline effects. Because germline segregation generates homozygous cell lineages within an initially heterozygous germline, genes are thereby exposed to selection on their recessive effects. Thus, germline segregation favors genetic variants with recessive effects that enhance cellular fitness (Hastings 1991; Otto and Hastings 1998). By contrast, *de novo* germline mutations are only subject to mutational selection on their recessive effects when they occur in germlines already heterozygous for a recessive allele.

Recessivity of somatic loss-of-function mutations minimizes organismal costs of these mutations (Orr 1995). In a Panglossian world, the combined effects of mutational and individual selection would favor genes with dominant germline effects (subject to haploid selection) but whose mutants have recessive somatic effects (subject to diploid selection).

Mutational selection and the frequency of de novo *mutations*

Most individuals with achondroplasic dwarfism possess a G to A transition at position 1138 of *FGFR3*, replacing glycine^{380} with arginine^{380} in FGFR3 protein. *De novo* mutations at this single nucleotide are observed in 1 in 20,000 births with almost all mutations inherited from fathers. This was originally interpreted as an extraordinarily high mutation rate at a single nucleotide, but arginine^{380} was subsequently shown to confer a proliferative advantage on male germ cells (Shinde et al. 2013; Arnheim and Calabrese 2016).

McCune-Albright syndrome occurs in perhaps one in a million births and, like achondroplasia, is caused by mutations in a single codon (Dumitrescu and Collins 2008). Most affected individuals possess C601T or G602A mutations in the gene that encodes the G protein α stimulatory subunit (Gαs) (Weinstein 2006). These mutations replace arginine^{201} with cysteine^{201} or histidine^{201} and are known only from mosaic individuals. This suggests that the mutations are lethal in early embryos but tolerated, albeit with substantial pathologies, in somatic mosaics (Happle 1986).

Mutational selection changes the frequency with which mutations are observed in zygotes and thereby changes the 'genetic load', the number of selective deaths of individuals required to eliminate a deleterious mutation. The mutations causing achondroplasic dwarfism are favored by mutational selection whereas those causing McCune-Albright syndrome are eliminated by mutational selection. Therefore, the genetic load is increased for achondroplasic dwarfism but decreased for McCune–Albright syndrome. A further consequence is that mutational selection alters the frequency of mutations at mutation–selection equilibrium.

Although mutational selection occurs only in generations in which a mutation occurs, a mutation can be subject to many generations of individual selection. Therefore, the effects of individual fitness dominate when germline and organismic fitness come into conflict. *FGFR3* G1138A, for example, is subject to strong positive selection in male germlines in the first generation it occurs but 'selfish' germline proliferation is limited to this single generation because individuals who inherit G1138A

possess the mutation in all their cells and no cells are thereby advantaged. The evolutionary fate of G1138A mutations is therefore determined by their effects on organismal fitness.

The effects of mutational selection will be most significant when a variant affects germline fitness but is almost neutral in its effects on organismal fitness. In such a scenario, selective events coincide with germline mutations. Mutational selection and germline mutation are evenly matched because the frequency of selective events is of the same order of magnitude as the mutation rate. Because a gene competes with its own mutants for limited places in the germline, mutational selection favors genes that place their own mutants at a selective disadvantage. Such genes evolve germline phenotypes that are fragile rather than robust to effects of mutation. They accumulate more and more features that cannot be changed without loss of germline viability. The genetic benefit of possessing a mutationally fragile phenotype is the extra gametes bearing a gene's copies that would otherwise have been occupied by the gene's mutations. At the limit, molecular evolution would come to a halt if every possible change to a gene resulted in cellular death in the germline.

The evolution of fragile germline phenotypes

The hypothesis that mutational selection favors easily-broken phenotypes has features in common with theories of 'constructive neutral evolution' (Stolztfus 1999) and 'irremediable complexity' (Gray et al. 2010). In these theories, a neutral mutation from *ab* to *Ab* is followed by another neutral mutation from *Ab* to *AB*, with the second mutation creating a dependency between *A* and *B* such that back-mutations from *A* to *a* are disfavored. Iteration of this process causes increasingly complex dependencies, *ABCDE*, to accumulate by a ratchet-like process.

The description of this process as neutral is somewhat misleading. Neutrality is a property of a difference between alternatives in a particular context. These models presume that the difference between *a* and *A* is neutral in the context of *b* but non-neutral in the context of *B*. The series of forward mutations from *abcde* to *ABCDE* are all neutral (by hypothesis) but, in the process, genetic differences that were once neutral become subject to selection. The system cannot drift backward. Increasing interdependency is predicted whether the initial mutations are positively selected or neutral. Dependency is another name for coadaptation of parts. Complexity increases as new dependencies prevent piecemeal reversions to a simpler past. Genes evolve sequences for which a higher proportion of mutations are rejected. Models of constructive neutral evolution typically assume that each

mutation in the series drifts to fixation before introduction of the next mutation. These models have not considered the consequences of sexual recombination.

Mutational selection and recombination are opposing forces with respect to mutational frailty. Suppose that *ab*, *Ab*, *AB* haplotypes are viable but *aB* haplotypes are inviable and eliminated by mutational selection (*B* depends on *A*, *a* depends on *b*, but neither *A* nor *b* depends on the variant at the other site). The *ab* and *AB* haplotypes can be considered 'frail' because either can mutate to an *aB* haplotype that is eliminated by mutational selection, whereas the *Ab* haplotype can be considered 'robust' because neither its *AB* nor *ab* mutant progeny are eliminated. Mutational selection thus favors 'frail' *ab* and *AB* haplotypes over 'robust' *Ab* haplotypes. Meiotic recombination in *ab*/*AB* heterozygotes destroys *ab* and *AB* haplotypes to create *Ab* and *aB* haplotypes, with the former surviving and the latter eliminated by individual selection. Thus, recombination favors the 'robust' *Ab* haplotype and disfavors 'frail' *ab* and *AB* haplotypes. Whether robustness or frailty is favored for *cis* interactions among linked sites will depend on whether recombination or mutation is a more important source of inviable interactions. An association of mutational robustness with increased recombination has been noted before (Gardner and Kalinka 2007; Desai et al. 2007; Klug et al. 2019).

Robustness is favored for interactions between recombining sites by individual selection. Two sites 'freely recombine' with respect to mutational selection if their variants are randomized with respect to each other in the generations between successive mutations. Thus, patterns of robustness and frailty of protein-coding sequences will be influenced by gene structure. Amino acids encoded in exons separated by large introns freely recombine on the time-scale of new mutations and are predicted to robustly interact whereas amino acids encoded within the same exon are predicted to form fragile dependencies. Greater mutational frailty is expected when a protein domain is encoded by a single large exon rather than several small exons. This may help to explain why large exons often encode intrinsically-disordered domains (Kawachi et al. 2021; Fukuchi et al. 2023). Another factor may be that individual selection is less effective at purging deleterious mutations within regions that rarely recombine (Felsenstein 1974).

Mutational selection favors mutational frailty at closely-linked sites but individual selection may favor more robust interactions. Firstly, proteins that are sensitive to mutational perturbation are probably also more prone to denature in response to environmental perturbations. Therefore, if denaturation involves a cost to individual fitness, individual selection favors genetic variants whose protein products are more robust to environmental perturbation. Secondly, mutations will inevitably

occur in somatic cells. Therefore, genetic variants in the germline may be favored whose protein products are more robust to the effects of somatic mutations if these mutations are costly to organismal fitness. A full treatment of this question would require understanding the interplay between selection and mutation in somatic cells. There may be little cost, for example, if a somatic mutation immediately causes cellular death. On the other hand, somatic mutations favored by cellular selection in the soma may have substantial organismal costs (as, for example, occurs in cancer).

The 'preservation of the frail' is a version of what Archetti (2009) called 'survival of the steepest'. In his model, juveniles competed with siblings for parental care. Alleles were favored whose phenotypes were sensitive to mutation because if a new mutation were transmitted to siblings, who thereby expropriated less care, extra resources would accrue to siblings without the mutation. Hamilton (1966) expressed a related idea when he argued that reproductive compensation during the period of parental care favors increased vulnerability of offspring at younger ages. Such ideas are easily generalized to competition between mutant and non-mutant cell lineages in germlines. The key factor is 'reproductive compensation' because the death of a mutant germ cell (or offspring) creates a new opening for a non-mutant germ cell (or offspring).

Synonymous constraints

Sequences subject to mutational selection are predicted to evolve complex dependencies that ensure more of their mutations are rejected. This process favors rampant pleiotropy and epistasis as genes accumulate more and more reasons why changes to their sequence will not work. Such genes are predicted to gather functional constraints that render themselves increasingly indispensable in the germline. They evolve to be 'easily broken' by mutation. One signature of mutational selection may be synonymous constraints within coding sequences. Such constraints suggest that a sequence has a nucleic acid phenotype subject to either individual or mutational selection in addition to the phenotype of its encoded protein (Shabalina et al. 2013; Savisaar and Hurst 2018). These constraints may include interactions of the nucleic acid sequence, either DNA or RNA with itself, and with various other RNAs and proteins. I looked for evidence of synonymous constraint in the immediate neighborhood of mutations known to be subject to strong mutational selection causing achondroplasic dwarfism and McCune–Albright syndrome.

I looked for evidence of synonymous constraint in the immediate neighborhood of two mutations known to be subject to strong mutational selection. Most individuals with achondroplasic dwarfism

possess a G to A transition at position 1138 in one of their *FGFR3* alleles. This mutation replaces glycine^{380} with arginine^{380} in the FGFR3 protein. *De novo* mutations at this single nucleotide are observed in 1 in 20,000 births with almost all mutations inherited from fathers. The mutation confers a proliferative advantage on male germ cells, thus amplifying the frequency of the mutation in spermatozoa (Shinde et al. 2013). 'Selfish' germline proliferation however is limited to a single generation because individuals who inherit such mutations possess the mutation in all cells and no cells are thereby advantaged.

Figure 1a

```
LeuSerTyrGlyValGlyPhe
cucagcuacgggugggcuuc Homo
.........a.......... achondroplasia
........cg.......... Pan
........cg.......... Gorilla
........u........... Callithrix
........cg......u.... Mus
........cg.......... Rattus
........cg..c........ Bos
........cg.......... Canis
........cg...ca...... Loxodonta
........cg.......... Dasypus
```

Figure 1b

```
AspLeuLeuArgCysArgValLeuThrSerGlyIlePheGluThrLysPheGlnValAspLysValAsnPhe
gaccugcuucgcugccguguccugacuucuggaaucuuugagaccaaguuccagguggacaaagucaacuuc Homo
.........cg.....a.................................................... McCune–Albright
.........cg....cg..................................................... Pan
.........cg....cgcg................................................... Gorilla
.........cg....cgcg................................................... Callithrix
.........cg....cg........c............................................. Mus
.........cg....cgcg......c............................................. Rattus
..u.....ccg....cg..................................................... Bos
..u.....cg.....cg..................................................... Canis
..u.....ccg....cg.............................a.............g......... Loxodonta
.........cg....cgcg................................................... Dasypus
..u..u..ca.a...a.a..au....a..a..g..u.....a.....a..u..a.....u.....a..u..u Monodelphis
.....u..ca.a...a.a..uu....a..a..c..u.....a........u..a...........a..u..u Ornithorhynchus
..u..u..ca.a...a.a..uu....a..a.....u.....a........u..a...........a..u..u Gallus
...u....gcga...cga..u..cg.a........a.....a........u..............a..u... Xenopus
```

Figure 1. Sequence alignments providing evidence of synonymous constraint in (a) *FGFR3* and (b) *GNAS*. Dots represent identical nucleotides relative to the human sequence, except that all CpG dinucleotides are shown in cyan. The species are human (*Homo*), chimpanzee (*Pan*), gorilla (*Gorilla*), marmoset (*Callithrix*), mouse (*Mus*), rat (*Rattus*), ox (*Bos*), dog (*Canis*), elephant (*Loxodonta*), armadillo (*Dasypus*), opossum (*Monodelphis*), platypus (*Ornithorhynchus*), chicken (*Gallus*), and frog (*Xenopus*). Mutations to the human sequences that cause achondroplasia (*FGFR3*) and McCune-Albright syndrome (*GNAS*) are also shown.

Figure 1a presents an alignment of RNA sequences encoding seven amino acids from the *FGFR3* genes of ten eutherian mammals along with the amino acid sequence of the human protein. Mutations that cause achondroplasia change the guanine in the first position of the codon that specifies glycine^{380}. The encoded amino acids are identical in seven of the species and there are three changes of a single

amino acid (one each in the mouse, cow, and elephant). Because of redundancies in the genetic code, there are 4608 different ways of specifying the seven amino acids of the human protein. In the seven species with identical amino acid sequence, there is one synonymous substitution in the marmoset. Clearly, some form of selection has rejected synonymous changes to the nucleic acid sequence.

Figure 1b presents an alignment of a sequence from exon 8 of *GNAS*. All sequences, from human to frog encode identical amino acids (apart from the single-nucleotide substitution in McCune-Albright syndrome). Given redundancies in the genetic code there are more than 3×10^{12} ways of encoding these 24 amino acids. Clearly, the nucleotide sequence exhibits strong conservation at synonymous sites. Few synonymous changes, and no non-synonymous changes, have been tolerated over the course of tetrapod evolution.

Non-coding constraints on the nucleic acid sequence of *GNAS* appear to have been rearranged in an early eutherian ancestor because the nine eutherian sequences form a highly similar cluster distinct from a cluster of non-eutherian sequences despite strict conservation of amino acid sequence between the clusters (Fig. 1b). The alignment shows CpG dinucleotides that are potential sites of cytosine methylation. The first ten sequences are from eutherians, followed by a marsupial (*Monodelphis*), monotreme (*Ornithorhynchus*), bird (*Gallus*), and amphibian (*Xenopus*). The eutherian and *Xenopus* sequences contain 2–3 CpG dinucleotides, two of which correspond to the first and second bases encoding arginine^{199} and arginine^{201}. The most variable position in eutherians is the third base of arginine^{201}: some eutherians (such as humans) use CGU for arginine^{201} whereas others (such as gorillas) use CGC which creates a third CpG dinucleotide in combination with the first base of valine^{202}. I conjecture this variability reflects mutation–selection balance: selection favors CGC; spontaneous deamination of methylcytosine to thymine favors CGU (Holliday and Grigg 1993). *Xenopus* uses CGA for both arginines. *Monodelphis*, *Ornithorhynchus*, and *Gallus* use AGA and lack CpG dinucleotides.

The mutations causing achondroplasia and McCune-Albright syndrome are both known to be subject to mutational selection and both are found here to reside in sequences subject to synonymous constraint. More such mutations need to be studied to determine whether this is a pattern. One way to become a highly-constrained sequence is to accumulate ‘essential’ interactions with many partners. Protein segments that interact with many partners are frequently intrinsically disordered (Haynes et al. 2006) and gene regions of synonymous constraint tend to encode intrinsically disordered protein

segments (Macossay-Castillo et al. 2014). Intrinsic disorder of encoded proteins may be an additional signature of genes subject to mutational selection.

Competition for dominance by elite alleles

A previous section raised the possibility of a non-cooperative outcome of mutational selection in which alleles compete for dominance of their mutant effects. This section explores that scenario in more detail.

Let allele *i* be represented by $\{x_i, y_i\}$ where x_i is its 'competence' in germline replication and y_i is the 'difficulty' or 'competence required' for its own replication. Competence is determined by gene products acting in *trans* whereas difficulty is determined by gene properties acting in *cis*. Genes sit for 'examinations' in diploid pairs, $\{x_1, y_1\}$ with $\{x_2, y_2\}$, where $x_1 \geq x_2$ by arbitrary assignment of subscripts. Each gene poses a problem, y_1 and y_2, and the pair passes the examination if both problems are solved. In this examination, x_1 is the 'effective competence' of the pair of alleles. Whichever of y_1 or y_2 is greater determines the 'effective difficulty' of the exam. Higher competence is dominant to lower competence. Higher difficulty is dominant to lower difficulty. The examination is passed if the allele of effective competence is able to solve both problems, $x_1 \geq y_1, y_2$. Competence is an antidote to difficulty.

An allele *i* is *competent* if $x_i - y_i \geq 0$, but *incompetent* if $x_i - y_i < 0$. If $\{x_1, y_1\}$ and $\{x_2, y_2\}$ are both competent, the examination is passed. If the allele of effective competence $\{x_1, y_1\}$ is incompetent, the examination is failed. If $\{x_1, y_1\}$ is competent but $\{x_2, y_2\}$ is incompetent, the examination is passed if $x_1 \geq y_2$. The difference between an allele's competence and difficulty, $x_i - y_i = \Delta_i$, is that 'surplus competence' is not needed to solve its problem. A competent allele has non-negative surplus competence, $\Delta_i \geq 0$, whereas an incompetent allele has negative surplus competence, $\Delta_i < 0$. An 'elite allele' is highly competent but with little surplus competence because it sets challenging problems for its own loss-of-function mutants.

Individual selection will be considered first. Genes sit for 'organismal examination' in pairs inherited from gametes, $\{x_1, y_1\}$ with $\{x_2, y_2\}$. The examination is passed if the allele of effective competence is able to solve both problems, $x_1 \geq y_1, y_2$. Otherwise, the individual fails the exam, and its germline is sterile. Thus, individual selection eliminates organisms with 'homozygous' genotypes containing two incompetent alleles and 'heterozygous' genotypes containing a competent allele that is unable to solve the problem posed by the incompetent allele (Figure 2).

Figure 2

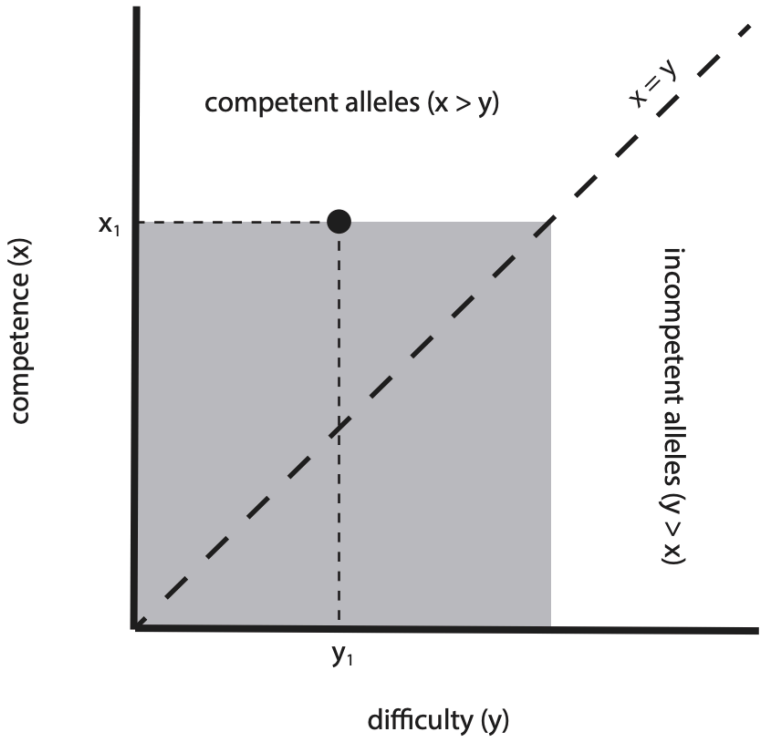

Figure 2. Alleles are defined by their competence (x) at solving a problem and the difficulty (y) of the problem they pose. Competent alleles (points above the diagonal) can solve their own problem. Incompetent alleles (points below the diagonal) cannot solve their own problem. The black dot represents the 'allele of effective competence' in a diploid cell, characterized by competence x_1 and difficulty y_1. It can solve the problems posed by all alternative alleles in the shaded area.

Mutational selection occurs in bodies with fertile germlines that have passed the organismal examination. Non-mutant cells in these bodies contain $\{x_1, y_1\}$ and $\{x_2, y_2\}$ alleles where, $x_1 \geq y_1, y_2$. Mutational selection eliminates mutant cells that fail 'cellular examinations'. Mutations will be assumed to occur one at a time in a subset of germline cells and change either competence or difficulty, but not both. For mutations that affect difficulty, y_m, mutational selection tolerates all $\{x_1,$

$y_m\}$ and $\{x_2, y_m\}$ where $y_m \leq x_1$ but rejects all $y_m > x_1$. Mutations that increase difficulty are tolerated up to the degree of difficulty x_1 determined by the nonmutant allele of higher competence, including new incompetent alleles $\{x_2, y_m\}$ where $x_1 \geq y_m > x_2$.

Mutations that change competence, x_m, will now be considered. Mutational selection tolerates all changes to the competence of the less-competent allele $\{x_m, y_2\}$ because $x_1 \geq y_2$, but rejects changes to the competence of the more-competent allele $\{x_m, y_1\}$ when $x_m < y_1, y_2 \leq x_1$. The latter requirement has two parts: (a) the more-competent allele must pose a problem that cannot be solved by the less-competent allele, $y_1 > x_2$; (b) the loss-of-competence mutation to the more-competent allele must create a new incompetent allele, $y_1 > x_m$. The smaller the surplus competence of the more competent allele, $\Delta_1 = x_1 - y_1$, the greater the range of reductions of competence that are subject to mutational selection. At the limit, $\Delta_1 = 0$, all $x_m < x_1$ are rejected. Thus, mutational selection is most effective at maintaining the competence of alleles of high competence that pose difficult problems for which they have minimal surplus competence.

Mutational selection will be absent in germlines homozygous for two elite alleles of equal competence, because each can solve the other's problems. This can result in incompetent alleles of high difficulty evading mutational selection. Such alleles will be a source of individual selection against competent alleles that pose problems of lower difficulty. Successful 'elite alleles' pose increasingly difficult problems for themselves while solving the problems posed by 'non-elite' alleles. In the process, elite alleles exempt less-competent alleles from the purifying effects of mutational selection. Other genes in the genome pay the cost of subsequent associations with alleles of lesser competence. Therefore, other genes would benefit from shifting the system from one in which only elite alleles are subject to mutational selection to one in which mutations to all alleles are haploinsufficient and subject to mutational selection. The collective opposes selfish elites.

In summary, mutational selection favors elite alleles of high competence, but minimal surplus competence, that pose challenging problems to themselves. It fails to eliminate loss-of-competence mutations to less-competent alleles and to elite alleles when these occur as homozygotes. Alleles that pose problems of higher difficulty are subject to stronger purifying selection for competence than alleles that pose problems of lower difficulty. Therefore, a positive association is predicted between the competence and difficulty of alleles because high difficulty can persist through many generations only if it is coupled in *cis* to high competence. Difficult problems hitchhike to high frequency with

alleles of high competence but high competence is maintained, in part, by its association with difficult problems.

In this model, mutational selection not only maintains the competence of elite alleles, but also tolerates incompetent alleles that can survive, for a while, in the gene pool in association with alleles of higher competence. Incompetent alleles are eliminated by individual selection in genotypes in which the allele of higher competence is unable to solve the problem of greater difficulty. Competent alleles of low competence and low difficulty are vulnerable to mutation to incompetent alleles either by further decreases of their competence or increases of their difficulty. Incompetent alleles are not subject to mutational selection for competence but their presence strengthens mutational selection on the competence of elite alleles. Incompetent alleles that pose problems of lower difficulty persist longer in the gene pool because their problems are solved by a higher proportion of competent alleles.

The model has features similar to the 'enhancer runaway' model of Fyon et al. (2015) in which stronger enhancers evolve because purifying selection purges deleterious mutations more efficiently for highly expressed alleles than for weakly expressed alleles. Through this process, stronger enhancers become associated with better-quality alleles. 'Coding sequences' and 'enhancers' in their model correspond to 'competence' and 'difficulty' in the model presented here. Their model predicted an escalation of ever-stronger enhancers as alleles competed with each other for dominance. The possibility of similar runaway selection for increased 'difficulty' may help to explain features of conserved fragile sites (discussed in a future section).

Mutational stress tests

Germline mutation rates in multicellular organisms are generally believed to be determined by the inability of natural selection to further reduce mutations rather than by selection against mutation rates that are too low. In this view, the negative effects of exposure to deleterious mutations become a weaker and weaker selective force as mutations become rarer and rarer. The lower bound occurs when mutations that further reduce the mutation rate are almost neutral in their effects on fitness relative to alleles causing slightly higher mutation rates (Lynch 2011).

Individual selection against 'mutator' alleles is surprisingly weak because the 'mutator' rapidly segregates away from most of the mutations it causes, thus sharing the costs of mutation with alternative alleles (Kimura 1967). This can be clearly seen in the case of recessive mutations. By the time that a recessive allele encounters another recessive allele, both alleles have been randomized by

recombination with respect to alleles at the mutator locus. A similar argument applies to all causes of weak selection in which a mutation only has effects every few generations. For this reason, it is principally dominant mutations of large effect and high penetrance that exert a downward selective pressure on the mutation rate via individual selection.

Mutational selection favors increased costs of mutations in the germline. For genes that influence the mutation rate, mutational selection favors variants whose loss-of-function leads to cellular death. During the generation in which it originates, a new mutator will be subject to all of the costs of the mutations it causes (in competition with other germline cells without the mutator). The previous section argued that elite alleles may maintain their own competence by posing difficult problems for themselves to solve. If high competence, translates into lower mutation rates affecting the rest of the genome, then mutational selection may be able to maintain higher fidelity of DNA replication than individual selection acting alone.

Faithful transcription, replication, and repair of the genome depend on the robust cooperation of many genes. I will call this set of genes 'the guild'. All guild members benefit from prompt and reliable elimination of deleterious mutations to other guild members, but this selection is relatively weak because a mutation that increases the mutation rate at other loci segregates away from most of the damage it causes within a few generations, thus sharing the cost of mutation with other alleles at its own locus. Each guild member has a particular interest in eliminating its own mutations and is predicted to pose particularly difficult problems for its loss-of-function mutations to solve. Therefore, the genomic sequence of each member of the guild is predicted to be associated with idiosyncratic features that test its own particular competence.

The members of the guild evolve to pose problems for the machinery of transcription, replication, and repair that constitute a 'stress test' of the efficiency and fidelity of that machinery. These problems involve complex interdependencies within genomic sequences that require precision work by multiple members of the guild. The rigor of the examination is maintained primarily by the benefits that accrue to each guild member from the elimination of its own mutants, but each member also gains short-term benefits from its association with functional rather than non-functional alleles at other loci.

Each member of the guild is predicted to accumulate difficult-to-replicate features that test its own competence. Because surplus competence is not subject to mutational selection, such genes are predicted to evolve to a point at which they are just able to solve the problem they pose to themselves. These self-examinations simultaneously test the competence of other guild members that participate

in replication and repair. The collective competence of the guild creates a nuclear environment in which difficult-to-replicate sequences are tolerated throughout the genome. The guild evolves to pose problems that must be passed for entry to the guild. A question for future work is whether these entry examinations simply maintain a 'closed shop' or whether they maintain a higher overall standard of work.

The unsettling effects of repetitive elements

The progress of mutational selection would be slow and sedate if every gene were present as a single copy per haploid genome and mutational change was restricted to point mutations in coding sequences. However, genomes also contain sequences present in multiple copies that replicate according to different rules, including transposable elements and satellite repeats. If the production of extra copies has been a recurrent event, not merely a random accident, then these sequences possess attributes that predispose them to preferential copying and that can be considered adaptations for intranuclear proliferation. Sequences with such properties proliferate because a variant that enters a zygote as a single copy can be transmitted to more than half of the gametes produced by the organism that develops from that zygote. This is especially true for copies that are dispersed to non-homologous loci because new copies segregate independently of source copies. Variants that succeed in intranuclear selection are transmitted to offspring. Therefore, intranuclear adaptations can be cumulative across many generations.

Tandem arrays of repeats are sites of replicative stress and chromosomal breakage which enables the repeats to recruit machinery of DNA repair to achieve amplification within the genome at double-strand breaks. Non-homologous end-joining (NHEJ) joins together broken ends and provides little opportunity for amplification of repeats. Homologous recombination (HR) requires replicative repair using a homologous template that allows repeats to proliferate when a replication fork copies a sequence already replicated in the same S-phase (roughly speaking a replication fork chases another replication fork or chases itself in a loop) (Haig 2021, 2022).

Intranuclear selection of multicopy sequences is the source of a mutational bias toward expansions of repeats and increased replicative stress because of their activities. Selfish repeats would increase without limit without countervailing selection from costs to cellular or organismal fitness. Changes in copy-number can be considered mutations that are subject to cellular selection within the germline before a variant's first transmission to a zygote followed by individual selection in subsequent

multicellular bodies. During the first phase of mutational selection, insertions and deletions will be tolerated at genomic locations where they do not disrupt cellular fitness but will be eliminated where they are incompatible with cellular survival. Those copy-number variants that pass this selective sieve will then be subject to individual selection on their somatic effects. One straightforward consequence is that repeats will accumulate in genomic regions where selection for cellular or organismal functions is relaxed. Another consequence is that they will accumulate in the introns of genes involved in the management of replicative stress. Expansions will be tolerated up to the replicative competence of the guild and help to maintain that competence.

The proliferative activities of repeats may contribute to the expansion of introns. This would favor robust, rather than fragile, interactions within proteins encoded by multiple exons that recombine frequently on the time-scale of mutational selection. If repeat expansions increase the difficulty of replication, then this creates a selective factor favoring increased competence of guild members without the need for direct selection to maintain difficulty. To the extent that repeat expansions within introns contribute to the self-examination of elite alleles, a division of labor is possible between exonic sequences determining 'competence' in *trans* and intronic sequences contributing to 'difficulty' in *cis*.

Fragile sites

Fragile sites are scattered throughout the genome with many of them evolutionarily conserved (Helmrich et al. 2006; Pentzold et al. 2015). I propose that their peculiar properties contribute to the replication stress test discussed in the previous section. The guild is ordinarily competent to solve the problems posed by fragile sites but they push this competence close to its limits and can become sites of chromosome breakage and instability under conditions of replicative stress.

Interference between RNA and DNA polymerases as they move past each other on common DNA templates poses difficulties for DNA replication (Gómez-Gonzalez and Aguilera 2019). This problem is particularly pronounced at common fragile sites (CFSs) associated with exceptionally large genes. Some of these genes are so large that transcription and processing of an mRNA may extend across more than one cell cycle. Not only is transcription prolonged but these genes also undergo prolonged replication with DNA synthesis extending into the G_2 phase of the cell cycle and sometimes not completed until after entry into mitosis (Glover et al. 2017).

Somatic breakage at CFSs has been variously ascribed to a paucity of replication origins that requires replication forks to travel long distances (Letessier et al. 2011); to various impediments to progress which cause forks to move slowly, including collisions between RNA and DNA polymerases (Helmrich et al. 2011); and to various DNA secondary structures that form difficult-to-replicate roadblocks (Irony-Tur Sinai et al. 2019). If the somatic fragility at CFSs is a pleiotropic side-effect of a rigorous 'stress test' in the germline, then CFSs would be expected to pose multiple problems at once. Each CFS tests the competence of its associated gene whilst at the same time testing the collective competence of the entire machinery of DNA replication and repair.

The CFS associated with *WWOX* will be used as an example chosen because of the gene's exceptionally large size (Krummel et al. 2002). *WWOX* encodes a highly conserved protein with pleiotropic roles in many processes (Abu-Odeh et al. 2014; Abu-Remaileh et al. 2015; Lee et al. 2021). Human WWOX protein contains 414 amino acids translated from a 2241-nucleotide mRNA that is processed from a pre-mRNA of more than 1.1 Mb. Thus, more than 99.8% of the pre-mRNA consists of intronic sequences that are discarded from the mature mRNA. *WWOX*'s two largest introns are 222,600 nucleotides (intron 5) and 778,900 nucleotides (intron 8) (Bednarak et al. 2000; Lee et al. 2021). The enormous size of the *WWOX* gene is deeply conserved in vertebrates. Although *WWOX* exonic sequences and some intronic sequences are deeply conserved, the bulk of the gene is comprised of highly repetitive intronic sequences that undergo rapid evolutionary flux.

DNA synthesis at *WWOX* continues into the G_2 phase of the cell cycle, with some sequences unreplicated at entry to mitosis (Palakodeti et al. 2004). Various impediments slow the progression of replication forks through the CFS including clashes with RNA polymerases and various secondary structures of DNA (Shah et al. 2010; Tubbs et al. 2018; Twayana et al. 2021). The stress-test hypothesis posits that *WWOX* will be associated with problems that challenge its own competence as a means of preserving this competence against mutational deterioration. Because of its immense size, *WWOX* appears particularly vulnerable to replication–transcription conflicts. Therefore, the stress-test hypothesis would be supported if WWOX protein performed functions that contributed to the resolution of such conflicts.

Of particular interest, WWOX promotes repair of double-strand breaks (DSBs) by non-homologous end-joining (NHEJ) rather than homologous recombination (HR) (Schrock et al. 2017). The WW1 domain of WWOX binds to BRCA1, inhibiting end-resection of DSBs, and thus promotes repair by the NHEJ pathway (Park et al. 2022). Collisions between RNA polymerase II and a replication fork could be

negotiated by formation of a DSB behind the replication fork, relieving torsional stress in the encounter zone, and allowing polymerase II to pass via the unbroken strand (Chappidi et al. 2019; Audoynaud et al. (2021). DNA replication is then resumed by repair of the DSB by a process that does not involve end-resection. An attractive hypothesis is that one of the many functions of WWOX is to resolve replication–transcription conflicts.

The enormous size of *WWOX*'s intron 8 means that allelic variation in exons 8 and 9 will be randomized with respect to each other in the generations that intervene between functional exonic mutations. Therefore, the stress-test hypothesis predicts mutational fragility and strong epistasis with respect to mutations within exons but robustness and weak epistasis with respect to interactions between amino acids encoded in distant exons.

BRCA1 was recently identified as an early-replicating fragile site (Deshpande et al. 2022) that "protects against its own fragility" (Martin and McVey 2022). That is, the functional copy of *BRCA1* undergoes a high frequency of mutations because of failures of HR in cells heterozygous for a mutated copy of *BRCA1*. *BRCA1* is much smaller than *WWOX* (81 kb vs. 1.1 Mb) but encodes a much larger protein (1863 vs. 414 amino acids). BRCA1 promotes end-resection at double-strand breaks with repair by the HR pathway (Chen et al. 2017). Several of *BRCA1*'s introns are abundantly populated by *Alu* repeats (Smith et al. 1996). Many pathogenic mutations at *BRCA1* involve homologous recombination among intronic Alu elements or between *BRCA1* and its neighboring pseudogene (Ewald et al. 2009; Caputo et al. 2021). Thus, the *BRCA1* locus has properties that render it sensitive to malfunction of the HR pathway in which it plays an active role. BRCA1 also protects stalled replication forks from degradation by nucleases (Daza-Martin et al. 2019).

BRCA1 protein interacts with WWOX but appears to favor DNA repair by the HR rather than the NHEJ pathway. Therefore, my initial prediction was that replication of *WWOX* would be associated with problems that required high performance of NHEJ whereas replication of *BRCA1* would be associated with problems that required high performance of HR. However, my analysis found that *BRCA1* also poses particularly challenging problems for the machinery of RNA splicing.

In addition to its functions in DNA replication and repair, BRCA1 also has roles in RNA processing (Hatchi et al. 2015; Zhao et al. 2017; Daza-Martin et al. 2019). The *BRCA1* gene exhibits several features that create difficulties for the correct processing of full-length mRNAs that can be translated as a functional protein.

A full-length transcript of human *BRCA1* consists of 22 coding exons. Exons 2–6 encode a RING domain, exons 12–13 encode a coiled-coil domain, and exons 15–24 encode two BRCT domains (Clark et al. 2012). More than 60% of the protein is encoded by the enormous 3.4 kb exon 11. This is one of the largest internal coding exons in the human genome (Raponi et al. 2014) and encodes an intrinsically-disordered protein (Mark et al. 2005). Amino acids encoded by exon 11 are essential for BRCA1's functions in HR (Tammaro et al. 2012; Zhao et al. 2017). Exon 19 (41 nucleotides) is the smallest exon.

Among other functions, BRCA1 recruits splicing factors to sites of DNA damage (Savage et al. 2008) and interacts with the machinery of mRNA 3' end cleavage and polyadenylation (Fontana et al. 2016). The assembly of a full-length *BRCA1* mRNA appears fraught with difficulties associated with splicing. Fifteen of the 20 internal coding exons of the canonical *BRCA1* transcript change phase and only five (exons 3, 5, 11, 18, 21) maintain phase and can be skipped without a frameshift in the encoded protein. Individual skipping of other exons results in nonsense-mediated decay of ephemeral transcripts.

A comprehensive analysis of alternative splicing during *BRCA1* transcription detected 63 distinct splicing events and proposed that these alternative events were randomly combined into hundreds of different *BRCA1* isoforms (Colombo et al. 2014). Every exon was excluded by some events. Some splice variants that delete multiple exons, but maintain an open reading frame, escape nonsense-mediated decay and are common variants. These have no known function. Since that study, the number of competent splicing events detected at *BRCA1* has expanded to more than a hundred (Li et al. 2019). Thousands of different RNAs, many of them highly unstable and difficult to detect, may be generated during the splicing of *BRCA1* mRNA.

Large exons pose problems for exon definition and selection of splice sites (Bolisetty and Beemon 2012). Exon 11 possesses weak donor and acceptor sites and contains 20 cryptic splice sites that must be suppressed for inclusion of the full-length exon (Mucaki et al. 2011). Exon 11's weak donor and acceptor sites result in the exon being skipped in a significant proportion of transcripts as well as variants transcripts that employ an alternative splice donor site within exon 11 (Δ11q) or a polyadenylation site within intron 11 (BRCA1-IRIS) (Tammaro et al. 2012; Raponi et al. 2014). The coding strand of exon 11 contains seven AATAAA polyadenylation sites whereas AATAAA occurs only twice on the non-coding strand. These polyadenylation sites must also be suppressed for a full-length mRNA to be assembled.

Finally, *BRCA1* resides in a very unusual part of the human genome. All human populations possess two major haplotypes with almost complete linkage disequilibrium extending more than 250 kb from

BRCA1 to the *RNU2* locus (Liu and Barker 1999). *RNU2* contains 5–82 copies of a 6.1 kb repeat that encodes the 188-bp U2 snRNA which resides at the heart of the spliceosome (Tessereau et al. 2014). *RNU2* repeats undergo concerted evolution, maintaining homogeneity within the array, while preserving linkage disequilibrium with flanking sequences (Liao et al. 1997; Tessereau et al. 2014). The presence of two major haplotypes in all populations suggests some form of balancing selection maintains both haplotypes. An attractive hypothesis is that *BRCA1* plays some role in the concerted evolution of the *RNU2* repeats which exhibit chromosomal fragility in *BRCA1* deficient cells (Pavelitz et al. 2008). With respect to mutational selection, the entire region of suppressed recombination may be evolving as a single functional unit.

Because of the enormous size of *WWOX*, WWOX protein could be considered a certificate, awarded at the end of a grueling marathon, granting permission to proceed down the next section of the germline. By comparison, the translation of BRCA1 protein from a full-length *BRCA1* mRNA appears to be a triumph of RNA gymnastics. Coding sequences of *WWOX* and *BRCA1* contain regions of strong synonymous constraint. The sequence in Figure 3a comes from the WW1 domain of *WWOX*. The sequences in Figure 3b come from *BRCA1* exon 11. Another region of synonymous constrain near the 5' end of *BRCA1* exon 11 has been analyzed by others (Hurst and Pál 2001; Lind et al. 2011; Macossay-Castillo et al. 2014).

Figure 3a

```
GlyTrpGluGluArgThrThrLysAspGlyTrp
ggcugggaggagagaaccaccaaggacggcugg Homo
.......................cg..... Pan
.......................cg..... Macaca
..u....................cg..... Callithrix
.......................cg..... Mus
.........c....g..........cg..... Bos
.......................cg..... Loxodonta
.......................u..... Dasypus
........a.....g..........cg..... Monodelphis
...........cgc..........cg..... Gallus
..........a...cg.........u...... Xenopus
.............gcg.........u...... Callorinchus
```

Figure 3b

```
GluAspLysIlePheGlyLysThrTyr   ProGluAspPheIleLysLys
gaagacaaaauauuugggaaaaccuau … ccugaggauuuuaucaagaaa Homo
.........................   ..................... Pan
.........................   ..................... Gorilla
.....u...................   ..................... Callithrix
.....u...................   ..................... Bos
.....u...................   ..................... Loxodonta
a....u...................   ..................... Dasypus
...............a.........   ..........c........g Monodelphis
....................a...   .....a..c..c..u.....g Gallus
.....uc....c.....a.....a...   .....a..c........a.c. Latimeria
......c....c............c   ..cg.a..c..c........g Callorinchus
```

Figure 3. Sequence alignments showing regions of synonymous constraint for (a) *WWOX* and (b) *BRCA1*. See legend to Figure 1 for details. Additional species in these alignments are coelacanth (*Latimeria*) and ghostshark (*Callorinchus*).

Conclusions

Mutation and mutational selection jointly determine which mutations are subject to individual selection. All heritable variation in the zygotic gene pool is filtered through their interaction. New mutations are tested against their progenitors on a common genetic background in small demes of cells. By this means, mutational selection changes the frequency of mutations making their zygotic debut. Individual selection then winnows the alleles inherited by zygotes with the survivors subject to further rounds of mutational selection.

If germline and organismal fitness were appropriately aligned, multicellular organisms would benefit from the elimination of deleterious mutations in the germline, by deaths of small numbers of replaceable cells, with most organismal functions performed by somatic cells that inherit high-quality alleles. In this ideal world, mutational selection would favor variants that promoted individual fitness;

beneficial mutations would obtain a boost in the germline that increased their chance of first representation in a zygote and harmful mutations would be eliminated before transmission to a zygote. 'Housekeeping' genes possibly approximate this ideal with germline and individual selection working in concert to maintain essential cellular functions (Hastings 1989, 1991). But it is not an ideal world. Mutational selection increases the frequency of some deleterious mutations that confer proliferative advantages on germ cells (Goriely and Wilkie 2012; Arnheim and Calabrese 2016) and germline selection is blind to mutations with exclusively somatic effects. Nevertheless, Otto and Hastings (1998) have argued that cellular and individual fitness should usually be aligned, with germline selection reducing the genetic load imposed on the population by germline mutations.

Mutational selection favors genetic sequences whose germline phenotypes are easily broken by mutation. This can lead to a runaway process in which genes evolve sequence-properties in *cis* that challenge the competence of their own gene products. This model proposes that conserved fragile sites in the genome have evolved to test the competence of the machinery of DNA replication and repair. Do these processes enhance organismal fitness? On the one hand, mutational fragility and replication stress-tests may strengthen purifying selection for essential cellular functions by enabling the more effective elimination of loss-of-function mutations. On the other hand, mutational fragility may be associated with somatic costs for organismal fitness. The complex interdependencies within a gene's sequence resulting from mutational selection could be considered barriers to market entry by potential competitors, including those with innovative new competencies. The variants that pass the germline examination may not be the best candidates for the somatic job.

A truism of medical genetics is that inherited loss-of-function mutations usually have recessive effects. These recessive alleles have escaped elimination in their germline of origin. One contributing factor may be that mutational selection is blind to mutations with exclusively somatic effects and almost blind to mutations with recessive germline effects. Another factor may be that, at some loci, elite alleles compete for germline dominance, preserving their own competence while seeding the gene pool with alleles of lesser competence.

Acknowledgements

Pavitra Muralidhar, Carl Veller, Marco Archetti, Yaniv Brandvain, and Ford Doolittle have made helpful comments on the manuscript.

References

- Abu-Odeh M, Salah Z, Herbel C, Hofmann TG, Aqeilan RI (2014) WWOX, the common fragile site FRA16D gene product regulates ATM activation and the DNA damage response. *Proceedings of the National Academy of Sciences, USA* 111: E4716–E4725.
- Abu-Remaileh M, Joy-Dodson E, Schuler-Furman O, Aqeilan RI (2015) Pleiotropic functions of tumor suppressor WWOX in normal and cancer cells. *Journal of Biological Chemistry* 290: 30728–30735.
- Archetti M (2009) Survival of the steepest: hypersensitivity to mutations as an adaptation to soft selection. *Journal of Evolutionary Biology* 22: 740–750.
- Arnheim N, Calabrese P (2016) Germline stem cell competition, mutation hot spots, genetic disorders, and older fathers. *Annual Review of Human Genetics* 17: 219–243.
- Audoynaud C, Vagner S, Lambert S (2021) Non-homologous end-joining at challenged replication forks: an RNA connection? *Trends in Genetics* 37: 973–985.
- Bednarak AK, Laflin KJ, Daniel RL, Liao Q, Hawkins KA, Aldaz CM (2000) WWOX, a novel WW domain-containing protein mapping to human chromosome 16q23.3–24.1, a region frequently affected in breast cancer. *Cancer Research* 60: 2140–2145.
- Bolisetty MT, Beemon KL (2012) Splicing of internal large exons is defined by novel cis-acting sequence elements. *Nucleic Acids Research* 40: 9244–9254.
- Caputo SM, Telly D, Briaux A, Sesen J, Ceppi M, Bonnet F, Bourdon V, et al. (2021) 5' region large genomic rearrangements in the BRCA1 gene in French families: identification of a tandem triplication and nine distinct deletions with five recurrent break points. *Cancers* 13: 3171.
- Chappidi N, Nascakova Z, Boleslavska B, Zellweger R, Isik E, Andrs M, Menon S, Dobrovolna J, Pogliano CB, Matos J, Porro A, Lopes M, Janscak P (2019) Fork cleavage-religation cycle and active transcription mediate replication restart after fork stalling at co-transcriptional R-loops. *Molecular Cell* 77: 528–541.
- Chen CC, Feng W, Lim PX, Kass EM, Jasin M (2017) Homology-directed repair and the role of BRCA1, BRCA2, and related proteins in genome integrity and cancer. *Annual Review of Cancer Biology* 2: 313–336.
- Clark SL, Rodriguez AM, Snyder RR, Hankins GDV, Boehning D (2012) Structure–function of the tumor suppressor BRCA1. *Computational and Structural Biotechnology Journal* 1: e201204005.

- Colombo M, Blok MJ, Whiley P, Santamariña M, Gutiérrez-Enríquez S, Romero A, Garre P, Becker A, Smith LD, De Vecchi G, Brandão RD, Tserpelis D, et al. (2012) Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Human Molecular Genetics* 23: 3666–3680.
- Daza-Martin M, Starowicz K, Jamshad M, Tye S, Ronson GE, MacKay HL, Chauhan AS, Walker AK, et al. (2019) Isomerization of BRCA1–BARD1 promotes replication fork protection. *Nature* 571: 521–527.
- Desai MM, Weissman D, Feldman MW (2007) Evolution can favor antagonistic epistasis. *Genetics* 177: 1001–1010.
- Deshpande M, Paniza T, Jalloul N, Nanjangud G, Twarowski J, Koren A, Zaninovic N, Zhan Q, Chadalaveda K, Malkova A, Khiabanian H, Madireddy A, Rosenwaks Z, Gerhardt J (2022) Error prone repair of stalled replication forks drives mutagenesis and loss of heterozygosity in haploinsufficient BRCA1 cells. *Molecular Cell* 82: 3781–3791.
- Dumitrescu CE, Collins MT (2008) McCune-Albright syndrome. *Orphanet Journal of Rare Diseases* 3: 12.
- Ewald IP, Ribeiro PLI, Palmero EI, Cossio SL, Giugliani R, Ashton-Prolla P (2009) Genomic rearrangements in BRCA1 and BRCA2: a feature literature review. *Genetics and Molecular Biology* 32: 437–446.
- Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78: 737–756.
- Fontana GA, Rigamonti A, Lenzken SC, Filosa G, Alvarez R, Calogero R, Bianchi ME, Barabino SML (2016) Oxidative stress controls the choice of alternative last exons via a Brahma–BRCA1–CstF pathway. *Nucleic Acids Research* 45: 902–914.
- Fukuchi S, Noguchi T, Anbo H, Homma K (2023) Exon elongation added intrinsically disordered regions to the encoded proteins and facilitated the emergence of the last eukaryotic common ancestor. *Molecular Biology and Evolution* 40: msac272.
- Fyon F, Cailleau A, Lenormand T (2015) Enhancer runaway and the evolution of diploid gene expression. *PLoS Genetics* 11: e1005665.
- Gardner A, Kalinka AT (2007) Recombination and the evolution of mutational robustness. *Journal of Theoretical Biology* 241: 707–715.
- Glover TW, Wilson TE, Arlt MF (2017) Common fragile sites in cancer: more than meets the eye. *Nature Reviews Cancer* 17: 489–501.

- Gómez-González B, Aguilera A (2019) Transcription-mediated replication hindrance: a major driver of genome instability. *Genes & Development* 33: 1008–1026.
- Goriely A, Wilkie AOM (2012) Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *American Journal of Human Genetics* 90: 175–200.
- Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF (2010) Irremediable complexity? *Science* 330: 920–921.
- Haig D (2021) Concerted evolution of ribosomal DNA: somatic peace amid germinal strife. *BioEssays* 43: 2100179.
- Haig D (2022) Paradox lost: concerted evolution and centromeric instability. *BioEssays* 44: 2200023.
- Hamilton WD (1966) The moulding of senescence by natural selection. *Journal of Theoretical Biology* 12: 12–45.
- Happle R (1986) The McCune-Albright syndrome: a lethal gene surviving by mosaicism. *Clinical Genetics* 29: 321–324.
- Hastings IM (1989) Potential germline competition in animals and its evolutionary implications. Genetics 123: 191–197.
- Hastings IM (1991) Germline selection: population genetic aspects of the sexual/asexual life cycle. *Genetics* 129: 1167–1176.
- Hatchi E, Skourti-Stathaki K, Ventz S, Pinello L, Yen A, Kamieniarz-Gdula K, Dimitrov S, Pathania S, McKinney KM, Eaton ML, Kellis M, Hill SJ, et al. (2015) BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Molecular Cell* 57: 636–647.
- Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Computational Biology* 2: e100.
- Helmrich A, Stout-Weider K, Hermann K, Schrock E, Heiden T (2006) Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. *Genome Research* 16: 1222–1230.
- Helmrich A, Ballarino M, Tora L (2011) Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Molecular Cell* 44: 966–977.
- Holliday R, Grigg GW (1993) DNA methylation and mutation. *Mutation Research* 285: 61–67.
- Hurst LD, Pál C (2001) Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends in Genetics* 17: 62–65.

- Irony-Tur Sinai M, Salamon A, Stanleigh N, Goldberg T, Weiss A, Wang YH, Kerem B (2019) AT-dinucleotide rich sequences drive fragile site formation. *Nucleic Acids Research* 47: 9685–9695.
- Kawachi T, Masuda A, Yamashita Y, Takeda J, Ohkawara B, Ito M, Ohno K (2021) Regulated splicing of large exons is linked to phase-separation of vertebrate transcription factors. *EMBO Journal* 40: e107485
- Kimura M (1967) On the evolutionary adjustment of spontaneous mutation rates. *Genetical Research* 9: 23–34.
- Klug A, Park SC, Krug J (2019) Recombination and mutational robustness in neutral fitness landscapes. *PLoS Computational Biology* 15: e1006884.
- Krummel KA, Denison SR, Calhoun E, Phillips LA, Smith DI (2002) The common fragile site FRA16D and its associated gene *WWOX* are highly conserved in the mouse at Fra8E1. Genes, Chromosomes & Cancer 34: 154–167.
- Lee CS, Choo A, Dayan S, Richards RI, O'Keefe LV (2021) Molecular biology of the *WWOX* gene that spans chromosomal fragile site *FRA16D*. Cells 10: 1637.
- Letessier A, Millot GA, Koundrioukoff S, Lachagès AM, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M (2011) Cell-type-specific replication initiation programs set the fragility of the FRA3B fragile site. *Nature* 470: 120–123.
- Li D, Harlan-Williams LM, Kumaraswamy E, Jensen RA (2019) BRCA1—no matter how you splice it. *Cancer Research* 79: 2091–2098.
- Liao D, Pavelitz T, Kidd JR, Kidd KK, Weiner AM (1997) Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the *RNU2* locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. *EMBO Journal* 16: 588–598.
- Lind MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M (2011) Locating protein-coding sequences under selection for additional, overlapping function in 29 mammalian genomes. *Genome Research* 21: 1916–1928.
- Liu X, Barker DF (1999) Evidence for effective suppression of recombination in the chromosome 17q21 segment spanning *RNU2–BRCA1*. *American Journal of Human Genetics* 64: 4372–1439.
- Lynch M (2011) The lower bound to the evolution of mutation rates. *Genome Biology and Evolution* 3: 1107–1118.
- Macossay-Castillo M, Kosol S, Tompa P, Pancsa R (2014) Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments. *PLoS Computational Biology* 10: e1003607.

- Maley CC, Tapscott SJ (2003) Selective instability: maternal effort and the evolution of gene activation and deactivation rates. *Artificial Life* 9: 317–326.
- Mark WY, Liao JC, Lu Y, Ayed A, Laister R, Szymczyna B, Chakrabarty A, Arrowsmith CH (2005) Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein–protein and protein–DNA interactions? *Journal of Molecular Biology* 345: 275–287.
- Martin SK, McVey M (2022) BRCA1 protects against its own fragility. *Molecular Cell* 82: 3757–3759.
- Mucaki EJ, Ainsworth P, Rogan PK (2011) Comprehensive analysis of mRNA splicing effects of BRCA1 and BRCA2 variants. *Human Mutation* 32: 735–742.
- Orr HA (1995) Somatic mutation favors the evolution of diploidy. *Genetics* 139: 1441–1447.
- Otto SP, Hastings IM (1998) Mutation and selection within the individual. *Genetica* 102/103: 507–524.
- Otto SP, Orive ME (1995) Evolutionary consequences of mutation and selection within an individual. *Genetics* 141: 1173–1187.
- Palakodeti A, Han Y, Jiang Y, Le Beau MM (2004) The role of late/slow replication of the FRA16D in common fragile site induction. *Genes, Chromosomes & Cancer* 39: 71–76.
- Park D, Gharghabi M, Schrock MS, Plow R, Druck T, Yungvirt C, Aldaz CM, Huebner K (2022) Interaction of Wwox with Brca1 and associated complex proteins prevents premature resection at double-strand breaks and aberrant homologous recombination. *DNA Repair* 110: 103264.
- Pavelitz T, Bailey RD, Elco CP, Weiner AM (2008) Human U2 snRNA genes exhibit a persistently open transcriptional state and promoter disassembly at metaphase. Molecular and Cellular Biology 28: 3573–3588.
- Pentzold C, Shah SA, Hansen NR, Le Tallec B, Seguin-Orlando A, Debatisse M, Lisby M, Oestergaard VH (2015) FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells. *Nucleic Acids Research* 46: 1280–1294.
- Raponi M, Smith LD, Silipo M, Stuani C, Buratti E, Baralle D (2014) BRCA1 exon 11 a model of long exon splicing regulation. *RNA Biology* 11: 351–359.
- Savage KI, Gorski JJ, Barros EM, Irwin GW, Manti L, Powell AJ, Pellagatti A, … Harkin DP (2008) Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability. *Molecular Cell* 54: 445–459.
- Savisaar R, Hurst LD (2018) Exonic splice regulation imposes strong selection at synonymous sites. *Genome Research* 28: 1442–1454.

- Schrock MS, Batar B, Lee J, Druck T, Ferguson B, Cho JH, Akakpo K, Hagrass H, Heerema NA, Xia F, Parvin JD, Aldaz CM, Huebner K (2017) Wwox–Brca1 interaction: role in DNA repair pathway choice. *Oncogene* 36: 2215–2227.
- Shabalina SA, Spiridonov NA, Kashina A (2013) Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Research* 41: 2073–2094.
- Shah SN, Opresko PL, Meng X, Lee MYWT, Eckert KA (2010) DNA structure and the Werner protein modulate human DNA polymerase delta-dependent replication dynamics within the common fragile site FRA16D. *Nuclei Acids Research* 38: 1149–1162.
- Shinde DN, Elmer DP, Calabrese P, Boulanger J, Arnheim N, Tiemann-Boege I (2013) New evidence for positive selection helps explain the paternal age effect observed in achondroplasia. *Human Molecular Genetics* 22: 4117–4126.
- Smith TM, Lee MK, Szabo CI, Jerome N, McEuen M, Taylor M, Hood L, King MC (1996) Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1. *Genome Research* 6: 1029–1049.
- Stoltzfus A (1999) On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* 49: 169–181.
- Tammaro C, Raponi M, Wilson DI, Baralle D (2012) BRCA1 exon 11 alternative splicing, multiple functions and the association with cancer. *Biochemical Society Transactions* 40: 768–772.
- Tessereau C, Lesecque Y, Monnet N, Buisson M, Barjhoux L, Léone M, Feng B, Goldgar DE, Sinilkova OM, Mousset S, Duret L, Mazoyer S (2014) Estimation of the *RNU2* macrosatellite mutation rate by *BRCA1* mutation testing. *Nucleic Acids Research* 42: 9121–9130.
- Twayana S, Bacolla A, Barreto-Galvez A, De-Paula RB, Drosopoulos WC, Kosiyatrakul ST, Bouhassira EE, Tainer JA, Madireddy A, Schildkraut CL (2021) Translesion polymerase eta both facilitates DNA replication and promotes increased human genetic variation at common fragile sites. *Proceedings of the National Academy of Sciences, USA* 118: e2016477118.
- Tubbs A, Sridharan S, van Wietmarschen N, Maman Y, Callen E, Stanlie A, Wu W, Wu X, Day A, Wong N, Yin M, Canela A, Fu H, Redon C, Pruitt SC, Jaszczyszyn Y, Aladjem MI, Aplan PD, Hyrien O, Nussenzweig A (2018) Dual roles of poly(dA:dT) tracts in replication initiation and fork collapse. *Cell* 174: 1127–1142.
- Weinstein LS (2006) Gsα mutations in fibrous dysplasia and McCune-Albright syndrome. *Journal of Bone and Mineral Research* 21 (supplement 2): P120–P124.
- Weismann A (1892) *Das Keimplasma. Eine Theorie der Vererbung*. Gustav Fischer, Jena.

- Zhao W, Steinfeld JB, Liang F, Chen X, Maranon DG, Ma CJ, Kwon Y, Rao T, Wang W, Sheng C, Song X, Deng Y, et al. (2017) BRCA1–BARD1 promotes RAD51-mediated homologous DNA pairing. *Nature* 550: 360–365.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.