Massively parallel functional testing of *MSH2* missense variants conferring Lynch Syndrome risk

Xiaoyan Jia[1], Bala Bharathi Burugula[1], Victor Chen[1], Rosemary M. Lemons[1], Sajini Jayakody[1], Mariam Maksutova[1], Jacob O. Kitzman[1,2,*]

[1]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA
[2]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA
*correspondence to kitzmanj@umich.edu

Introduction

The lack of functional evidence for the majority of missense variants limits their clinical interpretability, and poses a key barrier to the broad utility of carrier screening. In Lynch Syndrome (LS), one of the most highly prevalent cancer syndromes, nearly 90% of clinically observed missense variants are deemed 'variants of uncertain significance' (VUS). To systematically resolve their functional status, we performed a massively parallel screen in human cells to identify loss-of-function missense variants in the key DNA mismatch repair factor *MSH2*. The resulting sequence-function map is substantially complete, covering 94% of the 17,746 possible variants, and is highly concordant (≥96%) with existing functional data and expert clinicians' interpretations. The large majority (89%) of missense variants were functionally neutral, perhaps unexpectedly in view of its evolutionary conservation. These data provide ready-to-use functional evidence to resolve the ~1,300 extant missense VUSs in *MSH2*, and may facilitate the prospective classification of newly discovered variants in the clinic.

Main text

Lynch Syndrome (LS, OMIM:120435), the first discovered familial cancer syndrome[1], confers predisposition to colorectal and endometrial cancers due to the loss of DNA mismatch repair (MMR) and resulting mutagenic burden. Most affected individuals inherit a single loss-of-function variant in one of four MMR factors (*MSH2, MLH1, MSH6, PMS2*), followed by a somatic 'second hit' inactivating the remaining functional copy. Due to high population prevalence (≥1:300)[2,3], clear genetic etiology, and potential for prevention through intensified surveillance, MMR gene variants are considered clinically actionable[4].

Missense changes together comprise 20-30% of LS variants[5], and are particularly challenging to interpret: their functional impacts may range from minimal to profound, most are individually rare, and they exhibit incomplete penetrance[6]. Consequently, the overwhelming majority of LS gene missense variants listed in ClinVar (4,762/5,473, 87.0%) are deemed 'variants of uncertain significance', or VUS **(Supplementary Fig.1)**, and cannot be used to guide diagnosis.

To address these challenges, we performed a deep mutational scan[7] to systematically measure the functional impact of missense variants in the major Lynch Syndrome gene *MSH2.* We established a human cell system to model *MSH2* variant function using the mismatch repair proficient[8] cell line HAP1 **(Fig. 1a,c)**. First, to disrupt

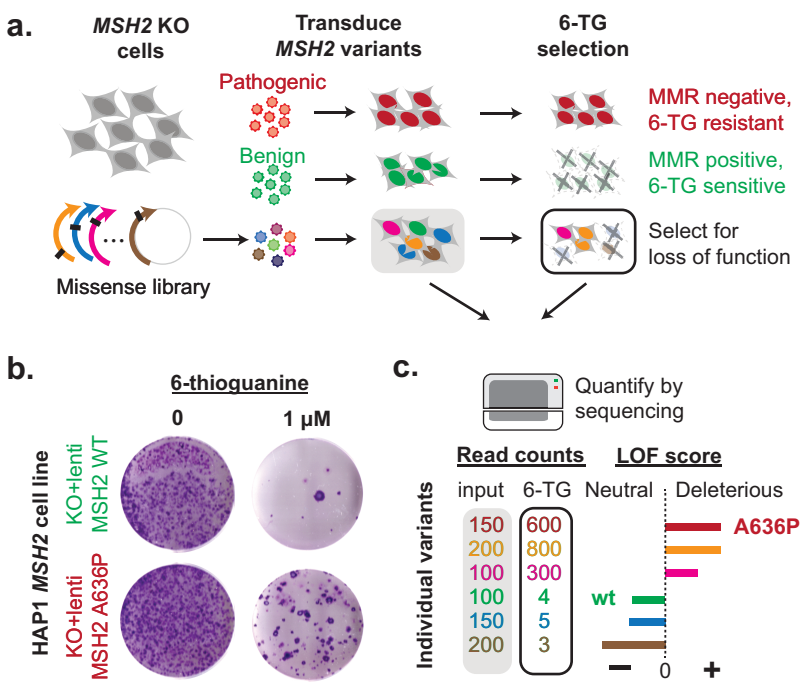

Fig. 1 | Overview of *MSH2* functional screen. a, Generating and testing *MSH2* variant function in isogenic human cells. Cells with pathogenic loss-of-function variants (top) are resistant to 6-thioguanine (6-TG), while complementation by a functional variant (middle) restores 6-TG sensitivity. In a pooled screen (bottom), a mixed library of *MSH2* missense variants is transduced, and 6-TG treatment selects for loss-of-function variants. **b,** 6-TG resistance reflects MMR function in human cells. Clonogenic assay of HAP1 *MSH2* KO cells transduced with either functional (wild-type) *MSH2* cDNA (upper row) or loss-of-function missense variant A636P (bottom row), without (left) or with 6-TG (1 μM, right). **c,** Readout using deep sequencing to quantify abundances of each *MSH2* allele before and after 6-TG selection. The resulting loss-of-function (LOF) score is positive for deleterious variants and negative for functionally neutral ones.

MMR, we derived clonal *MSH2* knockout cells bearing a 19-bp frameshift deletion in *MSH2* exon 6, and verified that it lacked detectable MSH2 protein expression **(Supplementary Fig. 2)**. To restore MMR, we stably reintroduced into cells either wild-type *MSH2* cDNA (KO+WT) or a pathogenic founder allele (KO+A636P) on inducible lentiviral constructs. After inducing expression with doxycycline, wild-type MSH2 protein reached near-endogenous levels, while A636P was only barely detectable, consistent with the known destabilizing effect of this variant[9] **(Supplementary Fig. 3)**.

As a readout for *MSH2* function, we leveraged selection with the purine analog 6-thioguanine (6-TG), which has been used to assay MMR activity in diverse systems[10,11]. Incorporation of 6-TG is selectively toxic to MMR-proficient cells, as it creates lesions that the MMR machinery recognizes but cannot repair, culminating in replication arrest[12]. As expected, KO cells showed increased 6-TG resistance compared to parental HAP1 cells, while 6-TG sensitivity was restored by reintroduction of wild-type *MSH2* but not the pathogenic A636P mutant **(Fig. 1b** and **Supplementary Fig. 3)**. To mimic simultaneously testing thousands of mutations, we prepared a mixed culture of KO+WT cells and KO+A636P cells, with the respective *MSH2* expression constructs bearing distinct, identifying barcode libraries *in cis* **(Supplementary Fig. 4).** We quantified barcode abundances by deep sequencing (**Methods**), and found barcodes linked to the pathogenic variant A636P were strongly enriched, relative to WT barcodes, by 6-TG treatment (median fold change: 179, $P<2.2\times10^{-16}$, Kolmogorov-Smirnov test). This result confirms the feasibility of highly multiplexed MMR functional tests in isogenic human cell mixtures.

To model known and novel *MSH2* variants, we applied single amino-acid saturation mutagenesis[13] to generate libraries comprising every possible missense, synonymous and nonsense variant (**Methods**) *MSH2* cDNA (2,802 bp) was divided into 21 short tiles, to enable tracking by amplicon sequencing, and to mitigate the impact of recombination during lentiviral replication[14]. The mutant cDNA library for each tile was cloned into the inducible lentiviral vector mentioned above, followed by a C-terminal 2A linker and a blasticidin resistance gene to permit selection for complete in-frame cDNA expression **(Supplementary Fig. 3)**. HAP1 *MSH2* KO cells were transduced at low multiplicity (<0.1) to yield a population in which each cell expressed a single mutant *MSH2* variant. Sequencing of integrated libraries from genomic DNA confirmed that nearly all single-codon mutations were present (97.3%) with highly uniform representation **(Supplementary Fig. 5**).

Each *MSH2* mutant cell pool was then selected *en masse* for MMR deficiency. To allow comparison across pools, each was spiked with barcoded control cells (KO+WT: 10% of cells; KO+A636P: 0.5%). Pools were then grown under selective (6-TG+blasticidin) or mock (blasticidin only) conditions for two passages **(Supplementary Fig. 6)**. As expected, the spiked-in *MSH2* A636P cells were consistently enriched over WT by 6-TG selection relative to mock treatment (median fold change: 105; **Supplementary Fig. 7**).

To track the functional status of each *MSH2* variant, we performed amplicon sequencing of each tile throughout the course of selection (**Supplementary Fig. 7-9**). Loss-of-function (LOF) scores for each variant were calculated as the $\log_2$-ratio of that variant's frequency after 6-TG treatment divided by its frequency after mock treatment, such that deleterious variants should score positively and neutral variants negatively **(Fig. 1c)**. LOF scores strongly correlated across replicate 6-TG selections (mean pairwise Pearson's r=0.77, **Supplementary Fig. 10**). As expected, mock treatment had little effect upon most variants' frequencies, except for truncating variants which terminate translation before the downstream blasticidin resistance marker and were therefore depleted (**Supplementary Fig. 11**).

The resulting sequence-function map was substantially complete, covering 94.4% of possible *MSH2* single-codon missense variants **(Fig. 2).** Missense variant scores were bimodal, with the large majority (89.4%) scoring negatively, suggesting a high degree of tolerance to single amino acid substitution (**Fig. 2g**). Indeed, for over half of MSH2 residues (510/934), substitution to any other amino acid was tolerated. As expected, synonymous variants predominantly scored negatively (820/841, 97.5%), with exceptions likely caused by second-site mutations outside the sequenced tile. Most of the scored missense variants were represented by multiple equivalent codons (on average, 3.0), and summarizing these internal

replicates i nto a single, amino-acid level LOF score res ulted in an overall false discovery rate of 0.95%. To verify these results were generalizable across cell types, we repeated the screen for two tiles (tile 14 and 15) in another MMR-proficient human cell line, HEK293, with largely concordant results (**Supplementary Fig. 12**).

We sought to validate these pooled measurements by comparison to traditional, low-throughput functional studies. We collated a set of 126 *MSH2* missense variants previously characterized as functionally deleterious or neutral by individual cell-based[10,11,15–19] and/or biochemical assays[20,21] (**Supplementary Tables 1-2**). For the great majority (121/126, 96.0%; **Fig. 3a**), our function scores agreed with earlier reports. Among 3 of the 5 discordant variants, our function scores were internally consistent across multiple codons encoding the same amino acid substitution, and with similar substitutions at the same residue (**Supplementary Fig. 13**). These may reflect false negatives on the part of the previous screens, or differences between this human cell model and previous yeast or *in vitro* assays.

Bioinformatic tools are frequently used to interpret clinical variants in the absence of other evidence[22]. We compared our LOF scores' classification performance with that of six computational predictors, taking the previously reported functional evidence as a truth set (**Fig. 3b**). Our LOF scores classified these variants with a Matthews Correlation Coefficient of 0.93, higher than all six tools' (range: 0.50-0.77). General-purpose, pathway-agnostic classifiers exhibited poor specificity, possibly reflecting the challenge of distinguishing functional effects between variants in highly conserved genes such as *MSH2.* Although the bioinformatic predictions were modestly correlated with our

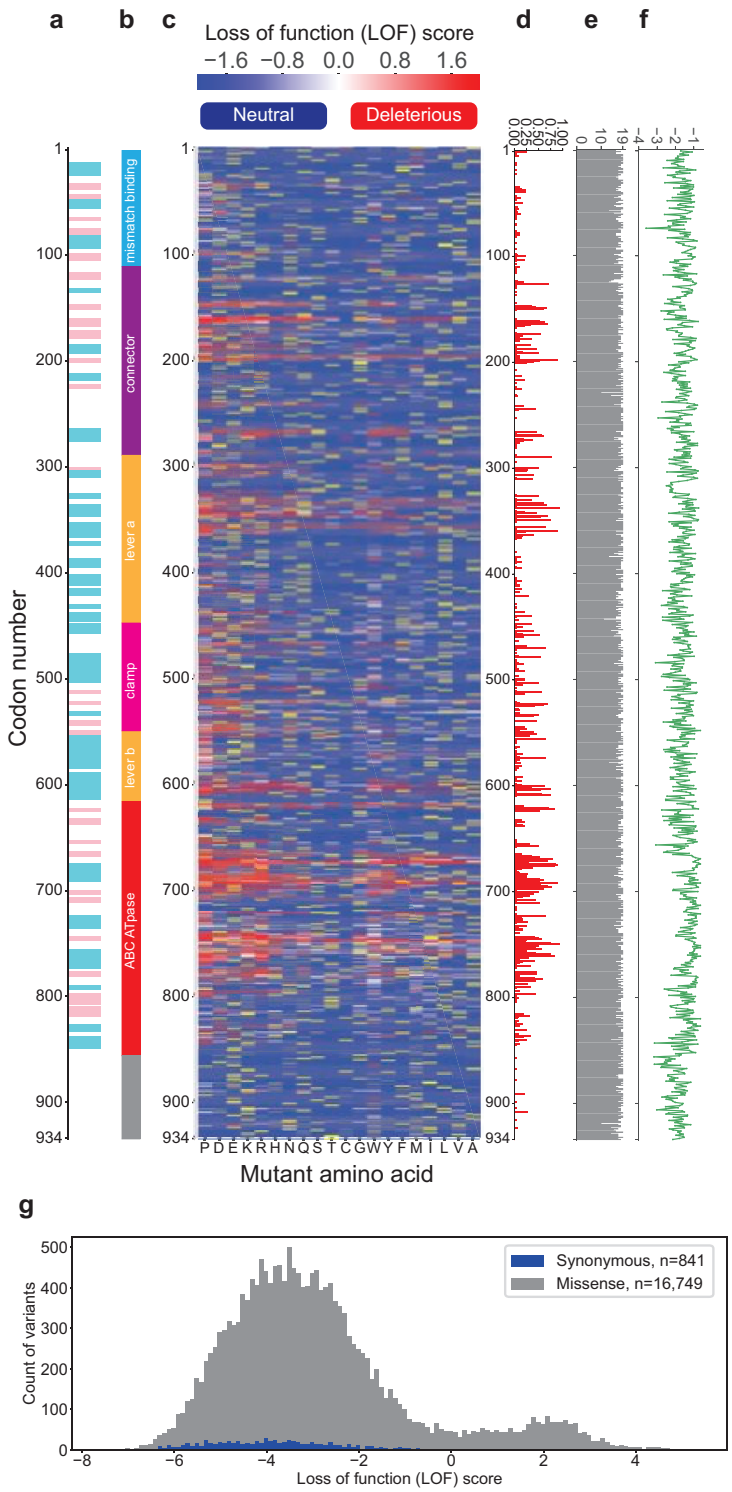

Fig. 2 | Sequence function map of *MSH2* missense variants. a, MSH2 secondary structure (cyan: alpha helices, pink: beta sheets), and **b**, protein domains. **c**, Heatmap of loss-of-function (LOF) scores after mutating each of 934 residues in MSH2 (rows) to each of 19 other possible amino acids (columns). Negative scores (blue) indicate functionally neutral variants, while positive scores (red) indicate deleterious ones; Yellow denotes wild-type amino acid, gray: no data. **d**, Fractions of substitutions which are disruptive (LOF score>0) by position. **e**, Number of scored missense variants by position (max. possible: 19). **f**, Evolutionary conservation score (PSIC) by position. **g**, Histogram of scores by variant class.

experimental measurements for variants in the functional validation set, they were markedly less concordant for other variants at those same residues, or throughout *MSH2* at large (**Supplementary Fig. 14**). As variant effect predictors are often trained on the limited number of known variants with available classifications, their divergence with our experimental measurements may reflect overfitting.

To assess the suitability of these scores for classifying variant pathogenicity, we intersected them with clinical variant databases. We retrieved all *MSH2* missense variants with consensus annotations from InSight[5] and/or expert review in ClinVar, totaling 75 variants. After excluding 9 variants previously known or predicted to impact splicing[23] (SpliceAI score>0.2), our LOF scores agreed with clinical annotations for 63 of the 66 remaining variants (**Fig. 3c**). These data provide evidence to facilitate interpretation of the 1,374 variants without classification (i.e., VUS) or with conflicting reports. Our scores predicted that 112/1,374 (8.2%) of these unresolved variants are functionally deleterious, similar to the overall rate of LOF among all missense variants reachable by a single base mutation (7.7%; P=0.51, two-sided binomial test). This result highlights the pervasiveness of benign missense variation in clinical databases and identifies a subset of variants likely to be causal for LS.

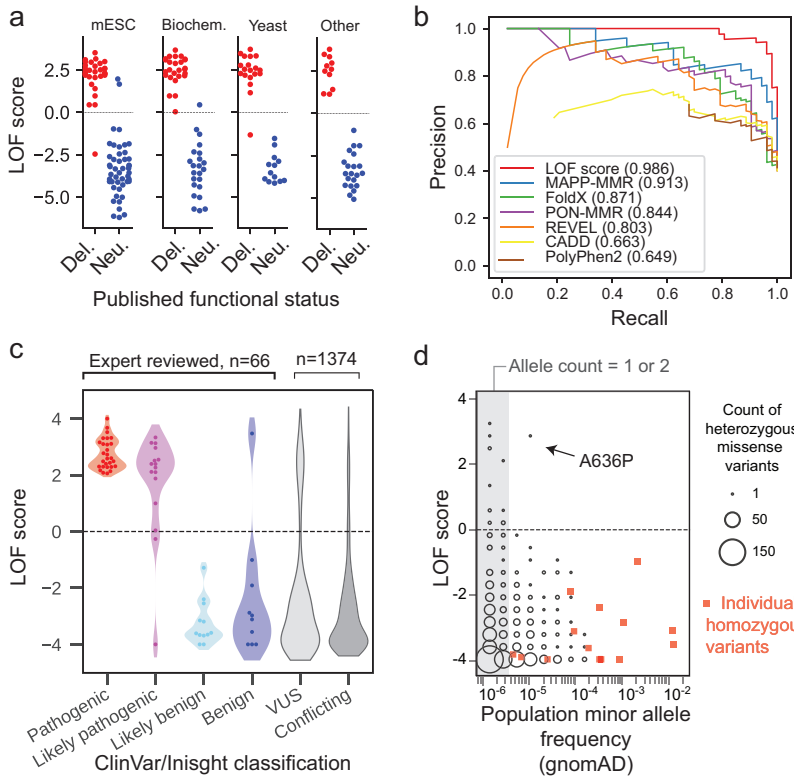

Fig. 3 | Concordance with existing functional and clinical classifications. a, Measured LOF scores (y-axis) for published *MSH2* missense variants (n=209 reports covering 126 distinct variants), previously characterized by individual cellular[10,11,15–19] or biochemical assays[20,21]. **b**, Precision-recall curves show LOF scores outperform all tested bioinformatic callers in predicting functional status of variants in (**a**), with area under the curve indicated for each. **c**, LOF scores for patient missense variants from ClinVar and InSight databases, by variant classification. **d,** LOF scores versus allele frequency for missense variants in the gnomAD database[24]; size of circle denotes number of heterozygous variants; red squares mark missense variants observed in homozygous state in ≥1 individual. Gray region denotes singletons/doubletons (gnomAD allele count ≤2 individuals).

We next examined how these function scores varied in apparently healthy populations (**Fig. 3d**). The gnomAD database[24] lists 744 of the scored *MSH2* missense variants, nearly all rare (742/744 with minor allele frequency < 0.5%). Among these, only 18 scored as deleterious in our assay (all singletons or doubletons except the founder allele A636P), a significant depletion (P=6.9x10^{-7}, two-sided binomial test) relative to randomly drawn, size-matched sets of single-base missense variants. All 14 *MSH2* missense variants observed in the homozygous state in gnomAD scored as functionally neutral, as expected given that inherited biallelic MMR loss causes pediatric-onset cancer syndromes. These results reflect survivorship bias which depletes adult cohorts of deleterious variants in genes underlying disorders such as LS, as previously noted for *MSH2* deletions[25].

Deep mutational scans can inform protein structure-function relationships[26]. On aggregate, chemically conservative substitutions across MSH2 were rarely deleterious (5.63%), compared to substitutions replacing hydrophobic with charged residues (28.0%; **Supplementary Fig. 15**). Mapping function scores onto the crystal structure[27] (**Fig. 4a**), we noted several regions of mutational intolerance corresponding

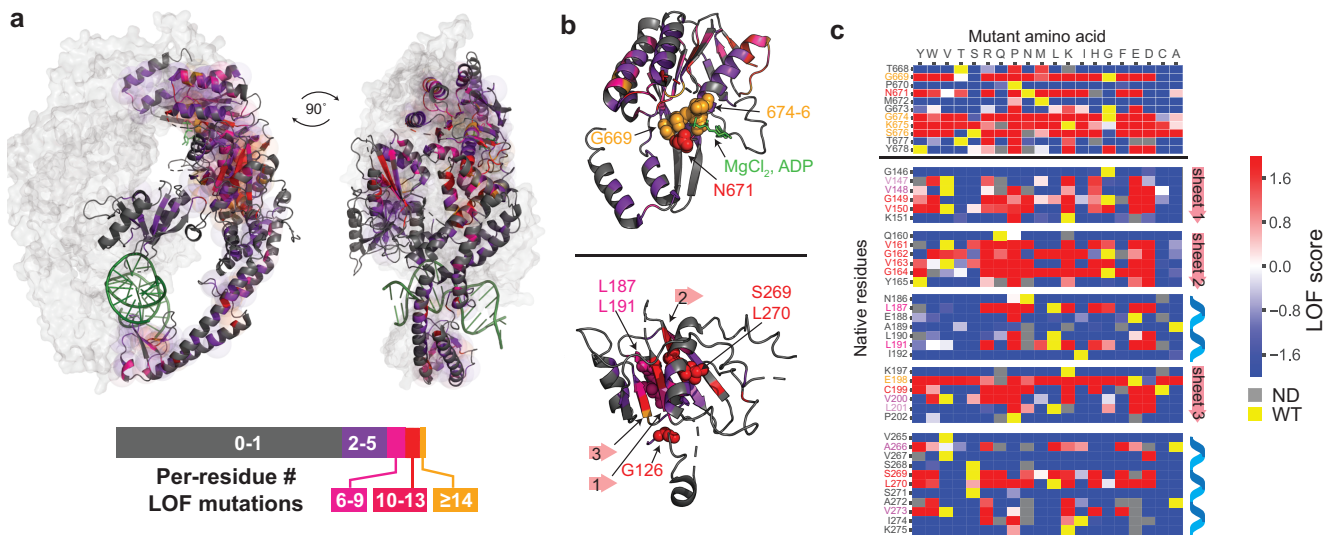

Fig. 4 | Structural insights from deep mutational scanning data. a, MutSα crystal structure (PDB:2O8E), with MSH6 surface in light gray, and MSH2 shaded by the number of deleterious missense variants at each residue (length of bar in legend proportional to count of residues within denoted range of LOF mutations). **b**, Extreme intolerance to substitution in MSH2 ATPase domain (upper), and connector domain (lower); selected residues or beta sheets shown with spheres or arrows, respectively. **c.** LOF score heatmap (as in **Fig. 2c**) showing mutational intolerance at Walker A motif of the ATPase domain (upper) and selected connector domain residues (lower) corresponding to (**b**).

to structural features. The phosphate-binding Walker A motif (residues 674-676) is required for the essential ATPase activity of MSH2[28] and was almost completely intolerant to mutation (50/55 mutations LOF, **Fig. 4b**). We also noticed strong constraint at glycine 126, possibly reflecting the requirement for flexibility to accommodate conformational shifts between the mismatch binding and connector domains (**Fig. 4c**)[27]. The connector domain's hydrophobic core, comprising three beta sheets and inward-facing residues of the surrounding helices, was also largely functionally intolerant to substitutions. Conversely, despite its name, the N-terminal mismatch binding domain (residues 1-124) was largely devoid of missense LOF variants, consistent with its dispensability in yeast[29], and observations that *MSH2* start codon loss is not strongly pathogenic in humans[30]. By contrast, prokaryotic mutS homodimerizes, with both N termini making extensive DNA contacts essential for mismatch recognition[31]. Several positively charged clamp domain residues (K512, R524, K546), which interact with the DNA backbone by hydrogen bonding[27], were strongly intolerant to substitution, especially to negatively charged residues. Although computational predictions of protein stability[32] showed modest overall correlation with LOF scores (Pearson's r=0.39, **Supplementary Fig. 14**), we noted a few exceptions including the ATPase Walker motifs, which showed strong constraint in our data but were not predicted to be destabilizing (**Supplementary Fig. 15**), suggesting mechanisms other than protein instability.

We also examined the scores of variants accepted by evolution. Given the strong pressure to maintain genome integrity, *MSH2* variants rising to fixation in extant lineages are expected to be functionally neutral, except for potential epistatic effects of other mutations *in cis* or trans[33]. We aligned the human MSH2 protein sequence with seven orthologs distributed across eukaryota, identifying 1707 amino acid substitutions relative to the human sequence. Of those scored, nearly all (1,627/1,639, 99.3%) had negative LOF scores (**Supplementary Fig. 16**), indicating these substitutions are functionally neutral in the context of the human MSH2 sequence, and supporting these scores' specificity.

We have described a nearly complete sequence-function map for the key Lynch Syndrome gene *MSH2*, and showed its predictive value in separating functionally intact variants from deleterious ones. This screen exploited the direct link between molecular dysfunction – the inability to correct replicative errors – and pathogenesis in Lynch Syndrome, and provides a blueprint for comprehensive functional annotation of the three other major LS genes. A limitation of our cDNA-based measurements is that they cannot model effects upon splicing, although among known LS variants and in other saturation screens[34],

248 splice disruption accounts for only a small fraction of pathogenic variants. Other genes with more varied
249 functions, or unknown pathogenic mechanisms, may require multiple assays for different aspects of
250 function, or may benefit from more generalized functional readouts such as protein stability[35],
251 localization[36], or target expression[37]. These functional scores, when rationally combined with other lines
252 of clinical evidence[22], will enable more accurate *MSH2* variant classification, and provide carriers of
253 deleterious variants the opportunity to learn their status and seek medical interventions[38].
254

Acknowledgements
We thank M. Meisler, A. Antonellis, E. Stoffel, S. Hemker, and A. Glazer for helpful discussions, J. Moran for providing sequencer access, and J. Moldovan for assistance with clonogenic assays. This work was supported by the National Institute of General Medical Sciences (5R01GM129123 to J.O.K.) and a Precision Health Fellowship from the University of Michigan (to X.J.).

Author contributions
J.O.K. conceived of and supervised the project. J.O.K. and X.J. designed the experiments. X.J. and B.B. executed the experiments with assistance from V.C., R.M.L., S.J., and M.M. X.J. and J.O.K. performed the data analysis. X.J. and J.O.K. prepared the manuscript with input from all other authors.

Competing interests
The authors declare no competing interests.

References

1. Lynch, H. T. & Krush, A. J. Cancer family 'G' revisited: 1895-1970. *Cancer* **27**, 1505–1511 (1971).
2. Win, A. K. *et al.* Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer. *Cancer Epidemiol. Biomarkers Prev.* **26**, 404–412 (2017).
3. Haraldsdottir, S. *et al.* Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2. *Nat. Commun.* **8**, 14755 (2017).
4. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2. 0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249 (2017).
5. Thompson, B. A. *et al.* Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat. Genet.* **46**, 107–115 (2014).
6. Hampel, H. & de la Chapelle, A. How do we approach the goal of identifying everybody with Lynch Syndrome? *Fam. Cancer* (2013).
7. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).
8. Harmsen, T., Klaasen, S., van de Vrugt, H. & Te Riele, H. DNA mismatch repair and oligonucleotide end-protection promote base-pair substitution distal from a CRISPR/Cas9-induced DNA break. *Nucleic Acids Res.* **46**, 2945–2955 (2018).
9. Foulkes, W. D. *et al.* The Founder Mutation MSH2*1906G→C Is an Important Cause of Hereditary Nonpolyposis Colorectal Cancer in the Ashkenazi Jewish Population. *Am. J. Hum. Genet.* **71**, 1395–1412 (2002).
10. Gammie, A. E. *et al.* Functional characterization of pathogenic human MSH2 missense mutations in Saccharomyces cerevisiae. *Genetics* **177**, 707–721 (2007).
11. Houlleberghs, H. *et al.* Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4128–4133 (2016).
12. Yan, T., Berry, S. E., Desai, A. B. & Kinsella, T. J. DNA mismatch repair (MMR) mediates 6-thioguanine genotoxicity by introducing single-strand breaks to signal a G2-M arrest in MMR-proficient RKO cells. *Clin. Cancer Res.* **9**, 2327–2334 (2003).
13. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–6, 4 p following 206 (2015).
14. Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. & Elledge, S. J. Sources of Error in Mammalian Genetic Screens. *G3* **6**, 2781–2790 (2016).
15. Rath, A. *et al.* Functional interrogation of Lynch syndrome-associated MSH2 missense variants via CRISPR-Cas9 gene editing in human embryonic stem cells. *Hum. Mutat.* **40**, 2044–2056 (2019).
16. Martinez, S. L. & Kolodner, R. D. Functional analysis of human mismatch repair gene mutations identifies weak alleles and polymorphisms capable of polygenic interactions. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5070–5075 (2010).
17. Mastrocola, A. S. & Heinen, C. D. Lynch syndrome-associated mutations in MSH2 alter DNA repair and checkpoint response functions in vivo. *Hum. Mutat.* **31**, E1699–E1708 (2010).
18. Wielders, E. A. L. *et al.* Functional analysis of MSH2 unclassified variants found in suspected Lynch syndrome patients reveals pathogenicity due to attenuated mismatch repair. *J. Med. Genet.* **51**, 245–253 (2014).
19. Brown, G. D. Biochemical and Cellular Studies of Human MSH2 ATPase Domain Mutants Support the Molecular Switch Model of Mismatch Repair In Vivo. (University of Connecticut, 2014).
20. Drost, M. *et al.* A rapid and cell-free assay to test the activity of lynch syndrome-associated MSH2 and MSH6 missense variants. *Human Mutation* vol. 33 488–494 (2012).
21. Drost, M. *et al.* A functional assay–based procedure to classify mismatch repair gene variants in Lynch syndrome. *Genet. Med.* **21**, 1486–1496 (2019).
22. Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).
23. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**,

535–548.e24 (2019).
24. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
25. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *Am. J. Hum. Genet.* **105**, 373–383 (2019).
26. Gray, V. E., Hause, R. J. & Fowler, D. M. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* **207**, 53–61 (2017).
27. Warren, J. J. *et al.* Structure of the Human MutSα DNA Lesion Recognition Complex. *Mol. Cell* **26**, 579–592 (2007).
28. Graham, W. J., Putnam, C. D. & Kolodner, R. D. The properties of Msh2–Msh6 ATP binding mutants suggest a signal amplification mechanism in DNA mismatch repair. *J. Biol. Chem.* (2018).
29. Kumar, C. *et al.* Multiple Factors Insulate Msh2–Msh6 Mismatch Repair Activity from Defects in Msh2 Domain I. *J. Mol. Biol.* **411**, 765–780 (2011).
30. Cyr, J. L., Brown, G. D., Stroop, J. & Heinen, C. D. The predicted truncation from a cancer-associated variant of the MSH2 initiation codon alters activity of the MSH2-MSH6 mismatch repair complex. *Molecular Carcinogenesis* vol. 51 647–658 (2012).
31. Obmolova, G., Ban, C., Hsieh, P. & Yang, W. Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* vol. 407 703–710 (2000).
32. Nielsen, S. V. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet.* **13**, e1006739 (2017).
33. Jordan, D. M. *et al.* Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* **524**, 225–229 (2015).
34. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
35. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
36. Penn, W. D. *et al.* Probing biophysical sequence constraints within the transmembrane domains of rhodopsin by deep mutational scanning. *Sci Adv* **6**, eaay7505 (2020).
37. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).
38. Hampel, H. & de la Chapelle, A. The search for unaffected individuals with Lynch syndrome: do the ends justify the means? *Cancer Prev. Res.* **4**, 1–5 (2011).

Methods

General

Except as noted, cell culture media and supplements were obtained from Gibco, PCR reactions used Kapa HiFi Polymerase kit (Kapa Systems), nucleic acid cleanup kits were from Zymo Research, chemicals were from Sigma, cloning enzymes were from NEB, and oligonucleotides were synthesized by IDT or Eurofins. Primer oligonucleotide sequences are listed in Supplementary Table 3 and PCR cycling programs are listed in Supplementary Table 4.

Cell culture

Human HAP1 cells (Horizon Discovery) were cultured in IMDM; HEK293 and 293T/17 cells (ATCC) were cultured in DMEM, each supplemented with 10% fetal bovine serum, 100 U/ml penicillin and 100 g/ml streptomycin. During routine passaging, cells were washed with PBS pH 7.4 and dissociated with TrypLE Express. Cell lines were tested monthly for mycoplasma; all were negative.

MSH2 knockout cell lines

Clonal knockout lines were derived by transient transfection with plasmids expressing SpCas9 and guide RNAs targeting *MSH2*. For HAP1, a single guide sequence (primers JBW0001/2) targeting *MSH2* exon 6 was cloned into pSpCas9(BB)-2A-GFP (from Feng Zheng lab, Addgene #48138) as described[39]. For HEK293, four guides (primers VC0019/20, VC0021/22, VC0023/24, VC0027/28), two flanking each side of the *MSH2* locus, were each likewise cloned into pSpCas9(BB)-2A-PuroV2 (Feng Zheng lab, Addgene #62988) and used in combination. Wild-type HAP1 or HEK293 cells were transfected with the respective plasmid(s) using Lipofectamine 3000 (Thermo). Clonal HAP1 *MSH2* knockout cells were isolated by sorting single GFP positive cells using a FACS Aria II instrument (BD), and colonies were screened for *MSH2* exon 6 frameshift mutations. Clonal HEK293 *MSH2* knockout cells were isolated by dilution, expanded, and screened by PCR for the presence of a full-gene knockout deletion junction(s), and the absence of an internal target in exon 12 (Supplementary Fig. 12).

MSH2 cloning

Wild-type *MSH2* cDNA was PCR amplified from a lentiORF plasmid (Dharmacon OHS5897), and fused in-frame by HiFi Assembly to a blasticidin resistance marker amplified from pLentiCas9Blast (Feng Zheng lab, AddGene #52962), separated by a P2A self-cleaving peptide linker. The p.Ala636Pro mutation was introduced by site-directed mutagenesis. Each resulting construct was subcloned the doxycycline-inducible lentiviral expression plasmid pCW57.1 (from David Root lab, Addgene #41393) by digestion with SalI and NheI and ligation with T4 DNA ligase.

Western blotting

Cells were seeded on 6-well plates, grown to confluence, and lysed by the addition of RIPA buffer and protease inhibitor cocktail (Sigma-Aldrich). Protein extracts were run for 30 min at 200 V on denaturing Bolt 4-12% Bis-Tris Plus Gels in 1× Bolt MOPS SDS buffer (Invitrogen), transferred to a nitrocellulose membrane for 1.5 h at 10 V, and blocked at 4˚C overnight in 1X TBST (20 mM Tris base, 0.15 M NaCl, 0.1% Tween-20, 2% milk). Blots were probed with a 1:2000 dilution of anti-MSH2 (FE11, Thermo) and as loading control, 1:5000 dilution of anti-beta actin (BA3R, Invitrogen), followed by a 1:1000 dilution of HRP-conjugated goat anti-mouse IgG (H+L) secondary antibody (#62-6520, Invitrogen). Detection was performed with SuperSignal West Pico Chemiluminescent substrate (Thermo) and ChemiDoc XRS system (BioRad).

Library mutagenesis

MSH2 cDNA was divided into 21 tiles (Supplementary Table 5), and each was subjected to programmed allelic series mutagenesis[13] modified as follows. Mutagenic PCR reactions used PrimeStar GXL polymerase (Takara), and 45mer mutagenesis primers (Supplementary Table 3) were individually synthesized. Each codon was randomized in an individual PCR (program MUT_1), with a common

321 forward primer (JOK_0184) upstream of *MSH2,* and a codon-specific degenerate (NNN) reverse primer.
322 Resulting partial-length mutant cDNA was subjected to 10 cycles of strand extension (program MUT_2)
323 with limiting template (250 pg) of wild-type cDNA plasmid. Finally, each full-length mutant cDNA was
324 PCR amplified with flanking outer primers (program MUT_3, primers JOK_0184 and JOK_0185). Per-
325 codon products were pooled within each tile and purified by SPRI bead cleanup[40]. PCR products digested
326 with DpnI, to deplete starting plasmid, and NheI-HF and SalI-HF, for cloning into the lentiviral vector as
327 described above. Ligation products were transformed into Endura electrocompetent *E. coli* (Lucigen). A
328 small aliquot of the transformant culture was plated to monitor library complexity, while the remainder
329 was expanded for 12-16 hours at 32ºC in Luria Broth with 100 ug/ml ampicillin (LB+amp). We required
330 total colony forming units ≥ $4x10^4$, for mean ≥14-fold coverage per variant. Lentiviral tile plasmid libraries
331 were isolated using ZymoPURE II Plasmid kit, and converted to shotgun libraries for deep sequencing
332 by Tn5 tagmentation[41], to assess mutational representation and uniformity (Supplementary Fig. 9).
333 Absent or under-represented codons were individually retrieved from the earlier mutagenesis steps,
334 pooled, cloned as above, and spiked into each tile library as needed to equalize coverage. Finally, these
335 mutant tile libraries, as well as control clones (WT and A636P), were modified to add trackable barcodes.
336 Each clone or library was linearized with SalI-HF (NEB) downstream of the *MSH2*-2A-bl^R ORF, and a
337 degenerate 20-mer (primer jklab0172) was inserted by HiFi Assembly (NEB). Resulting barcoded clone
338 pools were transformed, expanded, and isolated as described above.
339
340 **Lentiviral preparation and transduction**
341 Barcoded *MSH2*-2A-bl^R cDNA libraries were packaged into lentivirus by co-transfecting HEK293T/17
342 cells (ATCC) with the transfer plasmid pool plus envelope and packaging vectors pMD2.G and psPAX2
343 (from Didier Trono's lab, Addgene #12259 and #12260). For each pool, $4.4x10^6$ cells were plated in a
344 100 mm dish, then transfected with 17 ug total plasmid (2.0:1.0:1.3 transfer:envelope:packaging ratio),
345 using Lipofectamine 3000 (Thermo). Media was replaced at 6 h post transfection, and viral supernatants
346 were collected at 24 h, passed through an 0.45 micron filter (Millipore), and used immediately or stored
347 at -80ºC until use. Viral titer was estimated by transduction with a dilution series followed by puromycin
348 selection and CellTiterGlo (Promega) cell proliferation assay[42], and verified by counting unique barcodes
349 in the transduced population (barc-seq, described below). For each tile, *MSH2* knockout HAP1 or
350 HEK293 cells were transduced with mutant library at low multiplicity of infection (< 0.1), by applying 1.5
351 ml viral supernatant (with 8 µg mL^{-1} polybrene) to each of four 100mm dishes (Corning) containing
352 ~$7.5x10^6$ cells. After 48 h, transduced cells were selected by addition of 1 µg mL^{-1} puromycin, which was
353 included in all subsequent steps.
354
355 **Mismatch repair functional selection**
356 Cells expressing loss-of-function *MSH2* variants were enriched by adding 6-thioguanine (6-TG) to the
357 culture media. To find an optimal concentration, parental and *MSH2* KO cell lines were plated in
358 quadruplicate wells of a 96-well plate and treated with a range of 6-TG concentrations. Cell growth after
359 4 days was measured using the CellTiterGlo 2 kit (Promega) and a GloMax luminometer (Promega), and
360 survival was normalized to untreated wells (Supplementary Fig. 2).
361
362 For library selections, each mutant tile cell population was spiked with barcoded knock-in control cells
363 (*MSH2* WT: to 10% of the population, A636P: to 0.5%), then split and expanded in triplicate 10 cm dishes
364 (Supplementary Fig. 6) in media supplemented with 1 µg mL^{-1} doxycycline to induce *MSH2* expression.
365 From each baseline passage ('P0') culture, 10^7 cells were collected for gDNA extraction, and cells were
366 split into 6-TG treated and mock-treated cultures, with media containing 1 µg mL^{-1} puromycin, 1 µg mL^{-1}
367 doxycycline, 5 µg mL^{-1} blasticidin (to enforce *MSH2* expression), with and without 1 µM 6-TG,
368 respectively. Cells were grown to confluence for two rounds, at each round harvesting ≥10^7 cells for
369 gDNA extraction and seeding with $3x10^6$ cells to avoid population bottlenecks.
370
371 **Amplification and sequencing of integrated *MSH2***
372 After selections, genomic DNA was isolated from 1-$3x10^7$ cells using the Quick-DNA Midiprep Plus kit.
373 To provide sufficiently redundant sampling of integrated *MSH2* libraries, for each sample, 16 replicate
374 PCR reactions were assembled, each with 0.5 µg of gDNA as template (template copy number of $2.4x10^6$

genome equivalents per tile). Integrated inserts (~3.7 kbp), including *MSH2* cDNA and associated barcode, were PCR amplified with PrimeSTAR GXL (program MSH2gDNA). Each set of 16 replicate amplicons was pooled and purified with SPRI beads. Next, tileseq libraries were prepared from each pool of amplified integrated *MSH2* constructs. Tile-specific primers (Supplementary Table 3) were used to further amplify the tile mutagenized in each sample in a PCR reaction (program TILESEQ) with 10 ng full-length amplicon (≥ 2.6×10^9 copies) as template. Dual-indexed illumina adapters (sequences available upon request) were then added by a subsequent round of PCR. The resulting short amplicon libraries thus provide overlapping read coverage for most of each tile. Tile-seq libraries were pooled for paired end 150bp Illumina sequencing (≥ 1×10^6 paired reads per sample, n=189 libraries). Library complexity (the number of distinct transduced clones) was assessed by barc-seq. Barcodes were amplified from the full-length *MSH2* amplicon and converted to illumina amplicon libraries, as for tileseq, but using primers jklab0077 and jklab0078 (program BCSEQ) . Each tile's starting population ('P0') had ≥20,000 barcodes, indicating at least that number of unique initial transduction events (Supplementary Fig. 7). Catalogs of barcodes associated with each of the two *MSH2* control clones (WT, A636P), and for each tile library, were established by barc-seq (as described above) from their respective gDNAs at P0 (after transduction). Finally, tile-seq libraries were prepared from a fully wild-type *MSH2* plasmid DNA, to measure and correct by position-specific error rates, primarily single base substitutions which varied considerably by position and type (Supplementary Fig. 8)

Sequence data processing

Shotgun sequencing reads were aligned to a set of *MSH2* cDNA sequences with each codon sequentially replaced with 'NNN' using bbmap (https://jgi.doe.gov/data-and-tools/bbtools/) in semi-perfect mode, to identify reads containing at most one mutated codon with between one and three edits. Custom python scripts converted the resulting alignments to a per-sample table of counts for every possible codon mutation. For tile-seq libraries, paired reads were overlapped and error-corrected with PEAR v0.9.6 (ref. [43]); only perfectly matching pairs were used. For each tile, all possible single-codon mutation sequences were enumerated, and the count of reads exactly matching each mutant sequence was tabulated. Reads with additional mutations or sequencing errors were discarded. Barc-seq reads were clustered with starcode v1.3 (ref. [44]) to define and count unique barcodes. To determine the number of distinct barcodes present, robust to sequencing or amplification artifacts, we ranked clustered barcode groups by read count and determined the minimum number of barcodes for which all equally or more abundant barcodes cumulatively accounted for ≥90% of the total read counts.

Variant enrichment scoring

The raw frequency of each per-codon mutation *i* in each sample *s* was calculated, with a pseudocount of 0.9: $f'_{i,s} = \frac{(c_{i,s}+0.9)}{\sum_k (c_{k,s}+0.9)}$. Position specific error rates from wild-type *MSH2* plasmid tile-seq were subtracted and normalized to yield corrected frequencies: $f_{i,s} = \frac{max\,(f'_{i,s} - f'_{i,WT}, 0)}{\sum_k max\,(f'_{k,s} - f'_{k,WT}, 0)}$. Variants with frequency ≤1/200,000 in the initial transduced cell population (P0) were removed from further analysis. Next, variants substantially enriched or depleted after mock treatment (absolute $\log_2$ ratio > 2, after two passages blasticidin vs P0) were likewise removed; these variants may be *in cis* with a frameshift or premature truncating mutation beyond the sequenced tile. A loss-of-function (LOF) score was taken as the $\log_2$ratio of the frequency after 6-TG treatment over that after mock treatment. For amino acid substitutions represented by more than one equivalent codon substitution, the median was taken across those codons' LOF scores to yield an amino acid-level LOF score. Within each replicate, variants were removed when represented by exactly two LOF scores with opposite signs. Finally, the median score was taken for each amino acid LOF score across the three replicates; when only two replicate scores remained and had discordant signs, they were removed. False discovery rate was estimated by random sampling of synonymous variants' read counts, with replacement and keeping triplicates together. To mimic the number of codons corresponding to a given mutant amino acid, synonymous variants were sampled in groups of size one to six. Counts were processed to LOF scores and filtered as described above, combining first within each simulated missense amino acid and then across replicates; final LOF

427 scores greater than zero were taken as false discoveries. Overall FDR was estimated as the mean FDR
428 for amino acids with each number of different variant codons (1, 2, 3, 4, or 6 codons), weighted by the
429 fraction of variant amino acids with that number of codons.
430
431 **Clinical and population variants**
432 *MSH2* missense variants and classifications were obtained from ClinVar
433 (https://www.ncbi.nlm.nih.gov/clinvar/) on April 14, 2020. Redundant variants were removed, and likely
434 splice-disruptive variants with a SpliceAI[23] score of ≥0.2 were filtered out, along with those where the
435 ClinVar record supported a splicing disruptive mechanism (Supplementary Table 2). Only variants with
436 expert panel reviews were included in comparisons to LOF scores. *MSH2* missense variants and
437 classifications were accessed from InSiGHT (https://www.insight-group.org/variants/databases/). *MSH2*
438 missense population variation was obtained from the gnomAD database[24], combining whole-genome and
439 exome calls from versions 2.1.1 and 3.0. The expected number of LOF missense mutations was modeled
440 with a binomial distribution parameterized by n=744 (the number of scored missense variants in gnomAD)
441 and p taken as the sum of LOF missense point variants' relative likelihoods of occurring *de novo*[45],
442 equivalent to drawing sets of size 744 from the universe of possible missense SNVs with probabilities
443 scaled by their individual relative likelihoods.
444
445 **Bioinformatic classification**
446 All *MSH2* missense variants were scored by the following bioinformatic tools: CADD[46], PolyPhen-2[47],
447 REVEL[48], PON-MMR[49], MAPP-MMR[50], and FoldX (scores obtained from ref.[32]). For CADD and REVEL,
448 scores were available only for missense variants reachable by single-base variants (SNVs); for amino-
449 acid substitutions that could arise from more than one SNV, the mean of those SNVs' scores was taken.
450
451 **MSH2 structure and conservation**
452 The MSH2 crystal structure[27] was obtained from the Protein Databank (accession 2O8E) and rendered
453 with PyMOL. Amino acid secondary structure assignment was extracted from pre-calculated DSSP
454 files[51]. Surface accessibility was calculated using an online tool (http://cib.cf.ocha.ac.jp/bitool/ASA/).
455 MSH2 ortholog protein sequences were downloaded from Ensembl (https://useast.ensembl.org/) and
456 aligned with MUSCLE[52].
457
458 **Data availability**
459 Processed count and final LOF score data are available from NCBI Gene Expression Omnibus (GEO
460 accession XXXXX), and corresponding raw sequencing data are posted to SRA (accession XXXXX).
461

References

39. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
40. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
41. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
42. Joung, J. *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat. Protoc.* **12**, 828–863 (2017).
43. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
44. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913–1919 (2015).
45. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* **9**, 3753 (2018).
46. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894

(2019).
47. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
48. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
49. Ali, H., Olatubosun, A. & Vihinen, M. Classification of mismatch repair gene missense variants with PON-MMR. *Hum. Mutat.* **33**, 642–650 (2012).
50. Chao, E. *et al.* Accurate classification of missense variants in MLH1/MSH2 with MAPP-MMR. *Cancer Res.* **68**, LB–14–LB–14 (2008).
51. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Research* vol. 43 D364–D368 (2015).
52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).