OPEN ACCESS

Original research

Prevalence and spectrum of DNA mismatch repair gene variation in the general Chinese population

Li Zhang,[1] Zixin Qin ,[1] Teng Huang,[1] Benjamin Tam,[1] Yongsen Ruan,[2] Maoni Guo,[1] Xiaobing Wu,[1] Jiaheng Li,[1] Bojin Zhao,[1] Jia Sheng Chian,[1] Xiaoyu Wang,[1] Lei Wang,[1] San Ming Wang [1]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2021-107886).

[1]University of Macau, Taipa, Macau, China
[2]Sun Yat-Sen University, Guangzhou, China

Correspondence to
Professor San Ming Wang, University of Macau, Taipa 999078, Macau, China; sanmingwang@um.edu.mo

LZ and ZQ contributed equally.

Received 31 March 2021
Accepted 6 June 2021
Published Online First 25 June 2021

To cite: Zhang L, Qin Z, Huang T, *et al*. *J Med Genet* 2022;**59**:652–661.

ABSTRACT

Background Identifying genetic disease-susceptible individuals through population screening is considered as a promising approach for disease prevention. DNA mismatch repair (MMR) genes including *MLH1*, *MSH2*, *MSH6* and *PMS2* play essential roles in maintaining microsatellite stability through DNA mismatch repair, and pathogenic variation in MMR genes causes microsatellite instability and is the genetic predisposition for cancer as represented by the Lynch syndrome. While the prevalence and spectrum of MMR variation has been extensively studied in cancer, it remains largely elusive in the general population. Lack of the knowledge prevents effective prevention for MMR variation–caused cancer. In the current study, we addressed the issue by using the Chinese population as a model.

Methods We performed extensive data mining to collect MMR variant data from 18 844 ethnic Chinese individuals and comprehensive analyses for the collected MMR variants to determine its prevalence, spectrum and features of the MMR data in the Chinese population.

Results We identified 17 687 distinct MMR variants. We observed substantial differences of MMR variation between the general Chinese population and Chinese patients with cancer, identified highly Chinese-specific MMR variation through comparing MMR data between Chinese and non-Chinese populations, predicted the enrichment of deleterious variants in the unclassified Chinese-specific MMR variants, determined MMR pathogenic prevalence of 0.18% in the general Chinese population and determined that MMR variation in the general Chinese population is evolutionarily neutral.

Conclusion Our study provides a comprehensive view of MMR variation in the general Chinese population, a resource for biological study of human MMR variation, and a reference for MMR-related cancer applications.

INTRODUCTION

Cancer prevention is one of the ultimate goals in medicine.[1] Populational screening for cancer-causing genetic predisposition has been proposed as a promising approach to reach the goal as it allows comprehensive identification of the predisposition carriers for early prevention before cancer development.[2 3] The rapid progress of DNA sequencing technology is making populational screening for cancer prevention closer to reach the reality.[4–7] However, the scientific base of population screening needs to be well established.[8–10] One of the issues relating to the coming paradigm switching is the basic knowledge for cancer predisposition in the general population.

DNA mismatch repair (MMR) genes including *MLH1*, *MSH2*, *MSH6* and *PMS2* play essential roles in maintaining microsatellite stability through repairing DNA mismatch errors.[11] Pathogenic variants in MMR genes damage their normal function, resulting in microsatellite instability and increased risk for developing multiple types of cancer as best represented by the Lynch syndrome (LS), the cancer caused by MMR disorder affecting the gastrointestinal system.[12–14] Since the relationship between MMR variation and cancer was revealed, MMR variation has been extensively characterised in patients with cancer. The resulting MMR variation data are widely used to guide clinical diagnosis and treatment. Most of the MMR variation data currently available were derived from cancer samples, however, the information reflects mainly the status of MMR variation in patients with cancer. Knowledge of MMR variation in the general population without cancer is essential in order to prevent cancer development in the population. Referring the variation data from patients with cancer to the general population can be largely erratic, as MMR variation such as the spectrum, frequency and penetrance could be substantially different between patients with cancer and the general population.[15 16] Furthermore, the issue of ethnic-specific MMR variation further increases the complexity as the current MMR variation data were predominantly derived from the Caucasian populations.[17]

In the present study, we addressed MMR variation in the general population by using the ethnic Chinese population as a model, for which MMR variation has not been systematically characterised so far. We mined MMR variant data from 18 844 ethnic Chinese individuals, determined the prevalence and spectrum, and characterised the features of the variants including ethnic specificity, deleteriousness and evolution selection. Data generated from our study establish a scientific foundation for the use of population screening to prevent MMR-related cancer, a resource to study human MMR variation and a reference for MMR-related clinical applications.

MATERIALS AND METHODS

Samples and variant data collection

Genomic sequence data from 18 844 ethnic Chinese individuals were used in the study. These included the whole genome sequences from 10 588 Chinese

individuals by the ChinaMAP project (https://creativecommons.org/licenses/by/4.0/),[7] the whole genome sequences from 2657 Singapore ethnic Chinese by the Singapore SG10K project,[18] the whole genome sequences from 597 Chinese by the Chinese Academy of Sciences Precision Medicine Initiative project,[19] the whole genome sequences from 90 Chinese by the Han Chinese study,[20] the whole exome sequences from 610 normal Chinese control by the Chinese breast cancer study[21] and the MMR-targeted sequences from 4302 Macau Chinese by our own study (approved by the University of Macau Institutional Review Board, BESRE17-APP014-FHS) (table 1A).

Data analysis

The quality of sequence data was checked by FastQC and duplicates were removed by Trimmatic. Sequencing reads were mapped to human genome reference sequences (hg19) by Burrows-Wheeler Aligner.[22] Sequencing bias was marked by Base Quality Score Recalibration and sorted by Picard. Variants were called using GATK 4.1 following GATK best practices protocol,[23] annotated and classified by ANNOVAR.[24] Mutalyzer name checker and position converter tool were used for cDNA position and coding change with the following reference sequences: Genome: NC_000003.11 (*MLH1*), NC_000002.11 (*MSH2* and *MSH6*), NC_000007.13 (*PMS2*); cDNA: NM_000249.3 (*MLH1*), NM_000251.2 (*MSH2*), NM_000179.2 (*MSH6*), NM_000535.5 (*PMS2*); protein: NP_000240.1 (MLH1), NP_000242.1 (MSH2), NP_000170.1 (MSH6), NP_000526.2 (PMS2). The following databases were used for comparative analyses: gnomAD (http://gnomad.broadinstitute.org/),[25] ExAC (http://exac.broadinstitute.org/)[26] and database of 1000 Genomes Project (https://www.internationalgenome.org)[27] were used for comparing with non-Chinese general population; ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/),[28] InSiGHT (https://www.insight-group.org/variants/databases/),[29] COGR (http://opengenetics.ca/)[30] and UMD (http://www.umd.be)[31] were used for comparing with non-Chinese cancer cohort; dbSNP150 (https://www.ncbi.nlm.nih.gov/snp/)[32] and the databases listed above were used for novelty determination. The variants reported in any of the databases mentioned above were classified as known variants; the variants not reported were classified as novel variants. Power calculation was performed following procedures.[33]

Deleterious impact of variants on protein structural stability

Ramachandran Plot Molecular Dynamic Simulation (RPMDS) method was used following the procedures described in detail.[34] Briefly, the N terminus of MLH1 crystal structure (PDB:4P7A, 2.30 A, residues 1–340) was used as the template. Altered amino acid residues caused by missense variants were incorporated into the wild-type MLH1 structure to generate mutant structures by using the Chimera software.[35] Molecular dynamics simulation composed of five different programs of RMSD (root-mean-square-deviation),[36] RMSF (root-mean-square-fluctuations),[37] Rg (radius of gyration),[38] SASA (solvent accessible surface area)[39] and NH bond (hydrogen bond)[40] were applied to measure the structural changes of equilibrium state, flexibility, shape, hydrogen bond, surface accessibility and structural expansion caused by the variants. Ramachandran plot was used to qualify and quantify the changes. By comparing the data from known pathogenic and benign variants, the variants with deleterious impact on protein structure stability were identified.

Evolutionary analysis

Variants from the Singapore SG10K project were used for the analysis. Ka/Ks ratio was calculated by using dNdSloc model in the dNdScv R package.[41 42] Tajima's *D*,[43] the normalised Fay and Wu's *H* test,[44] the *DH* test and the *E* test[45] were performed by using the Readms package of *DH* software.

RESULTS

MMR variants identified in the general Chinese population

A total of 18 844 ethnic Chinese individuals were included in this study, including 11 885 (61.8%) mainland Chinese, 4302 (23.6%) Macau Chinese and 2657 (14.6%) Singapore Chinese (table 1A, figure 1A, online supplemental table S1). Online supplemental figure S1 outlines the analytical process. From the sequence data, we identified 17 687 distinct variants in MMR genes of *MLH1*, *MSH2*, *MSH6* and *PMS2*, of which 1166 (6.6%) were located in coding regions and 16 521 (93.4%) were located in non-coding regions (table 1, online supplemental tables S2, S3). Of the 1166 coding variants, 210 were in *MLH1*, 248 in *MSH2*, 493 in *MSH6* and 215 in *PMS2* (table 1B); 57.4% were singleton, 27.0% with 2–5 carriers, 6.1% with 6–10 carriers, 5.8% with 11–50 carriers, 0.4% with 51–100 carriers and 3.3% with >100 carriers (figure 1B, online supplemental table S2); of the 13 variation types, missense variant had the highest rate (50.3%), 24.1% remained as unclassified variants (table 1B) and 45.2% as novel variants (table 1C). Ts/Tv ratio was 3.53. G>A had the highest frequency of 18.8% (table 1D). Multiple variation hotspots were identified in MMR genes: 300th to 500th residues in *MLH1*, 400th to 600th residues in *MSH2*, and several hotspots in different functional domains in *MSH6* and *PMS2* (figure 2). Of the 16 521 non-coding variants, 3416 were in *MLH1*, 8629 in *MSH2*, 2052 in *MSH6* and 2424 in *PMS2*; 90.4% (14 941) were located in intron (table 1E), 69.8% (11 528) were not present in dbSNP, 62.9% (10 392) were singleton (online supplemental table S3) and 52.2% were within repetitive sequences (online supplemental table S4). Ts/Tv ratio was 3.38, lower than the 3.53 in the coding variants. C>T had the highest frequency of 20.7% (table 1D).

MMR variation between general population and cancer cohort

We compared the MMR coding variants between the general Chinese population and the Chinese patients with cancer generated by our previous study (online supplemental table S5).[17] We observed low overlapping rates between the two datasets: only 16 variants in *MLH1*, 18 in *MSH2*, 12 in *MSH6* variants and 23 in *PMS2* variants were shared in between (figure 1C, online supplemental table S6). We also searched the 1166 coding variants in ClinVar and InSiGHT databases. Of the 225 coding variants matched in the databases, colorectal cancer had the highest (142 variants) and LS was the second highest (100 variants) (online supplemental table S7), indicating that the coding variants were more oncogenic in colorectal tissue.[17]

MMR variation between Chinese and non-Chinese populations

We compared MMR data between the Chinese and non-Chinese populations. We first compared the data of general populations using the gnomAD, ExAC and 1000 Genome databases after removing the variants derived from ethnic Chinese in these databases. We observed that only 42.1% (491) of the 1166 coding variants and 1.9% (319) of the 16 582 non-coding variants were shared between the general Chinese and non-Chinese populations

Table 1 Summary of mismatch repair (MMR) variants identified in the general Chinese population

A. MMR variants identified

Data source	Cases	Variants identified				Total
		MLH1	*MSH2*	*MSH6*	*PMS2*	
ChinaMAP study[7]	10588	3065	6586	2140	2222	14013
Singapore 10k study[18]	2657	1006	3012	715	855	5588
597 Han individuals[19]	597	18	57	31	59	165
90 Han individuals[20]	90	154	547	89	169	959
Normal control[21]	610	145	701	169	176	1191
Macau individuals	4302	143	239	0	0	382
Total	18844	4531	11142	3144	3481	22298

B. Types of coding variants

Type	*MLH1* (%)	*MSH2* (%)	*MSH6* (%)	*PMS2* (%)	Total (%)
Missense	117 (10.0)	131 (11.2)	214 (18.4)	124 (10.6)	586 (50.3)
Synonymous SNV	44 (3.8)	48 (4.1)	85 (7.3)	54 (4.6)	231 (19.8)
Splice site	3 (0.3)	6 (0.5)	18 (1.5)	11 (0.9)	38 (3.3)
Stopgain	0 (0.0)	1 (0.1)	3 (0.3)	4 (0.3)	8 (0.7)
Frameshift insertion	0 (0.0)	1 (0.1)	3 (0.3)	2 (0.2)	6 (0.5)
Frameshift deletion	0 (0.0)	3 (0.3)	0 (0.0)	1 (0.1)	4 (0.3)
Splice acceptor	0 (0.0)	0 (0.0)	0 (0.0)	4 (0.3)	4 (0.3)
Nonframeshift insertion	0 (0.0)	0 (0.0)	2 (0.2)	0 (0.0)	2 (0.2)
Nonsense	0 (0.0)	2 (0.2)	0 (0.0)	0 (0.0)	2 (0.2)
Splice donor	0 (0.0)	1 (0.1)	0 (0.0)	1 (0.1)	2 (0.2)
Nonframeshift deletion	0 (0.0)	0 (0.0)	1 (0.1)	0 (0.0)	1 (0.1)
Stoploss	0 (0.0)	0 (0.0)	1 (0.1)	0 (0.0)	1 (0.1)
Unclassifiable	46 (3.9)	55 (4.7)	166 (14.2)	14 (1.2)	281 (24.1)
Total	210 (18.0)	248 (21.3)	493 (42.3)	215 (18.4)	1166 (100.0)

C. Novel variants

Item		*MLH1* (%)	*MSH2* (%)	*MSH6* (%)	*PMS2* (%)	Total (%)
Coding variants	Total	210	248	493	215	1166
	Novel	102 (8.7)	90 (7.7)	261 (22.4)	74 (6.3)	527 (45.2)
Non-coding variants	Total	3416	8629	2052	2424	16521
	Novel	2471 (14.5)	5867 (35.5)	1433 (8.7)	1690 (10.2)	11528 (69.8)
Total		2573 (15.1)	5957 (33.7)	1694 (9.6)	1764 (10.0)	12055 (68.2)

D. Ts/Tv ratios

Mutation type	Change	Coding (%)	Non-coding (%)	Total (%)
Transition	C>T	182 (16.9)	3068 (20.7)	3250 (20.4)
	G>A	202 (18.8)	2545 (17.2)	2747 (17.3)
	A>G	172 (16.0)	2171 (14.7)	2343 (14.7)
	T>C	131 (12.2)	1526 (10.3)	1657 (10.4)
Transversion	A>C	38 (3.5)	523 (3.5)	561 (3.5)
	A>T	31 (2.9)	492 (3.3)	523 (3.3)
	C>A	49 (4.6)	615 (4.2)	664 (4.2)
	C>G	92 (8.6)	1165 (7.9)	1257 (7.9)
	G>C	57 (5.3)	818 (5.5)	875 (5.5)
	G>T	61 (5.7)	855 (5.8)	916 (5.8)
	T>A	26 (2.4)	478 (3.2)	504 (3.2)
	T>G	35 (3.3)	561 (3.8)	596 (3.8)
Ts/Tv*		3.53	3.38	3.39

E. Location of non-coding variants

Non-coding variants	*MLH1* (%)	*MSH2* (%)	*MSH6* (%)	*PMS2* (%)	Total (%)
Intron	3351 (20.3)	7328 (44.4)	1904 (11.5)	2424 (14.7)	14941 (90.4)
UTR3	63 (0.4)	867 (5.2)	0 (0.0)	52 (0.3)	982 (5.9)
UTR5	2 (0.0)	0 (0.0)	148 (0.9)	14 (0.1)	316 (1.9)
Downstream	0 (0.0)	282 (1.7)	0 (0.0)	0 (0.0)	282 (1.7)
Total	3416 (20.7)	8629 (52.2)	2052 (12.4)	2424 (14.7)	16521 (100.0)

*Ts/Tv ratio was calculated by 2×Ts/Tv.

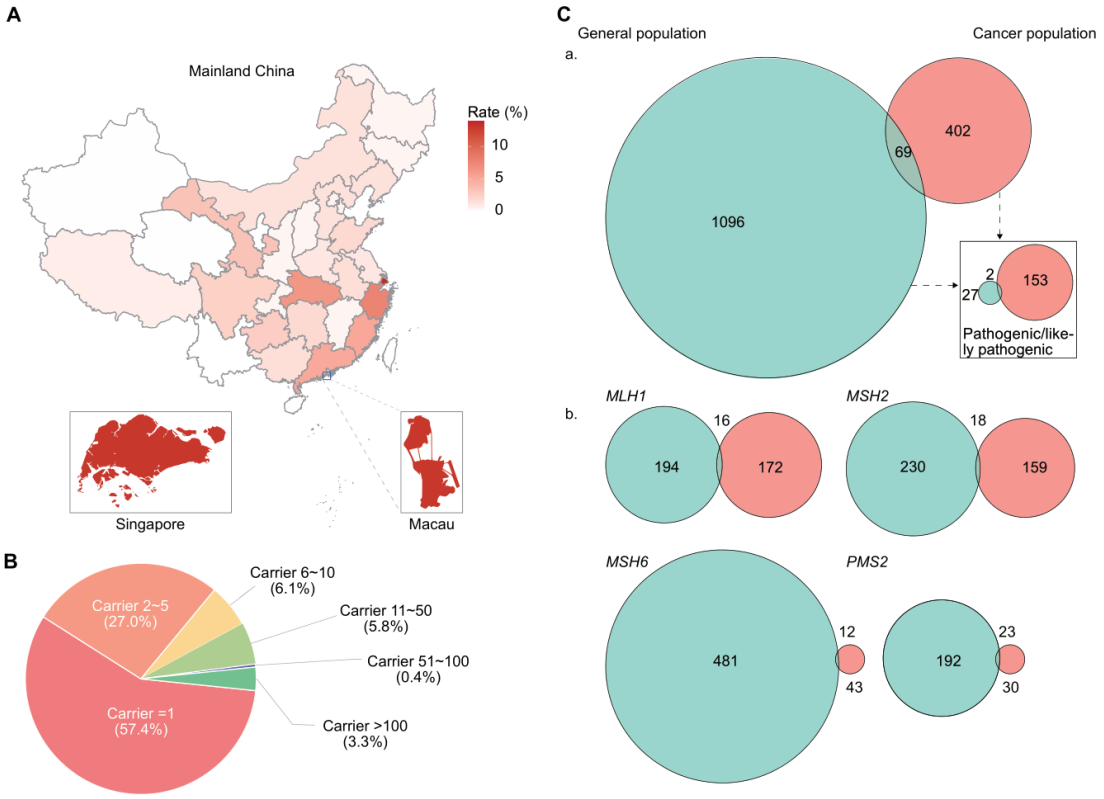


Figure 1 Mismatch repair (MMR) variants in ethnic Chinese population. (A) Sources of Chinese MMR variation data from mainland China, Macau and Singapore. (B) Distribution frequency of MMR variation. (C) Comparison of MMR variation between general Chinese population and patients with cancer. (a) Overall overlapping; (b) individual MMR gene overlapping.

(table 2A,B). Next, we compared the Chinese population with the non-Chinese cancer data from the ClinVar, InSiGHT, COGR and UMD databases after removing the variants derived from ethnic Chinese in these databases. We observed that only 47.4% (553) coding variants and 1.1% (181) non-coding variants in the Chinese population were shared with MMR variants derived from non-Chinese patients with cancer. The shared variants had the highest overlapping rate in ClinVar for both coding (624, 53.5%) and non-coding (176, 1.1%) variants, followed by InSiGHT database for coding (200, 17.2%) and non-coding (119, 0.7%) variants (table 2A,B). In total, 63% (735) of Chinese coding variants and 4.1% (688) of Chinese non-coding variants were shared with non-Chinese populations. Of the 37% (431) of coding variants present in Chinese only, 34.3% (148) were nonsynonymous variants and 2.6% (11) were loss-of-function variants (4 stopgain, 3 frameshift insertion, 2 frameshift deletion and 2 no frameshift insertion) (online supplemental table S2). The differences between the Chinese and non-Chinese population imply that MMR variation is highly ethnic-specific.

Pathogenic variants in the Chinese population

Using ClinVar as the reference, 898 of the 1166 (77.0%) coding variants matched in ClinVar were classified into different clinical classes including 24 (2.0%) as Pathogenic and Likely pathogenic, 507 (43.5%) as Variants of Unknown Significance (VUS), 282 (24.2%) as Benign and Likely benign, 83 (7.3%) as Conflicting classification and 268 (23.0%) as Unclassified (table 3A). There were 33 carriers for the 24 Pathogenic/Likely pathogenic variants in the 18 844 general Chinese individuals included in the study; 26 (78.8%) carriers were in *MSH6* and *PMS2* (table 3B). This resulted in the prevalence of 0.18% for MMR pathogenic variants in the tested general Chinese population, including 0.005% in *MLH1*, 0.032% in *MSH2*, 0.053% in *MSH6* and 0.085% in *PMS2* ($p<0.005$). Power calculation showed that screening 18 844 individuals at a prevalence of 0.18% provides a 97.8% probability of detecting all pathogenic variants in the studied population. The 0.18% prevalence implies the presence of 2.52 million MMR pathogenic variant carriers estimated in the 1.4 billion Chinese population. We performed the same analysis for the MMR data in the gnomAD database, and observed a 0.11% prevalence of MMR pathogenic variants in the general non-Chinese population, composed of 0.020% in *MLH1*, 0.015% in *MSH2*, 0.024% in *MSH6* and 0.052% in *PMS2*. Therefore, the general Chinese population had a higher prevalence than general non-Chinese populations in *MSH2*, *MSH6* and *PMS2* but lower prevalence in *MLH1*. In the Chinese population, the pathogenic variants c.3226C>T (*MSH6*), c.825A>G (*PMS2*) and c.82G>A (*MSH2*) had 5, 4 and 3 carriers, respectively, suggesting that they may be the potential founder mutations. We further collected cancer distribution information from the ClinVar database for the identified Pathogenic and Likely pathogenic variants. Similar to the distribution of MMR non-pathogenic variants, the 24 Pathogenic/Likely pathogenic variants had the highest frequency in colorectal cancer (table 3C).

Ethnic-specific pathogenic variants

Of the 1166 coding variants, 507 (43.5%) were classified as VUS and 268 (23.0%) remained as Unclassified (table 3A). The presence of 66.5% as functional unknown variants raises the question if these unclassified variants had any biological significance. We tested this possibility by measuring the impact of the unclassified variants on MLH1 protein structure stability using

Figure 2 Location and frequency of coding variants in the coding region of each mismatch repair gene. The map was constructed by using the ProteinPaint program.[56]

the structure-based RPMDS method.[34] By referring to the cut-off values from wild-type, known benign and known pathogenic variants to differentiate deleterious and non-deleterious variants, we tested the Unclassified variants (online supplemental table S8). We selected the Unclassified variants for the test under the following conditions: (1) the variants must be within the known MLH1 structure as the RPMDS method relies on the known protein structure in its analysis. Of the 756 amino acid residues in MLH1, the known structure (PDB:4P7A) covers 340 aa (1–340). (2) The variant must be missense as the RPMDS is designed for missense variant analysis. Under these conditions, we identified 16 Unclassified *MLH1* variants for the analysis. We identified six variants as deleterious that significantly disturbed MLH1 structure (online supplemental table S9). For example, K241T had deleterious effects as it destabilised the two alpha helixes formed between residues 234 and 340 in MLH1 (online supplemental figure S2). The results indicate that the Unclassified variants were enriched with Chinese-specific deleterious variants.

Evolutionary analysis of Chinese MMR variation

We performed evolutionary analysis for MMR variation in the Chinese population. We calculated Ka/Ks ratio for each MMR gene with 0.83 in *MLH1*, 0.92 in *MSH2*, 1.07 in *MSH6* and 1.17 in *PMS2* (figure 3A). We observed that A/S ratios were all close to 2.5 (2.08 in *MLH1*, 2.29 in *MSH2*, 2.80 in *MSH6* and 2.73 in *PMS2*). We validated the results by using multiple tests of Tajima

Table 2 Comparison of mismatch repair variants between Chinese and non-Chinese populations

A. Coding variants										
	Chinese									
	MLH1 (210)		*MSH2* (248)		*MSH6* (493)		*PMS2* (215)		Total (1166)	
Non-Chinese*	Total†	Reported (%)	Total	Reported (%)	Total	Reported (%)	Total	Reported (%)	Total	Reported (%)
General population										
gnomAD	1286	92 (43.8)	1685	109 (44.0)	2044	185 (37.5)	1531	121 (56.3)	6546	507 (43.5)
ExAC	949	72 (34.3)	1007	87 (35.1)	1378	136 (27.6)	916	98 (45.6)	4250	393 (33.7)
1000 genomes	720	1 (0.5)	1852	10 (12.9)	429	11 (2.2)	193	6 (2.8)	3194	28 (2.4)
Subtotal†	2258	84 (40.0)	2549	109 (44.0)	2646	181 (36.7)	1732	117 (54.4)	6927	**491 (42.1)**
Caner cohort										
ClinVar	2382	114 (54.3)	2961	142 (57.3)	3465	220 (44.6)	1791	148 (68.8)	10599	624 (53.5)
InSiGHT	1375	51 (24.3)	1352	42 (16.9)	506	55 (11.2)	414	52 (24.2)	3647	200 (17.2)
COGR	120	11 (5.2)	122	13 (5.2)	109	15 (3.0)	44	7 (3.3)	395	46 (3.9)
UMD	97	0 (0.0)	246	18 (7.3)	51	2 (0.4)	0	0 (0.0)	394	20 (1.7)
Subtotal†	3644	104 (49.5)	4312	121 (48.8)	4061	200 (40.6)	2165	128 (59.5)	10538	**553 (47.4)**
Total‡	5852	129 (61.4)	7383	168 (67.7)	6184	271 (55.0)	3983	167 (77.7)	23402	**735 (63.0)**
B. Non-coding variants										
	Chinese									
	MLH1 (3416)		*MSH2* (8629)		*MSH6* (2052)		*PMS2* (2424)		Total (16 521)	
Non-Chinese	Total	Reported (%)	Total	Reported (%)	Total	Reported (%)	Total	Reported (%)	Total	Reported (%)
General population										
gnomAD	1286	65 (1.9)	1685	81 (09)	2044	36 (1.7)	1531	58 (2.4)	6546	240 (1.4)
ExAC	949	46 (1.3)	1007	51 (0.6)	1378	26 (1.3)	916	5 (0.2)	4250	128 (0.8)
1000 genomes	720	18 (0.5)	1852	282 (3.3)	429	43 (2.1)	193	28 (1.2)	3194	371 (2.2)
Subtotal†	2258	81 (2.4)	2549	80 (0.9)	2646	76 (3.7)	1732	82 (3.4)	6927	**319 (1.9)**
Cancer cohort										
ClinVar	2382	55 (1.6)	2961	63 (0.7)	3465	39 (1.9)	1791	19 (0.8)	10599	176 (1.1)
InSiGHT	1375	33 (1.0)	1352	45 (0.5)	506	19 (0.9)	414	22 (0.9)	3647	119 (0.7)
COGR	120	4 (0.1)	122	7 (0.1)	109	4 (0.2)	44	3 (0.1)	395	18 (0.1)
UMD	97	0 (0.0)	246	0 (0.0)	51	0 (0.0)	0	0 (0.0)	394	0 (0.0)
Subtotal†	3644	51 (1.5)	4312	67 (0.8)	4061	38 (1.8)	2165	25 (1.0)	10538	**181 (1.1)**
Total‡	5852	95 (2.8)	7383	397 (4.6)	6184	99 (4.8)	3983	97 (4.0)	23402	**688 (4.1)**

The bold values refer to the sum in each and combined populations.
*254 MMR variants derived from ethnic Chinese in these databases were excluded for the comparison.
†Distinct variants in non-Chinese databases after combination.
‡Distinct variants after combination of all variants in each column.

D, Fay & Wu *H*, *DH* and Zeng *E* (figure 3B), and concluded that MMR genes in the Chinese population were evolutionarily neutral.

DISCUSSION

Several conclusions can be made from our current study:

1. Modest prevalence of MMR variation in the general population. Of the 1166 MMR coding variants identified in the general Chinese population, we identified 24 Pathogenic/Likely pathogenic variants with 33 carriers. Based on the data, we determined the 0.18% prevalence of MMR pathogenic variation in the 18 844 ethnic Chinese individuals, highlighting the presence of 2.52 million MMR pathogenic variant carriers in the 1.4 billion Chinese population. It is important to note that a group of autosomal dominant cancer predisposition mutations has much higher prevalence in the general population than these in many non-cancer hereditary diseases, such as spinocerebellar ataxia (SCA), in which there are only a few carriers per 100 000 individuals.[46] For example, the prevalence of pathogenic mutation in *BRCA1*/*BRCA2* (*BRCA*) is 0.26% in Japanese (1 in 384),[47] 0.38% in Chinese (1 in 265),[48] 0.38% in Mexicans (1 in 265),[49] 0.39% in Malaysians (1 in 556),[50] 0.53% in US population (1 in 189)[51] and 2.17% in the Ashkenazi Jewish (1 in 46).[52] The 0.18% prevalence in the Chinese population is the combination of four MMR genes of *MLH1*, *MSH2*, *MSH6* and *PMS2*. With 1 in every 556 Chinese individuals an MMR pathogenic variant carrier, it indicates the serious threat of MMR-related cancer risk for the public health and the importance of preventing MMR-related cancer in the general Chinese population. However, the 0.18% prevalence is lower than the 0.38% prevalence for *BRCA* pathogenic variation in the general Chinese population.[48] Although MMR and *BRCA* are the two groups of cancer predisposition genes with the most significant clinical value over other cancer predisposition genes,[53] priority will be given to *BRCA* screening first if only one choice can be made when planning a population screening for cancer prevention. Alternatively, MMR and *BRCA* can be combined as one panel for the screening with the expected outcome of identifying twice more of *BRCA* pathogenic variant carriers over MMR carriers.
2. Different spectra of MMR variation between the general population and patients with cancer. This is reflected by the low overlapping of MMR variants between the two cohorts,

Table 3 Clinical classification of coding variants

A. Classification

Class	Genes (%)				Total
	MLH1	*MSH2*	*MSH6*	*PMS2*	
Pathogenic	0 (0.0)	3 (1.2)	2 (0.4)	8 (3.7)	13 (1.1)
Likely pathogenic	1 (0.5)	1 (0.4)	4 (0.8)	5 (2.3)	11 (0.9)
VUS	113 (53.8)	118 (47.6)	193 (39.1)	83 (38.6)	507 (43.5)
Likely benign	48 (22.9)	50 (20.2)	89 (18.1)	53 (24.7)	240 (20.6)
Benign	7 (3.3)	11 (4.4)	8 (1.6)	17 (7.9)	42 (3.6)
Conflicting	8 (3.8)	22 (8.9)	24 (4.9)	31 (14.4)	85 (7.3)
Unclassified	33 (15.7)	43 (17.3)	173 (35.1)	19 (8.8)	268 (23.0)
Total	210 (100.0)	248 (100.0)	493 (100.0)	215 (100.0)	1166 (100.0)

B. List of Pathogenic/Likely pathogenic variants

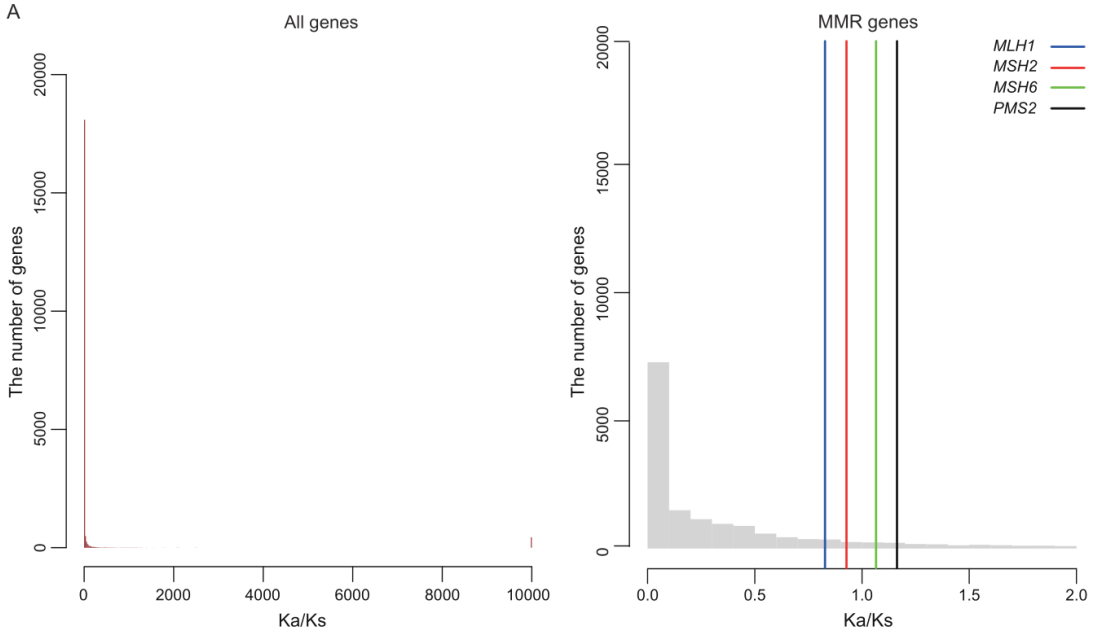
Base change	Amino acid change*	Variation type	Variant class	Carrier	Carrier rate (%)†
MLH1					
c.1984A>G	p.Thr662Ala	Nonsynonymous	Likely pathogenic	1	0.005
MSH2					0.032
c.28C>T	p.Gln10Ter	Nonsense	Pathogenic	1	
c.82G>A	p.Glu28Lys	Nonsynonymous	Likely pathogenic	3	
c.645+1G>A	–	Splice donor	Pathogenic	1	
c.1457_1460delATGA	p.Asn486fs	Frameshift deletion	Pathogenic	1	
MSH6					0.053
c.1807A>T	p.Lys603Ter	Stopgain	Likely pathogenic	1	
c.1838T>A	p.Leu613Ter	Stopgain	Pathogenic	1	
c.2095G>T	p.Glu699Ter	Stopgain	Likely pathogenic	1	
c.2906A>G	p.Tyr969Cys	Missense	Likely pathogenic	1	
c.3226C>T	p.Arg1076Cys	Missense	Likely pathogenic	5	
c.3253dupC	p.Thr1085fs	Frameshift insertion	Pathogenic	1	
PMS2					0.085
c.1A>G	p.Met1Val	Missense	Likely pathogenic	1	
c.2T>A	p.Met1Lys	Missense	Pathogenic	1	
c.164–1G>C	–	Splice acceptor	Likely pathogenic	1	
c.673G>T	p.Glu225Ter	Stopgain	Pathogenic	1	
c.825A>G	p.Gln275=	Synonymous	Likely pathogenic	4	
c.903+2T>C	–	Splice donor	Likely pathogenic	1	
c.993C>A	p.Cys331Ter	Stopgain	Pathogenic	1	
c.1240dupT	p.Asp414fs	Frameshift insertion	Pathogenic	1	
c.1687C>T	p.Arg563Ter	Stopgain	Pathogenic	1	
c.1731dupT	p.Arg578fs	Frameshift insertion	Pathogenic	1	
c.1864_1865delAT	p.Met622Glufs*5	Frameshift deletion	Pathogenic	1	
c.1959T>A	p.Cys653Ter	Stopgain	Pathogenic	1	
c.2276-2A>C	–	Splice acceptor	Likely pathogenic	1	

C. Distribution of Pathogenic/Likely pathogenic variants

Item	Genes (%)				Total (%)
	MLH1	*MSH2*	*MSH6*	*PMS2*	
Colon and rectal cancer	0 (0.0)	1 (7.1)	3 (21.4)	7 (50.0)	11 (78.6)
LS/HNPCC	0 (0.0)	2 (14.3)	2 (14.3)	1 (7.1)	5 (35.7)
Breast cancer	0 (0.0)	0 (0.0)	0 (0.0)	3 (21.4)	3 (21.4)
Endometrial cancer	0 (0.0)	0 (0.0)	2 (14.3)	2 (14.3)	4 (28.6)
Suspected LS	0 (0.0)	1 (7.1)	0 (0.0)	2 (14.3)	3 (21.4)
Ovarian cancer	0 (0.0)	0 (0.0)	1 (7.1)	1 (7.1)	2 (14.3)
Gastric cancer	0 (0.0)	0 (0.0)	1 (7.1)	0 (0.0)	1 (7.1)
Glioblastoma	0 (0.0)	0 (0.0)	1 (7.1)	0 (0.0)	1 (7.1)
Brain cancer	0 (0.0)	0 (0.0)	0 (0.0)	1 (7.1)	1 (7.1)
Duodenal cancer	0 (0.0)	0 (0.0)	0 (0.0)	1 (7.1)	1 (7.1)
Total	0 (0.0)	3 (21.4)	3 (21.4)	8 (57.1)	14 (100.0)

*Variants marked '–' were predicted by InterVar.
†
HNPCC, hereditary nonpolyposis colorectal cancer; LS, Lynch syndrome; VUS, variants of unknown significance.

A



B

Item	*MLH1*	*MSH2*	*MSH6*	*PMS2*	mergeGene
A/S* [44,45]	25/12	39/17	56/20	41/15	161/64
Singleton	24	28	44	31	127
d (pairwise distance) [46]	0.277	0.376	1.113	1.988	3.753
n (sample size)	2657	2657	2657	2657	2657
Tajima D (p-value) [46]	-2.271 (0)	-2.399 (0)	-2.288 (0)	-1.780 (0.003)	-2.366 (0)
Fay&Wu H (p-value) [46]	0.086 (0.359)	0.083 (0.354)	0.089 (0.359)	-0.923 (0.119)	-0.180 (0.265)
p-value of DH test	0.1482	0.1573	0.1571	0.0283	0.0993
Zeng E (p-value) [46]	-1.974 (0)	-2.109 (0)	-2.038 (0)	-0.845 (0.220)	-1.945 (1.0E-4)

*A: the number of asynonymous variants; S: the number of synonymous variants

Figure 3 Evolutionary analysis of mismatch repair (MMR) gene variation in the general Chinese population. (A) Ka/Ks ratio of MMR variation in the Chinese population. Most of the genes in the genomes were neutral (left part, Ka/Ks ratio<1), of which *MLH1* had 0.83, *MSH2* had 0.92, *MSH6* had 1.07 and *PMS2* had 1.17 (right part). (B) MMR neutral tests. The A/S ratios for each MMR gene variation were tested by methods of Tajima *D*, Fay & Wu *H*, DH and Zeng *E* tests, confirming that each MMR gene variation was evolutionarily neutral.

and most of the pathogenic variants in the general population were in *MSH6* and *PMS2* but in patients with cancer were in *MLH1* and *MSH2*. This implies that the MMR variation data derived from patients with cancer cannot be directly applied as the standard reference to judge MMR variation in the general population, as the former reflects the MMR pathogenic variation enriched in the cancer cohort whereas the latter reflects the genetic variation distributed in the general population.

3. Highly ethnic-specific MMR variation in the general population. This is reflected by the presence of 68.2% of novel MMR variants in the Chinese population (table 1C). A similar situation in *BRCA* variation was also present in the general Chinese population[48] and may also exist in other ethnic populations. This feature reminds the limitation of the MMR data currently available as they were mostly derived from the Caucasian populations, and highlights the need to collect variation information in cancer predisposition genes from different ethnic populations.
4. The challenge of classifying ethnic-specific MMR variants. Over 66% of Chinese MMR variants remain unclassified. It is difficult to classify the ethnic-specific variants and identify the potential ethnic-specific pathogenic variants due to the lack of clinical evidence, resources, expertise and references available in existing MMR databases. Substantial efforts need to be made to improve the situation.
5. Human MMR system is not under obvious evolution selection. Our study demonstrated that MMR genes in the Chinese population were neutral without obvious positive or negative selection. This is in contrast to the situation in human *BRCA1*, in which strong positive selection is present.[54] It is interesting to indicate that evolution selection can act differentially in different DNA damage repair genes/pathways for better fitness.
6. Potential value of MMR non-coding variants. Of the 17 687 variants identified in MMR genes, 16 521 (93.4%) were in non-coding regions mostly in introns, and 12 055 (68.2%) were not present in dbSNP. Their rich presence highlights the highly variable nature in MMR non-coding regions and further exploring their potential clinical relevance is warranted.

Certain limitations are present in our study, such as lack of sex and age information in the variation data as these were not available in the original data sources. For the variants detected only in single individuals, possibility exists that certain "singleton" MMR variants could be generated by sequencing errors instead of true genetic variants. In addition, the actual prevalence of MMR pathogenic variants could be higher than observed as genomic data used in our study were mainly collected by short sequences from the next-generation sequencing platform, which is sensitive in detecting single-base and small indel variants but lacks power to detect large structural variations. Further functional test of the rich MMR variants will help to identify the driver variants contributing to the oncogenic process.[55]

CONCLUSION

In summary, our study provides a populational view for MMR variation in an ethnic human population, a scientific basis in planning population screening for MMR-related cancer prevention, and a reference resource for biological study of human MMR variation and MMR-related clinical applications.

Acknowledgements We thank the late Dr Henry Lynch for his inspiration and encouragement in studying MMR-related cancer in the Chinese population. We thank the 'SG10K_Pilot Investigators' for providing the SG10K_Pilot data (EGAD00001005337). The data from the 'SG10K_Pilot Study' reported here were obtained from EGA. We are also thankful for the Information and Communication Technology Office (ICTO), University of Macau for providing the High-Performance Computing Cluster (HPCC) resource and facilities for the study.

Contributors LZ: method development, analysis, data interpretation, draft manuscript; LZ, ZQ, HT, BT, XWu, XWang, LW, YR, GM, JL, BZ, JSC: data acquisition, data interpretation; SMW: conception, design, analysis, data interpretation, manuscript revision and funding.

Funding This work was funded by Macau Science and Technology Development Fund (085/2017/A2, 0077/2019/AMJ), University of Macau (SRG2017-00097-FHS, MYRG2019-00018-FHS), Faculty of Health Sciences, University of Macau (FHSIG/SW/0007/2020P, Startup fund) (SMW).

Disclaimer This manuscript was not prepared in collaboration with the 'SG10K_Pilot Study' and does not necessarily reflect the opinions or views of the 'SG10K_Pilot Study'.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The MMR-targeted sequences in Macau Chinese was approved by University of Macau Institutional Review Board (BESRE17-APP014-FHS).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

ORCID iDs
Zixin Qin http://orcid.org/0000-0002-4435-8106
San Ming Wang http://orcid.org/0000-0002-2172-1320

REFERENCES

1 Murray MF, Evans JP, Angrist M, Uhlmann WR, Lochner Doyle D, Fullerton SM, Ganiats TG, Hagenkord J, Imhof S, Rim SH, Ortmann L, Aziz N, Dotson WD, Matloff E, Young K, Kaphingst K, Bradbury A, Scott J, Wang C, Zauber A, Levine M, Korf B, Leonard DG, Wicklund C, Isham G, Khoury MJ. A proposed approach for implementing genomics-based screening programs for healthy adults. *NAM Perspectives* 2018.
2 King M-C, Levy-Lahad E, Lahad A. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. *JAMA* 2014;312:1091–2.
3 Grzymski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, Rowan C, Slotnick N, Dabe S, Metcalf WJ, Lipp B, Reed H, Sharma L, Levin E, Kao J, Rashkin M, Bowes J, Dunaway K, Slonim A, Washington N, Ferber M, Bolze A, Lu JT. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat Med* 2020;26:1235–9.
4 Zlotogora J. Genetics and genomic medicine in Israel. *Mol Genet Genomic Med* 2014;2:85–94.
5 Qin Z, Kuok CN, Dong H, Jiang L, Zhang L, Guo M, Leong HK, Wang L, Meng G, Wang SM. Can population *BRCA* screening be applied in non-Ashkenazi Jewish populations? Experience in Macau population. *J Med Genet* 2021;58:587–91.
6 Manchanda R, Burnell M, Gaba F, Desai R, Wardle J, Gessler S, Side L, Sanderson S, Loggenberg K, Brady AF, Dorkins H, Wallis Y, Chapman C, Jacobs C, Legood R, Beller U, Tomlinson I, Menon U, Jacobs I. Randomised trial of population-based BRCA testing in Ashkenazi Jews: long-term outcomes. *BJOG* 2020;127:364–75.
7 Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R, Ye Z, Shi L, Tang X, Yan L, Gao Z, Chen G, Zhang Y, Chen L, Ning G, Bi Y, Wang W, Consortium C, ChinaMAP Consortium. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res* 2020;30:717–31.
8 Surbone A. Social and ethical implications of BRCA testing. *Ann Oncol* 2011;22 Suppl 1:i60–6.
9 Sie AS, Spruijt L, van Zelst-Stams WAG, Mensenkamp AR, Ligtenberg MJL, Brunner HG, Prins JB, Hoogerbrugge N. High satisfaction and low distress in breast cancer patients one year after BRCA-mutation testing without prior face-to-face genetic counseling. *J Genet Couns* 2016;25:504–14.
10 Manchanda R, Patel S, Gordeev VS, Antoniou AC, Smith S, Lee A, Hopper JL, MacInnis RJ, Turnbull C, Ramus SJ, Gayther SA, Pharoah PDP, Menon U, Jacobs I, Legood R. Cost-effectiveness of population-based BRCA1, BRCA2, RAD51C, RAD51D, BRIP1, PALB2 mutation testing in unselected general population women. *J Natl Cancer Inst* 2018;110:714–25.
11 Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology* 2010;138:2073–87.
12 Lynch HT, Lynch JF. Hereditary nonpolyposis colorectal cancer (Lynch syndromes I and II): a common genotype linked to oncogenes? *Med Hypotheses* 1985;18:19–28.
13 Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on HNPCC. *Anticancer Res* 1994;14:1661–4.
14 Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. *Nat Rev Cancer* 2015;15:181–94.
15 Bhaskaran SP, Chandratre K, Gupta H, Zhang L, Wang X, Cui J, Kim YC, Sinha S, Jiang L, Lu B, Wu X, Qin Z, Huang T, Wang SM. Germline variation in BRCA1/2 is highly ethnic-specific: evidence from over 30,000 Chinese hereditary breast and ovarian cancer patients. *Int J Cancer* 2019;145:962–73.
16 Heredia-Genestar JM, Marquès-Bonet T, Juan D, Navarro A. Extreme differences between human germline and tumor mutation densities are driven by ancestral human-specific deviations. *Nat Commun* 2020;11:1–9.
17 Zhang L, Bhaskaran SP, Huang T, Dong H, Chandratre K, Wu X, Qin Z, Wang X, Cao W, Chen T, Lynch H, Wang SM. Variants of DNA mismatch repair genes derived from 33,998 Chinese individuals with and without cancer reveal their highly ethnic-specific nature. *Eur J Cancer* 2020;125:12–21.
18 Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, Soon WWJ, Bertin N, Lin CB, Khor CC, DeGiorgio M, Cheng S, Bao L, Karnani N, Hwang WYK, Davila S, Tan P, Shabbir A, Moh A, Tan E-K, Foo JN, Goh LL, Leong KP, Foo RSY, Lam CSP, Richards AM, Cheng C-Y, Aung T, Wong TY, Ng HH, Liu J, Wang C, RSY F, CSP L, HH N, SG10K Consortium. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* 2019;179:736–49.
19 Du Z, Ma L, Qu H, Chen W, Zhang B, Lu X, Zhai W, Sheng X, Sun Y, Li W, Lei M, Qi Q, Yuan N, Shi S, Zeng J, Wang J, Yang Y, Liu Q, Hong Y, Dong L, Zhang Z, Zou D, Wang Y, Song S, Liu F, Fang X, Chen H, Liu X, Xiao J, Zeng C. Whole genome analyses of Chinese population and de novo assembly of a northern Han genome. *Genomics Proteomics Bioinformatics* 2019;17:229–47.
20 Lan T, Lin H, Zhu W, Laurent TCAM, Yang M, Liu X, Wang J, Wang J, Yang H, Xu X, Guo X. Deep whole-genome sequencing of 90 Han Chinese genomes. *Gigascience* 2017;6:1–7.
21 Zeng C, Guo X, Wen W, Shi J, Long J, Cai Q, Shu X-O, Xiang Y, Zheng W. Evaluation of pathogenetic mutations in breast cancer predisposition genes in population-based studies conducted among Chinese women. *Breast Cancer Res Treat* 2020;181:465–73.
22 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357.
23 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
24 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
25 Thompson BA, Spurdle AB, Plazzer J-P, Greenblatt MS, Akagi K, Al-Mulla F, Bapat B, Bernstein I, Capellá G, den Dunnen JT, du Sart D, Fabre A, Farrell MP, Farrington SM, Frayling IM, Frebourg T, Goldgar DE, Heinen CD, Holinski-Feder E, Kohonen-Corish M, Robinson KL, Leung SY, Martins A, Moller P, Morak M, Nystrom M, Peltomaki P, Pineda M, Qi M, Ramesar R, Rasmussen LJ, Royer-Pokora B, Scott RJ, Sijmons R, Tavtigian SV, Tops CM, Weber T, Wijnen J, Woods MO, Macrae F, Genuardi M. 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* 2014;46:107–15.
26 Béroud C, Collod-Béroud G, Boileau C, Soussi T, Junien C. UMD (universal mutation database): a generic software to build and analyze locus-specific databases. *Hum Mutat* 2000;15:86–94.
27 Lerner-Ellis J, Wang M, White S, Lebo MS, Canadian Open Genetics Repository Group. Canadian Open Genetics Repository Group. Canadian Open Genetics

Repository (COGR): a unified clinical genomics database as a community resource for standardising and sharing genetic interpretations. *J Med Genet* 2015;52:438–45.
28 Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46:D1062–7.
29 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Consortium EA, Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–91.
30 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG, Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43.
31 Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008;36:D13–21.
32 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature* 2015;526:68–74.
33 Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J* 2003;20:453.
34 Tam B, Sinha S, Wang SM. Combining Ramachandran plot and molecular dynamics simulation for structural-based variant classification: using *TP53* variants as model. *Comput Struct Biotechnol J* 2020;18:4033–9.
35 Meng EC, Pettersen EF, Couch GS, Huang CC, Ferrin TE. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* 2006;7:1–10.
36 Dong Y-W, Liao M-L, Meng X-L, Somero GN. Structural flexibility and protein adaptation to temperature: molecular dynamics analysis of malate dehydrogenases of marine molluscs. *Proc Natl Acad Sci U S A* 2018;115:1274.
37 Benson NC, Daggett V. A comparison of multiscale methods for the analysis of molecular dynamics simulations. *J Phys Chem B* 2012;116:8722–31.
38 Daidone I, Amadei A, Roccatano D, Nola AD. Molecular dynamics simulation of protein folding by essential dynamics sampling: folding landscape of horse heart cytochrome c. *Biophys J* 2003;85:2865–71.
39 Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external Bath. *J Chem Phys* 1984;81:3684–90.
40 Sheu S-Y, Yang D-Y, Selzle HL, Schlag EW. Energetics of hydrogen bonds in peptides. *Proc Natl Acad Sci U S A* 2003;100:12683–7.
41 Nekrutenko A, Makova KD, Li W-H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* 2002;12:198–202.
42 Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. Universal patterns of selection in cancer and somatic tissues. *Cell* 2017;171:1029–41.
43 Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–95.
44 Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. *Genetics* 2000;155:1405–13.
45 Zeng K, Fu Y-X, Shi S, Wu C-I, YX F, CI W. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 2006;174:1431–9.
46 Sullivan R, Yau WY, O'Connor E, Houlden H. Spinocerebellar ataxia: an update. *J Neurol* 2019;266:533–44.
47 Momozawa Y, Iwasaki Y, Parsons MT, Kamatani Y, Takahashi A, Tamura C, Katagiri T, Yoshida T, Nakamura S, Sugano K, Miki Y, Hirata M, Matsuda K, Spurdle AB, Kubo M. Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. *Nat Commun* 2018;9:4083.
48 Dong H, Chandratre K, Qin Y, Zhang J, Tian X, Rong C, Wang N, Guo M, Zhao G, Wang SM. Prevalence of *BRCA1*/*BRCA2* pathogenic variation in Chinese Han population. *J Med Genet* 2021;58:565–9.
49 Fernández-Lopez JC, Romero-Córdoba S, Rebollar-Vega R, Alfaro-Ruiz LA, Jiménez-Morales S, Beltrán-Anaya F, Arellano-Llamas R, Cedro-Tanda A, Rios-Romero M, Ramirez-Florencio M, Bautista-Piña V, Dominguez-Reyes C, Villegas-Carlos F, Tenorio-Torres A, Hidalgo-Miranda A. Population and breast cancer patients' analysis reveals the diversity of genomic variation of the BRCA genes in the Mexican population. *Hum Genomics* 2019;13:3.
50 Wen WX, Allen J, Lai KN, Mariapun S, Hasan SN, Ng PS, Lee DS-C, Lee SY, Yoon S-Y, Lim J, Lau SY, Decker B, Pooley K, Dorling L, Luccarini C, Baynes C, Conroy DM, Harrington P, Simard J, Yip CH, Mohd Taib NA, Ho WK, Antoniou AC, Dunning AM, Easton DF, Teo SH. Inherited mutations in *BRCA1* and *BRCA2* in an unselected multiethnic cohort of Asian patients with breast cancer and healthy controls from Malaysia. *J Med Genet* 2018;55:97–103.
51 Manickam K, Buchanan AH, Schwartz MLB, Hallquist MLG, Williams JL, Rahm AK, Rocha H, Savatt JM, Evans AE, Butry LM, Lazzeri AL, Lindbuchler D'Andra M, Flansburg CN, Leeming R, Vogel VG, Lebo MS, Mason-Suares HM, Hoskinson DC, Abul-Husn NS, Dewey FE, Overton JD, Reid JG, Baras A, Willard HF, McCormick CZ, Krishnamurthy SB, Hartzel DN, Kost KA, Lavage DR, Sturm AC, Frisbie LR, Person TN, Metpally RP, Giovanni MA, Lowry LE, Leader JB, Ritchie MD, Carey DJ, Justice AE, Kirchner HL, Faucett WA, Williams MS, Ledbetter DH, Murray MF. Exome sequencing-based screening for BRCA1/2 expected pathogenic variants among adult biobank participants. *JAMA Netw Open* 2018;1:e182140.
52 Gabai-Kapara E, Lahad A, Kaufman B, Friedman E, Segev S, Renbaum P, Beeri R, Gal M, Grinshpun-Cohen J, Djemal K, Mandell JB, Lee MK, Beller U, Catane R, King M-C, Levy-Lahad E. Population-based screening for breast and ovarian cancer risk due to BRCA1 and BRCA2. *Proc Natl Acad Sci U S A* 2014;111:14205–10.
53 Turnbull C, Sud A, Houlston RS. Cancer genetics, precision prevention and a call to action. *Nat Genet* 2018;50:1212–8.
54 Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, Hopper JL, Venter DJ. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat Genet* 2000;25:410–3.
55 Chung J, Maruvka YE, Sudhaman S, Kelly J, Haradhvala NJ, Bianchi V, Edwards M, Forster VJ, Nunes NM, Galati MA, Komosa M, Deshmukh S, Cabric V, Davidson S, Zatzman M, Light N, Hayes R, Brunga L, Anderson ND, Ho B, Hodel KP, Siddaway R, Morrissy AS, Bowers DC, Larouche V, Bronsema A, Osborn M, Cole KA, Opocher E, Mason G, Thomas GA, George B, Ziegler DS, Lindhorst S, Vanan M, Yalon-Oren M, Reddy AT, Massimino M, Tomboc P, Van Damme A, Lossos A, Durno C, Aronson M, Morgenstern DA, Bouffet E, Huang A, Taylor MD, Villani A, Malkin D, Hawkins CE, Pursell ZF, Shlien A, Kunkel TA, Getz G, Tabori U. DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. *Cancer Discov* 2021;11:1176–91.
56 Zhou X, Edmonson MN, Wilkinson MR, Patel A, Wu G, Liu Y, Li Y, Zhang Z, Rusch MC, Parker M, Becksfort J, Downing JR, Zhang J. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* 2016;48:4–6.