SUPPLEMENTARY METHODS

Germline Whole Genome Sequencing

Germline WGS was performed on peripheral blood-derived DNA from the 14 individuals with SLS and two *MSH2* pathogenic variant positive controls in this study using the Illumina TruSeq DNA library preparation guide, yielding an average insert size of 300 - 400 bp. The libraries were sequenced by Macrogen (South Korea) on an Illumina HiSeq 2000 Sequencer (USA) to an average 30x coverage.

Somatic MMR gene mutation testing and Whole Genome Sequencing

Formalin-Fixed Paraffin Embedded (FFPE) tissue was macrodissected to enrich for tumour cells. DNA extraction was performed using QIAamp DNA FFPE Tissue kit and standard protocols (Qiagen, Germany).

Briefly, FFPE-derived DNA was determined to be sufficiently degraded that library prep could proceed without further fragmentation. End repair and adapter ligation was performed with NEBNext Ultra II kits (New England Biolabs, USA) according to the manufacturer's protocols. Individually barcoded libraries were pooled in equimolar amounts and 1.1 nM of combined libraries were sequenced using a NovaSeq6000 S4-flowcell at 2 x 150bp reads with an average depth of 40x in the capture target regions.

AmpliSeq Targeted Tumour Sequencing

We designed a custom Ampliseq™ panel to screen the MMR genes from FFPE tumour DNA samples through generation of a PCR-based library (125-175bp) and sequencing on the IonTorrent. Our custom panel of 443 amplicons across 2 pools included the 4 core MMR genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) plus 6 additional genes involved in the mismatch repair pathway (*MSH3*, *MLH3*, *EPCAM*, *PMS1*, *EXO1, SETD2*) comprising ~35.07Kb total capture size (targeting exonic and splice sites only) and achieving 94.5% mean coverage of the 10 genes (ranging from *MLH1* – 96.9% > *PMS2* – 80.5%). DNA samples were obtained from archival FFPE tumour material with 10ng of dsDNA being used per sample per pool. Technical success was high for somatic mutation detection with 271/275 FFPE tumour DNA samples passing QC for library construction and sequencing and 84% of reads being on target. The mean read depth per sample was 2,800. Replicate testing provided identical results. Our reference group consisted of 10 MMR gene mutation carriers comprising 9 x CRCs and 1 x TVA polyp all of which demonstrated MMR-deficiency. Identified variants of interest with >10% allele fraction underwent Sanger sequencing to confirm their presence.

Sanger Sequencing

Validation of MMR gene mutations was performed using Sanger sequencing as previously described (Newcomb, Baron et al. 2007, Poynter, Siegmund et al. 2008, Walsh, Buchanan et al. 2010, Clendenning, Walsh et al. 2013, Buchanan, Tan et al. 2014). Amplicon specific primes are available on request.

Inversion-specific PCR

Primers were designed across the specific breakpoints of the *MSH2* exon 1-7 inversion where a 238bp product is amplified if the inversion is present. A control amplicon was designed for exon 3 of the TDG gene such that a 344bp product was amplified irrespective of whether the *MSH2* inversion is present or absent. The following primers were used:

Inversion Primers	**Primer Sequence**
Forward	GGGGCCATAATCCAGTCCTT
Reverse	atgtgtgcctgcatatgtgt
TDG Primers	**Primer Sequence**
Forward	CAAAACAACCAGTGGAACCCA
Reverse	acctcatgaagctgacacca

The inversion PCR amplifies a product with the following sequence:

238bp product

TGTGGAGCTGGGGCCATAATCCAGTCCTTATGTGATTACTGTGAAGTTATCCTTTTCC

CCCAACATCTACTTATGAATAAAGAGTTTATTAAGTGGTTAACTGCAAGGCAAGATT

GCTCACAGTACTCTCAATGACACTCCAGGCTCAATGGCCTAG[breakpoint]AGGACatata

tgtgtatatatacaCAtatatacgtatatatatacacacacacatatatatacacatatgcaggcacacat

UPPERCASE = Sequence 9.5Mb upstream of *MSH2* intron 7

LOWERCASE = *MSH2* intron 7 sequence

Inserted sequences that are not present in the reference sequence

Primers used to detect inversion

The PCR amplified bands from SLS12 who carried the *MSH2* exons 1-7 inversion were sequenced along with a positive control (which came from the Cardiff Molecular Genetics Laboratory in the UK) using the primers designed by Rhees *et al*[1]. The sequencing trace in **Supplementary Figure 2** confirms that the inversion carried by SLS12 is the same as the positive control.

Variant calling

For both germline and tumour DNA samples, raw FASTQ sequence quality control was confirmed using FastQC[2] and paired-end FASTQ files were aligned to the Human Reference Genome (hg19)[3] using BWA (v 0.7.12) [4]. Germline single nucleotide variants (SNVs) and short insertions and deletions (INDELS) were called using the GATK Best Practices Pipeline (v 3.4-46) [5]. Somatic single nucleotide variants (sSNVs) and short insertions and deletions (sINDELs) were called using Strelka (v 2.9.2)[6], Platypus (v 0.8.1) [7] and Mutect 2[8] using default settings. Each sSNV or sINDEL variant was annotated by the consensus of the three callers, and high-confidence calls were determined as those reported by at least 2 out of 3 callers to reduce false positives. Germline and somatic variants were annotated using the Ensembl Variant Effect Predictor[9] 77 (VEP) and SnpEff[10] 4.1, including *in silico* variant effect predictions from REVEL[11] and CADD[12], and population frequencies from gnomAD[13]. Pathogenicity classifications from InSiGHT[14] for the MMR genes and ClinVar[15] were associated with variants where available.

Germline structural variants (SVs) were detected with DELLY[16] 0.7.1, LUMPY[17] 0.2.11, and GRIDSS[18] 0.11.5, and somatic structural variants were detected with GRIDSS[18] 2.2.1 and Manta 1.5.0 [19]. In all cases, high-confidence SV calls were selected by applying the quality filters recommended in the documentation for each tool. The concordance between the calls made by each tool was computed, allowing a window of +/- 50bp uncertainty in break end coordinates. Germline and somatic copy number variants (CNVs) were detected using HMMCopy[20] 1.22.0.

SNVs were assessed for predicted effect on splicing using HumanSplicingFinder (HSF) 3.1[21] and MaxEntScan[22]. MSI was assessed computationally using MSIsensor 0.5[23] using recommended thresholds: >3.5 considered high level (MSI-H), 1 to 3.5 considered low level (MSI-L) and <1 considered microsatellite stable (MSS). Somatic mutational signatures were computed using the method described by DeconstructSigs[24], which estimates the relative contributions of each mutational process in a single tumour sample from a set of 30 standard profiles[25] as previously described[26]. Tumour mutation burden (TMB) was estimated by dividing the number of high-confidence sSNVs and sINDELs by the number of DNA bases in the coding region of the genome. Tumours with TMB >10 were considered hypermutated and with TMB >100 were considered ultra-hypermutated[27]. Tumour loss of heterozygosity (LOH) was assessed by our LOH[26] 0.3 tool which compares the allele fraction of germline variants within tumour samples and reports on regions with unexpected deviations from the germline state.

The selected thresholds for variant population frequency in gnomAD and predicted pathogenicity scores from CADD and REVEL were derived from the analysis of missense variants classified as pathogenic (class 5) or likely pathogenic (class 4) from the InSiGHT

database and published recommendations for the *in silico* prediction tools[11, 12], respectively. We annotated all likely pathogenic and pathogenic missense variants in the 4 main MMR genes in the InSiGHT database (accessed 18 September 2018) with their population frequencies from gnomAD and their predicted pathogenicity from CADD and REVEL to estimate suitable thresholds for variant filtering. We observed a mean population frequency of 2.9×10^{-5} (standard deviation 4.6×10^{-5}) with an extreme outlier maximum value of 1.7×10^{-4}, a mean CADD score of 29.3 (minimum 17.6 and standard deviation 3.8), and a mean REVEL score of 0.9 (minimum 0.3 and standard deviation 0.1).

All variants retained after filtering were manually inspected in the Integrative Genomics Viewer (IGV)[48] to assess the quality of read alignments and supporting evidence.

GATK detected the deletion of intron 6 in the Tier 2 gene *SMAD4* (c.787+1_788-1del) in SLS2 that was annotated as a high-impact splice-region variant by VEP. An association between large deletions in this gene and familial juvenile polyposis syndrome has previously been reported[28]. Inspection of SV calls in the same sample showed high-confidence deletions of 7 *SMAD4* introns, a pattern which typically arises from the presence of processed pseudogenes that are not present in the genome reference, and are therefore likely to be false positives[29]. GATK did not detect the other intron deletions identified by the SV callers most likely because they were longer than the sequencing read length.

TABLES

Supplementary Table 1. Three tiers of candidate genes used to prioritize germline variants from whole genome sequencing. Tier 1 genes underlie Lynch syndrome. Tier 2 genes assessed by the ClinGen Hereditary Colorectal Cancer and Polyposis Susceptibility Gene Curation Panel to have definitive, strong or moderate evidence supporting an association with hereditary CRC and/or polyposis or other syndromes with rare manifestation of CRC and/or polyposis. Tier 3 genes contain a curated list of DNA repair genes as described in the literature.

FIGURES

Supplementary Figure 1. (A) DNA sequencing read alignments for the 1928 bp deletion of *MSH2* exon 6 (chr2:47640695-47642623) in control sample C2. (B-C) DNA sequencing read alignments for the two breakends of the 9.5 Mb inversion encompassing exons 1-7 of *MSH2* (chr2:38121107-chr2:47669532) in mother-daughter pair of SLS11 and SLS12. Red bars in the alignment diagrams indicate discordant read pairs and illustrate that these large structural variants are readily detected in this type of sequencing data. (D) Cartoon diagram illustrating the position of the 9.5 Mb inversion encompassing exons 1-7 of *MSH2* on chromosome 2.

Supplementary Figure 2. Sequencing trace confirming that the inversion carried by SLS12 is the same as the positive control. The breakpoints are marked with a black line. The sequence between the lines is a novel inserted sequence.

References

1. Rhees J, Arnold M, Boland CR. Inversion of exons 1-7 of the MSH2 gene is a frequent cause of unexplained Lynch syndrome in one local population. Fam Cancer 2013.
2. Andrews S. FastQC: a quality control tool for high throughput sequence data. 0.10.1 ed, 2012.
3. Genome Reference C. hg19, GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1), 2009.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
5. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:11 10 1-33.
6. Saunders CT, Wong WS, Swamy S, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012;28:1811-7.
7. Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 2014;46:912-918.
8. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 2013;31:213-219.
9. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.
10. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80-92.
11. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 2016;99:877-885.
12. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310-5.
13. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 2019.
14. Thompson BA, Spurdle AB, Plazzer JP, 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 2014;46:107-15.
15. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062-D1067.
16. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012;28:i333-i339.
17. Layer RM, Chiang C, Quinlan AR, et al. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014;15:R84.
18. Cameron DL, Schroder J, Penington JS, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 2017;27:2050-2060.
19. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016;32:1220-2.
20. Ha G, Roth A, Lai D, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res 2012;22:1995-2007.

21. Desmet FO, Hamroun D, Lalande M, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 2009;37:e67.
22. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 2004;11:377-94.
23. Niu B, Ye K, Zhang Q, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 2014;30:1015-6.
24. Rosenthal R, McGranahan N, Herrero J, et al. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 2016;17:31.
25. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415-21.
26. Georgeson P, Walsh MD, Clendenning M, et al. Tumour Mutational Signature in Sebaceous Skin Lesions from Individuals with Lynch Syndrome. Molecular Genetics & Genomic Medicine 2019;provisionally accepted May 1st 2019.
27. Campbell BB, Light N, Fabrizio D, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell 2017;171:1042-1056 e10.
28. Aretz S, Stienen D, Uhlhaas S, et al. High proportion of large genomic deletions and a genotype phenotype update in 80 unrelated families with juvenile polyposis syndrome. J Med Genet 2007;44:702-9.
29. Watson CM, Camm N, Crinnion LA, et al. Characterization and Genomic Localization of a SMAD4 Processed Pseudogene. J Mol Diagn 2017;19:933-940.