medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Title: Evaluating the utility of tumour mutational signatures for identifying hereditary colorectal cancer and polyposis syndrome carriers

Peter Georgeson[1, 2], Bernard J. Pope[1,2,3], Christophe Rosty[1,2,4,5], Mark Clendenning[1,2], Romy Walker[1,2], Khalid Mahmood[1,2,3], Jihoon E. Joo[1,2], Ryan Hutchinson[1,2], Susan Preston[1,2], Julia Como[1,2], Sharelle Joseland[1,2], Aung K. Win[6,8], Finlay A. Macrae[6,7], John L. Hopper[8], Mark A. Jenkins[2,8], Ingrid M. Winship[6,9], Daniel D. Buchanan[1,2,6]

[1] Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, Victoria, Australia
[2] University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, Victoria, Australia
[3] Melbourne Bioinformatics, The University of Melbourne, Carlton, Victoria, Australia
[4] Envoi Pathology, Brisbane, Queensland, Australia
[5] University of Queensland, School of Medicine, Herston, Queensland, Australia
[6] Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Victoria, Australia
[7] Colorectal Medicine and Genetics, The Royal Melbourne Hospital, Parkville, Victoria, Australia
[8] Centre for Epidemiology and Biostatistics, The University of Melbourne, Carlton, Victoria, Australia
[9] Department of Medicine, The University of Melbourne, Parkville, Victoria, Australia

Correspondence:
Associate Professor Daniel Buchanan
Colorectal Oncogenomics Group
Department of Clinical Pathology
The University of Melbourne
Victorian Comprehensive Cancer Centre
305 Grattan Street
Parkville, Victoria, 3010 Australia
Ph: +61 385597004
Email: daniel.buchanan@unimelb.edu.au

Word count: 4003

Tables: 2

Figures: 7

Keywords: tumour mutational signatures, colorectal cancer, hereditary colorectal cancer, Lynch syndrome, DNA mismatch repair, *MUTYH*, *NTHL1*.

Abbreviations: area under the curve (AUC), Australasian Colorectal Cancer Family Registry (ACCFR), base excision repair (BER), colon adenocarcinoma (COAD), colorectal cancers

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(CRCs), depth of coverage (DP), formalin-fixed paraffin embedded (FFPE), Genetics of Colonic Polyposis Study (GCPS), indel (ID), Fisher's linear discriminant (LD), microsatellite instability (MSI), mismatch repair (MMR), MMR-deficient (MMRd), MMR-proficient (MMRp), *MUTYH*-associated polyposis (MAP), *NTHL1*-associated polyposis (NAP), pathogenic variant (PV), percentage points (pp), receiver operating curve (ROC), single base substitution (SBS), The Cancer Genome Atlas (TCGA), tumour mutational burden (TMB), tumour mutational signatures (TMS), variant allele fraction (VAF), variants of uncertain clinical significance (VUS), whole exome sequencing (WES)

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABSTRACT (250 words)

Objective: Germline pathogenic variants (PVs) in the DNA mismatch repair (MMR) genes and the base excision repair genes *NTHL1* and *MUTYH* underlie hereditary CRC and polyposis syndromes. We evaluate the robustness and discriminatory potential of tumour mutational signatures in colorectal cancers (CRCs) for identifying germline PV carriers.

Design: Whole exome sequencing of FFPE CRC tissue was performed on 14 MMR, 6 biallelic *MUTYH*, and 1 biallelic *NTHL1* PV carrier, 9 sporadic MMR-deficient CRCs (MMRd controls) and 18 sporadic MMR-proficient CRCs (MMRp controls). COSMIC V3 single base substitution (SBS) and indel (ID) mutational signatures were calculated and assessed for their ability to differentiate CRCs that developed in carriers and non-carriers.

Results: The combination of mutational signatures SBS18 and SBS36 contributing >23% of the signature profile was able to discriminate biallelic *MUTYH* carriers from MMRp and MMRd control CRCs with >99% confidence. Variant specific signatures SBS18 and SBS36 were identified for *MUTYH* p.Gly396Asp (p=0.015) and *MUTYH* p.Tyr179Cys (p=0.0012), respectively. SBS30 was significantly increased in a CRC from a biallelic *NTHL1* carrier compared with MMRp and MMRd control CRCs. The combination of ID2, ID7, SBS15 and SBS1 could discriminate the 14 MMR PV carrier CRCs from the MMRp control CRCs, however, SBS and ID signatures, alone or in combination, could not provide complete discrimination between CRCs from MMR PV carriers and sporadic MMRd controls.

Conclusion: Assessment of SBS and ID signatures can discriminate CRCs from *MUTYH*, *NTHL1* and MMR PV carriers from non-carriers, demonstrating utility as a potential diagnostic and variant classification tool.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

SIGNIFICANCE OF THE STUDY

What is already known about this subject?

- Identifying carriers of pathogenic variants (PVs) in moderate/high-risk CRC and polyposis susceptibility genes has clinical relevance for diagnosis, targeted screening and prevention strategies, prognosis, and treatment options, however, challenges still remain in the identification of carriers and the classification of rare variants in these genes.
- Previous studies have identified tumour mutational signatures that result from defective DNA repair including DNA mismatch repair (MMR) deficiency and base excision repair defects, DNA repair mechanisms that underlie the common hereditary CRC and polyposis syndromes.

What are the new findings?

- Single base substitution (SBS)-related mutational signatures derived from whole exome sequencing of FFPE-derived CRC tissue DNA could effectively discriminate CRCs that developed in biallelic *MUTYH* PV carriers and biallelic *NTHL1* PV carriers from CRC-affected non-carriers.
- CRCs that developed in MMR PV carriers (Lynch syndrome) could be effectively differentiated from sporadic MMR-proficient CRC by a combination of SBS and indel (ID) signatures, but they were less effective at discriminating Lynch syndrome-related CRC from sporadic MMR-deficient CRC resulting from *MLH1* gene promoter hypermethylation.
- The SBS and ID mutational signatures associated with hereditary CRC and polyposis syndrome carriers were robust to changes in experimental settings.
- We demonstrate the optimal variant filtering settings for calculating mutational signatures and define stringent thresholds for classifying CRC aetiology as hereditary or non-hereditary.

How might it impact on clinical practice in the foreseeable future?

- Deriving SBS- and ID-related mutational signatures from CRCs can identify carriers of PVs in hereditary CRC and polyposis susceptibility genes.
- The application of mutational signatures will improve the diagnosis of syndromic CRC and aid in variant classification, leading to improved clinical management and CRC prevention.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide and is a leading cause of cancer-related morbidity and mortality [1]. Currently, 5-10% of CRCs develop in individuals who carry a pathogenic variant (PV) in a known hereditary CRC and/or polyposis susceptibility gene, including the DNA mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*) and the base excision repair (BER) genes *MUTYH* [2] and *NTHL1* [3] (reviewed in [4]). Identifying carriers of PVs in these susceptibility genes has important implications for preventing subsequent primary cancers in the proband [5–7] and for the prevention of CRC in relatives through targeted screening approaches such as colonoscopy with polypectomy [8,9].

Currently, the most common strategy to identify MMR gene PV carriers (Lynch syndrome) starts with testing the tumour for microsatellite instability (MSI) and/or loss of MMR protein expression by immunohistochemistry [10,11]. However, loss of MMR protein expression (MMR-deficiency) in a CRC is not diagnostic for carrying a PV because MMR-deficiency can be caused by epigenetic inactivation of the MMR genes (hypermethylation of the *MLH1* gene promoter) or biallelic somatic mutations of MMR genes [12,13]. Moreover, no tumour-based approach is routinely applied to identify biallelic PV carriers in the *MUTYH* or *NTHL1* genes; currently germline testing is guided by the presence of associated phenotypic features [4], although the phenotype for biallelic *MUTYH* carriers has been shown to be variable, making this approach suboptimal [14]. Consequently, the classification of variants of uncertain clinical significance (VUS) still present challenges [15].

The determination of tumour mutational signatures is an emerging approach that integrates the somatic mutation landscape within a single tumour to identify patterns associated with distinct oncogenic pathways [16–19]. Each signature is derived from compositional changes of single base substitutions (SBS), indels (ID), and doublets. The most recent version of the predominant mutational signature framework published on the COSMIC website defines 95 signatures, of which a proposed aetiology is available for 63 (66%) signatures [20]. To date, seven signatures have been identified that relate to defective DNA repair including MMR (SBS6, SBS15, SBS20, SBS26), and BER defects caused by dysfunctional *NTHL1* (SBS30)[7] and *MUTYH* (SBS18 and SBS36) [21,22]. Therefore, determining signatures

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

presents a promising new strategy to identify PV carriers and to improve the current classification of rare variants and VUS.

Although mutational signatures show substantial promise in the translational setting, clinical adoption has been limited [23]. In this study, we evaluated the SBS and ID signature landscapes in CRCs caused by germline PVs in the MMR genes or by biallelic PVs in the *MUTYH* and *NTHL1* genes, to test the clinical utility of mutational signatures for predicting PVs in specific genes. We assessed the effect of experimental settings and quantified the discriminatory potential of signatures for identifying PV carriers from non-syndromic CRC.

MATERIALS AND METHODS

Study Cohort

CRC-affected individuals recruited to the Genetics of Colonic Polyposis Study (GCPS) [24] or the Australasian Colorectal Cancer Family Registry (ACCFR) [25,26] were selected for analysis in this study if they were carriers of germline PVs in *MLH1*, *MSH2* or *MSH6* or were biallelic or monoallelic carriers of PVs in the *MUTYH* or *NTHL1* genes. A single CRC-affected individual carrying a PV and a VUS in the *MUTYH* gene was also included (sample M12). Tumour MMR status was determined by immunohistochemistry with details of the tumour and germline characterisation undertaken described previously [7,27,28]. Two groups of CRC-affected individuals from the ACCFR [28] were selected as age-matched non-syndromic controls: 1) individuals who developed MMR-proficient (MMRp) CRC without a known germline mutation in a hereditary CRC/polyposis associated gene were included as "MMRp controls" (n=18), and 2) individuals who developed MMRd CRC resulting from somatic *MLH1* gene promoter hypermethylation were included as "MMRd controls" (n=9). In addition, 244 CRC tumours from The Cancer Genome Atlas (TCGA) Colon adenocarcinoma (COAD) study [29] were included as a validation set of non-syndromic CRCs, ranging in their age at CRC diagnosis of 34-90 years (**Supplementary Table 1**). Tumour MSI status [30] and DNA methylation analysis of the *MLH1* gene promoter were used to stratify the TCGA tumours into 218 MMRp ("TCGA MMRp controls") and 26 MMRd ("TCGA MMRd controls") cancers (**Supplementary Methods**).

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Whole Exome Sequencing (WES) and Analysis

Formalin-Fixed Paraffin Embedded (FFPE) CRC tissues were macrodissected and sequenced as tumour with matching peripheral blood-derived DNA sequenced as germline (**Supplementary Methods**). Somatic single-nucleotide variants and short insertion and deletions (indels) were called with Strelka 2.9.2 (recommended workflow) and used to calculate mutational signatures using the method described by DeconstructSigs [31] from the set of COSMIC version 3 signatures [20] (**Supplementary Methods**). The impact of experimental settings was explored by filtering variants based on depth of coverage (DP) and the variant allele fraction (VAF) in the tumour, then calculating signatures at each filter point. Details of the TCGA analysis are provided in the **Supplementary Methods**.

We assessed the ability of signatures to separate the specific syndromic CRC group from the non-syndromic CRC group using four methods, at each filtering setting, for each relevant signature: 1) Fisher's linear discriminant (LD) [32], or for groups with a single syndromic sample (N01), Grubbs outlier test [33], 2) absolute margin separating the specific syndromic CRC group from non-syndromic groups, 3) the difference in the means of the two groups, and 4) P-values were calculated using a one-sided t-test, or for groups with a single sample (N01), Grubbs outlier test [33] (**Supplementary Methods**).

We fitted the signatures of each group of CRCs to a beta distribution to enable the generation of percentiles for each group. Confidence thresholds that a given measurement belonged to the hereditary CRC syndrome of interest were also calculated. To find the best combination of signatures for identifying each hereditary syndrome, we applied forward selection while maximising the area under the curve (AUC) of the receiver operating curve (ROC) and the margin between the groups, with stringent requirements to reduce the likelihood of overfitting (**Supplementary Methods**).

Patient and Public Involvement

There has been no patient or public involvement in the generation of this research report.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

RESULTS

Overall Tumour Mutational Signature Results

The 60 CRCs analysed in this study were classified as syndromic CRCs (n=33) and non-syndromic CRC controls (n=27), comprising of 9 MMRd CRCs (mean age at diagnosis±SD 56.8±7.3 years), and 18 MMRp CRCs (44.1±12.4 years) (**Table 1**). The syndromic CRC group were not significantly different in their age at diagnosis compared with the non-syndromic CRCs (48.1±13.0 years versus 48.3±12.5 years, p=0.94) but were significantly younger when compared with the 244 TCGA tumours (66.0±12.6 years, $p=5x10^{-13}$). The characteristics of each individual and their CRC are provided in **Supplementary Table 1** and **Supplementary Table 2**.

Figure 1 illustrates the calculated SBS and ID signature compositions for each of the 60 CRCs (TCGA tumour SBS and ID signatures are shown in **Supplementary Figure 1**). The SBS-derived signatures that have been previously reported to be associated with BER defects were observed as the dominant signature in 9 out of 10 biallelic *MUTYH* related CRCs (SBS18 and SBS36) and for the single biallelic *NTHL1* related CRC (SBS30) (**Figure 1**). SBS6, SBS15, ID2 and ID7 were the dominant signatures in MMRd CRCs from both Lynch syndrome and *MLH1* methylated tumours (MMRd controls). The signatures observed for each CRC, their proportion and ranking are provided in **Supplementary Table 3**. The signatures associated with hereditary CRC syndromes are shown in **Figure 2**.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1. Summary of subtypes of hereditary CRC and polyposis syndromes investigated in this study, including their underlying gene defect, previously reported mutational signature associations and the number of individuals and CRCs tested by WES. Two sources of non-syndromic CRCs were included as control groups comprising CRCs from the ACCFR study and from the TCGA COAD study.

Syndrome	Defective Gene(s)	DNA repair mechanism	Associated Signatures	Sub-category	Indiv-iduals	No. of CRCs	CRC study IDs
Syndromic CRCs							
MUTYH-associated polyposis (MAP)	Biallelic *MUTYH*	Base excision repair	SBS18, SBS36	Biallelic *MUTYH* carrier	6	10	M01-M10
				Monoallelic carrier	3	3	M11, M13, M14
				VUS carrier	1	1	M12
NTHL1-associated polyposis (NAP)	Biallelic *NTHL1*	Base excision repair	SBS30	Bialllic *NTHL1* carrier	1	1	N01
				Monoallelic carrier	4	4	N02-N05
Lynch Syndrome	*MLH1*, *MSH2*, *MSH6*, *PMS2*	Mismatch repair	SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, SBS44, ID2, ID7, DBS7, DBS10	*MLH1* carrier	7	7	L07-L13
				MSH2 carrier	3	3	L01, L02, L14
				MSH6 carrier	4	4	L03-L06
Non-syndromic CRCs							
Sporadic MMR-deficient CRCs	*MLH1* methylation	Mismatch repair		MMRd controls	9	9	K01-K9
Sporadic MMR-proficient CRCs				MMRp controls	18	18	C01-C18
TOTAL					**56**	**60**	
TCGA Non-syndromic CRCs							
Sporadic MMR-deficient CRCs	*MLH1* methylation	Mismatch repair		TCGA MMRd controls	26	26	
Sporadic MMR-proficient CRCs				TCGA MMRp controls	218	218	
TOTAL					**244**	**244**	

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Identifying a biallelic *MUTYH* carrier mutational signature in CRC

For biallelic *MUTYH* carrier CRCs (M01-M10), only two signatures, SBS18 and SBS36, were significantly enriched when compared with non-*MUTYH* CRCs (mean±SD; 20.6±15.1% v. 0.6±1.7%; $p=6x10^{-12}$ and 25.5±15.6% v. 0.1±0.3%; $p=2x10^{-15}$, respectively; **Figure 2**). This significant enrichment of SBS18 and SBS36 was also observed when the biallelic *MUTYH* carrier CRCs (M01-M10) were compared with the TCGA tumours (mean±SD; 20.6±15.1% v. 2.8±5.1%; $p=1x10^{-18}$ and 25.5±15.6% v. 0.3±1.2%; $p=7x10^{-66}$, respectively; **Figure 2**).

The SBS18 and SBS36 mutational signatures differed by the underlying germline *MUTYH* mutation. The CRCs from homozygous or compound heterozygous carriers of the c.536A>G p.Tyr179Cys PV (n=7) showed significantly higher proportions of SBS36 (p=0.0012), while the CRCs from homozygous carriers of the c.1187G>A p.Gly396Asp variant (n=3) showed significantly higher proportions of SBS18 (p=0.015) (**Figure 3a** and **Figure 3b**). Tumour mutational burden (TMB) was also found to be significantly higher in homozygous or compound heterozygous carriers of the c.536A>G p.Tyr179Cys (15.22±3.65 mutations/Mb) relative to p.Gly396Asp homozygous carriers (6.30±0.52 mutations/Mb; p=0.0054).

We investigated the ability of mutational signatures to differentiate CRCs from biallelic *MUTYH* carriers from non-carriers. The combination of SBS18 and SBS36 provided the most effective discrimination of CRCs from biallelic *MUTYH* carriers from all other ACCFR CRCs in the study, achieving an AUC of 1.00, and separation of 9.0 percentage points (the difference between two percentages; pp) between the two groups (**Table 2**). The sum of SBS18 and SBS36 accounted for, on average, almost half of the total signature composition for biallelic *MUTYH* related CRCs (46.1±12.3%; range 18.2% to 57.0%) whereas these signatures had negligible contribution to all other CRCs in the study (ACCFR non-MUTYH CRCs: 0.6±1.8%; range 0.0% to 9.2%, $p=3x10^{-30}$; TCGA: 3.0±5.3%; range 0.0% to 29.3%, $p=4x10^{-65}$; **Figure 3b**).

We found that the robustness of mutational signatures is dependent on different minimum VAF and minimum DP variant filtering settings. A minimum VAF threshold between 0.05 and 0.20 (**Figure 4a** and **Figure 4b**), and a minimum DP threshold of at least 30 (**Figure 4c**

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

and **Figure 4d)** effectively separated biallelic *MUTYH* carrier and non-*MUTYH* carrier CRC groups. A minimum VAF of 0.1 and minimum DP of 50 were selected to maximise the capacity of signatures to identify CRCs from biallelic *MUTYH* carriers. Our analysis showed that a CRC with a combined SBS18 and SBS36 proportion of >23% or <9% has a >99% or <1% likelihood, respectively, of having a biallelic defect in *MUTYH* (**Figure 4c** and **Figure 4d**).

We applied these thresholds in two scenarios: 1) evaluating monoallelic carriers, and 2) classifying a VUS. The sum of SBS18 and SBS36 was evaluated in three CRCs from three monoallelic *MUTYH* PV carriers (M11, M13 and M14): both SBS18 and SBS36 were 0% for each of the three CRCs, suggesting that monoallelic *MUTYH* PVs are not related to defective BER. No identifiable second somatic "hit" (single nucleotide variant or loss of heterozygosity) in *MUTYH* was evident in any of the three CRCs (**Supplementary Figure 3**). We then calculated the combined SBS18 and SBS36 signatures for a CRC (M12) from a person who carried a heterozygous *MUTYH* c.1187G>A p.Gly396Asp PV and a heterozygous *MUTYH* c.912C>G p.Ser304Arg VUS [34]. The sum of SBS18 and SBS36 was 0.3%, (**Figure 4c** and **Figure 4d**) and therefore, according to our analysis of confirmed biallelic *MUTYH* variants, there is a <1% likelihood that this CRC has biallelic *MUTYH* inactivation. This further suggests that the c.912C>G p.Ser304Arg variant is not likely to be pathogenic or is *in cis* with the c.1187G>A p.Gly396Asp PV.

Identifying a biallelic *NTHL1* carrier mutational signature in CRC

The value of SBS30 was significantly higher in the CRC from the biallelic *NTHL1* carrier (N01) compared with the rest of the non-*NTHL1* ACCFR CRCs (46.5% v. 0.4±1.4%; $p=3x10^{-230}$) and compared with the TCGA tumours (46.5% v 0.0±2.5%; $p=1x10^{-77}$; **Figure 2**). SBS23 and SBS42 were also significantly higher in the biallelic *NTHL1* CRC compared with the rest of the ACCFR CRCs ($p=7x10^{-204}$ and $p=2x10^{-12}$, respectively; **Figure 2**). Forward selection identified SBS30 as the most effective mutational signature to discriminate the CRC from the biallelic *NTHL1* carrier from all other ACCFR CRCs in the study, achieving an AUC of 1.00, and a separation of 38.0% from the rest of the CRCs (**Table 2, Supplementary Table 4**), a result that showed a high degree of replication when the TCGA

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

tumours were utilised (Table 2), supporting the discriminatory strength of SBS30 for biallelic *NTHL1*.

Variant filtering settings were less critical for identifying the *NTHL1* carrier N01 (**Figure 5a** and **Figure 5b**), with SBS30 exhibiting discriminatory power at VAF thresholds between 0.05 and 0.375, and all measured DP thresholds. Using a VAF threshold of 0.1 and DP filter of 50, our analysis indicates that a CRC with a measured SBS30 value of >41% confers a likelihood of >99% that the tumour has a biallelic *NTHL1* defect. Similarly, if SBS30 is <38%, a biallelic *NTHL1* defect is <1% likely (**Figure 5c** and **Figure 5d**). The CRCs from the four monoallelic *NTHL1* PV carriers (N02, N03, N04 and N05) all exhibited an SBS30 value of 0%, suggesting these CRCs do not have defective BER related to *NTHL1*. No identifiable second somatic "hit" (single nucleotide variant or loss of heterozygosity) in *NTHL1* was evident in these four CRCs (**Supplementary Figure 3**).

Identifying a MMR gene (Lynch syndrome) carrier mutational signature in CRC

We investigated the utility of previously reported signatures for identifying Lynch syndrome-related CRCs compared to MMRp CRCs and sporadic MMRd CRCs. The Lynch syndrome-related CRCs (L01-L14) showed significantly higher levels of SBS1, SBS15, SBS20, SBS21, ID2 and ID7 ($p=5x10^{-6}$, $p=1x10^{-7}$, $p=1x10^{-8}$, $p=8x10^{-10}$, $p=7x10^{-16}$, $p=9x10^{-10}$ respectively) compared with the ACCFR MMRp control CRCs (**Figure 2**). SBS6 was seen at high levels in the Lynch syndrome CRCs (17.8% ± 8.4%) but was not considered significantly different when compared with MMRp CRCs (8.4±7.7%) after adjusting for multiple comparisons (p=0.0002). When comparing Lynch syndrome-related CRCs (L01-L14) to the MMRd control group of CRCs (K01-K09), SBS1 showed the most significant difference between the groups (26.3±8.7% v. 15.8±9.8%, p=0.009), however neither SBS1 nor any other signature were significantly different after adjustment for multiple comparisons (**Figure 2**).

When comparing the CRCs from *MLH1*, *MSH2* and *MSH6* PV carriers, no significant gene-specific difference in mutational signatures was found after multiple comparison adjustment (**Figure 3c** and **Figure 3d**). Signatures of interest were found to be consistent at VAF

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

thresholds between 0.00 and 0.15, and DP thresholds between 10 and 150 (**Figure 6, 7, Supplementary Figure 2**).

To first investigate a common molecular stratification in CRC, we applied forward selection to discriminate MMRd CRC (L01-L14 and K01-K09 CRCs) from all MMRp CRCs (C01-C18, M01-M14, N01-N05) in this study. This identified ID2, ID7, and SBS15 as the most informative combination of signatures, achieving an AUC of 1.00, and separation of 4.1pp between the groups (**Table 2**). We then applied forward selection to discriminate Lynch syndrome-related CRCs (L01-L14) from all MMRp CRCs in this study, identifying ID2, ID7, SBS1, and SBS15 as the most informative combination of signatures, achieving an AUC of 1.00, and separation of 33.2pp between the groups (**Table 2, Figure 3d**), a result that showed a high degree of replication when the TCGA MMRp controls were utilised (**Table 2, Figure 3d**). When considering the sum of ID2, ID7, SBS1 and SBS15, the likelihood of a Lynch syndrome-related CRC relative to a MMRp CRC is >99% and <1% at values of >105% and <76% (**Figure 7a** and **Figure 7b**).

A common diagnostic challenge for MMRd CRC is to differentiate inherited (Lynch syndrome) from sporadic MMRd CRC. Forward selection showed no significant (adjusted $P < 0.05$) mutational signature. SBS1 was the best performing signature that could discriminate Lynch syndrome-related CRCs (L01-L14) from the MMRd control CRCs (K01-K09) in this study. However, SBS1 only achieved an AUC of 0.76 and the groups overlap by 23% (**Table 2**). This result did not improve when a larger group of TCGA MMRd controls (n=26) were utilised (**Table 2**). Consequently, the Lynch syndrome and sporadic MMRd CRCs could not be separated with high confidence (**Figure 7c** and **Figure 7d**).

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 2. Individual mutational signatures identified to be associated with a hereditary CRC syndrome from the analysis of 60 CRCs. Each signature with an AUC>0.90 and mean difference >0.10 was included in the forward selection analysis. The best combination of signatures was determined using forward selection and adding signatures with the highest AUC, margin, and mean difference. Discrimination results with the ACCFR MMRp and MMRd control groups replaced with TCGA MMRp and MMRd controls to assess validation against a different set of tumours.

Comparison	Signatures	AUC	LD or Grubbs	Margin	Mean Diff	p-value
MUTYH Biallelic v. all other	SBS18	0.992	1.732	-0.040	0.200	1.3E-12
M01-M10 v. M11-M14 N K L C	SBS36	0.948	2.649	-0.019	0.254	2.6E-16
	SBS18, SBS36	**1.000**	**13.400**	**0.090**	**0.454**	**5.3E-32**
M01-M10 v. M11-M14 N L TCGA-MMRp TCGA-MMRd	***SBS18, SBS36***	***0.968***	***5.071***	***-0.270***	***0.389***	***1.0E-56***
	SBS30	1.000	560.054	0.379	0.461	7.0E-246
NTHL1 Biallelic v. all other	ID1	0.966	2.915	-0.071	0.131	7.9E-03
N01 v. N02-N05 M K L C	ID12	0.932	1.620	-0.141	0.194	3.6E-02
	SBS30	**1.000**	**560.054**	**0.379**	**0.461**	**7.0E-246**
N01 v. N02-N05 M L TCGA-MMRp TCGA-MMRd	***SBS30***	***1.000***	***98.561***	***0.098***	***0.503***	***4.4E-45***
	ID2	0.997	7.554	-0.074	0.406	1.68E-20
MMRd v. MMRp	ID7	0.976	1.993	-0.114	0.155	9.71E-12
L K v. M N C	SBS15	0.907	1.070	-0.107	0.196	4.02E-08
	ID2, ID7, SBS15	**1.000**	**9.423**	**0.041**	**0.756**	**1.58E-24**
L TCGA-MMRd v. M N TCGA-MMRp	***ID2, ID7, SBS15***	***0.994***	***13.426***	***-0.289***	***0.953***	***6.20E-87***
	ID2	0.996	7.456	-0.074	0.414	6.9E-16
Lynch v. MMRp	ID7	0.975	1.644	-0.114	0.158	8.8E-10
L v. M N C	**ID2, ID7, SBS15, SBS1**	**1.000**	**17.147**	**0.332**	**0.873**	**1.1E-23**
L v. M N TCGA-MMRp	***ID2, ID7, SBS15, SBS1***	***0.991***	***13.270***	***-0.367***	***0.919***	***3.4E-39***
Lynch v. *MLH1* methylated	SBS1	0.762	0.635	-0.228	0.105	9.3E-03
L v. K	**SBS1**	**0.762**	**0.635**	**-0.228**	**0.105**	**9.3E-03**
L v. TCGA-MMRd	***SBS1***	***0.683***	***0.208***	***-0.368***	***0.060***	***2.7E-02***

M= MMRp CRCs from ACCFR that are from biallelic *MUTYH* PV carriers (M01-M10) or from monoallelic *MUTYH* PV carriers (M11, M13, M14) or monoallelic *MUTYH* PV and VUS carrier (M12).
N= MMRp CRCs from ACCFR that are from biallelic *NTHL1* PV carrier (N01) or from monoallelic *NTHL1* PV carriers (N02-N05).
L= MMRd CRCs from the ACCFR from people with *MLH1*, *MSH2* or *MSH6* gene mutations (Lynch syndrome; L1-L14)
K= MMRd control CRCs from the ACCFR with evidence of *MLH1* gene promoter hypermethylation (K1-K9)
C= MMRp control CRCs from the ACCFR (C1-C18)

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

DISCUSSION

In this study, we compared mutational signature profiles in CRCs from carriers of PVs in CRC and polyposis susceptibility genes with those in CRCs from individuals who did not carry a PV in one of these genes. Germline PVs in the DNA MMR genes and biallelic PVs in the *MUTYH* and *NTHL1* genes result in a high risk of developing CRC and other extra-colonic cancers [8,9,35], highlighting the importance of identifying carriers [3,7]. Our analysis identified multiple SBS and/or ID signatures that were associated with each syndromic CRC, supporting germline inactivation of the DNA MMR and BER pathways as key drivers of CRC tumourigenesis. The sum of SBS18 and SBS36 provided the most effective combination to differentiate CRCs from biallelic *MUTYH* carriers from CRCs from non-carriers. The CRC from a biallelic *NTHL1* carrier demonstrated significantly increased proportions of SBS23, SBS30 and SBS42 compared with CRCs from non-carriers, however, the presence of SBS30 at >41% was sufficient to provide >99% likelihood that a CRC had developed in a person with biallelic inactivation of the *NTHL1* gene. These findings support previous observations of the significance of SBS30 in biallelic *NTHL1* deficient tumours [3,7,36]. CRCs that developed in MMR PV carriers (Lynch syndrome) could be effectively differentiated from sporadic MMR-proficient CRC by a combination of signatures (ID2, ID7, SBS1 and SBS15), but signatures were less effective at discriminating Lynch syndrome-related CRC from sporadic MMR-deficient CRC resulting from *MLH1* gene promoter hypermethylation (SBS1: AUC=76%). These findings were consistent when assessed against the larger TCGA MMRp and MMRd tumours, highlighting the robustness of these findings and the utility of deriving mutational signatures from CRCs to identify carriers of PVs in MMR, *MUTYH* and *NTHL1* genes.

SBS18 and SBS36 were individually associated with CRCs from biallelic *MUTYH* carriers in our cohort, but neither signature alone completely separated the carriers from the non-*MUTYH* CRCs. Combining SBS18 and SBS36 resulted in a 9.0pp separation from the non-*MUTYH* carrier CRCs. This was also observed for the Lynch syndrome-related CRCs where the combination of ID2, ID7, SBS15 and SBS1 resulted in a 33.2pp separation from the MMRp CRCs, while each individually associated signature alone overlapped with the MMRp CRCs, highlighting the benefit of combining signatures for improved discrimination.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

This study confirms previous reports of the association of SBS18 and SBS36 with CRCs from biallelic *MUTYH* carriers [21,22]. In addition, we observed significant differences between individual signatures and the underlying germline *MUTYH* PV. TMB in CRCs from p.Tyr179Cys PV carriers was also significantly elevated relative to the p.Gly396Asp PV carriers (p=0.0054) and to the MMRp controls (p=0.019), suggesting that the underlying germline mutation may have implications for immunotherapy-based treatment of CRC in these carriers. The reason for the observed differences by specific mutation is not yet known but it is possible that disruption of different functional domains (p.Tyr179Cys: N-terminal domain; p.Gly396Asp: C-terminal domain) may result in different somatic mutations and mutational spectrum. It remains to be determined if the signature composition will differ for carriers of *MUTYH* PVs other than the two common European PVs.

Monoallelic *MUTYH* PV carriers are reported to have an increased risk of CRC [27,37]. A previous study identified high levels of SBS18 in the CRCs from two monoallelic germline *MUTYH* PV carriers where loss of the wildtype allele was observed in the tumour [21]. In this study, SBS18 and/or SBS36 were not increased in the CRCs from the three monoallelic *MUTYH* carriers, nor did we find evidence of a second somatic hit in their CRCs. In addition, none of the CRCs from four monoallelic *NTHL1* PV carriers exhibited SBS30. Our findings support the notion that biallelic *MUTYH* or *NTHL1* inactivation is necessary to promote BER deficiency-related CRC tumourigenesis. This approach could be applied to rare variant classification, including VUSs. We explored this in a CRC (M12) from a carrier of a *MUTYH* PV (p.Gly396Asp) and a VUS (p.Ser304Arg). The absence of both SBS18 and SBS36 suggests that this VUS is not pathogenic and could be further supported by exclusion of the variant being *in cis* with the p.Gly396Asp PV through parent genotyping.

The current proposed mechanism for the accumulation of MMRd associated somatic mutations is based on polymerase slippage, particularly during replication of low complexity regions such as homopolymers, followed by defective repair of these errors [38]. The result is MSI, a phenomenon associated with Lynch syndrome. Our results reflect this underlying molecular mechanism, showing that ID signatures more effectively identify MMRd samples compared to SBS signatures. The most relevant signatures, ID2 and ID7, are primarily composed of 1bp homopolymer deletions, which corresponds to the expected aetiology. We

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

demonstrated that mutational signatures were able to differentiate MMRd CRC from MMRp CRC, a feat that is currently achieved by MMR immunohistochemistry and is recommended to be performed on all newly diagnosed CRCs to screen for Lynch syndrome-associated CRC [10]. As tumour sequencing becomes more widely implemented, calculating signatures on CRCs could supersede the need for MMR immunohistochemistry. When comparing Lynch syndrome MMRd CRCs (L01-L14) to the CRCs from the MMRd control group (K01-K09), SBS1 had the greatest capacity to separate the two groups, but was not significant after adjusting for multiple comparisons (AUC 76%; p=0.009). From our analysis we find that the current mutational signature framework cannot completely differentiate Lynch syndrome CRCs from *MLH1* methylated CRCs. Investigating novel approaches to derive mutational signatures that are focused on differentiating these two important MMRd subtypes may provide a better discrimination tool.

We assessed the robustness of mutational signatures to changes in variant filtering settings. Our results demonstrated that filtering VAF and DP too leniently (increased artefacts) or too stringently (reduced number of variants) resulted in substantial changes to the signature proportions, including loss of the hereditary CRC-associated signatures in the CRC. In our study, a minimum VAF threshold from 0.05 to 0.2 and minimum DP threshold from 25 to 100 resulted in stable somatic mutation counts and consequently increased robustness of the hereditary CRC-associated signatures. To overcome a potential limitation of selecting a set of fixed optimal filtering settings we generated progression graphs that illustrate the robustness of a tumour's signature profile across different filtering settings (**Figure 4a, 4b, 5a, 5b, 6a-f**), providing a visual confirmation of a single tumour's signature robustness when compared to presenting a tumour's profile as a "mutograph" (**Figure 1**). Although this approach demands increased computational requirements, this may resolve ambiguous cases, and reduces the likelihood of artefacts arising due to signature calculation at a single, potentially arbitrary, filtering threshold.

This study has some limitations. Only a single biallelic *NTHL1* carrier was included in the study. This CRC showed relatively high levels of SBS23, SBS30 and SBS42. SBS30 and SBS23 were previously identified in tumours from biallelic *NTHL1* carriers adding support to our findings [7]. SBS42 was not identified previously in this context as it was not part of the

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

previous version of signatures and therefore could not be reported. Doublet signatures were excluded from this study due to low numbers in WES data, and high reconstruction error. This study used pre-existing COSMIC signature definitions from which to generate mutational signatures for each CRC, rather than creating novel signatures specific for our cohort. The advantage of this approach is the existing associated phenotypes provided for the COSMIC definitions, but unrelated signatures are included in the calculation, which may result in reporting spurious signatures. Analysis of an independent set of syndromic and non-syndromic CRCs is needed to validate our findings.

CONCLUSIONS

Understanding the somatic mutational landscape can enhance precision oncology by enabling us to pinpoint biomarkers relevant to targeted treatment [39]. As access to tumour sequencing increases, the opportunity to derive mutational signatures, at minimal additional cost, can provide improved diagnostic yield and guide therapeutic options [23]. We have shown that CRCs from biallelic *MUTYH* carriers and from biallelic *NTHL1* carriers exhibit mutational signature profiles that distinguish them from CRCs from non-carriers, evidenced by the combination of SBS18 and SBS36, and by SBS30, respectively. These distinct mutational signature profiles have the potential to aid in rare variant classification for these genes. Furthermore, we identified a novel association between both the proportion of SBS18 and SBS36 and the TMB for specific *MUTYH* PVs. Our results highlight the additional utility of ID-derived signatures for determining defective DNA MMR where the combination of ID2 and ID7 with SBS1 and SBS15 effectively differentiated Lynch syndrome CRCs from MMRp CRCs. We have shown that mutational signatures generated from WES of FFPE-CRCs can effectively identify carriers of hereditary CRC and polyposis syndromes and provides a functional assay to aid in the clinical genetics of CRC. Further work is needed to distinguish germline MMR-deficiency from somatic MMR-deficiency in order to improve the diagnosis of Lynch syndrome.

Funding/Support: Funding by a National Health and Medical Research Council of Australia (NHMRC) project grant 1125269 (PI- Daniel Buchanan), supported the design, analysis and

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

interpretation of data. DDB is supported by a NHMRC R.D. Wright Career Development Fellowship and funding from the University of Melbourne Research at Melbourne Accelerator Program (R@MAP). PG is supported by an Australian Government Research Training Program Scholarship. BP is supported by a Victorian Health and Medical Research Fellowship

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA167551 and through a cooperative agreement with the Australasian Colorectal Cancer Family Registry (NCI/NIH U01 CA074778 and U01/U24 CA097735) and by the Victorian Cancer Registry, Australia. This research was performed under CCFR approved projects C-AU-0818-01, C-AU-1014-02, C-AU-0312-01, C-AU-1013-02.

Financial disclosure: DDB served as a consultant on the Tumour Agnostic (dMMR) Advisory Board of Merck Sharp and Dohme in 2017 and 2018 for Pembrolizumab.

Acknowledgments: We thank members of the Colorectal Oncogenomics Group for their support of this manuscript. We thank the participants and staff from the Colon-CFR in particular, Maggie Angelakos, Samantha Fox and Allyson Templeton for their support of this manuscript. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. We thank Melbourne Bioinformatics for their support of this work.

"The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Colon Cancer Family Registry (Colon-CFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the Colon-CFR."

Author contributions: DDB, PG, MC and IMW conceived the original study concept and design and designed the analysis. CR, FAM, IMW, AKW, JLH, MAJ contributed to the acquisition of study data. The sample curation and laboratory testing was performed by MC,

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

RW, RH, JJ, SP, SJ, JC. PG, BJP, KM implemented the bioinformatics analysis pipeline software. PG and DDB prepared the manuscript. All authors provided critical revisions to the manuscript for important intellectual content and have read and approved of the final manuscript.

Competing interests: The authors declare no competing interests.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figures

Figure 1. Measured (a) SBS and (b) ID tumour mutational signatures (TMS) according to the COSMIC v3 signatures across the 60 CRCs tested by WES, grouped by subtype. Signatures with values below 5% in all samples are excluded.

Figure 2. Heatmap showing the mean ± SD of each tumour mutational signature (TMS) that was significantly associated with a hereditary CRC syndrome when compared with the study matched MMRd and MMRp controls or when compared with the TCGA MMRd and MMRp controls. Associated signatures include both previously reported associations and significant novel associations (SBS23 and SBS42) discovered in this study. Red boxes highlight syndromes of interest and associated mutational signatures. SBS14, SBS26, and SBS44 have been reported to be associated with MMRd but were not significantly associated with Lynch syndrome related CRC in this study.

Figure 3. The investigation of gene and PV specific tumour mutational signatures (TMS): **(a)** The contribution of the biallelic MUTYH-associated signatures SBS18 and SBS36 to each of the 10 biallelic *MUTYH* related CRCs by their underlying germline PVs, **(b)** The mean of SBS18 was significantly higher in the 3 CRCs from homozygous p.Gly396Asp carriers (blue) compared with the 7 CRCs from homozygous or compound heterozygous carriers of the p.Tyr179Cys PV (green; p=0.015). Conversely, the mean of SBS36 was significantly higher in the 7 CRCs from the p.Tyr179Cys carriers (green) compared with the 3 CRCs from the p.Gly396Asp carriers (blue; p=0.0012); the combination of SBS18 and SBS36 effectively separates the biallelic *MUTYH* CRCs from both the MMRp and MMRd controls, and from the TCGA MMRp and MMRd controls, **(c)** The contribution of the Lynch syndrome-associated signatures for each of the 14 CRCs from the Lynch syndrome carriers by the germline MMR gene mutated, **(d)** The distribution of each of the Lynch syndrome-associated signatures for the 7 CRCs from *MLH1* carriers, 3 CRCs from *MSH2* carriers, and 4 CRCs from the *MSH6* carriers. No significant difference for any of the Lynch syndrome-associated signatures by MMR gene were observed after adjusting for multiple testing. The combination of ID2, ID7, SBS15, and SBS1 effectively separates Lynch syndrome CRCs from both the study matched MMRp controls and the TCGA MMRp controls.

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 4. Tumour mutational signature (TMS) assessment for *MUTYH*-biallelic CRCs. For a representative *MUTYH*-biallelic (M05), with minimum DP fixed at 50bp, *MUTYH*-associated polyposis specific signatures SBS18 and SBS36 are dominant at VAF thresholds between 0.075 and 0.25 (a), while SBS18 and SBS36 remain dominant at DP thresholds above 25bp when the VAF threshold is fixed at 0.1 (b). The sum of SBS18 and SBS36 varies depending on VAF, separating *MUTYH*-biallelic CRCs (light green) compared to monoallelic *MUTYH* (pink), MMRp controls (brown), and other CRCs (blue) when VAF lies between 0.075 and 0.175 (c) and when DP>25 (d). Calculated 99% (green) and 1% (red) probabilities reflect the likelihood that a sample above this level is *MUTYH*-biallelic.

Figure 5. Tumour mutational signature (TMS) assessment for the biallelic *NTHL1* CRC (N01) showing: (a) the proportions of each SBS-derived signature across changes in variant allele fraction (VAF) filtering in conjunction with the measured reconstruction error and number of somatic mutations, where SBS30 was stable between VAFs of 0.05 and 0.35, (b) the proportions of each SBS-derived signature across changes in DP in conjunction with the measured reconstruction error and number of somatic mutations, where SBS30 was stable at DP of >50 bp. The proportion of SBS30 across changes in VAF filtering (c) and DP (d) in the biallelic *NTHL1* CRC (dark green), monoallelic *NTHL1* CRCs (pink), MMRp controls (purple), and other CRCs (brown) where the proportion of SBS30 signature that gave 99% (green) and 1% (red) probability of observing a biallelic *NTHL1* CRC from all other CRCs studied.

Figure 6. Assessment of SBS and ID tumour mutational signatures (TMS) while varying the VAF threshold, for samples L01 (a, b), L04 (c, d), and L10 (e, f), representing tumours with germline PVs in *MSH2*, *MSH6*, and *MLH1* respectively. In all cases, ID signatures that have previously been associated with MMRd dominate at most VAF thresholds (b, d, f), while relevant SBS signatures are also present but less dominant, particularly at highly stringent settings (a, c, e).

Figure 7. The combination of SBS1, SBS15, ID2 and ID7 effectively separate Lynch syndrome CRCs (L01-L14) from all other MMRp samples at low AF thresholds (<0.125) and most DP thresholds (<150 bp) (a, b), but when separating Lynch syndrome CRCs from

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

MMRd controls (K01 to K09), mutational signatures do not effectively separate the two groups (c, d) at any filtering settings.

Bibliography

1 Ferlay J, Colombet M, Soerjomataram I, *et al.* Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer* 2019;**144**:1941–53. doi:10.1002/ijc.31937

2 Al-Tassan N, Chmiel NH, Maynard J, *et al.* Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* 2002;**30**:227–32. doi:10.1038/ng828

3 Weren RDA, Ligtenberg MJL, Kets CM, *et al.* A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet* 2015;**47**:668–71. doi:10.1038/ng.3287

4 Lorans M, Dow E, Macrae FA, *et al.* Update on hereditary colorectal cancer: improving the clinical utility of multigene panel testing. *Clin Colorectal Cancer* 2018;**17**:e293–305. doi:10.1016/j.clcc.2018.01.001

5 Win AK, Parry S, Parry B, *et al.* Risk of metachronous colon cancer following surgery for rectal cancer in mismatch repair gene mutation carriers. *Ann Surg Oncol* 2013;**20**:1829–36. doi:10.1245/s10434-012-2858-5

6 Win AK, Lindor NM, Young JP, *et al.* Risks of primary extracolonic cancers following colorectal cancer in lynch syndrome. *J Natl Cancer Inst* 2012;**104**:1363–72. doi:10.1093/jnci/djs351

7 Grolleman JE, de Voer RM, Elsayed FA, *et al.* Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* 2019;**35**:256–266.e5. doi:10.1016/j.ccell.2018.12.011

8 Win AK, Young JP, Lindor NM, *et al.* Colorectal and other cancer risks for carriers and noncarriers from families with a DNA mismatch repair gene mutation: a prospective cohort study. *J Clin Oncol* 2012;**30**:958–64. doi:10.1200/JCO.2011.39.5590

9 Win AK, Reece JC, Dowty JG, *et al.* Risk of extracolonic cancers for people with

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

biallelic and monoallelic mutations in MUTYH. *Int J Cancer* 2016;**139**:1557–63. doi:10.1002/ijc.30197

10 Hegde M, Ferber M, Mao R, *et al.* ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *Genet Med* 2014;**16**:101–16. doi:10.1038/gim.2013.166

11 Thompson BA, Goldgar DE, Paterson C, *et al.* A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. *Hum Mutat* 2013;**34**:200–9. doi:10.1002/humu.22213

12 Herman JG, Umar A, Polyak K, *et al.* Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci USA* 1998;**95**:6870–5. doi:10.1073/pnas.95.12.6870

13 Haraldsdottir S, Hampel H, Tomsic J, *et al.* Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology* 2014;**147**:1308–1316.e1. doi:10.1053/j.gastro.2014.08.041

14 Cleary SP, Cotterchio M, Jenkins MA, *et al.* Germline MutY human homologue mutations and colorectal cancer: a multisite case-control study. *Gastroenterology* 2009;**136**:1251–60. doi:10.1053/j.gastro.2008.12.050

15 Thompson BA, Spurdle AB, Plazzer J-P, *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. doi:10.1038/ng.2854

16 Ciriello G, Miller ML, Aksoy BA, *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* 2013;**45**:1127–33. doi:10.1038/ng.2762

17 Goncearenco A, Rager SL, Li M, *et al.* Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Res* 2017;**45**:W514–22. doi:10.1093/nar/gkx367

18 Alexandrov LB, Nik-Zainal S, Wedge DC, *et al.* Signatures of mutational processes in human cancer. *Nature* 2013;**500**:415–21. doi:10.1038/nature12477

19 Alexandrov L, Kim J, Haradhvala NJ, *et al.* The repertoire of mutational signatures in human cancer. *BioRxiv* Published Online First: 15 May 2018. doi:10.1101/322859

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

20 Wellcome Sanger Institute. COSMIC Signatures of Mutational Processes in Human Cancer. Signatures of Mutational Processes in Human Cancer. 2019.https://cancer.sanger.ac.uk/cosmic/signatures (accessed 31 May2019).

21 Pilati C, Shinde J, Alexandrov LB, *et al.* Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol* 2017;**242**:10–5. doi:10.1002/path.4880

22 Viel A, Bruselles A, Meccia E, *et al.* A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* 2017;**20**:39–49. doi:10.1016/j.ebiom.2017.04.022

23 Van Hoeck A, Tjoonk NH, van Boxtel R, *et al.* Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* 2019;**19**:457. doi:10.1186/s12885-019-5677-2

24 Buchanan DD, Clendenning M, Zhuoer L, *et al.* Lack of evidence for germline RNF43 mutations in patients with serrated polyposis syndrome from a large multinational study. *Gut* 2017;**66**:1170–2. doi:10.1136/gutjnl-2016-312773

25 Newcomb PA, Baron J, Cotterchio M, *et al.* Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* 2007;**16**:2331–43. doi:10.1158/1055-9965.EPI-07-0648

26 Jenkins MA, Win AK, Templeton AS, *et al.* Cohort profile: the colon cancer family registry cohort (CCFRC). *Int J Epidemiol* 2018;**47**:387–388i. doi:10.1093/ije/dyy006

27 Win AK, Cleary SP, Dowty JG, *et al.* Cancer risks for monoallelic MUTYH mutation carriers with a family history of colorectal cancer. *Int J Cancer* 2011;**129**:2256–62. doi:10.1002/ijc.25870

28 Buchanan DD, Clendenning M, Rosty C, *et al.* Tumor testing to identify lynch syndrome in two Australian colorectal cancer cohorts. *J Gastroenterol Hepatol* 2017;**32**:427–38. doi:10.1111/jgh.13468

29 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7. doi:10.1038/nature11252

30 Cortes-Ciriano I, Lee S, Park W-Y, *et al.* A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* 2017;**8**:15180. doi:10.1038/ncomms15180

medRxiv preprint doi: https://doi.org/10.1101/2019.12.11.19014597; this version posted March 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

31 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. doi:10.1186/s13059-016-0893-4

32 Webb AR. *Statistical Pattern Recognition.* 2nd ed. West Sussex, England: : Wiley 2002.

33 Grubbs FE. Sample Criteria for Testing Outlying Observations. *Ann Math Statist* 1950;**21**:27–58. doi:10.1214/aoms/1177729885

34 VCV000230638.1 - ClinVar - NCBI. ClinVar - Genomic variation as it relates to human health. 2019.https://www.ncbi.nlm.nih.gov/clinvar/variation/230638/ (accessed 11 Dec2019).

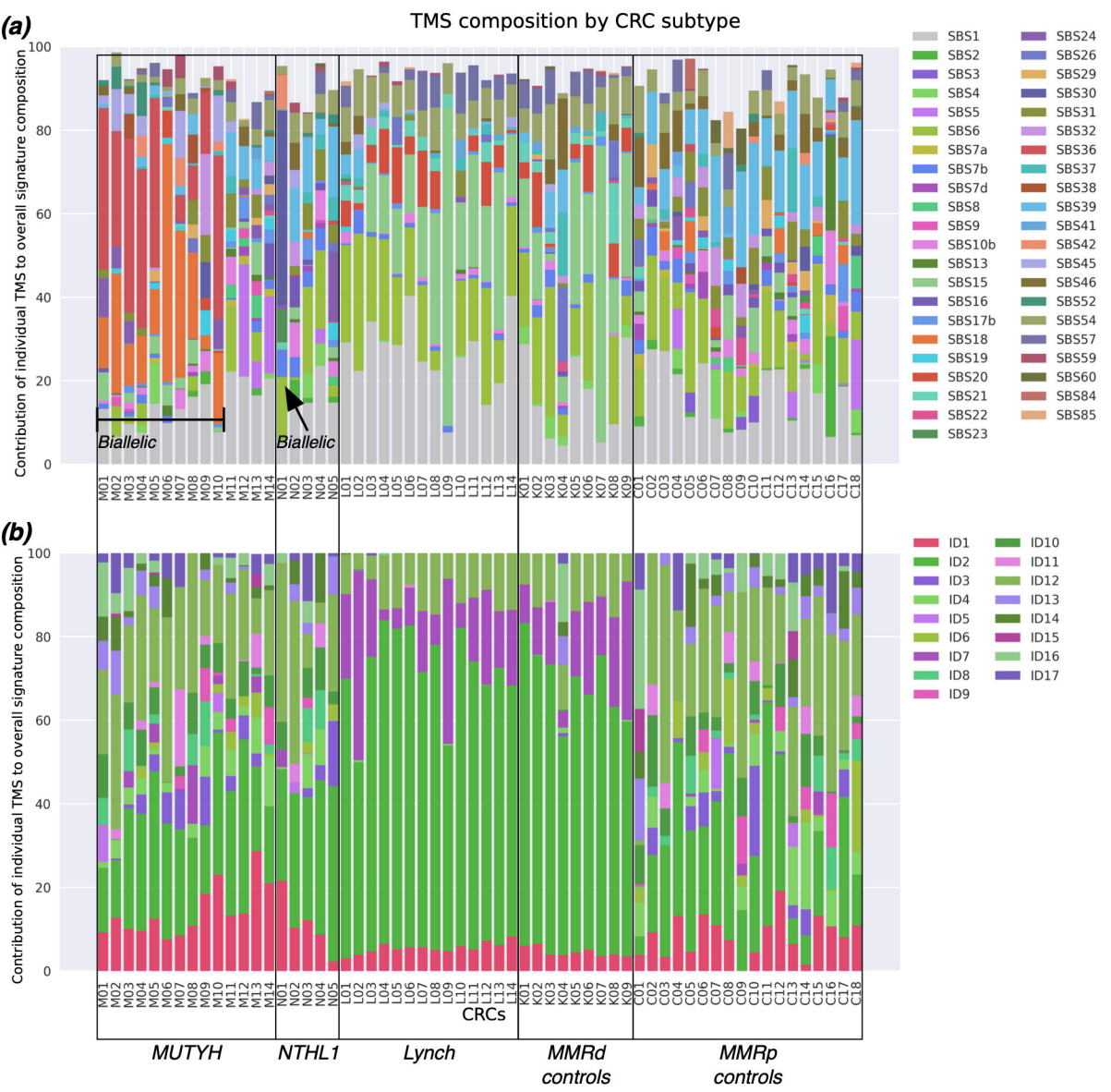
35 Møller P, Seppälä TT, Bernstein I, *et al.* Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database. *Gut* 2018;**67**:1306–16. doi:10.1136/gutjnl-2017-314057

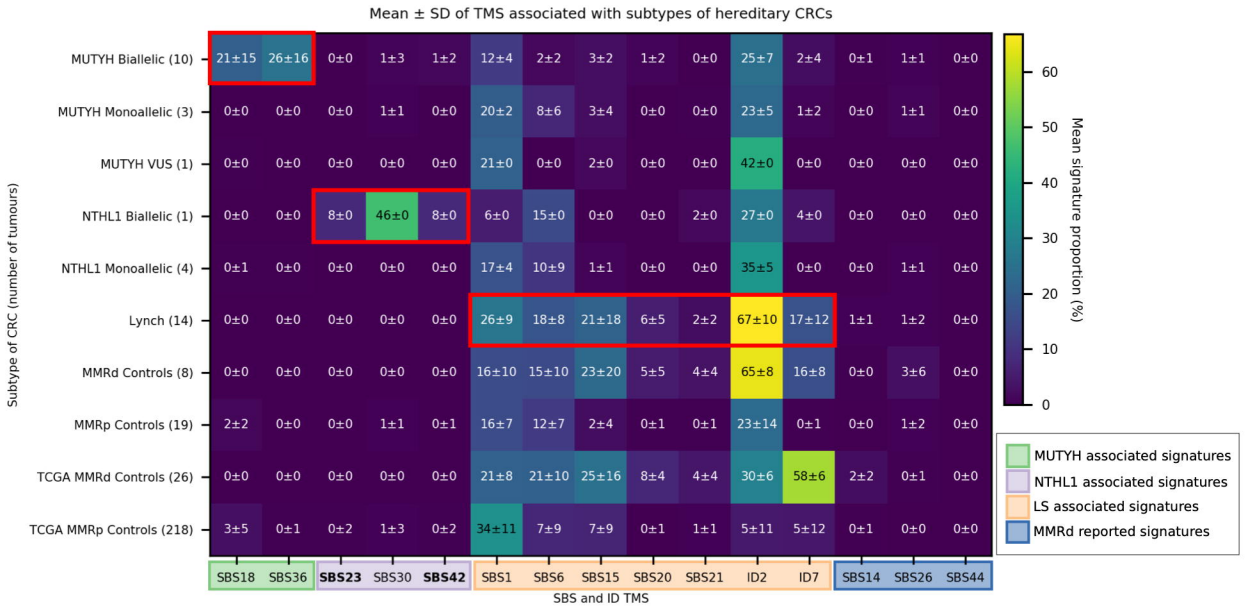
36 Drost J, van Boxtel R, Blokzijl F, *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* 2017;**358**:234–8. doi:10.1126/science.aao3130

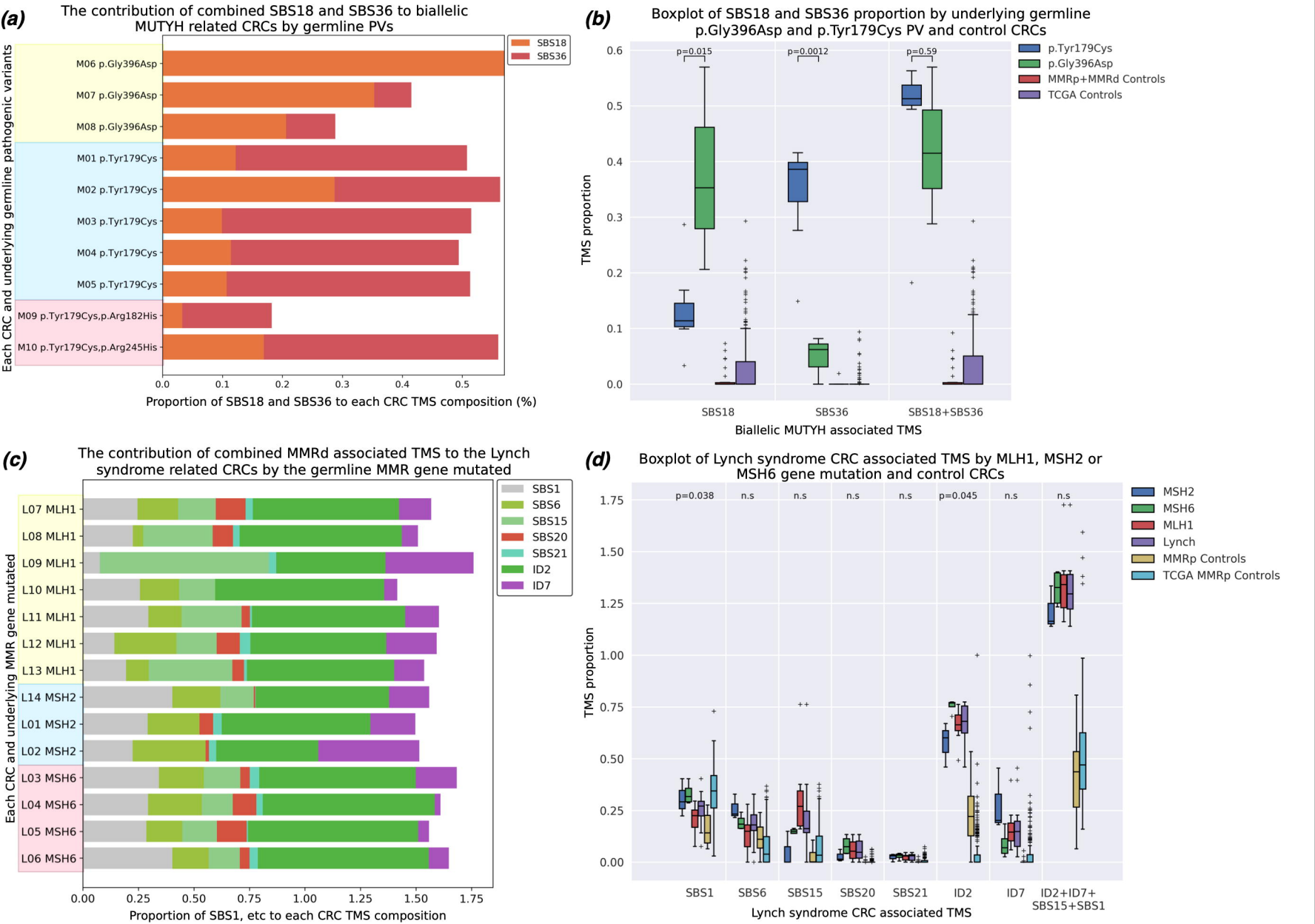
37 Win AK, Dowty JG, Cleary SP, *et al.* Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterology* 2014;**146**:1208–11.e1. doi:10.1053/j.gastro.2014.01.022

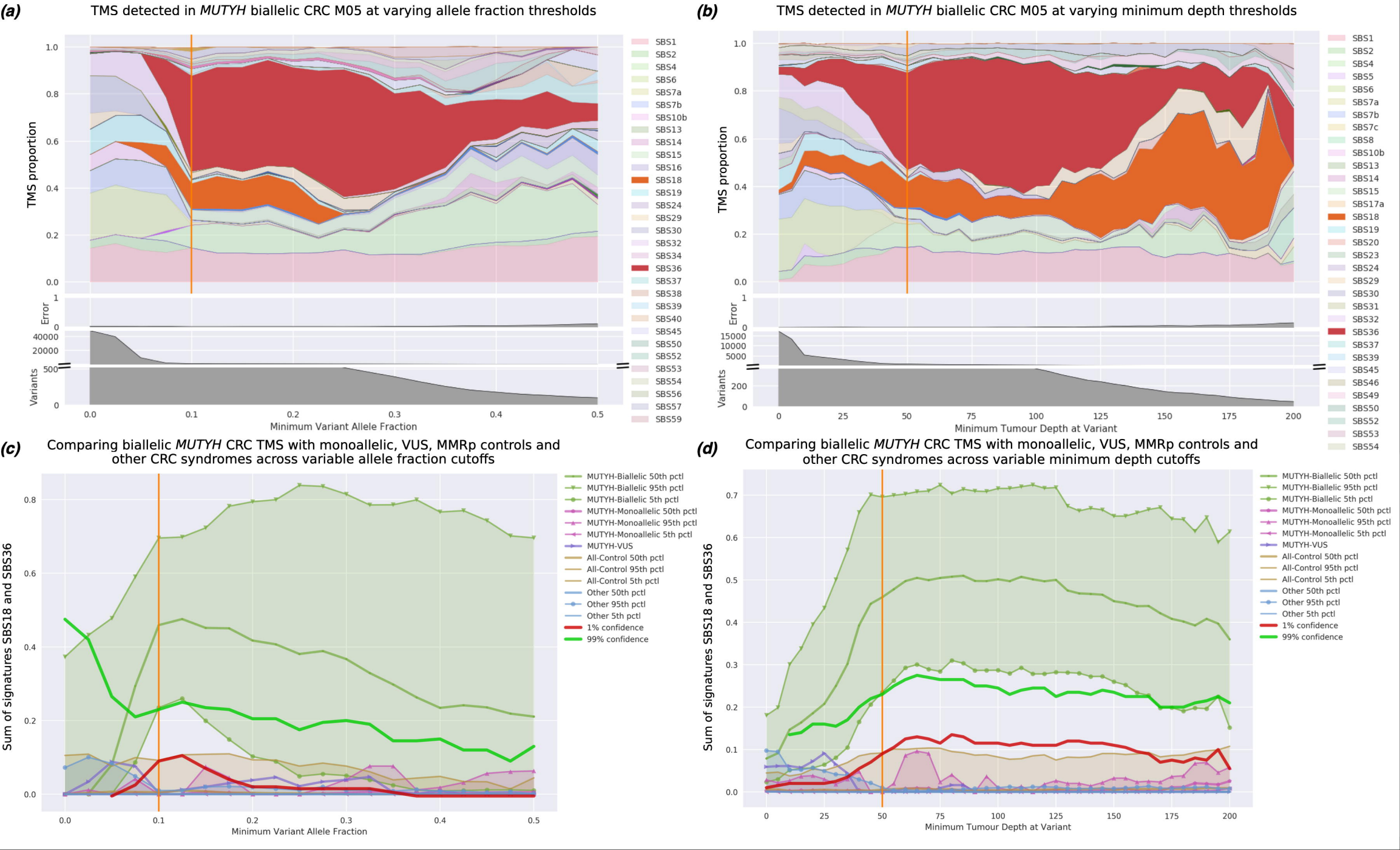
38 Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. *Science* 1993;**260**:816–9. doi:10.1126/science.8484122

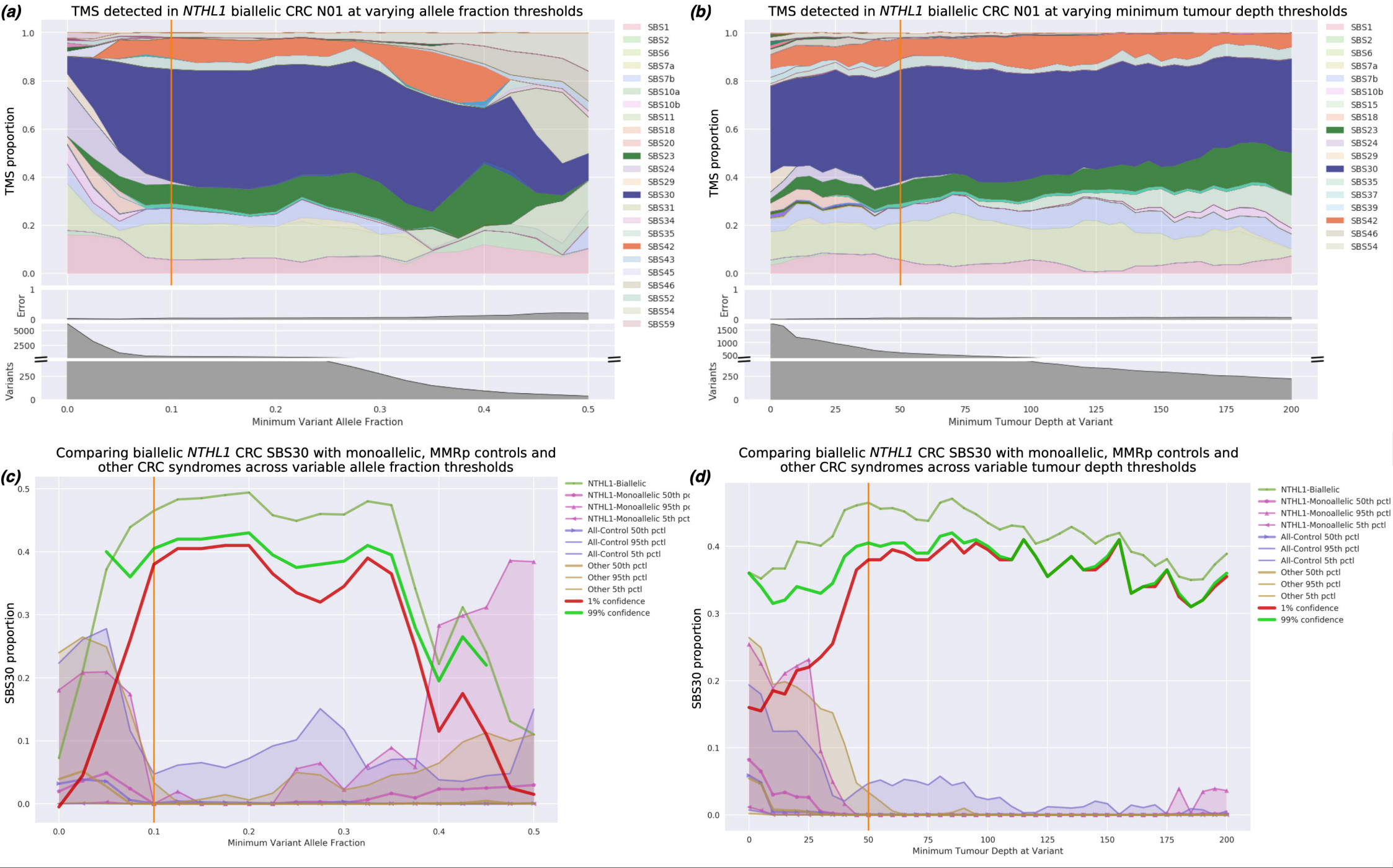
39 Ma J, Setton J, Lee NY, *et al.* The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat Commun* 2018;**9**:3292. doi:10.1038/s41467-018-05228-y

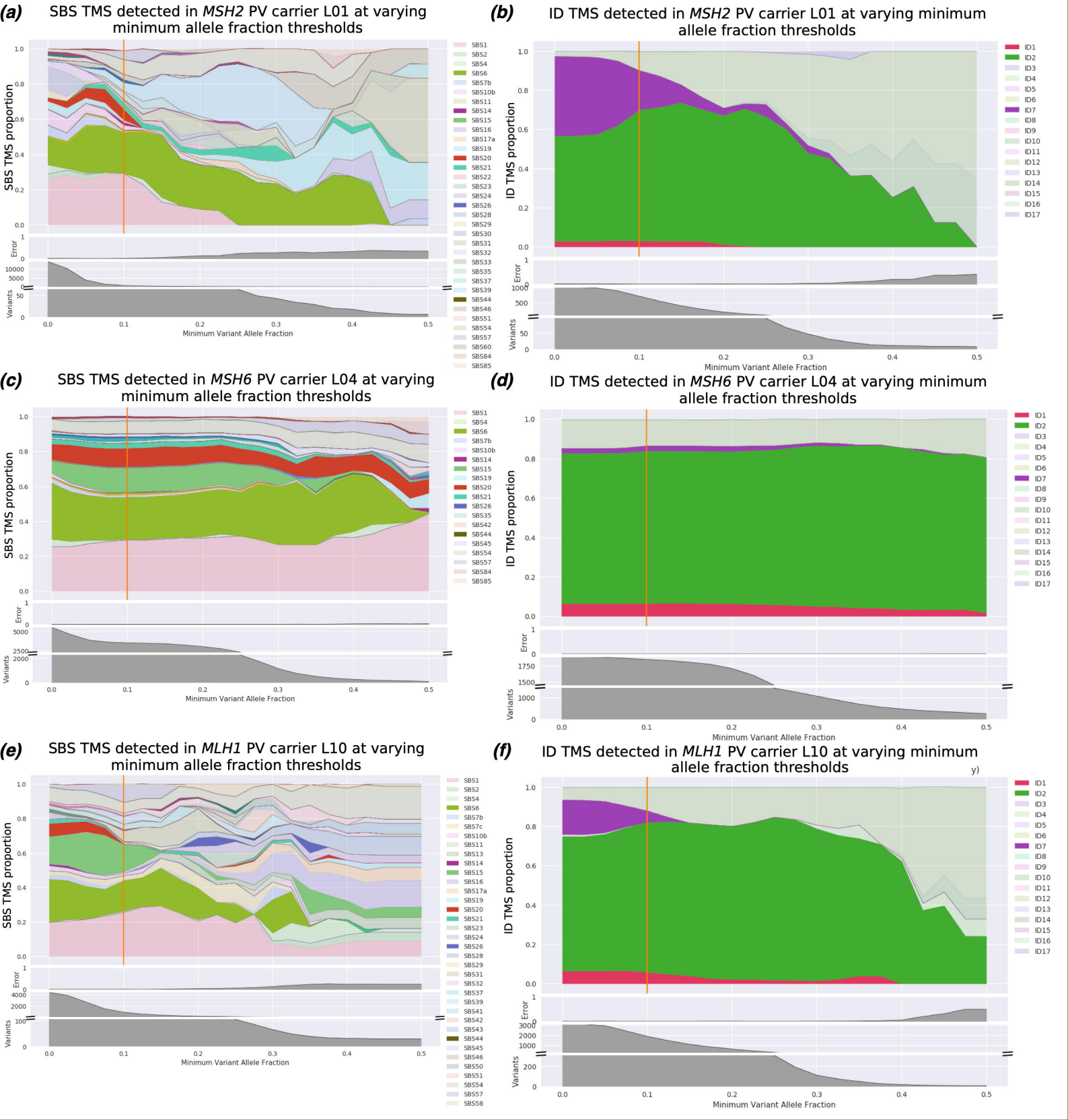

(a)
TMS composition by CRC subtype
Contribution of individual TMS to overall signature composition
Biallelic
Biallelic
SBS1
SBS2
SBS3
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7d
SBS8
SBS9
SBS10b
SBS13
SBS15
SBS16
SBS17b
SBS18
SBS19
SBS20
SBS21
SBS22
SBS23
SBS24
SBS26
SBS29
SBS30
SBS31
SBS32
SBS36
SBS37
SBS38
SBS39
SBS41
SBS42
SBS45
SBS46
SBS52
SBS54
SBS57
SBS59
SBS60
SBS84
SBS85
(b)
Contribution of individual TMS to overall signature composition
ID1
ID2
ID3
ID4
ID5
ID6
ID7
ID8
ID9
ID10
ID11
ID12
ID13
ID14
ID15
ID16
ID17
CRCs
MUTYH
NTHL1
Lynch
MMRd controls
MMRp controls

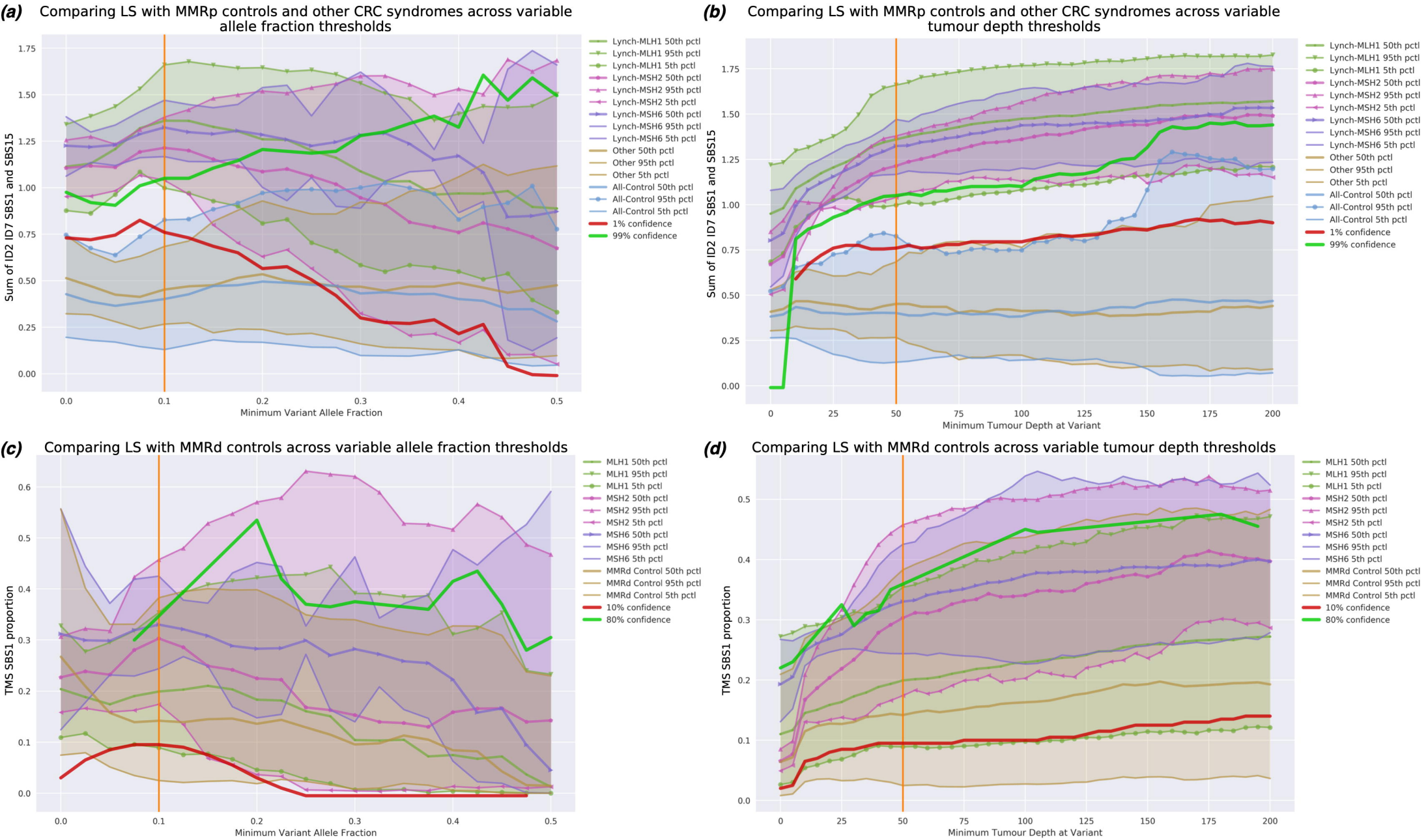

Mean ± SD of TMS associated with subtypes of hereditary CRCs
Subtype of CRC (number of tumours)
MUTYH Biallelic (10)
MUTYH Monoallelic (3)
MUTYH VUS (1)
NTHL1 Biallelic (1)
NTHL1 Monoallelic (4)
Lynch (14)
MMRd Controls (8)
MMRp Controls (19)
TCGA MMRd Controls (26)
TCGA MMRp Controls (218)
SBS18 SBS36 SBS23 SBS30 SBS42 SBS1 SBS6 SBS15 SBS20 SBS21 ID2 ID7 SBS14 SBS26 SBS44
SBS and ID TMS
Mean signature proportion (%)
0 10 20 30 40 50 60
MUTYH associated signatures
NTHL1 associated signatures
LS associated signatures
MMRd reported signatures


(a)
The contribution of combined SBS18 and SBS36 to biallelic MUTYH related CRCs by germline PVs
SBS18
SBS36
Each CRC and underlying germline pathogenic variants
M06 p.Gly396Asp
M07 p.Gly396Asp
M08 p.Gly396Asp
M01 p.Tyr179Cys
M02 p.Tyr179Cys
M03 p.Tyr179Cys
M04 p.Tyr179Cys
M05 p.Tyr179Cys
M09 p.Tyr179Cys,p.Arg182His
M10 p.Tyr179Cys,p.Arg245His
0.0
0.1
0.2
0.3
0.4
0.5
Proportion of SBS18 and SBS36 to each CRC TMS composition (%)
(b)
Boxplot of SBS18 and SBS36 proportion by underlying germline p.Gly396Asp and p.Tyr179Cys PV and control CRCs
p=0.015
p=0.0012
p=0.59
p.Tyr179Cys
p.Gly396Asp
MMRp+MMRd Controls
TCGA Controls
TMS proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
SBS18
SBS36
SBS18+SBS36
Biallelic MUTYH associated TMS
(c)
The contribution of combined MMRd associated TMS to the Lynch syndrome related CRCs by the germline MMR gene mutated
SBS1
SBS6
SBS15
SBS20
SBS21
ID2
ID7
Each CRC and underlying MMR gene mutated
L07 MLH1
L08 MLH1
L09 MLH1
L10 MLH1
L11 MLH1
L12 MLH1
L13 MLH1
L14 MSH2
L01 MSH2
L02 MSH2
L03 MSH6
L04 MSH6
L05 MSH6
L06 MSH6
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Proportion of SBS1, etc to each CRC TMS composition
(d)
Boxplot of Lynch syndrome CRC associated TMS by MLH1, MSH2 or MSH6 gene mutation and control CRCs
p=0.038
n.s
n.s
n.s
n.s
p=0.045
n.s
n.s
MSH2
MSH6
MLH1
Lynch
MMRp Controls
TCGA MMRp Controls
TMS proportion
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
SBS1
SBS6
SBS15
SBS20
SBS21
ID2
ID7
ID2+ID7+ SBS15+SBS1
Lynch syndrome CRC associated TMS


(a)
TMS detected in MUTYH biallelic CRC M05 at varying allele fraction thresholds
TMS proportion
Error
Variants
Minimum Variant Allele Fraction
SBS1
SBS2
SBS4
SBS6
SBS7a
SBS7b
SBS10b
SBS13
SBS14
SBS15
SBS16
SBS18
SBS19
SBS24
SBS29
SBS30
SBS32
SBS34
SBS36
SBS37
SBS38
SBS39
SBS40
SBS45
SBS50
SBS52
SBS53
SBS54
SBS56
SBS57
SBS59
(b)
TMS detected in MUTYH biallelic CRC M05 at varying minimum depth thresholds
TMS proportion
Error
Variants
Minimum Tumour Depth at Variant
SBS1
SBS2
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7c
SBS8
SBS10b
SBS13
SBS14
SBS15
SBS17a
SBS18
SBS19
SBS20
SBS23
SBS24
SBS29
SBS30
SBS31
SBS32
SBS36
SBS37
SBS39
SBS45
SBS46
SBS49
SBS50
SBS52
SBS53
SBS54
(c)
Comparing biallelic MUTYH CRC TMS with monoallelic, VUS, MMRp controls and other CRC syndromes across variable allele fraction cutoffs
Sum of signatures SBS18 and SBS36
Minimum Variant Allele Fraction
MUTYH-Biallelic 50th pctl
MUTYH-Biallelic 95th pctl
MUTYH-Biallelic 5th pctl
MUTYH-Monoallelic 50th pctl
MUTYH-Monoallelic 95th pctl
MUTYH-Monoallelic 5th pctl
MUTYH-VUS
All-Control 50th pctl
All-Control 95th pctl
All-Control 5th pctl
Other 50th pctl
Other 95th pctl
Other 5th pctl
1% confidence
99% confidence
(d)
Comparing biallelic MUTYH CRC TMS with monoallelic, VUS, MMRp controls and other CRC syndromes across variable minimum depth cutoffs
Sum of signatures SBS18 and SBS36
Minimum Tumour Depth at Variant
MUTYH-Biallelic 50th pctl
MUTYH-Biallelic 95th pctl
MUTYH-Biallelic 5th pctl
MUTYH-Monoallelic 50th pctl
MUTYH-Monoallelic 95th pctl
MUTYH-Monoallelic 5th pctl
MUTYH-VUS
All-Control 50th pctl
All-Control 95th pctl
All-Control 5th pctl
Other 50th pctl
Other 95th pctl
Other 5th pctl
1% confidence
99% confidence


(a)
TMS detected in NTHL1 biallelic CRC N01 at varying allele fraction thresholds
TMS proportion
Error
Variants
Minimum Variant Allele Fraction
SBS1
SBS2
SBS6
SBS7a
SBS7b
SBS10a
SBS10b
SBS11
SBS18
SBS20
SBS23
SBS24
SBS29
SBS30
SBS31
SBS34
SBS35
SBS42
SBS43
SBS45
SBS46
SBS52
SBS54
SBS59
(b)
TMS detected in NTHL1 biallelic CRC N01 at varying minimum tumour depth thresholds
TMS proportion
Error
Variants
Minimum Tumour Depth at Variant
SBS1
SBS2
SBS6
SBS7a
SBS7b
SBS10b
SBS15
SBS18
SBS23
SBS24
SBS29
SBS30
SBS35
SBS37
SBS39
SBS42
SBS46
SBS54
(c)
Comparing biallelic NTHL1 CRC SBS30 with monoallelic, MMRp controls and other CRC syndromes across variable allele fraction thresholds
SBS30 proportion
Minimum Variant Allele Fraction
NTHL1-Biallelic
NTHL1-Monoallelic 50th pc
NTHL1-Monoallelic 95th pc
NTHL1-Monoallelic 5th pct
All-Control 50th pctl
All-Control 95th pctl
All-Control 5th pctl
Other 50th pctl
Other 95th pctl
Other 5th pctl
1% confidence
99% confidence
(d)
Comparing biallelic NTHL1 CRC SBS30 with monoallelic, MMRp controls and other CRC syndromes across variable tumour depth thresholds
SBS30 proportion
Minimum Tumour Depth at Variant
NTHL1-Biallelic
NTHL1-Monoallelic 50th pctl
NTHL1-Monoallelic 95th pctl
NTHL1-Monoallelic 5th pctl
All-Control 50th pctl
All-Control 95th pctl
All-Control 5th pctl
Other 50th pctl
Other 95th pctl
Other 5th pctl
1% confidence
99% confidence


(a)
SBS TMS detected in MSH2 PV carrier L01 at varying minimum allele fraction thresholds
SBS TMS proportion
Error
Variants
Minimum Variant Allele Fraction
(b)
ID TMS detected in MSH2 PV carrier L01 at varying minimum allele fraction thresholds
ID TMS proportion
Error
Variants
Minimum Variant Allele Fraction
(c)
SBS TMS detected in MSH6 PV carrier L04 at varying minimum allele fraction thresholds
SBS TMS proportion
Error
Variants
Minimum Variant Allele Fraction
(d)
ID TMS detected in MSH6 PV carrier L04 at varying minimum allele fraction thresholds
ID TMS proportion
Error
Variants
Minimum Variant Allele Fraction
(e)
SBS TMS detected in MLH1 PV carrier L10 at varying minimum allele fraction thresholds
SBS TMS proportion
Error
Variants
Minimum Variant Allele Fraction
(f)
ID TMS detected in MLH1 PV carrier L10 at varying minimum allele fraction thresholds
ID TMS proportion
Error
Variants
Minimum Variant Allele Fraction


(a)
Comparing LS with MMRp controls and other CRC syndromes across variable allele fraction thresholds
Sum of ID2 ID7 SBS1 and SBS15
Minimum Variant Allele Fraction
(b)
Comparing LS with MMRp controls and other CRC syndromes across variable tumour depth thresholds
Sum of ID2 ID7 SBS1 and SBS15
Minimum Tumour Depth at Variant
Lynch-MLH1 50th pctl
Lynch-MLH1 95th pctl
Lynch-MLH1 5th pctl
Lynch-MSH2 50th pctl
Lynch-MSH2 95th pctl
Lynch-MSH2 5th pctl
Lynch-MSH6 50th pctl
Lynch-MSH6 95th pctl
Lynch-MSH6 5th pctl
Other 50th pctl
Other 95th pctl
Other 5th pctl
All-Control 50th pctl
All-Control 95th pctl
All-Control 5th pctl
1% confidence
99% confidence
(c)
Comparing LS with MMRd controls across variable allele fraction thresholds
TMS SBS1 proportion
Minimum Variant Allele Fraction
(d)
Comparing LS with MMRd controls across variable tumour depth thresholds
TMS SBS1 proportion
Minimum Tumour Depth at Variant
MLH1 50th pctl
MLH1 95th pctl
MLH1 5th pctl
MSH2 50th pctl
MSH2 95th pctl
MSH2 5th pctl
MSH6 50th pctl
MSH6 95th pctl
MSH6 5th pctl
MMRd Control 50th pctl
MMRd Control 95th pctl
MMRd Control 5th pctl
10% confidence
80% confidence