Framing the potential of public frameshift peptides as immunotherapy targets in colon cancer

Ide T. Spaanderman[1,2], Fleur S. Peters[2], Aldo Jongejan[3], Egbert J.W. Redeker[4], Cornelis J.A. Punt[1], Adriaan D. Bins[1,2*]

1. Dept. of Medical Oncology, Amsterdam University Medical Centers, University of Amsterdam, Cancer Center Amsterdam, The Netherlands
2. Dept. of Experimental Immunology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam Infection & Immunity, The Netherlands
3. Dept. of Bio-informatics, Amsterdam University Medical Centers, University of Amsterdam, The Netherlands
4. Dept. of Clinical Genetics, Amsterdam University Medical Centers, University of Amsterdam, The Netherlands

*a.d.bins@amsterdamumc.nl

Abstract

Approximately 15% of Colon Cancers are Microsatellite Instable (MSI). Frameshift Peptides (FPs) formed in MSI Colon Cancer are potential targets for immunotherapeutic strategies. Here we comprehensively characterize the mutational landscape of 71 MSI Colon Cancer patients from the cancer genome atlas (TCGA-COAD). We confirm that the mutations in MSI Colon Cancers are frequently frameshift deletions (23% in MSI; 1% in microsatellite stable), We find that these mutations cluster at specific locations in the genome which are mutated in up to 41% of the patients. We filter these for an adequate variant allele frequency, a sufficient mean mRNA level and the formation of a Super Neo Open Reading Frame (SNORF). Finally, we check the influence of Nonsense Mediated Decay by comparing RNA and DNA sequencing results. Thereby we identify a set of 20 NMD-escaping Public FPs (PFPs) that cover over 90% of MSI Colon Cancer patients and 3 out of 4 Lynch patients in the TCGA-COAD. This underlines the potential for PFP directed immunotherapy, both in a therapeutic and a prophylactic setting.

Introduction

Protein products derived from mutated regions of DNA are a source of neo-antigens that can drive immune activation against cancer and convey susceptibility to immunotherapy[1–4]. Specifically, frameshift insertions and deletions (INDELS) are able to drive the anti-tumor immune response[5]. INDELS in coding DNA result in frameshifted RNA containing neo open reading frames (NORFS) that, once translated, give rise to completely non-self, out-of-frame protein products. These out-of-frame proteins can be degraded into peptides and presented in the context of multiple MHC alleles. Recognition of these MHC-peptide complexes by CD4+ or CD8+ T-cells initializes an anti-tumor immune response and leads to an influx of T cells into the tumor microenvironment[1,6,7], Longer NORFS (so called super NORFS: SNORFS), have a higher chance to be recognized as a neo-antigens[8]. Importantly, NORFS are subject to degradation on a RNA level by the Nonsense Mediated Decay (NMD) pathway[9]. In line with this, the presence of NMD escaping SNORFS has been shown to be a strong predictor of immunotherapy response and survival in multiple cancer types[8].

Compared to other mutation signatures, INDELS are overrepresented in the exome of microsatellite instable (MSI) tumors[5,10]. Microsatellites are regions in the DNA with a repeated nucleotide motif. These regions are prone to polymerase slippage and subsequently incorrect re-annealing during DNA synthesis. This leads to single stand insertions and deletions in the nascent DNA strand[11]. In healthy cells, the mismatch repair (MMR) system recognizes and repairs these damages. However, MSI tumors have a deficient MMR system (dMMR) due to germline and/or acquired somatic mutations in one or multiple MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *EPCAM*) or by hyper methylation of the *MSH1* promotor[12,13]. Approximately 15% of colon cancers are considered MSI[14]. These dMMR colon cancers have high numbers of mutations, specifically INDELS in microsatellites.

Here we comprehensively asses frameshift mutations in colon cancer samples sequenced within the genome atlas (TCGA-COAD)[15]. We focus on the TGCA-COAD series as this contains a large subset of MSI tumors. Using a novel in silico approach (**figure 1**), we establish that colon cancer frameshift mutations cluster to specific loci in the colon cancer exome. Based on DNA and RNA sequencing data we investigate whether these mutations lead to SNORFS and subsequently whether these SNORFS are affected by NMD. Finally, we identify a set of frequently occurring frameshift peptides (FPs), that we name public frameshift peptides (PFPs). These PFPs are potential targets for immunotherapy strategies in an adjuvant setting, as they obviate the need for

time-consuming personalization of vaccines. Moreover, PFPs can also be found in all four (Lynch) patients with a germline MMR mutation in the TCGA-COAD. This highlights the possibility for PFP specific vaccination in a prophylactic setting as well.

Results

Frameshift mutations cluster to specific genomic loci in MSI colon cancer.

In line with previous reports in literature 71 out of 461 (15%) of TCGA-COAD patients are MSI[14]. As expected, MSI colon cancer patients have a higher mutational burden than microsatellite stable (MSS) colon cancer patients (MSI: median = 2974, n = 71; MSS: median = 218, n = 272; $p < 0.0001$) **(figure 2a)** and a larger percentage of their mutations consists of deletions or insertions (MSI: SNP = 72%, INS = 5%, DEL = 23%'; MSS: SNP = 97%, INS 1%, DEL = 1%) **(figure 2b)**. Almost all of these INDELS are frameshift mutations (INS: Frameshift 99%; DEL: Frameshift 98%) **(figure 2c)**. Contrary to single nucleotide polymorphisms (SNPs), these INDELS are preferentially located in coding microsatellites (cMS), pinpointing the role of the dysfunctional MMR system in the mutation signature of MSI colon cancer (INS: median = 7; DEL: median = 7; SNP: median = 1, $p < 0.0001$) **(figure 2d)**. As expected, both SNP and INDEL mutations cluster to specific loci in the MSI colon cancer exome. Similar to the well documented BRAF SNP at chr7:140753336[16], it seems probable that the most prevalent INDEL mutations also play a role in tumorigenesis, **(figure 3a)**. 13 of the 14 most prevalent mutations are frameshift mutations located in a cMS. Strikingly, the most frequently occurring frameshift mutation is present in 46% of the patients. 123 frameshift mutations occur in more than 10% of the patients **(figure 3b)**. Notably 82.5% of the deletions that occur in more than 10% of patients are listed in the candidate cancer gene atlas and 42.5% are listed as potential drivers in colon cancer[17] **(supplementary table 1)**. Interestingly, when comparing the tendency to cluster among DEL, INS and SNPs mutations, clustering is only clearly present for DEL mutations (median percentage of patients with a similar specific mutation; DEL = 14.08%, INS = 4.23%, SNP = 4.23%, $p < 0.0001$) **(figure3c)**.

Nonsense mediated decay escape

In light of specific immunotherapy approaches it is important to further investigate the identified public frameshift mutations. As previously mentioned, NORFS derived from INDELS are potentially subject to NMD. NMD is triggered by the formation of a premature termination codon (PTC), that is prone to occur in out-of-frame NORFS[9]. If such a PTC is located at least 50 base pairs before the last exon-exon junction, the mRNA is considered sensitive to degradation by NMD. If the PTC is

63 located behind this point the transcript is considered NMD resistant[18]. In the MSI cancers of the
64 TGCA-COAD this concept holds true, as the mean z-score for mutated transcripts with a NMD
65 sensitive PTC is significantly lower than that of a mutated transcripts with a NMD resistant PTC
66 (median SENS = -0.4655; RES = -0.0657; $p < 0.0001$) **(figure 4a)**. Nonetheless the NMD effect is
67 limited in an absolute sense and clearly heterogeneous as a large portion of potential sensitive
68 mutations seem to escape NMD based on the Z-score of mRNA transcript levels. However, the fact
69 that a single specific mutation can be mapped to multiple transcripts (isoforms), might limit the
70 sensitivity for detection of transcript specific NMD signals.

71 To negate this problem, we checked the Variant Allele Frequency (VAF) for each mutated locus in
72 the DNA and RNA. The VAF is the percentage of mutated reads in all reads on that specific locus. We
73 calculated the difference between the DNA VAF and RNA VAF: the VAF-ratio. This transcript
74 independent method is a much more sensitive and accurate measurement of NMD **(Supplementary**
75 **figure 1).** On average, for NMD sensitive mutations the variant allele frequency (VAF) in RNA
76 sequencing data is decreased by 65.7% compared to the DNA VAF (median VAF-ratio 34.3%) n =
77 904), which is significantly more than the 40.1% decrease in NMD resistant mutations (median VAF-
78 ratio 59.9%, n = 936; $p < 0.0001$) **(figure 4b)**. This clearly confirms that NMD is an important factor in
79 the formation of neo-antigens from NORFS in MSI colon cancer. Nonetheless roughly one third of
80 NMD sensitive NORFS still escape NMD, with a VAF-ratio of 50% or more. This presents a clear
81 opportunity for PFP formation.

82 **NMD escape by public frameshift peptides**

83 Besides the potential NMD impact of PTCs, PTCs also can limit NORF length to a size that prohibits
84 presentation in the context of MHC. In the NORFS identified by us in MSI cancers of the TGCA-COAD,
85 a PTC occurs on average after 14 amino acids **(Supplementary figure 2)**. Therefore, we selected
86 potential PFP by discarding NORFS with a length below 10 amino acids. Hereby, we identified 90
87 SNORFS occurring in more than 5% of MSI patients. Assuming that a set of 20 PFPs is feasible to
88 incorporate in a multivalent vaccine, we discarded the 70 least frequently occurring SNORFS **(table**
89 **1)**. The remaining 20 frequently occurring SNORFS have a mean NORF length of 28 amino acids, a
90 mean VAF-ratio of 67.93% and a mean mRNA level of 2.30 (Log2 TPM) for the most abundantly
91 available corresponding transcript (isoform). Most importantly, at least one of these 20 SNORFS is
92 present in the exome of 93% of MSI samples in the TGCA-COAD. A limited set of 10 SNOFRS, as has
93 been used in a personalized RNA vaccination setting[19], would still cover 74.6% of these patients
94 **(Supplementary figure 3)**. This highlights the potential for non-personalized immunotherapy that

95 targets PFP in MSI colon cancer patients. Notably, 25% of these 20 SNORFS occur are listed as
96 potential drivers in colon cancer[17].

97 **PFP set covers germline mismatch repair deficient tumors in TCGA-COAD.**

98 Lynch patients carry a heterozygous germline mutation in an MMR gene, resulting in a 50-80% life-
99 time chance to develop colon cancer[20]. MSI colon cancer tumorigenesis in Lynch patients may be
100 different than that of sporadic MSI colon cancer, leading to expression of different PFPs. At the
101 same time, prophylactic immunotherapy targeting PFP may be beneficial for these patients.
102 Therefore we evaluated whether the PFP set may potentially be applicable to patients with a
103 germline MMR mutation. Using normal blood DNA sequencing data, we identified 4 MSI TCGA-
104 COAD patients (5.6%) with a pathogenic germline dMMR mutation. This percentage is in line with
105 earlier reports of germline dMMR mutations in literature[20]. In overlap with sporadic MSI colon
106 cancers, three out of four of these germline mutated dMMR tumors encodes at least 1 PFP from the
107 20-valent PFP-set. Colon Cancer samples with a germline MMR mutation from two patients only
108 have 1 mutation in one of the identified hotspots, but 1 germline mutated MSI sample even has 4
109 hotspot mutations **(table 1)**. This supports the potential of the PFP-set in a prophylactic vaccination
110 setting for Lynch patients.

111 Discussion

112 Based on DNA and RNA sequencing data from MSI colon cancer patients in the TGCA we identified
113 a subset of frameshift mutations that cluster to distinct cMS in the genome, produce SNORFS that
114 evade NMD, and have sufficient length for presentation in MHC context. Based on this, we
115 established a PFP-set of 20 frameshift peptides that cover 93% of individual MSI patients. Such a
116 set of PFPs has high potential for non-personalized cancer-specific immunotherapy strategies
117 directed against MSI colon cancer, especially in the adjuvant setting. In addition, this set of PFPs
118 also covers three out of four Lynch patients in the TGCA-COAD, which supports potential
119 application in a prophylactic vaccination setting for patients with Lynch syndrome. Cleary, as the
120 sporadic MSI colon cancer mutation signature may differ from that of Lynch patients, target
121 discovery for Lynch patients requires a larger patient cohort than the 4 patients identified in the
122 TCGA-COAD.

123 We did not predict MHC binding for the selected PFPs. Currently MHC class I binding predictions
124 tools only achieve approximately 40% sensitivity[21]. Although new methods have recently been
125 published[22], especially the prediction of MHC class II binding remains very challenging, whilst MHC

126 class II presented neo-antigens seem to drive anti-tumor immunity[22,23]. Furthermore, the above
127 average length of the PFPs and the fact that they are non-self on any position makes adequate
128 presentation within a naturally polymorphic MHC context more likely. Also, the concept of non-
129 personalized immunotherapy would be compromised by MHC matching. In this context it is
130 noteworthy that the loss of MHC-I on colon cancer cells could potentially limit the effectiveness of
131 immunotherapy that target PFPs in the same way it can limit the effectiveness of checkpoint
132 inhibition.[24]

133 Finally, since this work is based on in silico analysis of sequencing data we could not determine the
134 occurrence of PFPs on a protein level. Nonetheless, based on our extensive filtering for NMD
135 signals in RNA sequencing data we deem it likely that a majority of the selected SNORFS will be
136 translated and lead to presentable PFPs. Alsco it has been shown that a first so-called pioneering
137 round of translation, can already be adequate for neo-antigen formation[25]. In that regard our higher
138 VAF threshold and strict NMD evasion filtering might even hinder identifying relevant PFPs.
139 Furthermore the availability of RNA sequencing data also negated the need for novel NMD
140 prediction algorithms[26]. In order to assess the FPF presentation on a protein level we are currently
141 collecting patients samples in the ATAPEMBRO trial[27], which we aim to screen for specific PFP T-
142 cell immunoreactivity by IFN-gamma Elispot.

143 In conclusion we identified a set of 20 DEL mutation hotspots in coding DNA microsatellites that
144 give rise to adequately transcribed, NMD-escaping SNORFS and subsequently can produce PFPs.
145 Likely, these SNORFs are involved in colon cancer tumorigenesis. The 20 most common PFPs
146 cover 93% of MSI patients and 3 out of 4 patients that harbor a pathogenic germline MMR mutation
147 in the TCGA-COAD. Therefore, these PFPs are potential targets for immunotherapy strategies in
148 MSI colon cancers.

149 Methods

150 **Data acquisition, storage and computational analysis**

151 All DNA and RNA sequencing data and corresponding patient characteristics, microsatellite panel
152 and *MLH1* promotor methylation data was collected from the Cancer Genome Atlas Colon
153 Adenocarcinoma Database (TCGA-COAD)[15]. Access to private data was granted by the National
154 Cancer Institute Data Access Committee (NCIDAC) from the U.S. National Institute of Health trough
155 the database of Genotypes and Phenotypes (dbGaP) portal. According to dbGaP guidelines private
156 data was stored on the pre-approved internal Amsterdam UMC data-infrastructure. Computations

157 on large public TCGA datasets were performed on the SurfSara high performance computing cloud
158 (grant provided by the Dutch government).

159 **Somatic mutation calling from DNA sequencing files**

160 Somatic mutations were called on aligned reads by Mutect[28] and Varscan[29] separately with default
161 quality control parameters. Those calls were independently filtered by empirical Bayesian
162 mutation filtering using EBFilter[30] to reduce the false positive rate. Finally, the overlapping filtered
163 mutations between these two methods for each individual patient were selected as true positive
164 somatic mutations. VAF was calculated as percentage of total reads that contains mutation. A VAF
165 cutoff of 20% was used.

166 **Nonsense mediated decay escape by RNA and DNA mutated reads comparison**

167 Counts for mutated reads were collected by performing Mpileup[31] on matched RNA and DNA tumor
168 aligned sequencing reads for each of the specific genomic position of the somatic mutations
169 selected prior. Mutations with less than 10 mutated reads in the DNA were discarded, although this
170 number was very limited due to the stringent selection of mutations. No lower limit was placed on
171 the RNA read counts since this would hinder the detection of complete nonsense mediated decay
172 (NMD). NMD is measured by the difference in percentage between the fraction of mutated reads in
173 the DNA compared to the RNA (or ΔVAF), were the fraction of mutated reads in the DNA is set as
174 100%.

175 **Transcript mRNA levels and Zscore**

176 Accurate transcript quantification from RNA sequencing data is calculated by RSEM[32] and mapped
177 to Ensembl[33] genome browser transcripts from reference genome GRCh38.p13. Absolute mRNA
178 transcript levels per million (TPM) are 2log transformed. The z-score for mRNA levels of each
179 mutated transcript in a specific patient is calculated based on the following equation for transcript
180 with a mean mRNA level of more than 1.0 in all patients, removing transcripts that are not
181 expressed in colon cancer tissue.

182 $$Zscore\ (mutated\ transcript\ x\ for\ patient\ y) = \frac{mRNA\ x(y) - \mu\ mRNA\ x}{\sigma\ mRNA\ x}$$

183 **Microsatellite and premature termination codon calculation**

184 A proprietary Perl script run with GNU parallel utilizes the Ensembl Perl API to collect transcript
185 sequences based on the position, allele, type and length of inputted true positive somatic
186 mutations. Using this sequence, it calculates the number of mono-, di- or multi-nucleotide repeats

187 on the mutation position. Subsequently it constructs the new reading frame after mutation and
188 checks the RNA sequence until the occurrence of a premature termination codon. Finally, it
189 records the position of this PTC in the transcript and relative to the mutational site.

190 **Germline mutation calling from normal tissue DNA sequencing files**

191 Aligned DNA read files were spliced to incorporated only genomic regions of the mismatch repair
192 genes *MLH1*, *MSH2*, *MSH6*, *PMS2* and *EPCAM*. Germline mutation calling on spliced aligned reads
193 was performed using the GATK germline short variant discovery pipeline and using the GATK[34] best
194 practices. The workflow subsequently utilizes Haplotypecaller, CNNScoreVariants and
195 FilterVariantTrances. Mutation calls with QD < 10 and read count < 10 were discarded. Mutational
196 calls were functionally annotated employing Funcotator[34]. Variants of unknown significance were
197 manually scored by a laboratory specialist clinical geneticist on a 1-5 scale according to the
198 American College of Medical Genetics and Genomics (ACMG) guidelines[35], based on mutation
199 characteristics, amino-acid change and taking patient clinical data, somatic mutations and MLH1
200 promotor methylation data into account.

201 **Downstream analysis**

202 Statistical analysis were performed using R 3.4.4. in Rstudio 1.2.1335 (Rstudio, INC) with the
203 tidyverse package enabled. Difference between means are calculated by unpaired t test for
204 comparison between 2 groups and ANOVA for more than 2 groups. Graphs were generated using
205 ggplot2.

References

1. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
2. Maletzki, C., Schmidt, F., Dirks, W. G., Schmitt, M. & Linnebacher, M. Frameshift-derived neoantigens constitute immunotherapeutic targets for patients with microsatellite-instable haematological malignancies: frameshift peptides for treating MSI+ blood cancers. *Eur. J. Cancer Oxf. Engl. 1990* **49**, 2587–2595 (2013).
3. Brown, S. D. *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.* **24**, 743–750 (2014).
4. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
5. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).
6. Rooij, N. van *et al.* Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. *J. Clin. Oncol.* **31**, e439–e442 (2013).
7. Gubin, M. M. *et al.* Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515**, 577–581 (2014).
8. Litchfield, K. *et al.* Escape from nonsense mediated decay associates with anti-tumor immunogenicity. *bioRxiv* 823716 (2019) doi:10.1101/823716.
9. Kervestin, S. & Jacobson, A. NMD: a multifaceted response to premature translational termination. *Nat. Rev. Mol. Cell Biol.* **13**, 700–712 (2012).
10. Lin, E. I. *et al.* Mutational profiling of colorectal cancers with microsatellite instability. *Oncotarget* **6**, 42334 (2015).
11. Boyer, J. C. *et al.* Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. *Hum. Mol. Genet.* **11**, 707–713 (2002).
12. Jiricny, J. The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.* **7**, 335–346 (2006).
13. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).
14. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
15. National Institute of Health. The Cancer Genome Atlas. *The Cancer Genome Atlas, National Institute of Health, Bethesda, U.S.* http://cancergenome.nih.gov/.".
16. Wu, M. *et al.* MSI status is associated with distinct clinicopathological features in BRAF mutation colorectal cancer: A systematic review and meta-analysis. *Pathol. - Res. Pract.* **216**, 152791 (2020).

17. Abbott, K. L. *et al.* The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* **43**, D844–D848 (2015).
18. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
19. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222–226 (2017).
20. Jasperson, K. W., Tuohy, T. M., Neklason, D. W. & Burt, R. W. Hereditary and Familial Colon Cancer. *Gastroenterology* **138**, 2044–2058 (2010).
21. Bonsack, M. *et al.* Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC-peptide binding dataset. *Cancer Immunol. Res.* (2019) doi:10.1158/2326-6066.CIR-18-0584.
22. Alspach, E. *et al.* MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* **574**, 696–701 (2019).
23. Kreiter, S. *et al.* Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* **520**, 692–696 (2015).
24. Garrido, F., Aptsiauri, N., Doorduijn, E. M., Garcia Lora, A. M. & van Hall, T. The urgent need to recover MHC class I in cancers for effective immunotherapy. *Curr. Opin. Immunol.* **39**, 44–51 (2016).
25. Apcher, S., Daskalogianni, C. & Fåhraeus, R. Pioneer translation products as an alternative source for MHC-I antigenic peptides. *Mol. Immunol.* **68**, 68–71 (2015).
26. Lindeboom, R. G. H., Vermeulen, M., Lehner, B. & Supek, F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat. Genet.* **51**, 1645–1651 (2019).
27. Pembrolizumab With Ataluren in Patients With Metastatic pMMR and dMMR Colorectal Carcinoma or Metastatic dMMR Endometrial Carcinoma: the ATAPEMBRO Study - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT04014530.
28. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
29. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinforma. Oxf. Engl.* **25**, 2283–2285 (2009).
30. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89–e89 (2013).

31. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
32. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
33. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
34. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
35. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).

Acknowledgments

This work is made possible by the Dutch Cancer Society (KWF) Grant 21923. We would like to thank Prof. dr. L. Vermeulen for proofreading the manuscript.

Author Contributions

I.T.S and A.D.B. conceived the main hypothesis. I.T.S, F.P and A.D.B designed the experiments, analyzed the data and wrote the paper. I.T.S. developed the code and performed the experiments. I.T.S. and A.J. set up the in silico testing environment. I.T.S and B.R identified samples with a germline MMR mutation. A.D.B and C.J.A.P. supervised the project. All authors read and commented on the manuscript.

Competing Interests

The authors declare no competing interests.

Code Availability

The proprietary Perl computer code used to calculate PTC position in coding microsatellites is available at https://github.com/ITSpaanderman/PTC_predictor, including run instructions and test samples. GATK pipeline information is available at https://gatk.broadinstitute.org/hc/en-us. RSEM computer code is made available by its author at https://github.com/deweylab/RSEM.

Figures

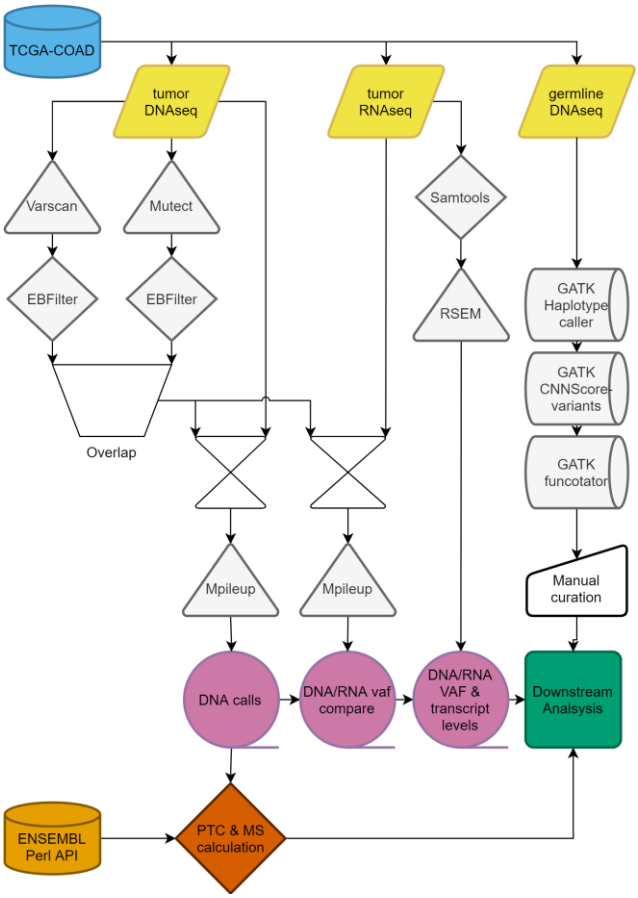


Figure 1. Schematic overview of in silico pipeline. TCGA-COAD: the cancer genome atlas colorectal adenocarcinoma database; Varscan: somatic variant calling; Mutect: somatic variant calling; EBFilter: Bayesian filtering for somatic variants; Mpileup: location specific read calling; RSEM: transcription quantification; VAF: variant allele frequency; ENSEMBL Perl API: Programmable interface for Ensembl genome browser; GATK: Genome analysis toolkit; Haplotypecaller: genetic variant calling; CNNScoreVariants: germline variant filtering; Functotator: Clinvar based annotator of genetic variants; PTC: premature termination codon; MS: microsatellite

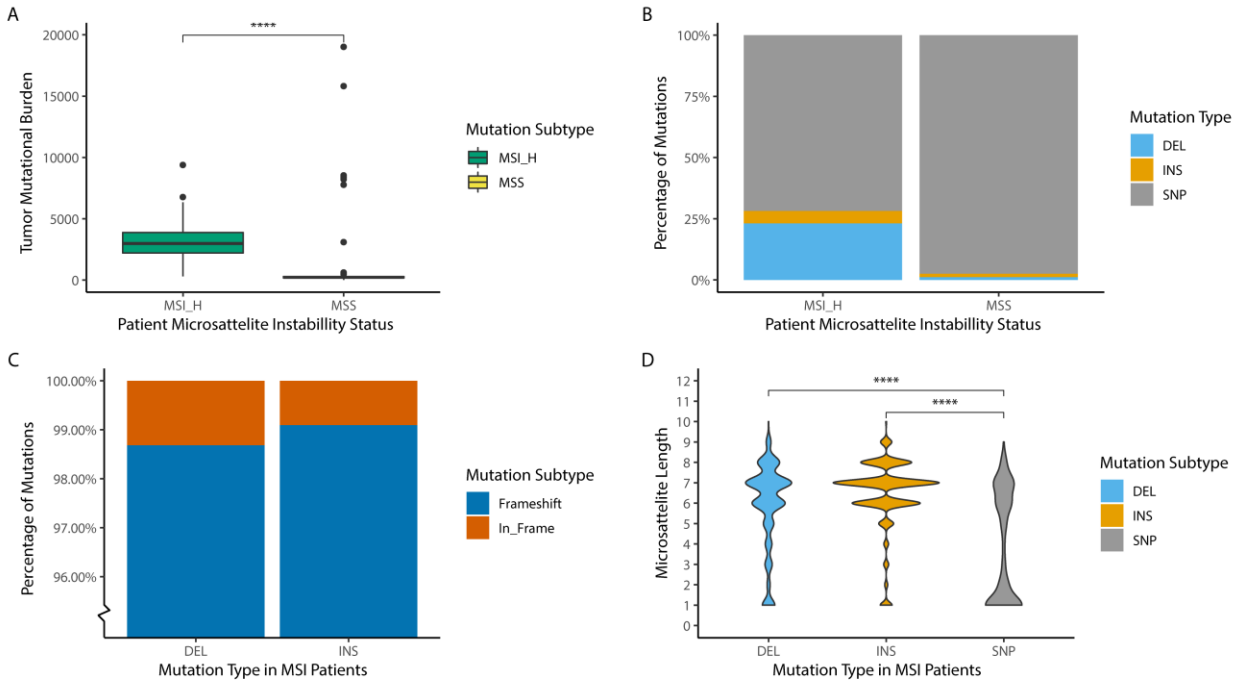


Figure 2. Mutational profile for MSS and MSI in TCGA-COAD. **A)** Difference in tumor mutational burden as number of mutations per patient between MSI (median = 2974, n = 71) and MSS patients (median = 218, n = 272, $p < 0.0001$). **B)** Mutation types as percentage of mutations between MSI (DEL = 23.15%, INS = 5.06%, SNP = 71.19%, n = 71) and MSS patients (DEL = 1.21%, INS = 1.34%, SNP = 97.4%, n = 218) **C)** Percentage of frameshift mutations for deletion (Frameshift = 98.69%, In_Frame = 1.31%, n = 53851) and insertions (Frameshift = 99.07%, In_Frame = 0.93%, n = 13172) in MSI patients (n = 71). **D)** Difference in microsatellite length in coding mononucleotide repeats on location of mutation in the genome between mutation subtypes, DEL (median = 7, n = 4960), INS (median = 7, n = 2384), SNP (median = 3, n = 1859, $p < 0.0001$) in MSI patients (n = 71). Microsatellite length of 3 mononucleotide repeats or less cannot be considered true microsatellites.

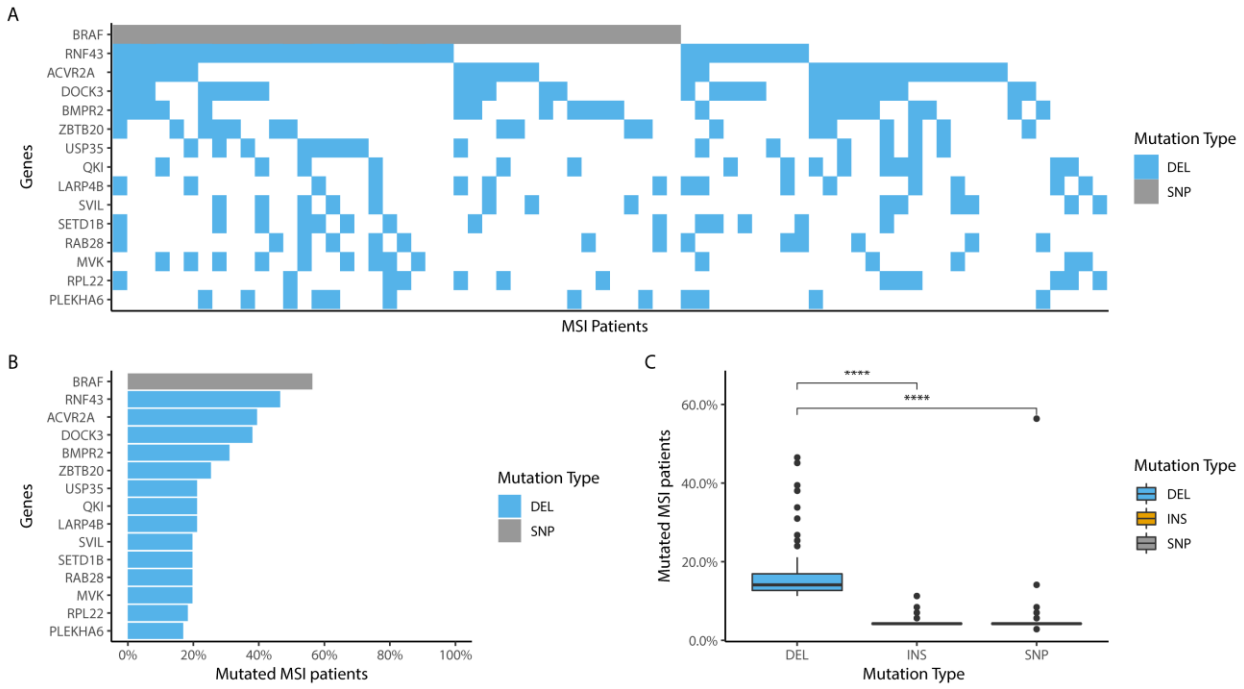


Figure 3. Mutation clustering in MSI TCGA-COAD. A) Waterfall graph depicting the 15 most frequently occurring mutation hotspots by Gene Identifier in MSI patients. Light grey mutations are SNPs. Light blue mutations are DELs. Other mutation types do not occur within the 15 most frequently occurring mutation hotspot. **B)** Percentage of most frequently occurring mutation loci. **C)** Difference in mutation frequency as percentage of patients with mutation on locus between mutation types for the top 100 most frequently occurring mutations, DEL (median = 14.08% , n = 100), INS (median = 4.23% , n = 100), SNP (median = 4.23% , n= 100 , $p < 0.0001$) in MSI patients (n = 71).

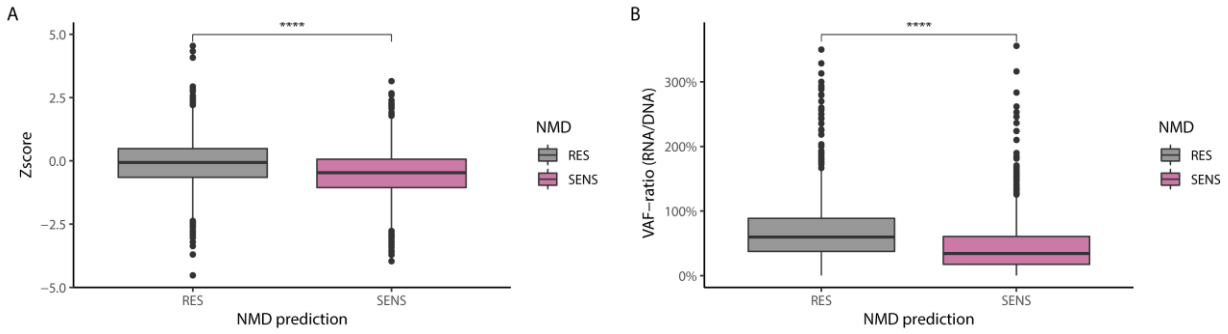


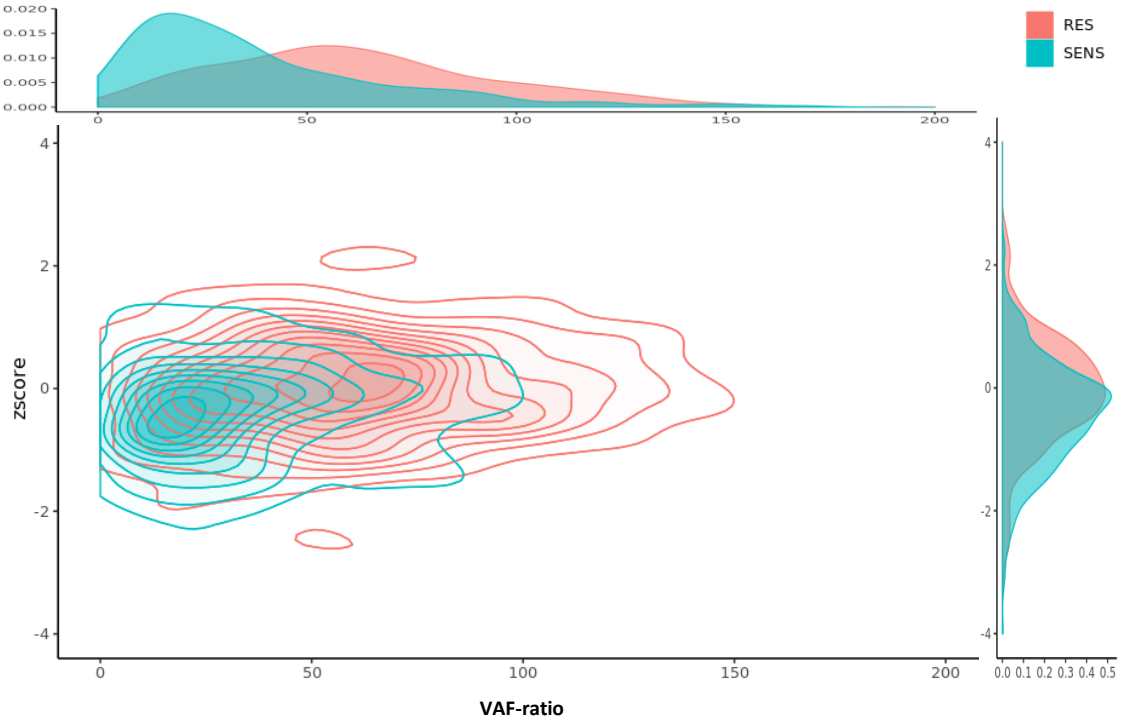
Figure 4. Nonsense Mediated Decay Escape in MSI TCGA-COAD. Approximation of Nonsense Mediated Decay(NMD) on mutated transcripts either predicted as NMD sensitive (SENS) or NMD resistant (RES), based on the location of the premature termination codon (PTC): RES PTC is located 50bp or less before the last exon-exon junction or in last exon **A)** NMD defined as a z-score of transcript levels for RES (median = -0.0657, n = 1357) and SENS (median = -0.4655, n = 1024, $p < 0.0001$). **B)** NMD defined as VAF-ratio: the allele frequency (VAF) as percentage RNA VAF of DNA VAF for RES (median = 59.9%, n = 936) and SENS (median = 34.3%, n = 904 , $p < 0.0001$).

Tables

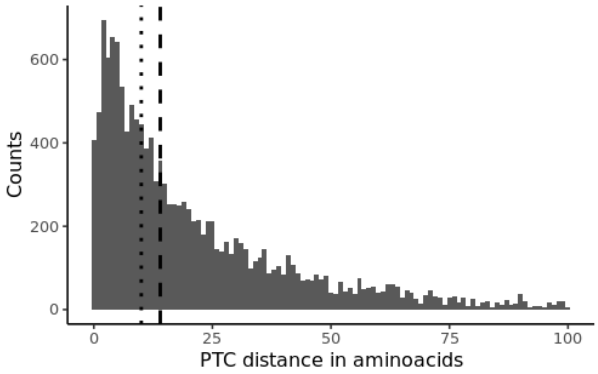
POSITION (HG38)	GENE	MS LENGTH	MS BASE	TYPE	NMD (50BP RULE)'	PTC DISTANCE (AMINOACIDS)	DELTAVAF (MEAN)	TRANSCRIPT MRNA LEVEL (LOG2 TPM)	CUMALITIVE MSI PATIENTS COVERED (N, %)	CUMALITIVE LYNCH PATIENTS COVERED (N, %)
17:58357799	*RNF43*	7	G	Frameshift_Del	SENS	39	85.73%	2.37	**33 (46.5%)**	
12:121804751	*SETD1B*	8	C	Frameshift_Del	SENS	28	51.65%	2.30	**36 (50.7%)**	
2:97083987	*FAHD2B*	7	C	Frameshift_Del	RES	23	57.86%	1.25	**39 (54.9%)**	
X:56565440	*UBQLN2*	7	C	Frameshift _Del	RES	71	103.06%	3.30	**44 (62.0%)**	1 (25%)
9:134053406	*BRD3*	8	C	Frameshift _Del	RES	84	111.84%	3.09	**45 (63.4%)**	
4:94252758	*SMARCAD1*	8	A	Frameshift _Del	SENS	24	77.61%	1.90	**47 (66.0%)**	
20:32202148	*PLAGL2*	7	C	Frameshift _Del	RES	32	53.33%	2.64	**47 (66.2%)**	
10:88922388	*STAMBPL1*	8	A	Frameshift _Del	RES	20	98.90%	2.44	**50 (70.4%)**	
X:130056035	*BCORL1*	7	C	Frameshift _Del	RES	20	129.17%	1.22	**52 (73.2%)**	
1:156745311	*HDGF*	8	T	Frameshift _Del	SENS	56	66.77%	6.11	**53 (74.6%)**	
4:2696449	*FAM193A*	7	A	Frameshift _Del	SENS	23	57.85%	2.53	**54 (76.1%)**	1 (25%)
12:49040708	*KMT2D*	5	C	Frameshift _Del	SENS	28	96.06%	2.29	**58 (81.7%)**	3 (75%)
21:14966008	*NRIP1*	8	A	Frameshift _Del	RES	22	66.73%	2.22	**58 (81.7%)**	3 (75%)
22:17818165	*MICAL3*	7	C	Frameshift _Del	SENS	104	68.38%	1.72	**59 (83.1%)**	
13:52474898	*CKAP2*	8	A	Frameshift _Del	RES	14	66.05%	3.24	**63 (88.7%)**	
14:23276100	*HOMEZ*	8	T	Frameshift _Del	RES	12	98.17%	2.11	**64 (90.1%)**	
19:13772153	*MRI1*	6	C	Frameshift _Del	RES	24	61.77%	2.15	**65 (91.5%)**	3 (75%)
16:14252442	*MRTFB*	7	C	Frameshift _Del	SENS	42	121.43%	1.94	**65 (91.5%)**	
1:27550357	*AHDC1*	7	A	Frameshift _Del	RES	54	52.91%	1.90	**66 (93.0%)**	
2:43225372	*ZFP36L2*	6	G	Frameshift _Del	RES	41	67.48%	6.38	**66 (93.0%)**	

Table 1. Public Frameshift Proteins (PFP). Properties of the 20 most frequently occurring mutations with SNORF of at least 10 Amino Acids and DeltaVAF of 50%. MS: microsatellite. Bold: cumulative numbers of MSI patient covered by mutation (n = 71). Germline patients n = 4. NMD: predicted NDM status based on PTC location in transcript. DeltaVAF: mean RNA-VAF percentage of DNA-VAF for all mutated patients. Expression: log2(TPM + 1) for most abundant corresponding transcript..

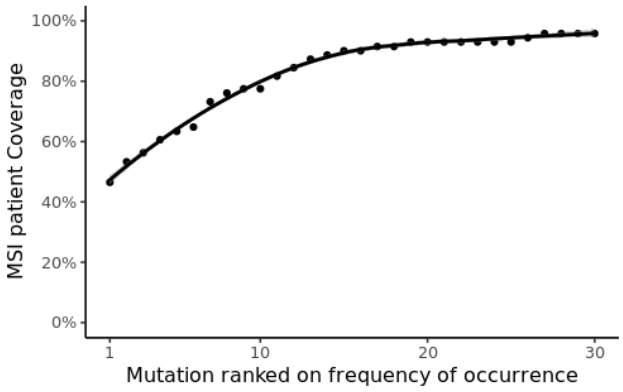
Supplementary Figures



Supplementary figure 1. Difference in NMD measurement. Density map showing difference between Z-score and VAF-ratio approach for the measurement of Nonsense Mediated Decay (NMD) in NMD RES and SENS mutations. Separation of NMD RES and SENS is clearer with VAF-ratio.



Supplementary figure 2. Premature Termination Codon (PTC) distance. PTC distance for mutation in MSI_H. Median is shown as dotted line (median = 14 Amino Acids). Dashed line shows 20 Amino Acids.



Supplementary figure 3. Patient coverage for top 30 frameshift mutations. Percentage of patients covered by n most frequently occurring frameshift mutations with adequate expression, a SNORF and NMD-escape.

Supplementary Tables

MOUSE SYMBOL	MOUSE ID	HUMAN SYMBOL	CANCER TYPE	RANK
Larp4b	217980	LARP4B	Blood Cancer	B
Larp4b	217980	LARP4B	Breast Cancer	B
Larp4b	217980	LARP4B	Colorectal Cancer	A
Larp4b	217980	LARP4B	Colorectal Cancer	A
Larp4b	217980	LARP4B	Colorectal Cancer	C
Larp4b	217980	LARP4B	Colorectal Cancer	D
Larp4b	217980	LARP4B	Gastric Cancer	B
Larp4b	217980	LARP4B	Liver Cancer	A
Larp4b	217980	LARP4B	Liver Cancer	A
Larp4b	217980	LARP4B	Nervous System Cancer	C
Larp4b	217980	LARP4B	Nervous System Cancer	D
Larp4b	217980	LARP4B	Nervous System Cancer	D
Larp4b	217980	LARP4B	Nervous System Cancer	D
Larp4b	217980	LARP4B	Nervous System Cancer	Not Ranked
Larp4b	217980	LARP4B	Pancreatic Cancer	B
Larp4b	217980	LARP4B	Pancreatic Cancer	B
Larp4b	217980	LARP4B	Skin Cancer	A
Larp4b	217980	LARP4B	Skin Cancer	C
Mecom	14013	MECOM	Blood Cancer	A
Mecom	14013	MECOM	Blood Cancer	A
Mecom	14013	MECOM	Blood Cancer	A
Mecom	14013	MECOM	Blood Cancer	B
Mecom	14013	MECOM	Blood Cancer	B
Mecom	14013	MECOM	Blood Cancer	D

Mecom	14013	MECOM	Blood Cancer	Not Ranked
Mecom	14013	MECOM	Blood Cancer	Not Ranked
Mecom	14013	MECOM	Colorectal Cancer	A
Mecom	14013	MECOM	Colorectal Cancer	D
Mecom	14013	MECOM	Colorectal Cancer	D
Mecom	14013	MECOM	Gastric Cancer	A
Mecom	14013	MECOM	Pancreatic Cancer	C
Mecom	14013	MECOM	Pancreatic Cancer	C
Mecom	14013	MECOM	Pancreatic Cancer	D
Acvr2a	11480	ACVR2A	Breast Cancer	C
Acvr2a	11480	ACVR2A	Colorectal Cancer	A
Acvr2a	11480	ACVR2A	Colorectal Cancer	B
Acvr2a	11480	ACVR2A	Colorectal Cancer	C
Acvr2a	11480	ACVR2A	Colorectal Cancer	C
Acvr2a	11480	ACVR2A	Colorectal Cancer	D
Acvr2a	11480	ACVR2A	Gastric Cancer	C
Acvr2a	11480	ACVR2A	Liver Cancer	C
Acvr2a	11480	ACVR2A	Liver Cancer	C
Acvr2a	11480	ACVR2A	Mixed	Not Ranked
Acvr2a	11480	ACVR2A	Pancreatic Cancer	A
Acvr2a	11480	ACVR2A	Pancreatic Cancer	B
Acvr2a	11480	ACVR2A	Pancreatic Cancer	B
Qk	19317	QKI	Blood Cancer	A
Qk	19317	QKI	Gastric Cancer	B
Qk	19317	QKI	Liver Cancer	A
Qk	19317	QKI	Liver Cancer	A

Qk	19317	QKI	Liver Cancer	B
Qk	19317	QKI	Liver Cancer	C
Qk	19317	QKI	Nervous System Cancer	A
Qk	19317	QKI	Nervous System Cancer	A
Qk	19317	QKI	Nervous System Cancer	B
Qk	19317	QKI	Nervous System Cancer	B
Qk	19317	QKI	Pancreatic Cancer	B
Qk	19317	QKI	Sarcoma	A
Qk	19317	QKI	Sarcoma	C
Zbtb20	56490	ZBTB20	Blood Cancer	A
Zbtb20	56490	ZBTB20	Liver Cancer	A
Zbtb20	56490	ZBTB20	Liver Cancer	B
Zbtb20	56490	ZBTB20	Liver Cancer	B
Zbtb20	56490	ZBTB20	Liver Cancer	B
Zbtb20	56490	ZBTB20	Nervous System Cancer	A
Zbtb20	56490	ZBTB20	Nervous System Cancer	A
Zbtb20	56490	ZBTB20	Nervous System Cancer	D
Zbtb20	56490	ZBTB20	Nervous System Cancer	D
Zbtb20	56490	ZBTB20	Nervous System Cancer	Not Ranked
Zbtb20	56490	ZBTB20	Pancreatic Cancer	A
Zbtb20	56490	ZBTB20	Pancreatic Cancer	B
Zbtb20	56490	ZBTB20	Skin Cancer	C
Son	20658	SON	Blood Cancer	D
Son	20658	SON	Breast Cancer	D
Son	20658	SON	Colorectal Cancer	B
Son	20658	SON	Colorectal Cancer	C

Son	20658	SON	Colorectal Cancer	D
Son	20658	SON	Gastric Cancer	D
Son	20658	SON	Liver Cancer	A
Son	20658	SON	Liver Cancer	A
Son	20658	SON	Nervous System Cancer	D
Son	20658	SON	Nervous System Cancer	D
Son	20658	SON	Pancreatic Cancer	C
Son	20658	SON	Pancreatic Cancer	D
Rnf43	207742	RNF43	Blood Cancer	B
Rnf43	207742	RNF43	Colorectal Cancer	A
Rnf43	207742	RNF43	Colorectal Cancer	B
Rnf43	207742	RNF43	Colorectal Cancer	B
Rnf43	207742	RNF43	Colorectal Cancer	D
Rnf43	207742	RNF43	Gastric Cancer	A
Rnf43	207742	RNF43	Liver Cancer	D
Rnf43	207742	RNF43	Liver Cancer	D
Rnf43	207742	RNF43	Pancreatic Cancer	A
Rnf43	207742	RNF43	Pancreatic Cancer	B
Tnrc6c	217351	TNRC6C	Blood Cancer	A
Tnrc6c	217351	TNRC6C	Colorectal Cancer	C
Tnrc6c	217351	TNRC6C	Liver Cancer	D
Tnrc6c	217351	TNRC6C	Liver Cancer	D
Tnrc6c	217351	TNRC6C	Nervous System Cancer	D
Tnrc6c	217351	TNRC6C	Nervous System Cancer	D
Tnrc6c	217351	TNRC6C	Pancreatic Cancer	C
Tnrc6c	217351	TNRC6C	Pancreatic Cancer	D

Tnrc6c	217351	TNRC6C	Sarcoma	B
Smarcad1	13990	SMARCAD1	Blood Cancer	D
Smarcad1	13990	SMARCAD1	Colorectal Cancer	B
Smarcad1	13990	SMARCAD1	Colorectal Cancer	D
Smarcad1	13990	SMARCAD1	Colorectal Cancer	D
Smarcad1	13990	SMARCAD1	Pancreatic Cancer	D
Smarcad1	13990	SMARCAD1	Sarcoma	D
Tmpo	21917	TMPO	Blood Cancer	C
Tmpo	21917	TMPO	Colorectal Cancer	B
Tmpo	21917	TMPO	Colorectal Cancer	C
Tmpo	21917	TMPO	Colorectal Cancer	D
Tmpo	21917	TMPO	Liver Cancer	C
Tmpo	21917	TMPO	Liver Cancer	D
Csmd3	239420	CSMD3	Gastric Cancer	D
Csmd3	239420	CSMD3	Liver Cancer	A
Csmd3	239420	CSMD3	Mixed	Not Ranked
Csmd3	239420	CSMD3	Nervous System Cancer	Not Ranked
Csmd3	239420	CSMD3	Sarcoma	A
Bmpr2	12168	BMPR2	Blood Cancer	B
Bmpr2	12168	BMPR2	Gastric Cancer	C
Bmpr2	12168	BMPR2	Nervous System Cancer	B
Bmpr2	12168	BMPR2	Nervous System Cancer	C
Brd3	67382	BRD3	Blood Cancer	A
Brd3	67382	BRD3	Liver Cancer	D
Brd3	67382	BRD3	Liver Cancer	D
Brd3	67382	BRD3	Pancreatic Cancer	D

Slc3a2	17254	SLC3A2	Blood Cancer	B
Slc3a2	17254	SLC3A2	Colorectal Cancer	D
Slc3a2	17254	SLC3A2	Colorectal Cancer	Not Ranked
Slc3a2	17254	SLC3A2	Liver Cancer	Not Ranked
Svil	225115	SVIL	Colorectal Cancer	D
Svil	225115	SVIL	Gastric Cancer	D
Svil	225115	SVIL	Liver Cancer	D
Svil	225115	SVIL	Lung Cancer	D
Hnrnpl	15388	HNRNPL	Colorectal Cancer	Not Ranked
Hnrnpl	15388	HNRNPL	Liver Cancer	D
Hnrnpl	15388	HNRNPL	Liver Cancer	D
Nap1l1	53605	NAP1L1	Blood Cancer	D
Nap1l1	53605	NAP1L1	Gastric Cancer	D
Nap1l1	53605	NAP1L1	Pancreatic Cancer	A
Wdtc1	230796	WDTC1	Blood Cancer	B
Wdtc1	230796	WDTC1	Colorectal Cancer	Not Ranked
Wdtc1	230796	WDTC1	Colorectal Cancer	Not Ranked
Camta2	216874	CAMTA2	Blood Cancer	D
Camta2	216874	CAMTA2	Colorectal Cancer	Not Ranked
Plekha6	240753	PLEKHA6	Colorectal Cancer	D
Plekha6	240753	PLEKHA6	Gastric Cancer	D
Card11	108723	CARD11	Blood Cancer	C
Ccr5	12774	CCR5	Blood Cancer	C
Cemip2	83921	CEMIP2	Blood Cancer	B
Dock3	208869	DOCK3	Blood Cancer	B
Fhod3	225288	FHOD3	Blood Cancer	D

Igf2r	16004	IGF2R	Nervous System Cancer	Not Ranked
Nr1h2	22260	NR1H2	Blood Cancer	D
Rab28	100972	RAB28	Blood Cancer	D
Rpl22	19934	RPL22	Liver Cancer	D
Xylt2	217119	XYLT2	Blood Cancer	D

Supplementary Table 1. Candidate cancer genes . List with genes described in the candidate cancer gene database that are part of the top 100 most occurring mutations in MSI colon cancer. Rank: Relative rank assigned to CIS in study