Genomic and transcriptomic heterogeneity of colorectal tumors arising in Lynch Syndrome

Hans Binder[1]*, Lydia Hopp[1]*, Michal R. Schweiger[2,3]*, Steve Hoffmann[1], Frank Jühling[1,4], Martin Kerick[2,3], Bernd Timmermann[5], Susann Siebert[2,3], Christina Grimm[2,3], Lilit Nersisyan[6], Arsen Arakelyan[6], Maria Herberg[1], Peter Buske[1], Henry Loeffler-Wirth[1], Maciej Rosolowski[7], Christoph Engel[7], Jens Przybilla[1], Martin Peifer[2], Nicolaus Friedrichs[2], Gabriela Moeslein[8], Margarete Odenthal[2], Michelle Hussong[2,3], Sophia Peters[9], Stefanie Holzapfel[9], Jacob Nattermann[10], Robert Hueneburg[10], Wolff Schmiegel[11], Brigitte Royer-Pokora[12], Stefan Aretz[9], Michael Kloth[2], Matthias Kloor[13], Reinhard Buettner[2#], Jörg Galle[1#], Markus Loeffler[7#]

* equally contributed
\# shared senior authorship

1 Interdisciplinary Centre for Bioinformatics, Leipzig University, Leipzig, Germany
2 Institute of Pathology, Centre for Integrated Oncology, University Hospital Cologne, Cologne, Germany
3 Translational Epigenomics, University of Cologne, Cologne, Germany
4 Inserm, U1110, Institut de Recherche sur les Maladies Virales et Hépatiques, Strasbourg, France; Université de Strasbourg, Strasbourg, France
5 Max Planck Institute for Molecular Genetics, Berlin, Germany
6 Group of Bioinformatics, Institute of Molecular Biology, National Academy of Sciences, Yerevan, Armenia
7 Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig, Germany
8 Department of Hereditary Tumor Syndromes, Surgical Center, HELIOS Clinic, University Witten/Herdecke; Wuppertal, Germany
9 Institute of Human Genetics, University Hospital Bonn, Center for hereditary tumor syndromes, University Bonn, Bonn, Germany
10 Department of Internal Medicine I, University Hospital Bonn, Center for hereditary tumor syndromes, University Bonn, Bonn, Germany
11 Department of Medicine, Hematology and Oncology, Ruhr-University of Bochum, Knappschaftskrankenhaus, Bochum, Germany.
12 Institute of Human Genetics and Anthropology, Heinrich-Heine University, Düsseldorf, Germany
13 Department of Applied Tumour Biology, Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany, and Clinical Cooperation Unit Applied Tumour Biology, DKFZ (German Cancer Research Center) Heidelberg, and Molecular Medicine Partnership Unit, University Hospital Heidelberg and EMBL Heidelberg, Germany

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/path.4948

Correspondence to; Hans Binder, Interdisciplinary Centre for Bioinformatics, Leipzig University, Haertelstraße 16 – 18, 04107 Leipzig, Germany; binder@izbi.uni-leipzig.de

Running title: Genomic and transcriptomic heterogeneity of Lynch Syndrome in colon

Conflict of interest: R.B. is a cofounder and scientific advisor of Targos Molecular Pathology, Kassel, Germany and has received research grants from Pfizer and Roche. All other authors declare no conflict of interests.

Abstract

Colorectal cancer (CRC) arising in Lynch Syndrome (LS) comprises tumors with constitutional mutations in DNA mismatch-repair genes. Whole-genome and transcriptome studies of LS-CRC are still missing to address questions about similarities and differences of mutation and gene expression characteristics between LS-CRC and sporadic CRC, about the molecular heterogeneity of LS-CRC and about specific mechanisms of LS-CRC genesis linked to dysfunctional mismatch-repair in LS colonic mucosa and the possible role of immune editing. Here, we provide a first molecular characterization of LS-tumors and of matched tumor-distant reference colonic mucosa based on whole-genome DNA- and RNA-sequencing analyses. Our data support two subgroups of LS-CRCs, G1 and G2, where G1 tumors show a higher number of somatic mutations, higher microsatellite slippage and a different mutation spectrum. The gene expression phenotypes support this difference. Reference mucosa of G1 shows a strong immune response associated with the expression of *HLA* and immune check point genes and the invasion of CD4+ T cells. Such immune response is not observed in LS tumors, G2-reference and normal (non-Lynch) mucosa and sporadic CRC. We hypothesize that G1 tumors are edited for an escape from a highly immunogenic microenvironment via loss of *HLA*-antigen presentations and T-cell exhaustion. In contrast, G2 tumors seem to develop in a less immunogenic microenvironment where tumor promoting inflammation parallels tumorigenesis. Larger studies on non-neoplastic

mucosa tissue of mutation carriers are required to better understand early phases of emerging tumors.

Key words: mismatch-repair, hereditary cancer, tumor heterogeneity, immune editing

Introduction

Lynch syndrome (LS) is one of the most frequently inherited cancer predisposition syndromes, contributing about 3% of all colorectal cancer (CRC) cases [1]. LS is defined by an autosomal dominant heterozygous constitutional mutation in one of four key mismatch-repair (MMR) genes, i.e. *MLH1* (about 60%), *MSH2* (about 30%), *MSH6* or *PMS2* [2,3]. LS colorectal cancers (LS-CRC) develop after acquisition of a second somatic hit in the respective gene, leading to microsatellite instability (MSI). In contrast, MSI in sporadic CRC (sCRC) most frequently results from promoter hyper-methylation of the *MLH1* gene [4,5]. Clinical studies demonstrated a better overall survival of LS-CRC patients compared to sCRC [6] which possibly links to local inflammation [7] in response to the production of highly immunogenic frameshift peptides transcribed from microsatellite-instable genes in LS mucosa [8]. Moreover, different modes of immune response are reported to edit tumor cells associated with different LS-phenotypes [9,10]. Whole-genome and transcriptome studies published in recent years provided deep insights into the molecular etiology of sCRC and revealed distinct molecular subtypes with diagnostic impact [4,11-13]. Despite recent discoveries about the molecular genetics of LS [14] such efforts are still missing for LS-CRC. We here present the first study of LS-CRC based on whole-genome DNA and RNA sequencing of tumor and matched tumor-distant reference samples of colonic mucosa in LS. Data for sCRC and matched healthy colonic mucosa were taken from the TCGA consortium [4] for comparison. In this study we asked about similarities and differences of mutation and gene expression characteristics between LS-CRC and sCRC, about the molecular heterogeneity of LS-CRC and about possible mechanisms of tumor genesis and progression

in LS linked to dysfunctional MMR in non-neoplastic colonic mucosa, and the possible role of immune editing.

Materials and Methods

Sample collection from LS-CRC patients and pathological diagnoses

Paired patient-matched specimens of tumor (adenoma and/or cancer) and tumor-distant non-neoplastic mucosa as reference samples were collected from LS-CRC patients (ten with approved constitutional MMR mutation, one with clinico-pathological characteristics) who were identified through the German HNPCC Consortium [15,16] (supplementary material, Table S1). The study had been approved by the Ethics committees of the Universities of Duesseldorf, Bochum and Bonn. Samples were used for pathological examination, whole-genome DNAseq and RNAseq analyses and immunohistochemical staining of CD4 (Mouse anti-human, 1:100, Thermo Scientific, Darmstadt, Germany; see [17]). For comparison we included sequencing data of microsatellite stable (MSS) and instable (MSI) sCRC cases and of healthy colonic mucosa taken from the TCGA repository (supplementary material, Table S2).

DNA and RNA whole-genome sequencing and primary data processing

Genomic DNA was extracted from fresh frozen tissue samples after histologic re-evaluation of representative sections as described elsewhere [18]. Libraries were prepared in duplicates according to the manufacturer´s protocol. Whole-genome sequencing was performed according to Illumina's TrueSeq Sample Preparation Kit followed by paired-end sequencing on the HiSeq 2500 platform (Illumina, San Diego, CA, USA). Total RNA was purified using the RNA MicroPrep Kit (Zymo Research, Irvine, CA, USA). Complementary DNA (cDNA) was generated using SuperScript II Reverse Transcriptase (Invitrogen, Waltham, MA, USA). Sequencing and library preparations were performed using the same techniques as described above for genomic DNA. The quality of libraries was checked using the FastQC tool [19]. Paired-end read sequencing data were mapped to the human genome (hg19) using the 'segemehl' algorithm with default settings using the split read mapping option [20]. Non-mapped RNAseq reads were remapped using the lack script in 'segemehl'. Mapped BAM files were realigned using the GATK suite [21]. Most BAM file manipulations were done using

either in house scripts or 'samtools' [22]. The data are available in the dbGaP data base (www.ncbi.nlm.nih.gov/gap) under accession number phs001407.

Mutation and microsatellite analysis

DNAseq data were used for mutation calling, mutation spectra analysis and MS slippage profiling. We applied genome analysis toolkit (GATK) to call SNVs and InDels. Somatic mutations were defined as the set of mutations that is present in tumor samples but not in matched reference tissue. Somatic InDels for frameshift analyses were accepted from the GATK InDel calls of tumor alignments when no hit was found within a window (±10 bp) around the corresponding position in reference sample. Mutation spectra were generated by calculating the ratios for every mutation type including the information of adjacent 3'/5' bases as described in [23]. For MS-slippage analysis we selected loci of A/T monorepeats of a length of 20 bp in the reference genome. Then, the actual length of these repeats in the LS genomes were determined (see supplementary material, Supplementary materials and methods) and analyzed in terms of their mean length (MSL) averaged over all repeats per sample, and their frequency distribution. These so-called 'slippage profiles' were calculated using repeats at loci which were detected in both, the tumor and reference samples. MS-specificity analysis was performed for all MS of length <21 bp (supplementary material, Supplementary materials and methods). Selected monorepeats were verified by Sanger sequencing (supplementary material, Table S3).

Gene expression analysis

Mapped RNAseq gene expression data were clustered using self-organizing map (SOM) machine learning as described previously [24]. The SOM method 'portraits' the expression state of each sample in terms of a two-dimensional 40x40 grid image where size and topology were chosen to allow robust identification of clusters of co-expressed genes [24,25]. Sample diversity was visualized using second level SOM which maps samples instead of genes [24]. For functional interpretation of the gene clusters we applied gene set analysis [26] in terms of gene set enrichment Z-score (GSZ) profiles and population maps of the set

genes. Gene sets were taken from the GSEA-repository and from literature for different functional categories defined by gene ontology (GO), ‘hallmarks of cancer’ and other sources [25,27,28]. The activity of selected pathways was estimated using the pathway signal flow method [29]. All methods were implemented in the R-package ‘oposSOM’ used for analysis [30].

Results

Mutation and microsatellite characteristics

Genome-wide DNA sequencings and RNA-Seq were performed from eleven LS-patients, each with tumor and corresponding reference tissues (supplementary material, Table S1). Patient's samples showed a large heterogeneity according to their genomic characteristics (**Figure 1**). We stratified them into two groups G1 and G2 for further analysis. G1 compared to G2 show higher mutation numbers, a higher fraction of recurrent frameshift mutations (**Figure 1**A, supplementary material, Table S4), higher MS slippage (**Figure 1**B, C) and a slightly different mutation spectrum (**Figure 1**D, E, supplementary material, Figure S1). Higher MS slippage is demonstrated by the frequency distribution of the length of 20meric repeats and the related mean MS length. Of the MS 77 were selected that separates both groups with high selectivity (supplementary material, Supplementary materials and methods and Figure S2). The number of somatic mutations in cancer and the mean length of the selected monorepeats anti-correlate in G1. *POLE* is mutated in some G1 samples (**Figure 1**A). Combination of LS constitutional mutations with somatic mutations of *POLE* seems to induce an additional mutation load due to *POLE* dysfunction in accordance with [31]. The mutation spectra resemble those previously attributed to the inactivation of MMR and impaired *POLE* function in CRC, respectively (spectra 6 and 10 in [23]). Mutation characteristics of LS-CRC were compared with those of sCRC samples available from TCGA. The mutation spectra and MS length distributions of G1 LS-CRC share similarity with MSI sCRC whereas G2 shares similarities with MSS sCRC (supplementary material, Figure S3). The MS length distributions in LS reference mucosa demonstrate nucleotide loss in contrast to normal mucosa in the TCGA samples. In summary, mutation numbers, slippage profiles and mutation spectra suport two groups of LS-CRC cases.

Recurrent mutations in LS-CRC

We identified 20 genes mutated in four of five G1 LS-CRC cases but not or only once in G2 (supplementary material, Table S5, Figure 1A, and supplementary material, Figure S4A). Among these genes were *ACVR2A*, *TGFBR2*, *CDC27*, *AIM2* and *PDS5B* which have been found frequently mutated also in MSI sCRC [4,12,32,33]. Other genes are known to be frequently mutated in MSS sCRC (*TP53*, *KRAS*) [4]. BRAF mutations, except one G1 adenoma, were absent, in agreement with [34-36]. Most of the genes recurrently affected by frameshifts in LS-CRC contain microsatellites [36] suggesting that frameshifts are directly caused by microsatellite instability. This applies also to existing instabilities of MS in the reference tissue evident as bp-loss documented in Figure 1B. In summary, most of the genes recurrently mutated in G1 are also frequently mutated in sCRC.

Heterogeneity of LS-CRC gene expression phenotypes

RNAseq data of LS-CRC samples were analyzed using self-organizing maps (SOM)-portrayal [24,37]. *Mean expression portraits* of each tissue type of G1 and G2 and *difference portraits* between cancer and reference from these reveal one unique cluster of genes up-regulated in cancers of both G1 and G2 (Figure 2A and supplementary material, Figure S5 for single sample portraits) which is enriched in genes that are related to proliferation and that are frequently activated in sCRC. In contrast, genes related to inflammatory processes have a high expression level specifically in G1 reference tissue only. The difference portrait between G1 and G2 indicates higher activity of proliferation, oxidative phosphorylation and energy metabolism in G1.

The *Euclidean* distance between the expression portraits of tumors and reference tissues were calculated as a measure for diverging transcriptional programs. In G1 tumors it increases proportional to the number of somatic mutations for most samples indicating that the accumulation of mutations in G1 tumors shifts the expression patterns (**Figure 2**B). Sample diversity maps revealed that all G1 cancer samples show high similarity and

accumulate in a narrow region of the map whereas the G1 reference samples are more diverse expressing either a mucosa or inflammatory signature (**Figure 2**C and supplementary material, Figure S6). Hence, heterogeneity of active transcriptional programs decreases during tumor formation especially in G1. The diversity of G2 tumors is comparable with that of the G2 reference tissues manifesting either a stromal or a mucosa signature. They show less pronounced cancer characteristics than G1. In summary, the diversity of gene expression phenotypes supports differentiation between G1 and G2.

Expression signatures of LS-CRC

For functional interpretation we calculated the expression profiles of selected gene sets in the LS-samples using gene set Z-score (GSZ) metrics [25] (**Figure 3**A). Sets related to *cell cycle activity*, *MYC target activation* and *CRC markers* are highly expressed in cancer samples whereas sets related to *immune response, mucosa* and *stroma* functions are overexpressed in reference samples of G1 and G2, respectively. The expression heatmap of genes with recurrent somatic mutations and with constitutional mutations in G1 reveals that many of these genes show reduced mRNA expression in the mutated state (**Figure 3**B) in agreement with previous results in MSI sCRC [12] and with their frequent functions as tumor suppressors (supplementary material, Table S5).

Next, we determined clusters of co-expressed genes in the expression landscape (**Figure 3**C). Six major clusters labeled A – F were identified and functionally assigned using gene set analysis, namely: A) *Immune Response*, B) *Stroma*, C) *Mesenchyme*, D) *Proliferation*, E) *Oxidative Phosphorylation (Oxphos),* and F) *Healthy Mucosa* (see also supplementary material, Table S6, Figure S7 and Figure S8). We found correspondence of the expression signatures of LS-CRC with such previously detected in studies on sCRC, and healthy and inflamed colon [38-43]. The functional characteristics of four consensus molecular subtypes of sCRC extracted recently [11] correspond to the functional context of our four main modules where our *proliferation* signature agrees with the signature of *Canonical CRC* (supplementary material, Figure S9). It is characterized by increased expression of *WNT-*

targets, among them *CD44*, *Birc5*, *CCND1*, *MYC* and *ASCL2*, all located in the proliferation spot D and it resembles the signature of the proliferative compartment in the lower crypt of healthy colon [42] (supplementary material, Figure S9C). Comparison with the TCGA sCRC samples shows that reference mucosa of LS-mutation carriers is more heterogeneous than normal (non-LS) mucosa which is exclusively characterized by the mucosa expression signature (supplementary material, Figure S10).

Pathway activity analysis suggests that mutation of *ACVRA*, *TCF7L2* and *TGFBR2* in G1 LS-CRC affect *MYC activity* (supplementary material, Figure S11). In summary, gene expression in the LS samples can be characterized by four main modules where the *proliferative module* is similar to the *Canonical CRC* signature, the *mucosa module* is also characteristic for *healthy, non-LS mucosa* in TCGA and the *inflammation* and *stroma* modules are specific for reference mucosa of G1 and G2, respectively.

Pre-oncogenic immune response in G1 mucosa

In reference mucosa of G1 the activity of expression signatures related to immune response exceeds that in G1 cancer ($p<0.02$, paired sample Wilcoxon test), that of all tissue types of G2 (**Figure 4**A) and of sCRC and normal 'non-LS' mucosa as well (supplementary material, Figure S10B). It is characterized by overexpression of *CD3*, *CD4* and *CD19* (local false discovery rate $fdr<0.01$) and to a lesser extent of *CD8* suggesting infiltration of T- and B-immune cells as observed by others for MSI sCRC [43,45,46]. This finding is supported by analysis of B- and T-cell receptor pathways, which reveals high activity of the *NFkappaB*, *MAPK* and *PIK3* branches in G1 reference mucosa (supplementary material, Figure S12 and Figure S13). Moreover, chemokine expression data indicate co-expression of special receptor-ligand pairs in the *Inflammation* expression signature (**Figure 4**B) and high activity of the branch B-cell adsorption in the 'Leukocyte trans-endothelial migration' pathway in G1 LS reference tissue (supplementary material, Figure S14). The inflammation signature (spot A) co-regulates with a gene set subsumed as 'coordinated immune response cluster' (CIRC) in sCRC [43] (supplementary material, Figure S9E) and with immune-checkpoint genes

(ICG) encoding T-cell inhibitors [47,48] (**Figure 4**C and supplementary material, Figure S9E). Another subset of chemokines antagonistically switches compared with the 'anti-tumoral' inflammatory signature and upregulates in LS-CRC (**Figure 4**B, spot D). This 'pro-tumoral' immune response includes the chemokines *CXCL1-3*, *8* (*IL-8*) and *10* which upregulate also in sCRC [49,50].

Genes associated with presentation of *HLA* class II antigens accumulate in spot A together with their regulators *CIITA* and *RFX5* (**Figure 4**C). *CD4*+ T-cells respond to *HLA* class II antigens that are expressed by various tumors of non-lymphoid origin, including a subset of sCRC [51]. Hence, *CD4* expression in reference tissue of G1 suggests infiltration by *CD4*+ T cells that recognize *HLA* class II antigens. Indeed, *CD4*+ staining of microscopic slides of the samples studied (**Figure 4**E, F) confirms that CD4+ cells accumulate throughout the stroma of G1 reference tissue preferentially close to the epithelium in contrast to the most de-differentiated parts of the G1 and G2 tumors and G2 reference mucosa which are nearly devoid of T-cells and which show a rather moderate *CD4*+ staining.

Both the number of neo-antigens predicted in G1 LS-CRC from the DNAseq data and the mean expression of all *HLA* class II genes in G1 reference tissue increase as a function of the mutational load in LS-CRC (**Figure 4**D). The immune response in reference tissue of G1 might be induced by cell damage and the expression of immunogenic frameshift peptide antigens [52]. We interpret the mutational load in LS-CRC as a proxy for the mutational activity in LS reference mucosa. This assumption is supported by the fact that more than 50% of the microsatellites selected as instable in G1-cancer are already instable in G1-reference mucosa (supplementary material, Figure S2A). Also *B2M*, a regulator of *HLA* class I antigens, co-regulates with the immune response cluster in spot A together with *HLA* class I antigens (supplementary material, Figure S15), suggesting also presentation of these antigens in reference mucosa. Strong down-regulation of the specific *Inflammation* signature (spot A) in cancers and adenoma of G1 suggests an immune-escape [46]. In summary, reference mucosa of G1 patients in contrast to G2 patients show a strong immune response that appears to suppress cancer development until the cells underwent an immune escape.

Discussion

We found remarkable heterogeneity of LS regarding: i) the mutation frequency and MS instability in LS-tumors, ii) the inflammatory state of pre-oncogenic mucosa and iii) the mechanism of tumorigenesis via immune-editing. We found it suggestive to stratify the LS-cases into two groups where, overall, G1 LS-CRC show strong similarities with the mutation characteristics of (MLH1 silenced) MSI sCRC, whereas G2 LS-CRC showed weaker MS-instabilities and fewer mutations.

Mechanisms affecting genetic heterogeneity in LS

G1 contained a higher percentage of *MLH1* mutation carriers than G2 which suggest an association between the constitutional mutation and the mutational characteristics. Regarding its molecular mechanism, impaired *MLH1* impedes proper localization of repair proteins near the mismatch and reduces the efficiency of MMR [53] with consequences for the resulting mutational load. Additional somatic mutations of *POLE* and *MSH3,* both observed in G1 LS-CRC and both involved in MMR, are expected to further increase the somatic mutations that accumulate during tumor progression. These mechanisms can explain higher mutation numbers and MS slippage observed in G1 LS-CRC. Interestingly, *MLH1* deficient haplotypes associate with the presence and family history of colonic inflammatory diseases, suggesting enhanced susceptibility of *MLH1* mutation carriers for inflammation [54,55].

The reason for the increased contribution of C>G substitutions observed in the mutation spectra of G1 cancers is unclear. One possible factor is an increased activity of cytidine deaminases of the APOBEC family found also in many cancers with local hyper-mutation patterns [23] (supplementary material, Figure S9D). The different C>T transition frequencies between G1 and G2 cancers possibly reflect differences in DNA-methylation between both groups [56]. Hereditary cancers can mimic DNA hyper-methylation patterns observed in

sporadic cancers [35,57]. Genes hyper-methylated in sCRC show differential expression between G1 and G2 in support of this hypothesis (supplementary material, Figure S9F).

Inflammation in G1 mucosa enforces immune escape

We found strong immune response in G1-reference mucosa that were not evident in G2. Inflammatory responses play an important role in the progression of a variety of solid tumors and particularly of CRC due to cumulative effect of sequential genetic alterations, leading to the expression of tumor associated antigens [58]. Thus, inflammation can act in two ways, either promoting or suppressing tumorigenesis [59]. This is also evident in LS where in G1 reference mucosa 'antitumor' inflammation can be assumed to suppress tumorigenesis, while 'pro-tumor' inflammation in both, G1 and G2 LS-CRC promotes it (**Figure 4**B). Notably, a specific immune response was detected in LS mutation carriers by means of peripheral blood analysis [60] and via the detection of premalignant MMR-deficient crypt foci [52] suggesting that an immunological interaction between emerging MMR lesions and the immune system occurs early during the course LS. Interestingly, *CTNNB1* was mutated in three out of five G1 cancers (60%) but not at all in adenoma (G1 and G2) and G2 cancer cases (**Figure 1**A) and relatively rarely in the TCGA sCRC cases studied (17% in MSI and 8% in MSS). *CTNNB1* mutations were suggested to associate with a subgroup of LS CRCs emerging from MMR-deficient crypt foci which give rise to non-polypous cancer precursors [61]. The question whether MMR inactivation in the analyzed tumors was an event secondary to polypous growth (e.g. caused by *APC* mutation) or the initial event as suggested by *CTNNB1*-mutant tumors described in [61] cannot be answered conclusively, particularly because *CTNNB1* and *APC* were co-mutated in part of our LS tumors, possibly because variants of unclear functional significance were included. Also *RNF43*, encoding a negative regulator of WNT-signaling, was mutated in three G1 cases, however this was mutually exclusive to *CTNNB1*. This observation is in agreement with [62] who found that *RNF43* mutations due to their association with MMR-deficiency and thus with mutational burden in LS commonly taking place before adenoma formation in Lynch syndrome.

We hypothesize that tumor-suppressing inflammation in reference mucosa of G1 is induced, among other factors, by accumulation of frameshift peptides and possibly DNA lesions due to the larger mutational load in this group. Then, the transition from non-neoplastic, inflamed mucosa into tumor states is characterized by antagonistic alterations of inflammatory and proliferative transcription programs suggesting an immune escape mechanism. Strikingly, *AIM2*, encoding an innate immunity sensor that recognizes double stranded DNA of microbial or host origin, is the only gene specifically and recurrently mutated in G1 which is activated in reference mucosa and deactivated in tumors. Because *AIM2* has been shown to activate several *HLA* class II genes [63], its mutation might be an effective trigger for emerging tumor cells to escape the immunogenic microenvironment in G1 LS mucosa. *AIM2* deficiency promotes uncontrolled proliferation of stem cells in the intestinal mucosa of mice and increases the likelihood of sCRC development [64]. Deactivation of *AIM2* also associates with tumor burden in colitis-associated colon cancer (CAC) [65]. CAC is also characterized by *KRAS* and *TP53* mutations, two important effectors of inflammatory pathways where mutations of *KRAS* appear as late events in CAC tumorigenesis, while loss of *TP53* function is an early event; usually not observed in 'normal' sCRC in this order [59,66,67]. We found recurrent somatic mutations of these genes in G1 LS-CRC which associate with deregulation of *KRAS* signaling and *TP53* target expression and decreased tumor-suppressing inflammation. A similar trend was found in *KRAS* mutated MSS sCRC [43] in contrast to MSI sCRCs which up-regulate inflammation upon tumorigenesis (supplementary material, Figure S10B).

Immuno-editing in LS-CRC

Immuno-editing describes how anti-tumor immune surveillance of the host leads to the selection of tumor cells that evade immune cell recognition and destruction [9,68]. Our results suggest that G1 tumor outgrowth is shaped by immune escape of emerging tumors cells in a highly immunogenic microenvironment in reference mucosa mainly via two main mechanisms [46,47]: (i) overexpression of immune checkpoint genes (ICG) in G1 reference

mucosa which inhibits local T-cell activity and, (ii) loss of HLA antigen presentation in G1 tumors as indicated by deactivation of *HLA* gene expression. Both mechanisms potentially allow to circumvent elimination of oncogenic lesions. In consequence, immune editing suppresses immune response in the resulting overt tumor in G1 compared with reference mucosa. Immune-editing by the loss of *HLA* I and by *HLA* II antigen presentation genes has been demonstrated to occur in LS-CRC after mutations of the regulators *B2M* [9] and *CIITA* and *RFX5* [46], respectively. These regulators are found to co-express with the respective HLA antigens but neither of these genes was found to be recurrently mutated in our G1 tumor samples which suggests an alternative regulatory mechanism, possibly via deactivating *AIM2* mutations as discussed above. Strikingly, reference mucosa of G1 LS seems to share similarities with MSI sCRC where high mutation rates activate inflammatory response and ICG expression [13]. Contrarily, low expression of ICG in G1 tumors shows analogy with MSS sCRC which in contrast to MSI sCRC do not benefit from the administration of immune-checkpoint inhibitors [69-71].

Limitations and outlook

The present study is limited by its small sample size. The need for higher quality DNA and RNA and sequencing from fresh frozen material restricted us to analyses of eleven patients. We were able to collect only this limited number of paired tissue specimen with high quality for our analysis. Larger series on such reference mucosa and tumor samples in LS-patients are clearly required to better resolve the molecular heterogeneity of LS in colon, to identify LS-specific mutational patterns including more detailed MS characteristics and their stratification into subgroups. Our study shows that special focus should be directed to 'non-neoplastic' mucosa of LS constitutional mutation carriers to better understand early phases of tumor development. This requires detailed genetic analyses in terms of somatic tissue specific mutations and also of expressed neo-antigens. Our study provided only first indications of the importance of the immune response in the colon of LS patients which clearly implies that more detailed analyses of the immune microenvironment is pertinent [72].

Finally, functional and clinical aspects must augment the molecular effects reported here, to address possible consequences for tumor diagnosis and prevention in LS. Hence our study contributes a working hypothesis to the field of carcinogenesis in LS mutation carriers which needs further investigations because of its potential clinical implications.

Acknowledgements

We very appreciate support of Mrs. Dr. Katja Rillich for coordinating this project. This publication is financially supported by the German BMBF (PTJ grant HNPCCSys 031 6065A). This work is further supported by the BMBF-project grants No. FFE-0034 (oBIG) and WTZ ARM II-010 (Pathway Maps) (HB, LH, HLW, AA, LN) and the Volkswagenstiftung (Lichtenberg program, MRS).

The results shown here are in part based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/.) Access has been granted under project number 7432 (HNPCC-Sys).

Authors contributions statement

Study design and coordination: ML, RB; data collection, analysis and interpretation: all authors; LS-sequencing studies: MRS; generation of most of the figures and writing of the manuscript: HB, JG, LH; All authors read and approved the final version of the manuscript.

SUPPLEMENTARY MATERIAL ONLINE

Supplementary materials and methods

Supplementary figure legends

Figure S1. Mutation spectra of all LS-CRC cancer and adenoma samples

Figure S2. Group selective MS profiling

Figure S3. Microsatellite length distributions, mutation spectra and principal component analysis from sCRC (TCGA data)

Figure S4. Recurrent mutated genes (all somatic SNV and indels were considered) in (A) LS-CRC and (B) sCRC

Figure S5. Self-organizing maps (SOM) gallery of single sample expression portraits of all LS specimen studied

References

1. de la Chapelle A, Hampel H. Clinical relevance of microsatellite instability in colorectal cancer. *Journal of Clinical Oncology* 2010; **28**: 3380-3387.
2. Lagerstedt Robinson K, Liu T, Vandrovcova J*, et al.* Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer) Diagnostics. *Journal of the National Cancer Institute* 2007; **99**: 291-299.
3. Kastrinos F, Stoffel EM, Balmaña J*, et al.* Phenotype Comparison of MLH1 and MSH2 Mutation Carriers in a Cohort of 1,914 Individuals Undergoing Clinical Genetic Testing in the United States. *Cancer Epidemiology Biomarkers & Prevention* 2008; **17**: 2044-2051.
4. TCGA-consortium. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330-337.
5. Toyota M, Ahuja N, Ohe-Toyota M*, et al.* CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci USA* 1999; **96**: 8681-8686.
6. Maccaroni E, Bracci R, Giampieri R*, et al.* Prognostic impact of mismatch repair genes germline defects in colorectal cancer patients: are all mutations equal? *Oncotarget* 2015; **6**: 38737-38748.
7. Takemoto N, Konishi F, Yamashita K*, et al.* The Correlation of Microsatellite Instability and Tumor-infiltrating Lymphocytes in Hereditary Non-polyposis Colorectal Cancer (HNPCC) and Sporadic Colorectal Cancers: the Significance of Different Types of Lymphocyte Infiltration. *Japanese Journal of Clinical Oncology* 2004; **34**: 90-98.
8. Linnebacher M, Gebert J, Rudy W*, et al.* Frameshift peptide-derived T-cell epitopes: A source of novel tumor-specific antigens. *International Journal of Cancer* 2001; **93**: 6-11.
9. Echterdiek F, Janikovits J, Staffa L*, et al.* Low density of FOXP3-positive T cells in normal colonic mucosa is related to the presence of beta2-microglobulin mutations in Lynch syndrome-associated colorectal cancer. *OncoImmunology* 2016; **5**: e1075692.
10. Bernal M, Ruiz-Cabello F, Concha A*, et al.* Implication of the β2-microglobulin gene in the generation of tumor escape phenotypes. *Cancer Immunology, Immunotherapy* 2012; **61**: 1359-1371.
11. Guinney J, Dienstmann R, Wang X*, et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; **21**: 1350-1356.
12. Kim T-M, Laird Peter W, Park Peter J. The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes. *Cell* 2013; **155**: 858-868.
13. Angelova M, Charoentong P, Hackl H*, et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol* 2015; **16**: 1-17.
14. Boland CR. Recent discoveries in the molecular genetics of Lynch syndrome. *Familial Cancer* 2016; **15**: 395-403.
15. Mangold E, Pagenstecher C, Friedl W*, et al.* Spectrum and frequencies of mutations in MSH2 and MLH1 identified in 1,721 German families suspected of hereditary nonpolyposis colorectal cancer. *International Journal of Cancer* 2005; **116**: 692-702.
16. Engel C, Rahner N, Schulmann K*, et al.* Efficacy of Annual Colonoscopic Surveillance in Individuals With Hereditary Nonpolyposis Colorectal Cancer. *Clinical Gastroenterology and Hepatology* 2010; **8**: 174-182.

17. Gullotti L, Czerwitzki J, Kirfel J*, et al.* FHL2 expression in peritumoural fibroblasts correlates with lymphatic metastasis in sporadic but not in HNPCC-associated colon cancer. *Lab Invest* 2011; **91**: 1695-1705.
18. Schweiger MR, Kerick M, Timmermann B*, et al.* Genome-Wide Massively Parallel Sequencing of Formaldehyde Fixed-Paraffin Embedded (FFPE) Tumor Tissues for Copy-Number- and Mutation-Analysis. *PLOS one* 2009; **4**: e5548.
19. Andrew S. FastQC: a quality control tool for high throughput sequence data. *available online: http://wwwbioinformaticsbabrahamacuk/projects/fastqc* 2010.
20. Hoffmann S, Otto C, Kurtz S*, et al.* Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. *PLoS Comput Biol* 2009; **5**: e1000502.
21. McKenna A, Hanna M, Banks E*, et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297-1303.
22. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011; **27**: 2987-2993.
23. Alexandrov LB, Nik-Zainal S, Wedge DC*, et al.* Signatures of mutational processes in human cancer. *Nature* 2013; **500**: 415-421.
24. Wirth H, Loeffler M, von Bergen M*, et al.* Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics* 2011; **12**: 306.
25. Wirth H, von Bergen M, Binder H. Mining SOM expression portraits: Feature selection and integrating concepts of molecular function *BioData Mining* 2012; **5:18**.
26. Toronen P, Ojala P, Marttinen P*, et al.* Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. *BMC Bioinformatics* 2009; **10**: 307.
27. Subramanian A, Tamayo P, Mootha VK*, et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 2005; **102**: 15545-15550.
28. Liberzon A, Birger C, Thorvaldsdóttir H*, et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* 2015; **1**: 417-425.
29. Nersisyan L, Loeffler-Wirth H, Arakelyan A*, et al.* Gene set- and pathway-centered knowledge discovery assigns transcriptional activation patterns in brain, blood and colon cancer - A bioinformatics perspective. *Journal of Bioinformatics Knowledge Mining* 2016; **4**: 46-70.
30. Löffler-Wirth H, Kalcher M, Binder H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. *Bioinformatics* 2015; **31**: 3225-3227.
31. Briggs S, Tomlinson I. Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. *The Journal of Pathology* 2013; **230**: 148-153.
32. Markowitz S, Wang J, Myeroff L*, et al.* Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* 1995; **268**: 1336-1338.
33. Staffa L, Echterdiek F, Nelius N*, et al.* Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters. *PLOS one* 2015; **10**: e0121980.

34. Oliveira C, Westra JL, Arango D, *et al.* Distinct patterns of KRAS mutations in colorectal carcinomas according to germline mismatch repair defects and hMLH1 methylation status. *Human Molecular Genetics* 2004; **13**: 2303-2311.
35. Nagasaka T, Koi M, Kloor M, *et al.* Mutations in both KRAS and BRAF may contribute to the methylator phenotype in colon cancer. *Gastroenterology* 2008; **134**: 1950-1960.e1951.
36. Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* 2010; **138**: 2073-2087.e2073.
37. Hopp L, Wirth H, Fasold M, *et al.* Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. *Systems Biomedicine* 2013; **1**: 99-121.
38. LaPointe LC, Dunne R, Brown GS, *et al.* Map of differential transcript expression in the normal human large intestine. *Physiological Genomics* 2008; **33**: 50-64.
39. Olsen J, Gerds TA, Seidelin JB, *et al.* Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflammatory Bowel Diseases* 2009; **15**: 1032-1038 1010.1002/ibd.20879.
40. Grade M, Hörmann P, Becker S, *et al.* Gene Expression Profiling Reveals a Massive, Aneuploidy-Dependent Transcriptional Deregulation and Distinct Differences between Lymph Node–Negative and Lymph Node–Positive Colon Carcinomas. *Cancer Research* 2007; **67**: 41-56.
41. Binder H, Hopp L, Lembcke K, *et al.* Personalized Disease Phenotypes from Massive OMICs Data. In: Big Data Analytics in Bioinformatics and Healthcare. Baoying W, Ruowang L, William P, (ed)^(eds). IGI Global: Hershey, PA, USA, 2015; 359-378.
42. Kosinski C, Li VSW, Chan ASY, *et al.* Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc Natl Acad Sci USA* 2007; **104**: 15418-15423.
43. Lal N, Beggs AD, Willcox BE, *et al.* An immunogenomic stratification of colorectal cancer: Implications for development of targeted immunotherapy. *OncoImmunology* 2015; **4**: e976052.
44. Whitfield ML, Sherlock G, Saldanha AJ, *et al.* Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Molecular Biology of the Cell* 2002; **13**: 1977-2000.
45. Drescher KM, Sharma P, Watson P, *et al.* Lymphocyte recruitment into the tumor site is altered in patients with MSI-H colon cancer. *Familial Cancer* 2009; **8**: 231-239.
46. Surmann E-M, Voigt AY, Michel S, *et al.* Association of high CD4-positive T cell infiltration with mutations in HLA class II-regulatory genes in microsatellite-unstable colorectal cancer. *Cancer Immunology, Immunotherapy* 2015; **64**: 357-366.
47. Llosa NJ, Cruise M, Tam A, *et al.* The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discovery* 2015; **5**: 43-51.
48. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* 2012; **12**: 252-264.
49. le Rolle A-F, Chiu TK, Fara M, *et al.* The prognostic significance of CXCL1 hypersecretion by human colorectal cancer epithelia and myofibroblasts. *Journal of Translational Medicine* 2015; **13**: 199.

50. Doll D, Keller L, Maak M, *et al.* Differential expression of the chemokines GRO-2, GRO-3, and interleukin-8 in colon cancer and their impact on metastatic disease and survival. *International Journal of Colorectal Disease* 2010; **25**: 573-581.
51. Sconocchia G, Eppenberger-Castori S, Zlobec I, *et al.* HLA Class II Antigen Expression in Colorectal Carcinoma Tumors as a Favorable Prognostic Marker. *Neoplasia (New York, NY)* 2014; **16**: 31-42.
52. Kloor M, Huth C, Voigt AY, *et al.* Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study. *The Lancet Oncology* 2012; **13**: 598-606.
53. Qiu R, Sakato M, Sacho EJ, *et al.* MutL traps MutS at a DNA mismatch. *Proc Natl Acad Sci USA* 2015; **112**: 10914-10919.
54. Annese V, Piepoli A, Andriulli A, *et al.* Association of Crohn's disease and ulcerative colitis with haplotypes of the MLH1 gene in Italian inflammatory bowel disease patients. *Journal of Medical Genetics* 2002; **39**: 332-334.
55. Pokorny RM, Hofmeister A, Galandiuk S, *et al.* Crohn's disease and ulcerative colitis are associated with the DNA repair gene MLH1. *Ann Surg* 1997; **225**: 718-723; discussion 723-715.
56. Lawrence MS, Stojanov P, Polak P, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013; **499**: 214-218.
57. Esteller M, Fraga MF, Guo M, *et al.* DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. *Human Molecular Genetics* 2001; **10**: 3001-3007.
58. Deschoolmeester V, Baay M, Van Marck E, *et al.* Tumor infiltrating lymphocytes: an intriguing player in the survival of colorectal cancer patients. *BMC Immunology* 2010; **11**: 19-19.
59. West NR, McCuaig S, Franchini F, *et al.* Emerging cytokine networks in colorectal cancer. *Nat Rev Immunol* 2015; **15**: 615-629.
60. Schwitalle Y, Kloor M, Eiermann S, *et al.* Immune Response Against Frameshift-Induced Neopeptides in HNPCC Patients and Healthy HNPCC Mutation Carriers. *Gastroenterology* **134**: 988-997.
61. Ahadova A, von Knebel Doeberitz M, Bläker H, *et al.* CTNNB1-mutant colorectal carcinomas with immediate invasive growth: a model of interval cancers in Lynch syndrome. *Familial Cancer* 2016; **15**: 579-586.
62. Sekine S, Mori T, Ogawa R, *et al.* Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis. *Mod Pathol* 2017.
63. Lee J, Li L, Gretz N, *et al.* Absent in Melanoma 2 (AIM2) is an important mediator of interferon-dependent and -independent HLA-DRA and HLA-DRB gene expression in colorectal cancers. *Oncogene* 2012; **31**: 1242-1253.
64. Man Si M, Zhu Q, Zhu L, *et al.* Critical Role for the DNA Sensor AIM2 in Stem Cell Proliferation and Cancer. *Cell* 2015; **162**: 45-58.
65. Wilson JE, Petrucelli AS, Chen L, *et al.* Inflammasome-independent role of AIM2 in suppressing colon tumorigenesis via DNA-PK and Akt. *Nat Med* 2015; **21**: 906-913.
66. Scarpa M, Castagliuolo I, Castoro C, *et al.* Inflammatory colonic carcinogenesis: A review on pathogenesis and immunosurveillance mechanisms in ulcerative colitis. *World Journal of Gastroenterology : WJG* 2014; **20**: 6774-6785.
67. Ullman TA, Itzkowitz SH. Intestinal Inflammation and Cancer. *Gastroenterology* **140**: 1807-1816.e1801.

68. Mittal D, Gubin MM, Schreiber RD*, et al.* New insights into cancer immunoediting and its three component phases — elimination, equilibrium and escape. *Current Opinion in Immunology* 2014; **27**: 16-25.
69. Angelova M, Charoentong P, Hackl H*, et al.* The colorectal cancer immune paradox revisited. *OncoImmunology* 2016; **5**: e1078058.
70. Topalian SL, Hodi FS, Brahmer JR*, et al.* Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *New England Journal of Medicine* 2012; **366**: 2443-2454.
71. Le DT, Uram JN, Wang H*, et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine* 2015; **372**: 2509-2520.
72. Bindea G, Mlecnik B, Tosolini M*, et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 2013; **39**.
73. Szolek A, Schubert B, Mohr C*, et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 2014; **30**: 3310-3316.
74. Hundal J, Carreno BM, Petti AA*, et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Medicine* 2016; **8**: 11.
75. Lundegaard C, Lamberth K, Harndahl M*, et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucl Acids Res* 2008; **36**: W509-W512.
76. Ang P, Loh M, Liem N*, et al.* Comprehensive profiling of DNA methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features. *BMC Cancer* 2010; **10**: 227.
77. Arakelyan A, Nersisyan L. KEGGParser: parsing and editing KEGG pathway maps in Matlab. *Bioinformatics* 2013; **29**: 518-519.
78. Arana E, Vehlow A, Harwood NE*, et al.* Activation of the Small GTPase Rac2 via the B Cell Receptor Regulates B Cell Adhesion and Immunological-Synapse Formation. *Immunity* 2008; **28**: 88-99.
79. George J, Lim JS, Jang SJ*, et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* 2015; **524**: 47-53.
80. Marisa L, de Reyniès A, Duval A*, et al.* Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Med* 2013; **10**: e1001453.
81. Budinska E, Popovici V, Tejpar S*, et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology* 2013; **231**: 63-76.
82. Lenz G, Wright G, Dave SS*, et al.* Stromal Gene Signatures in Large-B-Cell Lymphomas. *New England Journal of Medicine* 2008; **359**: 2313-2323.
83. ENCODE. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**: 799-816.

References 73- 83 are cited solely in the Supplementary Material

Figure Legends

Figure 1: Genomic characteristics of LS-CRC. A) Patient- and mutation characteristics: The LS-CRC cases cluster into two groups, G1 and G2, differing in the number of somatic mutations, the type of mutations (more frameshift in G1), the type of constitutional mutation carriers (83% *MLH1* in G1 and 40% in G2) and the recurrently mutated genes (see also Figure S 4a). B) Microsatellite (MS) slippage profiles: 20 bp-monorepeats in the reference genome shorten on the average and show a bimodal length distribution with different peak positions in G1 and G2 LS-CRC and already in reference tissue. C) Mean MS length (MSL) as a function of mutational load: MSL and the mutational load separate G1 and G2. D) Mean mutation spectra: The sequence context of the SNV is characterized using mutation spectra [23] which are different in G1 and G2 as revealed by their difference spectra and, E) Principal component analysis (PCA) of mutation spectra.

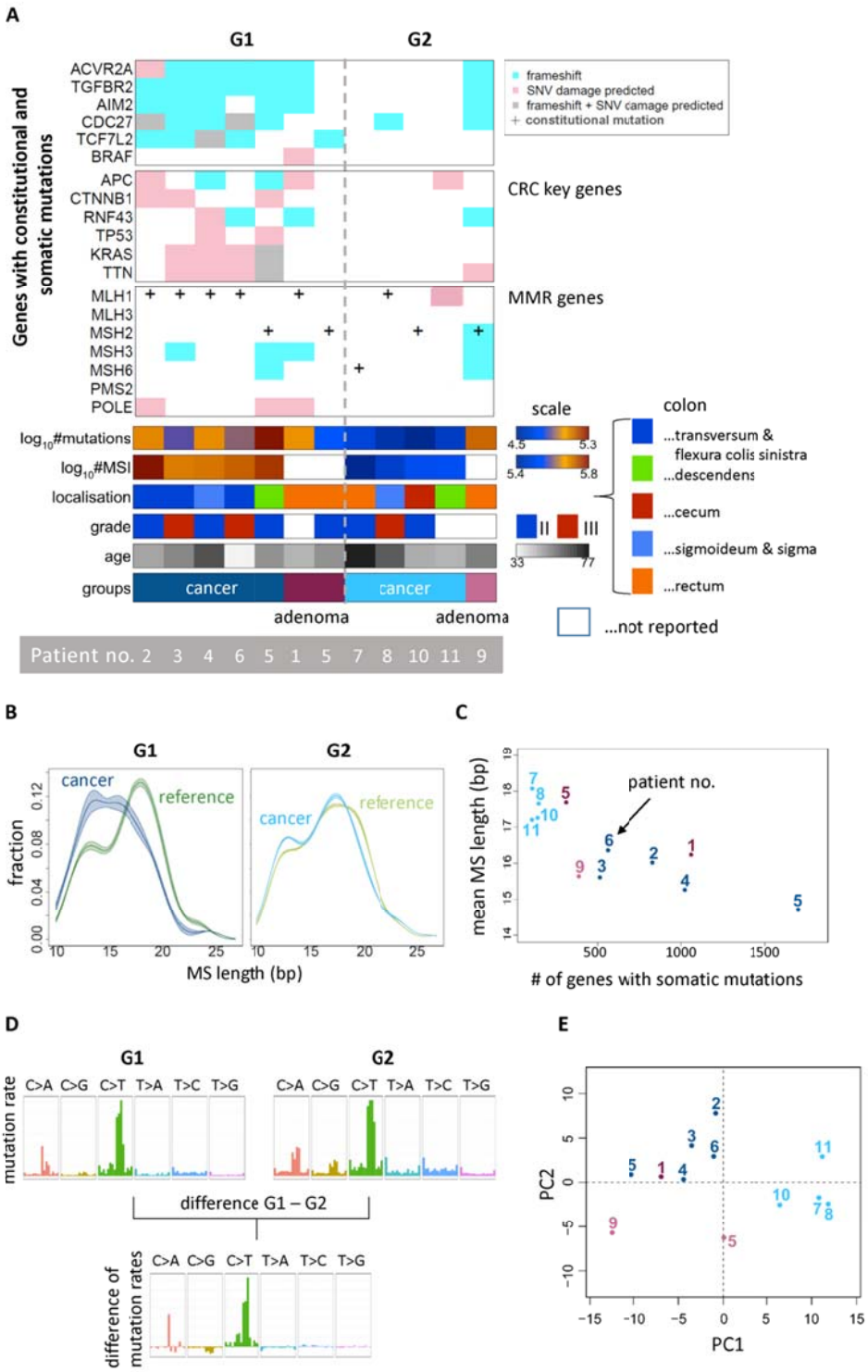

Figure 2: Heterogeneity of LS-CRC gene expression phenotypes: A) Mean SOM-portraits of reference mucosa, adenoma and cancer samples of each group. Red and blue 'spots' are clusters of over- and under-expressed genes, respectively. The difference portraits reveal clusters of differentially regulated genes. Their functional context is assigned

using gene set analysis. For lists of differentially expressed genes between the groups see Table S 7. B) A dose-response relationship is found between the Euclidian distance of the expression portraits of G1 cancers and their reference samples showing the inflammatory signature. C) 'Sample diversity maps of G1 (left) and G2 (right) specimen illustrate the heterogeneity of sample's transcriptomes in two-dimensional plots: Reference samples show distinct expression signatures ('mucosa', 'inflammation' and 'stroma', see text). This diversity decreases for adenoma and cancer of G1 which are characterized by 'oxphos' and 'proliferation' signatures, respectively. Tumor and reference samples collected from the same patient are connected by arrows.

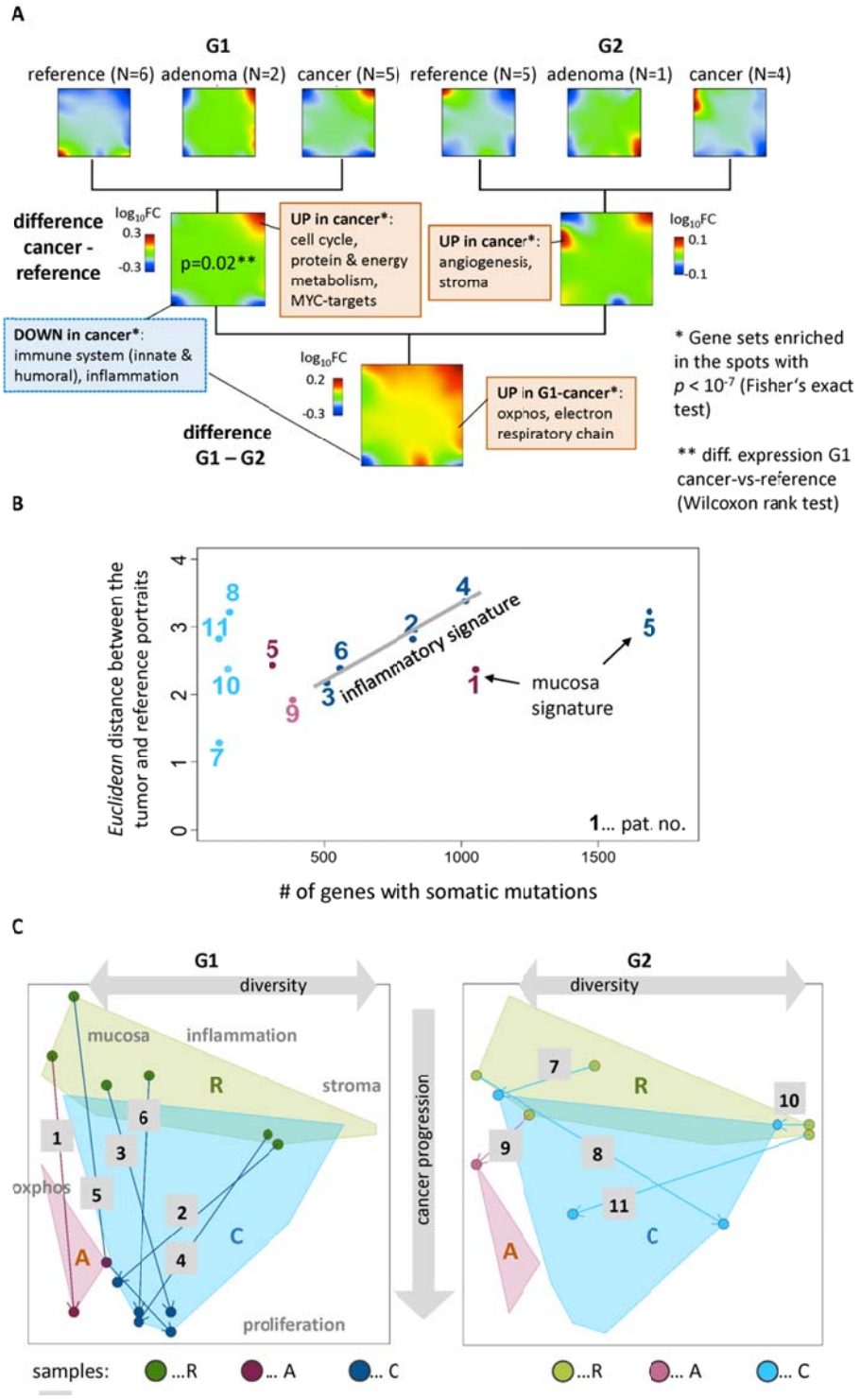

Figure 3: Expression signatures of LS-CRC: *A*) Gene set expression profiles associates samples of G1 and G2 with selected functional categories. Gene sets were taken from [24,27,28,40-42,44]. B) The expression heatmap of mutated genes shows that mutations reduce the expression level in most cases. C) The SOM-summary map provides an overview over all relevant spot clusters of co-expressed genes (red) observed in the SOM portraits (compare with Figure 2A). Spots are labelled A – F. Different regions of the map can be assigned to spots upregulated in the different tissue types as indicated. Each spot was functionally annotated using gene set analysis (see also Table S 6, Figure S 8). Four major expression modules (stroma, inflammation, mucosa and proliferation) were extracted.

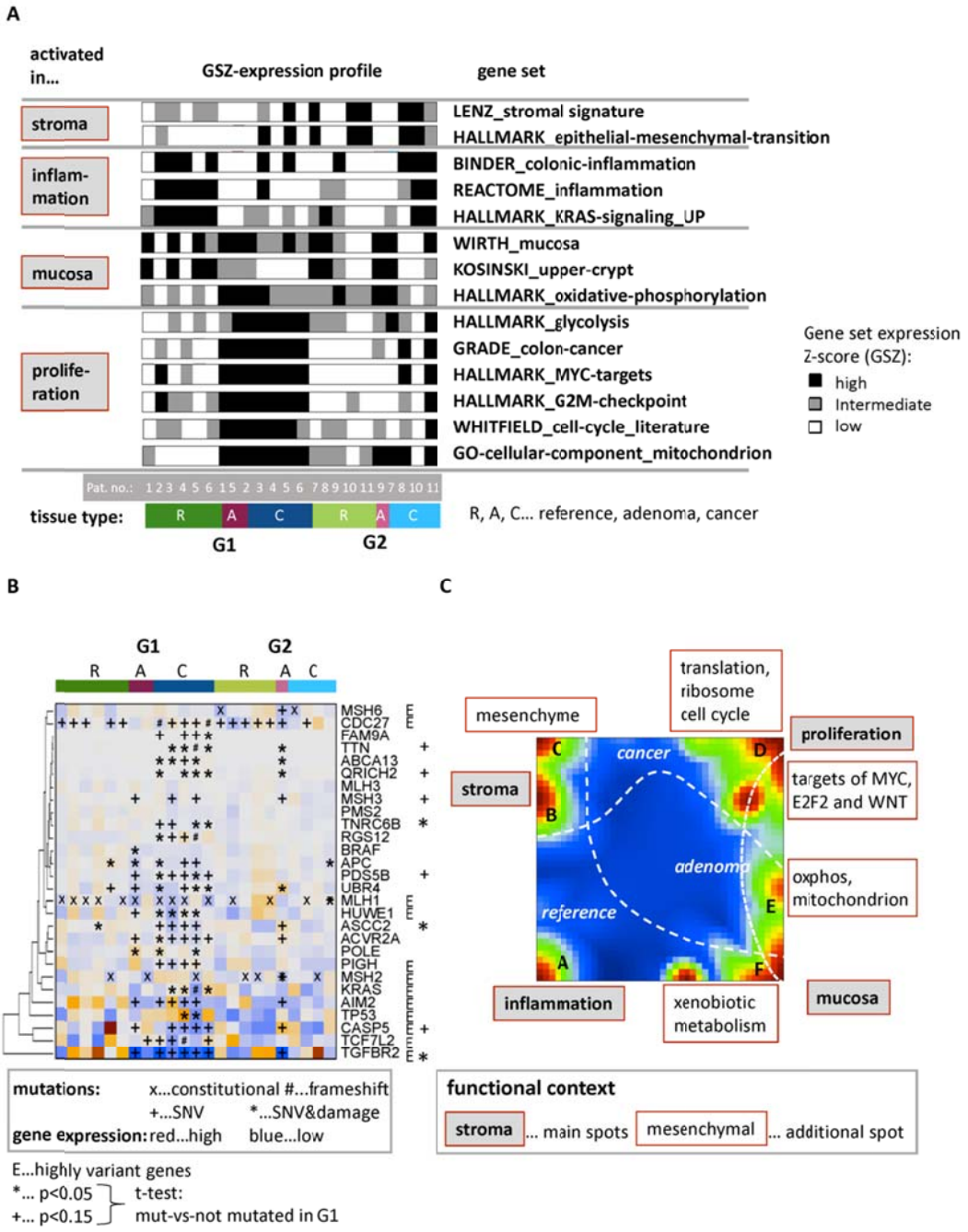

Figure 4: Inflammation characteristics of LS-CRC. A) Gene sets associated with immune response are highly expressed in reference tissue of G1 (dark green bars, colors are assigned in B). Overexpression with respect to tumors of G1 and with respect to all tissues of G2 were estimated using Wilcoxon test for paired and unpaired samples, respectively. The gene set maps indicate accumulation of these genes in spot A. B) Expression heatmap and SOM of chemokine receptors and ligands. They divide into clusters assigned to pro- and anti-tumoral inflammation as indicated by the accumulation of the respective genes in the spot modules A and D, respectively. Co-expression of chemokine receptor-ligand pairs in spot A suggests tumor suppressing inflammation by infiltration of CD4 and CD8 T-cells and of CD19 B-cells. C) *HLA* class II genes exclusively accumulate in spot A together with their regulators CIITA and RFX5, the inflammasome gene *AIM2* and immune-checkpoint genes thus indicating their co-regulation with the immune response signature and particularly their deactivation in G1 tumors. No recurrent somatic mutations were found in the *HLA*-regulators except A*IM2*. D) Expression of *HLA* class II genes and the number of neoantigenes in LS-CRC correlate with the mutation load in LS-CRC. E) (20x magnification) and F) (10x) CD4-immunohistochemistry of LS colon sections: G1 shows strong invasion of CD4+ T cells (brown color) in reference mucosa but not in tumor.

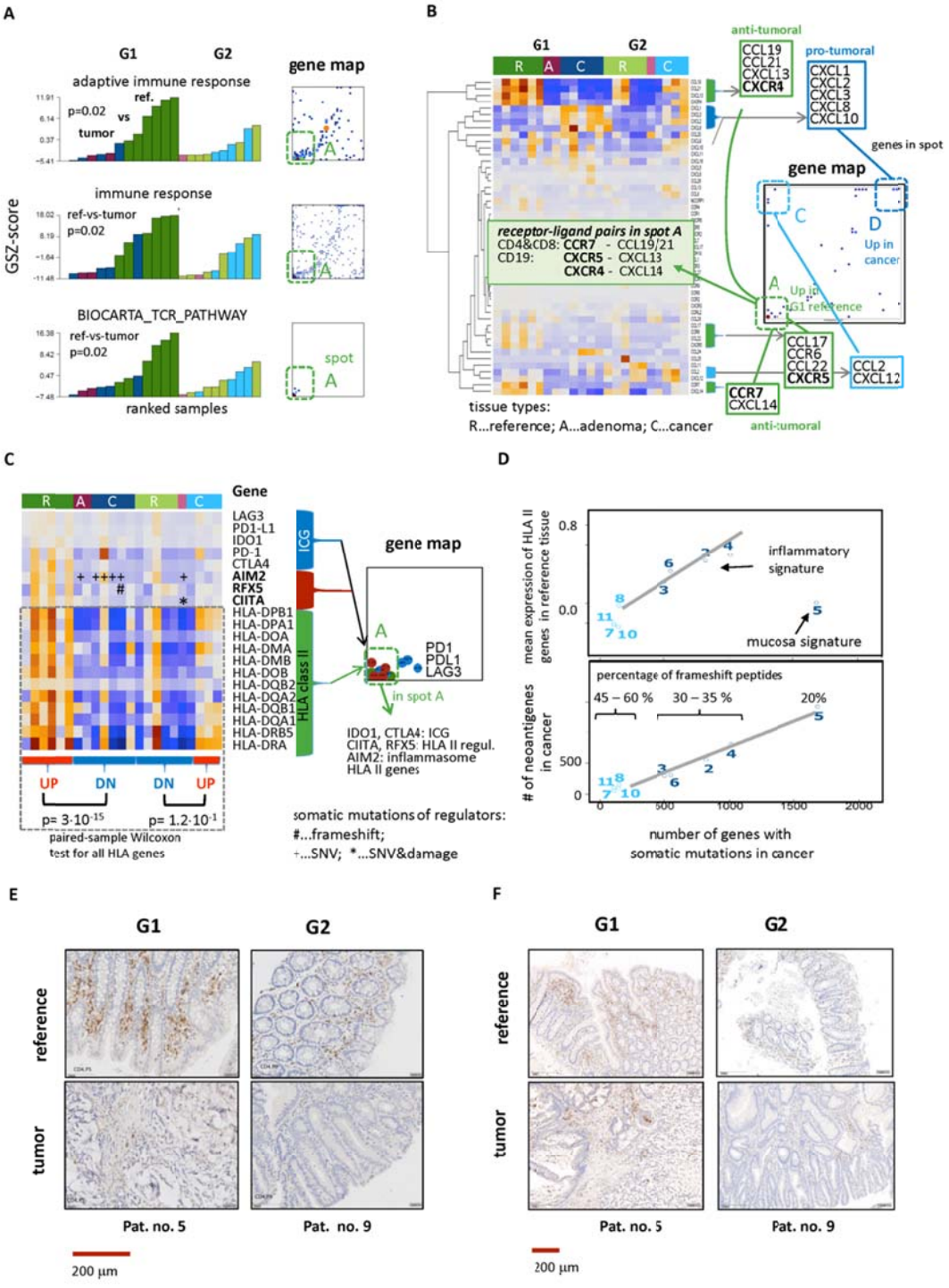

Figure 5: **Scheme illustrating immune-editing in the two LS groups arising in colon:** G1 LS-CRC cells occurring in the reference mucosa are edited for an escape from a highly immunogenic microenvironment by effectively suppressing immune response by loss of presentation of *HLA*-antigens and by T-cell exhaustion via expression of immune-checkpoint genes in LS-mucosa and their deactivation in tumor. In contrast, G2 LS-CRC cells develop in a less immunogenic microenvironment where tumor promoting inflammation parallels formation of overt tumor. Immunogenicity is moderate in G1 and G2 tumors.