Whole Exome Sequencing-Based Analysis Identifies DNA Damage Repair Deficiency as a Major Contributor to Gliomagenesis in Adult Diffuse Gliomas

Authors: Ege Ülgen MD[1], Özge Can PhD[2], Kaya Bilguvar MD[3], Yavuz Oktay PhD[4,5], Cemaliye B. Akyerli PhD[6], Ayça Erşen Danyeli MD[7], M. Cengiz Yakıcıer MD[8], O. Uğur Sezerman PhD[1], M. Necmettin Pamir MD[9], Koray Özduman MD[9]

Affiliations: Acibadem Mehmet Ali Aydinlar University, School of Medicine, Departments of Biostatistics and Medical Informatics[1], Medical Biology[6], Pathology[7], and Neurosurgery[9], Istanbul, Turkey

Acibadem Mehmet Ali Aydinlar University, Faculty of Engineering, Department of Medical Engineering, Istanbul, Turkey[2]

Yale University, School of Medicine, Department of Genetics, Yale Center for Genome Analysis, New Haven, CT, USA[3]

Izmir Biomedicine and Genome Center (IBG), Izmir, Turkey[4]

Dokuz Eylul University, Faculty of Medicine, Department of Medical Biology, Izmir, Turkey[5]

Acibadem Mehmet Ali Aydinlar University, School of Arts and Sciences, Department of Molecular Biology, Istanbul, Turkey[8]

Corresponding Author:

Koray Ozduman MD, Professor of Neurosurgery

Acıbadem Mehmet Ali Aydinlar University, School of Medicine

Department of Neurosurgery

Address: Altunizade Mahallesi, Yurtcan Sok. No:1

Üsküdar, Istanbul-34662-Turkey

Phone: +90(533)7673365

Email: Koray.ozduman@icloud.com

Keywords: glioma; mutational signatures; DNA repair; exome sequencing

Running title: DNA Damage Repair Deficiency Has a Critical Role in Gliomagenesis

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Portions of this work were presented in part at the 2018 American Association of Neurological Surgeons (AANS) Annual Scientific Meeting, April 28-May 2, 2018, New Orleans and received the AANS/CNS Joint Section on Tumors 2018 Brainlab Award.

Abstract word count = 299; text word count = 3951; number of references = 39; number of tables and figures = 6; number of supplementary tables and figures = 4

1. Abstract

Object

Processes that cause or contribute to cancer, such as aging, exposure to carcinogens or DNA Damage Repair deficiency (DDRd) create predictable and traceable nucleotide alterations in one's genetic code (termed as mutational signatures). Large studies have previously identified various such "mutational signatures" across cancers that can be attributed to the specific causative processes. To gain further insight into the processes in glioma development, we analyzed mutational signatures in adult diffuse gliomas (DG).

Methods

25 DG and paired blood samples were whole exome sequenced. Somatic mutational signatures were identified using two different methods. Associations of the signatures with age at diagnosis, molecular subset and mutational load were investigated. As DDRd-related signatures were frequently observed, germline and somatic DDR gene mutations as well as microsatellite instability (MSI) status were determined for all samples. For validation of signature prevalence, publicly-available The Cancer Genome Atlas (TCGA) data was used.

Results

Each tumor had a unique combination of signatures. Most common signatures were signature-1 (88%-aging related), signature-3 (52%-homologous recombination related) and signature-15 (56%-mismatch repair related). 84% of the tumors contained at least one DDRd signature. The findings were validated using public TCGA-data. Signature 1's weight positively correlated with age (r=0.43) while cumulative-weight of DDR deficiency signatures negatively correlated with age (r=-0.16). Each subject had at least one germline/somatic alteration in a DDR gene, the most common being the risk SNP rs1800734 in *MLH1.* rs1800734 AA genotype had higher cumulative DDRd weight as well as higher mutational load. TP53 was the most common somatically-altered DDR gene. MSI was observed in 24% of the tumors. No significant associations of MSI status with mutational load, rs1800734 or the cumulative weight of DDRd signatures were identified.

Conclusions

Current findings suggest that DDRd may act as a fundamental mechanism in gliomagenesis rather than being a random, secondary event.

Keywords: glioma; mutational signatures; DNA repair; exome sequencing

2. Introduction

Evidence on the cause of diffuse gliomas remains limited to this date. The only firmly-established risk factor is ionizing radiation exposure and the only widely-accepted protective factor is history of allergy/atopic disease[29]. No environmental carcinogens are proven to cause gliomas. The contribution of heredity is thought to be limited: Several rare Mendelian cancer-susceptibility syndromes such as Lynch, Li-Fraumeni, NF1, NF2, TS, melanoma-neural system tumor syndrome and Ollier's disease are linked to markedly increased glioma risk. However, these make up only a small fraction of glioma patients[27,31]. Additionally, several genetic polymorphisms, that are commonly observed in the general population, are known to modulate glioma risk but their ultimate contribution to gliomagenesis remains unknown[24,28]. In short, the cause of gliomas remains largely unaccounted for.

We hypothesized that analyzing the signatures of mutational processes in adult diffuse gliomas (DG) would provide further insight into gliomagenesis. The somatic mutations in a given cancer are the result of the DNA damage caused by multiple processes such as aging, exposure to carcinogens or DNA Damage Repair deficiency (DDRd), each of which has a different way of changing the genetic code. One or various of these mechanisms act in concert at various stages of a person's life and result in the cumulative genetic damage[6]. In certain cases, the cumulative damage inflicted on the inherited background is sufficient to cause cancer. Large-scale cancer genomics analyses have defined over 30 mutational signatures, each of which have been attributed to a specific process causing alterations in the genetic code[4,5]. Current bioinformatics tools can identify the contribution of each of these mutational signatures in any given cancer genome. Up to date, only two common mutational processes were detected in gliomas, including age-related and alkylating-agent-induced signatures[3-5]. However, these two processes alone fail to account for common clinical observations in gliomas, such as therapeutic response to DNA damaging therapies (radiotherapy or alkylating agents), mutator-phenotype, malignant transformation and high incidence of alkylating-agent-induced bone marrow toxicity.

Utilizing a curated list of signatures, we analyzed the paired whole exome sequencing (WES) results of 25 DG patients. Our hypothesis was that there would be one or more mutational processes common to most DGs, which would correlate with clinical factors.

3. Methods

Clinical Characteristics of the Cohort

The present study is an analysis of the WES results of 25 adult patients (median age 51; range 20-76), operated for DGs. Characteristics of the patients and their tumors are presented in Table 1.

Pathological Classification of Tumors

All pathological specimens were retrospectively re-reviewed by a single neuropathologist (AED) based on the 2016 World Health Organization (WHO) classification of tumors of the central nervous system criteria[23]. For all tumors, mutations in *IDH1/2*, *TERT* (228 and 250 promoter mutations) and *H3.3* (K27, G34 mutations) were tested using mini-sequencing or Sanger sequencing. For WHO grade II IDH-mutant tumors, 1p/19q co-deletion was assessed using microsatellite marker analysis. For cases with canonical glioblastoma molecular findings on WES/Sanger sequencing (including: (i) absence of IDH mutation, (ii) presence of TERT promoter mutations, (iii) chromosome 7 and/or EGFR amplifications and (iv) chromosome 10 and/or PTEN deletions), the diagnosis was determined as “Glioblastoma, IDH wild type”, even if morphological findings did not support a WHO grade IV tumor.

Ethics and Consent of Clinical Materials

The study was approved by Acibadem Mehmet Ali Aydinlar University’s institutional review board (ATADEK-2018/7). Written consent was obtained from all subjects prior to their inclusion in the study.

DNA Extraction

DNA was extracted from snap-frozen tumor and peripheral venous blood samples. For each tumor specimen submitted for whole-exome sequencing, sections were reviewed by a neuropathologist to confirm the diagnosis of DG and the adequacy for sequencing was assessed. DNA was extracted using the DNeasy Blood & Tissue Kit (QIAGEN).

Whole Exome Sequencing

Sequencing of the libraries were performed on Illumina HiSeq instruments using paired-end reads. We achieved mean target coverage of 122.924 and 84.08, for

tumor and blood, respectively. The average percentage of reads with at least 25X coverage was 89.78% and 82.15% for tumor and blood, respectively.

All tumor and blood WES data is available under the European Genome-Phenome Archive accession EGAS00001003035. The capture kit, read length, mean target coverage and percentage of reads with at least 25X coverage are provided in Supplementary Table S1.

The reads were aligned to the human reference genome (assembly hg19) using BWA – MEM algorithm (version 0.7.15-r1140)[20]. The mapped reads were cleaned with Picard – CleanSam. (Picard version 2.9.0-1 http://broadinstitute.github.io/picard/). Cleaned reads were sorted and mate information was fixed using Picard. MarkDuplicates algorithm from Picard was used for marking PCR duplicates. Next, base quality scores were recalibrated using the BQSR tool from the Genome Analysis Toolkit (GATK version 3.8-1-0-gf15c1c3ef).

Variant Calling

Germline single nucleotide variant (SNV)/insertion-deletion (INDEL) calling was performed using HaplotypeCaller from GATK. Somatic SNV/INDEL calling was performed using MuTect2 from GATK.

Somatic copy number alterations (SCNAs) were identified using ExomeCNV[34]. Genes located in SCNA segments were annotated using an in-house script.

Somatic mutational loads (mutations/Mb) were calculated as:

$$\frac{total\ number\ of\ somatic\ mutations}{total\ exome\ length\ (bases)} \times 10^6$$

Molecular Subsets

Molecular subset of each tumor was determined based on the presence of somatic mutations in the genes *TERT* and *IDH1/IDH2* as described by previous studies[2,11]. If a tumor contained any *TERT* promoter mutations (namely, C228T or C250T), it was classified as “TERT-only”. If the tumor carried a mutation in *IDH1* or *IDH2*, it was classified as “IDH-only”. If it didn’t contain any of these mutations, it was classified as “double-negative”.

Mutational Signatures

All statistical analyses were performed using R[37]. Transition/Transversion (Ti/Tv) frequencies were calculated and visualized using GenVisR[36].

For mutational signature analysis, “Signatures of Mutational Processes in Human Cancer” from COSMIC [13] was used. The weights of each mutational signature contributing to the total mutations of individual tumor samples were determined using DeconstructSigs[33]. DeconstructSigs is a method to determine the contribution of each mutational signature in a single tumor. It applies a multiple linear regression model with the caveat that any coefficient must be greater than 0. Through an iterative approach, the optimal weight for each signature is calculated. These weights are then normalized between 0 and 1. Finally, the weights of signatures with normalized weight < 0.06 are set to 0 because the majority of false positives had weights < 0.06 in the original study of DeconstructSigs.

The signature weights obtained using DeconstructSigs were validated by another approach called YAPSA (Huebschmann D, Gu Z, Schlesner M (2015). YAPSA: Yet Another Package for Signature Analysis. R package version 1.6.0) that uses linear combination decomposition to calculate the weights of signatures per individual tumor. The positive correlations of weights of signatures identified in this cohort by these tools are presented in Supplementary Figure 1 and Supplementary Table S2.

The signatures identified in each patient were then analyzed with regards to clinical parameters. To find any association between the mutational signature weights and the molecular subsets, multinomial logistic regression models were fitted using nnet[39].

Microsatellite Instability

Microsatellite instability (MSI) status of the tumors were determined using the MSIseq[17]. Traditionally, tumors in which ≥40% of the markers in a panel of repeat markers show somatic length mutations are termed MSI-high (MSI-H). Tumors in which no marker shows somatic length mutations are termed microsatellite stable (MSS). The remaining tumors are termed MSI-low (MSI-L). MSIseq provides a classifier determining the MSI status of a tumor as MSI-H or non-MSI-H. This classifier uses the following features: (i) count of INDELs/megabase in simple sequence repeats, (ii) count of SNVs/megabase and (iii) mutation count/megabase in simple sequence repeats.

Validation with The Cancer Genome Atlas Glioma Mutation and Clinical Data

The Cancer Genome Atlas (TCGA) – glioblastoma multiforme (GBM) and lower-grade glioma (LGG) somatic mutation annotation files and clinical data were obtained through the International Cancer Genome Consortium Data Portal release 26 (https://dcc.icgc.org/releases/release_26/Projects/GBM-US for GBM and https://dcc.icgc.org/releases/release_26/Projects/LGG-US for LGG). 2 subjects who were younger than 18 years old were excluded, resulting in a total number of 564 subjects. The somatic mutational signatures in the combined TCGA cohort were identified using DeconstructSigs. The overall percentages of all mutational signatures in the TCGA cohort and ours were then compared using Spearman's rank correlation test.

Germline and Somatic Alterations in DNA Damage Repair Genes

The list of human DDR genes and the corresponding annotations of functions were obtained from the curated table “Human DNA Repair Genes” (https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html)

Somatic non-synonymous variants (INDELs and “missense”, “non-sense” and “splice site” mutations), somatic homozygous losses (defined as Tumor/Normal ratio < 0.5) and germline risk SNPs in these DDR genes were identified and further investigated. The germline risk SNPs were identified from the list of SNPs annotated as “pathogenic” in NCBI - dbSNP (https://www.ncbi.nlm.nih.gov/snp).

4. Results

Characteristics of the 25 DG patients and their tumors are presented in Table 1. The median age at diagnosis was 51 with a wide range (20-76 years old). The majority of tumors were WHO grade IV DGs (n=20, 80%), followed by WHO grade II (n=3, 12%) and WHO grade III (n=2, 8%). Most tumors were in the TERT-only molecular subset (n=12, 48%) while the double-negative subset had 8 tumors (32%) and the IDH-only subset had 5 tumors (20%).

Somatic mutational load of the tumors ranged from 2.78 to 13.48 mutations/Mb and the median was 6 mutations/Mb. The Ti/Tv compositions of these somatic mutations were also markedly heterogenous (Figure 1A). There were no significant differences

in mean somatic mutational loads of the different molecular subsets (Analysis of Variance (ANOVA) p=0.43).

In total, 21 different mutational signatures were identified. The weights of these signatures in each tumor are presented in Figure 1B, along with the molecular subsets, WHO grades, pathological diagnoses, ages at diagnosis and mutational loads. Overall, each tumor had a unique combination of signatures.

We investigated whether any signature is associated with any molecular subsets. Using signature weights as predictor variables, two multinomial logistic regression models were fitted. To decrease the sparsity and prevent convergence issues, only signatures that had non-zero weight in at least 20% were used. In the first model, the odds of each subject belonging to IDH-only or double-negative vs. TERT-only were calculated. In the second model, the odds of each subject belonging to IDH-only or TERT-only vs. double-negative were calculated. It was identified that none of the selected signatures could be used to differentiate between TERT-only and IDH-only nor between double-negative and IDH-only cases. It was determined that one percent increase in the weight of signature-1 was associated with 13% increase in the odds of being in the TERT-only compared to being in the double-negative subset (p=0.04).

The percentages of subjects harboring each signature are presented in Table 2. Signatures-1, -15 and -3 were observed in the majority of the subjects (88%, 56% and 52%, respectively). Signature-1, associated with age at diagnosis, was frequently observed in each subset (91.67% in the TERT-only subset, 60% in the IDH-only subset, 100% in the double-negative subset). Signature-15, associated with defective DNA mismatch repair, was identified in 50% of the TERT-only subset, 40% of the IDH-only subset and 75% of the double-negative subset. Signature-3, associated with failure of DNA double-strand break-repair by homologous recombination, was present in 60% of the IDH-only subset, and half of the cases in each of the remaining subsets.

As 2 out of the 3 most common mutational signatures were associated with DDR deficiency (DDRd), DDRd signatures were further analyzed. The DDRd signatures in this cohort were signatures 3 (52%), 6 (16%), 15 (12%) and 26 (12%). The overall median cumulative weight of DDRd signatures was 0.2 (range=0-0.97). Median

cumulative weights of DDRd signatures for the double-negative, IDH-only and TERT-only subsets were 0.18 (range=0.13-0.51), 0.39 (range=0-0.6) and 0.21 (range=0-0.97), respectively. Overall, 84% of the tumors contained at least one DDRd signature. Individual analyses of the molecular subsets revealed 100% of the double-negative, 83.33% of the TERT-only and 60% of the IDH-only subset had at least one DDRd signature.

To validate our findings, mutational signatures in the TCGA cohort were determined. The percentages of patients that harbor a given mutational signature in our cohort and the TCGA cohort are presented as a heatmap in Figure 2A. Spearman's rank correlation test revealed that the percentages were positively correlated (ρ=0.35, p=0.058), indicating that most mutational signatures were observed in similar proportions. In addition to the DDRd signatures-3, -6, -15 and -26, TCGA contained signature-20, which is associated with defective DNA mismatch repair. Therefore, signature-20 was included in the DDRd signatures list for further analyses. Supporting our findings, it was observed that 70.5% of TCGA cases had at least one DDRd signature.

Investigating the association of signatures with age, it was discovered that the weight of signature-1 correlated positively with age at diagnosis in both our cohort and TCGA cohort (r=0.43 and p=0.031 for our cases, r=0.25 and p<0.001 for TCGA cases) (Figure 2B). Because the individual weights of DDRd signatures did not correlate with age at diagnosis significantly, except for signatures-15 and -20 in the TCGA cohort (Supplementary Table S3), the association between the cumulative weight of DDRd signatures and age at diagnosis was analyzed. This analysis yielded a negative correlation between the cumulative weight of DDRd signatures and age at diagnosis (r=0.16, and p=0.44 for our cases, r=-0.16 and p<0.001 for TCGA cases) (Figure 2C).

To detect any DNA alteration that might alter DDR mechanisms and take part in gliomagenesis, we determined the germline risk SNPs, somatic non-synonymous variants and somatic homozygous deletions in DDR genes in our 25 cases (Figure 3). We observed that all cases harbored at least one such alteration. 72% of cases (n=18) carried at least one germline risk polymorphism and 72% (n=18) had at least one somatic alteration.

The most frequently observed alteration was the germline risk SNP rs1800734 in *MLH1* (n=15, 60%). The allele frequency of rs1800734-A in gnomAD[19] was 0.23. Allelic counts of rs1800734 in gnomAD (n=30952) and our DG cohort (n=25) were compared to replicate the previously-shown association of rs1800734-A with DG [32]. Fisher's exact test revealed that rs1800734-A was indeed associated with DGs (odds ratio=4.94, 95% CI=2.07-12.30, p<0.001).

ANOVA indicated that the mean cumulative weights of DDRd signatures of rs1800734-GG (n=10, 40%), rs1800734-AG (n=12, 48%) and rs1800734-AA (n=3, 12%) of rs1800734 were different (p=0.021, Figure 4A). Post-hoc Tukey honest significant differences of means were calculated: AA – GG difference in cumulative weight was 0.40 (95% confidence interval (CI)=0.07-0.74, adjusted-p=0.01), AA – AG difference was 0.30 (95% CI=-0.03-0.63, adjusted-p=0.08) and AG – GG was 0.10 (95% CI=-0.11-0.32, adjusted-p=0.46). Therefore, homozygous risk subjects had a higher weight than both non-risk and heterozygous subjects.

ANOVA of mean mutational load in rs1800734-GG, rs1800734-AG and rs1800734-AA showed that there was a significant difference in the means (p=0.041, Figure 4B). Mean mutational load of AA subjects was 4.25/Mb higher than AG subjects (95% CI=0.27-8.23/Mb, adjusted-p=0.03). Mean mutational load of AA subjects was 2.86/Mb higher than the GG subjects, although this difference was not statistically significant (95% CI=-1.12-6.84/Mb, adjusted-p=0.19). Finally, there was no statistically significant difference in mean mutational load between AG and GG subjects (95%CI=-4.00-1.21/Mb, adjusted-p=0.39).

Alterations in the genes *TP53* (48%) and *POLG* (16%) were also relatively frequently observed. When t-tests were performed to compare the cumulative weights of DDRd signatures between subjects with and without germline/somatic variants in the genes *TP53* and *POLG*, no differences were identified (p=0.67 for *TP53*, p=0.81 for *POLG*). Additionally, no significant difference in mutational load was observed neither between subjects who harbored a variant in *TP53* and those who did not (p=0.34) nor between subjects who harbored a variant in *POLG* and those who did not (p=0.32).

We next determined the MSI status of the 25 tumors. 24% of the cohort was identified as MSI-H (Figure 3). The mean mutational load of non-MSI-H tumors was

6.08. The mean mutational load of MSI-H subjects was slightly higher, 7.88 (Wilcoxon-Mann-Whitney p=0.13). Pearson's Chi-squared test showed no significant association between different genotypes of rs1800734 and MSI status (p=0.81). 18.18% of wild-type subjects, 27.27% of heterozygous subjects and 33.33% of subjects homozygous for rs1800734 were MSI-H. Wilcoxon-Mann-Whitney test revealed there was no significant difference in the cumulative weights of DDRd signatures between MSI-H and non-MSI-H tumors (p=0.9). The mean cumulative weights of DDRd signatures were 0.26 and 0.33 for non-MSI-H and MSI-H, respectively.

5. Discussion

Despite strong individual variation, recurrent mutational signatures were identified

In this cohort, each tumor presented with a unique set of mutational signatures, where the weight of each signature varied considerably. However, there were also recurrent findings: Signature-1, signature-3 and signature-15 were the most commonly observed signatures in the cohort and each was present in more than 50% of the cases. Consistent with previous studies, the most commonly observed signature was Signature-1, which was causally related to spontaneous deamination of 5-methycytosine[15]. Our cohort size is limited to 25 but the incidence of signatures in the current cohort and those that were observed in the TCGA cohort of 564 cases (glioblastomas and lower-grade gliomas combined) were sufficiently correlated.
As glioma molecular subsets (based on the presence of *IDH* and *TERT* mutations) differ significantly in their patient and clinical characteristic as well as traits, we looked for specific mutational signatures or specific signature that could discriminate between the molecular subsets but failed to show any[2,11]. The three most common signatures of the whole cohort (signatures-1, signature-3 and signature-15) were also the most commonly observed signatures in each glioma molecular-subset (found in over 40% in each molecular subset).

DNA-Damage-Repair-related signatures are very common in diffuse gliomas

We noticed that the observed signatures were enriched in DDRd processes. 84% of the cases carried at least one DDRd signature. This was consistent in all molecular subsets: 60% of IDH-only, 83.3% of TERT-only and 100% of double-negative tumors

had at least one DDRd-signature. We confirmed this finding in the TCGA cohort where 70.5% of tumors had at least one DDRd signature. This finding is consistent with several common clinical observations in gliomas and support a possible role for DDRd in gliomagenesis including: (i) Markedly increased glioma incidence in Mendelian inherited DDRd syndromes such as Lynch or Li-Fraumeni, (ii) increased glioma risk after radiation exposure, (iii) therapeutic response to DNA damaging agents (alkylating agents, radiotherapy), (iv) quick acquisition of novel genetic alterations (mutator phenotype), and (v) almost universal malignant degeneration over time in lower-grade gliomas. Additionally, bone marrow suppression by alkylating agents in some but not all patients may also indicate a germline DDRd. The classic description of mutator phenotype proposed in late 1970's postulated that loss of DNA repair genes led to genomic instability[22,26]. With the widespread use of next generation sequencing, DDRd was documented to be induced by both germline and somatic aberrations in common solid cancers, such as breast and prostate cancers, and it was associated with therapeutic response[7,30]. Our findings present a similar observation in adult diffuse gliomas.

DNA-Damage-Repair-related signatures correlated negatively with age at diagnosis

Alexandrov et al.[3] showed that some mutational signatures correlated positively with age at diagnosis in most cancers including glioblastomas and lower-grade gliomas. The weight of signature 1 correlated positively with age at diagnosis both in our cohort and in the TCGA cohort. In this analysis, we also observed that the cumulative weight of DDRd signatures correlated negatively with age at diagnosis, also replicated in the TCGA cohort. As inherited DDRd results in increased cancer susceptibility and earlier cancer onset, this finding was not unexpected.

Both germline and somatic alterations in DNA Damage Repair genes are common in gliomas

Analysis of mutational signatures is a fairly novel approach and our knowledge on the subject is limited. Therefore, further evidence had to be provided to support the observed findings. To search for a causative factor underlying frequent DDRd signatures, we analyzed germline and somatic genetic alterations (single nucleotide changes or copy number changes) in a set of established DDR genes. For germline alterations, only those that are previously shown to have a significant pathogenic effect were determined. Using these criteria, we identified that every patient had at

least one germline or somatic alteration in a DDR gene. Nearly three-quarters of the subjects had at least one germline risk SNP in a DDR gene and nearly three-quarters of the subjects had at least one somatic variant in a DDR gene. The DDR genes that harbored these variants had various functions. While a gene may have multiple roles in DDR, the most prominent role of each DDR gene was indicated in this study. The most noteworthy DDR function was mismatch repair activity as the highest number of subjects with variants was observed in the *MLH1* gene. 60% of the subjects harbored the germline risk SNP rs1800734-A allele in *MLH1* in either heterozygous or homozygous fashion. This risk SNP is located in the 5' UTR of *MLH1* and was previously shown to be associated with *MLH1* CpG island hypermethylation and expression loss in colorectal cancer[35]. Additionally, AA or AG vs. GG genotype of rs1800734 was also shown to cause a heritable predisposition to epigenetic silencing of *MLH1*[9]. The risk SNP was identified as a glioblastoma susceptibility factor[32] and we replicated this association by showing AA or AG vs. GG genotype is associated with DG. We also demonstrated that the AA genotype had the highest cumulative weight of DDRd signatures as well as the highest mutational load.

Several Mendelian-inherited DDR genes are established glioma predisposition genes (such as TP53, MLH1, MSH2, MSH6, PMS2). Contribution of DDR-gene polymorphisms to polygenic-inherited glioma predisposition was also previously reported in the literature: Although only TP53 and MDM4 polymorphisms stand out in glioma-GWAS studies, several candidate gene analyses have indicated associations between numerous other DDR gene polymorphisms and adult glioma development[1,21,24]. In their analysis of coding pathogenic germline variants of the TCGA pan-cancer cohort, Huang *et al.*[16] reported that 6 of the 8 common pathogenic polymorphisms in GBM and 4 of the 6 common pathogenic polymorphisms in lower-grade gliomas were in DDR genes.

Somatic alterations in DDR genes were observed in almost three-quarters of the patients. None of the individual DDR gene alterations was common to all tumors but TP53 mutations were present in almost half of the tumors. The cumulative weights of DDRd signatures between subjects with and without *TP53* mutations were not significantly different. The mutational loads of tumors, which harbored a mutant *TP53* were not different from those with a wild type*TP53*.

Microsatellite instability was detected in nearly a quarter of subjects

MSI is a form of hypermutation caused by defective DNA mismatch repair, characterized by extensive alterations in the length of simple repeats[8,12]. High rates of SNVs are also observed in tumors with MSI[14]. MSI can arise due to germline or somatic mutations in MMR genes or due to epigenetic inactivation of MMR genes[10,38]. Because DNA mismatch repair (MMR) deficiency signatures were commonly observed in the cohort, and because several MMR genes were found to be frequently altered, we also tested for presence of MSI in individual tumors. An in-silico classifier method (MSIseq) was used to differentiate between tumors highly likely to have MSI (MSI-H) and those that are less likely to harbor MSI (non-MSI-H). In our analysis, 24% of the subjects were identified to be MSI-H. It was identified that the mean mutational load of MSI-H tumors was slightly higher than the mean of non-MSI-H tumors. Although no statistically significant association between the different genotypes and MSI status could be shown, there was an increasing percentage of MSI-H for subjects who had GG, AG and AA rs1800734 genotypes. The mean cumulative weight of DDRd signatures in MSI-H and non-MSI-H were not significantly different.

6. Conclusions

In this study, we identified that DDRd is a common finding in adult diffuse gliomas, represented by DDRd-related mutational signatures as well as germline/somatic genetic alterations in DDR genes. Based on the current findings as well as current literature we conclude that “DNA Damage Repair deficiency” may act as a fundamental mechanism in gliomagenesis rather than being a random event.

7. Acknowledgements

We are grateful to our patients who endure with courage.

8. Disclosure

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

9. References

1. Adel Fahmideh M, Schwartzbaum J, Frumento P, Feychting M: Association between DNA repair gene polymorphisms and risk of glioma: a systematic review and meta-analysis. **Neuro Oncol 16:**807-814, 2014
2. Akyerli CB, Yuksel S, Can O, Erson-Omay EZ, Oktay Y, Cosgun E, et al: Use of telomerase promoter mutations to mark specific molecular subsets with reciprocal clinical behavior in IDH mutant and IDH wild-type diffuse gliomas. **J Neurosurg:**1-13, 2017
3. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al: Clock-like mutational processes in human somatic cells. **Nat Genet 47:**1402-1407, 2015
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al: Signatures of mutational processes in human cancer. **Nature 500:**415-421, 2013
5. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR: Deciphering signatures of mutational processes operative in human cancer. **Cell Rep 3:**246-259, 2013
6. Alexandrov LB, Stratton MR: Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. **Curr Opin Genet Dev 24:**52-60, 2014
7. Alsop K, Fereday S, Meldrum C, deFazio A, Emmanuel C, George J, et al: BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. **J Clin Oncol 30:**2654-2663, 2012
8. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, et al: A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. **Cancer Res 58:**5248-5257, 1998
9. Chen H, Taylor NP, Sotamaa KM, Mutch DG, Powell MA, Schmidt AP, et al: Evidence for heritable predisposition to epigenetic silencing of MLH1. **Int J Cancer 120:**1684-1688, 2007
10. Cunningham JM, Christensen ER, Tester DJ, Kim CY, Roche PC, Burgart LJ, et al: Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. **Cancer Res 58:**3455-3460, 1998
11. Eckel-Passow JE, Lachance DH, Molinaro AM, Walsh KM, Decker PA, Sicotte H, et al: Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. **N Engl J Med 372:**2499-2508, 2015
12. Eshleman JR, Markowitz SD: Mismatch repair defects in human carcinogenesis. **Hum Mol Genet 5 Spec No:**1489-1494, 1996
13. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al: COSMIC: somatic cancer genetics at high-resolution. **Nucleic Acids Research 45:**D777–D783, 2017
14. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al: Patterns of somatic mutation in human cancer genomes. **Nature 446:**153-158, 2007
15. Helleday T, Eshtad S, Nik-Zainal S: Mechanisms underlying mutational signatures in human cancers. **Nat Rev Genet 15:**585-598, 2014
16. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al: Pathogenic Germline Variants in 10,389 Adult Cancers. **Cell 173:**355-370 e314, 2018

17. Huang MN, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG: MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. **Scientific Reports 5**, 2015
18. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al: Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. **Cell Rep 23:**239-254.e236, 2018
19. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al: Analysis of protein-coding genetic variation in 60,706 humans. **Nature 536:**285, 2016
20. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **arXiv**, 2013
21. Liu Y, Scheurer ME, El-Zein R, Cao Y, Do KA, Gilbert M, et al: Association and interactions between DNA repair gene polymorphisms and adult glioma. **Cancer Epidemiol Biomarkers Prev 18:**204-214, 2009
22. Loeb LA, Springgate CF, Battula N: Errors in DNA replication as a basis of malignant changes. **Cancer Res 34:**2311-2321, 1974
23. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al: The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. **Acta Neuropathol 131:**803-820, 2016
24. Melin BS, Barnholtz-Sloan JS, Wrensch MR, Johansen C, Il'yasova D, Kinnersley B, et al: Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. **Nat Genet 49:**789-794, 2017
25. Nguyen SA, Stechishin OD, Luchman HA, Lun XQ, Senger DL, Robbins SM, et al: Novel MSH6 mutations in treatment-naive glioblastoma and anaplastic oligodendroglioma contribute to temozolomide resistance independently of MGMT promoter methylation. **Clin Cancer Res 20:**4894-4903, 2014
26. Nowell PC: The clonal evolution of tumor cell populations. **Science 194:**23-28, 1976
27. Ohgaki H, Kim YH, Steinbach JP: Nervous system tumors associated with familial tumor syndromes. **Curr Opin Neurol 23:**583-591, 2010
28. Oktay Y, Ulgen E, Can O, Akyerli CB, Yuksel S, Erdemgil Y, et al: IDH-mutant glioma specific association of rs55705857 located at 8q24.21 involves MYC deregulation. **Sci Rep 6:**27569, 2016
29. Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, et al: The epidemiology of glioma in adults: a "state of the science" review. **Neuro Oncol 16:**896-913, 2014
30. Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, et al: Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. **New England Journal of Medicine 375:**443-453, 2016
31. Rahman N: Realizing the promise of cancer predisposition genes. **Nature 505:**302-308, 2014
32. Rodriguez-Hernandez I, Perdomo S, Santos-Briz A, Garcia JL, Gomez-Moreta JA, Cruz JJ, et al: Analysis of DNA repair gene polymorphisms in glioblastoma. **Gene 536:**79-83, 2014
33. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C: DeconstructSigs: delineating mutational processes in single tumors

distinguishes DNA repair deficiencies and patterns of carcinoma evolution. **Genome Biol 17:**31, 2016
34. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al: Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. **Bioinformatics 27:**2648-2654, 2011
35. Savio AJ, Bapat B: Modulation of transcription factor binding and epigenetic regulation of the MLH1 CpG island and shore by polymorphism rs1800734 in colorectal cancer. **Epigenetics 12:**441-448, 2017
36. Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith O, et al: GenVisR: Genomic Visualizations in R. **Bioinformatics 32:**3012-3014, 2016
37. Team RDC: R: A Language and Environment for Statistical Computing, in**.** Vienna, Austria: R Foundation for Statistical Computing, 2017
38. Veigl ML, Kasturi L, Olechnowicz J, Ma A, Lutterbaugh JD, Periyasamy S, et al: Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. **Proceedings of the National Academy of Sciences 95:**8698-8702, 1998
39. Venables WN, B.D. R: **Modern Applied Statistics with S, ed 4th.** New York: Springer, 2002

10. Figure Legends

Figure 1. (A) Stacked bar chart of the number of somatic Transition/Transversion SNVs. Each different bar color indicates a different substitution subtype (indicated in the upper right-hand legend). The subjects are grouped into their corresponding molecular subsets, indicated by the tiles of different colors at the bottom of the plot (colors for each subset are given in the lower right-hand legend). (B) Heatmap of weights of the mutational signatures per tumor. As the weight increases, the corresponding tile color changes from white to red. The subjects are grouped into their corresponding molecular subsets (the colors indicate the same molecular subsets as in Figure 1A). Annotations of molecular subsets, WHO grades, pathological diagnoses ages at diagnosis and mutational loads per subject is shown on top of the heatmap. (legends are displayed in the right-hand side) * indicates that the signature is associated with defective DNA mismatch repair. ** indicates that the signature is associated with failure of DNA double-strand break-repair by homologous recombination.

Figure 2. (A) Percentage of subjects that have the given mutational signatures in the current cohort (Current) and TCGA cohort (TCGA). As the percentage increases, the color of the corresponding tile changes from yellow to red. The percentage values are displayed in decreasing order according to the TCGA cohort percentages. (B) Scatter plots of weight of signature 1 vs. age at diagnosis for the current cohort and TCGA cohort separately. The red lines indicate the linear regression models fitted using these data, displaying the positive correlations. (C) Scatter plots of cumulative weights of DDR deficiency signatures (signatures 3, 6, 15, 20 and 26) vs. age at diagnosis for the current cohort and TCGA cohort separately. The red lines indicate the linear regression models fitted using these data, displaying the negative correlations.

Figure 3. Plot of germline and somatic variants in DDR genes in each subject. The plot is divided into two and MSI-High and non-MSI-High, as indicated by the bottom labels. Open circles indicate germline risk polymorphisms. Full dots represent somatic non-synonymous SNV/INDELs. Red triangles indicate somatic homozygous deletions. The DDR genes are grouped by function, annotated with the left-hand labels. On top, the presence of DDR deficiency signatures per subject is shown.

Figure 4. (A) Boxplots of cumulative weight of DDR deficiency signatures by carrier status of rs1800734. GG is wildtype, GA is heterozygous, and AA is homozygous for rs1800734. * indicates statistically significant difference. (B) Boxplots of mutational load (mutations/Mb) by carrier status of rs1800734. GG is wildtype, GA is heterozygous, and AA is homozygous for rs1800734. * indicates statistically significant difference.

Supplementary Figure S1. Scatter plots of signature weights obtained from YAPSA (x-axis) and DeconstructSigs (y-axis), demonstrating positive correlations of the

weights between the two tools. Each point corresponds to an individual. On the top left of each scatter plot, the Spearman correlation coefficient and the associated p-value is presented. “NA” indicates that the correlation test could not performed because the signature was not identified in any of the subjects by at least one tool. The red lines indicate the linear regression models fitted using these data, displaying the positive correlations.