# ULTRALOW-FREQUENCY MUTATIONAL VARIATIONS REFLECT SPATIAL HETEROGENEITY OF GLIOBLASTOMAS

By
Zhipeng Dong

A thesis submitted to the Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
August 2021

# Abstract

Glioblastoma multiforme (GBM) is the most common and aggressive type of brain tumor in adults, hallmarked by inter and intratumoral heterogeneity. Current treatment incorporates several biomarkers at the genomic and epigenomic levels. However, the efficacy of prognosis based on single biopsy was often undermined by the heterogeneous nature of GBM. Studies have highlighted the need for multi-sector biopsies to minimize the effect of intratumoral heterogeneity in clinical decision-making. In this project, we investigated mutations of geographically different regions of 20 primary glioblastoma specimens from seven patients for the selected regions of 13 genes using a novel targeted deep sequencing technology, Duplex Sequencing (DS). We have focused on subclonal (ultralow- and low-frequency) mutations that are not detectable by conventional next generation sequencing (NGS) methodologies but are accurately detectable by DS. Our findings indicate the heterogeneity of known GBM biomarkers, *TERT* promoter C228T mutation and *IDH1* nonsynonymous mutations (R132H, R132G) in codon 132, in different regions of the GBM. Intratumoral heterogeneity of subclonal mutations are mainly found in *EGFR, TERT, MSH6, PIK3CA*, and *PIK3R1* genes in most patients (six out of seven). Our results reveal that the similarity in mutation sequence context was not significantly higher in closely located specimens compared with distally located specimens. These findings could provide information on clinically relevant mutations that are unique to different regions of the tumors, and help guide future studies that seek to develop multi-sector biopsies for GBM prognosis.

Primary Reader and Advisor: Eun Hyun Ahn, MS, PhD

# Dedication

This thesis work is dedicated to my parents, Chunguang Dong and Zongting Li, for your love and support during the challenges of graduate school and life. You are my role models, inspiring me to strive for the things that I aspire to achieve.

# Acknowledgements

First and foremost, I am extremely grateful to my supervisor, Dr. Eun Hyun Ahn, for the invaluable advice, continuous support, and patience during my study. I would also like to thank Dr. Deok-Ho Kim and Dr. Jung Soo Suk for advising me on my committee. Furthermore, I want to express my gratitude to my labmates, Susan Kim, Sujin Kwon, Seung Hyuk Lee, Howard Nebeck, and Alex Wu, for their generous help during my entrance to the lab. Finally, I'd also like to thank Shunyao Lei, Mai Liu, and Joshua Ni for their collaboration.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| 1000g | The 1000 Genomes Project Consortium 2015 |
| ANNOVAR | Annotate Variation |
| CADD | Combined Annotation Dependent Depletion |
| CADD | Combined Annotation Dependent Depletion |
| CHASMplus | Cancer specific High-throughput Annotation of Somatic Mutations plus |
| CNV | Copy number variation |
| COSMIC | Catalogue of Somatic Mutations in Cancer |
| DCS | Duplex consensus sequences |
| DGIdb | The Drug Gene Interaction Database |
| DS | Duplex Sequencing |
| ESP6500 | NHLBI-ESP 6500 exomes project |
| ExAC | Exome Aggregation Consortium |
| GBM | Glioblastoma multiforme |
| *IDH1* | Isocitrate dehydrogenase 1 |
| MCS | Mutation sequence context spectra |
| ncRNA | Non-coding RNA |
| NGS | Next generation sequencing |
| OpenCRAVAT | Open Custom Ranked Analysis of Variants Toolkit |
| RT | Radiotherapy |
| SBS | Single base substitution |
| SNV | Single nucleotide variations |
| *TERT* | Telomerase reverse transcriptase |
| *TERTp* | *TERT* promoter |
| TMZ | Temozolomide |
| UTR3 | Three prime untranslated region |
| UTR5 | Five prime untranslated region |

# Introduction

Glioblastoma multiforme (GBM) is the most common and deadly form of brain tumors, with a median survival time of 14.6 months. The standard treatment of GBM involves maximal surgical resection followed by radiotherapy (RT) and chemotherapy temozolomide (TMZ) (1–3). The treatment of GBM is complicated by the intratumoral molecular and cellular heterogeneity, resulting in different cell populations reacting to therapy differently. Previous studies have demonstrated the heterogeneity of *EGFR* and *PDGFRA* amplification (4) and the mutational heterogeneity within different regions of primary GBM (5,6). However, previous studies have focused on the mutational heterogeneity at the clonal (high-frequency mutations) level rather than at the subclonal (low- and ultralow-frequency mutations) level. Subclonal mutations might account for the genetic heterogeneity of tumors that contributes to therapy resistance and tumor recurrence. However, subclonal mutations are not accurately detectable using conventional next generation sequencing (NGS) methods due to the high error rates ($10^{-2}$ to $10^{-3}$). Mutations with allele frequency below NGS's error frequency are confounded by the false positive variants. Duplex Sequencing (DS), a novel deep sequencing technology, improves the sequencing accuracy by sequencing both strands of DNA. DS only counts the mutations if the mutations are detected as complementary substitutions in both strands of the same DNA molecules (7–10). While the conventional sequencing technologies only investigate a single DNA strand, DS produce >10,000 fold more accurate results compared with other currently available high-throughput sequencing methods. The lowest error rate ($<5x10^{-8}$) among high-throughput DNA sequencing methodologies by DS ensures accurate detection of the subclonal mutations and clonal mutations.

This study presents evidence of intratumoral heterogeneity based on DS data for 20 primary GBM specimens from seven patients. Each primary GBM was sectioned into different geographical regions, and two to four specimens were obtained from each patient.

# Materials and methods

**Glioblastoma (GBM) specimens.** Twenty GBM biopsy specimens from seven patients were obtained in collaboration with Nameeta Shah, PhD,  Ralph Puchalski, PhD, and Charles S Cobbs, MD at the Ben and Catherine Ivy Center for Advanced Brain Tumor Treatment, Swedish Neuroscience Institute (Seattle, WA, USA). All specimens were primary GBMs which were not treated with GBM treatments. The study was approved by Western IRB (#20062252, #20091429, #20091563). The research was conducted under the guidance of the ethical principles as described in the report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research entitled "Ethical Principles and Guidelines for the Protection of Human Subjects of Research (Belmont Report)". All Patients' written consents were obtained for their respective study from Swedish Neuroscience Institute (Seattle, WA) in accordance with institutional guidelines.

Each patient's tumor was sectioned into different regions, and two to four sections, each approximately 1 $cm^3$, were obtained from each patient (Fig. S1). The specimen pairs were classified into adjacent or distally located based on the distance between the centers of specimens. Specimen pair was classified as closely located if the distance between the centers of specimens was less than 2 cm, and classified as distally located otherwise. The distance was measured using Fiji software (11).

**DNA extraction, adapter Synthesis, and DNA library preparation**. DNA was isolated using DNeasy kit (Qiagen Inc., Germantown, MD), based on a manufacturer's protocol with modifications. The DNA library preparation for Duplex Sequencing (DS) was carried out by the previously described protocol (10,12) with modifications. DNA library preparation for DS was designed to target selected regions of 13 genes, including *CDKN2A, CHEK2, EGFR, H3F3A, IDH1, MGMT, MSH6, PIK3CA, PIK3R1, PTEN, RB1, TERT*, and *TP53*, covering 41,515 bases (41483 bases after excluding the overlapping regions). DNA library for DS was sequenced for paired-end sequencing on an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA). These experiments and reviewing and planning of procedures were carried out by Eun Hyun Ahn, PhD, Kaitlyn J. Loubet-Senear, Kate Bayliss, MD, PhD, Joon Yup (Jason) Kim, and Seung Hyuk Lee.

**Duplex sequencing data processing and analysis.** Duplex consensus sequencing (DCS) data processing was carried out by Howard Nebeck, MS as described previously (12,13) with some modifications. Consensus sequencing-making script was used to merge sequencing data for both strands of DNA. GATK 3.7 was used to align the merged DCS files to the human reference genome GRCh37 (hg19). All reads with mapping quality scores below 40 or with 5% or more unreadable bases were filtered out. The leading seven bases at the 5' end and three bases at the 3' end of each read were removed to avoid potential artificial variants commonly present at the ends of each read. Variant-calling from BAM files was performed using SAMtools software with a base quality score of 13. Genome positions with duplex consensus sequencing (DCS) depth below 100 were filtered out. When comparing among specimens from the same patient, genome positions with DCS depth less than 100 in any of the specimens from the same patient were filtered out.

Subclonal mutation frequency was calculated by dividing the total number of DCS variant reads by the total number of DCS sequenced reads across the sequenced genome regions. In all other analyses (fraction of mutation types, mutation context spectra, and comparison of mutation positions), the number of unique mutations in each sample was used. This number was calculated by counting distinct variants at each genome position only once, regardless of the number of occurrences of that variant at that position. If different mutations (i.e. difference nucleotide change) occurred at the same genome position, the mutations were considered as different unique mutations. Mutations were classified into two classes based on the clonality, which was calculated by dividing the number of variants at a genome position by the total number of DCS reads at that position. The mutations were classified into subclonal (less than or equal to 10% clonality), and clonal (greater than 10% clonality) mutations.

**Mutation annotation and blood variant filtering.** Point mutations were annotated using the Annotate Variation (ANNOVAR) software version 2017 June 01 (annovar.openbioinformatics.org). Categories of annotations include protein codon change, point mutation type, presence of mutations in population databases including the 1000 Genome Project (14), NHLBI-ESP 6500 exomes (15), Exome Aggregation Consortium (16). Mutations present in these blood databases were filtered out. In specific, mutations that were present in 1000g or ESP6500, or mutations with greater than 5 occurrences in the ExAC database were excluded. Sujin Kwon and Seung Hyuk Lee contributed to the mutation annotation and filtering out mutations reported in the population blood databases.

Based on the functional annotation from ANNOVAR, the mutations were divided into 16 different categories (UTR5, UTR3, Upstream, Unknown, Synonymous-splicing, Synonymous, Splicing, Nonsynonymous-splicing, Nonsynonymous, Nonsense-splicing, Nonsense, ncRNA-splicing, ncRNA, Intronic, Intergenic, and Downstream). The number and fraction (%) of each mutational annotation category within each of the specimens were determined.

Using OpenCRAVAT (17), the mutations found privately in only one specimen within the same patient were annotated with predicted scores from CHASMplus-GBM (18), CADD Exome (19), and predicted drugs interaction scores from DGIdb (20). Mutations with pathogenicity scores with *p* value less than 0.01 by CHASMplus-GBM, or mutations with CADD-Exome Phred scores greater than 10, were kept for analysis. Mutations with drug interaction scores greater than four were kept for analysis.

**Cosine similarity and SigProfiler decomposition.** For mutation sequence context spectra (MCS) analysis, the single base substitutions (SBSs) were divided into 96 classes based on the six base changes (C>A, C>G, C>T, T>A, T>C, T>G) and two bases surrounding the mutated base, as described previously (13). The cosine similarity measures the similarity between two MCS datasets with a score ranging from 0 (completely dissimilar) to 1 (identical). Cosine similarities between the MCS of different specimens were calculated.

SigProfiler, a nonnegative matrix factorization method (21), was used to determine the contributions of each COSMIC signature to each cancer sample. Using SigProfiler, the MCS of our subclonal mutations was reconstructed from combinations of COSMIC v3.1 SBS mutational signatures, and the reconstructed MCS with the highest cosine similarity score to the original MCS was outputted.

**Statistical Analysis.** Differences in mutation frequencies, in the fraction (%) of mutation types, and in the telomerase reverse transcriptase (*TERT*) promoter mutation clonalities (%) between two GBM specimens were analyzed by performing 2-sample equality of proportions with continuity correction (also called Chi-Square test) using an R program (version 3.4.4). The Mann-Whitney U-test (Wilcoxon Rank-Sum test) was applied to compare the cosine similarity scores between the two groups. Differences between the two groups were considered significant if the *p* value was less than 0.05.

# Results

**The majority of mutations exist at the subclonal (low-frequency) level**. The number of unique mutations from each tumor sample was determined using Duplex Sequencing by counting the same nucleotide change at the same genome position only once. The majority of the mutations detected in the GBM specimens using DS were subclonal mutations. The average fractions of subclonal mutations out of the total mutations at all clonalities (0-100%) in all specimens are approximately 76.3%. Within 13 out of 20 specimens, >70% of the total mutations are subclonal mutations. Within 5 out of 20 specimens, >90% of the total mutations are subclonal mutations (Table. S1).

**Intratumoral mutational heterogeneity is mainly observed at the subclonal level.** Within each patient, the number of mutations exclusive to each specimen, as well as the number of mutations shared between specimens, was determined by comparing the mutation positions (Fig. S3). Most of the mutations shared among all specimens from the same patient were present as clonal (high-frequency, >10% to 100% clonality) mutations. The number of mutations exclusive to each specimen was mostly unaffected and the proportion of subclonal unique mutations exclusive to one specimen ranged between 86.8% and 96.9%, with a median of 93.8% (Fig. 3). Thus, most of the clonal mutations were shared between specimens within the same patient, and the subclonal mutations made up the majority of exclusive mutations. These results indicate that mutational heterogeneity is observed at the subclonal level in geographically different specimens within the same patient.

**The majority of mutational heterogeneity is found in missense and synonymous mutations.** Point mutations (single base substitutions) we identified were annotated into eight different categories using ANNOVAR. Missense and synonymous mutations made up the majority of mutations private to each specimen within the same patient. Combined together, missense and synonymous mutations made up 65.5% to 77.6% (average 72.1%) of the exclusive mutations in each patient (Fig. 4). Differences in the fractions of the mutation types were observed for most patients. For example, the section A from patient W4 had a higher ratio of missense mutations and a lower ratio of synonymous mutations than the other two specimens from the same patient (specimen B, F). The section A from patient W50 had a lower ratio of missense mutations and a higher ratio of synonymous mutations than the other specimens from the same patient (specimen B, I). For patient W33, sample F had a higher ratio of synonymous mutation than sample L. These results suggest that it is worthwhile to study not only the missense mutations but also the synonymous mutations that do not directly alter the amino acid sequence.

**Most of the private mutations present in only one specimen within the same patient are found in *EGFR, TERT, MSH6*.** The numbers of subclonal mutations exclusive to each specimen within the same patient for each of the selected 13 genes were determined using DS. The *EGFR* gene harbored the most exclusive mutations, followed by *TERT, MSH6, PIK3CA*, and *PIK3R1* (Fig. 5). While other genes had up to 58 exclusive mutations in one specimen, the *EGFR* gene had up to 502 exclusive mutations in one specimen. Section C from patient W3, section H from patient W22, section E from patient W48, and section C from patient W53 had much higher numbers of exclusive mutations compared to other specimens from the same patient. Looking further into these exclusive mutations gave us insight into the heterogeneity within the patient.

**Heterogeneity of subclonal mutations in TERT promoter region and IDH1 gene were observed.** *TERT* promoter (*TERTp*) mutations C>T at Chromosome 5 position 1295228 (C228T) and at position 1295250 (C250T) were often reported in GBM patients and were related to shorter survival periods (22). In our GBM specimens, *TERTp* mutation C228T was found in every specimen in six out of seven patients. However, C250T mutation was not found in any specimen. For patients W3, W4, and W22, the mutational clonality (%) of C228T was significantly different between any pairs of two specimens within the same patient (Table. 3, $p<10^{-5}$ by Chi-squared test). The significant difference in *TERTp* mutation clonality suggests that the generation of *TERTp* mutation during early tumorigenesis was distributed unevenly in different subclones of GBM cells.

*IDH1* mutation in codon 132 is often present in GBM patients and is used to stratify GBM patient diagnosis (23). In our study, the G>A mutation at Chromosome 2 position 209113112, R132H, was found in only section C of the patient W3, while R132G was found in all three specimens of the patient W50 (Table. 1). Our results are aligned with a previous study which reported that *TERTp* mutations (C228T, C250T) and *IDH1* mutations are inversely correlated (22).

**Heterogeneity of subclonal mutations are regional-specific.** After filtering out the population blood mutations present in 1000 genome, ESP6500, and ExAC databases, 165 subclonal mutations remained as exclusive mutations to each specimen within the same patient in all 20 specimens from seven patients. We queried these subclonal exclusive (private) mutations against previously reported driver mutations in GBM (24–33). The five published mutations were found in more than one specimen within the same patient, while 15 published mutations were exclusive to only one specimen within the same patient. Nine out of the 15 subclonal exclusive mutations were reported in GBM patients by the ClinVar database, and eight out of the nine subclonal mutations were reported to be pathogenic or likely-pathogenic (34). Out of the six subclonal mutations not previously reported in GBM by ClinVar, four had unknown effects, and two had pathogenic or likely-pathogenic effects in other diseases, suggesting that they were likely to contribute to GBM tumorigenesis.

Using OpenCRAVAT, the subclonal exclusive mutations were annotated with predicted pathogenicity scores (CHASMplus-GBM, CADD) and drug-interaction scores (DGIdb).

CHASMplus score ranges between 0 and 1, with higher scores reflecting a greater likelihood for a mutation to be driver. Combined Annotation Dependent Depletion (CADD) quantifies the deleteriousness of variants by integrating multiple databases. The Phred-scaled scores reflect the ranking of the CADD scores among ~8.6 billion single nucleotide variations (SNVs) from the hg19 reference, with the top 10% equivalent to CADD-Phred of 10, top 1% equivalent to CADD-Phred of 20. We found 22 subclonal mutations with significant CHASMplus-GBM scores ($p < 0.01$), and 69 mutations with CADD Phred scores greater than 10. The result shows that there is a high number of potentially deleterious mutations that are region-specific in glioblastomas. The exclusive subclonal mutations might be potential druggable targets. Using DGIdb, the drug interaction database, we found that 31 mutations could be targeted by Tertomotide (drug interaction score 9.88), five mutations druggable by Durvalumab (interaction score 9.47), five mutations druggable by O6-[3-(Aminomethyl)Benzyl]Guanine (interaction score 6.82), and 8 mutations druggable by Milciclib (PHA-848125AC, interaction score 5.16).

**Intratumor mutation sequence context (MCS) heterogeneity is manifested as low scores of cosine similarity.** The unique mutations determined using DS were divided into 96 classes to generate mutation sequence context spectra (MCS) for each of the 20 specimens (Fig. S5). Cosine similarity scores between the MCS of each pair of specimens were calculated. The score measures the similarity between mutational profiles of samples, and ranges from 0 (not similar) to 1 (identical). For a specific patient, we here defined the comparison between any two specimens from the same patient as ‘within a patient’, while the comparison between specimens from a patient and specimens from other patients as ‘across patients’. The ‘within a patient’ comparison results demonstrate that tumor specimens located closely to each other do not show significantly higher similarities in their MCS (Fig. 6). For example, within patient W22, the highest similarity score was found between sections F and H. However, these two specimens were not located closely within the tumor. The ‘within a patient’ cosine similarity scores indicate only moderate levels of similarity between specimens within the same patient. Furthermore, no statistically significant difference is observed between ‘within a patient’ cosine similarity scores and ‘across patients’ similarity scores (*p*>0.05, Mann-Whitney U test) for six out of seven patients (Fig. 6B). The ‘within a patient’ comparisons have significantly higher cosine similarity scores than ‘across patient’ comparisons for patient W3 (*p*=0.044, Mann-Whitney U test), and when all patients’ MCS are pooled together (Fig. 6C). However, this difference seen in the pooled data of all patients is not significant when the patient W3 is excluded (Fig. 6D). Taken together, the results indicate that GBM specimens do not necessarily carry significantly higher mutation sequence context similarity with other specimens from the same patient than with specimens from different patients.

We also compared our subclonal mutation sequence context with the Catalogue of Somatic Mutations in Cancer (COSMIC) SBS signatures. SigProfiler (21) was used to decompose the mutational context spectra data into the COSMIC SBS signatures (Fig. 7). The results support our findings from cosine similarity analysis (Fig. 6). Sections F and H from patient W22 have a high cosine similarity score, and the decompositions of these two specimens’ MCS have similar types and proportions of the SBS signatures. For pairs of specimens that have low cosine similarity scores, the decomposed signatures depict differences in the SBS signatures with proposed etiology. For example, signature 10b associated with polymerase epsilon exonuclease domain mutations, is related to only section A from patient W4, sections F and H from patient W22, section A from patient W48, and section L from patient W33. Signature 14 associated with defect DNA mismatch repair (MMR), was only related to section F from patient W33. Signature 18 associated with possible reactive oxygen species damage, was only related to section L from patient W33, and section A from patient W53. Signature 29 associated with tobacco chewing, is only related to section F from patient W4, only sections F and H from patient W22, section E from patient W48, and section C from patient W53.

# Discussion

In this study, we demonstrate that the majority of intratumor mutational heterogeneity is observed at subclonal levels. A previous whole exome sequencing study by Mahlokozera et al. (4) reported that 46% of the mutations were exclusive to only one section from the same primary GBM. In comparison, by using Duplex Sequencing, we here found that 86.8% to 96.9% of the subclonal mutations were are in only one specimen within the same patient. Among these subclonal (ultralow and low-frequency) mutations, we found more evidence of known GBM genetic markers as well as novel mutations that could potentially serve as druggable targets. Using DS with higher sensitivity and accuracy, we have found high numbers of mutations exclusive to only one specimen within the same patient. The exclusive mutations were mainly observed in the EGFR gene, with 6 specimens having more than 60 exclusive subclonal mutations (W22-A/F/H; W48-E; W53-A/C). A previous study showed that *EGFR* was among the most frequently mutated genes, resulting in the expression of diverse transcripts (35). The similar trend is observed in our study. Our results further confirm that *EGFR* is the most frequently mutated gene among the 13 genes we examined. Other frequently mutated genes in our results include *MSH6, PIK3CA, PIK3R1*, and *TERT*.

Certain specimens are more highly mutated than other specimens from the same patient. For example, three specimens (W3-C, W22-H, and W48-E) carry higher number of subclonal unique mutations compared with the other specimens (W3-B/E, W22-A/E/F, and W48-A/G) from the same patients. The three specimens have greater numbers of mutations by > 4-fold compared with the specimen with the lowest number of mutations from the same patient. This difference is not confounded by differences of sequencing depth because the ratio of average DCS sequencing depth to the lowest depth within the same patient was not proportionally higher for the three specimens (W3-C, W22-H, and W48-E) (Fig S4). This result suggests that each of the three patients exhibits a putative hypermutation phenotype in the 13 genes in one region of the GBM. Our finding is consistent with an observation by Mahlokozera et al. (4), which reported that the majority of mutational load in hypermutated GBM was region-specific.

We demonstrated the heterogeneity of well-known mutations of the promoter regions of *TERT* and isocitrate dehydrogenase 1 (*IDH1*) genes in glioblastoma. *IDH1* mutations were reported in ~90% of GBM and known as a marker of better prognosis (36). In our study, we report the regional heterogeneity of *IDH1* mutation: *IDH1* mutation R132H is found in only section C from patient W3 while a less common allelic change, R132G, is found in all three sections from patient W50. *TERT* promoter mutations C228T and C250T were reported to be prognostic biomarkers for poorer survival (22). We have found the C228T mutation with varying degrees of clonalities in different specimens from the same patient (Table. 3). This finding suggests that the regional heterogeneity of the well-known mutations of *TERTp* and *IDH1* genes could confound the molecular diagnosis based on a single biopsy.

Our subclonal mutation context spectra analysis indicate that the primary GBM samples were dissimilar to samples from the same patient. In only one patient (W22), cosine similarity scores are significantly higher in 'within a patient' than 'across patients'. The MCS between specimens that are spatially located closely is not significantly more similar than specimens that are distally located within the same tumor. Our results suggest that multiple biopsies do not need to occur at distal regions of the tumor sample to detect mutational heterogeneity. Decomposing the MCS into the COSMIC SBS signatures also supports that different tumor regions could be associated with different DNA damage, repair, or replication mechanisms. Our results suggest that a single biopsy is not sufficient for a clinical decision-making based on selected oncogenetic targets.

The current study did not examine the heterogeneity of copy number variation (CNV) within the primary GBM. Common CNV present within GBM includes EGFR amplification and alpha-type platelet-derived growth factor receptor (PDGFRA) amplification, with EGFR amplification present in ~40% of primary GBM (37) and PDGFRA amplification present in about 15% of tumors (38). Future experiments with proper control samples such as blood from the matching patients will help explore the heterogeneity of CNV in primary GBM in support of the findings by previous groups (39).

# Conclusion

We present evidence of intratumor mutational heterogeneity in primary glioblastomas. Our results demonstrate that the majority of mutational heterogeneity occurs at the subclonal (ultralow and low frequency) levels. Investigating these low-frequency mutations using targeted deep sequencing methods provides insight into the clonal evolution of GBM, and elucidates the mechanisms of treatment resistance based on oncogenetic targets. Our results suggest that the subclonal mutation context differences between different regions of the same tumor do not correlate with the spatial distance between the specimens within the same patient.

# References

1. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJB, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352:987–96.
2. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJB, Janzer RC, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10:459–66.
3. Osuka S, Van Meir EG. Overcoming therapeutic resistance in glioblastoma: the way forward. Journal of Clinical Investigation. 2017;127:415–26.
4. Mahlokozera T, Vellimana AK, Li T, Mao DD, Zohny ZS, Kim DH, et al. Biological and therapeutic implications of multisector sequencing in newly diagnosed glioblastoma. Neuro Oncol. 2018;20:472–83.
5. Kumar A, Boyle EA, Tokita M, Mikheev AM, Sanger MC, Girard E, et al. Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. Genome Biol. 2014;15:530.
6. Parker NR, Hudson AL, Khong P, Parkinson JF, Dwight T, Ikin RJ, et al. Intratumoral heterogeneity identified at the epigenetic, genetic and transcriptional level in glioblastoma. Sci Rep. 2016;6:22477.
7. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc. 2014;9:2586–606.
8. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci USA. 2012;109:14508–13.
9. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nature Methods. 2015;12:423–5.
10. Ahn EH, Lee SH. Detection of Low-Frequency Mutations and Identification of Heat-Induced Artifactual Mutations Using Duplex Sequencing. IJMS. 2019;20:199.
11. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9:676–82.
12. Ahn EH, Hirohata K, Kohrn BF, Fox EJ, Chang C-C, Loeb LA. Detection of Ultra-Rare Mitochondrial Mutations in Breast Stem Cells by Duplex Sequencing. PLoS ONE. 2015;10:e0136216.
13. Kwon S, Kim S, Nebeck H, Ahn E. Immortalization of Different Breast Epithelial Cell Types Results in Distinct Mitochondrial Mutagenesis. IJMS. 2019;20:2813.
14. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
15. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493:216–20.
16. 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. 2016;536:285–91.
17. Pagel KA, Kim R, Moad K, Busby B, Zheng L, Tokheim C, et al. Integrated Informatics Analysis of Cancer-Related Variants. JCO Clin Cancer Inform. 2020;4:310–7.
18. Tokheim C, Karchin R. CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. Cell Syst. 2019;9:9-23.e8.

19. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
20. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Research. 2021;49:D1144–51.
21. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
22. Simon M, Hosen I, Gousias K, Rachakonda S, Heidenreich B, Gessi M, et al. TERT promoter mutations: a novel independent prognostic factor in primary glioblastomas. Neuro Oncol. 2015;17:45–52.
23. Sanson M, Marie Y, Paris S, Idbaih A, Laffaire J, Ducray F, et al. Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. J Clin Oncol. 2009;27:4150–4.
24. Hunter C, Smith R, Cahill DP, Stephens P, Stevens C, Teague J, et al. A Hypermutation Phenotype and Somatic *MSH6* Mutations in Recurrent Human Malignant Gliomas after Alkylator Chemotherapy. Cancer Res. 2006;66:3987–91.
25. Cahill DP, Levine KK, Betensky RA, Codd PJ, Romany CA, Reavie LB, et al. Loss of the Mismatch Repair Protein MSH6 in Human Glioblastomas Is Associated with Tumor Progression during Temozolomide Treatment. Clin Cancer Res. 2007;13:2038–45.
26. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8.
27. Yip S, Miao J, Cahill DP, Iafrate AJ, Aldape K, Nutt CL, et al. MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance. Clin Cancer Res. 2009;15:4622–9.
28. Frattini V, Trifonov V, Chan JM, Castano A, Lia M, Abate F, et al. The integrated landscape of driver genomic alterations in glioblastoma. Nature Genetics. 2013;45:1141–9.
29. Vinagre J, Almeida A, Pópulo H, Batista R, Lyra J, Pinto V, et al. Frequency of TERT promoter mutations in human cancers. Nat Commun. 2013;4:2185.
30. Liu R, Xing M. TERT promoter mutations in thyroid cancer. Endocrine-Related Cancer. 2016;23:R143–55.
31. Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DIS, Zairis S, et al. Clonal evolution of glioblastoma under therapy. Nature Genetics. 2016;48:768–76.
32. Körber V, Yang J, Barah P, Wu Y, Stichel D, Gu Z, et al. Evolutionary Trajectories of IDHWT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. Cancer Cell. 2019;35:692-704.e12.
33. Tanaka S, Batchelor TT, Iafrate AJ, Dias-Santagata D, Borger DR, Ellisen LW, et al. PIK3CA activating mutations are associated with more disseminated disease at presentation and earlier recurrence in glioblastoma. acta neuropathol commun. 2019;7:66.
34. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.
35. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The Somatic Genomic Landscape of Glioblastoma. Cell. 2013;155:462–77.
36. Cohen AL, Holmen SL, Colman H. IDH1 and IDH2 mutations in gliomas. Curr Neurol Neurosci Rep. 2013;13:345.
37. Hurtt MR, Moossy J, Donovan-Peluso M, Locker J. Amplification of epidermal growth factor receptor gene in gliomas: histopathology and prognosis. J Neuropathol Exp Neurol. 1992;51:84–90.

38. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17:98–110.

39. Sottoriva A, Spiteri I, Piccirillo SGM, Touloumis A, Collins VP, Marioni JC, et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc Natl Acad Sci U S A. 2013;110:4009–14.

# Appendices – Figures and Tables

**Table 1.** Clinical features of glioblastoma patients. The status of *TERT* promoter mutations C>T at Chromosome 5 position 1295228 (C228T) and T>C at position 1295394 (T349C) and *IDH1* G>A at Chromosome 2 position 209113112 (R132H) and G>C at position 209113112 (R132G) nonsynonymous mutations was examined using Duplex Sequencing.

| **Patient** | **Gender** | **Age at diagnosis (yr)** | **Surv days** | **RT+ TMZ[1]** | **Other Chemo Therapies** | **Tissue section** | ***TERT* promoter mutations** | ***IDH1* mutations** |
|---|---|---|---|---|---|---|---|---|
| W3 | F | 66 | 980 | Y | Y | B | C228T | - |
| | | | | | | C | C228T | R132H |
| | | | | | | E | C228T | - |
| W4 | F | 51 | 541 | Y | N | A | C228T,T349C | - |
| | | | | | | B | C228T,T349C | - |
| | | | | | | F | C228T,T349C | - |
| W22 | F | 53 | 2038 | Y | N | A | C228T,T349C | - |
| | | | | | | E | C228T,T349C | - |
| | | | | | | F | C228T,T349C | - |
| | | | | | | H | C228T,T349C | - |
| W48 | M | 52 | 455 | Y | N | A | C228T,T349C | - |
| | | | | | | E | C228T,T349C | - |
| | | | | | | G | C228T,T349C | - |
| W50 | M | 27 | 1327 | Y | N | A | - | R132G |
| | | | | | | B | - | R132G |
| | | | | | | I | - | R132G |
| W33 | M | 61 | 2464 | Y | N | F | C228T,T349C | - |
| | | | | | | L | C228T,T349C | - |
| W53 | M | 55 | Alive[a] (>2050) | Y | Y | A | C228T,T349C | - |
| | | | | | | C | C228T,T349C | - |

(1) Treatment given after the first surgery (2) Last confirmed January 27, 2020

Abbreviations used are: RT, radiotherapy; TMZ, temozolomide; -, mutation is not present in the respective category

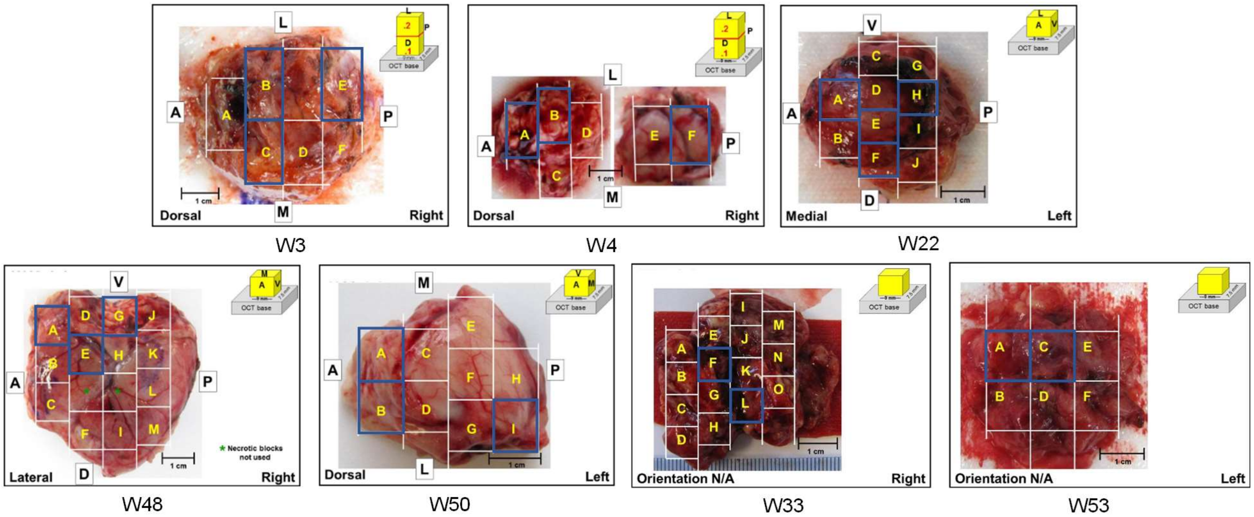

**Figure S1**. Tissue sections of primary glioblastomas from seven patients examined in the current study are highlighted with blue boxes.

**Table. S1**: The number of total and subclonal mutations and fraction (%) of subclonal mutations in glioblastomas were identified using Duplex Sequencing.

| Patient | Tissue section | No. of subclonal muts (0-10% clonality) | No. of total muts (0-100% clonality) | Fraction (%) of subclonal muts |
|---|---|---|---|---|
| W3 | B | 78 | 104 | 75.0 |
|  | C | 379 | 405 | 93.6 |
|  | E | 88 | 114 | 77.2 |
| W4 | A | 91 | 114 | 79.8 |
|  | B | 33 | 57 | 57.9 |
|  | F | 36 | 57 | 63.2 |
| W22 | A | 179 | 196 | 91.3 |
|  | E | 59 | 86 | 68.6 |
|  | F | 142 | 170 | 83.5 |
|  | H | 484 | 512 | 94.5 |
| W48 | A | 48 | 68 | 70.6 |
|  | E | 752 | 772 | 97.4 |
|  | G | 48 | 68 | 70.6 |
| W50 | A | 38 | 65 | 58.5 |
|  | B | 43 | 69 | 62.3 |
|  | I | 56 | 84 | 66.7 |
| W33 | F | 51 | 85 | 60.0 |
|  | L | 92 | 126 | 73.0 |
| W53 | A | 188 | 215 | 87.4 |
|  | C | 491 | 519 | 94.6 |
| Average |  | 168.8 | 194.3 | 76.3 |

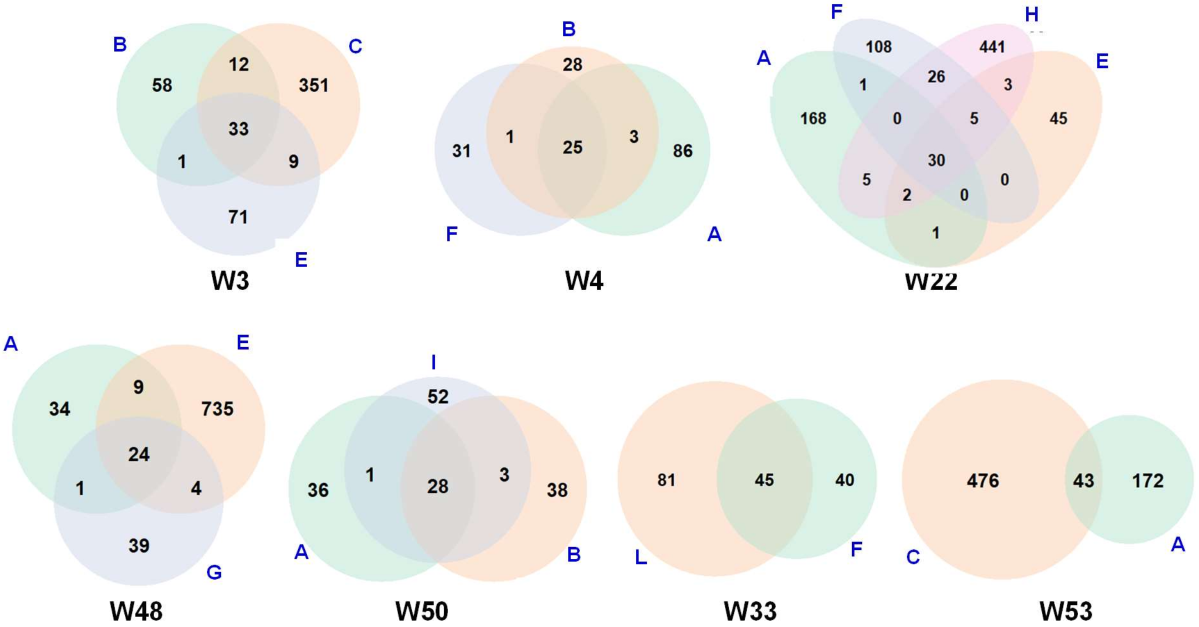

**Figure S2.** Number of unique mutations at any clonalities (0-100%) shared between specimens or exclusive to each specimen within the same patient for 20 GBM specimens from seven patients were determined using DS.

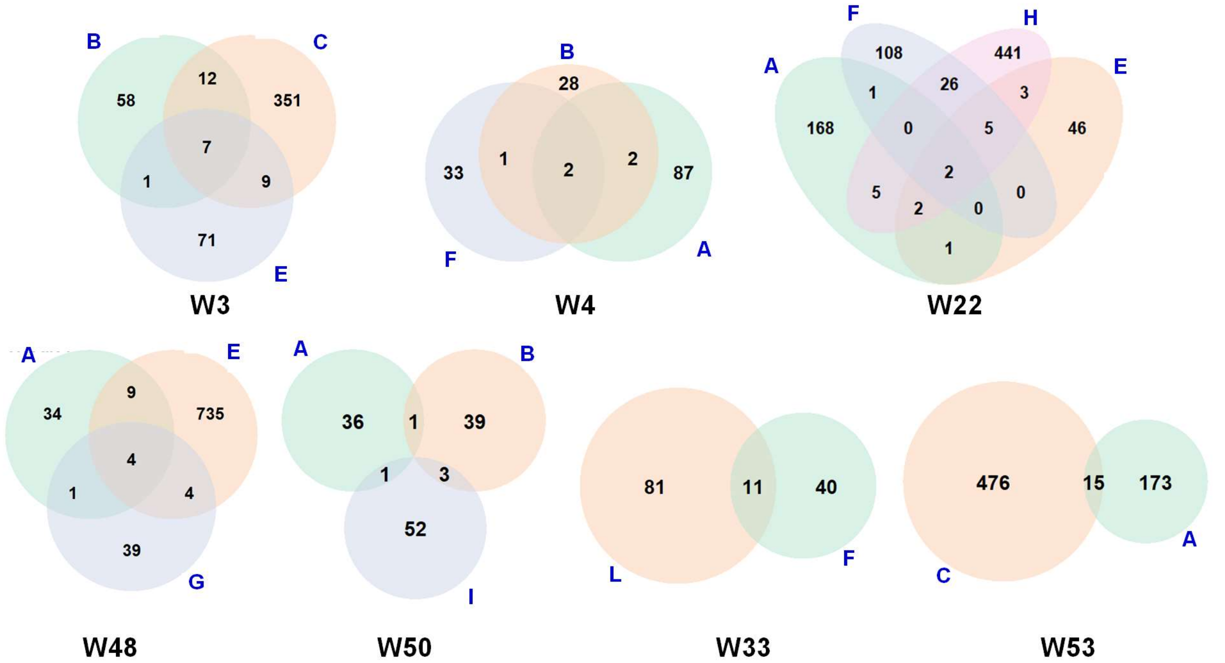

**Figure 1**. Numbers of subclonal unique mutations shared between specimens or exclusive to each specimen within the same patient for 20 GBM specimens from seven patients were determined using Duplex Sequencing.

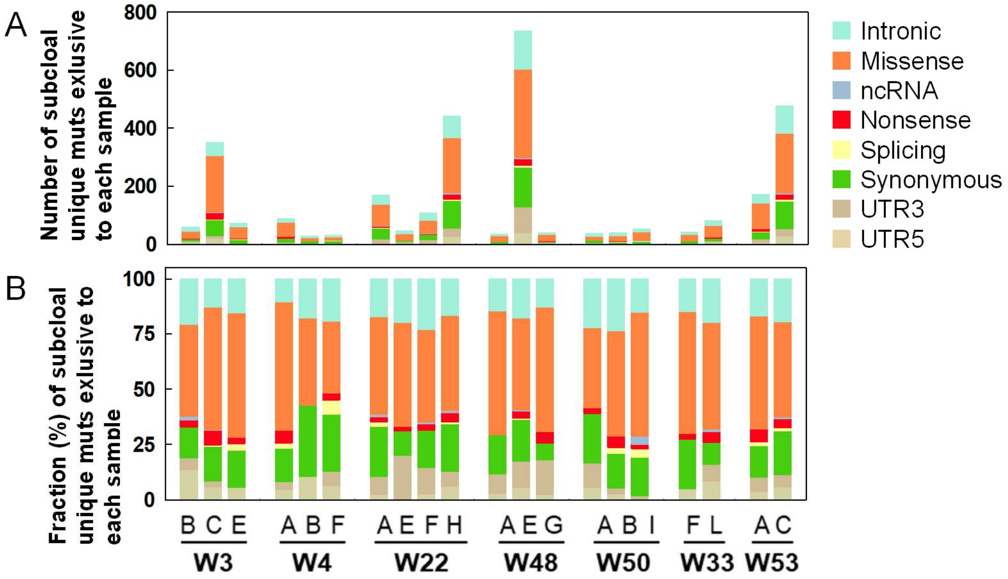

**Figure 2.** Subclonal unique mutations in different mutation annotation categories exclusive to each specimen within the same patient were determined using Duplex Sequencing. Numbers (A) and fractions (%) (B) of the subclonal mutations in each mutation annotation category for all 13 genes of each GBM specimen.

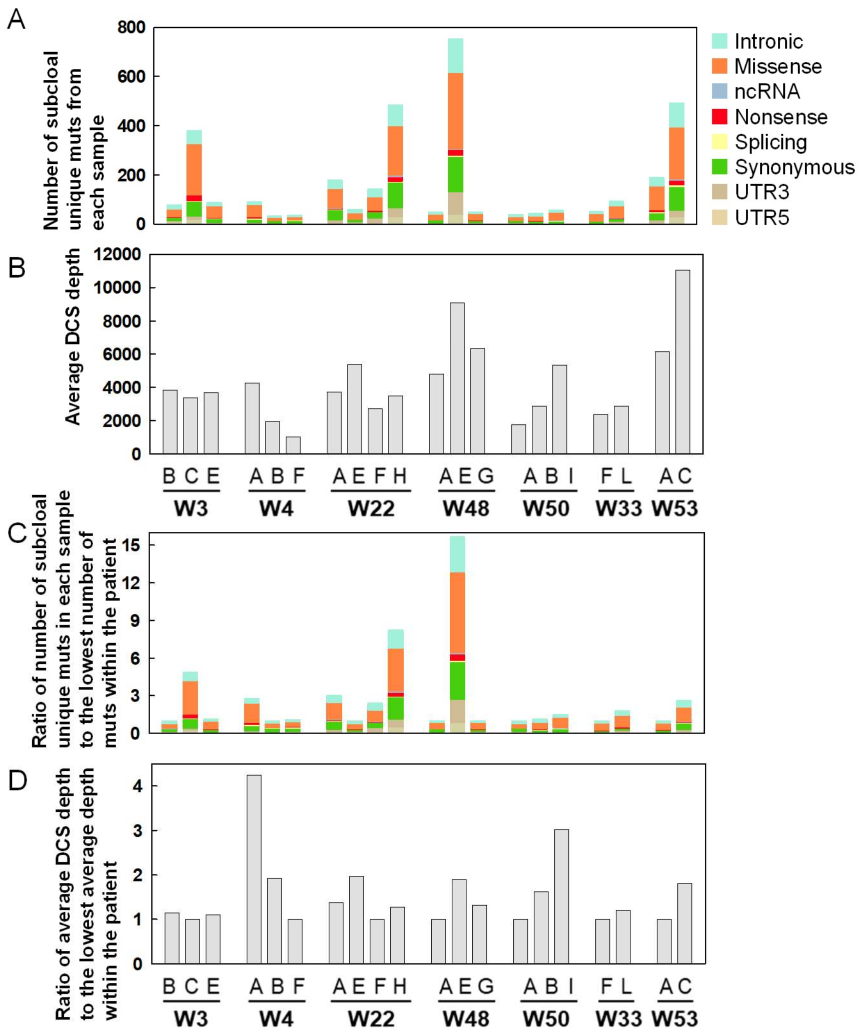

**Figure S3.** Subclonal unique mutations in different mutation annotation categories in glioblastomas were determined using Duplex Sequencing. Numbers (A), average DCS depths (B), ratios of the number of subclonal mutations in each mutation annotation category to the lowest number of total subclonal mutations of a specimen within the same patient (C) and ratios of the average DCS depths for each specimen to the lowest depth within the same patient (D).

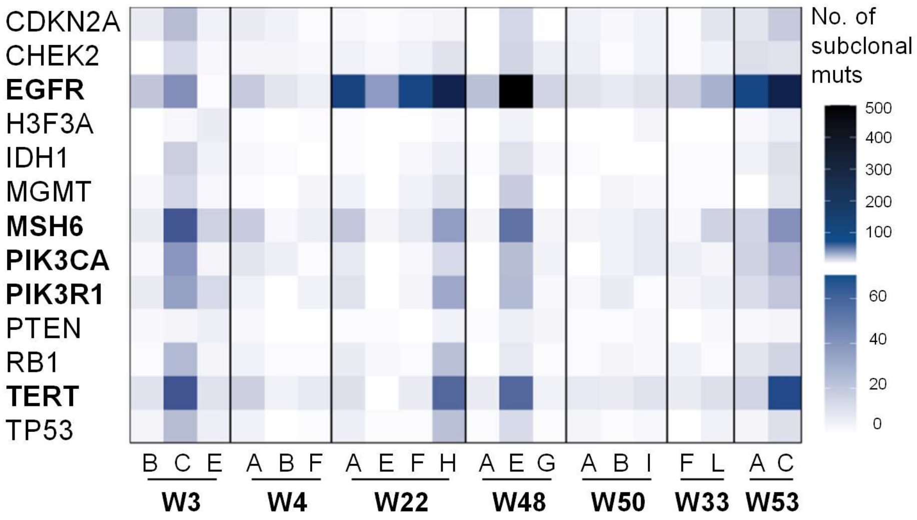

**Figure 3.** Number of subclonal unique mutations exclusive to each specimen within the same patient for each of the selected regions of 13 genes were determined using Duplex Sequencing for 20 GBM specimens from seven patients

**Table 2.** Differences of the clonalities of *TERT* promoter mutation C>T at Chromosome 5 position 1295228 (C228T) between specimens within the same patient were examined (*$p$ < 0.05, **< $5x10^{-4}$, ***< $5x10^{-10}$ by the Chi-Square test).

| **Patient** | **Tissue section comparison within a patient** | ***p* value** | **Significance** |
|---|---|---|---|
| W3 | B vs C | $6.06*10^{-10}$ | ** |
| | B vs E | $5.03*10^{-20}$ | *** |
| | C vs E | $1.14*10^{-2}$ | * |
| W4 | A vs B | $3.71*10^{-39}$ | *** |
| | A vs F | $2.98*10^{-4}$ | ** |
| | B vs F | $8.51*10^{-29}$ | *** |
| W22 | A vs E | $1.55*10^{-22}$ | *** |
| | A vs F | $5.92*10^{-3}$ | * |
| | A vs H | $4.78*10^{-9}$ | ** |
| | E vs F | $2.73*10^{-5}$ | ** |
| | E vs H | $2.17*10^{-4}$ | ** |
| | F vs H | $1.56*10^{-1}$ | ns |
| W48 | A vs E | $1.80*10^{-2}$ | * |
| | A vs G | $1.34*10^{-1}$ | ns |
| | E vs G | $3.61*10^{-1}$ | ns |
| W33 | F vs L | $1.93*10^{-1}$ | ns |
| W53 | A vs C | $8.65*10^{-1}$ | ns |

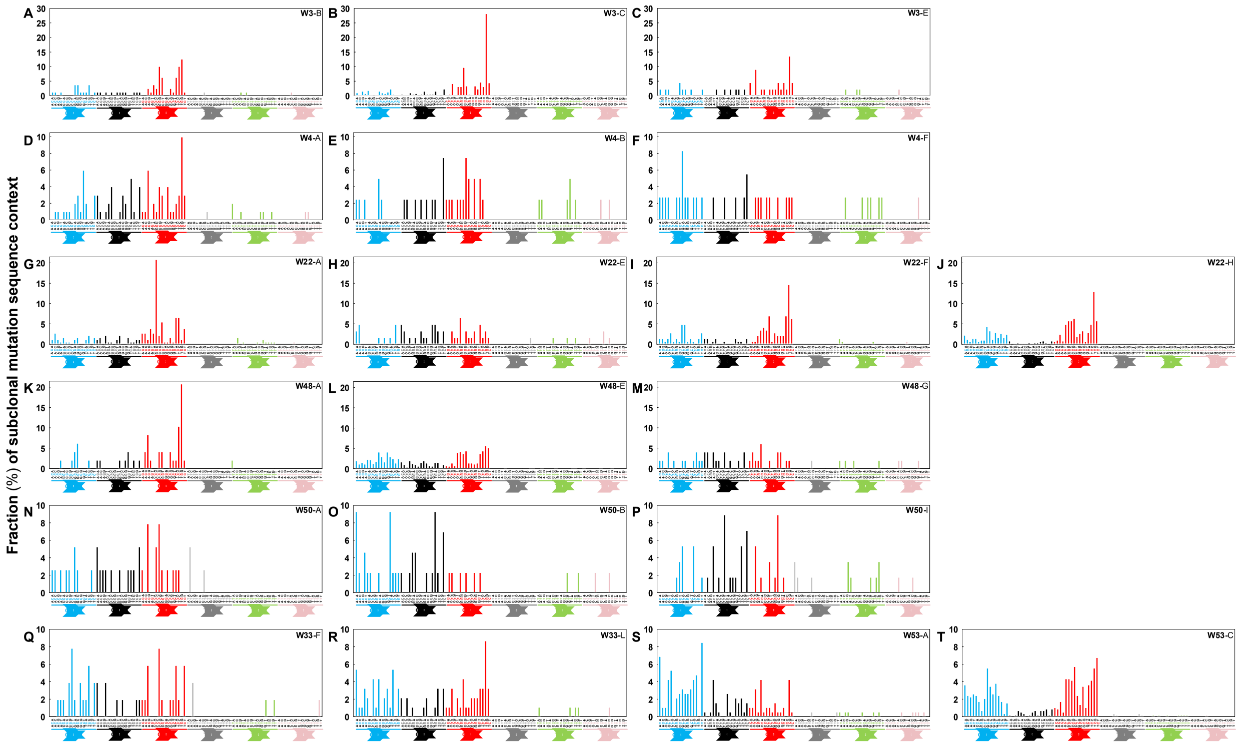

**Figure S4.** Fractions (%) of the subclonal mutation sequence context spectra (MCS) were determined using Duplex Sequencing for 20 GBM specimens from seven patients. Trinucleotide contexts are mutated bases surrounded by all possible combinations to its flanking 5' and 3' bases. To keep the graph concise, these point mutation trinucleotides are complemented as necessary to always depict the reference base as the pyrimidine of its pair.

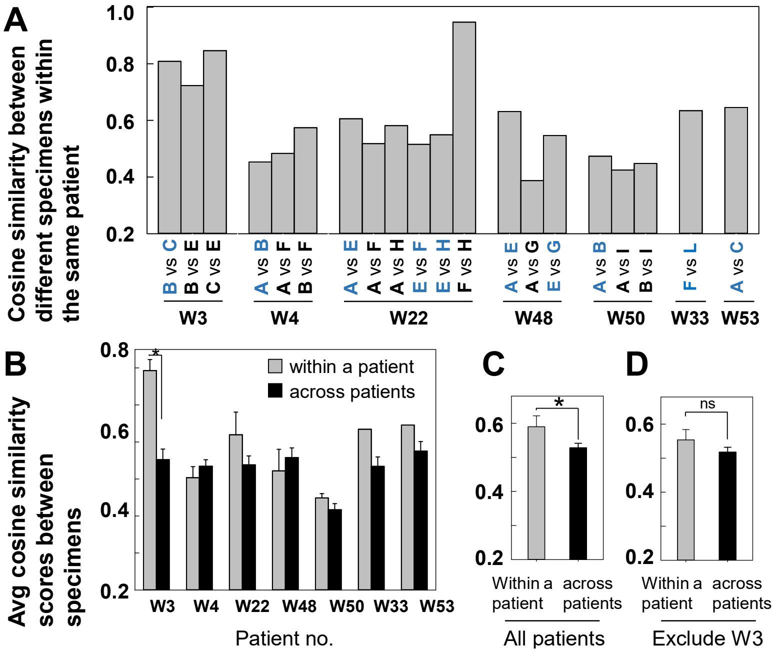

**Figure 4.** The sequence context spectra of subclonal mutations between specimens were compared within a patient and across patients using Cosine similarity. (A) Cosine similarity scores between each pair of specimens within the same patient. The specimens located closely to each other are highlighted in blue. (B-D) Average (Avg) cosine similarity scores 'within a patient' reflect comparisons between each pair of specimens within the same patient (gray bars). Avg cosine similarity scores 'across patients' indicate comparisons between a pair of specimens from two different patients (black bars). Avg cosine similarity scores for within a patient and across patients from all the seven patients (C) or from six patients except a patient W3 (D). **p* < 0.05 by the Mann-Whitney U test.

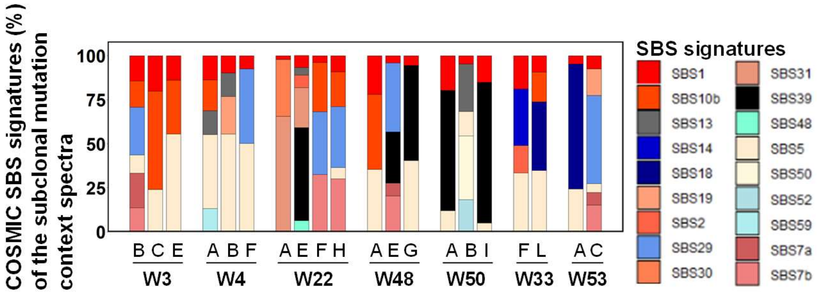

**Figure 5.** Fraction (%) of COSMIC v3.1 signatures. The sequence context spectra of subclonal unique mutations were decomposed to compare with the single base substitution (SBS) signatures of the COSMIC v3.1. Fractions (%) of major SBS signatures identified by SigProfiler for 20 GBM specimens of seven patients are displayed.