Genomics 105 (2015) 159–167

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Rapid and convergent evolution in the Glioblastoma multiforme genome

Max Shpak [a,b,c,*], Marcus M. Goldberg [b], Matthew C. Cowperthwaite [a,b,d]

[a] NeuroTexas Institute, St. David's Healthcare, Austin, TX 78705, United States
[b] Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, TX 78712, United States
[c] Fresh Pond Research Institute, Cambridge, MA 02140, United States
[d] Texas Advanced Computing Center, University of Texas at Austin, Austin, TX 78758, United States

ARTICLE INFO

Article history:
Received 18 August 2014
Accepted 31 December 2014
Available online 7 January 2015

Keywords:
Glioblastoma multiforme
Somatic evolution
Missense mutation
Convergent evolution
Neutral evolution

ABSTRACT

Determining which mutations drive tumor progression is a defining question in cancer genomics. We analyzed sequence evolution in Glioblastoma multiforme (GBM) by computing the number of parallel mutations and by estimating $\omega = dN/dS$, a measure of the strength and direction of selection. The ω values of almost all 7617 mutated genes in GBM are much higher than in germline genes. We identified only 21 genes under significant positive selection in GBM, as well as 29 genes under significant purifying selection, including several zinc finger proteins. Therefore, most of the high ω values in the GBM genome are due to weaker purifying selection rather than positive selection. We also found multiple recurrent mutations in GBM, several of which are associated with patient survival time. Our results suggest that convergence and neutral evolution play a significant role in GBM, and that sites with recurrent mutations can serve as molecular diagnostics of the clinical course of GBM tumors.

1 . Introduction

Glioblastoma multiforme (GBM, a WHO grade IV glioma or grade IV astrocytoma), is the most common type of primary brain cancer, with approximately 17,000 new cases diagnosed annually in the USA alone. It is also the most aggressive and lethal; following diagnosis, the median survival time is approximately 14 months even with intensive treatment, and fewer than 5% of patients survive beyond the third year [21,23]. Like most cancers, GBM develops through a complex suite of genetic alterations [2,12,13], rather than being induced by any single key mutation. Cancer progression is driven by a process of somatic selection [15,32,33], and recent studies have focused on genes relevant to GBM progression by identifying those with the highest density of missense mutations [41,4,47,48]. However, a common assumption that the genes with the greatest number of missense substitutions are the ones under the strongest positive selection in cancer is not necessarily correct. Higher missense mutation counts could also be associated with higher silent mutation counts and be indicative of overall increases in the mutation rate or relaxed purifying selection (neutral evolution) rather than positive selection. Consequently, in this study we use the ratio of missense to silent substitutions as an indicator of selection acting on genes in GBM; we also analyze the frequency of parallel mutations (somatic mutations appearing in multiple independent tumors) as an additional indicator of selection and convergent evolution.

During the process of somatic evolution, the genomes of cancer cells diverge from those of normal tissues through the accumulation of mutations via natural selection and genetic drift. The rate at which substitutions of non-synonymous mutations occur relative to synonymous mutations between the tumor and normal sequences provides an estimate of the strength and direction of natural selection. We use the value $\omega = dN/dS$ as a measure of the strength of selection, where dN is the fraction of non-synonymous substitutions relative to the number of non-synonymous sites and dS is the fraction of synonymous substitutions relative to the number of synonymous sites in the gene [37,38,54, 10]. Typically, one infers purifying selection for $\omega << 1$ and positive selection when $\omega >> 1$. In fact, the value of ω can deviate from one due to nucleotide-specific mutation biases, therefore inferences of selection should be made with respect to appropriate background values (e.g. those computed by comparing human gene sequences to homologs in outgroup species).

Many cancer genomics efforts have focused on rapidly evolving genes and have not used methods designed to detect essential tumor genes with low substitution rates, including activated oncogenes or compromised tumor-suppressor genes. Another approach to the discovery of functionally significant genes under positive selection in GBM is to identify parallel (or recurrent) mutations, which are identical nucleotide substitutions found at the same site in tumors from different patients. Because the independent random fixation of the same mutation in different patients is highly improbable, parallel mutations can provide powerful evidence of positive directional selection on GBM genes, as well as a means of identifying specific sites. For example, a recent study [34] found identical missense mutations in the *IDH1* gene

[*] Corresponding author.
E-mail address: shpak.max@gmail.com (M. Shpak).

http://dx.doi.org/10.1016/j.ygeno.2014.12.010
0888-7543/© 2015 Elsevier Inc. All rights reserved.

in a significant fraction of GBM tumors. An advantage to this approach is that it can identify biologically significant genes that have small overall numbers of mutations and would therefore be missed by approaches based on mutation counts, including estimates of ω.

This study also investigates the association between the number of missense mutations in genes with significant mutational recurrence and post-diagnosis survival times of patients, with the working hypothesis that genes under strong selection relate to the rate of tumor growth and resistance to immune response and therapies, which in turn determines the time until recurrence and the duration of patient survival.

2 . Materials and methods

2.1 . Identification of somatic mutations and mutated genes

A summary of the data processing and analysis is shown schematically in Fig. 1. All of the bam format files used in this analysis were obtained from the GBM data set in the Cancer Genome Atlas (TCGA), https://tcga-data.nci.nih.gov/tcga/. The sequence data are whole-exome from the Agilent HumanSureSelect 50 Mb capture chip, which also contains extensive non-coding flanking regions and introns for many of the genes. In the case of duplicate samples from the same patient, those with the highest mean read depth were selected. Recurrent tumors were excluded from the analysis, because they typically were sampled from patients who had undergone radiation or chemotherapy prior to sampling (i.e. levels of genetic diversity in recurrent tumors are not representative because of effective bottlenecks on cell lineages following therapy).

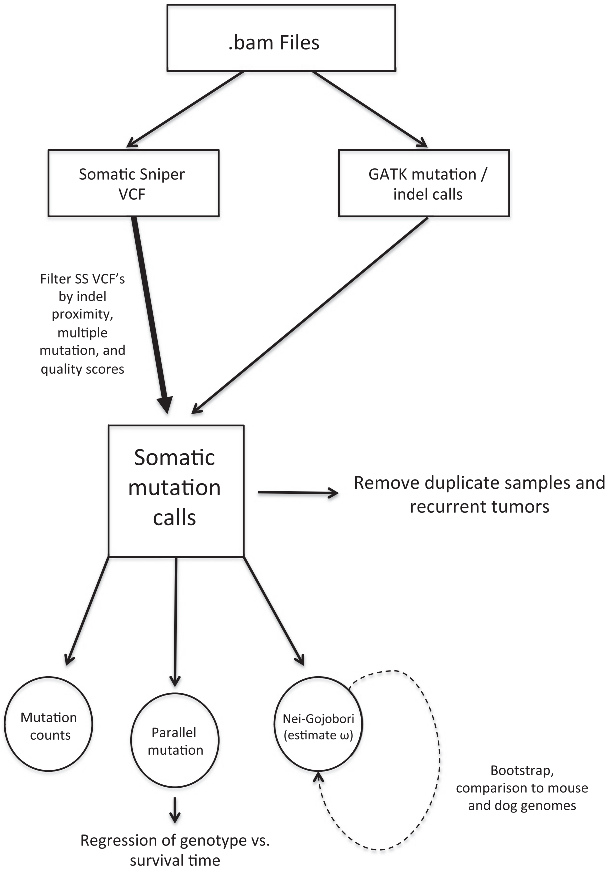
.bam Files
Somatic Sniper VCF
GATK mutation / indel calls
Filter SS VCF's by indel proximity, multiple mutation, and quality scores
Somatic mutation calls
Remove duplicate samples and recurrent tumors
Mutation counts
Parallel mutation
Nei-Gojobori (estimate ω)
Bootstrap, comparison to mouse and dog genomes
Regression of genotype vs. survival time

Fig. 1. Flow chart with a schematic of the data analysis pipeline used in this study, starting with nextgen sequence data.bam files.

Somatic mutations were called by SomaticSniper [25], a program which uses joint information from blood and tumor samples from the same individual to estimate conditional probabilities of tumor and blood. The quality scores of mutation calls are log-scale likelihood functions of this conditional probability. GATK [30] was also used to cross-validate genotype calls in certain instances, and to identify loci in indel regions.

We scored the following blood/tumor genotypes as somatic mutations (where 0 represents reference and 1 variant nucleotide): blood:{0/0} vs. tumor:{0/1,1/1} or blood:{1/1} vs. tumor:{0/1,0/0} (the latter correspond to those rare instances where the patient's genotype is homozygous variant with respect to reference genome, while mutations in the tumor happen to revert back to reference nucleotide). Because of the different genetic processes (e.g. mitotic recombination, gene conversion [3]) and selective pressures that generate and maintain them, loci with instances of loss of heterozygosity (LOH) were excluded from this study.

Somatic mutation calls were filtered so that only sites with a read depth of ≥10 in both tumor and blood genotypes contribute to somatic mutation calls, in order to reduce the probability of calling homozygosity through sampling error at a heterozygous locus. Similarly, a genotype quality score of ≥20 was required for both tumor and blood at each site called as a somatic mutation. In addition (following [25]), the loci with any of the following characteristic were excluded in order to avoid errors due to sampling or misreads: *a*. all sites identified as SNP loci on dbSNP, *b*. all mutation calls within 10 bp of an indel site called by GATK with a quality score ≥ 50, and *c*. all mutations located within 10 kb of at least 3 additional mutations.

Mutation counts in each sample were tallied to identify samples with extensive read errors or contamination. Specifically, samples with mutation counts more than 2 standard deviations above the per-sample mutation count mean were removed (including several derived from whole-genome amplification sequences). A list of TCGA accession numbers the 283 retained paired samples is provided in the supplementary document S7.

2.2 . Mutation counts and parallel mutation

Variant nucleotides at somatic mutation loci were mapped onto coding sequences using the reference human genome CDS and coordinates from the UCSC Genome Browser https://genome.ucsc.edu/cgi-bin/hgTables?command=start (2009 assembly, version GRCh37/hg17) to generate matched pairs of reconstructed normal and GBM exomes for every sample. Convergent substitutions across samples were tallied from the tumor sequences. Bootstrapped permutation of variant nucleotides in a CDS (see below) was used to determine the minimum number of parallel mutations that is statistically significant. Unless otherwise indicated, the mutation counts and other analyses were written and executed in Python 2.7.2, with code available from the corresponding author upon request. The mapped mutations were characterized as missense, silent (synonymous), or nonsense. This data, together with gene identification and mutation coordinates, were used to construct Mutation Annotation Files (MAF), as input for identifying significantly mutated genes with MutSig1.4 [26]. MutSig uses background mutation frequencies to compute the beta binomial probabilities of an observed number of missense mutations in a gene given background mutation frequencies across samples. As a heuristic, we consider "highly mutated genes" (genes of interest) as those with unadjusted $p < 0.05$, and "significantly mutated" genes as those with Benjamin–Hochberg FDR-adjusted $q < 0.05$. These gene sets will be compared to those under significant directional selection and to those with recurrent mutation.

Mutations in non-coding regions (defined as those regions of chromosome captured on the Agilent SureSelect human whole exome.bed file that are located outside annotated UCSC exon coordinates) were tallied separately for comparison.