Genome-wide Analyses Identify Recurrent Amplifications of Receptor Tyrosine Kinases and Cell Cycle Regulatory Genes in Diffuse Intrinsic Pontine Glioma

Paugh, et al

Supplemental Materials and Methods

Tumor Samples:

Brainstem low-grade gliomas (WHO grade I or II) (n= 8 for copy number analysis), were resected from patients with median age 4.2 years, range 2-11.4 years. Diagnoses were 7 pilocytic astrocytomas and one low-grade astrocytoma. Brainstem low-grade gliomas evaluated for expression analyses included 4 of the samples above and 2 additional LGG samples. The age range for this cohort was 2-13 years, median age, 9.2 age, diagnoses, 5 pilocytic astrocytomas, 1 ganglioglioma.

Non-brainstem low-grade (WHO grade I or II) gliomas (n=66) were resected from children or young adults aged 1 - 19 years (mean = 7.9 years; median = 6.5 years). Tumor site was varied: cerebral cortex - 27; cerebellum - 24, diencephalon - 12, spinal cord - 3. Pathological diagnoses were: pilocytic / pilomyxoid astrocytoma - 39, diffuse astrocytoma - 14, mixed glioma - 8, pleomorphic xanthoastrocytoma - 2, angiocentric glioma - 2, subependymal giant cell astrocytoma - 1.

Copy Number Analysis

The copy number of the tumor samples as well as the available matched normal samples was profiled using Affymetrix SNP6.0 arrays. These arrays were analyzed using the following steps:

1) ***Normalization***: First CHP files were generated using APT/1.10-64bit (Affymetrix) with birdseed algorithm[3]. Only normal samples used in this study were used to generate the CHP files. Chip intensities were extracted using dChipSNP[4,5], and SNP probes and CN probes were separated for later use. SNP probes were used to select a set of diploid chromosome(s) for each sample, with consideration of both SNP intensity as well as SNP calls. With the selected diploid chromosomes, SNP probes and CN probes were normalized separately using reference-based normalization method [6].

2) ***Ratio calculation and segmentation***: To calculate the ratio for each sample, we selected either 1 or 5 nearest normal samples based on Euclidean distance of SNP probes. In detail, for each sample, the Euclidean distance between the sample and all the normal samples were calculated using the SNP probes only (chrX and chrY were excluded). Then we sorted the Euclidean distance from smallest value to largest value and calculated the standard deviation among the distances. Next we compared the smallest value and the second smallest value to decide if one or 5 normal samples should be used. If the second smallest value is larger than the sum of the smallest value and the standard deviation, we chose one sample, which is the matched normal, to calculate the ratio. Otherwise, we chose five normal samples with smaller distance to calculate the ratio. This includes tumors with no matched normal samples or tumors with matched normal but run in two different batches. The log2

ratios were then calculated for all samples for both SNP probes and CN probes, which were then merged together and sorted based on chromosomal locations. Circular binary segmentation (CBS) algorithm (DNAcopy, Bioconductor,[7]) was then applied to the sorted $\log_2$ ratio data. The segmentation on the smoothed $\log_2$ ratios was performed using a *p*-value threshold of 0.01 (significance level a = 0.01). We assigned gain or loss status to the segments with at least 8 probes and the absolute $\log_2$ ratios higher than 0.2, with which the normal samples have few gain or loss regions.

3) ***Copy Number Variant Identification***: Before applying GISTIC analysis or focal lesion identification, we first removed the genomic regions which are associated with copy-number variations (CNVs). Identification of excluded regions was similar to the approach used by the TCGA. The regions include:

    a. CNVs found in a SNP6.0 analysis of all HapMap normals;
    b. CNVs identified in at least two independent publications listed in the Database of Genomic Variants (DGV, http://projects.tcga.ca/variation, version 3);
    c. CNVs found in the collected normal samples by manual inspection, as described[1] ;
    d. CNVs indentified in previous SNP array studies[8];
    e. Regions with more than 90% of the bases located in repeat regions.

4) ***Global changes***: We compared chromosome arm changes between pediatric diffuse intrinsic pontine gliomas (DIPGs) and pediatric glioblastomas (GBMs) arising outside of the brainstem and adult GBMs. If more than half of the markers on a chromosome arm had copy number gain or loss, then the entire arm was classified as gain or loss. The pediatric supratentorial GBM data are from[1] and for adult GBMs we used level 3 (segmented) SNP6.0 TCGA data downloaded from the TCGA portal (http://cancergenome.nih.gov/dataportal/data/about/) in February 2009.

5) ***Candidate targets of focal gain or loss***: We derived minimum common regions for recurrent focal gains (copy number > 2.3) or recurrent focal deletions (copy number < 1.7) found in at least two tumors or were classified as a single focal gain or deletion. Regions associated with known CNVs were removed as described earlier[1] and above. All remaining regions with less than 60 genes were manually inspected for cancer/glioma-related genes, and candidate targets of focal gain or deletion were selected. Large regions of copy number imbalance were determined by merging the segmented data to the neighboring segment if the segment loss or gain call was concordant. Large regions of copy number imbalance were defined as larger than 25 Mbp with copy number > 2.3 or copy number < 1.7.

6) ***LOH analysis***: LOH analysis was done using dChipSNP [4,5], only for samples with paired normal samples. The LOH call was based on the paired normal samples.

Expression Profiling

Expression data was analyzed using Affymetrix Microarray Suite software. Gene expression signals were scaled to a target intensity of 500. Probe sets lacking present calls for any samples were excluded. Signals were then variance-stabilized by adding 25 and $\log_2$ transformed for subsequent analysis.

Unsupervised Hierarchical Clustering analysis and identification and functional annotation of signature genes

We calculated the median absolute deviation (MAD) score for each probe set using the $\log_2$ transformed data and selected the top 1000 most variable probe sets for unsupervised hierarchical clustering analysis (UHC). UHC was carried out using GeneMaths software (Applied Maths, Inc., Austin, TX), using Pearson correlation as the similarity coefficient and Ward as the clustering method. Three major tumor subgroups were identified.

We then identified probe sets that are most up-regulated in each subgroup by comparing one group against the other two. The *limma* (Linear Models for Microarray Analysis) [9] and empirical Bayes *t*-test implemented in Bioconductor [10] (www.bioconductor.org) were used to identify differentially expressed probe sets at Benjamini-Hochberg false discovery rate (q-value) of <0.1 and fold change of >2. The heatmap was generated using 150 probe sets from each group.

Gene Set Enrichment Analysis (GSEA) of known HGG signature genes in pediatric DIPG

GSEA implemented in R (www.r-projects.org) was used to assess enrichment of previously identified adult and pediatric HGG signature genes[1,11,12] in pediatric DIPG. The upregulated genes in each subgroup from the supplemental tables were used to define the gene sets. We applied the collected gene sets with GSEA to the pediatric DIPG, with one subgroup against the other two. The results of statistical analysis were listed in Table S5, with the gene set showing highest significance for each subgroup shown in Figure 3.

Principal Component Analysis

The 27 pediatric DIPG samples, 2 pre-treatment DIPGs previously published, and 51 non-brainstem pediatric HGGs [1], or with additional 6 brainstem LGGs and 66 non-brainstem LGGs were pooled together for principal component analysis (PCA) using GeneMaths. Based on the median absolute deviation (MAD) score for each probe set using the $\log_2$ transformed data, we selected the top 1000 most variable probe sets for PCA analysis.

Differentially expressed genes between pediatric DIPG and pediatric glioblastomas arising outside the brainstem and functional annotation

We applied Limma/Bioconductor to identify differentially expressed genes between pediatric DIPG and HGG. We selected 1480 probe sets using a cut-off of q-value <0.01 and minimum fold change of 2. To further characterize the set of genes, we did Gene Ontology analysis using DAVID Bioinformatics Resources (http://david.abcc.ncifcrf.gov).

Real-Time Quantitative Polymerase Chain Reaction

We validated inferred copy number analyses using quantitative real-time polymerase chain reaction (PCR) for 10 loci (Table S1). Primers and probes (Table S1) were designed using Primer 3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). GAPDH was used as internal standard to normalize the data. Two nanograms of DNA from the DIPG samples or control normal DNA were amplified using a Taqman 7900 Real-Time PCR system and 7900 System Software (Applied Biosystems). Standard curves for each locus tested were generated from three-fold serial dilutions of control human WBC DNA. Quantitative real-time PCR for each primer set was performed in triplicate, and means are reported. The concordance with the inferred SNP copy number was analyzed using Pearsons correlation. For *AKT3*, *BRCA2*, *CDK6* and *MET*, a Fast SYBR Green kit (Applied Biosystems) was used with the following PCR conditions: 95°C for 20 seconds, then 40 cycles of 95°C for 5 seconds and 60°C for 30 seconds. At the end of the PCR, samples were subjected to a melting analysis to confirm specificity of these amplicons. For *CDK4*, *CDKN2A*, *DLK1*, *PDGFRA*, *PTEN* and *TP53* fast-mode quantitative PCR was performed using TaqMan Fast Universal PCR Master Mix from Applied Biosystems (95°C for 20 seconds, then 40 cycles of 95°C for 5 seconds and 60°C for 30 seconds).

Fluorescent In Situ Hybridization

Where formalin-fixed paraffin-embedded material was available (28 DIPG samples), fluorescent in situ hybridization for probes identifying PDGFRA (4q12), MET (7q31), EGFR (7p12) and IGF1R (15q26.3), was performed as described [13]. A probe directed against PDGFRA (a pool of BAC clones RP11-231C18 and RP11-601I5) was labeled with rhodamine (Roche) and control probe for chromosome 4p (a pool of BAC clones CTD2057N12 and CTD2588A19) was labeled with AlexaFluor488 (Invitrogen). MET and IGF1R probes labeled with Platinumbright 550 and SE7 and 15q11 control probes labeled with Platinumbright 495 were obtained from Kreatech (Amsterdam, the Netherlands). EGFR probe (a pool of BAC clones RP11-148P17 and RP11-1083E20) was labeled with AlexaFluor488 and the control probe for chromosome 7q (a pool of BAC clones RP11-460J21 and CTB-133K23) was labeled with rhodamine. An additional probe for MET (BAC clone RP11-163C9) was labeled with AlexaFluor488 for co-hybridization with the PDGFRA probe labeled with rhodamine. Images were captured using the AI Cytovision software (Applied Imaging, Santa Clara, CA).

Statistical Analyses

A Kruskal-Wallis Test was performed to determine if there was an association between age at diagnosis or overall survival and the following molecular characteristics: PI3K pathway alterations, RB pathway alterations, PI3K and RB pathway alteration, focal gain of PDGFRA, MET, or IGF1R, large scale gain of

1q, 2p, 2q, 8q or 9q, large scale loss of 10q, 11p, 13q, 14q, 16q, 17p or 20p, or the gene expression subgroups Proliferative, Proneural and Mesenchymal. Associations between the same molecular markers and the pattern of disease progression at first recurrence (local only versus local plus additional sites) was tested by Fisher’s exact test and also by logistic regression with pattern of progression set as the dependent variable to explore associations between it and each molecular feature. A Bonferroni adjusted p-value threshold for determining statistical significance based on a family-wise error rate of 0.1 would be approximately 0.001. There were no significant associations between molecular features and age at diagnosis, overall survival or pattern of disease progression at first recurrence based on this threshold.

References

1. Paugh BS, Qu C, Jones C, et al: Integrated molecular genetic profiling of pediatric high-grade gliomas reveals key differences with the adult disease. J Clin Oncol 28:3061-8, 2010
2. TCGA: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:1061-8, 2008
3. Korn JM, Kuruvilla FG, McCarroll SA, et al: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40:1253-60, 2008
4. Lin M, Wei LJ, Sellers WR, et al: dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20:1233-40, 2004
5. Zhao X, Li C, Paez JG, et al: An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64:3060-71, 2004
6. Pounds S, Cheng C, Mullighan C, et al: Reference alignment of SNP microarray signals for copy number analysis of tumors. Bioinformatics 25:315-21, 2009
7. Olshen AB, Venkatraman ES, Lucito R, et al: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5:557-72, 2004
8. Mullighan CG, Phillips LA, Su X, et al: Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. Science 322:1377-80, 2008
9. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3, 2004
10. Gentleman RC, Carey VJ, Bates DM, et al: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80, 2004
11. Phillips HS, Kharbanda S, Chen R, et al: Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9:157-73, 2006
12. Verhaak RG, Hoadley KA, Purdom E, et al: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17:98-110, 2010
13. McManamy CS, Pears J, Weston CL, et al: Nodule formation and desmoplasia in medulloblastomas-defining the nodular/desmoplastic variant and its biological behavior. Brain Pathol 17:151-64, 2007