Translational Oncology
www.transonc.com

Volume 12 Number 7 July 2019 pp. 908–916 908

Host Genome Variation is Associated with Neurocognitive Outcome in Survivors of Pediatric Medulloblastoma[1]

Benjamin I Siegel[*], Tricia Z King[†], Manali Rupji[‡], Bhakti Dwivedi[‡], Alexis B Carter[§], Jeanne Kowalski[‡,¶] and Tobey J MacDonald[*,#]

[*]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA; [†]Department of Psychology and Neuroscience Institute, Georgia State University, Atlanta, GA; [‡]Winship Cancer Institute of Emory University, Atlanta, GA; [§]Department of Pathology and Laboratory Medicine, Children's Healthcare of Atlanta, Atlanta, GA; [¶]Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, GA; [#]Aflac Cancer & Blood Disorders Center, Children's Healthcare of Atlanta, Atlanta, GA

Abstract

Host genome analysis is a promising source of predictive information for long-term morbidity in cancer survivors. However, studies on genetic predictors of long-term outcome, particularly neurocognitive function following chemoradiation in pediatric oncology, are limited. Here, we evaluated variation in host genome of long-term survivors of medulloblastoma and its association with neurocognitive outcome. Whole-genome sequencing was conducted on peripheral blood of long-term survivors of pediatric medulloblastoma who also completed neuropsychological testing. Cognitively impaired and less impaired survivors did not differ in exposure to chemoradiation therapy or age at treatment. Unsupervised consensus clustering yielded two distinct variant clusters that were significantly associated with neurocognitive outcome. Interestingly, 34 of the 36 significant variants were found in noncoding DNA regions with unknown regulatory function. A separate unsupervised cluster analysis of variants within DNA repair genes identified discrete variant groups that were not associated with neurocognitive outcome, suggesting that variations in genes corresponding to a single functional group may be insufficient to predict long-term outcome alone. These findings are supportive of the presence of a genetic diathesis for treatment-related neurocognitive morbidity in medulloblastoma that may be driven by variation in noncoding regulatory elements.

Translational Oncology (2019) 12, 908–916

Introduction

Advances in the treatment of medulloblastoma, the most common central nervous system malignancy in children, have led to a substantial increase in survival [1,2]. Despite this improvement, survivors are at a high risk of long-term neurocognitive impairment, largely driven by core cognitive abilities of processing speed, working memory, and attention [3–6]. There is a considerable degree of heterogeneity in neurocognitive outcome that cannot be entirely explained by molecular tumor subtype, cranial radiation dose, or age at treatment. A growing body of evidence suggests the presence of genetic determinants that predispose some brain tumor survivors to experience marked cognitive impairment following treatment, whereas others experience only mild deficits [7].

Address all correspondence to: Tricia Z. King, PhD, Professor, P.O. Box 5010, Atlanta, GA 30302-5010, USA. E-mail: tzking@gsu.edu

[1] Funding: This work was supported by the Aflac Cancer & Blood Disorders Center Pediatric Hematology-Oncology Research Grant (T.Z.K. and T.J.M.); the American Cancer Society [#RSGPB-CPPB-114044, T.Z.K.]; the Pediatric Research Alliance Center for Neurosciences Research and Children's Healthcare of Atlanta (#00060319, T.J.M, B.I.S., T.Z.K.); and the Developmental Neuropsychology Across the Lifespan graduate students who received fellowships from GSU Brains & Behavior (S.N., M.F., E.S, R.J., A.P, J.M.), 2CI Neuroimaging (R.B.), 2CI Neurogenomics (R.K.), and Language & Literacy (A.A., K.S.) Initiatives.
Received 18 January 2019; Revised 20 March 2019; Accepted 26 March 2019

1936-5233/19
https://doi.org/10.1016/j.tranon.2019.03.004

Studies examining neurocognitive outcome in medulloblastoma have focused primarily on tumor genome rather than host genome and particularly on the four molecular subgroups of medulloblastoma: WNT, SHH, group 3, and group 4. These subgroups are biologically distinct and correlate with response to treatment and overall prognosis [8]. One retrospective study using a broad range of cranial radiation exposure found that SHH patients had less decline in processing speed compared to other subgroups, although the difference was small and did not fully explain the heterogeneity in cognitive outcome [9].

To date, studies investigating host genetic predictors of treatment-related toxicity outcomes in cancer survivors have targeted specific genes or gene groups. Early reports identified specific host genetic single nucleotide polymorphisms (SNPs) in DNA repair genes associated with radiation toxicity in adults with breast, lung, prostate, and head and neck cancers [10–14]. Variation in DNA repair genes also has been specifically linked to neurocognitive outcome in medulloblastoma survivors [15–17]. In recent years, SNPs identified from genes involved in other functional pathways including intrinsic cognitive function, neurotransmitter production, and inflammatory pathways have been shown to predict neurocognitive outcome in brain tumor survivors [18–20].

A relatively new area of investigation suggests that examining whole genome SNP variation, rather than targeting individual SNPs, may allow for more robust prediction of outcome. This approach has been employed successfully using tumor samples, including medulloblastoma [21,22], to predict long-term survival and response to treatment but has not been applied to the evaluation of the relationship between host genome and long-term cognitive outcome. Besides detecting novel target genes, genome sequencing data can be used to identify clusters of variants that collectively increase risk for poor outcome even if each variant has a small individual effect [23–25].

The heterogeneity of adverse long-term neurocognitive outcomes, coupled with emerging evidence of genetic determinants of long-term cancer survival, suggests that there may be distinct genomic profiles influencing how patients respond to the intensive chemoradiation therapy involved in medulloblastoma treatment. Identifying robust variant profiles that are predictive of cognitive outcomes has the potential to both provide patients with personalized prognostic information at the time of diagnosis and to facilitate clinicians' development of targeted interventions to offset long-term neurocognitive morbidity.

In the present study, we conducted whole genome sequencing on serum from long-term medulloblastoma survivors. We then performed two sets of unsupervised hierarchical cluster analyses to assess whether discrete gene variant profiles were associated with neurocognitive outcome. In the first analysis, variants from all disease-associated genes were included. Then, to test the hypothesis that differences in DNA repair specifically predispose to radiotoxicity, a second analysis limited to genes within this pathway was performed. Finally, we compared allelic frequencies of significant variants in our sample to those of the general population using sequencing data aggregated from over 100,000 individuals.

Materials and Methods

Study Participants and Data Acquisition

The study was approved by the Institutional Review Board of the Georgia State University/Georgia Tech Joint Center for Advanced Brain Imaging, Emory University, and Children's Healthcare of Atlanta (CHOA). Informed consent was obtained from all participants or their legal guardians where appropriate. The study took place at the Center for Advanced Brain Imaging in accordance with the relevant guidelines and regulations. All participants were long-term survivors of childhood medulloblastoma. Long-term survivorship was defined as being at least 5 years from completion of therapy and without any clinical evidence of residual or recurrent tumor. Individuals were excluded from the study if they had a history of moderate to severe traumatic brain injury, major psychotic disorders, neurofibromatosis, cancer predisposition syndromes, or recurrent or progressive medulloblastoma. Clinical and demographic information was obtained by review of participants' electronic medical records. All medulloblastoma survivors treated at CHOA who met inclusion criteria were invited to participate in the study. Molecular subgroup classification was performed on tumor samples by NanoString assay according to established methods [26,27]. Genomic sequencing data are available in the European Genome-Phenome Archive under accession number EGAD00001004115.

Neuropsychological Measures and Cognitive Impairment Classification

Eighteen participants completed a neuropsychological evaluation. The battery consisted of widely used clinical tests with well-developed norms. Given the large number of cognitive performance measures available, we employed a composite neuropsychological score based on key cognitive components of neurodevelopmental models of long-term outcomes of childhood brain tumors [4–6]. The composite score computed the average z scores for the following performance measures: oral processing speed (Oral Symbol Digit Modality Test) [28], working memory (Auditory Consonant Trigram) [29,30], attention span (Digit Span Forward subtest from Wechsler Memory Scale) [31], Verbal IQ (Wechsler Abbreviated Scale of Intelligence Vocabulary & Similarities) [32], Performance IQ (WASI Block Design & Matrix Reasoning) [32], and reading academic achievement (Letter Word Identification subtest of the Woodcock Johnson Tests of Achievement, Third Edition) [31,33]. Consistent with the literature, clinically significant cognitive impairment was defined as an average z score of $-1.5z$ or lower [34].

Genome Sequencing and Bioinformatics analysis

Whole genome DNA sequencing was performed on blood samples from 22 pediatric medulloblastoma survivors on the Illumina HiSeq X platform as described in Johnston et al. (2017) [35]. Following sequencing, all base calling was performed using standard Illumina software to generate the final FastQC files for each sample. The quality of raw reads generated from Illumina sequencing was assessed using FastQC [36]. Reads were filtered and trimmed using the Trimmomatic tool [37]. BWA aligner was used to map post-quality filtered reads against the human reference genome (hg19) [38]. The alignment quality was evaluated using SAMtools [39] and Picard-Tools (http://picard.sourceforge.net). The mean target coverage was 30×, and 95% of the targeted bases have a coverage of 10× or greater. Potential PCR duplicates were removed with Picard-Tools. Somatic variants (SNV and Indel) were called using SAMTools [39] with Varscan2 [40] and annotated using ANNOVAR [41]. Variants with low-quality read depth (<6×) were excluded from the analysis. A variant proportion was estimated for each gene variant for each sample. Here, variant proportion is defined as the reads supporting

the variants divided by the total number of reads supporting the variant and the reference allele, hence ranging from 0 to 1. A value of 0 means no reads supporting the variant have been identified, a value of 0.5 means half of the reads support variant and half support reference allele, and a value of 1 means all reads are supporting the variant allele. Genomewide variants were subsetted into variants identified to be associated with disease as reported in the COSMIC v81 and ClinVar2017 databases using a 10× coverage threshold. Variants were also subsetted into those associated with DNA repair function as reported in the REPAIRtoire (http://repairtoire.genesilico.pl/), Human DNA Repair Genes (http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html), and repairGenes (http://www.repairgenes.org/) databases.

Statistical Analysis

To test the association between gene variant profile and cognitive outcome, two sets of unsupervised hierarchical agglomerative clustering analyses were performed: first on variants within all disease-associated genes and then on variants within DNA repair genes alone. To test the association of the disease-associated gene variants with the cognitive outcome, disease-associated gene variants were identified using the cosmic 81 and Clinvar 2017 databases. Among them, the top most variable gene variants that had an interquartile range of greater than or equal to 99.5th percentile were identified and used for downstream analysis. Variant proportions (p) were transformed using $\log2((1 + p)/(1 - p))$, and a two-sided t test between the impaired and less impaired samples was performed to identify clinically significant gene variants. Fold-change was calculated as the difference in the mean transformed proportions between the two sample groups. Statistically significant variants (false discovery rate, FDR < 0.05) were defined as core gene variants and used to generate the heatmap. Unsupervised (within the context of samples) hierarchical agglomerative heatmap clustering using the original variant proportions was carried out using euclidean distance and ward.D clustering. Heatmap clustering analysis was conducted using NOJAH [23].

For the DNA repair gene analysis, we first identified the topmost variable gene variants in DNA repair genes identified from the REPAIRtoire, Human DNA Repair Genes, and repairGenes databases using an interquartile range of greater than 99th percentile. We then identified sample clusters using ConsensusClusterPlus R package with 1-Pearson correlation distance, ward.D agglomerative hierarchical clustering, 80% item resampling, 80% gene resampling, and 1000 resamplings [42]. To define a set of core gene variants, the topmost variable gene variant proportions were first transformed into binary values: variants with gene variant proportion ≤ 0.5 were coded to 0, and those with variants proportion> 0.5 were coded to 1. Core gene variants were defined as those that were significantly (P value < .05; FDR < 0.326) associated with the core sample clusters based on a Fisher's exact test. Unsupervised (in the context of the samples) hierarchical agglomerative clustering was performed on the original variant proportion values for the core gene variants and core samples using a 1-Pearson correlation distance and ward.D clustering [43,44].

In order to assess whether the frequency of the disease-associated variants in the study sample differed from the general (nonstudy) population, allele frequencies (AFs) were calculated in aggregate from three online human genomic variation databases: 1000 genomes project (phase 3), the NHLBI GO Exome Sequencing Project, and the Genome Aggregation Database (gnomAD) [45–47]. Variants were referenced by their dbSNP identification number [48]. For each of the 36 variants identified as significantly different in variant proportion between the cognitively impaired and less impaired individuals, between 2667 and 131,771 control individuals were sequenced. AFs from these databases are herein referenced as "expected." Observed AFs were derived for all study participants who completed neuropsychological assessment and for the less impaired subgroup. A minimum of two and a maximum of three alleles were identified at each variant site. The expected AFs were compared to each of the observed AFs by the χ^2 Yates-corrected method, where numbers of alleles in the study population permitted. Statistical comparisons between observed and expected alleles were not performed for the impaired participants because of low sample size or for the variants where the expected or observed AF was very low.

Results

Clinical Characteristics

Genome sequencing was conducted on 22 long-term medulloblastoma survivors. Of these, 18 completed the neuropsychological evaluation (Table 1). The composite neurocognitive score was calculated by averaging the standardized scores from the six cognitive

Table 1. Clinical Characteristics by Cognitive Impairment Group*,†,‡,§

	Impaired	Less Impaired	$P^{\dagger}$
N	4	14	
Age at diagnosis, years, *M* (SD)	6.3 (3.7)	8.8 (3.9)	.26
Latency, years,‡ *M* (SD)	18.3 (10.6)	12.3 (6.5)	.17
Molecular subtype			
WNT	1	1	.81
SHH	0	4	.65
Group 3	1	2	1.00
Group 4	2	7	1.00
CSI dose, Gy, *n*			
18	0	1	1.00
23.4	2	10	.81
35-36	2	3	.60
Total PF dose, Gy, *n*			
30.6	0	1	1.00
37.8	1	0	.44
54-56	3	13	.81
Chemotherapy regimen, *n*			
Average risk			
CCG 9961	1	4	1.00
ACNS 0331	1	3	1.00
CCG 9892	0	1	1.00
CHP 693	0	1	1.00
High risk			
CCG 99701	0	1	1.00
CCG 99703	0	1	1.00
ACNS 0332	0	1	1.00
Unknown	2	2	.39
Neurologic complications, *n*			
Hydrocephalus	3	11	1.00
Cerebellar mutism	1	2	1.00
Radiation necrosis	0	2	1.00
Secondary tumor§	1	1	.81
Endocrine dysfunction, *n*			
GHD	4	10	.65
Hypothyroid	4	10	.65
AI	1	0	.44
HPG	2	5	1.00

CSI, craniospinal irradiation; *PF*, posterior fossa; *GHD*, growth hormone deficiency; *AI*, adrenal insufficiency; *HPG*, hypothalamic-pituitary-gonadal dysfunction (e.g., primary ovarian insufficiency, precocious puberty).

* Impaired defined as composite cognitive score of less than −1.5 *z*.

† *P* value by Student *t* test for continuous variables or Fisher's exact test for categorical variables.

‡ Latency between treatment completion and neuropsychological evaluation.

§ Secondary tumor was meningioma in both cases.

measures. The composite scores ranged from −2.64 to 0.22 *z* (mean = −0.91, SD 0.72). Four participants had scores less than −1.5 and were categorized as impaired (range = −2.64 to −1.54). The less impaired group ranged from an average *z* score of 0.22 to −1.4.

The age range at time of medulloblastoma diagnosis was 2 to 16 years (mean 8.2, SD 3.9 years). One participant in each cognitive group was diagnosed and received radiation treatment before age 5 years. Molecular subgroup distribution ascertained by NanoString assay was 2 WNT, 4 SHH, 3 group 3, and 9 group 4, which is representative of the medulloblastoma population treated at CHOA. All participants were at least 5 years from completion of therapy at the time of study enrollment (mean time since treatment 13.7, SD 7.6 years). All participants were treated with surgery, chemotherapy, and radiation. Chemotherapy protocols included CCG 9961 (n = 5), CCG 9892 (n = 1) CCG 99703 (n = 1), CCG 99701 (n = 1), CHOP 693 (n = 1), ACNS 0331 (n = 4), ACNS 0332 (n = 1), and unknown (n = 4). All but two participants received a total posterior fossa radiation dose of between 54 and 56 Gy. Two participants (one in each cognitive group) received a reduced posterior fossa dose (30.6 and 37.8 Gy), one because of shunt infection and one for an unknown reason. Three participants had postoperative cerebellar mutism, two had radiation necrosis, and two had secondary meningiomas. Endocrine dysfunction was highly prevalent, with all but one participant diagnosed with an endocrine disorder. Growth hormone deficiency and hypothyroidism were most common, followed by hypothalamic-pituitary-gonadal dysfunction and adrenal insufficiency. Clinical characteristics were not significantly different between the cognitively impaired and less impaired groups.

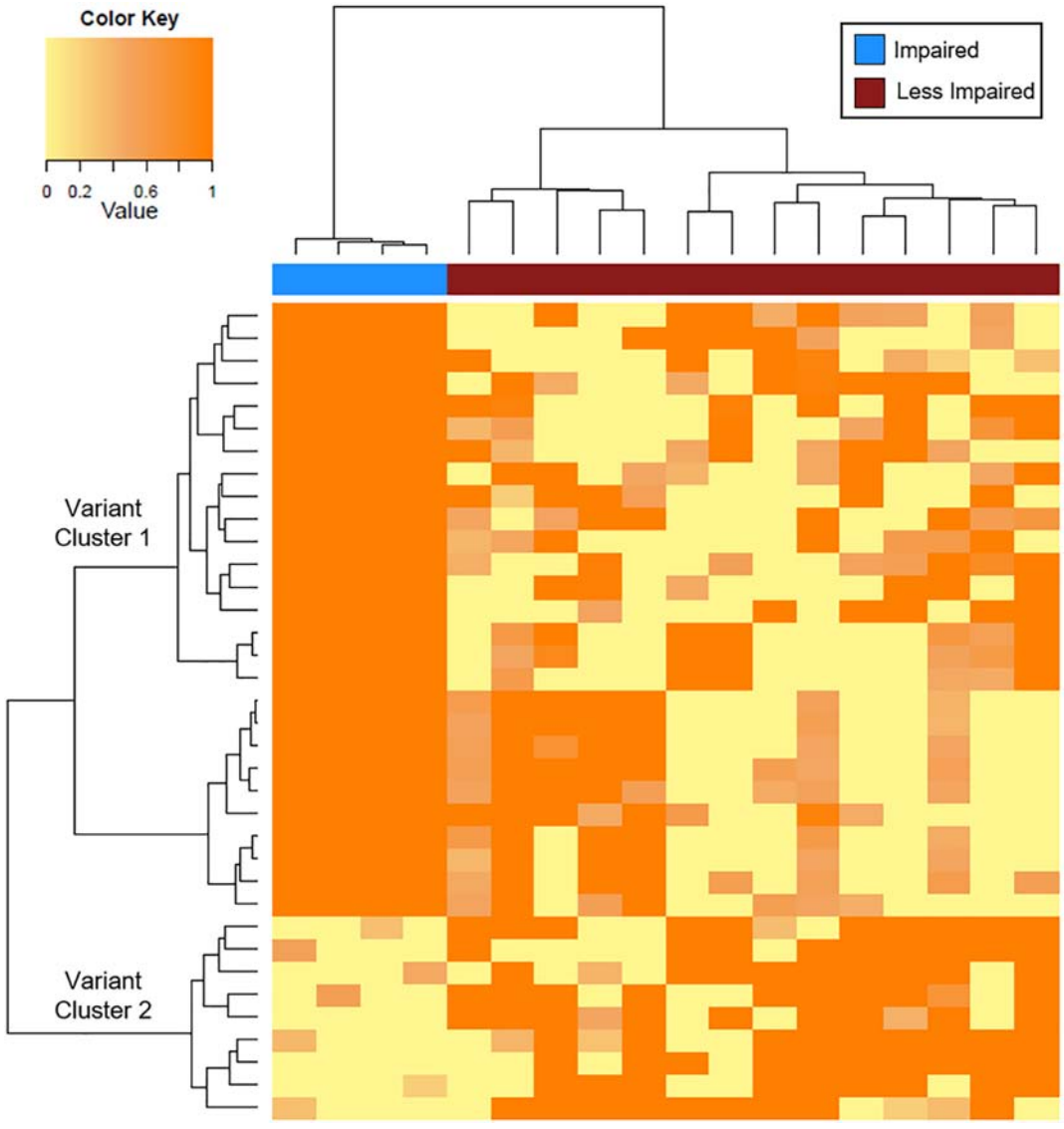

Figure 1. Disease-associated gene variants unsupervised clusters and association with cognitive impairment. Heatmap derived from hierarchical clustering analysis of relative variant expression between impaired vs. less impaired survivors. Each column is a single participant, and each row is a single nucleotide variant. The variant proportion is represented by scale on the top left, where dark orange signifies a higher proportion relative to the reference and yellow signifies a lower proportion. In this heatmap, cognitively impaired survivors exhibit a higher proportion of gene variants in cluster 1 and a lower proportion of gene variants in cluster 2 relative to less-impaired survivors.

Variant Cluster Analysis of Disease-Associated Genes

A total of 1,172,762 disease-associated gene variants were identified using the cosmic 81 and Clinvar 2017 databases. Of these, 6540 topmost variable variants were identified, 36 of which were found to be significantly different in prevalence between the impaired and less impaired survivors using a FDR-adjusted *P* value of .05 (Figure 1). Among the significant variants, 2 were exonic and 34 were located in noncoding regions: 10 intronic, 1 untranslated region 3, and 23 intergenic (Table 2).

Of additional interest was whether the variants included in the cluster analysis were found in or near genes that have been previously reported to be associated with neurocognitive function. Thirty-three genes of interest were identified post hoc from the literature, including genes associated with cognitive outcome in survivors of brain tumors *(APOE4, BDNF, COMT, IRS1, ERCC4, ABCC1, IL16, PPARD, NOS1, POLE, MSR1, SLC22, GSTT1, GSTM1, SOD2,* and *DTNBP1*) [16–20,49], leukemia (*MS, MTHFR, GSTP1, MAOA, NOS3, SLCO2A1, HFE, TSER,* and *CBS*) [50–52], and traumatic brain injury (*GAD1, ADORA1, APOE, ACE, ANKK1, WWC1, DBH,* and *GRIN2A*) [53]. Of these genes, nine contained variants that met filtering criteria for inclusion in the unsupervised hierarchical agglomerative cluster analysis (Supplemental Table 1). None of the variants were significantly different in variant proportion between the impaired and less impaired groups, although variants near two genes (ACE and ERCC4) had fold-change values of less than −4.0 and FDR-adjusted *P* values < .10.

Variant Cluster Analysis of DNA Repair Genes

A total of 158,754 variants were identified within the 409 genes associated with DNA repair pathways. Within the 1583 topmost variable gene variants, 242 variants among 21 genes were identified as core gene variants. Additionally, 17 core samples were identified among the 22 patient samples. Unsupervised consensus clustering using the two identified patient groups with distinct variant profiles is depicted in the heatmap in Figure 2. Fisher's exact tests (results not shown) were performed to assess whether there were differences between the patient groups in prevalence of endocrine disorders, sleep impairment, cerebellar mutism, radiation necrosis, or secondary tumors. Similarly, Student *t* tests (results not shown) were performed to assess differences in age at diagnosis, radiation dose, and cognitive

Table 2. Frequency of Mutated Samples Among Cognitively Impaired and Less Impaired Survivors

SNP	Gene/Flanking Gene	Allele	Region	Hom	Het	NC	Hom	Het	NC
				Impaired			Less Impaired		
				n = 4			*n* = 14		
Coding DNA									
rs227368	MANBA	C/T	EX	4	0	0	4	4	6
rs3740199	ADAM12	C/G	EX	4	0	0	3	3	8
Noncoding DNA									
rs12723918	LINC01221, NR5A2	C/G	IG	0	1	3	9	2	3
rs1509038	LINC01492, LOC101928523	C/T	IG	0	1	3	9	2	3
rs9347870	QKI, MEAT6	T/C	IG	0	1	3	10	1	3
rs2662780	LINC01492, LOC101928523	C/A	IG	0	1	3	9	1	4
rs364288	LINC01492, LOC101928523	G/C	IG	0	1	3	9	0	5
rs372046	LINC01492, LOC101928523	G/T	IG	4	0	0	4	4	6
rs378466	LINC01492, LOC101928523	T/C	IG	4	0	0	5	3	6
rs418119	LINC01492, LOC101928523	A/G	IG	4	0	0	4	4	6
rs13161948	FLT4, OR2Y1	C/T	IG	4	0	0	4	4	6
rs2507304	ANKRD20A3, MIR4477A	A/C	IG	4	0	0	6	2	6
rs400549	LINC01492, LOC101928523	G/A	IG	4	0	0	3	5	6
rs412741	LINC01492, LOC101928523	A/G	IG	4	0	0	3	5	6
rs419472	LINC01492, LOC101928523	A/T	IG	4	0	0	4	4	6
rs4585689	PODXL, LOC101928782	G/T	IG	4	0	0	4	3	7
rs7861436	LOC103908605, FAM27C	T/C	IG	4	0	0	3	4	7
rs9273206	HLA-DQA1, HLA-DQB1	T/C	IG	4	0	0	4	3	7
rs10148510	LOC101927620, MIR5580	G/C	IG	4	0	0	7	0	7
rs10005153	CLNK, MIR572	T/G	IG	4	0	0	2	5	7
rs10279849	PMS2P9, CCDC146	A/C	IG	4	0	0	4	3	7
rs13236623	ARL4A, ETV1	T/A	IG	4	0	0	4	3	7
rs2803191	LCAL1, LCA5	T/C	IG	4	0	0	5	1	8
rs659494	FAM35A, NUTM2A	T/A	IG	4	0	0	4	2	8
rs7938520	ALX4, CD82	C/A	IG	4	0	0	3	3	8
rs1482089	ARHGAP24	C/T	IN	0	0	4	9	2	3
rs2062100	ARHGAP24	T/A	IN	0	1	3	10	0	4
rs2806429	NFIA	C/T	IN	0	0	4	10	0	4
rs370593786	CFAP54	G/A	IN	0	1	3	9	0	5
rs4693720	ARHGAP24	C/T	IN	4	0	0	5	2	7
rs10743823	CLECL1	C/T	IN	4	0	0	4	3	7
rs4908277	COL11A1	C/T	IN	4	0	0	2	5	7
rs10422502	ZNF71	A/C	IN	4	0	0	4	3	7
rs2326797	LAMA2	A/G	IN	4	0	0	4	2	8
rs138306877	LINC00836	G/A	IN	4	0	0	3	3	8
rs30886	PDE6A	C/T	UTR3	4	0	0	4	4	6

For each variant, homozygosity was defined as AF ≥ 0.70, heterozygosity as AF ≥ 0.20, and no call as AF = 0.
Hom, homozygous; *Het*, heterozygous; *NC*, no call; *IG*, intergenic; *IN*, intronic; *EX*, exonic; *UTR*, untranslated region.

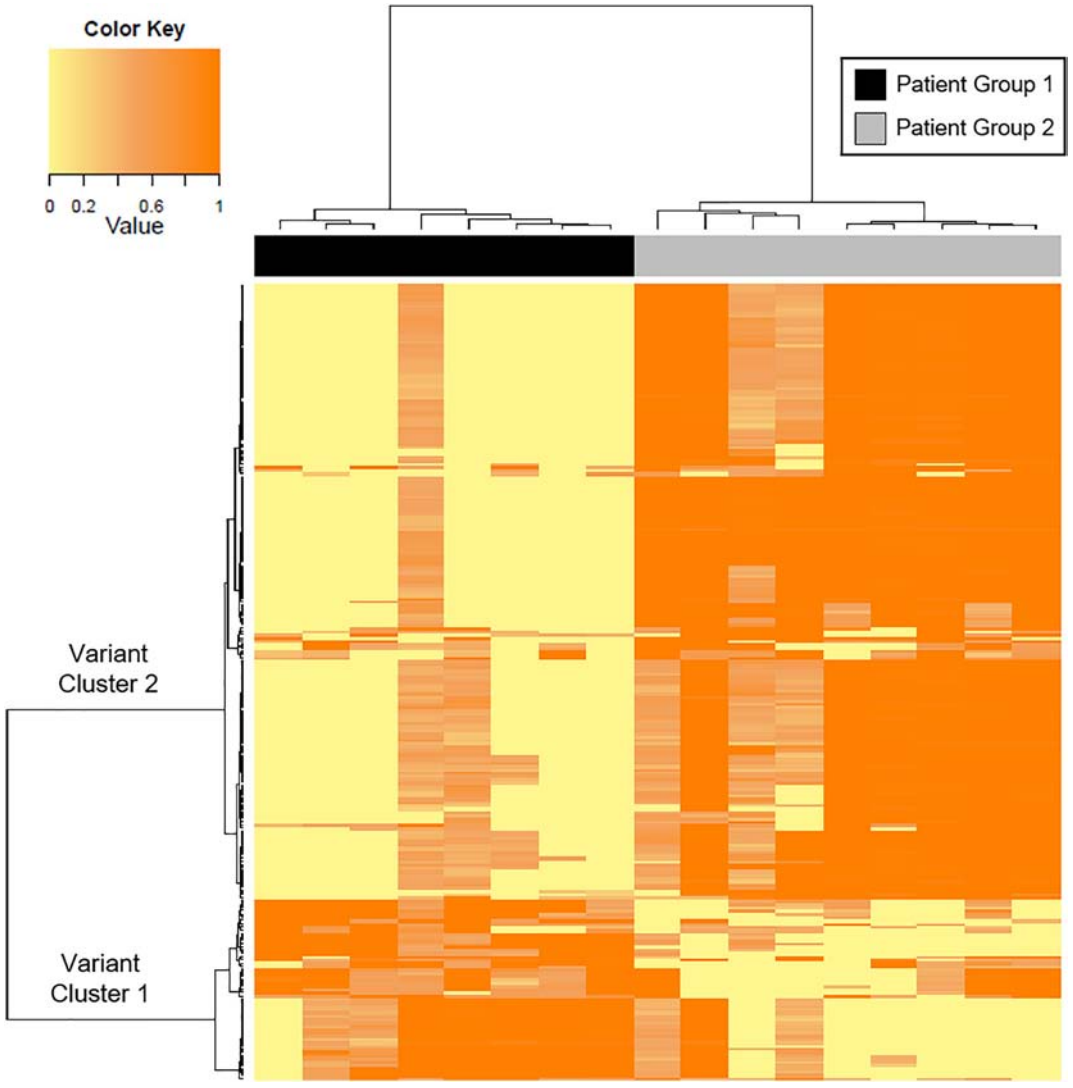

Figure 2. Heatmap derived from unsupervised cluster analysis of DNA repair gene variants.Each column is a single participant, and each row is a single nucleotide variant. The variant proportion is represented by scale on the top left, where dark orange signifies a higher proportion relative to the reference and yellow signifies a lower proportion. In this heatmap, gene variants in cluster 1 are overrepresented in patient group 1 and underrepresented in patient group 2. Likewise, variants in cluster 2 are overrepresented in patient group 2.

assessment scores. There were no significant differences between patient groups along any of these parameters.

Allelic Frequency Analysis

For each of the 36 disease-associated variants identified in the cluster analysis, observed AFs in the study sample were compared to expected AFs derived from the general population (Supplemental Table 2). Seven variants, all of which were in noncoding DNA, had significant differences between the expected and observed AF. Of these, four variants (rs2326797, rs13236623, rs138306877, and rs9347870) were more prevalent in the study sample than the general population, and three variants (rs7861436, rs659494, and rs9273206) were less prevalent. For rs7861436 and rs659494, this difference was driven by low AFs in the impaired participants, while for rs9273206, it was driven by low AF in the less impaired participants. Additionally, within the less impaired subgroup, rs364288 had a significantly lower AF and rs2507304 had a significantly higher AF compared to the general population.

Discussion

In the present study, we conducted whole-genome sequencing of host blood in a cohort of long-term medulloblastoma survivors to identify genomic variants associated with neurocognitive morbidity. We found that cognitively impaired survivors did not differ from less impaired survivors in terms of exposure to chemoradiation or age at diagnosis but did have differences in host genome profile. Unsupervised analysis of all genome-wide disease-associated variants demonstrated that the cognitive groups have distinct variant profiles. The survivors also segregated into a separate set of two groups with distinct DNA repair gene profiles by unsupervised consensus clustering. These DNA repair profiles were not associated with cognitive outcome, suggesting that variation in genes corresponding to a single functional group may be insufficient to predict long-term cognitive outcomes alone.

In recent years, efforts have been made to deescalate radiotherapy with a goal of reducing long-term impairment in medulloblastoma patients. However, deescalation may not be a viable option in most

cases. The most recent Children's Oncology Group clinical study for average-risk medulloblastoma (ACNS 0331) attempted to reduce the craniospinal dose from 24 to 18 Gy but found that patients receiving the lower radiation dose had an unacceptable rate of tumor relapse and overall survival [54]. Findings from this phase III randomized trial indicate that late toxicity from craniospinal irradiation will continue to be a major clinical problem for the majority of future medulloblastoma survivors. As a result, identifying clinically predictive genetic profiles to provide individualized prognostic information is an important area of research.

Prior genetic studies in survivors of childhood CNS tumors have assessed the effect of specific candidate SNPs on neurocognitive outcome, identified a priori or by pathway-oriented methods [16,17]. In contrast, the approach taken by the present study is novel in three ways. First, it incorporated host whole-genome sequencing data, allowing for detection of clinically meaningful but rare variants in both coding (genic) and noncoding (intergenic and intronic) DNA regions. Second, it employed a hierarchical cluster analysis of variant proportion data, a technique that identifies coherent subpopulations within an immense amount of sequencing data and has previously been applied in other populations but not in cancer survivors [25,55]. Third, the allelic frequencies of the identified variants in the study sample were compared to those of the general population using an aggregate of three large, well-validated human genome sequencing databases. This allowed for an assessment of whether the identified variants are specific to individuals with medulloblastoma and are therefore more likely to be clinically meaningful.

Notably, 94% of the variants identified in the disease-associated analysis are located in the noncoding DNA regions. Noncoding DNA makes up 98.8% of the entire human genome and has been previously dismissed as "junk DNA" with no function [56]. However, more recent studies indicate that noncoding DNA is responsible for gene regulation and facilitating complex temporal and spatial gene expression through combinatorial interactions with other gene regulatory elements, with the major regions involved in gene regulation being the 5′ and 3′ untranslated regions and introns [57]. These potentially important regulatory DNA sequences would be missed in the absence of whole genome analysis.

The precise function of the noncoding DNA regions identified in this cluster analysis is not known [58]. However, allelic frequency analysis demonstrated several highly statistically significant differences between the study population and the general population, as well as specific differences between the less impaired and impaired participants. Taken together, these differences suggest that the identified variants may be involved both in tumor development in patients without a known cancer predisposition syndrome and in vulnerability to neurocognitive radiotoxicity. They also provide a key starting point for mechanistic studies employing combinatorial *in silico* and experimental methods to examine cause-and-effect relationships between noncoding DNA variation and patient outcomes.

This study has several strengths. First, the neurocognitive data were obtained by standardized performance measures rather than by self-report and examined using a composite neuropsychological score based on key components of empirically and theoretically derived neurodevelopmental models of long-term outcomes of childhood brain tumors [4–6]. Second, the participants all had medulloblastoma and had a distribution of molecular subgroups that is representative of the larger medulloblastoma population at our institution, thus reducing the likelihood of potential confounds that may occur when examining cognition in survivors with diverse tumor sites and characteristics. Participants in the cognitively impaired and less impaired groups were also found to be similar in terms of clinical features and comorbidities known to influence cognition, including age at diagnosis, radiation dose, and cerebellar mutism [59]. In sum, the homogeneity in tumor characteristics and even distribution of plausible confounds between cognitive groups made differences in long-term cognitive outcome more easily attributable to differences in host response to treatment rather than to differences in clinical presentation, course, or the treatment itself.

The current study must be considered within the context of a limited sample size and relatively small number of impaired survivors relative to less impaired survivors, which restricted the ability to detect statistically significant differences in variant proportion between the cognitively impaired and less impaired groups among genes previously identified to be involved in neurocognitive outcome or between DNA repair gene profiles. Variants within these genes could be identified as associated with cognitive impairment in a larger sample.

The current study contributes to the identification of genetic influences on outcome in medulloblastoma by employing complementary genome-wide and pathway-specific approaches coupled with the innovative technique of hierarchical clustering. The robust segregation of our cohort into genetically distinct clusters suggests that these methods represent a previously untapped avenue for identifying genetic risk factors for cognitive impairment many years following the complex chemoradiation treatment. Likewise, similar methods could be used to identify SNPs associated with resiliency or only mild cognitive difficulties following the same treatment. Future multisite studies using a longitudinal case-control genetic association method (genome-wide and candidate gene) with functional validation in independent cohorts are needed to confirm these findings and translate them to clinical practice. This study establishes an evidence base to justify such larger scale investigations and provides a blueprint for independent replication.

Acknowledgements

We are indebted to the individuals and their families who gave willingly of their time to make this research possible. We thank the students in Dr. King's Developmental Neuropsychology Across the Lifespan Laboratory who assisted with neuropsychological data acquisition, scoring, and management. We appreciate the assistance of Patty Church, RN, with participant recruitment and blood draw. Research reported in this publication was supported in part by the Emory Integrated Genomics Core Shared Resource and the Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and National Institutes of Health/National Cancer Institute under award number P30CA138292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tranon.2019.03.004.

References

[1] Weil AG, Wang AC, Westwick HJ, Ibrahim GM, Ariani RT, Crevier L, Perreault S, Davidson T, hong Tseng C, and Fallah A (2017). Survival in pediatric medulloblastoma: a population-based observational study to improve prognostication. *J Neurooncol* **132**, 99–107. http://dx.doi.org/10.1007/s11060-016-2341-4.

[2] Coluccia D, Figuereido C, Isik S, Smith C, and Rutka JT (2016). Medulloblastoma: tumor biology and relevance to treatment and prognosis paradigm. *Curr Neurol Neurosci Rep* **16**43. http://dx.doi.org/10.1007/s11910-016-0644-7.

[3] Palmer SL, Reddick WE, and Gajjar A (2007). Understanding the cognitive impact on children who are treated for medulloblastoma. *J Pediatr Psychol* **32**, 1040–1049. http://dx.doi.org/10.1093/jpepsy/jsl056.

[4] Palmer SL (2008). Neurodevelopmental impact on children treated for medulloblastoma: a review and proposed conceptual model. *Dev Disabil Res Rev* **14**, 203–210. http://dx.doi.org/10.1002/ddrr.32.

[5] Wolfe KR, Madan-Swain A, and Kana RK (2012). Executive dysfunction in pediatric posterior fossa tumor survivors: a systematic literature review of neurocognitive deficits and interventions. *Dev Neuropsychol* **37**, 153–175. http://dx.doi.org/10.1080/87565641.2011.632462.

[6] King TZ, Ailion AS, Fox ME, and Hufstetler SM (2019). Neurodevelopmental model of long-term outcomes of adult survivors of childhood brain tumors. *Child Neuropsychol* **25**, 1–21. http://dx.doi.org/10.1080/09297049.2017.1380178.

[7] Wefel JS, Noll KR, and Scheurer ME (2016). Neurocognitive functioning and genetic variation in patients with primary brain tumours. *Lancet Oncol* **17**, e97–e108. http://dx.doi.org/10.1016/S1470-2045(15)00380-0.

[8] Taylor MD, Northcott PA, Korshunov A, Remke M, Cho Y-J, Clifford SC, Eberhart CG, Parsons DW, Rutkowski S, and Gajjar A, et al (2012). Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol* **123**, 465–472. http://dx.doi.org/10.1007/s00401-011-0922-z.

[9] Moxon-Emre I, Taylor MD, Bouffet E, Hardy K, Campen CJ, Malkin D, Hawkins C, Laperriere N, Ramaswamy V, and Bartels U, et al (2016). Intellectual outcome in molecular subgroups of medulloblastoma. *J Clin Oncol* . http://dx.doi.org/10.1200/JCO.2016.66.9077.

[10] Damaraju S, Murray D, Dufour J, Carandang D, Myrehaug S, Fallone G, Field C, Greiner R, Hanson J, and Cass CE, et al (2006). Association of DNA repair and steroid metabolism gene polymorphisms with clinical late toxicity in patients treated with conformal radiotherapy for prostate cancer. *Clin Cancer Res* **12**, 2545–2554. http://dx.doi.org/10.1158/1078-0432.CCR-05-2703.

[11] Lopez Guerra JL, Wei Q, Yuan X, Gomez D, Liu Z, Zhuang Y, Yin M, Li M, Wang L-E, and Cox JD, et al (2011). Functional promoter rs2868371 variant of HSPB1 associates with radiation-induced esophageal toxicity in patients with non-small-cell lung cancer treated with radio(chemo)therapy. *Radiother Oncol* **101**, 271–277. http://dx.doi.org/10.1016/j.radonc.2011.08.039.

[12] Guo C-X, Wang J, Huang L-H, Li J-G, and Chen X (2016). Impact of single-nucleotide polymorphisms on radiation pneumonitis in cancer patients. *Mol Clin Oncol* **4**, 3–10. http://dx.doi.org/10.3892/mco.2015.666.

[13] Alsbeih G, El-Sebaie M, Al-Harbi N, Al-Hadyan K, Shoukri M, and Al-Rajhi N (2013). SNPs in genes implicated in radiation response are associated with radiotoxicity and evoke roles as predictive and prognostic biomarkers. *Radiat Oncol* **8**, 125. http://dx.doi.org/10.1186/1748-717X-8-125.

[14] Borchiellini D, Etienne-Grimaldi M-C, Thariat J, and Milano G (2012). The impact of pharmacogenetics on radiation therapy outcome in cancer patients. A focus on DNA damage response genes. *Cancer Treat Rev* **38**, 737–759. http://dx.doi.org/10.1016/j.ctrv.2012.02.004.

[15] Oyefiade A, Erdman L, Goldenberg A, Malkin D, Bouffet E, Taylor MD, Ramaswamy V, Scantlebury N, Law N, and Mabbott DJ (2019). PPAR and GST polymorphisms may predict changes in intellectual functioning in medulloblastoma survivors. *J Neurooncol* **142**, 39–48. http://dx.doi.org/10.1007/s11060-018-03083-x.

[16] Barahmani N, Carpentieri S, Li X-N, Wang T, Cao Y, Howe L, Kilburn L, Chintagumpala M, Lau C, and Okcu MF (2009). Glutathione S-transferase M1 and T1 polymorphisms may predict adverse effects after therapy in children with medulloblastoma. *Neuro Oncol* **11**, 292–300. http://dx.doi.org/10.1215/15228517-2008-089.

[17] Brackett J, Krull KR, Scheurer ME, Liu W, Srivastava DK, Stovall M, Merchant TE, Packer RJ, Robison LL, and Okcu MF (2012). Antioxidant enzyme polymorphisms and neuropsychological outcomes in medulloblastoma survivors: a report from the Childhood Cancer Survivor Study. *Neuro Oncol* **14**, 1018–1025. http://dx.doi.org/10.1093/neuonc/nos123.

[18] Correa DD, Satagopan J, Cheung K, Arora AK, Kryza-Lacombe M, Xu Y, Karimi S, Lyo J, Deangelis LM, and Orlow I (2016). COMT, BDNF, and DTNBP1 polymorphisms and cognitive functions in patients with brain tumors. *Neuro Oncol* **18**, 1425–1433. http://dx.doi.org/10.1093/neuonc/now057.

[19] Correa DD, Satagopan J, Baser RE, Cheung K, Richards E, Lin M, Karimi S, Lyo J, DeAngelis LM, and Orlow I (2014). APOE polymorphisms and cognitive functions in patients with brain tumors. *Neurology* **83**, 320–327. http://dx.doi.org/10.1212/wnl.0000000000000617.

[20] Liu Y, Zhou R, Sulman EP, Scheurer ME, Boehling N, Armstrong GN, Tsavachidis S, Liang FW, Etzel CJ, and Conrad CA, et al (2015). Genetic modulation of neurocognitive function in glioma patients. *Clin Cancer Res* **21**, 3340–3346. http://dx.doi.org/10.1158/1078-0432.CCR-15-0168.

[21] Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, Heisler LE, Beck TA, Simpson JT, and Tonon L, et al (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun* **6**10001. http://dx.doi.org/10.1038/ncomms10001.

[22] Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger N, Remke M, Shih D, and Northcott PA, et al (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71. http://dx.doi.org/10.1016/j.cell.2011.12.013.

[23] Rupji M, Dwivedi B, and Kowalski J (2019). NOJAH: not just another heatmap for genome-wide cluster analysis. *BioRxiv* **14**e0204542. http://dx.doi.org/10.1371/journal.pone.0204542.

[24] Tusher VG, Tibshirani R, and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116–5121. http://dx.doi.org/10.1073/pnas.091062498.

[25] Yu C, Baune BT, Licinio J, and Wong M-L (2017). Single-nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data. *J Hum Genet* **62**, 1–4. http://dx.doi.org/10.1038/jhg.2017.2.

[26] Kulkarni MM (2011). Digital multiplexed gene expression analysis using the NanoString nCounter system. *Curr Protoc Mol Biol* . http://dx.doi.org/10.1002/0471142727.mb25b10s94 [Chapter 25, Unit25B.10].

[27] Northcott PA, Shih DJH, Remke M, Cho Y-J, Kool M, Hawkins C, Eberhart CG, Dubuc A, Guettouche T, and Cardentey Y, et al (2012). Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. *Acta Neuropathol* **123**, 615–626. http://dx.doi.org/10.1007/s00401-011-0899-7.

[28] Smith A (1982). Symbol digits modalities test. Los Angeles, CA: West. Psychol. Serv.; 1982 .

[29] Stuss DT, Stethem LL, and Pelchat G (1988). Three tests of attention and rapid information processing: an extension. *Clin Neuropsychol* **2**, 246–250. http://dx.doi.org/10.1080/13854048808520107.

[30] Strauss E, Sherman E, and Spreen O (2006). A compendium of neuropsychological tests: administration, norms, and commentary. . 3rd ed.Oxford University Press; 2006 .

[31] Wechsler D (1997). Wechsler memory scale manual. . 3rd ed.San Antonia, TX: The Psychological Corporation; 1997 .

[32] Wechsler D (1999). Wechsler abbreviated scale of intelligence. San Antonia, TX: The Psychological Corporation: Harcourt Brace & Company; 1999 .

[33] McGrew KS, Dailey DEH, and Schrank FA (2007). Score differences: what the user can expect and why (Woodcock-Johnson III assessment service bulletin no. 9). Rolling Meadows, IL: Riverside Publishing; 2007 .

[34] Lezak M, Howieson D, Bigler E, and Tranel D (2012). Neuropsychological assessment. . fifth ed.Oxford: Oxford University Press; 2012 .

[35] Johnston HR, Chopra P, Wingo TS, Patel V, Epstein MP, Mulle JG, Warren ST, Zwick ME, and Cutler DJ (2017). PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci* **114**, E1923-1932. http://dx.doi.org/10.1073/pnas.1618065114.

[36] Andrews S (2012). FastQC. Cambridge, UK: Babraham Bioinformatics; 2012 .

[37] Bolger AM, Lohse M, and Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

[38] Li H and Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. http://dx.doi.org/10.1093/bioinformatics/btp698.

[39] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.

[40] Koboldt DC, Zhang Q, Larson DE, Shen D, Mclellan MD, Lin L, Miller CA, Mardis ER, Ding L, and Wilson RK (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–576. http://dx.doi.org/10.1101/gr.129684.111.

[41] Wang K, Li M, and Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, 1–7. http://dx.doi.org/10.1093/nar/gkq603.

[42] Wilkerson MD and Hayes DN (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573. http://dx.doi.org/10.1093/bioinformatics/btq170.

[43] Eisen MB, Spellman PT, Brown PO, and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863–14868. http://dx.doi.org/10.1073/pnas.95.25.14863.

[44] Kohonen T (1997). Self-organizing maps. . 2nd ed.Berlin Heidelberg: Springer-Verlag; 1997 .

[45] Birney E and Soranzo N (2015). Human genomics: the end of the start for population sequencing. *Nature* **526**, 52–53. http://dx.doi.org/10.1038/526052a.

[46] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, and Cummings BB, et al (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291. http://dx.doi.org/10.1038/nature19057.

[47] Exome variant server. NHLBI GO Exome Seq. Proj. (n.d.) http://evs.gs.washington.edu/EVS, Accessed date: 22 December 2017.

[48] dbSNP: short genetic variations. Natl. Cent. Biotechnol. Inf. (n.d.) https://ncbi.nlm.nih.gov/projects/SNP/snp_tableList.cgi, Accessed date: 14 January 2018.

[49] Ellenberg L, Liu Q, Gioia G, Yasui Y, Packer RJ, Mertens A, Donaldson SS, Stovall M, Kadan-Lottick N, and Armstrong G, et al (2009). Neurocognitive status in long-term survivors of childhood CNS malignancies: a report from the Childhood Cancer Survivor Study. *Neuropsychology* **23**, 705–717. http://dx.doi.org/10.1037/a0016674.Neurocognitive.

[50] Cole PD, Finkelstein Y, Stevenson KE, Blonquist TM, Vijayanathan V, Silverman LB, Neuberg DS, Sallan SE, Robaey P, and Waber DP (2015). Polymorphisms in genes related to oxidative stress are associated with inferior cognitive function after therapy for childhood acute lymphoblastic leukemia. *J Clin Oncol* **33**, 2205–2211. http://dx.doi.org/10.1200/JCO.2014.59.0273.

[51] Krajinovic M, Robaey P, Chiasson S, Lemieux-Blanchard E, Rouillard M, Primeau M, Bournissen FG, and Moghrabi A (2005). Polymorphisms of genes controlling homocysteine levels and IQ score following the treatment for childhood ALL. *Pharmacogenomics* **6**, 293–302. http://dx.doi.org/10.1517/14622416.6.3.293.

[52] Krull KR, Bhojwani D, Conklin HM, Pei D, Cheng C, Reddick WE, Sandlund JT, and Pui CH (2013). Genetic mediators of neurocognitive outcomes in survivors of childhood acute lymphoblastic leukemia. *J Clin Oncol* **31**, 2182–2188. http://dx.doi.org/10.1200/JCO.2012.46.7944.

[53] Kurowski BG, Treble-Barna A, Pitzer AJ, Wade SL, Martin LJ, Chima RS, and Jegga A (2017). Applying systems biology methodology to identify genetic factors possibly associated with recovery after traumatic brain injury. *J Neurotrauma* . http://dx.doi.org/10.1089/neu.2016.4856.

[54] Michalski J, Vezina G, Burger P, Gajjar A, Pollack I, Merchant T, Fitzgerald T, Booth T, Tarbell N, and Shieh I, et al (2016). Phase III trial of involved field radiotherapy (IFRT) and low dose craniospinal irradiation (LD-CSI) with chemotherapy in average risk medulloblastoma: a report from the Children's Oncology Group. *Neuro Oncol* **18**, iii122.

[55] Monti S, Tamayo P, Mesirov J, and Golub T (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* **52**, 91–118. http://dx.doi.org/10.1023/A:1023949509487.

[56] Taft RJ, Pheasant M, and Mattick JS (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**, 288–299. http://dx.doi.org/10.1002/bies.20544.

[57] Kim Y-J, Lee J, and Han K (2012). Transposable elements: no more "junk DNA". *Genomics Inform* **10**, 226–233. http://dx.doi.org/10.5808/GI.2012.10.4.226.

[58] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. http://dx.doi.org/10.1038/nature11247.

[59] Rizzo D, Peruzzi L, Attinà G, Triarico S, and Ruggiero A (2017). Neurocognitive outcomes in pediatric brain tumors survivors. *Med One* **2**, 1–5. http://dx.doi.org/10.20900/mo.20170015.