OPEN ACCESS Freely available online

PLoS GENETICS

Phenotype Restricted Genome-Wide Association Study Using a Gene-Centric Approach Identifies Three Low-Risk Neuroblastoma Susceptibility Loci

Lễ B. Nguyễn[1,2,3], Sharon J. Diskin[1], Mario Capasso[4,5], Kai Wang[6], Maura A. Diamond[1], Joseph Glessner[6], Cecilia Kim[6], Edward F. Attiyeh[1,7], Yael P. Mosse[1,7], Kristina Cole[1,7], Achille Iolascon[4,5], Marcella Devoto[8], Hakon Hakonarson[6,7,8], Hongzhe K. Li[3], John M. Maris[1,2,5,9]*

1 Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **2** Genomics and Computational Biology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America, **3** Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America, **4** Department of Biochemistry and Medical Biotechnology, University of Naples Federico II, Naples, Italy, **5** CEINGE Biotecnologie Avanzate, Naples, Italy, **6** The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **7** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America, **8** Division of Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **9** Abramson Family Cancer Research Institute, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America

Abstract

Neuroblastoma is a malignant neoplasm of the developing sympathetic nervous system that is notable for its phenotypic diversity. High-risk patients typically have widely disseminated disease at diagnosis and a poor survival probability, but low-risk patients frequently have localized tumors that are almost always cured with little or no chemotherapy. Our genome-wide association study (GWAS) has identified common variants within *FLJ22536*, *BARD1*, and *LMO1* as significantly associated with neuroblastoma and more robustly associated with high-risk disease. Here we show that a GWAS focused on low-risk cases identified SNPs within *DUSP12* at 1q23.3 ($P = 2.07 \times 10^{-6}$), *DDX4* and *IL31RA* both at 5q11.2 ($P = 2.94 \times 10^{-6}$ and 6.54×10^{-7} respectively), and *HSD17B12* at 11p11.2 ($P = 4.20 \times 10^{-7}$) as being associated with the less aggressive form of the disease. These data demonstrate the importance of robust phenotypic data in GWAS analyses and identify additional susceptibility variants for neuroblastoma.

Citation: Nguyễn LB, Diskin SJ, Capasso M, Wang K, Diamond MA, et al. (2011) Phenotype Restricted Genome-Wide Association Study Using a Gene-Centric Approach Identifies Three Low-Risk Neuroblastoma Susceptibility Loci. PLoS Genet 7(3): e1002026. doi:10.1371/journal.pgen.1002026

Editor: Nicholas J. Schork, University of California San Diego and The Scripps Research Institute, United States of America

Received November 24, 2010; **Accepted** January 31, 2011; **Published** March 17, 2011

Funding: This work was supported in part by NIH Grants R01-CA124709 (JMM), NIH grants ES009911 and CA127334 (HKL), U10-CA98543 and U10-CA98413 to the Children's Oncology Group and the Giulio D'Angio Endowed Chair (JMM), the Alex's Lemonade Stand Foundation (JMM), the Rally Foundation (JMM), Andrew's Army Foundation (JMM), the Abramson Family Cancer Research Institute (JMM), the Associazione Italiana per la Ricerca sul Cancro and the Associazione Italiana per la Lotta al Neuroblastoma (AI), and an Institutional Development Award to the Center for Applied Genomics from the Children's Hospital of Philadelphia (HH). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Maris@email.chop.edu

Introduction

Neuroblastoma is a pediatric cancer of the developing sympathetic nervous system and is the most common childhood solid tumor outside the central nervous system [1,2]. Its broad spectrum of clinical behaviors is the basis for ways to categorize neuroblastoma into three risk groups: high-risk, intermediate-risk and low-risk. The approximately 50% of cases classified as high-risk show an aggressive clinical course with widespread metastases to bone and bone marrow present at diagnosis [3]. Despite intensive multimodal therapy, the long-term survival rate is less than 50% for children with high-risk neuroblastoma [1]. On the other hand, substantial portions of neuroblastoma patients show favorable clinical features including spontaneous regression of disease, and are classified as low-risk. Low-risk neuroblastoma patients have a greater than 95% survival probability with minimal, if any, chemotherapy [1]. Intermediate-risk cases are the most heterogeneous, and also the smallest subset using current definitions, comprising about 15% of all neuroblastoma patients.

We have recently performed a neuroblastoma GWAS by applying single marker analyses and identified three distinct loci significantly associated with neuroblastoma. Each of these SNP associations was within genes and particularly enriched in the high-risk group of patients: *FLJ22536* at chromosome 6p22 [4], *BRCA1* associated *RING* domain 1 (*BARD1*) at 2q35 [5], and *LIM* domain only 1 (*LMO1*) at 11p15 [6]. A similar approach was utilized to identify a common copy number variation (CNV) at chromosome 1q21 within the *NBPF23* gene that is also robustly associated with neuroblastoma [7].

In this study, we report that by adapting statistical methods to analyze genotype data, we discovered, and successfully replicated, three distinct loci as associated with the low-risk group of neuroblastoma. Furthermore, we report several gene sets as enriched in all risk groups of neuroblastoma.

Author Summary

Neuroblastoma is the most common solid tumor outside the central nervous system and is accountable for 10% of the mortality rate of all children's cancers. It has distinctive clinical behaviors and is categorized into different risk groups: high-risk, intermediate-risk, and low-risk. Genome-wide association studies have reported a number of genetic variations predisposing to high-risk neuroblastoma. This study focuses on the low-risk neuroblastoma group and identifies four novel genes (*DUSP12*, *DDX4*, *IL31RA*, and *HSD17B12*) at three distinct genomic positions that harbor disease-causing variants. This study also reports several gene sets that are enriched in overall neuroblastoma as well as in both high-risk and low-risk groups. Also of importance is that this study adopts a new computational method that identifies genes, instead of only one single nucleotide polymorphism, as disease-causing variants. Shown to have superior power of detection genome-wide association signals for neuroblastoma, the methodology presented in this study has great potential applications in case-control association studies in other diseases.

Results

Gene-centric method identifies three low-risk neuroblastoma susceptibility loci

As we are interested in studying disease causal variants that have a high likelihood of impacting protein-encoding genes, we developed a gene-centric computational method to test for association signals at the gene level. This method adapted the global test [8], developed to test association of genes groups using microarray expression data, to analyze our genotype data. Our method computes an aggregated test score based on genotype data of all SNPs on a region extending 10 kilo-bases upstream and downstream of a gene. We applied this method using a discovery set containing 1627 cases and 2575 control subjects, aimed at analyzing association to 15,885 genes annotated in the UCSC Genome Browser [9] (Materials and Methods). The replication dataset contained 398 cases and 1507 control subjects. Our methodology correctly identified the three significant genes already reported (*FLJ22536*, *BARD1* and *LMO1*). In addition, our method also identified the dual-specificity phosphatase 12 gene (*DUSP12*) at chromosome band 1q23.3 (Table 1) as significantly associated with neuroblastoma.

We next sought to determine if association signals discovered in our unbiased scan would be further enriched, or diminished, when we restricted our analyses to the divergent phenotypes of low-risk or high-risk neuroblastoma. We first analyzed a subset of 678 high-risk neuroblastoma cases from the original discovery case series, again matched to 2575 control subjects. This analysis reconfirmed that all three previously reported signals were truly associated with high-risk neuroblastoma (Table 1), but *DUSP12* did not show a strong association signal in the high-risk disease case series ($P = 4.56\times10^{-04}$). In parallel, we analyzed a subset of 574 low-risk cases and 1722 matched control subjects and a replication set of 124 cases and 496 matched control subjects (Materials and Methods). This analysis confirmed *DUSP12* and three novel genes as associated with low-risk neuroblastoma: DEAD (Asp-Glu-Ala-Asp) box polypeptide 4 isoform (*DDX4*) and interleukin-31 receptor A precursor (*IL31RA*) both at the same locus within chromosome band 5q11.2, and hydroxysteroid (17-beta) dehydrogenase 12 (*HSD17B12*) at chromosome band 11p11.2 (Table 1). All signals had significant discovery p-values using Bonferroni correction over 15,885 genes ($P<3.15\times10^{-6}$), and replication p-values less than 0.05.

Our gene-centric method was able to detect *DUSP12* and *HSD17B12*, the only two genes containing at least one SNP that passed the Bonferroni correction in single marker (SNP) analysis of the low-risk neuroblastoma (Figure 1 and Figure 2) using association testing as implemented in PLINK [10]. The fact that our gene-centric results were compatible with the single marker results supported the effectiveness of our method. In addition, we were able to detect two gene-level association signals located at a single locus for *DDX4* and *IL31RA* even though these genes did not contain any significant SNPs in the single marker analysis (Figure 1 and Table 2). These genes, however, contained several SNPs with moderate signals (Figure 2), and our gene-centric method was able to combine these effects and detected the overall significance of these two gene's signals. Being independently replicated in our study ($P = 7.20\times10^{-3}$ and 1.48×10^{-2} respectively), these two signals offered indications that our gene-centric method was more

Table 1. Summary of gene-centric analysis results for different phenotypic neuroblastomas.

Gene Symbols	Chromosome	Start- Stop	N° of SNP	Overall Discovery P-values	Overall Replication P-values	High-risk Discovery P-values	High-risk Replication P-values	Low-risk Discovery P-values	Low-risk Replication P-values
BARD1	2q35	215301519-215382673	28	**9.92×10^{-11}**	**2.19×10^{-03}**	**$<1.00\times10^{-30}$**	**3.00×10^{-03}**	1.62×10^{-01}	6.49×10^{-01}
FLJ44180	6p22.3	22243164-22255401	8	**$<1.00\times10^{-30}$**	**1.94×10^{-04}**	**$<1.00\times10^{-30}$**	**5.45×10^{-03}**	1.40×10^{-03}	3.66×10^{-02}
LMO1	11p15.4	8202432-8246758	29	**1.80×10^{-07}**	**1.59×10^{-03}**	**2.51×10^{-08}**	**2.82×10^{-02}**	1.40×10^{-02}	5.46×10^{-02}
DUSP12	1q23.3	159986204-159993576	4	**1.16×10^{-07}**	**3.30×10^{-02}**	4.56×10^{-04}	1.97×10^{-01}	**2.07×10^{-06}**	**2.92×10^{-02}**
DDX4	5q11.2	55070534-55148362	11	2.81×10^{-05}	3.11×10^{-03}	2.95×10^{-02}	2.67×10^{-01}	**2.94×10^{-06}**	**7.20×10^{-03}**
IL31RA	5q11.2	55183090-55254434	18	2.75×10^{-04}	5.74×10^{-02}	2.88×10^{-01}	7.28×10^{-01}	**6.54×10^{-07}**	**1.48×10^{-02}**
HSD17B12	11p11.2	43658718-43834745	22	1.29×10^{-04}	3.05×10^{-02}	6.82×10^{-02}	3.82×10^{-01}	**4.20×10^{-07}**	**5.37×10^{-02}**

Bold-faced p-values indicate significant association signals with Bonferroni correction over 15,885 genes.

doi:10.1371/journal.pgen.1002026.t001

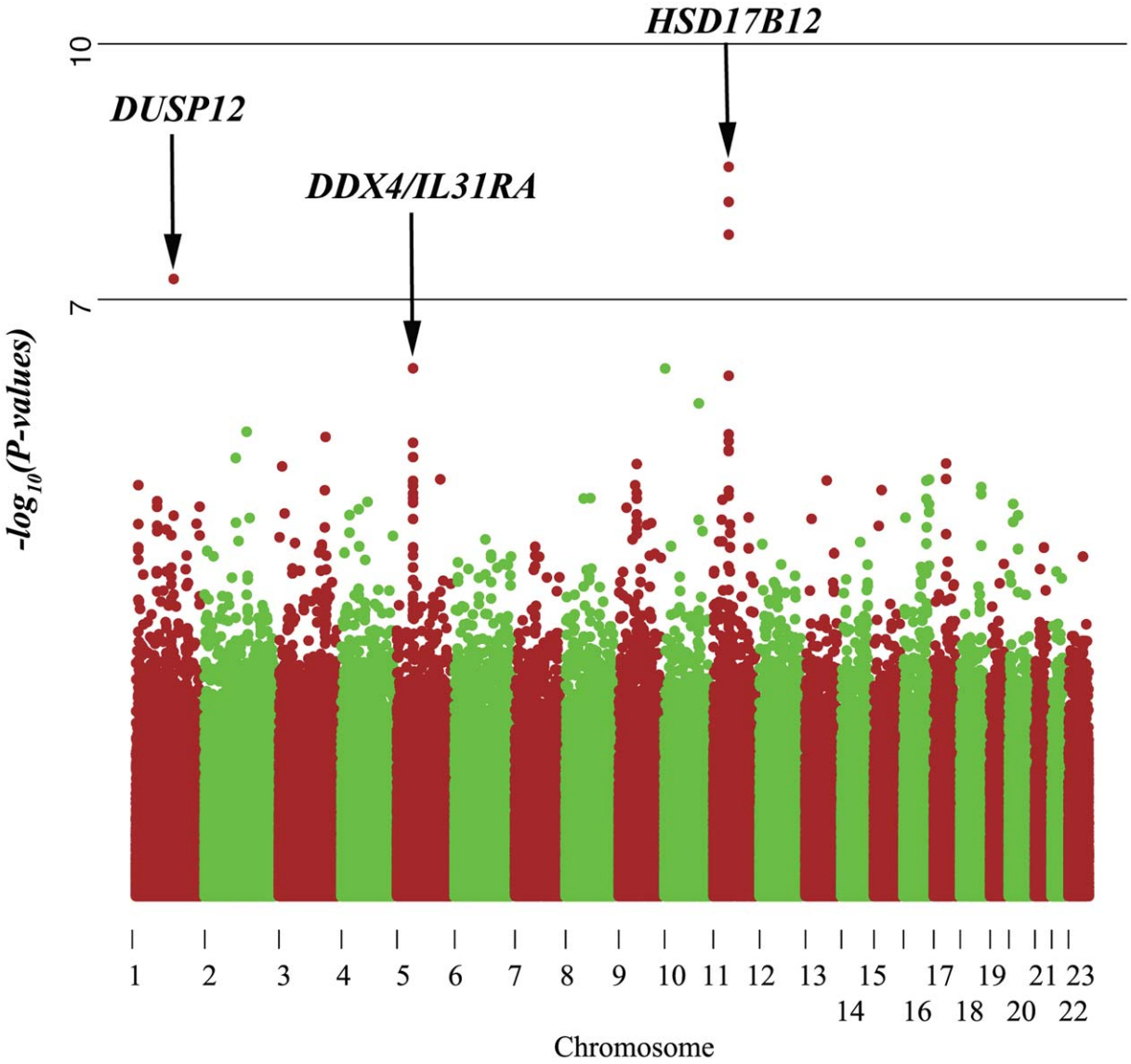

Figure 1. Manhattan plot of single marker analysis of the low-risk neuroblastoma data set. Even though the genes *DDX4* and *IL31RA* do not contain significantly associated SNPs ($P = 1.0\times10^{-07}$), the combined effect of moderately associated SNPs drives these two genes to be significant in our gene-centric analysis (genome-wide gene centric threshold p-values for significance is $P<3.15\times10^{-6}$).
doi:10.1371/journal.pgen.1002026.g001

effective than single marker analysis in detecting gene-level association signals. Indeed, our power computation, adjusting for 15,885 tests, indicated that our method performed far better than the single SNP method in both our discovery and replication case series (Figures S1, S2, S3, S4).

To further confirm the validity of our discovery, we computed a randomization p-value for each of the four newly discovered genes. For each of these genes, we computed a separate null test statistic distribution by calculating the gene-centric test statistics of one million randomly selected pseudo-genes. These pseudo-genes were selected to contain the same number of SNPs as were contained in the referenced gene. This method of selecting pseudo-genes was based on our observation that the average gene-based test statistic was strongly correlated with the number of SNPs in these genes (Figure S5), Using these null distributions to compare against the observed test statistics of the four newly discovered genes, we arrived at the randomization p-values (Table 2). These p-values (range 2.0×10^{-5}–1.0×10^{-6}) were compatible with the p-values asymptotically computed by our gene-centric method, and notably strengthened the credibility of the discovery of four novel disease causal genes associated with low-risk neuroblastoma.

To assess the joint impact on disease risk of these genes, we estimated the two-locus genotype odd ratios for all pairs amongst the four most significant SNPs within these four genes (Table 3). For each SNP pair tested, the independently contributed disease risks for carriers of risk alleles at only one locus were overall slightly stronger than the disease risks of each SNP when analyzed separately. In all but one case, the odd ratios of disease risks for carriers of both risk alleles increased markedly (odd ratios range from 2.505 to 3.435). However, no significant interaction between these SNP pairs was detected (P ranges from 0.459 to 0.909). Further, we computed all SNP pair interactions amongst all four genes and again noticed no significant SNP-

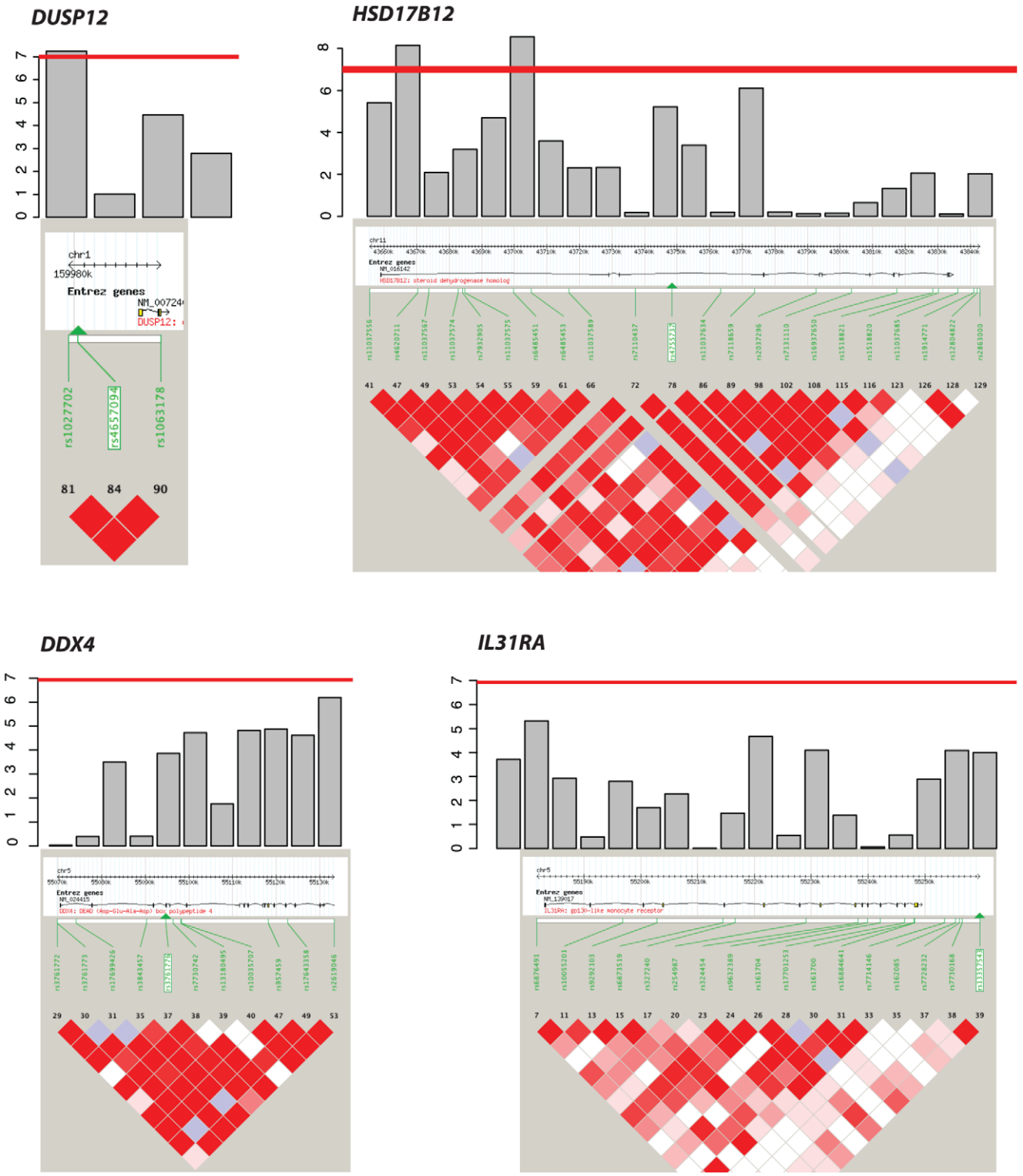

Figure 2. Haplotype view of the 4 genes significantly associated with low-risk neuroblastoma. Red line indicates $P<1.0\times10^{-7}$. Only *DUSP12* and *HSD17B12* contain SNPs with significant single-marker p-values in neuroblastoma low-risk subset. While *DDX4* and *IL31RA* do not contain significant SNPs, our gene-centric method was able to detect these genes as associated with low-risk neuroblastoma.
doi:10.1371/journal.pgen.1002026.g002

SNP interaction signals (the best SNP pair signals' P ranges from 0.108 to 0.523, Table S1). In the special case of *DDX4* and *IL31RA*, the modest disease risk for carriers of both risk alleles implicated the true association signal encompassed both genes though they are 38 kilo-bases apart from each other (Figure S6).

Table 2. Additional summaries of gene-centric analysis results for low-risk neuroblastoma.

Gene Symbols	Chromosome	Start- Stop	N° of SNP	Gene Randomization P-values	Most significant SNP	Most significant SNP P-values	Single SNP Replication P range
DUSP12	1q23.3	159986204-159993576	4	2.00×10^{-05}	rs1027702	**5.74×10^{-8}**	2.32×10^{-4}–8.32×10^{-2}
DDX4	5q11.2	55070534-55148362	11	1.00×10^{-06}	rs2619046	6.41×10^{-7}	1.15×10^{-3}–9.50×10^{-1}
IL31RA	5q11.2	55183090-55254434	18	1.00×10^{-06}	rs10055201	4.80×10^{-6}	7.09×10^{-4}–9.28×10^{-1}
HSD17B12	11p11.2	43658718-43834745	22	1.00×10^{-06}	rs11037575	**2.77×10^{-9}**	1.49×10^{-6}–8.28×10^{-1}

Bold-faced p-values indicate significant signal in single marker analysis using Bonferroni correction over 479,811 SNPs.
doi:10.1371/journal.pgen.1002026.t002

Lastly, we sought to further replicate our results in an independent cohort in Italy of 115 low-risk cases and 680 controls. We selected to genotype the three most significant SNPs (rs1027702, rs2619046, rs11037575) in the three loci that contain *DUSP12*, *DDX4/IL31RA*, and *HSD17B12* respectively. We analyzed these SNPs data sing various statistical tests listed in Table 4. Interestingly, rs1027702 showed strong replication signals for allele frequency association test as well as dominant model association test (P = 0.031 and 0.008 respectively). On the other hand, both rs2619046, rs11037575 showed strong significant signals for homozygous association test (P = 0.042 and 0.028 respectively) as well as recessive model association test (P = 0.047 and 0.037 respectively). Overall, these replication results provide unambiguous evidence to confirm these three loci as significantly associated with low-risk neuroblastoma.

Gene-set analyses identify enriched gene sets in all phenotypes

We are also interested in gene set analyses to identify specific pathways and gene sets that are enriched in neuroblastoma. To perform this analysis, we adapted the random-set approach, which was developed to analyze gene set analysis using gene expression data. This method [11] was suitable for our purpose since it required gene-level scores, which were conveniently obtained by taking the logarithm transformation of our gene-centric p-values.

We applied this random-set procedure using the overall, high-risk and low-risk data sets described earlier over 4734 gene sets obtained from the Gene Set Enrichment Analysis site [12], and selected enriched gene sets based on Bonferroni correction criterion ($P<1.05\times10^{-5}$). Additionally selected based on replication p-value threshold of 0.05, three Gene Ontology [13] sets were associated with all cases of neuroblastoma: *Nuclear Ubiquitin Ligase Complex*, *Negative Regulation of Intracellular Transport*, and *Regulation of Phosphorylation* (Table 5). The first two gene sets were also significantly enriched in high-risk neuroblastoma ($P=1.275\times10^{-09}$ and 6.332×10^{-07} respectively) with significant replication p-values (0.030 and 0.036 respectively). The third gene set appeared to be enriched in low-risk neuroblastoma ($P=1.678\times10^{-06}$); however, we were unable to replicate this result (P = 0.96). Furthermore, we identified and successfully replicated an additional gene set that was exclusively enriched in low-risk neuroblastoma: *Cytokine and Chemokine Mediate Signaling Pathway* (discovery $P=8.175\times10^{-06}$ and replication P = 0.040). The identification of these gene sets may elucidate biological pathways that are important in the biology of neuroblastoma.

Discussion

Taken together, this study implicates *DUSP12*, *DDX4*, *IL31RA*, and *HSD17B12* as neuroblastoma susceptibility genes, with particular relevance for those at low-risk for malignant progression and death from disease. Methodologically, we suggest that the

Table 3. Estimates of low-risk neuroblastoma odd ratios by genotype between the most significant SNPs.

Gene 1	Most significant SNP 1	Single marker SNP1 OR (95% CI) *P*	SNP1 carrier & SNP2 non-carrier OR (95% CI) *P*	Gene 2	Most significant SNP 2	Single marker SNP2 OR (95% CI) *P*	SNP2 carrier & SNP1 non-carrier OR (95% CI) *P*	SNP1 & SNP2 carrier OR (95% CI) *P*	Interaction P
DUSP12	rs1027702	**2.012** (1.47–2.79) 3.381×10^{-06}	**2.373** (1.48–3.98) 1.217×10^{-04}	*DDX1*	rs2619046	**1.477** (1.21–1.79) 5.702×10^{-05}	**1.826** (0.97–3.49) 5.018×10^{-02}	**3.435** (2.13–5.76) 1.123×10^{-08}	0.904
DUSP12	rs1027702	**2.012** (1.47–2.79) 3.381×10^{-06}	**2.108** (1.35–1.39) 4.308×10^{-04}	*IL31RA*	rs10055201	**1.494** (1.23–1.81) 3.848×10^{-05}	**1.622** (0.87–3.04) 0.132	**3.140** (2.00–5.07) 2.276×10^{-08}	0.627
DUSP12	rs1027702	**2.012** (1.47–2.79) 3.381×10^{-06}	**2.018** (1.11–3.93) 1.753×10^{-02}	*HSD17B12*	rs11037575	**1.674** (1.35–2.08) 1.075×10^{-06}	**1.715** (0.87–3.57) 0.122	**3.379** (1.90–6.47) 3.148×10^{-06}	0.778
DDX1	rs2619046	**1.477** (1.21–1.79) 5.702×10^{-05}	**1.346** (0.85–2.08) 0.170	*IL31RA*	rs10055201	**1.494** (1.23–1.81) 3.848×10^{-05}	**1.288** (0.58–2.68) 0.451	**1.561** (1.27–1.91) 1.193×10^{-05}	0.459
DDX1	rs2619046	**1.477** (1.21–1.79) 5.702×10^{-05}	**1.546** (1.07–2.24) 1.828×10^{-02}	*HSD17B12*	rs11037575	**1.674** (1.35–2.08) 1.075×10^{-06}	**1.732** (1.27–2.39) 3.645×10^{-04}	**2.534** (1.85–3.49) 6.632×10^{-10}	0.728
IL31RA	rs10055201	**1.494** (1.23–1.81) 3.848×10^{-05}	**1.485** (1.02–2.15) 3.453×10^{-02}	*HSD17B12*	rs11037575	**1.674** (1.35–2.08) 1.075×10^{-06}	**1.665** (1.24–2.26) 4.805×10^{-04}	**2.505** (1.84–3.42) 5.091×10^{-10}	0.909

Odd Ratios (OR), Confident Intervals (CI) and P-values (P) were computed from Fisher's exact test. No significant interaction was detected between any pairs of most significant SNPs.
doi:10.1371/journal.pgen.1002026.t003

Table 4. Single SNP replication results in Italian cohort (n = 115 low-risk neuroblastoma and 680 controls).

Genes	SNP	Discovery Single Marker TREND Test	Replication Allele Frequency Test	Replication Homozygous Model Test	Replication Dominant Model Test	Replication Recessive Model Test
DUSP12	rs1027702	5.74×10^{-08}	**0.031**	0.102	**0.008**	0.490
DD×4/IL31RA	rs2619046	6.41×10^{-07}	0.129	**0.042**	0.343	**0.047**
HSD17B12	rs11037575	2.77×10^{-09}	0.053	**0.028**	0.194	**0.037**

Bold-faced p-values indicate significant replication P-values<0.05.
doi:10.1371/journal.pgen.1002026.t004

gene-centric method has stronger power of detection of association signals compared to the single marker method (Figures S1, S2, S3, S4). Not only was the gene-centric method able to detect the two genes harboring genome-wide significant SNPs (*DUSP12* and *HSD17B12*), but also it was able to detect 2 genes that would have been missed by the single marker analysis (*DDX4* and *IL31RA*). Since this method was originally developed to analyze gene expression data, its limitation is the lack of ability to take into account the haplotype effect in computing gene level test statistics. However, our efforts to replicate the discovery with two independent cohorts unequivocally verify association signals at these loci. Further studies will be required to determine if these common variations tag cis- or trans-acting disease causal variations. Interestingly, the segregation of gene-level association signals and gene set enrichment scores between high-risk and low-risk neuroblastoma (Table 1 and Table 5) supports the view that common variation in the human genome can predisposed not only to a particular disease, but also to a clinically relevant disease subsets, thus demonstrating the power of robust phenotypic data in GWAS efforts.

Materials and Methods

Subjects and quality control

Study subjects. The neuroblastoma patients in this study were children registered through the North American-based Children's Oncology Group (COG) and were diagnosed with neuroblastoma or ganglioneuroblastoma. Blood samples from the neuroblastoma cases were identified through the COG neuroblastoma repository for specimen collection at time of diagnosis. All specimens were annotated with clinical and genomic information (Table S2). Samples were assigned into three risk groups (low-risk, intermediate-risk and high-risk) based on the COG risk assignment algorithm [1], that includes patient age at diagnosis [14], International Neuroblastoma Staging System (INSS) stage [2], tumor histopathology [15], DNA index [16], and *MYCN* amplification status [17]. The only eligibility criterion for genotyping was availability of 1.5 μg of high quality DNA from a tumor-free source such as peripheral blood or bone marrow cells uninvolved with tumor. Since neuroblastoma in the United States is demographically a disease of Caucasian of European descendent, we limited our analyses to this ethnic group to minimize genetic heterogeneity. Summaries of clinical and genomic information of our discovery and replication cohorts are provided in Table S2.

The control group in this study included 2575 children of self-reported Caucasian ancestry who were recruited and genotyped by the Center for Applied Genomics at the Children's Hospital of Philadelphia (CHOP). Eligibility criteria for control subjects were: 1) self-reported Caucasian; 2) availability of 1.5 μg of high quality DNA from peripheral blood or mononuclear bone marrow cells; and 3) no known medical disorder, including cancer, based on self-reported intake questionnaire and/or clinician-based assessments.

The CHOP Institutional Review Board approved this study.

Genotyping and quality control for discovery cohort. SNP genotyping was performed using the Illumina Infinium II BeadChip (Illumina, San Diego, CA, USA) according to methods detailed elsewhere [4], [5]. Since a portion of the individuals in the discovery cohort was genotyped by the HumanHap550 v1 array (n = 859) while others were genotyped by the v3 array (n = 768), our analysis only concerned the markers shared by the v1 and v3 array. The HumanHap550 v1 array contains 555,175 markers, while the v3 array contains 561,288 markers, including 544,902 markers that are shared by the two arrays. We filtered out 8,749 SNP markers with call rate less than 95%. We also excluded 5,415 SNP markers whose Hardy-Weinberg Equilibrium p-values were less than 0.001. Finally, we excluded additional 50,869 SNP markers whose minor allele frequency is less than 5%.

Table 5. Summary of gene set analysis results for all, high-risk, and low-risk neuroblastoma.

Gene Set Names	N° of genes	Overall Discovery p-values	Overall Replication p-values	High-risk Discovery p-values	High-risk Replication p-values	Low-risk Discovery p-values	Low-risk Replication p-values
Nuclear Ubiquitin Ligase Complex	10	$\mathbf{1.084\times10^{-09}}$	$\mathbf{6.620\times10^{-03}}$	$\mathbf{1.275\times10^{-09}}$	3.024×10^{-02}	0.469	0.753
Negative Regulation of Intracellular Transport	10	$\mathbf{5.692\times10^{-07}}$	$\mathbf{1.160\times10^{-02}}$	$\mathbf{6.332\times10^{-07}}$	3.610×10^{-02}	0.361	0.184
Regulation of Phosphorylation	42	$\mathbf{6.020\times10^{-07}}$	$\mathbf{4.142\times10^{-02}}$	2.940×10^{-02}	0.925	1.678×10^{-06}	0.960
Cytokine and Chemokine Mediate Signaling Pathway	18	0.109	0.756	0.813	0.928	$\mathbf{8.175\times10^{-06}}$	$\mathbf{4.027\times10^{-02}}$

Bold face p-values of different gene sets at different risk groups (overall, high-risk, low-risk) indicate significant enrichment of that gene set in that risk group.
doi:10.1371/journal.pgen.1002026.t005

A total of 96 cases were removed from our data set due to their low genotype call rate (<95%). Furthermore, we used Multi-Dimensional Scaling (MDS) as implemented in the PLINK [10], for inferring population structure (Figure S7). Comparing self-identified ancestry with MDS-inferred ancestry confirmed 1642 neuroblastoma patients of European ancestry. Finally, we calculated genome-wide identity-by-state (IBS) estimates for all pair-wise comparisons among all case subjects and control subjects to detect cryptic relatedness and potential duplicated genotype within our data set. This step further excluded 15 neuroblastoma patients from our analyses.

After all quality control steps, our discovery data set contained 1627 neuroblastoma case subjects of European ancestry, each of which contained 479,811 SNP markers. To correct the potential effects of population structure, 2575 matching control subjects of European ancestry were selected based on their low IBS estimates with case subjects. The genomic control inflation factor for this data set was 1.08.

Five hundred and seventy four (574) low-risk cases, selected from the above 1627 cases, were included for all low-risk neuroblastoma analyses. To keep the genomic inflation factor low, three best matching control subjects were selected for each case, based on IBS estimates, making a total of 1722 control subjects included for analyses. The genomic control inflation factor for this data set was 1.07.

Genotyping and quality control for initial replication cohort. SNP genotyping was performed using the Illumina Human610-Quad array that includes both SNP and CNV markers. The Human610-Quad array contains 620,901 SNPs. We filtered out 48,831 SNP markers with call rate less than 95%. We also excluded 13,305 SNP markers whose Hardy-Weinberg Equilibrium p-values were less than 0.001. Finally, we excluded additional 49,057 SNP markers whose minor allele frequency was less than 5%. A total of 15 cases were removed from our data set due to their low genotype call rate (<95%). After all quality control steps, our replication data set contained 398 neuroblastoma case subjects of European ancestry, each of which contained 509,708 SNP markers. To correct the potential effects of population structure, 1507 matching control subjects of European ancestry were selected based on their low IBS estimates with case subjects.

One hundred and twenty four (124) low-risk cases, selected from the above 398 cases, were included for all low-risk neuroblastoma replication analyses. For each case, four best matching control subjects were selected based on IBS estimates, making a total of 496 control subjects included for analyses.

Genotyping of second replication cohort. One hundred and fifteen (115) low-risk neuroblastoma subjects for Italy and six hundred and eighty (680) control Italian subjects were selected to be genotyped at three SNPs: rs1027702, rs2619046, and rs11037575. All samples were genotyped by Taqman SNP Genotyping Assay by Applied Biosystems.

Statistical analyses

Gene-centric analysis. Our gene-centric analysis adopted the global test method [8], developed to test association of a group of genes using microarray data. First, to mirror gene expression data, we quantified our SNP genotype data by counting the number of minor alleles for each sample at each SNP. Second, due to the analogy in relative relationship of the two concepts in global test and in our study, we substituted the concepts of "genes" and "group of genes" from global test with "SNPs" and "genes" respectively.

This method adopted the generalized linear model framework to model the relationship between Y, a vector of clinical outcomes, and X, the $n\times m$ matrix of genotypic data of n subjects and m SNPs. In this model, α is the intercept, β is a length m vector of regression coefficients, and h is a general link function such as the logit function

$$E(Y_i|\beta)=h^{-1}\left(\alpha+\sum_{j=1}^{m}x_{ij}\beta_j\right)$$

Testing association between genotypic data and clinical outcomes is equivalent to testing the null hypothesis H_0: $\beta_1=\beta_2=\ldots=\beta_m=0$. Since the number of SNP is much larger than the number of subjects, it is not possible to test this hypothesis in a classical way. Instead, we could test H_0 if we assume $\beta_1,\ldots,\beta_m$ to be samples from a common distribution with expectation zero and variance τ^2. The null hypothesis becomes simply H_0: $\tau^2=0$. If we rewrite the model in terms $r_i=\sum_j x_{ij}\beta_j$, with $i=1,\ldots,n$, then r_i is the linear predictor, the total effect of all covariates for subject i. Let $\boldsymbol{r}=(r_1,\ldots,r_n)$, then $\boldsymbol{r}$ is a random vector with expectation zero and $Cov(\boldsymbol{r})=\tau^2XX'$. The original model simplifies to

$$E(Y_i|r_i)=h^{-1}(\alpha+r_i)$$

This is a simple random effect model. Under the null hypothesis, the test statistic

$$Q=\frac{(Y-\mu)'R(Y-\mu)}{\mu_2}$$

has expectation $E(Q)=trace(R)$ and variance:

$$Var(Q)=2trace\left(R^2\right)+\left(\frac{\mu_4}{\mu_2^2}-3\right)\sum\nolimits_i R_i^2$$

where $R=(1/m)XX'$ is an $n\times n$ matrix proportional to the covariance matrix of the random effects $\boldsymbol{r}$, $\mu=h^{-1}(\alpha)$ is the expectation of Y under H_0, and μ_2 and μ_4 are the second and fourth central moments of Y under H_0.

The test statistic Q could be rewritten as

$$Q=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\mu_2}\left[X_i'(Y-\mu)\right]^2$$

where X_i is the length n vector of genotype of SNP i. The expression $Q_i=(1/\mu_2)[X_i'(Y-\mu)]^2$ would be exactly the test statistic of SNP i if it were the only SNP on the gene of interested; or we could interpret that Q_i is the "contribution" of SNP i to the overall test statistic. This means that the overall test statistic is simply the average of the statistics $Q_1,\ldots,Q_m$ of m individual SNPs. Notably, the averaging is over a squared covariance between genotype and clinical outcomes, SNPs with large variance (i.e. strong association signals) have stronger influences on the outcome of the test statistic Q than those with weaker association signals.

Using this method, we computed an aggregated effect of all SNPs that are located from 10-kilo bases upstream to 10-kilo bases downstream for the gene being tested, and computed asymptotic p-value for each gene. We performed the global test on 15,885

annotated, unannotated and predicted genes downloaded from the UCSC Genome Browser [9], and used strict Bonferroni correction criterion ($P<3.15\times10^{-6}$) to determine whether a gene was associated with neuroblastoma. True association signals were further selected based on replication p-value less than 0.05.

Randomization p-value computation. For each significant gene, we computed randomization p-values by comparing its test statistic and its respective null distributions. A null distribution of a gene was composed of test statistics of one million pseudo-genes having the same number of SNPs as the referenced genes. The SNPs of these pseudo-genes were randomly selected across the genome.

Odd ratios estimation. We used the Fisher's exact test to estimate the odd ratios using genotype data as well as the 95% confident interval and p-values.

SNP-SNP interaction estimation. Single marker interaction scores were computed using the general linear model to compute interaction effect between two SNPs.

Gene set analysis. To analyze the significance of gene sets, we adopted the random-set method [11] since it allows us to utilize the gene-centric results to compute enrichment score for each gene set. In this analysis, we used the logarithm transformation of our gene-centric method as gene-level scores to detect gene sets enriched in neuroblastoma.

We performed three separate gene set analyses for overall, high-risk and low-risk data sets over 4734 gene sets downloaded from the Broad Institute MsigDb [12]. These gene sets include five categories: *positional gene sets*, *curated gene sets* (chemical and genetic perturbations, and canonical pathways), *motif gene sets* (microRNA targets, and transcription factor targets), *computational gene sets* (cancer modules, and cancer gene neighborhoods), and *GO gene sets* (GO cellular components, GO biological process, and GO molecular function). Strict Bonferroni correction criterion was used to select gene sets that are enriched in neuroblastoma.

Data deposition

The genotypic and phenotypic information from this study is deposited in dbGAP (www.ncbi.nlm.gov/gap) under accession number phs000124.v2.p1.

Ethics statement

The Children's Hospital of Philadelphia Institutional Review Board approved this study.

Supporting Information

Figure S1 Power calculation of single SNP analysis of the low-risk neuroblastoma discovery set, adjusting for 500,000 tests.
(TIF)

Figure S2 Power calculation of single SNP analysis of the low-risk neuroblastoma discovery set, adjusting for 15,885 tests.
(TIF)

Figure S3 Power calculation of single SNP analysis of the low-risk neuroblastoma replication set, adjusting for 10 SNPs (average number of SNPs in a gene).
(TIF)

Figure S4 Power calculation of single SNP analysis of the low-risk neuroblastoma replication set with no multiple testing adjustments.
(TIF)

Figure S5 The correlation between average test statistic of genes and the number of SNPs on those genes.
(TIF)

Figure S6 Single SNP association signals of 610 kilo-base region encompassing *DDX4* and *IL31RA*: blue color indicates the SNPs mapped to these two genes respectively.
(TIF)

Figure S7 Multi-dimensional scaling plot. Circled area denotes Caucasian cluster which includes HapMap CEU subjects as well as the cases and controls used in this study.
(TIF)

Table S1 Summary of the most significant SNP pair interaction signals amongst the four genes associated with low-risk neuroblastoma.
(DOC)

Table S2 Summary of clinical and genomic information of neuroblastoma cases.
(DOC)

Author Contributions

Performed analyses: LBN KW MD HH HKL JMM. Performed genotyping and validation: LBN SJD MC MD JG CK EFA YPM KC AI. Conceived and designed the experiments: JMM LBN HKL SJD HH MC. Performed the experiments: LBN SJD MC KW MAD JG CK EFA YPM KC AI MD HH HKL JMM. Analyzed the data: LBN SJD MC KW MAD JG CK EFA YPM KC AI MD HH HKL JMM. Contributed reagents/materials/analysis tools: JMM HH MC AI. Wrote the paper: LBN SJD HH HKL JMM.

References

1. Maris JM (2010) Recent advances in neuroblastoma. N Engl J Med 362: 2202–2211.
2. Brodeur GM, Pritchard J, Berthold F, Carlsen NL, Castel V, et al. (1993) Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment. J Clin Oncol 11: 1466–1477.
3. Ambros PF, Ambros IM, Brodeur GM, Haber M, Khan J, et al. (2009) International consensus for neuroblastoma molecular diagnostics: report from the International Neuroblastoma Risk Group (INRG) Biology Committee. Br J Cancer 100: 1471–1482.
4. Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, et al. (2008) Chromosome 6p22 Locus Associated with Clinically Aggressive Neuroblastoma. N Engl J Med 358: 2585–2593.
5. Capasso M, Devoto M, Hou C, Asgharzadeh S, Glessner JT, et al. (2009) Common variations in BARD1 influence susceptibility to high-risk neuroblastoma. Nat Genet 41: 718–723.
6. Wang K, Diskin SJ, Zhang H, Attiyeh EF, Winter C, et al. (2011) Integrative genomics identifies LMO1 as a neuroblastoma oncogene. Nature 469: 216–220.
7. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, et al. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459: 987–991.
8. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93–99.
9. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34: D590–598.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
11. Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahquist P (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 1: 85–106.
12. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics 23: 3251–3253.

13. Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Res 36: D440–4.
14. Breslow N, McCann B (1971) Statistical estimation of prognosis for children with neuroblastoma. Cancer Res 31: 2098–2103.
15. Shimada H, Ambros IM, Dehner LP, Hata J, Joshi VV, et al. (1999) The International Neuroblastoma Pathology Classification (the Shimada system). Cancer 86: 364–372.
16. Look AT, Hayes FA, Shuster JJ, Douglass EC, Castleberry RP, et al. (1991) Clinical relevance of tumor cell ploidy and N-myc gene amplification in childhood neuroblastoma: a Pediatric Oncology Group study. J Clin Oncol 9: 581–591.
17. Brodeur GM, Seeger RC, Schwab M, Varmus HE, Bishop JM (1984) Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. Science 224: 1121–1124.