

Inferring a Tumor Progression Model for Neuroblastoma From Genomic Data

Sven Bilke, Qing-Rong Chen, Frank Westerman, Manfred Schwab, Daniel Catchpole, and Javed Khan

From the Oncogenomics Section, Pediatric Oncology Branch, Advanced Technology Center, National Cancer Institute, Gaithersburg, MD; Department of Tumor Genetics, German Cancer Research Center, Heidelberg, Germany; and Tumour Bank, The Children's Hospital at Westmead, Westmead, Australia.

Submitted June 27, 2005; accepted July 18, 2005.

Supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research.

S.B. and Q.-R.C. contributed equally to this work.

Terms in [blue](#) are defined in the glossary, found at the end of this issue and online at www.jco.org.

Authors' disclosures of potential conflicts of interest are found at the end of this article.

Address reprint requests to Javed Khan, MD, National Cancer Institute, Advanced Technology Center, Room 134E, 8717 Grovemont Cir, Bethesda, MD 20892-4605; e-mail: khanjav@mail.nih.gov.

0732-183X/05/2329-7322/\$20.00

DOI: 10.1200/JCO.2005.03.2821

A B S T R A C T

Purpose

The knowledge of the key genomic events that are causal to cancer development and progression not only is invaluable for our understanding of cancer biology but also may have a direct clinical impact. The task of deciphering a model of tumor progression by requiring that it explains (or at least does not contradict) known clinical and molecular evidence can be very demanding, particularly for cancers with complex patterns of clinical and molecular evidence.

Materials and Methods

We formalize the process of model inference and show how a progression model for neuroblastoma (NB) can be inferred from genomic data. The core idea of our method is to translate the model of clonal cancer evolution to mathematical testable rules of inheritance. Seventy-eight NB samples in stages 1, 4S, and 4 were analyzed with array-based comparative genomic hybridization.

Results

The pattern of recurrent genomic alterations in NB is strongly stage dependent and it is possible to identify traces of tumor progression in this type of data.

Conclusion

A tumor progression model for neuroblastoma is inferred, which is in agreement with clinical evidence, explains part of the heterogeneity of the clinical behavior observed for NB, and is compatible with existing empirical models of NB progression.

J Clin Oncol 23:7322-7331.

INTRODUCTION

Knudson's two-hit model¹ describes the deactivation of both alleles of a **tumor suppressor gene** as the initiating event of oncogenesis. The subsequent progression toward an aggressive malignancy is a multi-step process with the reduction of the cell's dependence on growth signals as well as suppression of apoptotic pathways as hallmarks. The linear progression of colorectal cancer is probably the best-characterized genetic model of tumorigenesis²: the inactivation of a gate-keeper gene,³ adenomatous polyposis coli (*APC*), initiates colorectal neoplasia and proceeds through the mutation of oncogenes such as *KRAS* and

apoptosis-related genes such as *TP53* to a carcinoma.⁴ In general, one can assume that the activity of several genes needs to be changed in order to develop any cancer. The mutation process underlying these changes is inherently random and undirected. However, it seems unlikely that the necessary alterations could happen simultaneously by chance alone, particularly when more than a few genes need to be mutated. This has led to the model of a **clonal evolution**,⁵ which guides the random mutation process by selection of alterations providing a growth advantage.

In this article, we argue that the clonal evolution process should leave characteristic signatures of inheritance along the

pathways of progression and present a method to infer models of tumor progression by an identification of these signatures in genome-wide data of mutations. One mutational process that can be monitored is changes of DNA copy number by high-resolution comparative genomic hybridization (CGH). The distribution of copy-number changes in a given cancer type is nonrandom, and changes occur at recurrent locations.⁶ This suggests the presence of tumor suppressors or oncogenes in lost or gained regions, respectively. For a broad range of cancers, as well as for subtypes of the same cancer, characteristic patterns of recurrent alterations have been observed.⁷ The direct impact of DNA copy-number changes on gene transcript levels was demonstrated by simultaneous measurements of DNA copy numbers and mRNA levels.^{8,9} Consequently, DNA copy-number changes contribute to the mRNA expression profile and ultimately to the behavior of the tumor cell.

Neuroblastoma (NB) is well known for its pronounced clinical heterogeneity, and several studies¹⁰⁻¹³ indicate that the characteristic patterns of genomic alterations correlate with the different phenotypic stages of the disease. This makes NB well suited as a test set for the methods presented here. NB is one of the most common pediatric solid tumors and accounts for 7% to 10% of all childhood cancers.¹⁴ The prognosis of patients with NB varies according to the stage and *MYCN* amplification status.¹⁴ Stage 1 disease is essentially curable, whereas patients with stage 4 disease, in particular those with *MYCN* amplification, remain largely incurable despite advances in cancer therapeutics.¹⁴ Stage 4S tumors represent an enigmatic group of metastatic tumors with a small, localized primary tumor. This type, which is associated with an excellent prognosis and spontaneous regression in the majority of the cases, has a unique pattern of dissemination primarily to the liver and skin in infants younger than 1 year. These diverse biologic behaviors, which are often associated with particular genetic changes, makes NB a paradigm for the investigation of genomic alterations associated with progression models. Genomic alterations in NB have been investigated by cytogenetic, and molecular methods including *spectral karyotyping* and *metaphase CGH (M-CGH)*.^{11,13,15-18} On the basis of these studies, several genomic alterations have been reported to correlate with prognosis, including amplification of the *MYCN* oncogene (found in 30% of NB),^{14,19} gains of 17q (> 50%) and loss of 1p36 (30% to 35%).^{14,20-22} Other recurrent changes including losses of 3p, 4p, 9p, 11q, and 14q, as well as frequent gain of chromosome 7, have also been suggested to have relevance to the development and progression of these tumors.²²⁻²⁵

Currently no gold-standard molecular model of NB progression exists. The pronounced clinical heterogeneity of NB indicates a nonlinear progression, unlike the

development of colorectal cancer. The contrast between the highly malignant and benign stages in NB is so extreme that some authors have noted that the two groups of tumors seem to reflect different diseases.^{26,27} Aneuploidy was found to be an important prognostic marker for survival in children younger than approximately 18 months,²⁸ but it loses its predictive power for older patients. More recent reports have indicated that near ditetraploidy²⁹ is a factor that indicates poor prognosis in NB. The ploidy-changing process is commonly considered as a distinctive and early event in NB development,^{21,27,30-32} and hypothetical models of NB development center around this process. One popular model³² reflects the older distinction between diploid and aneuploid tumors. It classifies NB into benign, hyperdiploid variants with mitotic dysfunction and aggressive variants characterized by gain of 17q. A more recent speculative model²⁷ incorporates the observation that tetraploidy is a strong marker for bad outcome and uses a hypothesis by Kaneko and Knudson³¹ that suggested that all levels of ploidy in NB result from the same molecular process, namely a multipolar division of tetraploid cells. Here we have utilized high-resolution genomic copy number data generated from array-based CGH (aCGH) to infer a model of the progression of neuroblastoma.

MATERIALS AND METHODS

Model Selection

The principle used to select the one tumor-progression model compatible with genomic data from all possible tumor progression models (applying the biologic assumptions outlined in Results) is fairly straightforward: each theoretical model has a one-to-one correspondence to occupation pattern of the common, shared (between two or more sets) and specific patterns of mutations. Therefore it is sufficient to identify which sets are occupied and which sets are empty. Possible outcomes of such experiments can diagrammatically be represented as occupations of a Venn diagram (Fig 3). For two stages, such a diagram has three distinct sets, one representing common alterations and two sets representing alterations specific to A and B, respectively. Each set can either be empty or occupied; therefore there are $2^3 = 8$ possible experimental outcomes. Of these, only three map to the progression models I and II in that figure. The remaining five possible experimental outcomes are incompatible with an evolutionary progression of the disease involving the observed genomic alterations. For example an experiment that does not detect any recurrent alterations at all (all sets empty), is in violation of rule 1 (presence of a progression signature) and rule 3 (signature of a common disease origin).

Mutations are typically not present in all tumors of a given stage and they are also typically not exclusive to that stage. However, the frequency of mutations is often significantly different in distinct stages. Consequently we use the frequency of mutations as the primary observable. In what follows, we present the rules to identify common, shared and specific mutations specifically for DNA copy number changes. It should not be difficult to perform similar calculations for other types of mutations.

We define a genomic alteration as common to all stages if the alteration is recurrent for each stage individually. Recurrent means that the frequency v of an alteration at genomic position x is higher ($v > \phi$) than expected by random chance. A value for the threshold ϕ can be estimated by analyzing the null-hypothesis, namely that the probability $\Pi(x) = \Pi$ for a mutation at position x is independent of x and approximately constant for the whole genome. The random process is binomial and the probability to find $\omega > n$ out of N samples with a genomic imbalance is given by

$$P(\omega \geq n | N, \Pi) = \sum_{\omega'=n}^N \binom{N}{\omega'} \Pi^{\omega'} (1-\Pi)^{N-\omega'}. \quad (1)$$

The smallest $v = n/N$ for which this P value is $P(\omega \geq n) < \alpha$ defines the threshold ϕ for recurrent regions. The desired significance level α needs to be adjusted for multiple comparisons. We use $\alpha = .05/L \cong 2.5 \times 10^{-6}$ where $L \cong 20,000$ is the number of probed genomic locations. An estimate of the position independent probability Π can be obtained from the data set by counting the number I of genomic imbalances on the whole genome in the N samples: $\Pi \cong I/(LN)$. In our NB data set, we find empirically $\Pi \cong 0.07$ and typically have $N = 20$ for each phenotype. With equation 1, one finds for these parameters that a region can be called recurrent when an imbalance is observed in more than 50% of the samples. An imbalance is defined as common to all stages if the region is recurrent for each stage individually and we use as the criterion

$$P_{\Sigma} = \sum_{\text{stages}} P_s(\omega \geq n_s | N_s, \Pi) < \theta; \quad (2)$$

the definition of unique or shared alterations is based on significant differences in the frequency of imbalances in one (unique), two, or three (shared) phenotypes as compared to the remaining phenotypes. The term "significantly more frequent" can be analyzed using the null hypothesis

$$\Pi(x|\text{stage A}) = \Pi(x|\text{stage B}) = \Pi(x). \quad (3)$$

The probability for a gain (or loss) at a location x does not depend on the phenotype of the disease. Under this assumption, the hypergeometric distribution can be used to estimate the probability to observe more than n_a genomic imbalances in a subset of N_a samples, whereas n_b out of the total N_b samples in a different set:

$$P_{A,B}(\omega \geq n_A | N_A, N_B, n_B) = \sum_{\omega'=n_A}^{n_A+n_B} \frac{\binom{N_A}{\omega'} \binom{N_B}{n_A+n_B-\omega'}}{\binom{N_A+N_B}{n_A+n_B}}. \quad (4)$$

An imbalance unique to stage A is defined by

$$P_{\Sigma}^A = \sum_{B \neq A} P_{A,B} < \theta_u, \quad (5)$$

the observed frequency in stage A is significantly higher than in any other stage. Similarly,

$$P_{\Sigma}^{[A,B]} = \sum_{C \notin \{A,B\}} P_{A,C} + P_{B,C} < \theta_s \quad (6)$$

a significantly higher frequency in both A and B compared with all other phenotypes defines shared imbalances. The definition of

shared between two for four end points is given by equation 6 with the indices $P_{B,A}$ swapped (ie, a genomic change is much less frequent in a stage A than in any other stage). In order to understand the distribution of the composite values (equations 2, 5, and 6), which are a sum of non-independently distributed terms, we simulated the distributions in a random permutation test with 20,000 re-labelings providing an approximation of the relevant distributions.

The thresholds θ_s, θ_u for shared and unique regions need to be adjusted for multiple comparisons. Some care is needed at this step, because type I errors caused by a too-large threshold may cause an otherwise empty set to appear occupied, whereas too-small thresholds (type II errors) may change the outcome of the model selection by making an occupied set appear empty. Therefore, instead of using, to a degree, arbitrary thresholds, we probe a range of thresholds and compare the number of selected locations with the number obtained in a random permutation test. We define a set to be empty if in the whole range of thresholds the number of selected locations never exceeds the number found in the random permutation test by more than a few percent.

Data Analysis

Fluorescence ratios were normalized for each microarray by setting the average log ratio for each subarray element equal to zero (commonly referred to as pin normalization). The data were quality-filtered by removing those clones that had poor quality measurement⁴⁴ (quality < 0.5) in more than 20% of all the samples. For the clones that passed this filter, the fluorescence ratio of low-quality spots for the individual samples was replaced by the average ratio value of the remaining good measurements for that clone. The clones were then assigned to UniGene clusters (February 2005). For the UniGene clusters represented by multiple clones, mean fluorescence ratios of those clones are used. After these processes we had 17,692 unique UniGene clusters remaining from the initial 42,591 clones. Map positions for the clusters were assigned by Blat searches against the Golden Path genome assembly (<http://genome.ucsc.edu/>; May 2004 Freeze). Throughout this article, all genomic coordinates are given with the respect to this assembly. Finally, the clusters were sorted according to their starting position of sequence on each individual chromosome.

Detection of Genomic Changes and Frequency Estimation

Systematic errors make the detection of low-level DNA copy-number changes with cDNA arrays difficult³⁵ and may reduce the reliability⁴⁵ of the data. We used topological statistics³⁵ to reduce systematic errors and obtain P values for the presence of gains and losses in individual samples. This algorithm is a generalization of the sliding-window smoothing filter that uses data from self-self hybridization to reduce systematic errors in cDNA aCGH data. It calculates, for each sample and each chromosomal position, a P value for the presence of a gain or a loss, respectively. To deal with the limited sensitivity of cDNA microarrays and to reduce type II (false-negative) statistical errors, we estimate the frequency of genomic alterations³⁵ from the average P value denoted as \bar{p} within S1, S4S, S4-, and S4+. This value is proportional to the frequency

$$v = \frac{N_w}{N_t} \quad (7)$$

of their occurrence, where N_w is the number of samples in a subgroup with a given genomic imbalance and N_t is the overall

number of samples in that subgroup. This is valid as follows: For all loci in which there are no genomic imbalances, the observed P value P_{gc} will follow the flat theoretical distribution with a mean (expectation value) $\langle P_{gc} \rangle = 0.5$. When one is sure of a genomic imbalance, P_{gc} is close to zero (for example $P_{gc} < .001$). With these considerations one can write

$$\bar{p} \approx \frac{(0.5 \times N_{\text{no change}}) + (0 \times N_w)}{N_t} = (1 - v) \times 0.5; \quad (8)$$

that is, the lower P , the higher the frequency of a genomic change in that locus.

Tumors

The data used in this study include and extend data published earlier and we refer to Chen, Bilke, and Wei¹³ for details left out here to save space, including the technical details about the microarray experiments. Seventy-eight snap frozen NB specimens were obtained from 20 patients with S1, 15 S4S samples without and two with *MYCN* amplification, and 39 patients with stage 4, of which 18 were *MYCN*-amplified S4+ and 21 were *MYCN* single-copy S4- tumors. The original histologic diagnoses were made at tertiary hospitals with extensive experience in diagnosis and management of NB.

RESULTS

PCA Visualization of Genomic Alterations in NB of Different Stages

In this study we focused our analysis on NB tumors of stages 1(S1), 4S (S4S), 4 without *MYCN* amplification (S4-) and 4 with *MYCN* amplification (S4+). To demonstrate the stage dependence of DNA copy-number changes in NB, we selected measurements indicating differential DNA copy-number levels ($P < .01$) with a one-way ANOVA³³ (Analysis of Variance) analysis. Next we used principal component analysis³⁴ to visualize the data for these clones (Fig 1). A moderate separation between the different stages of NB is evident in the copy-number profiles projected on the second and third principal component. Although S4-, S4+, and S1 + S4S form three well-defined, separate clusters, we cannot discriminate between S1 and S4S with this simple analysis. Interestingly, the aberration patterns for the two *MYCN*-amplified S4S tumors appear to be closer to S4+ than to the non-*MYCN* amplified tumors in the same stage. The much more aggressive phenotype and the observed altered expression profiles of the *MYCN*-amplified S4S tumors suggest that they form a distinct biologic subgroup. However, the only two samples available in this study did not suffice to draw statistically significant conclusions. The two samples were therefore excluded from the subsequent analysis.

Frequency of Genomic Alterations in NB of Different Stages

Next we used topological statistics³⁵ on the complete data set to obtain P values for the presence of DNA copy-

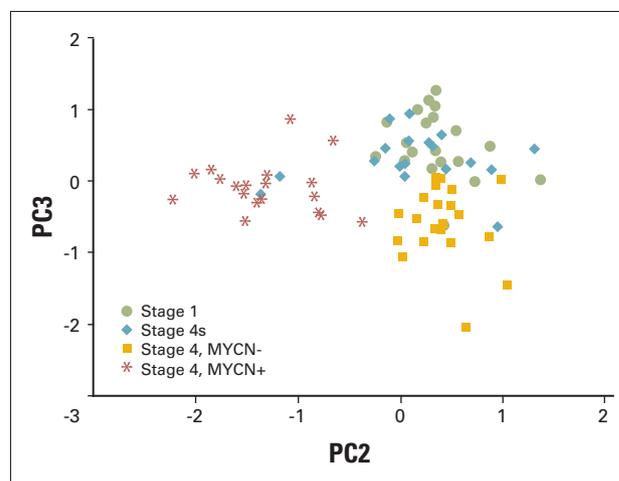


Fig 1. Principal component analysis of the DNA copy-number data indicating differential copy numbers ($P < .01$) between stages 1, 4-, 4+, and 4S in a one-way Analysis of Variance analysis. Each point represents one patient, the coordinates were calculated by projecting the DNA alteration pattern on the second (PC2) and third (PC3) principal component. With these coordinates a separation of the different stages, with the exception of stages 1 and 4S, becomes visible. The two S4S samples within the S4+ samples were the only two *MYCN*-amplified samples. These two samples were then removed in the subsequent analysis.

number changes for each tumor. On average, each NB tumor in our analysis gained or lost (with $P < .001$) approximately 7.5% of the genome. One can expect that the majority of changes present in only one tumor (or a small number of tumors) do not carry a high level of biologic significance. Of interest are recurrent genomic alterations occurring with a higher frequency than expected by chance. One way to estimate the frequency of genomic alterations from noisy data is to calculate the average P value p for copy-number changes (see Methods section and Bilke et al³⁵). The result of this analysis for the four stages of NB is depicted in Figure 2. Interestingly our high-resolution analysis indicated that a small region extending from 118 to 119 Mbp on 11q is lost in all S4- and a large fraction of S1 and S4S tumors. Proximal to this region a larger loss of heterozygosity (LOH) extending over 20MBp was observed most frequently for S4- tumors. A gain in another small region on 2p23 (31 to 33 Mbp) was identified for all S4S and most S1 tumors. We also confirm that gains on chromosome 17 are mostly limited to 17q for S4- and S4+, whereas in S1 and S4S, gains of the whole chromosome 17 are frequent.

Inferring a Tumor Progression Model of NB for Four Stages

We next utilized the frequency of genomic alterations for each of the four subgroups to determine the best fitting model of genetic evolution for NB. The inference procedure is built around the three following widely accepted principles of genetic evolution: (1) All changes found in a parent genotype must be present in the offspring

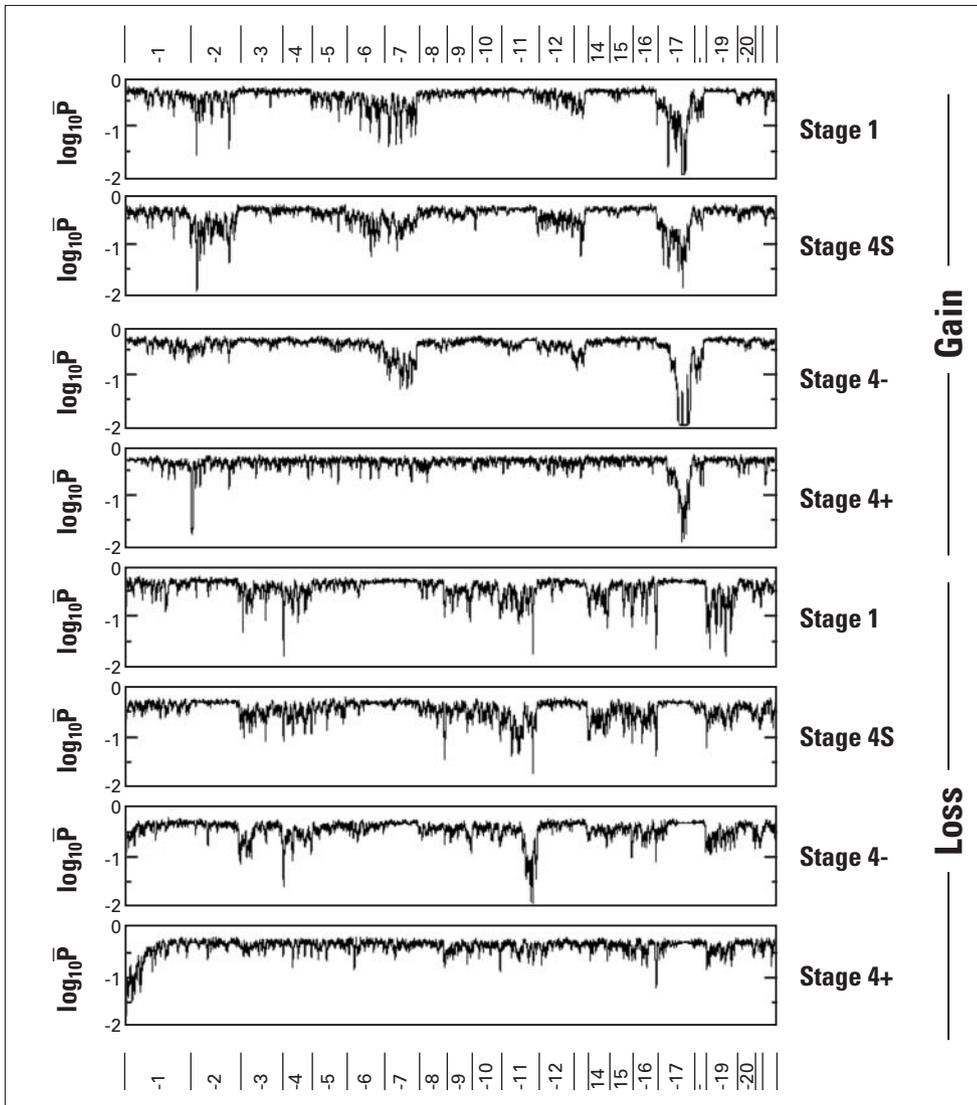


Fig 2. The average P value \bar{p} for gains and losses detected by CGH analysis for neuroblastoma is proportional to the frequency of genomic alterations (Eq.9). Lower \bar{p} indicates higher frequencies (ie, more samples show significant [small] P values). The presence of a gain or loss in all samples is indicated by a very low \bar{p} . With the number of tumors used in this study it is safe to use $\log_{10}(\bar{p}) \leq -2$ as an indication of an alteration present in all samples. Chromosomal positions are indicated on the top and the bottom of the diagram. The position of the centromere is indicated by the small line within each chromosome.

occurring with a similar frequency. The daughter generation acquires additional genomic imbalances. (2) Unobserved intermediate genotypes are possible, but the model with the smallest number of genotypes (observed + unobserved) is utilized (Occam's Razor³⁶). (3) All tumor stages belonging to the same diagnostic group arise from a common ancestor (ie, the phylogeny has a root).

The first rule, the **signature of inheritance**, is the major key for model inference: Progression from one stage to a later stage manifests itself by a set of shared mutations present in both the parent and the offspring, plus changes specific to the offspring generation. To see how this can be used to identify models of tumor progression from genomic data, consider as an example the situation with only two stages, A and B (Fig 3). Only two distinct progression models compatible with the above rules are possible in this case: (I) linear progression from A to B (and, of course, the inverse B to

A, which is not counted as a distinct model because it is a mere re-labeling of stages) and (II) a progression from a common ancestor denoted C in Figure 3. Rule 1 predicts different patterns of genomic alterations for the two distinct models. In model I, all recurrent changes in A are present also in B and thus the changes in B are a true super-set of those in A. In model II, both A and B have recurrent alterations specific for each type but also alterations common to both types. In the latter case the common changes are associated with an unobserved stage C by rule 2. A hypothetical experiment designed to discriminate between the two models would need to identify whether both stages A and B have genomic alterations specific to the two stages.

For more than two stages, the basic principle of model inference remains the same: Each of the possible tumor progression models generates a unique distinct pattern of common, shared (between two or more stages) and

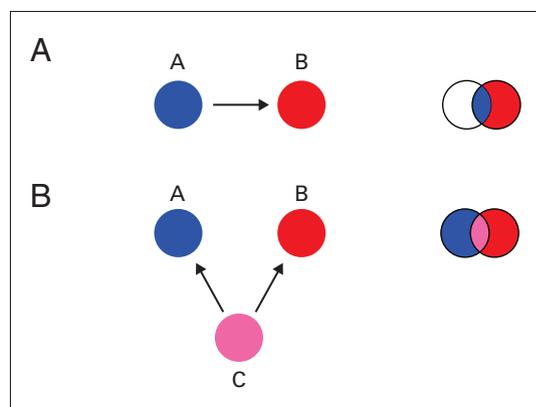


Fig 3. The two-tumor progression models for two observed phenotypes compatible with the assumption of an evolutionary progression. Arrows indicate accumulation of mutations, nodes observed, and unobserved phenotypes, respectively. Each of the progression models generates a unique pattern of mutations present in one or both phenotypes. Diagrammatically the distinct configurations can be depicted by the Venn diagrams shown on the right part of the diagram.

unique genomic alteration. The number of progression models grows quickly with the number N of stages and the number S of distinct sets in the (abstract) Venn diagram is given by

$$S = \sum_{n=1}^N \frac{N!}{(N-n)!n!} \quad (9)$$

generating 2^S possible experimental outcomes. The number of topologically distinct models compatible with evolutionary progression is smaller because of symmetries and experimental outcomes incompatible with evolutionary progression.

For the four NB stages the (abstract) Venn diagram has 14 distinct sets: four specific to each type, six shared between two stages, three shared between three stages and one set containing common alterations. In order to test whether these sets are occupied or empty in our NB data set, we used the statistical model described in the Materials and Methods section. In brief, P values were calculated for each clone to be a member of one of the 14 sets with a random permutation test. If the observed number of clones with $P < \theta$ in one of the 14 sets was found to be considerably larger than what one would expect by chance, the respective set was defined occupied. In order to avoid a strong dependence on the choice of the threshold θ , we repeated this step for various θ (Fig 4). We found that S4-, S4S, and S4+ have unique alterations, while alterations specific for S1 were not found (Fig 4A). Besides the specific alterations and the changes common to all tumors, S4+ had no further alterations in common with any of the other stages. At the same time, S1, S4-, and S4S had common alterations not present in S4+ (Fig 4C). This allows us to conclude that S4+ resides in a separate branch of the NB progression model. The alterations in S4S were found to be a true super-set of the S1 alterations (Figs 4A and 4B), which

allows one to conclude that S4S progresses from S1. Some of the alterations common to S1 and S4S were absent in S4-, which had its own specific changes. This allows one to conclude that S4- is an independent sub-branch of the group S1, S4-, and S4S. Figure 5 summarizes these findings and depicts the final inferred model of tumor progression in NB. A list of the major genomic regions and their genomic positions appears in Table 1.

DISCUSSION

In this article, we have presented an unbiased approach to learn models of tumor progression from genomic data. Here we have used DNA copy-number data in 76 NB samples to infer a progression model involving S1, S4S, and 4 with and without *MYCN* amplification. The key to our analysis that translated the flat data to a rich biologic model was the integration of the concept of clonal evolution of cancer in the data analysis. This permitted the use of the biologic principles of inheritance to establish a link between possible theoretical tumor progression models and the experimental observation whether recurrent genomic alterations were specific to, shared between a few, or common to all subtypes of a specific cancer. This is analogous to the analysis of the phylogeny of species, similarity of features (here, mutations) in different cancer species (phenotypes) were used to establish inheritance (progression). One important difference, though, is that individual cancer specimens are in fact part of different evolutionary processes: Each tumor has developed independently in each patient. The justification to nonetheless apply the concept of inheritance is (1) that the individual evolutionary processes start from an identical (or at least similar) population of normal cells and (2) that recurrent genomic alterations exist. The latter fact indicates that very similar selective forces act in the individual processes of clonal evolution and it is exactly these selective forces (mutation of gene X provides growth advantage) that defines a model of tumor progression.

Clonal evolution is an inherently random process, often with multiple alternative mutations affecting the same intracellular pathway. For example, in colorectal cancer, one of the best-studied systems, the *APC* gene was found to be mutated in 85% of the cases, whereas in half of the remaining specimens, *β -catenin*, a downstream target of *APC* in the same pathway, is mutated.^{4,37} This inherent biologic variability needs to be incorporated in the analysis of tumor progression for any cancer. Consequently we used a probabilistic language; the analysis of the frequency, or probability, of genomic alterations allowed us to identify the dominant pathways of genomic alterations. To define frequencies in a sensible way, it was necessary to categorize the cancers into classes. We used staging and *MYCN* status for this purpose. A discovery of subclassifications from the data is therefore not possible with our

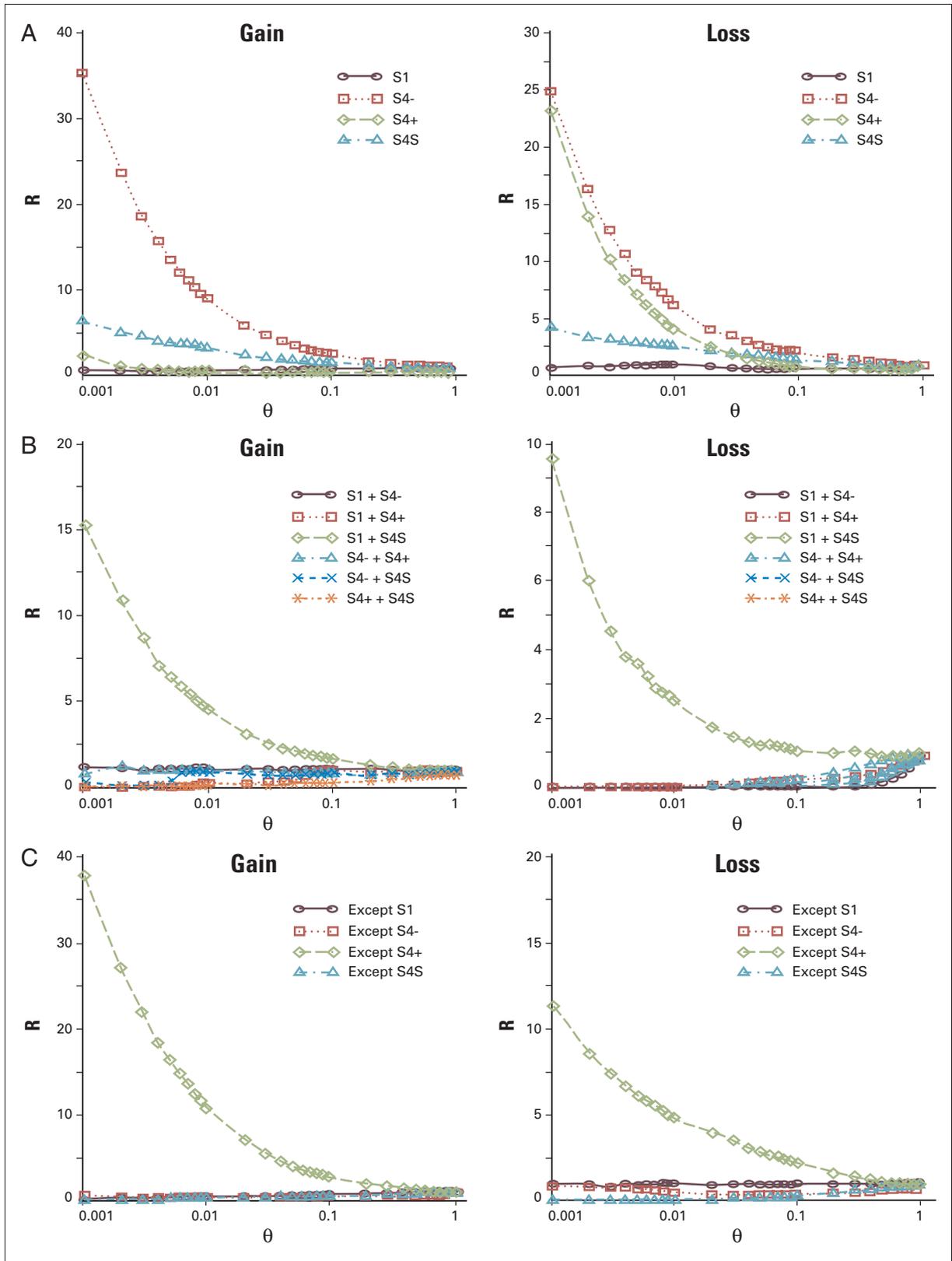


Fig 4. Illustrates of the method used to identify recurrent genomic alterations that are unique or shared between two or three out of the four phenotypes. Only if the observed number of such regions is larger than the expected by the false discovery rate (FDR), the corresponding subset in the Venn diagram is said to be occupied. Shown is the ratio R (observed/expected by FDR) for several P value thresholds θ . A ratio R considerably larger than one for any θ indicates that a particular set is occupied: (A) recurrent regions unique to a stage; (B) regions shared between two stages; and (C) regions shared between three stages.

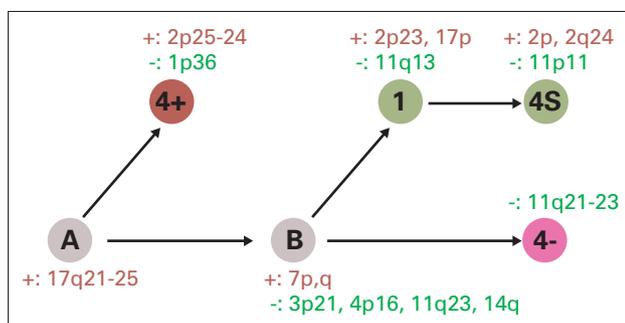


Fig 5. Graphical representation of the inferred tumor progression model for neuroblastoma. Stage 1, 4S, 4-, and 4+ and two intermediate stages A and B are depicted as circles. Arrows indicate accumulation of mutations. Major affected regions are shown in the figure where a + sign indicates gain and - sign loss of DNA material.

method. **Unsupervised methods** such as those presented in Desper, Jiang, and Kallioniemi³⁸ and von Heydebreck, Gunawan, and Fuzesi³⁹ are better suited for this purpose; however, typically they require a much larger data set in order to get robust results.

We found pronounced, distinct patterns of genomic alterations associated with different stages of NB. Our analysis confirmed many of the known regions as well as their association with the different NB stages. Loss of 1p36, for example, has been associated²⁰ with advanced stages and most frequently occurs together with *MYCN* amplification. We found that loss of 11q is inversely correlated with *MYCN* amplification in agreement with various studies (eg, Guo et al¹⁰ and Plantaz et al¹¹). Also, the losses on 3p, 4p, and 14q were found earlier to be correlated with the loss of 11q and exclusively in tumors without *MYCN* amplification.¹²

Table 1. Major Regions Found to Be Associated to One of the 14 Distinct Sets of the Venn Diagram for the Four Stages in NB

| Type | Classification | Cytoband | Start | End |
|------|----------------|-------------|-------|-----|
| LOH | Specific 4+ | 1p36 | 4 | 12 |
| LOH | Specific 4+ | 1p36-1p34 | 20 | 40 |
| GAIN | Specific 4+ | 2p25 | 11 | 21 |
| LOH | Specific 4- | 11q21-11q23 | 92 | 117 |
| GAIN | Specific 4S | 2p22 | 33 | 37 |
| GAIN | Specific 4S | 2p21 | 45 | 48 |
| GAIN | Specific 4S | 2q24 | 151 | 154 |
| LOH | Specific 4S | 11p11 | 45 | 48 |
| GAIN | Shared S1, S4S | 2p23 | 31 | 33 |
| GAIN | Shared S1, S4S | 17p13 | 3 | 11 |
| GAIN | Shared S1, S4S | 17p11 | 26 | 29 |
| LOH | Shared S1, S4S | 11q13 | 65 | 67 |
| GAIN | All but 4+ | 7p22-14 | 5 | 37 |
| GAIN | All but 4+ | 7p11-q11 | 54 | 65 |
| GAIN | All but 4+ | 7q21-q31 | 77 | 123 |
| GAIN | All but 4+ | 7q34-ter | 138 | 158 |
| LOH | All but 4+ | 3p21 | 39 | 50 |
| LOH | All but 4+ | 4p16 | 0 | 7 |
| LOH | All but 4+ | 11q23 | 117 | 121 |
| LOH | All but 4+ | 14q11 | 19 | 24 |
| LOH | All but 4+ | 14q24-33 | 75 | 102 |
| GAIN | Common | 17q21-25 | 32 | 77 |

NOTE. Shown here are regions with $P < 0.001$ covering at least 10 sequential clones on the DNA array. Start and end positions are given in Megabase with respect to the *p*-terminus.
Abbreviations: NB, neuroblastoma; LOH, loss of heterozygosity.

The loss on 4p16, which we found frequently in S1, S4S, and S4-tumors, may deserve special attention. Several studies suggested the presence of a tumor-suppressor gene.^{23,25} A weak linkage of LOH on 4p16 to familial NB predisposition has been described,²³ and *Phox2b* located close by on 4p15 has been identified as a marker for neuroblastoma.⁴⁰

The observed association of recurrent DNA copy number changes with the different stages of NB allowed us to map these data to identify the one tumor progression model that summarizes our data (Fig 5). In this diagram, nodes represent the different phenotypes (S1, S4S, S4-, and S4+) and two intermediate, unobserved stages denoted A and B. Arrows represent accumulation of genomic alterations. They do not necessarily imply a specific temporal order of events. Technically we cannot draw a conclusion on a specific sequence of events from our data, and the represented graph should therefore be interpreted as a decision tree. Biologically, it nevertheless seems reasonable that the shared genomic alterations represented by the intermediate, unobserved stages tend to occur early in NB progression. An interesting speculation is that the first node represents the genomic changes found in the neuroblastic modules, which resemble NB in situ that commonly occurs in infants younger than 3 months who die of other causes.⁴¹ Unless additional hits like *MYCN* amplification and loss of 1p36 occur, these in situ genomic alterations would ordinarily result in apoptosis during normal development, explaining why these phenotypes are not observed. It is encouraging to note that the inferred model predicts clinical as well as pathologic features of NB, even though it was solely derived from molecular data. With the exception of S1, all phenotypes in this study are end points of the progression model. This implies that a fully developed NB clone does not progress to a more aggressive tumor. This prediction is in very good agreement with clinical evidence, where it was found that NB rarely, if ever⁴² progresses to a more aggressive tumor. Our model is thus in agreement with a hypothesis suggested by several authors^{27,31} that the aggressive stages of NB (stage 4 with and without *MYCN* amplification) are created as advanced stages. The markedly different behavior of the favorable-prognosis, benign, metastatic S4S disease compared with the other metastatic variants of stage four NB is partially explained by the predicted model: The S4S disease is not a variant of the other metastatic stages but rather progresses from the benign S1 disease. This prediction of our model supports the hypothesis that NB consists of at least two distinct clinical-biologic types.²⁶ Of note is also that the pathologic description of stage S4S as a localized primary tumor (as defined for stage 1, 2A or 2B) with limited dissemination seems to be a possible consequence predicted by our model. The dominant pattern of ploidy in NB²⁸ with near triploid clones predominantly in less aggressive tumors and

diploid tumors in more aggressive stages is reflected in this model: the dominantly triploid stages S1 and S4S appear in a common branch of the model, even though we did not explicitly incorporate ploidy information.

As mentioned in the Introduction, ploidy is considered to play a prominent role in NB biology and therefore deserves some more attention. In principle the ploidy information can be treated in the same way as the DNA copy-number information by classifying it as common or specific for the different tumor stages. Unfortunately, CGH techniques do not allow inference of ploidy and we did not have available the ploidy data for all of our tumors to formally use it in the model-building process. Nevertheless, it is interesting to see that the inferred model is compatible with the empirical models built around ploidy by using their implicit ploidy pattern. One popular variant³² essentially classifies NB in aneuploid lower-stage tumors with overexpressed *TrkA* and aggressive diploid tumors with overexpressed *TrkB*. In this case, aneuploidy is shared by S1 and S4S, and thus ploidy change would occur at node S1. Another, more recent speculative model²⁷ describes ploidy instability as an event common to all tumors. This would lead to the conclusion that ploidy change happens in the first intermediate stage A. A third speculative alternative is that ploidy change occurs at the intermediate stage B. This situation would generate the testable hypothesis that tetraploid tumor cells are significantly more frequent in S4- tumors without *MYCN* compared with S4+ with *MYCN* amplification.

DNA copy-number changes alone cover only a fraction of possibilities for cancer cells to acquire mutations. Few-nucleotide mutations or aberrant patterns of promoter methylation are technically more difficult to monitor but are an important pathways of mutation. The probabilistic analysis presented in this article can in principle be used for these more subtle mutations. However, currently no sufficiently high-throughput method is available to actually identify these mutations on a genome-wide scale. One example in which DNA copy-number data alone are not sufficient to infer a tumor progression model is colorectal cancer. In this case, no association between the stage of the tumor and a particular pattern of gains and losses could be identified.⁴³ An important question is how progression models would change when new information becomes available. As with any model, new information may make changes necessary. However, this new information would not invalidate currently identified

shared alterations and rather add additional shared or unique alterations. Additional data would consequently leave the basic structure of the model unchanged and at most induce additional arrows or unobserved phenotypes.

In summary, we have described and tested a method that formalizes the inference of models of tumor progression from genomic data. The task to develop such models intuitively gets increasingly complex with an increased amount of information available and with a certain level of complexity a formalized method for model inference may become a necessity. We have tested our method on DNA copy number data for neuroblastoma, but it is limited neither to cytogenetic data nor to this specific cancer type. The inherent platform independence allows one to integrate existing epidemiologic data in larger studies. This will be useful in the next step in our analysis of the NB progression model to include the intermediate clinical stages left out in this study. Here we had limited our analysis to the least and most aggressive stages of neuroblastoma, expecting that genotypic differences between these types would be most pronounced. We conclude that our progression model inferred from genomic data is compatible with currently proposed progression models centered around ploidy changes. The model reflects the heterogeneity of clinical behavior found in different NB stages. Our genomic data did not support a linear model of tumor progression from the least to most aggressive phenotypes except for S4S tumor for which we speculate an evolution or progression from S1 tumors. Additionally we believe the identified common, shared, and unique regions will harbor genes that will help to get important clues as to the causal reasons for tumor phenotype and progression.

Acknowledgment

We thank John Maris, MD, Children's Hospital of Philadelphia (Philadelphia, PA), Steven Qualman, MD, at the Cooperative Human Tissue Network and the Children's Oncology Group for several of the tumors. We thank Peter Ambros, MD, and Craig Whiteford, MD, for helpful discussions. Jun Wei, MD, and Braden Greer provided technical support.

Authors' Disclosures of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

REFERENCES

1. Knudson AG: Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68:820-823, 1971
2. Fearon ER, Vogelstein B: A genetic model for colorectal tumorigenesis. *Cell* 61:759-767, 1999
3. Kinzler KW, Vogelstein B: Gatekeepers and caretakers. *Nature* 386:761-763, 1997
4. Morin PJ, Sparks AB, Korinek V: Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* 275:1787-1790, 1997
5. Nowell PC: The clonal evolution of tumor cell populations. *Science* 194:23-28, 1976
6. Mittelman A: Tumor etiology and chromosome pattern. *Science* 176:1340-1341, 1962
7. Forozan F, Karhu R, Kononen J, et al: Genome screening by comparative genomic hybridization. *Trends Genet* 13:405-409, 1997
8. Pollack JR, Perou CM, Alizadeh AA, et al: Genome-wide analysis of DNA copy-number

changes using cDNA microarrays. *Nat Gen* 23:41-46, 1999

9. Hyman E, Kauraniemi P, Hautaniemi S, et al: Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res* 62:6240-6245, 2002

10. Guo C, White PS, Weiss MJ, et al: Allelic deletion at 11q23 is common in *MYCN* single copy neuroblastomas. *Oncogene* 18:4948-4957, 1999

11. Plantaz D, Vandesompele J, van Roy N, et al: Comparative genomic hybridization (CGH) analysis of stage 4 neuroblastoma reveals high frequency of 11q deletion in tumors lacking *MYCN* amplification. *Int J Cancer* 91:680-686, 2001

12. Oude Luttikhuis MEM, Powell JE, Rees SA, et al: Neuroblastomas with chromosome 11q loss and single copy *MYCN* comprise a biologically distinct group of tumors with adverse prognosis. *Br J Cancer* 85:531-537, 2001

13. Chen QR, Bilke S, Wei JS, et al: cDNA Array-CGH profiling identifies genomic alterations specific to stage and *MYCN*-amplification in neuroblastoma. *BMC Genomics* 5:70, 2004

14. Brodeur GM: Neuroblastoma: Biological insights into a clinical enigma. *Nat Rev Cancer* 3:203-216, 2003

15. Brodeur GM, Azar C, Brother M, et al: Neuroblastoma: Effect of genetic factors on prognosis and treatment. *Cancer* 70:1685-1694, 1992

16. Brinkschmidt C, Christiansen H, Terpe HJ, et al: Comparative genomic hybridization (CGH) analysis of neuroblastomas—An important methodological approach in pediatric tumor pathology. *J Pathol* 181:394-400, 1997

17. Cohen N, Betts DR, Trakhtenbrot L, et al: Detection of unidentified chromosome abnormalities in human neuroblastoma by spectral karyotyping (SKY). *Genes Chromosomes Cancer* 31:201-208, 2001

18. Schleiermacher G, Janoueix-Lerosey I, Combaret V, et al: Combined 24-color karyotyping an comparative genomic hybridization analysis indicates predominant rearrangements of early replicating chromosome regions in neuroblastoma. *Cancer Genet Cytogenet* 141:32-42, 2003

19. Schwab M, Varmus HE, Bishop JM: Human N-myc gene contributes to neoplastic

transformation of mammalian cells in culture. *Nature* 316:160-162, 1985

20. Brodeur GM, Green AA, Hayes FA, et al: Cytogenetic features of human neuroblastomas and cell lines. *Cancer Res* 41:4678-4686, 1981

21. Maris J, Matthay KK: Molecular biology of neuroblastoma. *J Clin Oncol* 17:2264-2279, 1999

22. Vandesompele J, Speleman F, van Roy N, et al: Multicentre analysis of patterns of DNA gains and losses in 204 neuroblastoma tumors: How many genetic subgroups are there? *Med Pediatr Oncol* 36:5-10, 2001

23. Perri P, Longo L, Cusano R, et al: Weak linkage at 4p16 to predisposition for human neuroblastoma. *Oncogene* 21:8356-8360, 2002

24. Perri P, Bachetti T, Longo L, et al: *PHOX2B* mutations and genetic predisposition to neuroblastoma. *Oncogene* 24:3050-3053, 2005

25. Caron H, van Sluis P, Buschman R, et al: Allelic loss of the short arm of chromosome 4 in neuroblastoma suggests a novel tumor suppressor gene locus. *Hum Genet* 97:834-837, 1996

26. Woods WG, Tuchman M, Bernstein ML, et al: Screening for neuroblastoma in North America: 2-year results from the Quebec Project. *Am J Pediatr Hematol Oncol* 14:312-319, 1992

27. Westermann F, Schwab M: Genetic parameters of neuroblastomas. *Cancer Lett* 184:127-147, 2002

28. Look AT, Hayes FA, Shuster JJ, et al: 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, 1991

29. Ladenstein R, Ambros IM, Potschger U: Prognostic significance of DNA di-tetraploidy in neuroblastoma. *Med Pediatr Oncol* 36:83-92, 2001

30. Ambros IM, Zellner A, Roald B, et al: Role of ploidy, chromosome 1p, and Schwann cells in the maturation of neuroblastoma. *N Engl J Med* 334:1505-1511, 1996

31. Kaneko Y, Knudson AG: Mechanism and relevance of ploidy in neuroblastoma. *Genes Chromosomes Cancer* 29:89-95, 2000

32. Brodeur GM: The concept of tumorigenesis in neuroblastoma, in Brodeur GM, Sawada T, Tsuchida Y et al (eds): *Neuroblastoma*. Amsterdam, the Netherlands, Elsevier Science BV, 2000

33. Churchill GA: Using ANOVA to analyze microarray data. *Biotechniques* 37:173-177, 2004

34. Alter O, Brown PO, Botstein D: Singular Value Decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97:10101-10106, 2000

35. Bilke S, Chen QR, Whiteford CC, et al: Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. *Bioinformatics* 21:1138-1145, 2005

36. Smith TF: Occam's Razor. *Nature* 285:620, 1980

37. Rajagopalan H, Nowak MA, Vogelstein B, et al: The significance of unstable chromosomes in colorectal cancer. *Nat Rev Cancer* 3:695-701, 2003

38. Desper R, Jiang F, Kallioniemi OP, et al: Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol* 6:37-51, 1999

39. von Heydebreck A, Gunawan B, Fuzesi L: Maximum likelihood estimation of oncogenetic tree models. *Biostatistics* 5:545-556, 2004

40. Son CG, Bilke S, Davis S, et al: Database of mRNA gene expression profiles of multiple human organs. *Genome Res* 15:443-450, 2005

41. Beckwith J, Perrin E: In situ neuroblastomas: A contribution to the natural history of neural crest tumors. *Am J Pathol* 43:1089-1104, 1963

42. Brodeur GM, Maris JM, Yamashiro DJ, et al: Biology and genetics of human neuroblastomas. *Pediatr Hematol Oncol* 19:93-101, 1997

43. Nakao K, Mehta KR, Fridlyand J, et al: High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis* 25:1345-1357, 2004

44. Chen Y, Dougherty ER, Bittner ML: Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 2:364-374, 1997

45. Bilke S, Breslin T, Sigvardsson M: Probabilistic estimation of microarray data reliability and underlying gene expression. *BMC Bioinformatics* 4:40, 2003