Construction and analysis of tree models for chromosomal classification of diffuse large B-cell lymphomas

Hui-Yong Jiang, Zhong-Xi Huang, Xue-Feng Zhang, Richard Desper, Tong Zhao

Hui-Yong Jiang, Xue-Feng Zhang, Department of General Surgery, General Hospital of Shenyang Military Region, Shenyang 110016, Liaoning Province, China

Zhong-Xi Huang, Tong Zhao, Department of Pathology, Nanfang Hospital, Southern Medical University, Guangzhou 510515, Guangdong Province, China

Richard Desper, Department of Biology, University College London, Darwin Building, Gower Street, London, WC1E 6BT, United Kingdom

Supported by Science and Technology Project of Guangzhou, No. 2002Z3-E4016; No. B30101, China

Co-correspondence to: Tong Zhao

Correspondence to: Richard Desper, Department of Biology, University College London, Gower Street, London WC1E 6BT, United Kingdom. r.desper@ucl.ac.uk

Telephone: +86-20-62787274 Fax: +86-20-61642381

Received: 2006-11-23   Accepted: 2007-03-15

Abstract

AIM: To construct tree models for classification of diffuse large B-cell lymphomas (DLBCL) by chromosome copy numbers, to compare them with cDNA microarray classification, and to explore models of multi-gene, multi-step and multi-pathway processes of DLBCL tumorigenesis.

METHODS: Maximum-weight branching and distance-based models were constructed based on the comparative genomic hybridization (CGH) data of 123 DLBCL samples using the established methods and software of Desper *et al*. A maximum likelihood tree model was also used to analyze the data. By comparing with the results reported in literature, values of tree models in the classification of DLBCL were elucidated.

RESULTS: Both the branching and the distance-based trees classified DLBCL into three groups. We combined the classification methods of the two models and classified DLBCL into three categories according to their characteristics. The first group was marked by +Xq, +Xp, -17p and +13q; the second group by +3q, +18q and +18p; and the third group was marked by -6q and +6p. This chromosomal classification was consistent with cDNA classification. It indicated that -6q and +3q were two main events in the tumorigenesis of lymphoma.

CONCLUSION: Tree models of lymphoma established from CGH data can be used in the classification of DLBCL. These models can suggest multi-gene, multi-step and multi-pathway processes of tumorigenesis. Two pathways, -6q preceding +6q and +3q preceding +18q, may be important in understanding tumorigenesis of DLBCL. The pathway, -6q preceding +6q, may have a close relationship with the tumorigenesis of non-GCB DLBCL.

Jiang HY, Huang ZX, Zhang XF, Desper R, Zhao T. Construction and analysis of tree models for chromosomal classification of diffuse large B-cell lymphomas. *World J Gastroenterol* 2007; 13(11): 1737-1742

http://www.wjgnet.com/1007-9327/13/1737.asp

INTRODUCTION

Evidence has indicated that tumorigenesis is a multi-step process in humans. It is believed that genetic alterations occur in tumor genomes at multiple locations progressively from subtle mutations to alterations of entire chromosomes. Desper *et al*[1] have proposed tree models to define multi-step and multi-pathway processes of tumorigenesis based on analysis of comparative genomic hybridization (CGH) data of chromosome alterations. These methods have been applied successfully on several tumors. However, to our knowledge, there has been no report about tree models of diffuse large B-cell lymphoma (DLBCL). DLBCL is the most common type of malignant lymphoma, accounting for approximately 30%-59%[2-4] in adult NHL. By establishing tree models, we have tried to classify DLBCL using chromosome copy number alterations, and explored the multi-gene, multi-step and multi-pathway processes of DLBCL tumorigenesis.

Based on the differences of gene expression profiles, DLBCL can be divided into two classes: germinal center B-cell-like (GCB) and non-GCB. The latter class includes activated B-cell-like (ABC) and type 3 lymphomas. The prognosis of the CGB subtype is better than non-GCB[5,6]. We compared the tree model classification of DLBCL using chromosome copy number alterations with a classification by gene expression profiling.

MATERIALS AND METHODS

Materials

CGH data published in four previous studies were

analyzed. The selection criteria are as follows: (1) all the specimens are primary tumors; (2) all the specimens were fresh frozen tissues; and (3) a normal ratio range of chromosome fluorescence signal (0.8-0.85, 1.15-1.25) was predefined. CGH data of 173 DLBCL cases[7-10] from these studies were collected, including 138 cases of primary tumors. Karnan et al[7] analyzed 26 cases of CD5+ and 44 cases of CD5- tumors. CD5+ tumors covered 10%-20% of DLBCL. We randomly selected 11 cases of CD5+ tumors and 44 cases of CD5- tumors so as to make the proportions of the data reflect the real status. We collected 123 cases of CGH data, of which 116 had chromosome abnormalities.

Selection of nonrandom events

Chromosomal alteration data of various regions of the genome were recorded according to chromosome arms. Abnormalities from chromosomes 19, 22 and Y, and chromosomal arms 1p, 13p, 14p, 15p, 16p and 21p, were excluded because of the high false positive rates in GCH analysis. We used the established method of Brodeur et al[11] to select nonrandom events. This method presumes a prior probability distribution, whereby the gains and losses of each overlap region occur independently with individual probabilities proportional to the lengths of the various chromosome arms. Random simulations were used to generate 10 000 replicates of the distribution. For each replicate, a score was computed, relating each event to its prior probability, and the maximal scores were recorded for each event. Scores were also computed for each of the events in the data set, and an event was considered nonrandom if its score in the real data exceeded the 95th percentile of the 10 000 maximal scores from the random replicates.

Maximum-weight branching models

Desper et al[1,12] proposed the construction of a tree model of oncogenesis based on a computational technique known as maximum-weight branching, and we refer to the models based on this method as "maximum-weight branching tree models". A model of this type introduces an artificial node to root the branching (an event occurs in all samples), and the hereditary events are represented by the other nodes of the branching. An edge (i, j) in a tree model represents a probabilistic cause-and-effect relationship between a pair of events: the occurrence of event i makes the occurrence of j more likely. The branching tree model is a generalization of the Vogelstein path model. A tree model allows multiple edges to come from each node representing different possible pathways for oncogenesis while a path model only has one edge from each node. In the branching construction process, edges were included into the model tree based on a weight function which takes into account the number of occurrences of each event and the number of co-occurrences of each pair of events. The weight function used is $w_{ij} = \log \hat{p}_{ij} - \log(\hat{p}_i + \hat{p}_j) - \log \hat{p}_j$, where $\hat{p}_i$ is the observed probability of event i, $\hat{p}_j$ is the observed probability of event j, and $\hat{p}_{ij}$ is the observed probability of the co-occurrence of events i and j in the same tumor.

Desper et al[1] found that, under plausible assumptions about the stochastic process of oncogenesis and with enough samples, the tree produced by the maximum weight branching method converges to the correct tree model. Although in practice we generally do not have enough tumor samples for the convergence result, we would still expect that the maximum weight branching tree is correct in most of the edges. We used the program oncotrees designed by Desper to establish the maximum-weight branching and the associated tree models.

Distance-based trees

In addition to the maximum-weight branchings, Desper et al[1] constructed tree models of oncogenesis using methods borrowed from the field of phylogenetics. They are known generally as distance methods, since each branch in the tree has a length assigned to it derived from the relative probabilities of the nodes on the edge. A distance-based tree model regards all the events as leaves of a tree, while the internal nodes of the tree represent latent and unknown events. This approach is similar to the techniques used in the construction of phylogenetic tree models, which regard the known species as leaves while the inferred common ancestors are represented by internal nodes. In the figures showing distance trees, the horizontal distance of each edge is proportional to its distance.

The first step in the construction of such a tree is to define a distance matrix on the set of events (including the artificial "root" event in each sample). Given two events i and j, the distance between event i and event j was set to:

$$d(i, j) = -2 \log \hat{p}_{ij} + \log \hat{p}_i + \log \hat{p}_j$$

where $\hat{p}_i$ $\hat{p}_j$ and $\hat{p}_{ij}$ are defined as in the previous section. Next, established phylogenetic algorithms were used to search the space of possible trees to find a tree whose associated metric fits best the input matrix.

The algorithms used included the programs Fitch[13] and Neighbor[14] from the PHYLIP package (3.5c edition), and the minimum evolution software FastME[15]. The results were further optimized using dynamic programs. The resulting models may provide a statistically robust answer to the basic questions: (a) Which is the early event? Those events are near the root; (b) Which event(s) marks subclasses of tumors? Those events cluster together in subtrees.

We also used a resampling method to test the persistence of the classification signal measured by the distance methods. We resampled the data by including each tumor in each replicate with probability 0.5, with all choices across tumors and replicates being made independently. This sampling was done without replacement, and did not guarantee the resulting sample size. Each resampling produces a subset of the tumors containing approximately half the original set. We performed the distance analysis on each subset, producing a distance tree for each replicate. We then used the PHYLIP program consense to form a consensus tree.

Likelihood trees
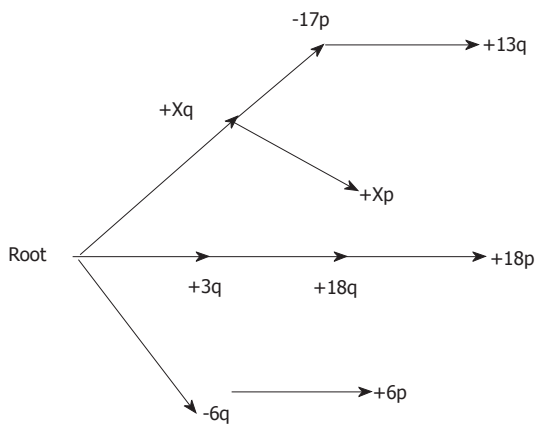
We applied the method of von Heydebreak et al[16] to find

Figure 1 Branching for small clique.

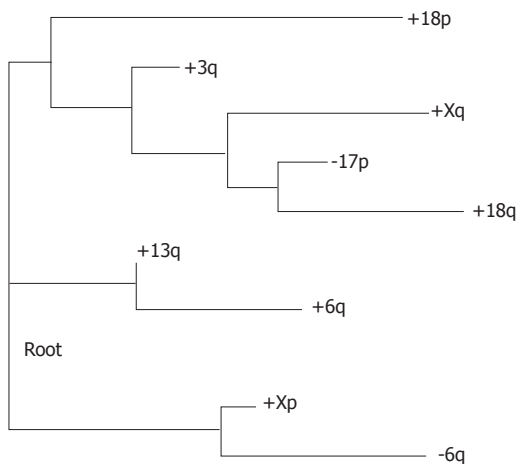**Figure 2** Consensus distance tree.

Figure 3 Maximum likelihood tree.

the maximum likelihood oncogenetic tree model for the data. The tree search method has been implemented as a software package for the statistical programming language R. The method uses a heuristic to search across topologies, seeking to maximize the likelihood of the data, given the tree topology and parameters. If the data is generated by an oncogenetic tree process, the resulting tree should be similar or identical to the tree produced by the distance methods.

RESULTS

Gains in DLBCL by CGH were most often seen on chromosome arms Xq (46.6%), Xp (44%), 3q (28.4%), 18q (27.6%), 1q (25%), 13q (21.6%), 12q (20.7%), 7q (19.8%), 3p (18.1%), 6p (18.1%), 7p (17.2%), 12p (15.5%), 11q (15.5%), 5q (14.7%), 9p (14.7%), 5p (8.6%), 2p (8.6%), 18p (7.8%) Losses were most often seen at 17p (36.2%), 6q (25.9%), 9q (15.5%), 8p (13.8%). It is common to model the occurrence of random chromosomal aberrations as chance events that happen with equal likelihood at any point in the genome. Thus, because the probability of a random aberration on a chromosome is related to its length, we conclude that the events most often seen in DLBCL were not all nonrandom events. The method of
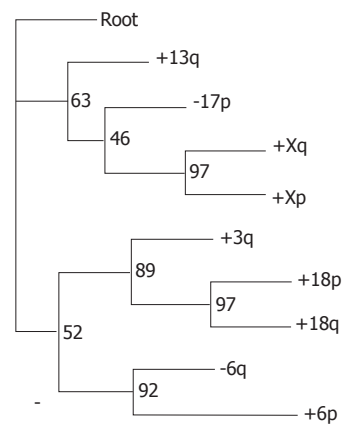
Brodeur *et al*[11] was used to select the following thirteen events as nonrandom in DLBCL cases: -6q, -8p, -17p, +3q, +6p, +7p, +9p, +12p, +13q, +18p, +18q, +Xp and +Xq. A plus symbol (+) indicates a gain of a chromosomal region and a minus symbol (-) a loss. We further confined our attention to non-random events, eliminating events -8p, +7p, +9p and +12p, which occurred less than 20% of the time. The gains +18p and +6p were retained to examine their relationship to +18q and -6q, respectively.

Both the branching tree and the distance tree classified DLBCL into subgroups. In branching tree (Figure 1), the first group was marked by +Xq, +Xp, -17p and +13q; the second group by +3q, +18q and +18p; and the third group was marked by -6q and +6p. The distance tree shows the same subdivisions (Figure 2), and an additional classification is closely related between the second and third groups. The values at the internal nodes represent the number of replicates that contained the subtree to the right of the node out of 100. Several subtrees have respectably high numbers, especially the three subtrees corresponding to pairs of events from the same respective chromosomes, as well as the +3q/+18 subtree.

The likelihood tree (Figure 3) did not show this classification. Curiously, the likelihood tree did not group together the three pairs of events from the same respective chromosomes. This absence suggests that the likelihood tree should be kept as an indication that restraint is needed for any conclusions by the high-weight branching and the distance tree in stead of suggesting any alternative hypothesis.

DISCUSSION

CGH[17] and microarray[18] techniques can be applied to examine the alterations and abnormalities throughout the whole genome. In tumor cells, the abnormalities of various chromosomes may tend to occur in certain orders, or to co-occur in certain groups, as suggested in Vogelstein's[19] linear path model. However, in late stage cancers, abnormalities of chromosomes may occur extensively, and many of these abnormalities appear to occur randomly and independently from the progression of the cancers because of the general phenomenon known as genome instability. For these reasons, Desper *et al*[11,12] used the well-
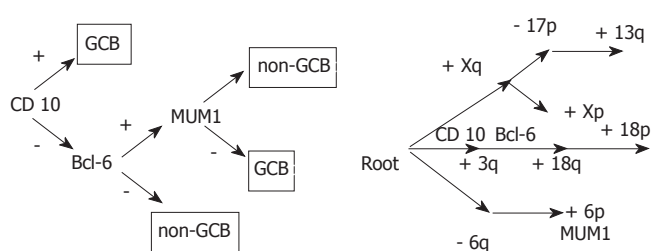
Figure 4 Distribution of genes in tree models used in molecular classification in lymphoma by IHC. Diagram in upper left is the classification of DLBCL by IHC. Adapted from Hans CP *et al*. Blood 2004; 103: 275-282.

known method of Brodeur *et al*[11] to select abnormalities of chromosomes that are associated with progression of cancers, and developed two methods, distance-based trees and branching trees, to determine the co-occurrence relationships and the typical orders in which various events occur. Tree models of the multi-step and multi-pathway processes of oncogenesis were constructed based on the individual and pairwise probabilities of events. These methods have previously been used to construct models for renal cell carcinoma[20], breast cancer[21], bladder cancer[22], and head-and-neck carcinomas[23]. Early events and chromosomal classification of oncogenesis of these carcinomas were analysed. Yin Z *et al*[24] established tree models of chromosome arms of nasopharyngeal carcinoma (NPC) and colorectal carcinoma. Based on the models of chromosome arms, we[25] established the tree models of chromosome "overlap regions", analyzed CGH data of NPC and constructed tree models of the multi-step and multi-pathway process of NPC carcinogenesis, and classified NPC based on CGH data. To date, there has been no other established tree model of DLBCL.

With the development of gene chips, DLBCL study at the molecular level has become a very active field since 2000. Alizadeh *et al*[5] first found that, using a cDNA microarray, DLBCL can be divided into two prognostically distinct subgroups: germinal center B-cell-like (GCB) and activated B-cell-like (ABC) lymphomas. The GCB subgroup has a significantly higher survival rate than the ABC subgroup. Rosenwald *et al*[6] found a third group (Type 3) that had a poor outcome in a manner similar to the ABC group according to statistical analysis. Another study using an oligonucleotide array conducted by Shipp *et al*[26] demonstrated that DLBCL could be divided into 2 molecularly distinct populations (cured and fatal/refractory). Although both fresh and frozen specimens can be used to facilitate the extraction of enough RNA in microarray detection, this technique cannot be applied widely because of the high price of chips. Using the cDNA microarray results as a gold standard, Hans *et al*[27] found that the expression of CD10, bcl-6 and MUM1 could form a combination to divide cases of DLBCL into GCB and non-GCB subgroups, with an outcome closely related to prognosis.

Using the method of Desper *et al*[1] , we have established tree models of chromosomal classification of DLBCL, which also divided DLBCL into three types. Does the chromosomal classification relate to molecular

classifications of DLBCL? We analyzed the distributions of CD10 (3q25.1-q25.2), BCL-6(3q27) and MUM1 (6p25-p23) in tree models adopted by Hans *et al*[27] in lymphoma classification (Figure 4).

In our study, chromosomal regions with the genes (CD10，MUM1，BCL-6) crucial to the classification of DLBCL are present in the tree model. 3q25.1-q25.2 and 3q27, the corresponding chromosomal regions of CD10 and BCL-6, are in the second category of the tree model. Chromosomal amplifications have characteristics of this type. 6p, the chromosomal region of MUM1, is in the third category. 6p showed amplification in this type. This indicated that there was correlation between tree model classifications and DLBCL subtypes.

Rosenwald *et al*[6,28] selected three genes closely correlated with Germinal-center B-cells using cDNA chips. DLBCLs with high expressions of these genes had better prognosis. They were bcl-6 (3q27) and the other two gene clones: IMAGE clone 1334260 and IMAGE clone 814622. Genes defined from the latter two clones are named GCET1 and GCET2[29,30], respectively. GCET1 is in chromosome 14q32 and GCET2 is located in chromosome 3q13.2. Though GCET1 is absent from the tree models, the other two genes both correspond to the second type in the chromosomal classification.

As indicated in both the branching and distance based trees, -6q is an early event in DLBCL. But the correlation between this chromosomal mutation and DLBCL is rarely reported and needs further examination. Our tree models seem to indicate that this chromosomal segment is of some importance to the carcinogenesis of non-GCB lymphoma. Detection of -6q may become a marker in evaluating the prognosis of lymphoma.

Identification of regions of minimal cytogenetic deletions (RCDs) has proven to be a useful predictor of candidate tumor suppressor genes (TSGs). Cigudosa *et al*[31] reported 11 RCDs in chromosomes 1, 3, 6, and 7, and suggested that these regions contain candidate TSGs. Chromosome 6 was the chromosome most frequently deleted in their series. Offit *et al*[32] proposed that deletions of 6q are associated with a lower CR and survival rate. It is widely believed that patients of non-GCB have a lower CR and survival rate[6]. The proposal that both deletions of 6q and patients of non-GCB are associated with a lower CR and survival rate is supported by the correlation between -6q and type non-GCB.

In 1990, Vogelstein proposed a chain of four events in colorectal carcinogenesis: first was the deletion of the fap gene in chromosome 5q; second, the mutation of K-ras in chromosome 12q; third, the deletion of the DCC (deleted in colorectal carcinoma) gene in 18q; and lastly, the mutation or deletion of p53 in chromosome 17p. Although similar pathways have long been sought in other types of tumors, such searches have been essentially fruitless because of the difficulty in getting pathological materials in each step. Our investigation suggests that deletions of tumor suppressor genes resulting from deletions of chromosome 6q may cause non-GCB DLBCL. As a result, a search of this chromosomal region may be of great help to find important lymphoma-related genes. Offit *et al*[32] reached the same conclusion by karyotyping.

By spectral karyotyping (SKY), Nanjangud] *et al*[33] found that, gain of 3 (53%), 7q (65%), 18q21 (41%) and loss of 6q11-13 (59%) were significantly more frequent in DLBCL, suggesting that gain and/or loss of genetic material might play a more proximal role in the development of this tumor. By a tree model analysis, we also found that +3q, +18q, -6q play some role in the tumorigenesis of DLBCL. Tree models can indicate successive relationships of these mutations. As a result, the pathway of -6q preceding +6p may be important in non-GCB. Further studies may identify tumor-related genes involved in this pathway.

Recent research by Silvia Bea *et al*[34] indicated that changes of gene copy numbers of DLBCL were related to molecular subtypes. ABC-DLBCL had frequent trisomy 3, gains of 3q and 18q21-q22, and losses of 6q21-q22, whereas GCB-DLBCL had frequent gains of 12q12, and PMBCL had gains of 9p21-pter and 2p14-p16. Parallel analysis of CGH alterations that -6q was regarded as the characteristic criterion of non-GCB classification was consistent with our tree model. Their results also proved that +3q, +18q, -6q can be used as evidences in DLBCL classification. Defects exist in the current tree model classification based on chromosomal data. First, the minimum units of classification are chromosome arms. While the methods are amenable to a finer level of resolution in the genome, such a step would require considerably more data. Second, regions of loss or gains with length less than 10Mbp are too small to be detected by CGH. And third, gene aberrations caused by factors such as mutations or translocations without copy number changes cannot be detected. In the research of DLBCL, decreased major histocompatibility class II (MHCII) expression is associated with poor survival in diffuse large B-cell lymphoma[6]. Lisa *et al*[35] show that loss of MHCII in diffuse large B-cell lymphoma is highly coordinated not due to chromosomal deletions. As a result, classifications of tree models are not totally consistent with molecular classifications of DLBCL. Further experimental evidence is needed to explore tree models in the classification of chromosomal CGH data.

ACKNOWLEDGMENTS

COMMENTS

Background

Desper *et al* have proposed tree models to define multi-step and multi-pathway processes of tumorigenesis based on analysis of comparative genomic hybridization (CGH) data of chromosome alterations. Diffuse large B-cell lymphoma (DLBCL) is the most common type of malignant lymphoma. Statistical data show that DLBCL accounts for approximately 30%-59% in adult NHL. Recently, based on differences of gene expression profiles, DLBCL can be divided into two classes: germinal center B-cell-like (GCB) and non-GCB.

Research frontiers

Tree model methods have been applied successfully on several tumors, However, tree models of DLBCL have not been reported.

Innovations and breakthroughs

The authors constructed tree models of DLBCL for classification by chromosome copy numbers, to compare with cDNA microarray classification, and to explore models of multi-gene, multi-step and multi-pathway process of DLBCL tumorigenesis, and found the tree models of lymphoma established from CGH data can be used in the classification of DLBCL.

Applications

These models can suggest multi-gene, multi-step and multi-pathway processes of tumorigenesis. Two pathways, -6q preceding +6q and +3q preceding +18q, may be important in understanding tumorigenesis of DLBCL. The pathway, -6q preceding +6q, may have a close relationship with the tumorigenesis of non-GCB DLBCL. Tree modelling may be used in the classification for DLBCL.

Peer review

The authors constructed tree models of DLBCL for classification by chromosome copy numbers. The classifications of tree models are not totally consistent with molecular classifications of DLBCL. Further experimental evidence is needed in order to explore the scientific entity of tree models in the classification of chromosomal CGH.

REFERENCES

1   **Desper R**, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schäffer AA. Distance-based reconstruction of tree models for oncogenesis. *J Comput Biol* 2000; **7**: 789-803

2   **Yin HF**, Li T, Li JX. Retrospective analysis of 304 cases of malignant lymphomas in pathology: study and practice of the WHO classification of lymphoid neoplasms. *Zhonghua Yixue Zazhi* 2003; **83**: 1556-1560

3   **Ohshima K**, Suzumiya J, Sato K, Kanda M, Haraoka S, Kikuchi M. B-cell lymphoma of 708 cases in Japan: incidence rates and clinical prognosis according to the REAL classification. *Cancer Lett* 1999; **135**: 73-81

4   **A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma**. The Non-Hodgkin's Lymphoma Classification Project. *Blood* 1997; **89**: 3909-18

5   **Alizadeh AA**, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**: 503-511

6   **Rosenwald A**, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, López-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002; **346**: 1937-1947

7   **Karnan S**, Tagawa H, Suzuki R, Suguro M, Yamaguchi M, Okamoto M, Morishima Y, Nakamura S, Seto M. Analysis of chromosomal imbalances in de novo CD5-positive diffuse large-B-cell lymphoma detected by comparative genomic hybridization. *Genes Chromosomes Cancer* 2004; **39**: 77-81

8   **Monni O**, Joensuu H, Franssila K, Knuutila S. DNA copy number changes in diffuse large B-cell lymphoma--comparative genomic hybridization study. *Blood* 1996; **87**: 5269-5278

9   **Stokke T**, DeAngelis P, Smedshammer L, Galteland E, Steen HB, Smeland EB, Delabie J, Holte H. Loss of chromosome 11q21-23.1 and 17p and gain of chromosome 6p are independent prognostic indicators in B-cell non-Hodgkin's lymphoma. *Br J Cancer* 2001; **85**: 1900-1913

10 **Berglund M**, Enblad G, Flordal E, Lui WO, Backlin C, Thunberg U, Sundström C, Roos G, Allander SV, Erlanson M, Rosenquist R, Larsson C, Lagercrantz S. Chromosomal imbalances in diffuse large B-cell lymphoma detected by comparative genomic hybridization. *Mod Pathol* 2002; **15**: 807-816

11 **Brodeur GM**, Tsiatis AA, Williams DL, Luthardt FW, Green AA. Statistical analysis of cytogenetic abnormalities in human cancer cells. *Cancer Genet Cytogenet* 1982; **7**: 137-152

12 **Desper R**, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schäffer AA. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol* 1999; **6**: 37-51

13 **Fitch WM**, Margoliash E. Construction of phylogenetic trees. *Science* 1967; **155**: 279-284

14 **Saitou N**, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406-425

15 **Desper R**, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol* 2002; **9**: 687-705

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

17 **Kallioniemi A**, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 1992; **258**: 818-821

18 **Schena M**, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**: 467-470

19 **Fearon ER**, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61**: 759-767

20 **Jiang F**, Desper R, Papadimitriou CH, Schäffer AA, Kallioniemi OP, Richter J, Schraml P, Sauter G, Mihatsch MJ, Moch H. Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. *Cancer Res* 2000; **60**: 6503-6509

21 **Kainu T**, Juo SH, Desper R, Schaffer AA, Gillanders E, Rozenblum E, Freas-Lutz D, Weaver D, Stephan D, Bailey-Wilson J, Kallioniemi OP, Tirkkonen M, Syrjäkoski K, Kuukasjärvi T, Koivisto P, Karhu R, Holli K, Arason A, Johannesdottir G, Bergthorsson JT, Johannsdottir H, Egilsson V, Barkardottir RB, Johannsson O, Haraldsson K, Sandberg T, Holmberg E, Grönberg H, Olsson H, Borg A, Vehmanen P, Eerola H, Heikkila P, Pyrhönen S, Nevanlinna H. Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus. *Proc Natl Acad Sci USA* 2000; **97**: 9603-9608

22 **Schäffer AA**, Simon R, Desper R, Richter J, Sauter G. Tree models for dependent copy number changes in bladder cancer. *Int J Oncol* 2001; **18**: 349-354

23 **Huang Q**, Yu GP, McCormick SA, Mo J, Datta B, Mahimkar M, Lazarus P, Schäffer AA, Desper R, Schantz SP. Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression. *Genes Chromosomes Cancer* 2002; **34**: 224-233

24 **Yin ZH**, Huang ZX, Liu TF, Li H, Yao KT. Exploration of multigene, multistep and multipathway model of nasopharyngeal and colorectal carcinogenesis. *Zhonghua Zhongliu Zazhi* 2004; **26**: 135-138

25 **Huang Z**, Desper R, Schäffer AA, Yin Z, Li X, Yao K. Construction of tree models for pathogenesis of nasopharyngeal carcinoma. *Genes Chromosomes Cancer* 2004; **40**: 307-315

26 **Shipp MA**, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002; **8**: 68-74

27 **Hans CP**, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, Müller-Hermelink HK, Campo E, Braziel RM, Jaffe ES, Pan Z, Farinha P, Smith LM, Falini B, Banham AH, Rosenwald A, Staudt LM, Connors JM, Armitage JO, Chan WC. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* 2004; **103**: 275-282

28 **Rosenwald A**, Wright G, Leroy K, Yu X, Gaulard P, Gascoyne RD, Chan WC, Zhao T, Haioun C, Greiner TC, Weisenburger DD, Lynch JC, Vose J, Armitage JO, Smeland EB, Kvaloy S, Holte H, Delabie J, Campo E, Montserrat E, Lopez-Guillermo A, Ott G, Muller-Hermelink HK, Connors JM, Braziel R, Grogan TM, Fisher RI, Miller TP, LeBlanc M, Chiorazzi M, Zhao H, Yang L, Powell J, Wilson WH, Jaffe ES, Simon R, Klausner RD, Staudt LM. Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma. *J Exp Med* 2003; **198**: 851-862

29 **Pan Z**, Shen Y, Du C, Zhou G, Rosenwald A, Staudt LM, Greiner TC, McKeithan TW, Chan WC. Two newly characterized germinal center B-cell-associated genes, GCET1 and GCET2, have differential expression in normal and neoplastic B cells. *Am J Pathol* 2003; **163**: 135-144

30 **Lossos IS**, Alizadeh AA, Rajapaksa R, Tibshirani R, Levy R. HGAL is a novel interleukin-4-inducible gene that strongly predicts survival in diffuse large B-cell lymphoma. *Blood* 2003; **101**: 433-440

31 **Cigudosa JC**, Parsa NZ, Louie DC, Filippa DA, Jhanwar SC, Johansson B, Mitelman F, Chaganti RS. Cytogenetic analysis of 363 consecutively ascertained diffuse large B-cell lymphomas. *Genes Chromosomes Cancer* 1999; **25**: 123-133

32 **Offit K**, Jhanwar SC, Ladanyi M, Filippa DA, Chaganti RS. Cytogenetic analysis of 434 consecutively ascertained specimens of non-Hodgkin's lymphoma: correlations between recurrent aberrations, histology, and exposure to cytotoxic treatment. *Genes Chromosomes Cancer* 1991; **3**: 189-201

33 **Nanjangud G**, Rao PH, Hegde A, Teruya-Feldstein J, Donnelly G, Qin J, Jhanwar SC, Zelenetz AD, Chaganti RS. Spectral karyotyping identifies new rearrangements, translocations, and clinical associations in diffuse large B-cell lymphoma. *Blood* 2002; **99**: 2554-2561

34 **Bea S**, Zettl A, Wright G, Salaverria I, Jehn P, Moreno V, Burek C, Ott G, Puig X, Yang L, Lopez-Guillermo A, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Gascoyne RD, Connors JM, Grogan TM, Braziel R, Fisher RI, Smeland EB, Kvaloy S, Holte H, Delabie J, Simon R, Powell J, Wilson WH, Jaffe ES, Montserrat E, Muller-Hermelink HK, Staudt LM, Campo E, Rosenwald A. Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* 2005; **106**: 3183-3190

35 **Rimsza LM**, Roberts RA, Campo E, Grogan TM, Bea S, Salaverria I, Zettl A, Rosenwald A, Ott G, Muller-Hermelink HK, Delabie J, Fisher RI, Unger JM, Leblanc M, Staudt LM, Jaffe ES, Gascoyne RD, Chan WC, Weisenburger DD, Greiner T, Braziel RM, Miller TP. Loss of major histocompatibility class II expression in non-immune-privileged site diffuse large B-cell lymphoma is highly coordinated and not due to chromosomal deletions. *Blood* 2006; **107**: 1101-1107