Published in final edited form as:
Expert Opin Med Diagn. 2011 November 1; 5(6): 539–550. doi:10.1517/17530059.2011.618185.

Risk Factors of Non-Hodgkin Lymphoma

Yawei Zhang, M.D. Ph.D., **Ying Dai, Ph.D.**, **Tongzhang Zheng, D.Sc.**, and **Shuangge Ma, Ph.D.***
School of Public Health, Yale University, 60 College ST, New Haven, CT 06520, USA

Abstract

Introduction—Despite decades of intensive research, Non-Hodgkin Lymphoma (NHL) remains poorly understood and is largely incurable. NHL is a heterogeneous group of malignancies with multiple subtypes, each of which has distinct morphologic, immunophenotypic, and clinical features. Identifying the risk factors for NHL may improve our understanding of the underlying biological mechanisms and have an impact on clinical practice.

Areas covered—This article provides a review of several aspects of NHL, including epidemiology and subtype classification, clinical, environmental, genetic, and genomic risk factors identified for etiology and prognosis, and available statistical and bioinformatics tools for identification of genetic and genomic risk factors from the analysis of high-throughput studies.

Expert opinion—Multiple clinical and environmental risk factors have been identified. However, they have failed to provide practically effective prediction. Genetic and genomic risk factors identified from high-throughput studies have suffered a lack of reproducibility. The identification of genetic/genomic risk factors demands innovative statistical and bioinformatics tools. Although multiple analysis methods have been developed, there is still room for improvement. There is a critical need for well-designed, prospective, large-scale pangenomic studies.

Keywords

NHL; etiology; prognosis; risk factors; bioinformatics analysis

1. Introduction

Lymphomas are types of cancer derived from lymphocytes. Non-Hodgkin lymphoma (NHL) includes all lymphomas except for Hodgkin's lymphoma. During the past three decades, there have been consistent reports of an increase in the incidence of NHL worldwide [1]. In the United States, the age-adjusted incidence rate has almost doubled since the 1970s, from 11.07 per 100,000 in 1975 to 20.20 per 100,000 in 2008 [2]. The more developed regions have higher incidence rates compared with less developed areas. The incidence rates are about 1.5 times higher in men than in women. The average age at diagnosis is about the sixth decade of life, although certain subtypes of NHL, such as Burkitt lymphoma and lymphoblastic lymphoma, have been diagnosed at a younger age. The mortality rates of NHL have shown a parallel increase as the incidence rates before the later 1990s. After then, there has been a decrease in NHL mortality. The five-year relative survival for the time period of 1975 to 1995 was stable at about 50%, but was reported to have increased to 65% for the time period of 1996 to 2007. NHL survival is better in Caucasians than in African Americans and better in females than in males.

*Shuangge Ma, Ph.D. (Corresponding Author), School of Public Health, Yale University, 60 College ST, New Haven, CT 06520, USA, Tel: 203-785-3119; Fax: 203-785-6912; Shuangge.ma@yale.edu.

NHL represents a heterogeneous group of malignancies ranging from very indolent forms to aggressive ones. It has been enormously challenging to categorize NHL subtypes. During the past 50 years, numerous classification schemes have been proposed and revised [3,4]. The new WHO classification, which incorporates morphology, immunophenotype, cytogenetic and molecular features, clinical behavior, and some known aspects of etiology and pathogenesis into classification of NHL, has become an international standard for both clinical practice and research [5] .There are more than 30 NHL subtypes, with the two most common subtypes -- diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) -- accounting for about 30% and 20% of all NHL cases, respectively [6]. The survival patterns vary greatly by NHL subtypes. It has also been indicated in both heterogeneity and commonality in the etiology of NHL subtypes [7].

Identifying NHL risk factors is an extremely complex process. The design and execution of clinical, genetic, and genomic NHL studies are highly nontrivial and warrant careful investigations. However, due to the limited scope of this article, we will not be able to address all of the issues involved in the identification of NHL risk factors. The main objectives of this article are two-fold. First, the risk factors for developing NHL are reviewed in Section 2. The risk factors for prognosis are reviewed in Section 3. Clinical, environmental, genetic, and genomic risk factors are discussed separately. Second, in Section 4, the available statistical and bioinformatics tools for analyzing NHL studies and identifying risk factors are reviewed. The most recent developments are reviewed as well, and their pitfalls are discussed.

2. Etiology

2.1 Clinical and environmental risk factors

The etiology of NHL is poorly understood, with the only established risk factors being infection and immune dysregulation. The supportive evidences include elevated incidence rates in immunosuppressed populations (those who had organ transplant, immunosuppressive medical treatment, and HIV/AIDS) and among individuals with certain auto-immune diseases (i.e., rheumatoid arthritis, systemic lupus erythematosis, psoriasis, Sjogren's syndrome, celiac disease, etc.) and established links between specific infectious agents and rare NHL subtypes (Epstein-Barr virus and Burkitt lymphoma, human T-cell leukemia/lymphoma virus 1 and adult T-cell leukemia/lymphoma, human herpes virus 8 and primary effusion lymphoma, *Helicobacter pylori*, and gastric mucosa-associated lymphoid tissue (MALT) lymphoma) [8]. However, the aforementioned conditions and factors are relatively uncommon in general populations and can explain only a small portion of NHL cases.

Epidemiological studies have suggested that certain environmental and occupational exposures and lifestyle factors may be associated with the risk of NHL. However, some of the results have been inconsistent. A number of pooled analyses using data from the International Lymphoma Epidemiology Consortium (InterLymph) have investigated some of the factors in greater detail. For example, cigarette smoking was associated with an increased risk of follicular lymphoma, particularly among current smokers compared with former smokers [9]. Alcohol consumption was associated with a reduced risk of NHL, with the lowest protective effect for Burkitt lymphoma [10]. Increased risks of FL and chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL) were found among women who started using hair dyes before 1980, while an increased risk of FL among women who started using hair dyes in 1980 and after cannot be excluded [11]. The association between ultraviolet radiation and the risk of NHL has been controversial [12,13,14]. Although the InterLymph pooled analysis suggested a protective effect of recreational ultraviolet radiation exposure on NHL, the heterogeneity between studies involved in this analysis was

significant [15]. The InterLymph pooled analysis also suggested that severe obesity increased the risk of DLBCL [16]. A recent meta-analysis of prospective studies also indicated a positive association between BMI and the risk of DLBCL [17]. Occupational exposure to organic solvents has been suggested to be associated with an increased risk of NHL [18,19,20,21]. However, the associations between individual solvents (i.e., benzene, toluene, styrene, trichloroethylene, tetrachloroethylene, etc.) and specific NHL subtypes have been inconsistent [22,23,24]. Higher serum levels of organochlorine compounds (i.e., polychlorinated biphenyls (PCB) and p,p′-dichlorodiphenyldichloroethylene (p,p′-DDE)) have been reported to be associated with an increased risk of NHL by some studies [25,26] but not others [27,28]. Several studies have also explored dietary intake and the risk of NHL. High protein and high fat intake was associated with an increased risk of NHL, whereas high vegetable and fruit intake was associated with a reduced risk of NHL [29,30]. High intake of one-carbon nutrients has been suggested to be associated with a reduced risk of NHL [31].

2.2 Genetic and genomic risk factors

There is increasing evidence of common genetic polymorphisms altering NHL risk [32,33,34]. The InterLymph Consortium demonstrated that specific polymorphisms in Th1/Th2 cytokine pathway genes (TNF and IL10) increased risk of B-cell lymphomas compared to control subjects [35]. A study from Germany found that toll-like receptor genes such as *TLR2* and *TLR4* were associated with the risk of NHL subtypes [36]. A pooled analysis of three population-based case-control studies conducted in the United States and Australia reported that genetic variations in the TLR10-TLR1-TLR6 region were associated with the risk of NHL and suggested that TLR2 variants influenced susceptibility to marginal zone B-cell lymphoma [37]. A recent hospital-based study examined polymorphisms in 1,253 immune and inflammation genes in relation to NHL risk and found 12 genes (*TRAF1, RIPK3, BAT2, TLR6, MAP3K5, DUSP2, CREB1, B3GNT3, SELPLG, LSP1, FGG, ITGB3*) with possible roles in tumorigenesis [38]. Reactive oxygen species (ROS) propagates pro-inflammatory cytokines that signal molecules for proliferation of lymphocytes and tumor cells. In addition, ROS, via activation of transcription factors, responds to environmental stress and regulates many cellular actions, including apoptosis, cell differentiation, and proliferation [39]. Several epidemiologic studies reported an increased risk of NHL associated with SNPs (single nucleotide polymorphisms) in genes of *GPX1, NOS2A, SOD2, AKR1A1*, and *CYBA* [34,40].

Chromosomal translocations, such as t(3, 22) translocation in DLBCL and t(14,18) translocation in FL, are a hallmark of NHL [41]. Chromosome aberrations are a consequence of unrepaired or misrepaired DNA strand breaks. Polymorphisms in DNA repair genes may modify repair capacity and alter the risk of NHL. Several population-based case-control studies have found an increased risk of NHL associated with genetic polymorphisms in DNA repair genes, including *RAG1, LIG4, ERCC5, WRN, MGMT,* and *XRCC1* [42]. One-carbon metabolism plays a key role in DNA synthesis and methylation in mammalian cells. The lymphatic system has a great demand for rapid availability of nucleotide precursors because its function depends on proliferation and cell division in response to a foreign stimulus [43]. Alteration of one-carbon metabolism causes alteration of immune function and results in lymphogenesis. This has been supported by epidemiological studies that have consistently reported a link between NHL risk and genetic variations in one carbon metabolic pathway genes (e.g., *TYMS, MTHFR, MTR, BHMT, CBS, FPGS, FTHFD,* and *SHMT1*) [44]. Recent genome-wide association studies have identified three variants associated with follicular lymphoma at 6p21.32 (rs10484561, rs7755224, and rs2647012) [45,46].

3. Prognosis

3.1 Clinical risk factors

In 1993, the International NHL Prognostic Factors Project developed a predictive model and determined five factors: age, tumor stage, serum concentration of lactate dehydrogenase (LDH), performance status, and number of sites of extranodal disease as the International Prognostic Index (IPI) [47]. The IPI has been used to design therapeutic trials for NHL patients and in the selection of appropriate treatment strategies for individual patients. However, prognostic models based on clinical characteristics have not been very successful in determining the best initial treatment and overall survival. Furthermore, these clinical characteristics describe the disease burden but not disease pathogenesis. Therefore, the model is limited in its capacity in describing the role of lymphomagenesis and disease progression, as well as identifying new therapeutic targets.

3.2 Environmental risk factors

Battaglioli et al. first linked smoking and alcohol drinking to NHL survival in a population-based study with 1,138 Italian patients [48]. They found that, compared with those with a lower cumulative exposure to tobacco smoking, those who had smoked for more than 31 pack-years had a worse survival rate (HR(hazard ratio)=1.60, 95% CI=1.18–2.18). When analyzed by subtype, the association was mainly observed for FL. Two subsequent studies also demonstrated the link between cigarette smoking and NHL survival [49,50]. However, one study observed a strong link with CLL/SLL and the other study with FL. Smoking was found to be associated with a high prevalence of translocation t(14; 18) [51]. Furthermore, smoking contains polycyclic aromatic hydrocarbons (PAHs). PAHs induce mutations in tumor suppressor gene *p53* [51], which is a negative prognostic factor in NHL, as it increases the risk of transformation from low-grade NHL towards intermediate/high-grade NHL [52].

In addition, all three studies found that heavy drinkers had a poorer survival rate and higher risk of death [48,49,50]. Battaglioli et al. reported that drinkers had a 40% higher risk of death (HR=1.41, 95% CI=1.10–1.81) compared to non-drinkers, while Talamini et al. found that patients who drank more than 4 drinks a day experienced a higher risk of death (HR=1.69, 95% CI=1.04–2.76) in comparison to drinkers of less than 2 drinks a day [50]. Geyer et al. reported that those drinking more than 43.1g a week had poorer survival rates (HR=1.55, 95% CI=1.06–2.27) compared to never-drinkers [49]. When analyzed by NHL subtypes, the observed risks were elevated in common subtypes. Another study among women found that wine drinkers experienced better survival, while liquor drinkers experienced poorer survival compared with never-drinkers [53]. High alcohol consumption is considered to be one of the factors interfering with the host's immune surveillance system through impairing both humoral and cell-mediated immunity and reducing the functional activity of NK-cells [54]. This is characterized by decreased inflammatory response, altered cytokine production, abnormal reactive oxygen intermediate generation, and impaired antigen-specific immune response. Moreover, high levels of alcohol consumption induce liver cirrhosis and pancreatitis, which may influence mortality and compliance with the chemotherapy regime [54].

One study has linked being overweight to poor outcomes of NHL in a high-dose chemotherapy cohort with 121 patients [55]. It was found that overweight patients (BMI≥28) had significantly shorter overall and disease-free survival compared to patients with BMI<28 (P<0.002 for each). One population-based study found that NHL patients with BMI ≥30 experienced a poorer survival rate compared to patients with BMI 20–24.9 [49]. While an inappropriate production of growth-promoting factors in obese patients, such as

insulin-like growth factor, could render tumor cells more aggressive and hence less sensitive to chemotherapy, a slight reduction of drug delivery to neoplastic tissues in overweight patients may also be an explanation for poorer outcomes.

Except for smoking, alcohol consumption, and obesity, no other environmental and lifestyle factors have been confirmed to be associated with NHL survival. Several putative risk factors identified in etiological studies such as UV radiation, dietary, and occupational exposures are worth pursuing in NHL prognostic studies. For example, UV radiation was found to induce expression of *IL10*, *TNF-α*, and *IL6*, while the elevation of serum level of these mediators has been suggested as a negative prognostic factor and predictor of poor NHL survival [56].

3.3 Somatic alterations

Alizadeh et al. [57] conducted cDNA microarray experiments and showed that DLBCL patients with significantly longer overall survival had higher levels of expression of genes in the germinal-center B-cells. Two genes specifically expressed in the germinal-center B-cells, BCL6 and HGAL, were demonstrated to be predictors of overall survival independent of IPI in two independent groups of DLBCL patients [58,59]. Using supervised analysis of microarray data, Shipp et al. [60] derived a 13-gene model to predict survival of DLBCL independent of IPI. Two genes (*NOR1* and *PDE4B*) could be confirmed independently in the dataset used by Alizadeh et al. [57]. Rosenwald et al. [61] identified a 17-gene outcome predictor including GC-like genes, major histocompatibility complex (MHC) class II genes, a lymph-node signature, and a proliferation signature, each of which was differentially expressed in the three subtypes of DLBCL. Particularly, the proliferation signature and bone morphogenetic protein 6 (BMP6) were up-regulated in the ABC-like DLBCL subtype. The GC-like signature was up-regulated in the GC-like DLBCL subtype. Among the 17 identified genes, nine (*FN1, PLAU, HLA-DQA1, HLA-DRA, EEF1A1L4, NPM3, MYC, BCL6, HGAL*) were associated with survival in independent data analyses. Lossos et al. [58] measured the expressions of 36 genes in DLBCL patients to build a predictive model and then validated the model using two independent microarray datasets from Shipp et al. [60] and Rosenwald et al. [61]. This model contained 6 genes which occur in the germinal center B-cell signature (*LMO2* and *BCL6*), the lymph-node signature (*FN1*), and activated B-cell signature (*CCND2, SCYA3*, and *BCL2*). The expression of *LMO2, BCL6*, and *FN1* is associated with favorable survival, while the expression of *CCND2, SCYA3*, and *BCL2* is correlated with poor survival.

The second most common NHL subtype is FL. The vast majority of FLs have at least one karyotypic abnormality in addition to the t(14:18) translocation [62]. The poor prognosis is correlated with a greater number of abnormalities [63]. An *in vitro* study demonstrated that stromal cells and stimulation of the CD40 receptor in combination with cytokine cocktails (i.e., IL10, IL6, IL15, TNF) were necessary to sustain tumor cell growth in addition to acquired proliferative capacity [64]. Gene expression studies also showed that immune environment was very important in FL prognosis [65]. Results from a recent study showed that the strongest predictors of FL prognosis were the gene expression signatures of non-malignant, tumor-infiltrating immune cells, including T-cells, macrophages, and dendritic cells, suggesting the importance of an interaction between specific molecular alterations in functional B-cells and immunologic regulatory network factors in FL [66].

3.4 Germline polymorphisms

Lech-Maranda et al. [67] found that the *IL10*(−1082G) allele predicted longer disease-free survival (HR=0.76, p=0.00035) and overall survival (HR=0.78, p=0.0015) among 199 DLBCL patients. However, Berglund et al. [68] could not replicate these results in their

population of 244 DLBCL patients. Domingo et al. [69] included 234 NHL cases and found that the *IL10* (3575) and (−1082) G-A/G-A diplotype was an independent prognostic factor for survival (HR=0.26, p=0.003), with patients of this combined genotype having longer overall survival. Kube et al. [70] did not find that survival was associated with *IL10* (−1087AG) in a German study involving 406 aggressive NHL cases. Lee et al. [71] investigated the prognostic role of *IL10* in 108 T-cell lymphoma patients and found that the ATA haplotype was a favorable prognostic factor compared to non-ATA haplotype (HR=2.1, p=0.037).

In a study involving 273 NHL patients [72], a haplotype analysis showed that the presence of at least two *TNF* or *LT-α* high-producer alleles was significantly associated with shorter progression-free survival and overall survival. In another study using 488 clinical trial patients with childhood or adolescent NHL, Seidemann et al. [73] found that high-producer haplotype *TNF*-308/LT-α+252 was associated with a 2.34-fold increase in risk of events (relapse, death in continuous complete remission, and second malignancy) in two NHL subtypes -- pediatric Burkitt's lymphoma and B-cell acute lymphoblastic leukemia. Juszczynski et al. [74] found that both *TNF* (−308A) and *HLA-DRB1**2 alleles were associated with shorter progression-free survival and overall survival (p=0.004 and p=0.005, respectively) in 204 NHL patients. Nowak et al. [75] found that the *HLA8.1* haplotype (AH) was an important contributor to progression-free survival and overall survival in 154 NHL patients. Fitzgibbon et al. [76] did not find an association with *TNF* and *LT-α* in 121 FL patients.

Gemmati et al. [77] investigated polymorphisms in a one-carbon metabolic pathway gene, *MTHFR*, in relation to survival in 110 high-grade NHL patients. They found a lower probability of event-free survival at five years for 677T-carriers with log-rank p values of 0.05 in the whole group and 0.07 in the methotrexate (MACOP-B)-treated group. The results suggested that *MTHFR* gene variants played a critical role in NHL outcomes, possibly by interfering with the action of methotrexate with significant effects on toxicity and survival. It may be that genetic polymorphisms in folate pathway genes are useful through the reduction of chemotherapy toxicity and/or to improve survival. Hohaus et al. [78] investigated polymorphisms in detoxification enzymes of the glutathione S-transferase (GST) family in relation to disease prognosis in 89 patients with FL. Both *GSTM1* and *GSTT1* deletions were found to be significantly associated with poor event-free survival independent of IPI. The survival was even worse when patients carried a double negative genotype (p=0.01). Hu et al. [79] found that polymorphisms of *BCRP* (G34A) and (C421A) were associated with survival of DLBCL in a Chinese cohort with 156 DLBCL patients. Patients with 34AA genotypes had worse survival compared with those with GG/GA genotypes (HR=3.69, p=0.001). A significant association between 421CC genotypes and poor survival of DLBCL was only observed among patients diagnosed at age 50 or younger.

All studies mentioned above were hospital-based and investigated a very small number of genes. One population-based study has investigated the role of germline mutations in FL and DLBCL survival [80]. This study evaluated 73 SNPs from 44 candidate immune genes in 278 FL and 365 DLBCL patients with a median follow-up of 4.9 years (range 2.3–6.5 years). The researchers found that SNPs in IL8 (rs4073; HR=2.14, 95% CI=1.26–3.63), IL2 (rs2069762; HR=1.80, 95% CI=1.06–3.05), IL12B (rs3212227; HR=1.83, 95% CI=1.06–3.06), and IL1RN (rs454078; HR=1.93, 95% CI=1.11–3.34) were significantly associated with overall survival of FL and that SNPs in IL8RB (rs1126580; HR=2.11, 95% CI=1.28–3.50), IL1A (rs1800587; HR=1.90, 95% CI=1.26–2.87), TNF (rs1800629; HR=1.44, 95% CI=0.95–2.18), and IL4R (rs2107356; HR=1.97, 95% CI=1.01–3.83), as well as an IL10 haplotype (global p=0.03), were the strongest predictors of overall survival of DLBCL. These results supported a role for germline variation in immune genes, particularly genes

involved in a pro-inflammatory state, as predictors of improved survival. Another population-based study identified six SNPs in four metabolism genes (*CYP2E1*, *GSTP1*, *GSTT1*, and *NAT1*) associated with NHL survival [81].

4. Statistical and Bioinformatics Analysis

4.1 Identification of clinical and environmental risk factors

Research on risk factors of NHL started with clinical measurements and environmental exposures. Examples of such risk factors have been provided in Sections 2.1, 3.1, and 3.2. In such studies, a small number of covariates (clinical measurements and/or environmental exposures) were measured on a relatively large number of samples. Standard statistical analysis techniques, including exploratory analysis (examination of marginal and joint distributions, detection of outliers and influential observations, and dichotomization and transformation of covariates), univariate analysis (evaluation of marginal association, marginal regression with individual covariates), and multivariate analysis (regression analysis with multiple covariates), are usually straightforwardly applicable. Commonly adopted models include linear regression models, (polytomous) generalized linear models, and Cox models. Such techniques are relatively mature and have been described in many publications [82].

4.2 Identification of genetic and genomic risk factors

As has been demonstrated in multiple studies, clinical and environmental risk factors do not have satisfactory predictive power. In recent studies, more and more attention has been shifted towards the identification of genetic and genomic risk factors [83].

4.2.1 Candidate-gene and pangenomic approaches—In terms of study design, genetic and genomic studies may take a candidate-gene approach or a pangenomic approach. In candidate-gene-based studies, investigators select a small number of potential risk factors (SNPs, genes, pathways) for profiling. For example, in the study described in [84], a total of 1,764 SNPs were profiled. Among them, there were 1,462 tag SNPs from 201 candidate genes related to immune response. In addition, 302 SNPs in 143 candidate genes previously genotyped by Taqman assay were also profiled. In pangenomic studies, a large number of potential risk factors are profiled without any strict pre-selection. Many existing microarray gene expression studies took this approach. For example, the study reported in [60] profiled the expressions of 6,817 genes. The study conducted by an NIH group collected gene expression data of 7,399 probes [61].

The candidate-gene approach was widely adopted early on. It made possible for researchers to produce in-house printed chips from their own laboratories. This approach has the advantage of being able to be customized for each experiment or for the area of interest of one's own laboratory. This was important in the early days of genomic studies, as the alternative of commercially printed arrays had relatively few probe sets, without full genome coverage. In addition, in some early studies, financial limitation had been a main reason for adopting the candidate-gene approach. In more recent studies, with the fast development of commercial arrays, the pangenomic approach has gained more popularity. Compared with the candidate-gene approach, it has some obvious advantages. First, our knowledge of NHL genetic/genomic risk factors is still very limited. The pangenomic approach makes it possible to explore regions of chromosomes not previously known to be associated with NHL and can lead to the discovery of new markers. Second, in the literature, there have been considerable discrepancies in the identified risk factors. Among the multiple possible causes is the incomparability of different profiling platforms. Commercially available pangenomic arrays may have less technical variation and hence lead to better reproducibility

for identified risk factors. In addition, nowadays, the pangenomic arrays can be as affordable as in-house-built arrays.

4.2.2 Exploratory analysis—With high-throughput genetic/genomic measurements, exploratory analysis may also include examination of marginal distributions and detection of outliers, although examination of joint distributions and dichotomization and transformation of covariates are less frequently conducted because of the high dimensionality.

An important exploratory analysis tool is clustering. Clustering includes supervised and unsupervised clustering, depending on whether the response variable is used in clustering. An area where clustering has been especially useful is the identification of genetic/genomic risk factors that can be used to discriminate subjects with and without cancer and subjects with different subtypes of cancer [85]. In the microarray study conducted by Alizadeh et al. [57], unsupervised clustering was used to group genes on the basis of similarity of their expressions over 96 samples. It was suggested that there might exist a subgroup of DLBCL resembling normal germinal centre B cells. In the microarray study conducted by Rosenwald et al. [61], which included 240 cases of DLBCL all treated with pre-rituximab, unsupervised clustering was performed in the gene expression space of 100 genes selected from a preliminary study predictive of ABC-like or GC-like DLBCL. This procedure identified groups with gene expression profiles typical of either the ABC-like or GC-like subtypes, together with a third group that did not express like either subtype. The level of heterogeneity in the third group indicated that it might consist of additional subtypes of DLBCL. The GC-like DLBCL subtype was characterized by the t(14; 18) translocation and amplification of REL, which were exclusive to this subtype, suggesting that a limited number of genomic features might be able to accurately identify this subtype.

In statistics literature, a large number of clustering methods have been proposed, many of which have been applied to genetic/genomic studies [85]. Different clustering methods may lead to different clustering results. Part of the difference is caused by the fact that the validity of different clustering methods relies on different data and model assumptions. For example, one of the most popular clustering methods is K-means clustering, which assumes that the covariates have a multivariate normal distribution. Such assumptions rarely hold and are practically impossible to verify with high-dimensional data. Partly to tackle this problem, Monti et al. [86] proposed to investigate the underlying biology of DLBCL by the identification of molecularly distinct subsets of DLBCL that were sufficiently robust to be captured by multiple clustering methods. Three different clustering approaches, including hierarchical clustering, self-organizing maps, and probabilistic clustering, were employed, together with consensus clustering, to identify the top 5% of genes with the highest reproducibility across duplicate samples and largest variation across tumors. The most robust resultant substructure across all the three clustering algorithms was one of three clusters, the biological nature of which was determined by gene set enrichment analysis. The first cluster (OxPhos cluster) was enriched in oxidative phophorylation genes, the second (BCR/proliferation cluster) in cell cycle regulatory genes, and the third (HR cluster) in genes associated with the HR.

4.2.3 Univariate and multivariate analysis—Consider the NHL prognosis study as an example. Let T denote the survival time of interest and $X = (X_1,\ldots,X_p)^T$ denote the p-dimensional covariates (genetic, genomic measurements). The following two types of analysis have been conducted.

The first type of analysis is to identify which X_js(j=1…p) are significantly associated with T in a univariate sense [87]. This type of analysis consists of the following steps: (1) For j = 1…p, describe the relationship between covariate X_j and survival time T using model T ~

$f(X_j\beta_j)$ where β_j is the unknown regression coefficient and f is the known link function. f can be parametric or semiparametric; and (2) A statistic measuring the strength of association for X_j is computed. Examples of the statistic include the magnitude of the estimate of β_j, significance level (p-value) of the estimate, likelihood, and others. Of note, as the statistical models have only a single covariate, standard estimation and inference approaches as described in many statistics textbooks are directly applicable and can be realized using multiple software packages. The p potential risk factors can then be ranked based on the statistics obtained above. This procedure provides a relative ranking of potential risk factors. It is useful when researchers are interested in investigating a fixed number of top-ranked covariates. For rigorous statistical inference, a commonly adopted approach is to take p-value as the ranking statistic. When there are a large number of potential risk factors, the identification problem becomes a multiple-comparisons adjustment problem. A “classic” approach is the Bonferroni approach, which adjusts the cutoff by dividing the number of covariates and may be overly conservative. More recent, less conservative approaches target to control the false discovery rate (FDR). Such approaches target to control the proportion of false positives, as opposed to the probability of a single false positive [88]. There are also open software packages that can easily implement the FDR approaches.

In the second type of analysis, a multivariate regression model $f(X_1,\ldots,X_p)$ is constructed to describe the association between survival and all genetic/genomic measurements and predict the survival time or failure risk of a new subject based on the genetic/genomic profile [89]. In “classic” regression analysis, where the dimension of covariates is much smaller than the sample size, multivariate regression analysis is straightforward and can be implemented using many software packages. With genetic/genomic measurements, particularly pangenomic measurements, usually the data characteristic is reversed. For example, the study described in [90] profiled the expressions of 8,810 genes on 92 MCL patients. Straightforwardly fitting multivariate models leads to saturated models. In addition, in pangenomic studies, only a small number of measurements are expected to be real risk factors, with the rest being “noises”. With genetic/genomic measurements, it is usually necessary to conduct simultaneous model construction and risk factor identification. Traditional statistical methods that can accommodate multi-dimensional covariates and select risk factors include the stepwise, best subset, AIC, BIC, and others. With high-dimensional covariates, those methods have been shown to be unsatisfactory, with major drawbacks including extremely high computational cost and lack of stability. In recent statistics/bioinformatics literature, many advanced data mining methods have been developed to more effectively carry out risk factor identification. Available methods can be classified as dimension reduction, variable selection, and hybrid methods. Detailed reviews have been provided in [91] and references therein.

The first type of analysis is popular in early research. Its most significant advantage is its computational simplicity. On the other hand, it has an obvious drawback. The statistical formulation conflicts with the fundamental biology of NHL. As with other types of cancers, the development and progression of NHL are caused by the joint effects of multiple genetic mutations or defects, as opposed to the disturbance of a single gene. Marginal analysis may easily miss genomic risk factors with important joint but weak marginal effects. On the other hand, risk factors with significant marginal effects may not have predictive power in the joint models. Although conceptually the second type of analysis can solve this problem, it may suffer high computational cost and complexity. Our recent studies were among the few conducting joint analysis of a large number of potential risk factors [92,93]. Two computationally affordable data mining approaches, based on gradient thresholding regularization and penalization, respectively, were proposed. Data analysis suggested that marginal analysis and joint analysis describe different aspects of genetic/genomic risk factors and cannot replace each other. Our literature review suggests that a large number of

effective data mining methods have been developed in statistics and bioinformatics literature. However, many of them have not been applied to NHL studies.

4.2.4 Individual marker-based analysis and gene set-based analysis—Consider, for example, the identification of risk factors from the analysis of a genetic association study. When taking the hierarchical structure of genomic measurements into consideration, analysis can be conducted on at least three different levels, namely the SNP level, gene level, and pathway level [94,95]. To avoid confusion of terminology, consider SNP-level and pathway-level analysis. In SNP-level analysis, the functional units are SNPs, and conclusions are drawn on the effects of individual SNPs. Most results discussed in Sections 2 and 3 were obtained from SNP-level analysis. In the identification of genetic/genomic risk factors of multiple cancers, it has been shown that there are very few “mountains” but many “hills”. That is, the effects of most SNPs are moderate to small. They may be easily missed in individual SNP-based analysis. In addition, the development and progression of NHL is a complex process, caused by the coordinated effects of multiple SNPs. Pathway-based approaches conduct analysis at a higher level of the hierarchical structure. They target identifying the combined effects of multiple SNPs (genes) with coordinated biological functions. Thus, the conclusions are on whether a functional set of SNPs with coordinated biological functions are risk factors. The most popular pathway-based analysis is perhaps the gene set enrichment analysis (GSEA) [96]. It targets identifying pathways enriched with SNPs marginally significantly associated with disease outcomes or phenotypes. Popular alternatives include the maxmean approach [97], global test [98], and others. In a recent study [93], five pathways were identified as predictive for the prognosis of DLBCL, including selenoamino acid metabolism, Type II diabetes mellitus, Glycine, serine, and threonine metabolism, TGF-beta signaling pathway, and insulin signaling pathways. Two pathways were identified as predictive for the prognosis of FL, including endometrial cancer and melanogenesis pathways. Three pathways were identified as predictive for the prognosis of all subtypes combined, including drug metabolism-other enzymes, drug metabolism cytochrome P450, and caffeine metabolism pathways. It was also shown that the risk factors identified from SNP-level analysis and pathway-level analysis were significantly different.

In recent studies [99,100], the weighted co-expression network was used to describe the interplay among genomic risk factors. The network structure was constructed using the connectivity similarity measurements. A dynamic tree cut approach was applied to cut the clustering tree (dendrogram), and the resulting branches were identified as modules. Here, the modules were composed of tightly connected genes and taken as the functional units in analysis. Analysis of DLBCL and MCL gene expression data showed that the predictive models constructed by accounting for the network structure had significantly better predictive performance than those from gene-based analysis.

5. Discussion

Identification of risk factors is a critical step in NHL research and clinical practice. It is the foundation of predictive model-building and treatment regimen selection. NHL has multiple subtypes, with different subtypes having different risk factors, prognosis patterns, and treatment strategies. As a scientific question, it is probably of interest to identify risk factors associated with all subtypes combined. From a clinical point of view, a more realistic strategy is to study different subtypes separately.

For better clarity, we provided separate discussions on clinical/environmental risk factors and genetic/genomic risk factors. From an analysis point of view, the former have a low dimensionality and can be analyzed using standard statistical techniques, whereas the latter have an extremely high dimensionality and demand advanced data mining techniques.

Clinical and environmental risk factors have been investigated in many epidemiologic studies, and some of the findings have been confirmed in independent studies. However, those risk factors alone have not been successful for predicting the risk of NHL or its progression path. High-throughput genetic and genomic studies were conducted in the past two decades, searching for risk factors with independent predictive power. A few possible risk factors were suggested. Bioinformatics investigations and cell line experiments were conducted, showing that some of the suggested risk factors had sound pathological implications. Because of our limited knowledge, we are only able to provide a partial review of existing studies and risk factors they identified. As NHL research is a fast-moving field, the present review may need to be updated in the near future.

Data analysis and risk factor identification with high-throughput genetic/genomic studies are extremely challenging. Such analysis can be classified as exploratory analysis and regression analysis. Regression analysis includes both univariate and multivariate analysis. In regression analysis, both individual marker-based and gene set-based analyses have been conducted. It is worth noting that different analysis strategies may lead to different sets of identified risk factors. It is still unclear how to select the best analysis strategy.

6. Expert Opinion

Research on the identification of NHL risk factors has a long history. Clinical and environmental risk factors have been studied in a large number of epidemiological studies. Some of the findings are consistent, while others are still being debated and investigated. Predictive models constructed using clinical/environmental risk factors alone do not have sufficient predictive power for the risk, prognosis, and initial treatment selection of NHL. Recent studies suggest that genetic and genomic risk factors may have independent predictive power. Existing genetic and genomic studies may have the following limitations. First, the findings have been inconsistent, and many of the identified risk factors have failed to be reproduced in independent studies. Multiple factors may contribute to the low reproducibility, including inherent technical variations (for example, batch effects) in high-throughput profiling techniques, heterogeneity in study cohorts (for example, different age and race groups), the extremely noisy nature of genetic/genomic measurements, limitations of analysis techniques, and limited power of individual studies. Second, most existing studies have been focused on individual marker-based analysis, which contradicts with the fact that the etiology and prognosis of NHL are associated with the interplay of multiple genetic mutations and defects. Even some pathway-based analysis methods suffer from this limitation. For example, the GSEA approach attempts to draw pathway-level conclusions based on separate analysis of individual SNPs (genes). Such analysis can be limited or even misleading considering the difference between marginal and joint analysis. Third, the focus has been on the estimation significance of markers in statistical models. In clinical practice, it is prediction significance (as opposed to estimation significance) that matters more. With high-throughput data, the estimation significance of markers can be a significantly different measure from prediction significance. In a recent study, a random sampling-based method was proposed to quantify the prediction significance of pathways in genetic association studies [93]. It was shown that the risk factors (pathways) identified as with predictive power differed significantly from those obtained using GSEA and other analysis methods.

Our recommendations for future development include the following. First, there is a need for a comprehensive examination of existing studies. This is particularly true for genetic/genomic studies. Individual studies have suffered insufficient power because of small sample sizes caused by limited resources. Meta-analysis or integrative analysis needs to be conducted, increasing statistical power and improving reproducibility of identified markers. Second, pending on available resources, well-designed, prospective, large-scale,

pangenomic studies should be conducted. Many existing NHL genomic studies were retrospective in nature, often without well-defined patient selection criteria. This has led to the great heterogeneity among multiple studies and contributed to the low reproducibility. Quite a few studies, including some reviewed in this article, took a candidate-gene approach. Such studies might miss important risk factors and failed to provide a comprehensive description of the genomic profile. Third, more attention needs to be paid to the statistical and bioinformatics analysis. For example, in most prognosis studies, the Cox proportional hazards model has been adopted as the default model. In a recent study [89], we showed that risk factor identification results might highly depend on the underlying prognosis models, which, unfortunately, are extremely hard to select in practical data analysis. In a second study [93], we demonstrated that different analysis methods might lead to the identification of dramatically different prognosis pathways. Even though quite a few effective analysis techniques have been developed for mining high-throughput data, there is still a lag between the conduct of NHL experiments and effective analysis of such studies. The development and progression of NHL involve the complex interplay of multiple types of risk factors. There is a lack of systematic approaches that can quantify the contributions of different risk factors and effectively integrate multiple types of risk factors into models that can predict an individual's disease development.

Article Highlights

- Despite tremendous effort, the etiology and prognosis of NHL remain poorly understood, and the disease is still largely incurable.
- A few clinical and environmental risk factors have been identified. However, they are only able to explain a small percentage of the development and progression of the disease.
- Quite a few genetic and genomic risk factors have been suggested. There is considerable inconsistency in the findings, and the identified risk factors still need to be comprehensively validated.
- Novel statistical and bioinformatics tools are needed to analyze high-throughput genetic and genomic studies.
- Overall, the identification of NHL risk factors is still at an immature stage. More extensive investigations will be needed before clinically effective predictive models are available.

Acknowledgments

Declaration of Interest

This paper was partly supported by NSF grant DMS-0904181 and NIH grant CA-142774

Bibliography

1. Eltom MA, Jemal A, Mbulaiteye SM, et al. Trends in kaposi's sarcoma and Non-Hodgkin's lymphoma incidence in the united states from 1973 through 1998. Journal of the National Cancer Institute. 2002; 94:1204–1210. [PubMed: 12189223]
2. Howlader N, Noone AM, Krapcho M, et al. SEER Cancer Statistics Review, 1975–2008. Bethesda, MDNational Cancer Institutebased on November 2010 SEER data submission **It includes important statistics on incidence and survival.
3. Aisenberg AC. Historical review of lymphomas. Br. J. Haematol. 2000; 109:466–476. [PubMed: 10886191]

4. Cogliatti SB, Schmid U. Who is WHO and what was REAL? Swiss Med Wkly. 2002; 132:607–617. [PubMed: 12587044] *It reviews classification of NHL subtypes.
5. Jaffe, ES.; Harris, NL.; Stein, H., et al. Lyon, France: IARC Press; 2001. Organization Classification of Tumours/Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues.
6. Morton LM, Wang SS, Devesa SS, et al. Lymphoma incidence patterns by who subtype in the United States, 1992–2001. Blood. 2006; 107:265–276. [PubMed: 16150940]
7. Morton LM, Wang SS, Cozen W, et al. Etiologic heterogeneity among Non-Hodgkin Lymphoma subtypes. Blood. 2008; 15:5150–5160. [PubMed: 18796628]
8. Hartge, P.; Wang, SS.; Bracci, PM., et al. Non-Hodgkin Lymphoma. In: Schottenfeld, D.; Fraumeni, JFJ., editors. Cancer epidemiology and prevention. Vol. 2006. New York: Oxford University Press; p. 898-918.
9. Morton LM, Hartge P, Holford TR, et al. Cigarette smoking and risk of non-Hodgkin lymphoma: a pooled analysis from the international lymphoma epidemiology consortium (InterLymph). Cancer Epidemiol Biomarkers Prev. 2005; 14:925–933. [PubMed: 15824165]
10. Morton LM, Zheng T, Holford TR, et al. Alcohol consumption and risk of non-Hodgkin lymphoma: a pooled analysis. Lancet Oncol. 2005; 6:469–476. [PubMed: 15992695]
11. Zhang Y, Sanjose SD, Bracci PM, et al. Personal use of hair dye and the risk of certain subtypes of non-Hodgkin lymphoma. Am J Epidemiol. 2008; 167:1321–1331. [PubMed: 18408225]
12. Hughes AM, Armstrong BK, Vajdic CM, et al. Sun exposure may protect against non-Hodgkin lymphoma: a case-control study. Int J Cancer. 2004; 112:865–871. [PubMed: 15386383]
13. Smedby KE, Hjalgrim H, Melbye M, et al. Ultraviolet radiation exposure and risk of malignant lymphomas. J Natl Cancer Inst. 2005; 97:199–209. [PubMed: 15687363]
14. Zhang Y, Holford TR, Leaderer B, et al. Ultraviolet radiation exposure and risk of non-Hodgkin's lymphoma. Am J Epidemiol. 2007; 165:1255–1264. [PubMed: 17327216]
15. Kricker A, Amstrong BK, Hughes AM, et al. Personal sun exposure and risk of non-Hodgkin lymphoma: a pooled analysis form the InterLymph Consortium. Int J Cancer. 2008; 122:144–154. [PubMed: 17708556] *It investigates the association between UV and NHL.
16. Willett EV, Morton LM, Hartge P, et al. Non-Hodgkin lymphoma and obesity: a pooled analysis from the InterLymph Consortium. Int J Cancer. 2008; 122:2062–2070. [PubMed: 18167059]
17. Larsson SC, Wolk A. Body mass index and risk of non-Hodgkin’s lymphoma: a meta-analysis of prospective studies. Eur J Cancer. 2011 Jul 4. [Epub ahead of print].
18. Dryver E, Brandt L, Kauppinen T, et al. Occupational exposures and non-Hodgkin’s lymphoma in Southern Sweden. Int J Occup Environ Health. 2004; 10:13–21. [PubMed: 15070021]
19. Kato I, Koenig KL, Watanabe-Meserve H, et al. Personal and occupational exposure to organic solvents and risk of non-Hodgkin’s lymphoma (NHL) in women (United States). Cancer Causes Control. 2005; 16:1215–1224. [PubMed: 16215872]
20. Fritschi L, Benke G, Hughes AM, et al. Risk of non-Hodgkin lymphoma associated with occupational exposure to solvents, metals, organic dusts and PCBs (Australia). Cancer Causes Control. 2005; 16:599–607. [PubMed: 15986116]
21. Vineis P, Miligi L, SenioriCostantini A. Exposure to solvents and risk of non-Hodgkin lymphoma: clues on putative mechanisms. Cancer Epidemiol Biomarkers Prev. 2007; 16:381–384. [PubMed: 17337640]
22. Miligi L, SenioriCostantini A, et al. Occupational exposure to solvents and the risk of lymphomas. Epidemiology. 2006; 17:552–561. [PubMed: 16878041]
23. Cocco P, t’Mannetje A, Fadda D, et al. Occupational exposure to solvents and risk of lymphoma subtypes: results from the Epilymph case-control study. Occup Environ Med. 2010; 67:341–347. [PubMed: 20447988]
24. Wang R, Zhang Y, Lan Q, et al. Occupational exposure to solvents and risk of non-Hodgkin lymphoma in Connecticut women. Am J Epidemiol. 2009; 169:176–185. [PubMed: 19056833]
25. Viel JF, Floret N, Deconinck E, et al. Increased risk of non-Hodgkin lymphoma and serum organochlorine concentrations among neighbors of a municipal solid waste incinerator. Environ Int. 2011; 37:449–453. [PubMed: 21167603] *One of the most recent studies on occupational exposures and NHL.

26. Bertrand KA, Spiegelman D, Aster JC, et al. Plasma organochlorine levels and risk of non-Hodgkin lymphoma in a cohort of men. Epidemiology. 2010; 21:172–180. [PubMed: 20087190]
27. Laden F, Bertrand KA, Altshul L, et al. Plasma organochlorine levels and risk of non-Hodgkin lymphoma in the Nurses' Health Study. Cancer Epidemiol Biomarkers Prev. 2010; 19:1381–1384. [PubMed: 20406963]
28. Cocco P, Brennan P, Ibba A, et al. Plasma polychlorobiphenyl and organochlorine pesticide level and risk of major lymphoma subtypes. Occup Environ Med. 2008; 65:132–140. [PubMed: 17699548]
29. Chang ET, Balter LM, Torrang A, et al. Nutrient intake and risk of non-Hodgkin's lymphoma. Am J Epidemiol. 2006; 164:1222–1232. [PubMed: 17005624]
30. Zheng T, Holford TR, Leaderer B, et al. Diet and nutrient intakes and risk of non-Hodgkin's lymphoma in Connecticut women. Am J Epidemiol. 2004; 159:454–466. [PubMed: 14977641]
31. Lim U, Wang SS, Hartge P, et al. Gene-nutrient interactions among determinants of folate and one-carbon metabolism on the risk of non-Hodgkin lymphoma: NCI-SEER case-control study. Blood. 2007; 109:3050–3059. [PubMed: 17119116]
32. Skibola CF, Nieters A, Bracci PM, et al. A functional tnfrsf5 gene variant is associated with risk of lymphoma. Blood. 2008; 111:4348–4354. [PubMed: 18287517]
33. Lan Q, Zheng T, Chanock S, et al. Genetic variants in caspase genes and susceptibility to non-Hodgkin lymphoma. Carcinogenesis. 2007; 28:823–827. [PubMed: 17071630]
34. Lan Q, Zheng T, Shen M, et al. Genetic polymorphisms in the oxidative stress pathway and susceptibility to non-Hodgkin lymphoma. Hum genet. 2007; 121:161–168. [PubMed: 17149600]
35. Rothman N, Skibola CF, Wang SS, et al. Genetic variation in tnf and il10 and risk of non-Hodgkin lymphoma: a report from the Interlymph consortium. Lancet oncol. 2006; 7:27–38. [PubMed: 16389181] **Investigation of genetic variant and NHL risk from InterLymph.
36. Nieters A, Beckmann L, Deeg E, et al. Gene polymorphisms in toll-like receptors, interleukin-10, and interleukin-10 receptor alpha and lymphoma risk. Genes Immun. 2006; 7:615–624. [PubMed: 16971956]
37. Purdue MP, Lan Q, Wang SS, et al. A pooled investigation of Toll-like receptor gene variants and risk of non-Hodgkin lymphoma. Carcinogenesis. 2009; 30:275–281. [PubMed: 19029192]
38. Cerhan JR, Ansell SM, Fredericksen ZS, et al. Genetic variation in 1253 immune and inflammation genes and risk of non-Hodgkin lymphoma. Blood. 2007; 110:4455–4463. [PubMed: 17827388] **Investigation of the association between a large number of genes and NHL.
39. Sauer H, Wartenberg M, Hescheler J. Reactive oxygen species as intracellular messengers during cell growth and differentiation. Cell Physiol Biochem. 2001; 11:173–186. [PubMed: 11509825]
40. Lightfoot TJ, Skibola CF, Smith AG, et al. Polymorphisms in the oxidative stress genes, superoxide dismutase, glutathione peroxidase and catalase and risk of non-Hodgkin's lymphoma. Haematologica. 2006; 91:1222–1227. [PubMed: 16956821]
41. Chaganti RS, Nanjangud G, Schmidt H, et al. Recurring chromosomal abnormalities in non-Hodgkin's lymphoma: biologic and clinical significance. Semin Hematol. 2000; 37:396–411. [PubMed: 11071361]
42. Shen M, Purdue MP, Kricker A, et al. Polymorphisms in DNA repair genes and risk of non-Hodgkin's lymphoma in New South Wales, Australia. Haematologica. 2007; 92:1180–1185. [PubMed: 17666372]
43. Nauss KM, Newberne PM. Effects of dietary folate, vitamin b12 and methionine/choline deficiency on immune function. Adv Exp Med Biol. 1981; 135:63–69. [PubMed: 7010963]
44. Kim HN, Lee IK, Kim YK, et al. Association between folate-metabolizing pathway polymorphism and non-Hodgkin lymphoma. Br J Haematol. 2008; 140:287–294. [PubMed: 18042267]
45. Conde L, Halperin E, Akers NK, et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat Genet. 2010; 42:661–664. [PubMed: 20639881]
46. Smedby KE, Foo JN, Skibola CF, et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. PLoS Genet. 2011; 7 e1001378.

47. A predictive model for aggressive non-Hodgkin's lymphoma. The international non-Hodgkin's lymphoma prognostic factors project. N Engl J Med. 1993; 329:987–994. [PubMed: 8141877] *It describes a predictive model using clinical risk factors.
48. Battaglioli T, Gorini G, Costantini AS, et al. Cigarette smoking and alcohol consumption as determinants of survival in non-Hodgkin's lymphoma: a population-based study. Ann oncol. 2006; 17:1283–1289. [PubMed: 16728483]
49. Geyer SM, Morton SM, Habermann TM, et al. Smoking, alcohol use, obesity, and overall survival from non-Hodgkin lymphoma: a population-based study. Cancer. 2010; 116:2993–3000. [PubMed: 20564404] **Clinical and environmental risk factors for NHL prognosis.
50. Talamini R, Polesel J, Spina M, et al. The impact of tobacco smoking and alcohol drinking on survival of patients with non-Hodgkin lymphoma. Int J Cancer. 2008; 122:1624–1629. [PubMed: 18059029]
51. Bell DA, Liu Y, Cortopassi GA. Occurrence of bcl-2 oncogene translocation with increased frequency in the peripheral blood of heavy smokers. J Natl Cancer Inst. 1995; 87:223–224. [PubMed: 7707410]
52. Lo coco F, Gaidano G, Louie DC, et al. P53 mutations are associated with histologic transformation of follicular lymphoma. Blood. 1993; 82:2289–2295. [PubMed: 8400281]
53. Han X, Zheng T, Foss FM, et al. Alcohol consumption and non-Hodgkin lymphoma survival. J Cancer Surviv. 2010; 4:101–109. [PubMed: 20039144]
54. Diaz LE, Montero A, Gonzalez-gross M, et al. Influence of alcohol consumption on immunological status: a review. Eur J Clin Nutr. 2002; 56 suppl 3:s50–s53. [PubMed: 12142963]
55. Tarella C, Caracciolo D, Gavarotti P, et al. Overweight as an adverse prognostic factor for non-Hodgkin's lymphoma patients receiving high-dose chemotherapy and autograft. Bone Marrow Transplant. 2000; 26:1185–1191. [PubMed: 11149729]
56. El-Far M, Fouda M, Yahya R, El-Baz H. Serum il-10 and il-6 levels at diagnosis as independent predictors of outcome in non-Hodgkin's lymphoma. J Physiol Biochem. 2004; 60:253–258. [PubMed: 15957243]
57. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature. 2000; 403:503–511. [PubMed: 10676951] **It is one of the earliest gene expression studies on NHL.
58. LossosI S, Alizadeh AA, Rajapaksa R, et al. Hgal is a novel interleukin-4-inducible gene that strongly predicts survival in diffuse large b-cell lymphoma. Blood. 2003; 101:433–440. [PubMed: 12509382]
59. Lossos IS, Jones CD, Warnke R, et al. Expression of a single gene, bcl-6, strongly predicts survival in patients with diffuse large b-cell lymphoma. Blood. 2001; 98:945–951. [PubMed: 11493437]
60. Shipp MA, Ross KN, Tamayo P, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002; 8:68–74. [PubMed: 11786909]
61. Rosenwald A, Wight G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N Engl J Med. 2002; 346:1937–1947. [PubMed: 12075054]
62. Horsman DE, Connors JM, Pantzar T, et al. Analysis of secondary chromosomal alterations in 165 cases of follicular lymphoma with t(14;18). Genes Chromosomes Cancer. 2001; 30:375–382. [PubMed: 11241790]
63. Yunis JJ, Frizzera G, Oken MM, et al. Multiple recurrent genomic defects in follicular lymphoma. A possible model for cancer. N Engl J Med. 1987; 316:79–84. [PubMed: 3537802]
64. Eray M, Postila V, Eeva J, et al. Follicular lymphoma cell lines, an in vitro model for antigenic selection and cytokine-mediated growth regulation of germinal centre b cells. Scand J Immunol. 2003; 57:545–555. [PubMed: 12791092]
65. Bohen SP, Troyanskaya OG, Alter O, et al. Variation in gene expression patterns in follicular lymphoma and the response to rituximab. Proc Natl Acad Sci USA. 2003; 100:1926–1930. [PubMed: 12571354]
66. De Jong D. Molecular pathogenesis of follicular lymphoma: a cross talk of genetic and immunologic factors. J Clin Oncol. 2005; 23:6358–6363. [PubMed: 16155020]

67. Lech-Maranda E, Baseggio L, Bienvenu J, et al. Interleukin-10 gene promoter polymorphisms influence the clinical outcome of diffuse large b-cell lymphoma. Blood. 2004; 103:3529–3534. [PubMed: 14701701]
68. Berglund M, Thunberg U, Roos G, et al. The interleukin-10 gene promoter polymorphism (−1082) does not correlate with clinical outcome in diffuse large b-cell lymphoma. Blood. 2005; 105:4894–4895. [PubMed: 15933064]
69. Domingo-Domenech E, Benavente Y, Gonzalez-Barca E, et al. Impact of interleukin-10 polymorphisms (−1082 and −3575) on the survival of patients with lymphoid neoplasms. Haematologica. 2007; 92:1475–1481. [PubMed: 18024395]
70. Kube D, Hua TD, Kloss M, et al. The interleukin-10 gene promoter polymorphism - 1087AG does not correlate with clinical outcome in non-Hodgkin's lymphoma. Genes Immun. 2007; 8:164–167. [PubMed: 17215862]
71. Lee JJ, Kim DH, Lee NY, et al. Interleukin-10 gene polymorphism influences the prognosis of t-cell non-Hodgkin lymphomas. Br J Haematol. 2007; 137:329–336. [PubMed: 17408400]
72. Warzocha K, Ribeiro P, Bienvenu J, et al. Genetic polymorphisms in the tumor necrosis factor locus influence non-Hodgkin's lymphoma outcome. Blood. 1998; 91:3574–3581. [PubMed: 9572991]
73. Seidemann K, Zimmermann M, Book M, et al. Tumor necrosis factor and lymphotoxinalfa genetic polymorphisms and outcome in pediatric patients with non-Hodgkin's lymphoma: results from Berlin-Frankfurt-Munster trial nhl-bfm 95. J Clin Oncol. 2005; 23:8414–8421. [PubMed: 16293872]
74. Juszczynski P, Kalinka E, Bienvenu J, et al. Human leukocyte antigens class ii and tumor necrosis factor genetic polymorphisms are independent predictors of non-Hodgkin lymphoma outcome. Blood. 2002; 100:3037–3040. [PubMed: 12351419]
75. Nowak J, Kalinka-Warzocha E, Juszczynski p, et al. Association of human leukocyte antigen ancestral haplotype 8.1 with adverse outcome of non-Hodgkin's lymphoma. Genes Chromosomes Cancer. 2007; 46:500–507. [PubMed: 17311253]
76. Fitzgibbon J, Grenzelias D, Matthews J, et al. Tumour necrosis factor polymorphisms and susceptibility to follicular lymphoma. Br J Haematol. 1999; 107:388–391. [PubMed: 10583231]
77. Gemmati D, Ongaro A, Tognazzo S, et al. Methylenetetrahydrofolatereductase c677t and a1298c gene variants in adult non-Hodgkin’s lymphoma patients: association with toxicity and survival. Haematologica. 2007; 92:478–485. [PubMed: 17488658]
78. Hohaus S, Mansueto G, Massini G, et al. Glutathione-s-transferase genotypes influence prognosis in follicular non-Hodgkin's lymphoma. Leuk Lymphoma. 2007; 48:564–569. [PubMed: 17454600]
79. Hu LL, Wang XX, Chen X, et al. Bcrp gene polymorphisms are associated with susceptibility and survival of diffuse large b-cell lymphoma. Carcinogenesis. 2007; 28:1740–1744. [PubMed: 17494054]
80. Cerhan JR, Wang S, Maurer MJ, et al. Prognostic significance of host immune gene polymorphisms in follicular lymphoma survival. Blood. 2007; 109:5439–5446. [PubMed: 17327408]
81. Han X, Zheng T, Foss FM, et al. Genetic polymorphisms in the metabolic pathway and non-Hodgkin lymphoma survival. Am J Hematol. 2010; 85:51–56. [PubMed: 20029944]
82. Armitage, P.; Berry, G.; Mathews, JNS. Statistical Methods in Medical Research. Wiley-Blackwell; 2001.
83. Orsborne C, Byers R. Impact of gene expression profiling in lymphoma diagnosis and prognosis. Histopathology. 2011; 58:106–127. [PubMed: 21261687] **Review of gene expression studies and results on NHL diagnosis and prognosis.
84. Zhang Y, Lan Q, Rothman N, et al. A putative exonic splicing polymorphism in the BCL6 gene and the risk of non-Hodgkin lymphoma. JNCI. 2005; 97:1616–1618. [PubMed: 16264183]
85. MacCuish, JD.; MacCuish, NE. Clustering in Bioinformatics and Drug Discovery. Chapman and Hall/CRC; 2010.

86. Monti S, Savage KL, Kutok JL, et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood. 2005; 105:1851–1861. [PubMed: 15550490]
87. Ma S, Song X. Ranking prognosis markers in cancer genomic studies. Briefings in Bioinformatics. 2011; 12:33–40. [PubMed: 21087949]
88. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Annals of Statistics. 2003; 31:2013–2035.
89. Ma S, Huang J, Shi M, et al. Semiparametric prognosis models in genomic studies. Briefings in Bioinformatics. 2010; 11:385–393. [PubMed: 20123942]
90. Rosenwald A, Wright G, Wiestner A, et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell. 2003; 3:185–197. [PubMed: 12620412]
91. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Briefings in Bioinformatics. 2008; 9:392–403. [PubMed: 18562478]
92. Ma S, Zhang Y, Huang J, et al. Identification of Non-Hodgkin's lymphoma prognosis signatures using the CTGDR method. Bioinformatics. 2010; 26:15–21. [PubMed: 19850755] *It conducts simultaneous analysis of a large number of SNPs on NHL prognosis.
93. Han X, Li Y, Huang J, et al. Identification of predictive pathways for non-Hodgkin lymphoma prognosis. Cancer Informatics. 2010; 9:281–292. [PubMed: 21245948]
94. Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. European Journal of Human Genetics. 2010; 18:111–117. [PubMed: 19584899]
95. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genome wide association studies. American Journal of Human Genetics. 2007; 81:1278–1283. [PubMed: 17966091]
96. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005; 102:15545–15550. [PubMed: 16199517] *It describes one of the most popular gene-set based analysis techniques.
97. Efron B, Tibshirani R. On testing the significance of sets of genes. Annals of Applied Statistics. 2007; 1:107–129.
98. Goeman JJ, van de Geer S, de Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004; 20:93–99. [PubMed: 14693814]
99. Ma S, Shi M, Li Y, et al. Incorporating gene co-expression network in identification of cancer prognosis markers. BMC Bioinformatics. 2010; 11:271. [PubMed: 20487548]
100. Ma S, Kosorok MR, Huang J, et al. Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. BMC Medical Genomics. 2011; 4:5. [PubMed: 21226928]