Translational Research 265 (2024) 1–16

Contents lists available at ScienceDirect

Translational Research

journal homepage: www.elsevier.com/locate/trsl

Tumor molecular landscape of Epstein-Barr virus (EBV) related nasopharyngeal carcinoma in EBV-endemic and non-endemic areas: Implications for improving treatment modalities

Deborah Lenoci [a], Carlo Resteghini [b], Mara S. Serafini [a], Federico Pistore [b], Silvana Canevari [d,1], Brigette Ma [e], Stefano Cavalieri [b,c], Salvatore Alfieri [b], Annalisa Trama [f], Lisa Licitra [b,c], Loris De Cecco [a,*]

[a] Molecular Mechanisms Unit, Experimental Oncology Department, Fondazione IRCCS Istituto Nazionale dei Tumori, Via GA. Amadeo, 42-20133 Milano, Italy
[b] Head and Neck Medical Oncology Department, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, 20133 Milan, Italy
[c] Department of Oncology and Hemato-Oncology, University of Milan, 20122 Milan, Italy
[d] Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, 20133 Milan, Italy
[e] Department of Clinical Oncology, The Chinese University of Hong Kong, Hong Kong SAR, China
[f] Evaluative Epidemiology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

ARTICLE INFO

Keywords:
Nasopharyngeal carcinoma
Epstein-Barr virus
Gene expression
Meta-analysis
Molecular subtypes

ABSTRACT

Epstein-Barr virus (EBV) related- nasopharyngeal carcinoma (NPC) is a squamous carcinoma of the nasopharyngeal mucosal lining. Endemic areas (EA) are east and Southeast Asia, were NPC was recorded with higher incidence and longer estimated survival than in non-endemic area (NEA) such as Europe, We analyzed the gene expression and microenvironment properties of NPC in both areas to identify molecular subtypes and assess biological and clinical correlates that might explain the differences in incidence and outcome between EA- and NEA-NPCs.

Six EA-NPC transcriptomic datasets, including tumor and normal samples, were integrated in a meta-analysis to identify molecular subtypes using a ConsensusClusterPlus bioinformatic approach. Based on the biological/functional characterization of four identified clusters were identified: Cl1, *Immune-active*; Cl2, *defense-response*; Cl3, *proliferation*; Cl4, *perineural-interaction/EBV-exhaustion*. Kaplan–Meier survival analysis, applied to the single dataset with available disease-free survival indicated Cl3 as the cluster with the worst prognosis ($P = 0.0476$), confirmed when applying four previously published prognostic signatures. A Cl3 classifier signature was generated and its prognostic performance was confirmed ($P = 0.0368$) on a validation dataset. Prediction of treatment response suggested better responses to: radiotherapy and immune checkpoint inhibitors *immune-active* and *defense-response* clusters; chemotherapy *proliferation* cluster; cisplatin *perineural-interaction/EBV-exhaustion* cluster. RNA sequencing for gene expression profiling was performed on 50 NEA-NPC Italian samples.

In the NEA cohort, Cl1, Cl2 and Cl3 were represented, while *perineural-interaction/EBV-exhaustion* was almost absent. The immune/biological characterization and treatment-response prediction analyses of NEA-NPC partially replicated the EA-NPC results.

List of abbreviations: NPC, Nasopharyngeal carcinoma; EA, endemic area; NEA, non-endemic area; EBV, Epstein-Barr; RNAseq, RNA sequencing; GE, gene expression; GSEA, Gene Set Enrichment analyses; RT, Radiotherapy; CRT, chemoradiotherapy; PNI, Perineural invasion; GEO, GE Omnibus; RMA, Robust Multi-Array Average; FPKM, fragments per kilobase per million; TPM, transcripts per kilobase million; INT, IRCCS Istituto Nazionale dei Tumori; EBER 1, EBV-Encoded RNA 1; UMI, unique molecular identifiers; PCR, Polymerase Chain Reaction; TMM, trimmed mean of M-values; log2-CPM, log2-counts per million; CDF, cumulative distribution function; DSGA, Disease-Specific Genomic Analysis; SOM, self-organizing map; LASSO, least absolute shrinkage and selection operator; DFS, disease free survival; HR, hazard ratios; 95 % CI, 95 % confidence intervals; PFS, progression-free survival; OS, overall survival; PS, performance status; ROC, Receiver operating characteristics; AUC, area under the curve; DOR, diagnostic odds ratio; RSI, radiosensitivity index; TCIA, The Cancer Immunome Database; IPS, immunophenoscore; DC, Dendritic cells; NKT, Mast cells and natural killer.

* Corresponding author.
E-mail address: loris.dececco@istitutotumori.mi.it (L. De Cecco).

[1] Dr. Canevari is presently retired.

https://doi.org/10.1016/j.trsl.2023.10.004
Received 27 December 2022; Received in revised form 14 October 2023; Accepted 27 October 2023
Available online 8 November 2023

Well characterized EA- and NEA-NPC retrospective and prospective cohorts are needed to validate the obtained results and can help designing future clinical studies.

Brief Commentary

Background

Nasopharyngeal carcinoma (NPC) is a squamous carcinoma of the head and neck showing great variation in geographical and ethnic distribution. As a matter of fact, NPC is a huge health burden in Epstein-Barr virus (EBV) endemic areas, especially in the developing world, while in non endemic areas the same disease has lower incidence rates but with worst survival rates.

Translational significance

Our study provides a relevant biological overview of EBV related NPC, proving the role of immune microenvironment stemming from the viral etiology of this malignancy. By dissecting the molecular landscape of NPC from EBV-endemic and non endemic areas, 4 molecular subtypes were identified (Cl1= Immune-active; Cl2= Defense-response; Cl3= Proliferation; Cl4= Perineural-interaction/EBV-exhaustion) with different prognosis and treatment sensitivity. The immune/biological characterization of NPC could help designing future clinical studies to improve therapeutic strategies and differentiate treatment modalities in endemic and non-endemic areas.

Introduction

Epstein-Barr virus (EBV)-related nasopharyngeal carcinoma (NPC) is an epithelial malignancy arising from the nasopharyngeal mucosal lining. A high incidence of EBV-related NPC has been recorded in EBV-endemic areas (EA) such as East and Southeast Asia,[1] with a high prevalence in males.[2] The incidence of NPC in Europe, a non-endemic area (NEA), is low ($1/10^5$/year); however, the estimated survival rate in that region is much lower than that recorded in Asia (5-year age-standardized relative survival = 54–57 % vs. 74 %).[3] Risk factors of NPC include genetic, ethnic, and environmental factors.[4] Differences in incidence and survival rates between EA and NEA NPCs could involve several factors, including EBV-related factors, genetic susceptibility of different populations to EBV infections, and environmental factors such as local diet and pollution.[5–8] Nevertheless, all proposed models of NPC pathogenesis are based on data derived from EA in Asia. Furthermore, clinical, pathogenic, and microenvironmental characteristics may play additional roles in differences of incidence and outcome observed between EA- and NEA-NPC.

EBV-related NPC in EA has already been characterized using genomic and transcriptomic data analysis.[9] Moreover, prognostic and predictive signatures of EA-NPC have been reported in the literature.[10] Several gene expression datasets for NPC are available, and genome-wide molecular profiling investigations are being conducted on NPC. However, the sample sizes of these studies were relatively small. Furthermore, published/available gene expression analysis data[11] on NEA-NPC remains limited.

Comparing gene expression data from EA- and NEA-NPC allows the recognition of similarities and differences among diseases arising in different geographical areas. We investigated if the transcriptomic patterns involved in EA-NPC could be verified to an Italian cohort for which tumor tissue and clinical data were available. The immune and biological/functional characterization of EA- and NEA-NPC could help in identifying new therapeutic strategies. It has been described in literature that suppressor immune cells, such as myeloid-derived suppressor cells (MDSCs) could be a promising target to enhance the efficacy of cancer immunotherapy.[12] Currently, the treatment for localized NPC includes radiotherapy, which is often combined with platinum-based chemotherapy, especially for locally advanced cancer. Neoadjuvant chemotherapy with cisplatin and gemcitabine was administered in the case of high-risk disease.[13–14] Immunotherapy with checkpoint inhibitors has shown clinical efficacy in recurrent/metastatic advanced NPC and studies aiming to determine its mechanism of action are underway.[15]

Our study aimed to dissect the gene expression and microenvironment of NPC, leading to the identification of the molecular subtypes of EA- and NEA-NPC. We also aimed to elucidate the biological/ functional differences within EA-NPC and between EA-and NEA-NPC to eventually provide new insights into novel treatment strategies.

Methods

Public datasets

The workflow followed that of our previous work.[16] A survey of gene expression data on NPC that were accessible as of May 31, 2022, was conducted. Datasets were chosen based on the following eligibility criteria: i) primary lesions of NPC; ii) endemic areas only; iii) at least 15 tumor samples; iv) Minimum Information About a Microarray Experiment (MIAME)-compliant data,[17] with raw data posted on publicly available repositories and comprehensive gene annotation (GenBank accession or EntrezID). After a literature review, six datasets, with a total of 314 tumors and 35 normal samples, were retrieved from the NCBI Gene Expression Omnibus (GEO) database.[18] Supplementary Table S1 and Fig. S1 contain details of the datasets, including their accession numbers and methods described in the original papers.

Regarding the Affymetrix data (microarray: GSE12452, GSE34573, and GSE132112) signal intensity was normalized within each dataset using a robust multi-array average (RMA) tool; Agilent data (microarray: GSE53819) signal intensity was processed using quantile normalization. For RNA sequencing (RNA-seq) datasets (GSE68799 and GSE102349), the fragments per kilobase per million (FPKM) values were transformed into transcripts per kilobase per million (TPM). To reduce the likelihood of systemic nonbiological technical experimental biases causing batch effects, the ComBat algorithm was used to adjust the data.[19] The redundancy of probes mapping the same EntrezID was eliminated by selecting the probe with the highest variance across samples using the collapse Row R function available in the WGCNA R package.[20] Moreover, we collected available clinical data related to the selected datasets. For validation purposes, the gene expression matrix with computed RPKM values, along with the anonymized clinical annotations from Tay et al.,[21] were retrieved from the Supplementary Materials accompanying the original paper.

Non-endemic study population: recruitment and clinical characteristics

Clinical data and formalin-fixed paraffin-embedded (FFPE) samples of primary lesions were retrospectively collected from patients with histologically confirmed NEA-NPC treated at the Fondazione IRCCS Instituto Nazionale dei Tumori (INT) in Milan from 2009 to 2018. Clinical data were collected from patients' clinical charts. Inclusion criteria were: 1) signed informed consent; 2) histologically confirmed diagnosis of EBV-related NPC; 3) evidence of EBV infection in the nasopharyngeal tumor tissue defined as positive in in situ hybridization

analysis designed for EBV-encoded RNA 1 (EBER 1); 4) clinical stage I–IV according to the AJCC staging system VIIth ed.; 5) availability of archived pre-treatment primary tumor specimen (i.e., macrodissected areas from FFPE sections of good quality) with >70 % of tumor content and no/minor signs of necrosis; and 6) age ≥ 18 years.

Patients with a history of any previous malignancy treated with surgery or radiation in the head and neck region were excluded to avoid radiation-induced tumors, whose biology is expected to be different from that of de novo-arising NPCs. This study was approved by the Institutional Ethics Committee of the INT of Milan (study number: INT188/19).

Gene expression profiling

RNA-seq of tumor samples from patients with NEA was performed using the QuantSeq 3' mRNA Library Prep Kit FWD for Illumina (Lexogen, Vienna, Austria) and NextSeq sequencing in accordance with the manufacturer's protocol for low-input/FFPE samples. Briefly, to make Illumina-compatible libraries, 100 ng of total RNA was reverse transcribed by incubation for 60 min at 42 °C to generate first-strand cDNA, and the RNA was removed. During the synthesis of second-strand cDNA, unique molecular identifiers (UMI) were added, followed by purification of the double-stranded cDNA and amplification by polymerase chain reaction (PCR). The number of PCR cycles was adjusted for each sample based on the RNA input amount and quality using a PCR add-on kit for Illumina, and 16–22 cycles were used for library amplification. The libraries were pooled together in an equal molar ratio, denatured, and diluted to produce a 2.0 pM DNA solution before sequencing on a NextSeq 500 (112 bp single read). Primary demultiplexed data were processed using the BlueBee Genomics Platform. FastQC (version 0.11) and MultiQC (version 1.7) were used in the quality assurance process.

Raw counts obtained from the QuantSeq 3mRNA-Seq were processed using the voom/limma pipeline. First, the dataset was filtered to remove genes with <10 reads in >95 % of samples. We then performed trimmed mean of M-value (TMM) normalization using the limma and edgeR packages to estimate a scale factor to decrease technical bias between samples resulting from differences in library size.[22–23] Finally, we applied voom transformation to convert the raw counts into log2-counts per million (log2-CPM) and calculated the respective observation-level weights for differential expression analysis.[24] RNA-seq data were obtained from the GEO repository (accession number: GSE208281).

Bioinformatics analyses

Unsupervised tumor subtype identification was performed using ConsensusClusterPlus,[25] with k-means clustering and 1-Pearson correlation as the distance matrix. In addition, 1000 resampling interactions were applied to the data by randomly selecting a fraction of the samples, and the presence of $2 < k < 5$ clusters was tested. To identify the number of clusters with the highest stability, empirical cumulative distribution function (CDF) plots displaying the consensus distribution for each k were assessed. To assess the topological connections and distances between each NPC subtype and normal tissues, we performed disease-specific genomic analysis (DSGA).[26,18] DSGA is a computational bioinformatics approach based on data decomposition using the equation $T = NcT + DcT$, where T is the log-transformed tumor gene expression data, NcT is the normal component, and DcT is the disease component. NcT represents the best approximation for the "normal-like" cell state. DcT is the deviation from the normal-like state, thereby highlighting the "aberrant" cell state, which was retained for further analyses. Correlation networks were built based on the metagene data. Metagenic analysis of the gene expression data was performed as previously described.[27] Gene expression data were clustered using self-organizing map (SOM) machine learning. We used SOM, as implemented in the "oposSOM" R package.[28] The SOM translated the original meta-analysis data matrix consisting of 314 NPC cases into a data matrix with a reduced dimensionality of $K = 256$ meta-profiles. Each meta-profile was the mean profile averaged over the expression of all genes in the respective meta-cluster. To assess how gene expression and the identified clusters reflected the diversity of the samples, similarity networks were calculated based on the correlation coefficients of all pairwise combinations of samples.[27] Diversity analysis of the samples and their relationships with the subtypes were performed in terms of correlation networks; correlations connecting samples with $r > 0.5$ were retained for drawing a network map.

For bioinformatic analyses, we used R software version 4.2.0, Bioconductor version 3.15,[29] and NCI BRB-ArrayTools v4.6.1.[30] Plots were produced using the R package ggplot2[31] along with its extension ggrepel (https://ggrepel.slowkow.com/) and were successively assembled into panels using the free and open-source vector graphics editor InkScape.[32] As the true number of NPC subtypes is unknown, we must consider that there may be more than four subtypes, with certain subtypes underrepresented in our cohort.[33,34] A quality check was conducted using silhouette plots to determine the extent to which samples belonged to a specific cluster. The silhouette width values[35] for all samples were calculated (R-package: cluster), and the plot function was used to visualize the partition object. The performance of cluster stratification was assessed using silhouette scores, an estimate of the optimal cluster membership of samples with positive values indicating a preference for the actual cluster chosen and negative values for alternative cluster memberships.

A literature survey of prognostic signatures in EA-NPC yielded four studies, the details of which are presented in Supplementary Table S2. Data, including the gene list and coefficients used to assess signature scores, were retrieved from original papers. The genes included in the signatures were manually curated and implemented in the hacksig package, a unified framework for obtaining single-sample scores.[36]

To construct a classifier capable of identifying Cl3 cases, the method of Friedman et al.[37] using the least absolute shrinkage and selection operator (LASSO) was applied to fit a logistic regression on the gene expression data to predict the binary class stratification, including the cluster of interest, and compare it with the other clusters. The algorithm applies an L1 penalized maximum likelihood method to generate parsimonious models. Diagnostic odds ratio (DOR), a measure of the effectiveness of a diagnostic test that enables correct classification with respect to misclassification, was used to identify the stratification threshold. In this instance, the cutoff was selected to maximize the chance of correct classification relative to the chance of misclassification (maxDOR). Receiver operating characteristic (ROC) and area under the curve (AUC) analyses were performed as per standard methodology using the continuous score generated by LASSO. Calculations and plotting were performed using the easyROC v1.3.1 available (http://www.biosoft. hacettepe. edu. tr/easyROC/).

The established classifier was used to predict the worst cluster membership in gene expression data reported by Tay et al.[21] A survey of the literature from January 2021 to December 2022 for analyses of genes deregulated by EBV infection enabled us to select 11 genes[21,38,39] whose expression was analyzed in the EA-NPC data analysis.

Statistical analysis

Survival analysis and visualization were performed using the Survminer R package and ggsurvplot function (https://rpkgs.datanovia.com/survminer/index.html). The endpoint of the analyses was disease-free survival (DFS), defined as the time from the date of diagnosis to the date of objective tumor progression, excluding clinical deterioration without evidence of disease (local, regional, or distant). We estimated the signature stratification capability using the Kaplan–Meier method and compared curves using the log-rank test. The results of the Cox analyses were reported as hazard ratios (HR) with their corresponding 95 % confidence intervals (95 % CI). We applied a Fleming-Harrington test for censored data based on permutations to assess the behavior/

distortion potentially affecting a small sample size, imposing rho=1, lambda=1.

Tumor microenvironment analysis

Immune and stromal cell prevalences in NPC were estimated using the xCell tool.[40] xCell applies single-sample gene set enrichment analysis (GSEA) to estimate the abundance of 64 cell types, including adaptive and innate immune cells, hematopoietic progenitors, epithelial cells, and extracellular matrix cells, using a compendium of 489 gene sets.

Prediction of treatment response

The multiple cancer signature radiosensitivity index (RSI), directly proportional to radioresistance, was applied to the EA- and NEA-NPC clusters, following the linear model provided in the original study.[41]

Drug sensitivity was evaluated using the pRRophetic R package,[42] according to the guidelines of the authors. This program includes public Cancer Genomic Project data, comprising baseline gene expression and drug sensitivity data for approximately 700 cell lines and 130 compounds currently in clinical use or under investigation. To construct a ridge regression model to predict the AUC value according to the cell line gene expression profile of two databases, we focused our analysis on 'upper aerodigestive' cell lines selected for consistency with NPC. Gene annotations were mapped to the official GeneSymbol, and the cell line and meta-analysis data matrices were homogenized using the ComBat function. For data processing, 20 % of the genes with the lowest variability were eliminated. The homogenized dataset was fitted using a linear ridge regression model to estimate drug sensitivity for each tumor.

The immunophenogram available in The Cancer Immunome Database (TCIA) (https://tcia.at/)[43] was used to visualize the immunophenotypes of the tumor samples. The immunophenogram allows the computation of an aggregated score (i.e., immunophenoscore, IPS) based on the expression profiles of major histocompatibility complex (MHC) molecules, immunomodulators, effector cells [activated CD8 + and CD4 + T cells, effector memory (Tem) CD8 + and CD4 + cells], and suppressor cells [regulatory T cells (Tregs), and MDSCs. It was used to determine the cellular composition of neoantigens in the two cytolytic subsets of skin melanomas. The IPS ranged from 0 to 10, with higher scores associated with increased immunogenicity. We analyzed the IPS, which imposes a clinical benefit when IPS $\geq$ 8, in data of patients with melanoma.[44]

Results

Meta-analysis and identification of clusters in endemic NPC

From the six selected datasets, a data matrix containing 11640 unique genes was constructed and analyzed. The expression data led to the identification of coherent molecular patterns. Using an unsupervised clustering method, the samples were stratified into four molecular clusters (Fig. 1A). Optimal clustering was demonstrated by a positive silhouette score (Si) in all clusters (average Si: 0.91, 0.92, 0.89, and 0.93 for Cluster (Cl)1 Cl2, Cl3, and Cl4, respectively), except for one sample (Fig. 1B). An alluvial diagram was used to exclude technical biases owing to the different platforms employed for expression profiling (Supplementary Fig. 1). No significant stratification was observed, corroborating the biological value of the four-subtype clustering.

We used a correlation spanning tree on deconvoluted data using DSGA to depict the similarity network of the four tumor clusters compared to healthy tissue. The resulting tree showed that most samples segregated into well-localized clusters, reflecting their mutual similarities and divergences from healthy tissues (Fig. 1C). Cl4 was the cluster most distant from the normal tissue samples, whereas Cl1 was the closest. Cl2 and Cl3 formed branches that were not directly interconnected with the normal samples, highlighting similarities between those 2 clusters. The network map depicts the extent of similarities among the tumor clusters, confirming the presence of well-defined clusters among the NPC expression data: Cl1 shows similarities with Cl2 to a certain extent, but not with Cl3 or Cl4 (Supplementary Fig. 2). *The careful meta-analysis of six gene expression datasets allowed to identify 4 well-defined clusters in EA- NPC.*

Prognostic analyses of endemic NPC

A total of 88 cases belonging to only one of the six selected datasets (GSE102349) that reported survival data were stratified as Cl1 (n = 23; 26.1 %), Cl2 (n = 25; 28.4 %), Cl3 (n = 34; 38.6 %), and Cl4 (n = 6; 6.8 %). The 2-year DFS rates were 90.6 %, 90.0 %, 69.1 %, and 100 % for the Cl1, Cl2, Cl3, and Cl4 clusters, respectively (Fig. 2A). The predicted Cl3 group had significantly worse outcomes than the other clusters (log-rank test, P = 0.0476).

Four prognostic signatures were retrieved from the literature (Supplementary Table 2): Tang2018_NPC, Si2022_NPC, and Lu2020_NPC, which included 13, 4, and 3 genes, respectively, and were developed to predict distant recurrence; Zou2020_NPC, a 10-gene prognostic signature, was based on tumor-infiltrating immune cells and microenvironment-relevant genes (Fig. 2B). When our four clusters were challenged against these prognostic signatures, Tang2018_NPC was unable to determine prognosis pertaining to each of the four clusters (P = 0.0492), while the other signatures identified differences among clusters. Si2022_NPC (P = 9.2e-09), Lu2020_NPC (P = 2.6e-06), and Zou2020_NPC (P < 2.2e-16) signatures clearly attributed the best prognosis to Cl1 and the worst to Cl3 and Cl4.

Based on the Kaplan–Meier survival and available prognostic gene signature analyses, Cl3 was associated with the worst prognosis; thus, a classifier signature was developed using the meta-analysis dataset. A 10-fold cross-validated LASSO logistic regression model, following 10 permutations in the 314 cases of the meta-analysis (Cl3, n = 105 vs. other Cls n = 209), generated a classifier signature of 51 genes. The list of genes included in our model is shown in the heatmap (Fig. 3A), and the coefficients are listed in Supplementary Table 3.

The prognostic performance of the classifier was assessed, and the threshold to impute Cl3 membership was selected when the DOR was maximal, corresponding to a value $\geq$ -0.1023 (Fig. 3A). The binary prediction accuracy of the signature (i.e., Cl3 vs. other clusters) was evaluated using ROC, and the AUC was 0.972. (Fig. 3B).

DFS data from 47 cases in the Tay validation set[20] were stratified and dichotomized based on the identified threshold in 28 Cl3 cases and 19 cases belonging to other clusters (clinical characteristics are shown in Supplementary Table 4). Kaplan–Meier curves confirmed the significantly different DFS for the two groups (log-rank test, P = 0.0368 and Fleming-Harrington test p-value= 0.0213), corresponding to a DFS of 77.4 % and 94.7 % for Cl3 and the other clusters, respectively (Fig. 3C). Based on survival and available prognostic gene signature analyses, Cl3 was associated with the worst prognosis and the diagnostic performance of the developed classifier signature (51 genes) was very high.

Immune, biological and functional characterization of endemic NPC

The immune score and cell components were inferred using an in silico approach with the xCell tool to evaluate heterogeneity in the tumor microenvironment (see Supplementary Fig. 3 for all available data about the immune and microenvironment cells in EA-NPC samples). Immune and microenvironment scores exhibited the same trend, with Cl1 expressing the highest immune score (P < 2.2e-16), followed by Cl2, Cl4, and Cl3 (Fig. 4A). Among different cell types, B cells, CD4+ memory T cells, CD8+ T cells, dendritic cells (DC), mast cells, and natural killer T (NKT) cells were considered for further analyses because of their abundance and association with better prognosis.[45] Cl1 showed the highest numbers of B cells, CD4+ memory T cells, CD8+ T cells, DC,

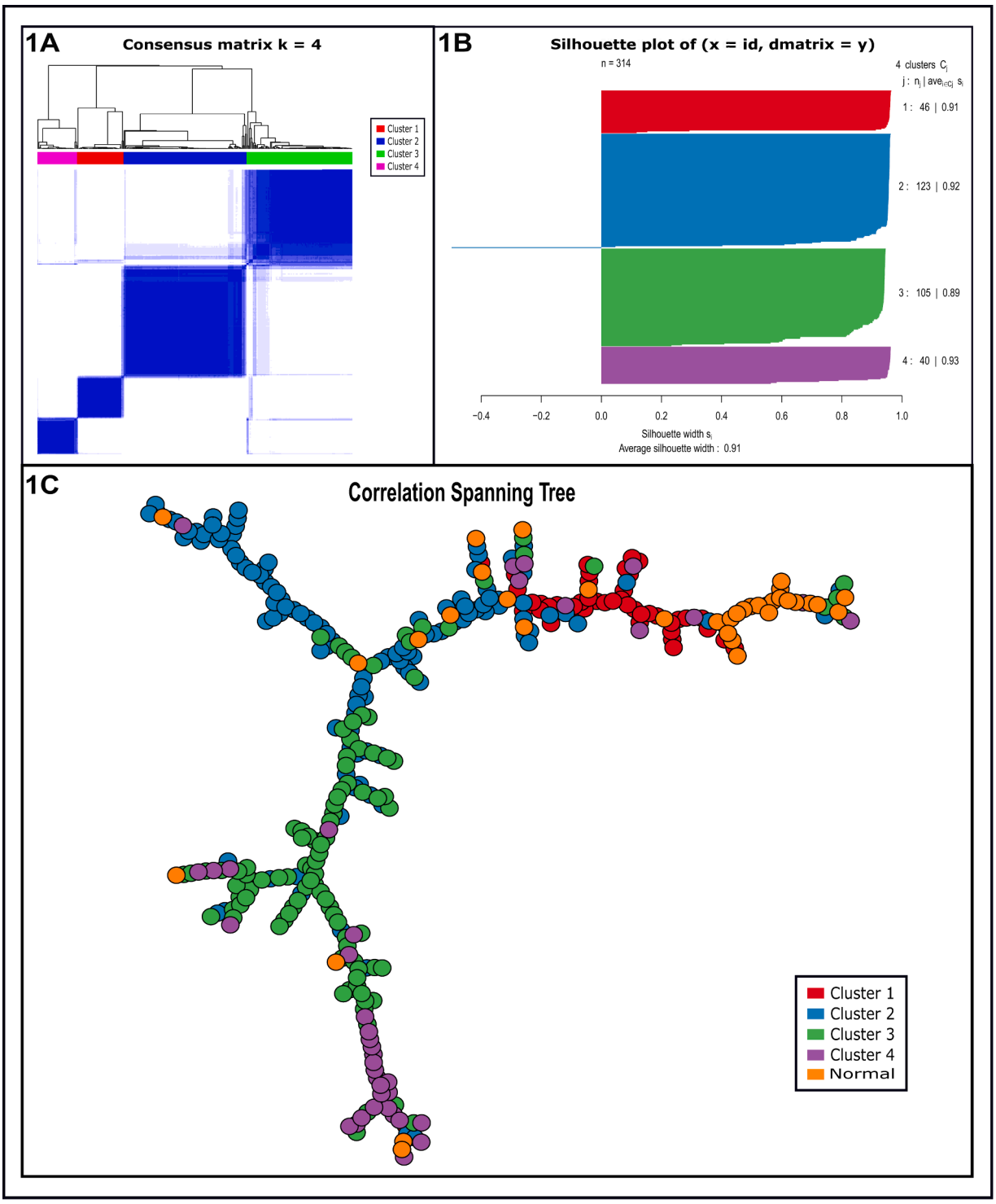

Fig. 1. *Identification of four molecular clusters in EA-NPC.* A. Unsupervised clustering analysis. Gene-expression data revealed four molecular clusters: Cl1 ($n = 46$), Cl2 ($n = 123$), Cl3 ($n = 105$), and Cl4 ($n = 40$) accounting for 14.6 %, 39.2 %, 33.4 %, and 12.7 %, of the cohort, respectively. Correlation matrix showed a high correlation within each subtype, while a low correlation was observed among the different clusters (blue = high correlation; white = low correlation). B. Silhouette plot analysis. Silhouette widths (S_i) have a range of [-1, 1]: coefficients near +1 indicate that the sample is far away from the neighboring clusters; a value of 0 indicates that the sample is on or very close to the boundary between two neighboring clusters; negative values indicates that the samples may have been incorrectly assigned to a cluster. C. Correlation spanning tree. To disclose the connections between normal tissue and tumor molecular subtypes, DSGA was applied to EA-NPC data. Disease state (DcT) representing the distance from the healthy component (NcT) was computed and retained to draw the correlation spanning tree. The modules in the graph are nodes (i.e., samples) connected to a spanning tree of maximal mutual correlation between connected nodes. Thus, the tree summarizes the extent to which each sample belonging to a subtype is similar to the normal tissue.

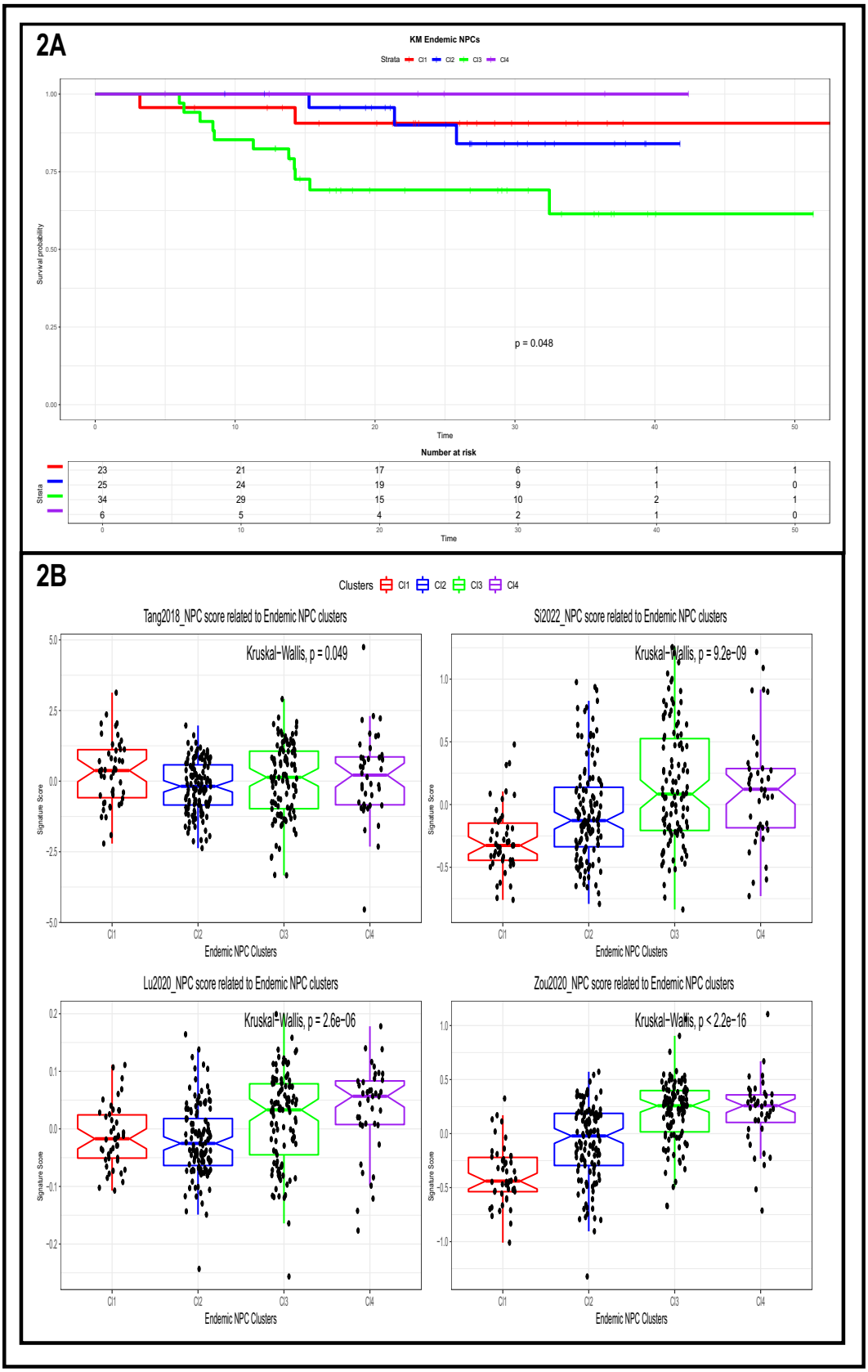

Fig. 2. *Prognostic analyses of endemic NPC clusters.* A. Kaplan–Meier survival curves applied to the GSE102349 dataset. Selecting DFS as the clinical endpoint, prognosis was different in the four clusters (log-rank $P = 0.0476$), with the worst outcome for the Cl3 subtype. B. Application of prognostic signatures to the four clusters: these signatures were imputed following the methods in the original papers, and the association of signatures scores and the four clusters was assessed. Box plots depict the scores of the four prognostic signatures for each cluster: Tang2018_NPC, $P = 0.0492$; Si2022_NPC, $P = 9.2e\text{-}09$; Lu2020_NPC, $P = 2.6e\text{-}06$; and Zou2020_NPC, $P < 2.2e\text{-}16$.

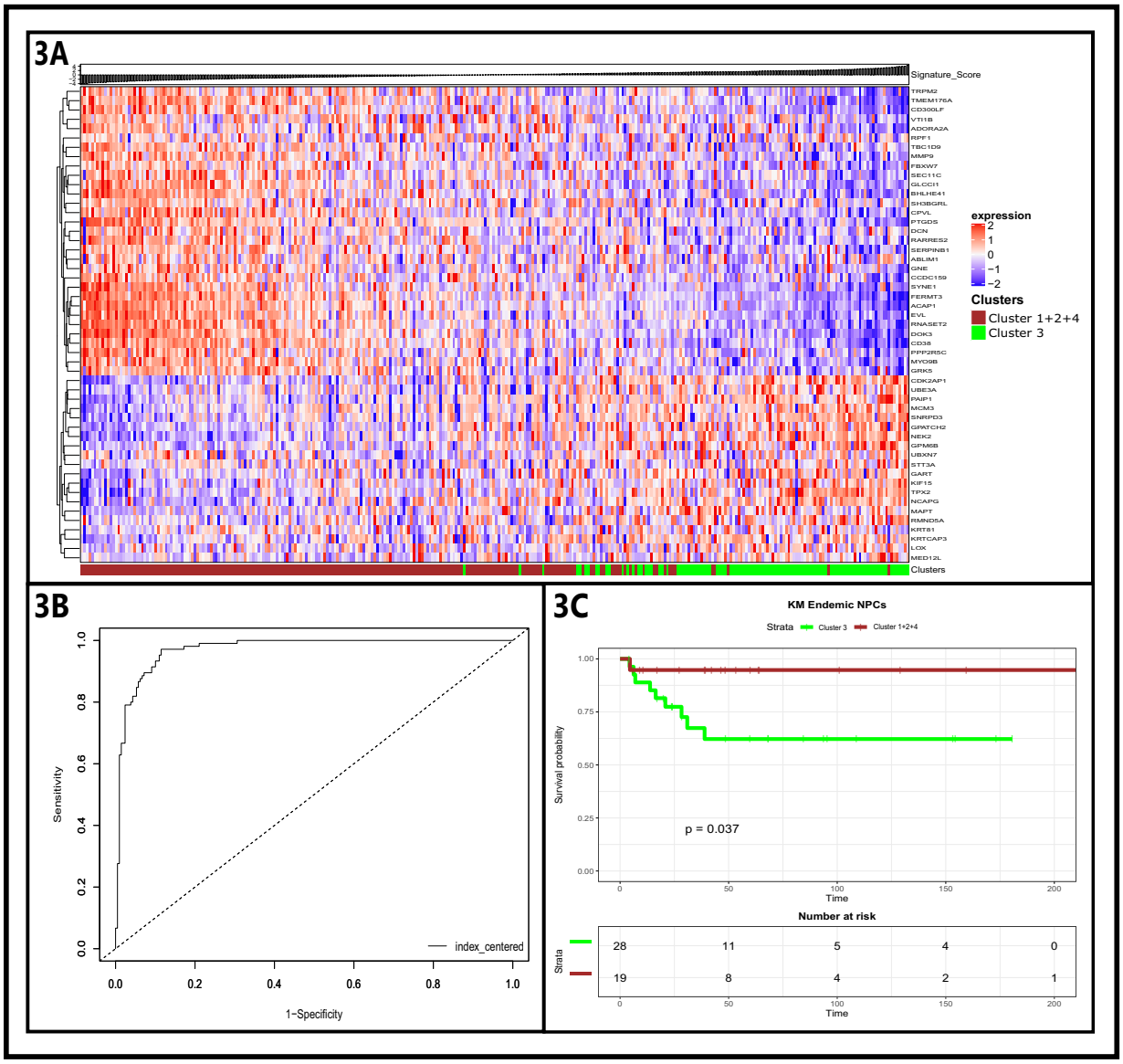

Fig. 3. *Generation and evaluation of gene expression classifier to separate Cl3 from others.* A. Heatmap of the classifier. The heatmap depicts the expression of 51 genes selected by LASSO regression, ordered by rank score value. B. Accuracy of classifier signature by ROC and AUC. AUC was calculated to compare Cl3 with the other clusters, resulting in AUC = 0.972, 95 % CI: 0.956–0.988. C. Kaplan–Meier survival curves of the validation dataset. Patients predicted as belonging to Cl3 ($n = 28$) or other clusters ($n = 19$). Cl3 patients presented a shorter DFS than those belonging to the other clusters [log-rank test, $P = 0.0368$; hazard ratio (HR) = 6.72, 95 % (CI): 0.85–53.17].

and mast cells. Cl3 showed the lowest abundance of CD8+ T cells, DC, and NKT cells. Cl4 contained the highest number of NKT cells but the lowest number of B cells, CD4+ memory T cells, and mast cells (Fig. 4B). In addition, Cl4 showed the highest number of myeloid suppressor cells (MSC, p = 5.9e−07). (Supplementary Fig. 3)

In agreement with the meta-analysis data, Cl3 of the validation set was associated with a low immune score and poor prognosis (Supplementary Fig. 4).

To characterize the biology of EA-NPC clusters, hallmark gene sets were analyzed using GSEA, comparing each cluster with the others (the enriched pathway list is available in Supplementary Table 5): Cl1 was enriched in six of seven immune categories; Cl2 was enriched in all categories except for cellular components, particularly in four of seven immune and two of three DNA damage categories; Cl3 expressed four out of six proliferation categories; and Cl4 did not express any hallmarks (Fig. 4C). These results confirmed the xCell analyses and supported the immune contents of Cl1 and Cl2. In association with a worse outcome, Cl3 showed enrichment of proliferation pathways.

Even after exploiting a holistic approach using GSEA and hallmark gene sets, Cl4 remains largely uncharacterized. To further investigate the biology of Cl4, we compared the expression of 11 single genes (whose expression has already been reported to be deregulated in EA-NPC) in tumors with that in normal tissues. As summarized in Table 1 (see Supplementary Fig. 5 for each gene boxplot), all genes were significantly downregulated in the Cl4 group, except for SSTR2, whose expression was marginally increased. Considering the other clusters, the expression of each gene (upregulation of *TRAF2, LDHA, HIF1A*, and *GPX4* and downregulation of *GSKB*) was in line with previously published data (Table 1).

An indirect comparison between the literature data of pathway regulation (*FGFR1/FGFR2* and *NKB1/NFKB2*) and the gene expression we documented in EA-clusters confirmed the significant downregulation of most genes of interest in Cl4. Moreover, we retrieved four gene sets associated with EBV-related NPC (Liu_NPC, Wood EBV EBNA1 Targets

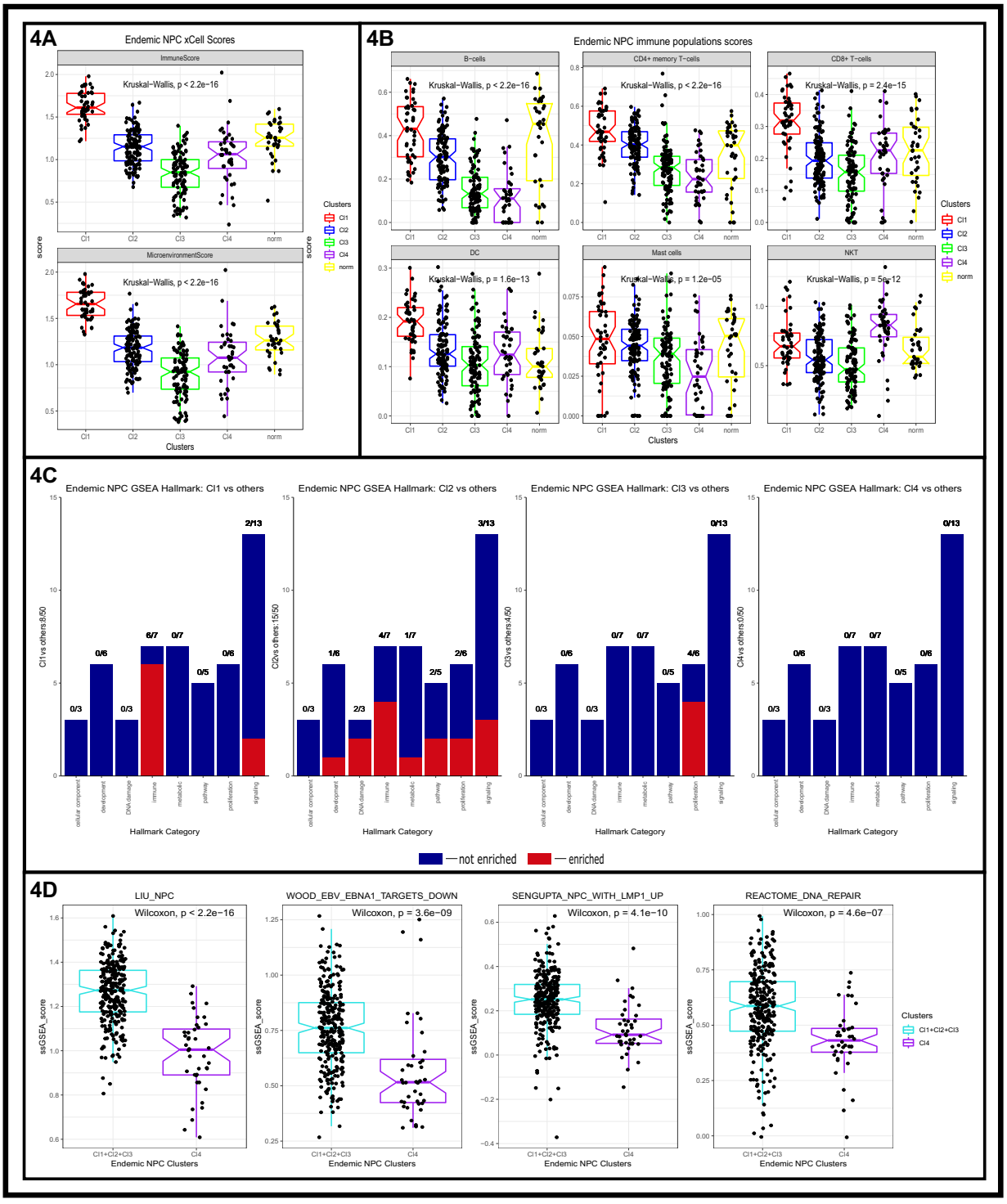

Fig. 4. *Immune and biological characterization of endemic NPC clusters.* A. Immune and microenvironment scores. Immune ($P < 2.2e-16$) and microenvironment ($P < 2.2e-16$) components calculated by xCell for each tumor cluster and normal tissue. B. Expression of immune cells. The abundance of B cells ($P < 2.2e-16$), CD4+ memory T cells ($P < 2.2e-16$), CD8+ T cells ($P = 2.4e-15$), dendritic cells (DC) ($P = 1.2e-13$), mast cells ($P = 1.2e-05$), and natural killer T (NKT) cells ($P = 5e-12$) calculated by xCell for each tumor cluster and normal tissue. C. GSEA hallmark analyses of clusters. Hallmarks were divided into eight categories: cellular component, development, DNA damage, immune, metabolic, pathway, proliferation, and signaling. Gene expression analyses were performed for each cluster vs. others to analyze which component one cluster is enriched in compared to others (see Supplementary Table 5 for the list of enriched pathways). D. ssGSEA of four NPC-related gene sets. Gene sets were analyzed in Cl4 vs others: Liu_NPC ($P < 2.2e-16$), Wood EBV EBNA1 Targets Down ($P = 3.6e-09$), Sengupta NPC_with LMP1 UP ($P = 4.1e-10$), and REACTOME DNA Repair ($P = 4.6e-07$).

Table 1
Single-Gene Expression in EA-NPC clusters: genes/pathways selected according to recent literature data.

Genes/ pathways regulated by LMP1	Literature Data	Reference Author/ year DOI	Gene expression levels in Tumor Cluster vs Normal Cl-1	Cl-2	Cl-3	Cl-4
GSK3B	**GE ↓**	Yang T/2022	=	+	+	- -
TRAF2	**GE ↑**	10.3389/	=	++	+	-
FGF1	**GE ↑**	fcimb.2022.935205	+	++	++	+
LDHA°	**GE ↑**		=	++	++	-
HIF1A°	**GE ↑**		++	++	=	-
SSTR2°°	**GE ↑**	Tay JK/2022 10.1126/sciadv. abh2445	+	++	++	+
NFKB1	**Pathway activation**	Yang T/2022 10.3389/	+	+	=	-
NFKB2	**Pathway activation**	fcimb.2022.935205	=	+	=	-
FGFR1	**Pathway activation**		=	=	=	- -
FGFR2	**Pathway activation**		=	=	+	- -
Genes regulated by EBV infection	**Literature Data**	**Reference Author/ year DOI**	**Gene expression levels in Tumor Cluster vs Normal Cl-1**	**Cl-2**	**Cl-3**	**Cl-4**
GPX4	**GE ↑**	Yuan L/2022 10.1038/s41418-022-00939-8	-	++	++	- -

The LDHA and HIF1A gene expression is regulated by FGRF1/FGFR2 pathways activation.
The SSTR2 gene expression is regulated by NFKB1/NFKB2 pathways activation.

Down, Sengupta NPC with LMP1 UP, and REACTOME DNA Repair) and compared their expression in Cl4 with that in the other clusters (Fig. 4D). The Cl4 cluster was associated with the downregulation of all four selected gene sets.

The presence of neuronal components was also investigated. Starting from the perineural invasion literature data, we analyzed the presence of four pathways associated with neuronal differentiation and activity in Cl4 and compared the results with those of other clusters; the results suggested the presence of neuronal-tumor cell interactions only in Cl4 (Fig. 5).

Considering the above reported immune, biological and functional characterization of the 4 EA-NPC identified clusters, we arrived at the following designation of the corresponding molecular subtypes: Cl1 = *Immune-active*, Cl2 = *Defense-response,* Cl3 = *Proliferation,* and Cl4 = *Perineural-interaction/EBV-exhaustion.*

Prediction of treatment response in endemic NPC

A significant relationship was found between our stratification and RSI. Specifically, *immune-active* and *defense-response* clusters displayed the lowest RSI scores, which predicted their radiosensitivity. In contrast, *proliferation* and *perineural-interaction/EBV-exhaustion* clusters exhibited the highest RSI scores compared with the other two, predicting their intrinsic radioresistance ($P = 5.87 \times 1e{-}05$) (Fig. 6A).

We tested the chemosensitivity of the four clusters based on gene expression data from our meta-analysis. Our findings demonstrated a statistically significant difference in drug sensitivity for patients belonging to different clusters. The *proliferation* cluster showed greater chemo-sensitivity than the others, suggesting that patients could benefit from intensified treatment [cisplatin AUC = 0.653, 95 % CI (0.586–0.721); gemcitabine AUC = 0.739, 95 % CI (0.681–0.796]) (Fig. 6B). The GPX4 expression levels reported in Table 1 were concordant with the chemosensitivity to gemcitabine (Table 1).

A greater benefit from immune checkpoint inhibitors was predicted for *Immune-active* and *defense-response* clusters as compared with *Proliferation* and *Perineural-interaction/EBV-exhaustion* ($\chi 2$ test = 63.6, P < 10e-07) (Fig. 6C).

On the whole, prediction of treatment response in endemic NPC suggested better responses to radiotherapy and immune checkpoint inhibitors in immune-active and defense-response clusters and to chemotherapy in the Proliferation cluster.

Molecular clusters, immune, biological and functional characterization of non-endemic NPC

Based on an institutional historical database, we collected FFPE samples from the primary tumors of 50 EBV-related NPC cases. The clinical characteristics of the NEA (Italian) cohort are presented in Supplementary Table 6. When the four EA-NPC four clusters were applied to the gene expression data of NEA-NPC cohort, the prevalence of the respective molecular subtypes was as follows: *immune-active* (n = 16, 32 %), *defense-response* (n = 11, 22 %), and *proliferation* (n = 22, 44 %); the *perineural-interaction/EBV-exhaustion* subtype was found in only one sample (2 %) (Fig. 7A). Due to the low number of cases in the *perineural-interaction/EBV-exhaustion* cluster, it was not considered for further analyses. The prognostic signatures were tested in NEA cases [Supplementary Table 2]. Applying the Tang2018_NPC signature yielded no statistically significant results (P = 0.58). However, the other signatures identified differences among clusters, as they did in EA-NPC: Si2022_NPC (P = 0.0011), Lu2020_NPC (P = 0.021), and Zou2020_NPC (P = 3.8e-05). These last 3 signatures attributed the best prognosis to Cl1 and the worst to Cl3 [Supplementary Fig. 6].

The expression of the immune-related score in NEA-NPC was higher in the *immune-active* than in the other two clusters, with the lowest expression observed in the *proliferation* cluster (P = 6.4E-09) (Fig. 7B). Furthermore, the microenvironment score showed the same trend as in the EA cohort. These results reflected the overall expression profile of the immune cell population already observed in EA-NPC.

Immune population characterization among the clusters (Fig. 7C) indicated that: *immune-active* expressed a major amount of B cells and mast cells, and the population of CD8+ T cells and NKT cells was comparable with that in *defense-response; defense-response* showed a highest expression of CD4+ memory T-cells and DC; *proliferation* presented the lowest expression of immune cells, similar to that in EA-NPC (Fig. 4B). At variance from deregulation of MSC in EA-NPC, no significant difference in the levels of these cells was recorded in NEA-NPC (see Supplementary Fig. 7 for all available data about the immune and microenvironment cells in NEA-NPC samples).

Hallmark analysis of NEA-NPC (a list of the enriched pathways is presented in Supplementary Table 7) indicated that *immune-active* cluster did not express immune pattern in the immune-active cluster of EA NPC, and no immune pathways were active. The allograft rejection pathway (immune) was enriched in the *defense response* cluster and the *proliferation* cluster showed expression of five out of six proliferation pathways.

Characterization of NEA-NPC indicated that: three out of four clusters of EA-NPC were identified in NEA-NPC while the perineural-interaction/EBV-exhaustion cluster was not expressed; the immune and microenvironment scores showed the same trend between EA and NEA while the expression of some immune populations differed.

Prediction of treatment response in non-endemic NPC

RSI did not differ significantly among the three NEA-NPC clusters (Fig. 8A). We tested the sensitivity to cisplatin and gemcitabine, and our findings revealed a statistically significant difference in drug sensitivity: the *proliferation* cluster showed better sensitivity than the other clusters [cisplatin AUC = 0.774, 95 % CI (0.638–0.911); gemcitabine AUC = 0.889, 95 % CI (0.794–0.984]) (Fig. 8B). When analyzing the IPS, all clusters showed potential responsiveness to immunotherapy (Fig. 8C).

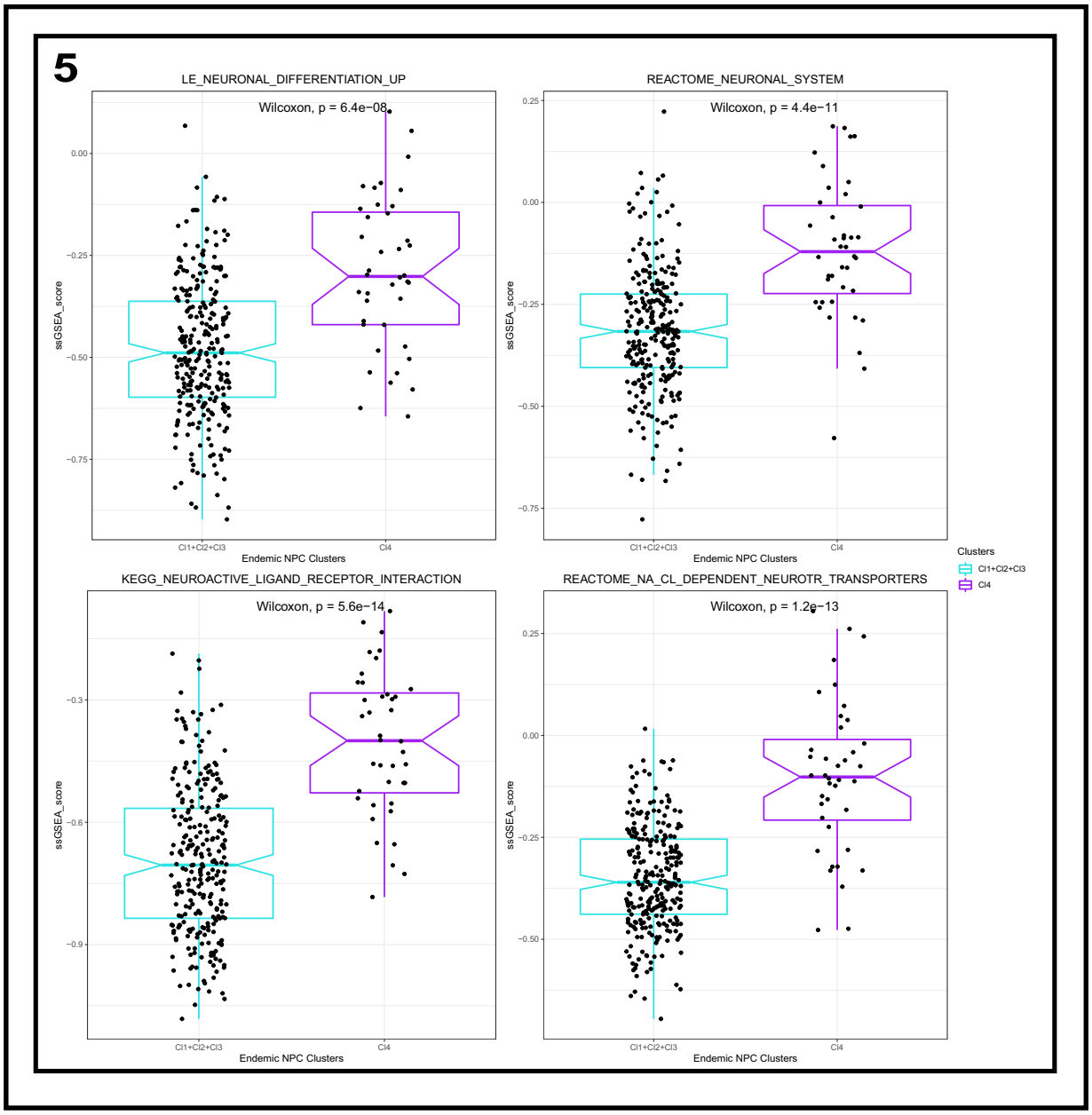

Fig. 5. *Perineural invasion pattern in CL4.* Four perineural invasion gene sets were analyzed in Cl4 vs. others. Results showed that Cl4 was enriched in LE_NEURONAL_DIFFERENTIATION_UP ($P = 6.4e-08$), REACTOME_NEURONAL_SYSTEM ($P = 4.4e-11$), KEGG_NEUROACTIVE_LIGAND _RECEPTOR_INTERACTION ($P = 5.6e-14$), and REACTOME_NA_CL_DEPENDENT NEUROTRANSMITTER_TRANSPORTERS ($P = 1.2e-13$).

Discussion

In this study, we characterized EBV-related EA-NPC and we identified four clusters with unique biological characteristics: *immune-active, defense-response, proliferation*, and *perineural-interaction/EBV-exhaustion*. The workflow of our analyses and main results are shown in Fig. 9.

The EA *immune-active* cluster presented a good prognosis and was characterized by the highest expression of the immune score, immune cells, immune hallmarks, and EBV-related pathways. In NEA cases, the immune activity was similar to that of its EA counterpart; however, the *immune active* cluster expressed more B cells and mast cells than the other clusters (see Fig. 7C) and did not express immune hallmarks (see Fig. 7D). The absence of the expression of immune hallmarks and low levels of T cells and NKT in NEA-NPC might be explained by the level of mast cells, according to literature exert an immune suppressive role.[46]

The EA *defense-response* cluster presented a good prognosis and was similar to *immune-active* in terms of the immune score, immune cells, and immune hallmarks. It also expressed other hallmarks, such as DNA damage, proliferation, and signaling, and EBV-related pathways. The NEA *defense-response* cluster showed a good immune cell expression profile with the presence of many immune cell types, but only one immune hallmark (allograft rejection pathway).

In agreement with worst prognosis, the EA *proliferation* cluster exhibited lower immune score and immune cell levels, and increased expression of proliferation hallmarks and EBV pathways. Similar to its counterpart, the NEA *proliferation* cluster showed low levels of immune score and immune cells and increased expression of proliferation hallmarks. Overall, the NEA *proliferation* cluster appeared to share the hallmarks of the EA *proliferation* and *defense-response* clusters; the NEA-NPC small sample size could explain this partial similarity.

The EA *perineural-interaction/EBV-exhaustion* cases considered in the Kaplan–Meier analysis (six of 40 patients with available clinical data)

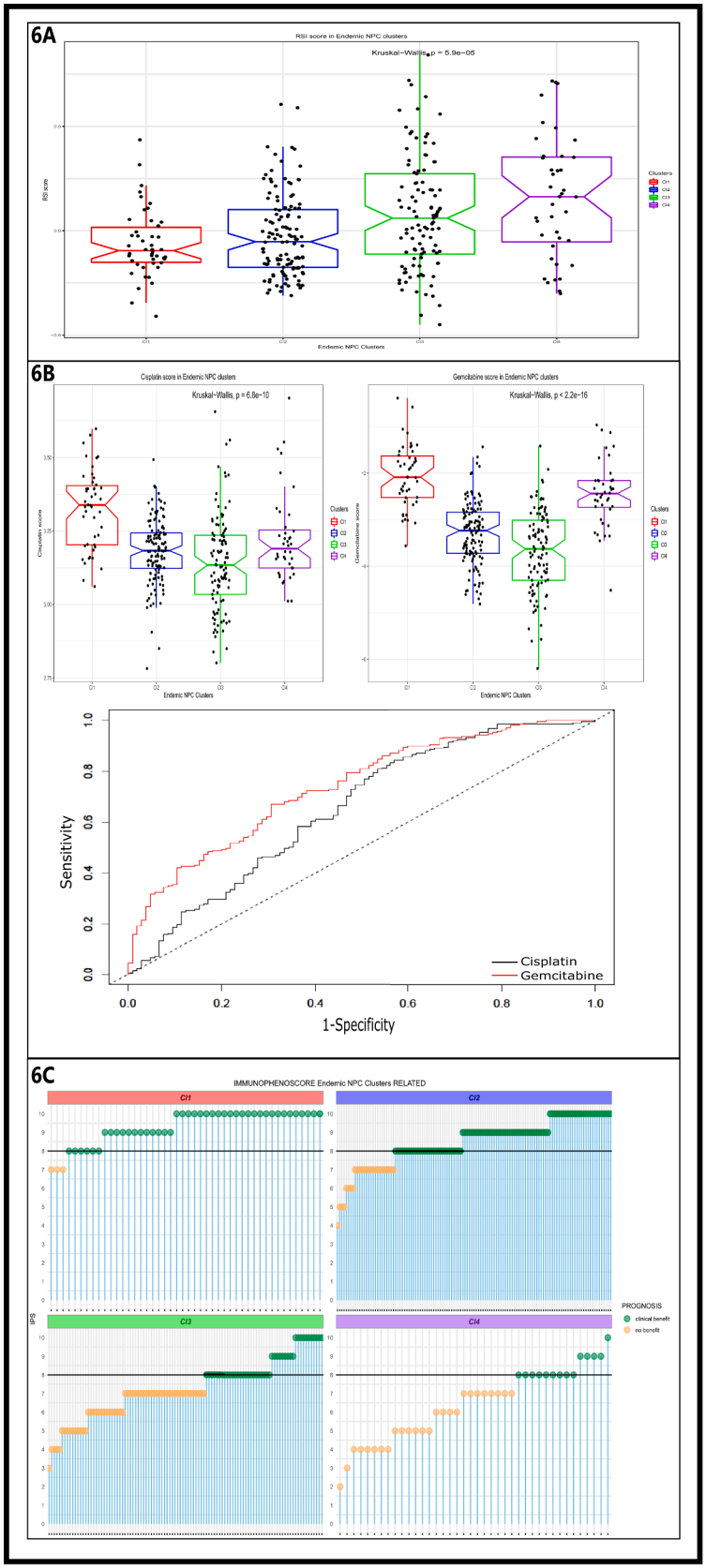

Fig. 6. *Prediction of treatment response in endemic NPC.* A. Radiosensitivity index. RSI was evaluated and found to be directly proportional to radioresistance (high index = radioresistance). Stratification by RSI reached $P = 5.9e\text{-}05$. B. Drug sensitivity was predicted for each case in the meta-analysis dataset. Two therapeutic agents were investigated (i.e., cisplatin and gemcitabine). Boxplots depict the predicted drug sensitivity in the four clusters ($P = 6.8e\text{-}10$ and 2.2e-16 for cisplatin and gemcitabine, respectively); the ROC curves estimate the prediction accuracy of the most sensitive subtype against the others. P value by Kruskal–Wallis test. C. IPS. The plots show the IPS score distributions for the four clusters.

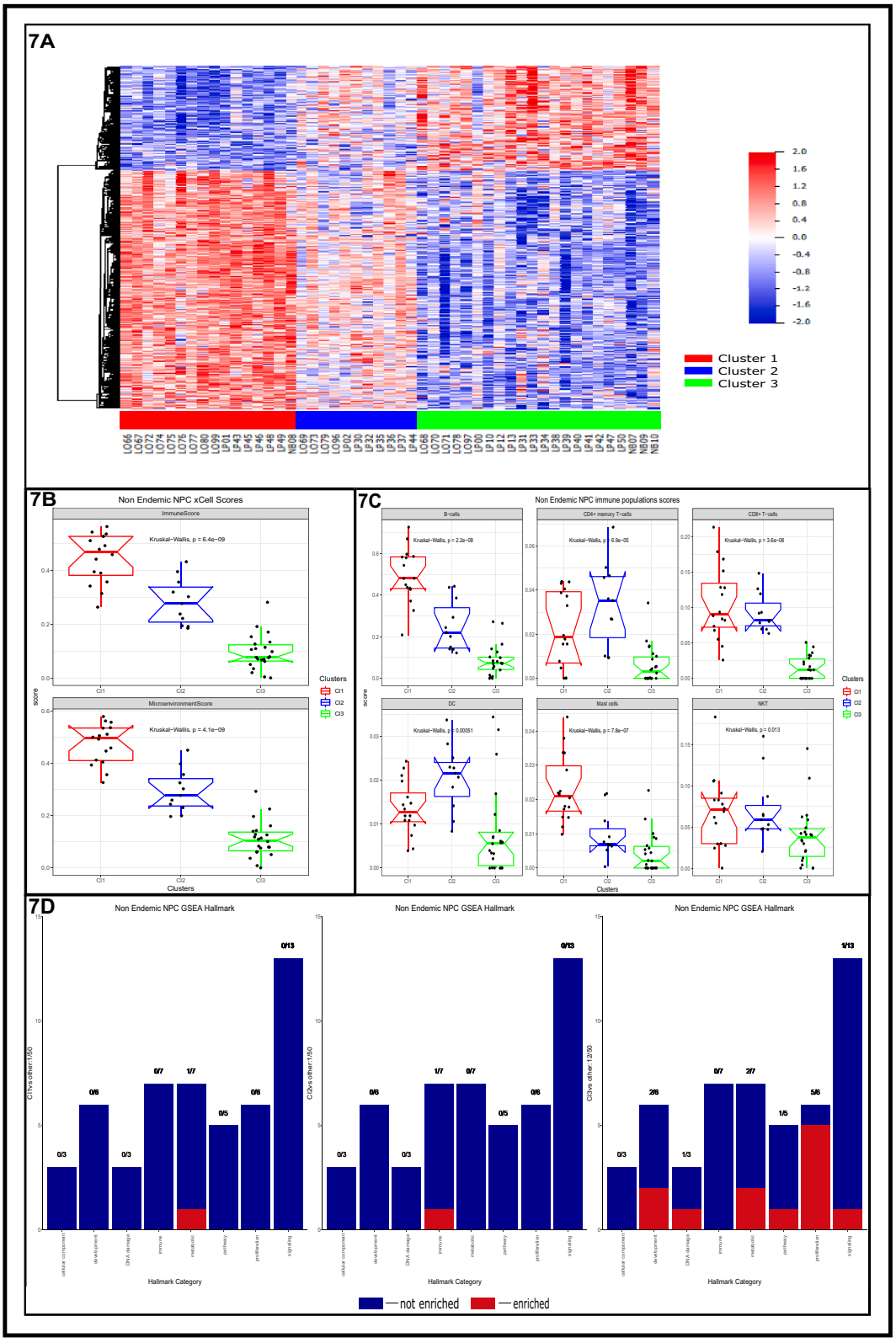

Fig. 7. *Identifying clusters and immune, and biological characteristics of non-endemic NPC.* A. Cluster prediction in non-endemic NPC. Cluster prediction disclosed three molecular clusters: Cl1 ($n = 16$), Cl2 ($n = 11$), Cl3 ($n = 22$), and Cl4 ($n = 1$) accounting for 32 %, 22 %, 44 %, and 2 % of the cohort, respectively. B. Immune and microenvironment scores. Immune ($P = 6.4e\text{-}09$) and microenvironment ($P = 4.1e\text{-}09$) components calculated by xCell for each tumor cluster. C. Expression profiles of immune cells. The abundance of B cells ($P = 2.2e\text{-}08$), CD4+ memory T cells ($P = 6.9e\text{-}05$), CD8+ Tcells ($P = 3.6e\text{-}08$), dendritic cells (DC) ($P = 0.00051$), mast cells ($P = 7.8e\text{-}07$), and natural killer T (NKT) cells ($P = 0.013$) were calculated using xCell for each tumor cluster. D. GSEA hallmark analyses of clusters. Hallmarks were divided into eight categories: cellular component, development, DNA damage, immune, metabolic, pathway, proliferation, and signaling. Gene expression analyses were performed for each cluster vs. others to analyze in which component one cluster is enriched in compared to others (see Supplementary Table 7 for the list of enriched pathways).

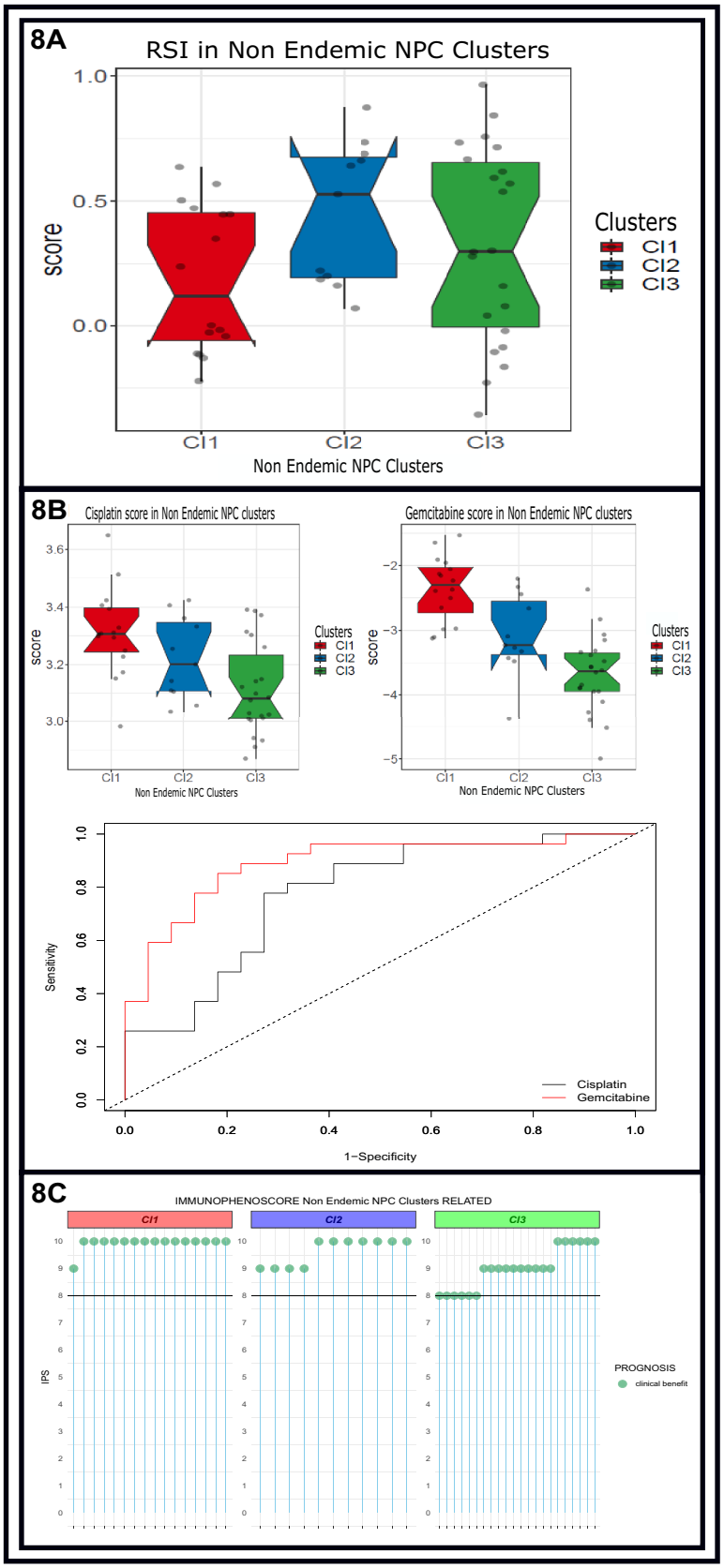

Fig. 8. *Prediction of treatment response in non-endemic NPC.* A. Radiosensitivity index. Stratification by RSI; $P = 0.0924$. B. Cisplatin and gemcitabine sensitivity were predicted for each case in the meta-analysis dataset. Boxplots depict the predicted drug sensitivity in the *immune-active, defense-response*, and *proliferation* clusters ($P = 0.00268$ and 1.46e-06 for cisplatin and gemcitabine, respectively); the ROC curves estimate the prediction accuracy of the most sensitive subtype against the others. P-value by Kruskal–Wallis test. C. IPS. The plots show the IPS score distribution in the three clusters.

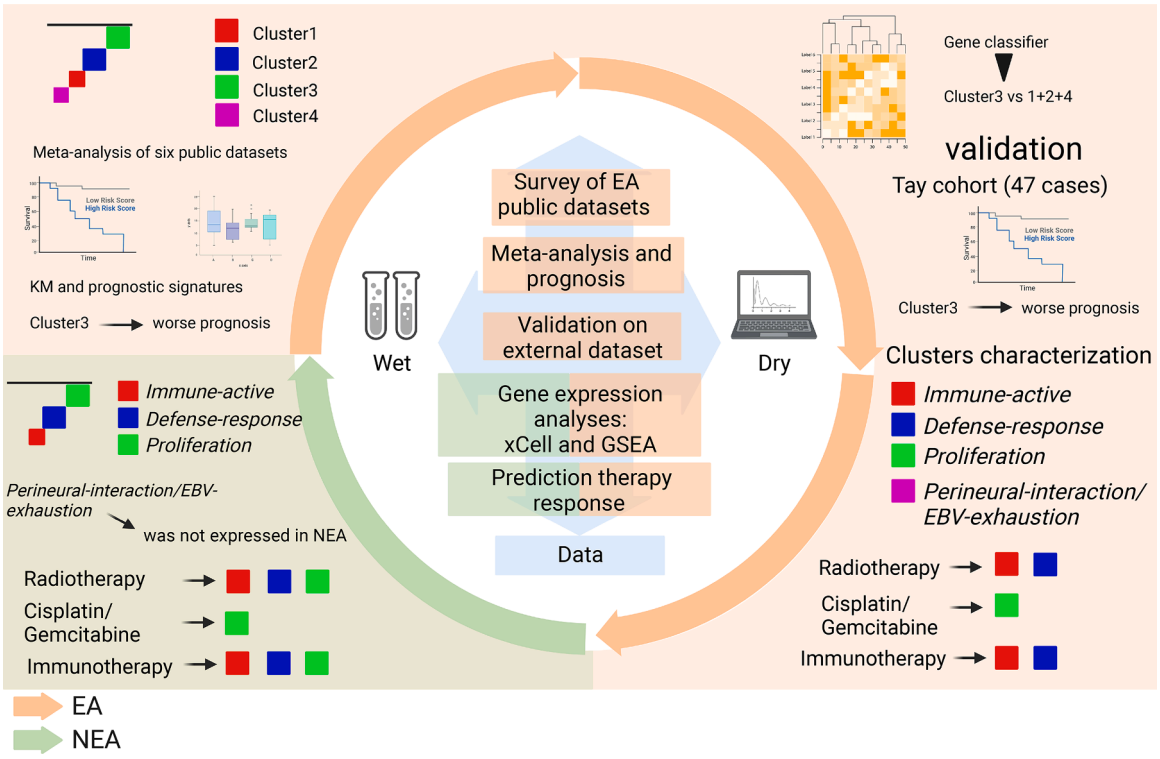

Fig. 9. *Biological/functional characterization of NPC: summary of the main steps and major results*. In the central area, the main applied techniques are shown and in the external quadrants, each step of the study is shown. EA-NPC (up left-quadrant): identification of four clusters in the meta-analysis of 6 datasets and cluster 3 as that with worst prognosis; EA-NPC (up right-quadrant): validation of worst prognosis of cluster 3 in an independent dataset; EA-NPC (down right-quadrant): identification of the main biological/functional characteristics of the clusters and their predicted sensitivity to different therapeutic strategies. NEA-NPC (down left-quadrant): identification of three clusters in a new Italian cohort of patients and their sensitivity to different therapeutic strategies. NPC: Nasopharyngeal Cancer; EA: Endemic Area (light pink); NEA: Non-Endemic Area (light gray). Created with BioRender.com.

presented the best prognosis. However, the biological/functional characterization of the *perineural-interaction/EBV-exhaustion* cluster suggested the worst prognosis; in fact, it express a low level of immune scores and anti-tumoral immune cells, a high level of MDSCs and exhibited expression of perineural pathways.

MDSCs suppress T-cell activity contributing the immune escape of tumors.[12] Perineural invasion, characterized by the invasion of cancer cells into the space surrounding a nerve, has been observed in many cancer types, including head and neck cancer and NPC, and is associated with poor prognosis in head and neck cancer.[47–48]

Although only EBV-positive patients were included in the EA meta-analysis, the *perineural-interaction/EBV-exhaustion* cluster did not show an upregulation of EBV pathways. The spanning tree analysis indicated that the *perineural-interaction/EBV-exhaustion* cluster was the most distant from normal phenotypes and the network map confirmed that the data cloud of this cluster was more distinct than that of the others. These results suggest that in the *perineural-interaction/EBV-exhaustion* different oncogenetic mechanisms of EBV exhaustion may be at play. We hypothesize a virus-related "hit and run theory", according to which such agents promote oncogenesis by inducing the accumulation of mutations, resulting in increased genomic instability; in this scenario, the virus would no longer be necessary for tumor maintenance and would be lost after tumor cell proliferation.[49–50] *Perineural-interaction/EBV-exhaustion* was not identified in the NEA cohort.

Similarly to patients with other cancer types, not all patients with NPC exhibit the same clinical outcomes after therapy. Risk stratification in patients with NPC is urgently needed to select better treatment plans (i.e., the addition of neoadjuvant or adjuvant chemotherapy), especially for NEA-NPC. According to the literature, upregulated GPX4 associated with EBV infection[39] leads to chemoresistance.[51] On the other hand, the data on the significant downregulation of the majority of a selected genes, including GPX4 (see Table 1), and of the four genesets all associated to an active EBV infection strongly suggested the switching off of EBV infection and support the final designation of Cl4 as *perineural-interaction/EBV-exhaustion.* Hence, *perineural-interaction/EBV-exhaustion*, which is present only in EA- NPC, is expected to lead to a worse prognosis owing to its biological and functional features, but it could be successfully treated using gemcitabine. The NEA-NPC treatment response analysis partially confirmed the EA-NPC results (see Fig. 9).

Although our analyses offer a tool for investigating NPC prognosis, further validation is needed to apply this finding in a clinical setting. This study has both strengths and limitations. Its first strength is the comprehensive meta-analysis of 314 EA-NPC cases belonging to six different datasets, four of which also included 35 normal tissue samples. This normal subset enabled us to assess the topological connections and biological distances between each NPC tumor subtype and the normal tissue. Additional strengths include the availability of previously published prognostic signatures and an external validation EA cohort with good clinical annotation.

The Italian cohort of NEA-NPC described herein could be considered the first available gene expression dataset, analysis focusing specifically on NEA-NPCs; in fact, Tay et al.[21] included in analyzed dataset of NPC (47 cases) both 12 NEA-NPC cases, but they did not analyze them separately from 35 EA-NPC cases they could also characterize. In addition, we collected samples from a unique NEA-NPC cohort with good clinical annotation, which enabled the preliminary analysis of the similarities and divergencies between EA-NPC and NEA-NPC gene-expression patterns.

The limitations of this study include the small number of available samples and few survival events that prevented survival prediction analysis. Considering the limitations EA-NPC, we should recognize that i) the amount of clinical information of EA datasets used for meta-

analysis is limited, and ii) the validation cohort consists of is mixed EA/NEA NPC. These limitations reduced the accuracy of the analysis of prognostic outcomes for each of the four clusters identified in EA-NPC. This limited the interpretation of the contradictory prognostic results obtained for the *perineural-interaction/EBV-exhaustion* cluster.

Our future goals differ between the incidence areas (EA and NEA), based on the reported data and their limitations. Regarding EA-NPC, access to a larger dataset with complete clinical variables would enable the validation of our data and confirmation of the presence and characteristics of the *perineural-interaction/EBV-exhaustion* cluster and its prognosis. In the NEA-NPC population, we are expanding sample collection to provide deeper insights into the molecular landscape and regulatory networks of gene expression pertaining to these tumors.

Conclusions

Our study provides a relevant biological overview of EBV-related NPC in patients from both EA and NEA. The immune microenvironment plays a critical role in NPC owing to the viral etiology of this malignancy. A *perineural-interaction/EBV-exhaustion* cluster in EA-NPC suggests an inactive EBV infection but more precise molecular analyses are needed to confirm this suggestion. Well characterized EA- and NEA-NPC retrospective and prospective cohorts would enable a validation of the results obtained herein.

CRediT authorship contribution statement

Deborah Lenoci: Formal analysis, Writing – original draft. **Carlo Resteghini:** Data curation, Formal analysis. **Mara S. Serafini:** Formal analysis. **Federico Pistore:** Formal analysis, Visualization, Software. **Silvana Canevari:** Investigation, Writing – original draft. **Brigette Ma:** Investigation. **Stefano Cavalieri:** Formal analysis, Investigation, Writing – review & editing. **Salvatore Alfieri:** Data curation, Investigation. **Annalisa Trama:** Conceptualization, Investigation, Writing – review & editing. **Lisa Licitra:** Conceptualization, Funding acquisition, Investigation, Supervision, Writing – review & editing. **Loris De Cecco:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgements

The authors gratefully thank Arianna Micali for technical assistance.

Availability of data and materials

The datasets of endemic NPC analyzed during the current study are available on the GE Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/).

Microarray data of NEA-NPC were compliant to MIAME (Minimum Information about a Microarray Experiment) and the GE profiles of our non-endemic NPC cohort were deposited on GEO (accession number GSE208281).

Consent for publication

Not applicable.

Ethics approval and consent to participate

This study was approved by the Institutional Ethical Review Board of the IRCCS Istituto Nazionale dei Tumori, and all patients provided a signed informed consent.

Funding

This work was supported by 5 × 1000 funds (2017)- Italian Ministry of Health, financial support for healthcare research).

This work was partially supported by Associazione Italiana Ricerca Cancro (AIRC IG23573 to L.D.C.).

Conflict of Interest

The authors declare that no competing interests exist.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.trsl.2023.10.004.

References

1. Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J. Nasopharyngeal carcinoma. *Lancet*. 2019;394(10192):64–80. https://doi.org/10.1016/S0140-6736(19)30956-0. Epub 2019 Jun 6PMID: 31178151.
2. Chang ET, Ye W, Zeng YX, Adami HO. The evolving epidemiology of nasopharyngeal carcinoma. *Cancer Epidemiol Biomarkers Prev*. 2021;30(6):1035–1047. https://doi.org/10.1158/1055-9965.EPI-20-1702. Epub 2021 Apr 13. PMID: 33849968.
3. Gatta G, Capocacia R, Botta L, et al. Burden of centralised treatment in Europe of rare tumours: results of RARECAREnet-a population-based study. *Lancet Oncol*. 2017;18:1022–1039.
4. Lo KW, To KF, Huang DP. Focus on nasopharyngeal carcinoma. *Cancer Cell*. 2004;5(5):423–428. https://doi.org/10.1016/s1535-6108(04)00119-9. PMID: 15144950.
5. Russo A, Crosignani P, Berrino F, et al. Incidence of cancer in migrants: data of the Lombardy tumor registry. *Epidemiol Prev*. 1994;18:125–132.
6. Murata T, Sato Y, Kimura H. Modes of infection and oncogenesis by the Epstein-Barr virus. *Rev Med Virol*. 2014.
7. Farrow DC, Vaughan TL, Berwick M, Lynch CF, Swanson GM, Lyon JL. Diet and nasopharyngeal cancer in a low-risk population. *Int J Cancer*. 1998;78(6):675–679. https://doi.org/10.1002/(sici)1097-0215(19981209)78:6<675::aid-ijc2>3.0.co;2-j. PMID: 9833758.
8. Wang Y, Zhang Y, Ma S. Racial differences in nasopharyngeal carcinoma in the United States. *Cancer Epidemiol*. 2013;37(6):793–802. https://doi.org/10.1016/j.canep.2013.08.008. Epub 2013 Sep 12. PMID: 24035238; PMCID: PMC3851929.
9. Tsao SW, Tsang CM, Lo KW. Epstein-Barr virus infection and nasopharyngeal carcinoma. *Philos Trans R Soc Lond B Biol Sci*. 2017;372, 20160270. https://doi.org/10.1098/rstb.2016.0270, 1732PMID: 28893937; PMCID: PMC5597737.
10. Islam KA, Chow LK, Kam NW, et al. Prognostic Biomarkers for survival in nasopharyngeal carcinoma: a systematic review of the literature. *Cancers*. 2022;14(9):2122. https://doi.org/10.3390/cancers14092122. PMID: 35565251; PMCID: PMC9103785.
11. Fountzilas G, Psyrri A, Giannoulatou E, et al. Prevalent somatic BRCA1 mutations shape clinically relevant genomic patterns of nasopharyngeal carcinoma in Southeast Europe. *Int J Cancer*. 2018;142(1):66–80. https://doi.org/10.1002/ijc.31023. Epub 2017 Sep 30. PMID: 28857155.
12. Wu Y, Yi M, Niu M, Mei Q, Wu K. Myeloid-derived suppressor cells: an emerging target for anticancer immunotherapy. *Mol Cancer*. 2022;21(1):184. https://doi.org/10.1186/s12943-022-01657-y. PMID: 36163047; PMCID: PMC9513992.
13. Bossi P, Chan AT, Licitra L, ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org; EURACAN. Nasopharyngeal carcinoma: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Ann Oncol*. 2021;32(4):452–465. https://doi.org/10.1016/j.annonc.2020.12.007. Epub 2020 Dec 25. PMID: 33358989.
14. Zhang Y, Chen L, Hu GQ, et al. Gemcitabine and cisplatin induction chemotherapy in nasopharyngeal carcinoma. *N Engl J Med*. 2019;381(12):1124–1135.
15. Xu JY, Wei XL, Wang YQ, Wang FH. Current status and advances of immunotherapy in nasopharyngeal carcinoma. *Ther Adv Med Oncol*. 2022;14. https://doi.org/10.1177/17588359221096214. PMID: 35547095; PMCID: PMC9083041.
16. De Cecco L, Nicolau M, Giannoccaro M, et al. Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data. *Oncotarget*. 2015;6(11):9627–9642. https://doi.org/10.18632/oncotarget.3301. PMID: 25821127; PMCID: PMC4496244.
17. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*. 2001;29(4):365–371. https://doi.org/10.1038/ng1201-365. PMID: 11726920.
18. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*. 2013;41(Database issue):D991–D995.
19. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–127.
20. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf*. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.
21. Tay JK, Zhu C, Shin JH, et al. The microdissected gene expression landscape of nasopharyngeal cancer reveals vulnerabilities in FGF and noncanonical NF-κB signaling. *Sci Adv*. 2022 Apr 8;8(14). https://doi.org/10.1126/sciadv.abh2445. eabh2445Epub 2022 Apr 8. PMID: 35394843; PMCID: PMC8993121.

22. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. https://doi.org/10.1093/nar/gkv007.
23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616. Epub 2009 Nov 11. PMID: 19910308; PMCID: PMC2796818.
24. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15:R29. https://doi.org/10.1186/gb-2014-15-2-r29.
25. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26:1572–1573.
26. Nicolau M, Tibshirani R, Børresen-Dale AL, Jeffrey SS. Disease-specific genomic analysis: identifying the signature of pathologic biology. *Bioinformatics*. 2007;23(8): 957–965. https://doi.org/10.1093/bioinformatics/btm033. Epub 2007 Feb 3. PMID: 17277331.
27. Locati LD, Serafini MS, Iannò MF, et al. Mining of self-organizing map gene-expression portraits reveals prognostic stratification of HPV-positive head and neck squamous cell carcinoma. *Cancers*. 2019;11(8):1057. https://doi.org/10.3390/cancers11081057. PMID: 31357501; PMCID: PMC6721309.
28. Löffler-Wirth H, Kalcher M, Binder H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. *Bioinformatics*. 2015;31(19):3225–3227. https://doi.org/10.1093/bioinformatics/btv342. Epub 2015 Jun 10. PMID: 26063839.
29. 21R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2022.
30. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-ArrayTools. *Cancer Inform*. 2007;4(3):11–27. PMID: 19455231; PMCID: PMC2675854.
31. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4:1686. https://doi.org/10.21105/joss.01686.
32. Draw Freely | Inkscape. Accessed 1 June, 2022; Available online: https://inkscape.org/.
33. Zhao Lan, Lee Victor H F, Ng Michael K, Yan Hong, Bijlsma Maarten F. Molecular subtyping of cancer: current status and moving toward clinical applications. *Briefings Bioinf*. March 2019;20(2):572–584. https://doi.org/10.1093/bib/bby026.
34. Li Y, Chung G, Lui V, et al. Exome and genome sequencing of nasopharynx cancer identifies NF-κB pathway activating mutations. *Nat Commun*. 2017;8:14121. https://doi.org/10.1038/ncomms14121.
35. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Compu Appl Math*. 1987;20:53–56.
36. Carenzo A, Pistore F, Serafini MS, Lenoci D, Licata AG, De Cecco L. hacksig: a unified and tidy R framework to easily compute gene expression signature scores. *Bioinformatics*. 2022;38(10):2940–2942. https://doi.org/10.1093/bioinformatics/btac161. PMID: 35561166; PMCID: PMC9113261.
37. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1).
38. Yang T, You C, Meng S, Lai Z, Ai W, Zhang J. EBV Infection and its regulated metabolic reprogramming in nasopharyngeal tumorigenesis. *Front Cell Infect Microbiol*. 2022;12, 935205. https://doi.org/10.3389/fcimb.2022.935205. PMID: 35846746; PMCID: PMC9283984.
39. Yuan L, Li S, Chen Q, et al. EBV infection-induced GPX4 promotes chemoresistance and tumor progression in nasopharyngeal carcinoma. *Cell Death Differ*. 2022;29(8): 1513–1527. https://doi.org/10.1038/s41418-022-00939-8. Epub 2022 Feb 1. PMID: 35105963; PMCID: PMC9346003.
40. Aran D, Hu Z, Butte AJ. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Boil*. 2017;18:220. https://doi.org/10.1186/s13059-017-1349-1.
41. Torres-Roca JF, Eschrich S, Zhao H, et al. Prediction of radiation sensitivity using a gene expression classifier. *Cancer Res*. 2005;65(16):7169–7176. https://doi.org/10.1158/0008-5472.CAN-05-0656. PMID: 16103067.
42. Geeleher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One*. 2014;9(9), e107468. https://doi.org/10.1371/journal.pone.0107468. PMID: 25229481; PMCID: PMC4167990.
43. Charoentong P, Finotello F, Angelova M, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep*. 2017;18(1):248–262. https://doi.org/10.1016/j.celrep.2016.12.019. PMID: 28052254.
44. Liu J, Xu J, Luo B, et al. Immune landscape and an RBM38-associated immune prognostic model with laboratory verification in malignant melanoma. *Cancers*. 2022;14(6):1590. https://doi.org/10.3390/cancers14061590. PMID: 35326741; PMCID: PMC8946480.
45. Su ZY, Siak PY, Leong CO, Cheah SC. Nasopharyngeal carcinoma and its microenvironment: past, current, and future perspectives. *Front Oncol*. 2022;12, 840467. https://doi.org/10.3389/fonc.2022.840467. PMID: 35311066; PMCID: PMC8924466.
46. Komi DEA, Redegeld FA. Role of mast cells in shaping the tumor microenvironment. *Clin Rev Allergy Immunol*. 2020;58(3):313–325. https://doi.org/10.1007/s12016-019-08753-w. PMID: 31256327; PMCID: PMC7244463.
47. Erin N, Shurin GV, Baraldi JH, Shurin MR. Regulation of carcinogenesis by sensory neurons and neuromediators. *Cancers*. 2022;14(9):2333. https://doi.org/10.3390/cancers14092333. PMID: 35565462; PMCID: PMC9102554.
48. Bakst RL, Glastonbury CM, Parvathaneni U, Katabi N, Hu KS, Yom SS. Perineural invasion and perineural tumor spread in head and neck cancer. *Int J Radiat Oncol Biol Phys*. 2019;103(5):1109–1124. https://doi.org/10.1016/j.ijrobp.2018.12.009. Epub 2018 Dec 15. PMID: 30562546.
49. Ferreira DA, Tayyar Y, Idris A, McMillan NAJ. A "hit-and-run" affair - A possible link for cancer progression in virally driven cancers. *Biochim Biophys Acta Rev Cancer*. 2021;1875(1), 188476. https://doi.org/10.1016/j.bbcan.2020.188476. Epub 2020 Nov 10. PMID: 33186643.
50. Mundo L, Del Porro L, Granai M, et al. Frequent traces of EBV infection in Hodgkin and non-Hodgkin lymphomas classified as EBV-negative by routine methods: expanding the landscape of EBV-related lymphomas. *Mod Pathol*. 2020;33(12): 2407–2421. https://doi.org/10.1038/s41379-020-0575-3. Epub 2020 Jun 1. Erratum in: Mod Pathol. 2020 Jun 29;: PMID: 32483241; PMCID: PMC7685982.
51. Yuan L, Li S, Chen Q, et al. EBV infection-induced GPX4 promotes chemoresistance and tumor progression in nasopharyngeal carcinoma. *Cell Death Differ*. 2022;29: 1513–1527. https://doi.org/10.1038/s41418-022-00939-8.