Integration of Transcriptional and Mutational Data Simplifies the Stratification of Peripheral T-Cell Lymphoma

Francesco Maura[1-3]*, Luca Agnelli[3-4]*#, Daniel Leongamornlert[2], Niccolò Bolli[2,3,6], John Chan[5], Anna Dodero[6], Cristiana Carniti[6], Tayla B. Heavican[7], Alessio Pellegrinelli[8], Giancarlo Pruneri[8], Adam Butler[2], Shriram G Bhosle[2], Annalisa Chiappella[9], Alice Di Rocco[10], Pier Luigi Zinzani[11], Francesco Zaja[12], Roberto Piva[13], Giorgio Inghirami[13,14], Wenyi Wang[15], Teresa Palomero[16], Javeed Iqbal[7], Antonino Neri[3-4], Peter J Campbell[2] and Paolo Corradini[3,5#]

[1]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[2]Cancer, Ageing and Somatic Mutation Programme, Wellcome Trust Sanger Institute, Hinxton, United Kingdom;
[3]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy;
[4]Hematology, Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy;
[5]Department of Pathology, City of Hope National Medical Center, Duarte, CA;
[6]Department of Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy;
[7]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE;
[8]Department of Pathology and Laboratory Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy;
[9]Department of Hematology, Azienda Ospedaliera Città della Salute e della Scienza, Torino, Italy;
[10]Sapienza University of Rome, Rome, Italy;
[11]Institute of Hematology, University of Bologna, Bologna, Italy
[12]Clinical Ematologica, DAME, University of Udine, Udine, Italy;
[13]Center for Experimental Research and Medical Studies, University of Torino, Department of Molecular Biotechnology and Health Sciences, Torino, Italy;
[14]Pathology and Laboratory Medicines, Weill Cornell Medical College, NY;
[15]The University of Texas MD Anderson Cancer Center, Houston, TX;
[16]Institute for Cancer Genetics, Columbia University, New York, NY.

*The first two authors equally contributed

Running title: Peripheral T-Cell Lymphoma transcriptional stratification

Key words: PTCL, AITL, ALCL, gene expression, CIBERSORT

Corresponding Author:
Luca Agnelli
Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy
F. Sforza, 35, 20122 Milan Italy
luca.agnelli@unimi.it
Tel (+39)0255033328
Fax: (+39)0255034571

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ajh.25450

Abstract

The histological diagnosis of peripheral T-cell lymphoma (PTCL) can represent a challenge, particularly in the case of closely related entities like angioimmunoblastic T-lymphoma (AITL), PTCL-not otherwise specified (PTCL-NOS), and ALK-negative anaplastic large-cell lymphoma (ALCL). Although gene expression profiling and next generations sequencing have been proven to define specific features recurrently associated with distinct entities, genomic-based stratifications have not yet led to definitive diagnostic criteria and/or entered into the routine clinical practice.

Herein, to improve the current molecular classification between AITL and PTCL-NOS, we analyzed the transcriptional profiles from 503 PTCLs stratified according to their molecular configuration and integrated them with genomic data of recurrently mutated genes (*RHOA*^{G17V}, *TET2*, *IDH2*^{R172}, *DNMT3A*) in 53 cases (39 AITLs and 14 PTCL-NOSs) included in the series. Our analysis unraveled that the mutational status of *RHOA*^{G17V}, *TET2* and *DNMT3A* poorly correlated, individually, with peculiar transcriptional fingerprints. Conversely, in *IDH2*^{R172} samples a strong transcriptional signature was identified that could act as a surrogate for mutational status. The integrated analysis of clinical, mutational and molecular data led to a simplified 19-gene signature that retains high accuracy in differentiating the main nodal PTCL entities. The expression levels of those genes were confirmed in an independent cohort profiled by RNA-sequencing.

Introduction

Peripheral T-cell lymphomas (PTCL) represent a heterogeneous group of nodal and extra-nodal mature T-cell Non-Hodgkin lymphomas (T-NHL) accounting for approximately 10-15% of all lymphoma in the Western countries.[1] Histological diagnosis of the various PTCL subtypes can still represent a challenge and difficulties occur in particular for those samples with borderline features between angioimmunoblastic T-cell lymphoma (AITL), follicular T-cell lymphoma and PTCL not otherwise specified (PTCL-NOS).[1,2] Previous studies have shown that these entities might bear distinct transcriptional and mutational profiles.[3-8] Gene expression profiling has the potential to represent the gold standard for classification, but its clinical use is still limited due to technical limits and to the absence of a manageable and practical short consensus gene signature. Recent advances in next generation sequencing (NGS) allowed the discovery of recurrently mutated genes (*RHOA*, *TET2*, *DNMT3A*) in approximately 60-70% of AITL and in 20-30% of PTCL-NOS, changing in part this landscape.[6,9-12] Notably, 20-30% of AITL cases can carry hotspot *IDH2*^{R172} mutations that are virtually absent in PTCL-NOS.[9] Nevertheless, these findings have not yet significantly impacted diagnosis in daily clinical practice, which largely relies on morphological and immunophenotypic features of tumor cells.[1] Moreover, albeit some mutations appear to be linked to distinct transcriptional signature(s),[6] the full potential of an integrated genotypic-transcriptomic analysis has not been thoroughly tested in PTCLs. Herein, we collected a large gene expression profiling dataset of PTCLs, and performed an integrative analysis with available mutational data to improve our understanding of the underlying structure of sample clusters, with potential implications for disease classification particularly at the interface between of AITL and PTCL-NOS lymphomas.

Methods

Dataset

We analysed 503 PTCL, univocally acquired from 8 studies (GSE6338, GSE14879, GSE19067, GSE19069, GSE58445 and GSE65823 at http://www.ncbi.nlm.nih.gov/geo/;

ETABM702 and ETABM783 at https://www.ebi.ac.uk/arrayexpress, **Supplementary Figure 1**). Normalized data were extracted using RMA procedure and the annotation available at http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/21.0.0/entrezg.asp. A batch-effect correction was applied using *ComBat* function in *sva* package for R software. The whole data set with all available clinical and genomic information acquired was uploaded to https://github.com/emacgene/PTCL.

Transcriptional and statistical analysis

The statistical models that allow measuring the association between mutations and gene expression was firstly described in Gerstung *et al*.[13] and here adapted to 39 AITLs and 14 PTCL-NOSs for whom mutational data for *IDH2*^{R172}, *DNMT3A*, *TET2* and *RHOA*^{G17V} were available.[6]

ConsensusClusterPlus package for R [14] was used to determine the significance and robustness of natural grouping of patients based on selected transcriptional data, using Ward and Euclidean as linkage and distance metrics, respectively.

CIBERSORT analysis was performed as previously described, using standard procedure and LM22 signature.[15] The CIBERSORT different contribution for each signature was then tested by *pairwise.wilcox.test* R function. Benjamini-Hochberg correction was used for multiple testing adjustment. The pathway enrichment analysis was performed using different modalities. The *tmod* R package was used on *limma*-derived signatures to decipher whether clusters deregulate blood cell-associated transcriptional modules described by Chaussabel *et al.*[16] and by Li *et al.*,[17] according to the procedure described by Weiner et al.[18] The full analysis process written in R is provided in **Supplementary Data** at https://github.com/emacgene/PTCL.

Data from 34 previously published RNAseq samples have been imported upon obtainment of accession to dbGap dataset #phs000689.v1.p1.[9] RNAseq mapping and

read counts were processed using iRAP pipeline,[19] and the RNAseq raw expression data were normalized as previously described.[20]

Results

According to the most recent updates in the T-cell lymphoma classification[1] and the gene-expression based classification criteria as in Iqbal *et al*,[21] we generated the whole transcriptional profile of a dataset including 127 AITL, 144 PTCL-NOS, 56 *ALK*+ (Anaplastic Large-Cell Lymphoma) ALCL, 96 *ALK*- ALCL, 21 Adult T-Cell Lymphoma (ATLL), 59 NK/T-cell lymphomas (**Figure 1a**, from here on named as "molecular classification"). Both unsupervised hierarchical clustering and principal component analysis on the most variable genes (exceeding the mean an average 2-fold across the dataset) showed that the known entities, such as *ALK*+ ALCLs and ATLL were associated with markedly distinct signatures; notably, the transcriptional portrait of AITL and PTCL-NOS displayed a considerable overlap (**Figure 1b-c**). For completeness, the stability of the identified clusters was tested to unravel the most relevant overlapping and to describe the phenotypes characterized by overall uniform transcriptional pattern; the whole confusion matrix was reported in **Supplementary Data**. Overall, our meta-analysis of the largest gene-expression profiling dataset tested to date showed that the consensus between transcriptomic analysis and histology is still imperfect, opening the field for the search of additional features that could improve diagnostic accuracy.

Definition of gene signature associated with molecular classification and mutational status

To search for additional features that could better define the AITL and PTCL-NOS entities, we investigated whether recurrent mutations or clinical features might correlate with a specific transcriptional pattern that could be used for stratification purposes. To this aim, we adapted a recently published analysis[13] to a set of 39 AITL and 14 PTCL-NOS cases for which mutational data for *IDH2*^{R172}, *DNMT3A*, *TET2* and *RHOA*^{G17V} were available.[6] This analysis allowed the creation of a linear model that associated gene-by-gene expression to

putative predictor variables, namely the molecular classification histotype, mutations, age and gender of each patient. We identified 221 modulated genes across the training dataset at false discovery rate (FDR) <1% (**Figure 1a**), among which 30 of them emerged as significantly associated to one distinct variable (**Figure 2b**). Gender selectively impacted 14 genes located on the X or Y chromosomes, which were discarded from further analysis. Histotype and *IDH2*^{R172} mutations specifically impacted the expression of 13 and 3 genes, respectively, while notably the mutational status of *RHOA*^{G17V}, *TET2* and *DNMT3A* (**Figure 2b**) were not associated with any distinct gene expression change. The 3 genes associated with *IDH2*^{R172} mutation were *ID2*, *NETO2* and *SLC5A3*. Three out of 13 histotype-associated genes were also reported in a previous large gene expression signature that discriminates PTCL-NOS and AITL (*ROBO1*, *ARHGEF10* and *EFNB2*)[3]. Based on these findings, we wondered whether a minimal signature including these 16 genes might effectively stratify AITL and PTCL-NOS cases. A leave-one-out cross validation (LOOCV) procedure run using linear discriminant analysis[22] supported the robustness of this 16-gene model, indicating overall 86.4% accuracy (sensitivity 86.9%; specificity 85.9%) in discriminating AITL from PTCL-NOS.

Prompted by the evidence that this small 16-gene signature robustly stratified these two entities, we hypothesized that combining this signature with the previously described 3-gene *ALK*- ALCL signature[22] may lead to a 19-gene model able to improve the stratification of all major PTCL entities (*ALK*- ALCL, PTCL-NOS and AITL). Again, a LOOCV procedure indicated that the 19-gene model was robust (overall accuracy 80.1%).

Assessment of the 19-gene signature expression level in an independent cohort

To test the reproducibility of the signature, we analyzed the expression levels of the genes for which data were available in 30 PTCL-NOSs cases (11 AITL, 11 PTCL-NOS, 8 *ALK*- ALCL) from a previously published RNAseq dataset.[9] To this aim, we first classified the samples according to the described molecular classification using a LOOCV procedure and the published gene signature[3] that has been used to test build the expression/mutation

model (**Supplementary data**). Of the 19 genes included in our model, 17 were mapped and detected in the 30 samples, overall retaining significant difference in the group distribution (**Supplementary Figure 2**).[3,22]

Patients' stratification according to the 19-gene signature

To further investigate the significance and robustness of natural grouping of patients based on selected 19-gene transcriptional data, we applied the *ConsensusClusterPlus* package to our original cohort, using Ward and Euclidean as linkage and distance metrics, respectively. This analysis led to the recognition of a minimum of five distinct subsets at the highest significance (**Figure 2c**), whose features are summarized in **Table 1**. The first group (C-1; 87 cases) mostly included AITL samples (93%; 81/87) enriched for *IDH2*^{R172} (16/36; 44%) and *RHOA*^{G17V} (23/29; 79%) mutations. The second (C-2; 103 samples), where the 19-gene expression pattern resembled that in C-1 but at lower levels, included samples annotated either as PTCL-NOS (64/103; 62%) or AITL (33/103; 32%). Interestingly, in the C-2 group, despite the *IDH2*^{R172} mutation was detected in only one out of 23 cases where the mutational status was known (p=0.001 if compared to C-1), a very high prevalence of *RHOA*^{G17V} mutations was observed (15/29; 52%). *RHOA*^{G17V} mutation occurrence was, moreover, significantly higher than what observed in cluster 3 and 4 (p<0.005). Despite their high prevalence, *RHOA*^{G17V} mutations had a lower impact on transcription (as evidenced by **Figure 2b**), likely because they were prevalently detected at sub-clonal level in a significant fraction of these cases, and this may explain their specific lower impact on transcription (as evidenced by **Figure 2b**). The third cluster (C-3; 32 samples) included patients who showed the lowest 19-gene expression levels overall, and an equal admixture of the three main PTCL entities. Group 4 (C-4; n=63) included mainly PTCL-NOS (52/63; 82%), showing low prevalence of *RHOA*^{G17V} and *TET2* mutations compared either to C-1 (p<0.0001 and p=0.0002, respectively) or C-2 (p=0.005 and 0.03 respectively). In the last cluster (C-5; n=55) *ALK*- ALCL cases (47/69; 68%) were over-represented, confirming the strong

association between these lymphomas and the previously published expression pattern of *TNFRSF8*, *BATF3* and *TMOD1*.[22]

Functional annotation, gene-expression based estimation of microenvironment composition and clinical relevance of 19-gene associated groups

To decipher whether these clusters might be characterized by common transcriptional behavior, we first investigated if they showed pathway enrichments. To this aim, *tmod* R package was used on *limma*-derived signatures to query whether clusters deregulated specific transcriptional modules associated with distinct blood cells. This analysis suggested significant enrichment of B-cell related pathways in C-1 and C-2, whereas C-5 showed enrichment for stimulated CD4+ T cells associated pathways, and significantly lower involvement of T-cell differentiation and T-cell activation pathways (**Supplementary Figure 3**). To understand whether this might be the consequence of a somehow unbalanced microenvironment composition, we then applied CIBERSORT,[23] an analytical tool developed to provide an estimation of the abundances of specific cell types in a mixed cell population based on gene expression data. CIBERSORT analysis was performed as previously described, using standard procedure and the LM22 signature (**Supplementary Figure 4**).[23] Notably, we observed a prevalence of pathways associated with plasma cells in C-1 and C-2, in agreement with the notion that AITL are often enriched in B-cells/plasma cells.[3] Also of interest, CIBERSORT analysis evidenced a slight prevalence of activated NK cell-like profiles in the C-3 cluster, overall suggesting that this cluster might be considered a spurious entity due to partial contamination of NK-associated signatures; and higher presence of activated memory T-CD4 lymphocytes in the *ALK*- ALCL cluster (C-5), supporting previous evidences that CD4+ expression characterizes almost all the anaplastic lymphomas,[24] and is not exclusively dependent on the ectopic expression of the ALK protein.

Finally, we tested whether the five clusters were associated to different prognosis in the 239 patients for whom outcome data were available. No significant global differences in

prognosis could be appreciated; however, a marginally poorer prognosis was identified in C-3 patients (*p*=0.0614, **Supplementary Data** and **Supplementary Figure 5**).

Discussion

Gene expression has emerged as one of the most robust and reliable approaches to differentiate human lymphomas. This has also been the case of T-cell lymphomas, where gene expression-based strategies have allowed the distinction of closely related entities like AITL and PTCL-NOS. [3,21,22,25-28] However, gene expression profiling from formalin fixed paraffin embedded samples has known technical artifacts[29] and analysis is not standardized. Thus, the molecular/expression stratification of PTCL has not entered yet into routine clinical practice. Recently, novel technologies (e.g. nanoString) have provided reproducible and feasible quantification of specific transcripts from FFPE samples, particularly for diffuse large B-cell lymphoma.[30-32] However, this approach has not yet been tested in PTCL, likely because a short and robust list of differentially expressed genes has not emerged yet.

To bridge this gap, we performed a meta-analysis of a large PTCL series, assessing the molecular profile of the main PTCL subgroups and defining a manageable list of significant differentially expressed genes. By integrating transcriptional and NGS data, we were able to discover the transcriptional impact of the recurrent mutations in AITLs and PTCL-NOSs. Interestingly, *TET2*, *DNMT3A*, *RHOA*^{G17V} mutations were not associated with a distinct gene expression signature. In particular, *RHOA*^{G17V} mutations were detected at sub-clonal level in a significant fraction of these cases (43%)[6] and this may explain its low impact on transcription[6]. On the other hand, *TET2* and *DNMT3A* mutations are likely clonal mutations arising from hematopoietic stem cells involved in clonal hemopoiesis;[33-35] therefore, these mutations could induce gene expression signatures shared by other PTCL subtypes and masked by the molecular classification, subsequently limiting its extraction through our statistical process. Conversely, in line with a previous report,[6] the *IDH2*^{R172} mutation significantly correlated with a distinct expression signature independently from the molecular subgroup. Through the integration pathological data, transcriptional and NGS

analyses we then defined a short list of 19 genes whose differential expression divided the samples into 5 clusters, strongly associated with distinct PTCL entities. Specifically, C-1 and C-2 were characterized by most of the AITL hallmarks. In addition, despite their limited gene expression impact, the great majority of *RHOA*^{G17V} and *TET2* mutations were grouped together, confirming their potential utility in the diagnostic process of these lymphomas. A significant fraction of PTCL-NOSs were included in these two "AITL clusters". Most of these cases showed high prevalence of *RHOA*^{G17V} and *TET2* mutations and expressed some distinct, and in part previously reported, AITL-associated genes. Unfortunately, missing paired molecular/genomic data in a significant portion of the dataset prevented the possibility to define a robust and definitive molecular classifier able to dissect this "grey zone" of PTCL classification.

Nonetheless, 6 of the 16-gene signature were previously described as T Follicular Helper (TFH)-phenotype related[8]. Unfortunately, no cases could be *a priori* classified as nodal PTCL-NOS with (TFH) phenotype, since WHO has introduced this category after the original studies were performed[1] and phenotypic data were not available. However, these 6 genes were significantly overexpressed in C-1 and C-2, in line with the recent WHO classification where PTCL TFH and AITL are grouped under an unique umbrella category.

Finally, based both on the present and recently published[3,21,36] data, PTCL-NOS cases in cluster C-4 likely may represent a distinct biologically entity. In line with previous findings,[4] cases within this cluster had a distinct genotypic and transcriptional pattern, mostly characterized by significant enrichment in downregulated genes.

Taken together, our analyses based on a small 19-gene model offered a more insightful perspective of the dissection of the complex nodal PTCL entities. We suggest that, in the future, the integration of transcriptional and genotypic data may improve the identification of clinic-pathological entities, contributing to the current diagnostic approach of PTCLs and laying the basis for effective treatments through the identification of specifically and recurrently dysregulated pathways.

Acknowledgment

NB is funded by A.I.R.C. (Associazione Italiana per la Ricerca sul Cancro) through a MFAG #17658. FM is supported by A.I.L. (Associazione Italiana Contro le Leucemie-Linfomi e Mieloma ONLUS). AN is funded by A.I.R.C. Investigator grants #16722 and #10136. PC is funded by A.I.R.C. Investigator grant #14346.

Authorship Contributions:

F.M., L.A. and P.C.: designed the study, collected and analysed the data, and wrote the paper; D.L., W.W., P.J.C., and S.G.B. analysed the data.; J.C., T.H., A.D., C.C., A.P., G.P., A.B., A.C., A.D.R., P.L.Z., F.Z., U.V., R.P., J.I., T.P. and J.I. collected the data; N.B., A.N., and J.I. critically revised the paper

Disclosure of Conflicts of Interest:

No conflict of interests to declare.

References

1. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization (WHO) classification of lymphoid neoplasms. *Blood.* 2016.

2. Maura F, Dodero A, Carniti C, Bolli N. Biology of peripheral T cell lymphomas – Not otherwise specified: Is something finally happening? *Pathogenesis.* 2016;3(1):9-18.

3. Iqbal J, Wright G, Wang C, et al. Gene expression signatures delineate biological and prognostic subgroups in peripheral T-cell lymphoma. *Blood.* 2014;123(19):2915-2923.

4. Piccaluga PP, Fuligni F, De Leo A, et al. Molecular profiling improves classification and prognostication of nodal peripheral T-cell lymphomas: results of a phase III diagnostic accuracy study. *J Clin Oncol.* 2013;31(24):3019-3025.

5. Piva R, Agnelli L, Pellegrino E, et al. Gene expression profiling uncovers molecular classifiers for the recognition of anaplastic large-cell lymphoma within peripheral T-cell neoplasms. *J Clin Oncol.* 2010;28(9):1583-1590.

6. Wang C, McKeithan TW, Gong Q, et al. IDH2R172 mutations define a unique subgroup of patients with angioimmunoblastic T-cell lymphoma. *Blood.* 2015;126(15):1741-1752.

7. Wang T, Feldman AL, Wada DA, et al. GATA-3 expression identifies a high-risk subset of PTCL, NOS with distinct molecular and clinical features. *Blood.* 2014;123(19):3007-3015.

8. Dobay MP, Lemonnier F, Missiaglia E, et al. Integrative clinicopathological and molecular analyses of angioimmunoblastic T-cell lymphoma and other nodal lymphomas of follicular helper T-cell origin. *Haematologica.* 2017;102(4):e148-e151.

9. Palomero T, Couronne L, Khiabanian H, et al. Recurrent mutations in epigenetic regulators, RHOA and FYN kinase in peripheral T cell lymphomas. *Nat Genet.* 2014;46(2):166-170.

10. Sakata-Yanagimoto M, Enami T, Yoshida K, et al. Somatic RHOA mutation in angioimmunoblastic T cell lymphoma. *Nat Genet.* 2014;46(2):171-175.

11. Schatz JH, Horwitz SM, Teruya-Feldstein J, et al. Targeted mutational profiling of peripheral T-cell lymphoma not otherwise specified highlights new mechanisms in a heterogeneous pathogenesis. *Leukemia.* 2015;29(1):237-241.

12. Yoo HY, Sung MK, Lee SH, et al. A recurrent inactivating mutation in RHOA GTPase in angioimmunoblastic T cell lymphoma. *Nat Genet.* 2014;46(4):371-375.

13. Gerstung M, Pellagatti A, Malcovati L, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun.* 2015;6:5901.

14. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics.* 2010;26(12):1572-1573.

15. Gentles AJ, Newman AM, Liu CL, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med.* 2015;21(8):938-945.

16. Chaussabel D, Quinn C, Shen J, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity.* 2008;29(1):150-164.

17. Li S, Rouphael N, Duraisingham S, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol.* 2014;15(2):195-204.

18. Weiner J, 3rd, Parida SK, Maertzdorf J, et al. Biomarkers of inflammation, immunosuppression and stress with active disease are revealed by metabolomic profiling of tuberculosis patients. *PLoS One.* 2012;7(7):e40221.

19. Fonseca NA, Petryszak R, Marioni J, Brazma A. iRAP - an integrated RNA-seq Analysis Pipeline. *bioRxiv.* 2014.

20. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res.* 2015;4:1070.

21. Iqbal J, Weisenburger DD, Greiner TC, et al. Molecular signatures to improve diagnosis in peripheral T-cell lymphoma and prognostication in angioimmunoblastic T-cell lymphoma. *Blood.* 2010;115(5):1026-1036.

22. Agnelli L, Mereu E, Pellegrino E, et al. Identification of a 3-gene model as a powerful diagnostic tool for the recognition of ALK-negative anaplastic large-cell lymphoma. *Blood.* 2012;120(6):1274-1281.

23. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015;12(5):453-457.

24. Kesler MV, Paranjape GS, Asplund SL, McKenna RW, Jamal S, Kroft SH. Anaplastic large cell lymphoma: a flow cytometric analysis of 29 cases. *Am J Clin Pathol.* 2007;128(2):314-322.

25. Iqbal J, Weisenburger DD, Chowdhury A, et al. Natural killer cell lymphoma shares strikingly similar molecular features with a group of non-hepatosplenic gammadelta T-cell lymphoma and is highly sensitive to a novel aurora kinase A inhibitor in vitro. *Leukemia.* 2011;25(2):348-358.

26. Piccaluga PP, Agostinelli C, Califano A, et al. Gene expression analysis of angioimmunoblastic lymphoma indicates derivation from T follicular helper

cells and vascular endothelial growth factor deregulation. *Cancer Res.* 2007;67(22):10703-10710.

27. Piccaluga PP, Agostinelli C, Califano A, et al. Gene expression analysis of peripheral T cell lymphoma, unspecified, reveals distinct profiles and new potential therapeutic targets. *J Clin Invest.* 2007;117(3):823-834.

28. Scarfo I, Pellegrino E, Mereu E, et al. Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts. *Blood.* 2016;127(2):221-232.

29. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin Chem.* 2015;61(1):64-71.

30. Reddy A, Zhang J, Davis NS, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell.* 2017;171(2):481-494 e415.

31. Scott DW, Mottok A, Ennishi D, et al. Prognostic Significance of Diffuse Large B-Cell Lymphoma Cell of Origin Determined by Digital Gene Expression in Formalin-Fixed Paraffin-Embedded Tissue Biopsies. *J Clin Oncol.* 2015;33(26):2848-2856.

32. Veldman-Jones MH, Lai Z, Wappett M, et al. Reproducible, Quantitative, and Flexible Molecular Subtyping of Clinical DLBCL Samples Using the NanoString nCounter System. *Clin Cancer Res.* 2015;21(10):2367-2378.

33. Genovese G, Kahler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med.* 2014;371(26):2477-2487.

34. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med.* 2014;371(26):2488-2498.

35. Tiacci E, Venanzi A, Ascani S, et al. High-Risk Clonal Hematopoiesis as the Origin of AITL and NPM1-Mutated AML. *N Engl J Med.* 2018;379(10):981-984.

36. Inghirami G, Chan WC, Pileri S, malignancies AxcG-dtmol. Peripheral T-cell and NK cell lymphoproliferative disorders: cell of origin, clinical and pathological implications. *Immunol Rev.* 2015;263(1):124-159.

Figure Legends.

Figure 1. (a) Molecular composition of the gene expression cohort (503 tumor cases); unsupervised hierarchical clustering (b) and Principal Component Analysis (c) on the entire series.

Figure 2. (a) Distribution of the variance of expression levels across genes explained by clinical, molecular and genetic alterations (f-test; FDR<1%; 221 samples). (b) Statistically significant mutation expression interaction terms (f-test; FDR<1%), for each alteration and clinical variable. The associated logarithmic expression fold change is indicated by colour. (c) Heatmap of the 19-genes model including all PTCL-NOS, AITL and *ALK*- ALCL cases (340 samples), stratified according to the cluster determined by the *ConsensusClusterPlus* R function.

Table 1. Clinical and molecular characterization of the 5 clusters extracted by *ConsensusClusterPlus*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Histology					
AITL	81 (93%)	33 (32%)	8 (25%)	5 (8%)	0
ALCL *ALK*-neg	3 (3.5%)	6 (6%)	7 (22%)	6 (9.5%)	47 (88%)
PTCL-NOS	3 (3.5%)	64 (62%)	17 (53%)	52 (82.5%)	8 (12%)
Gender					
Female	26 (37%)	31 (41%)	4 (33%)	20 (41%)	8 (26%)
Male	45 (63%)	43 (59%)	8 (67%)	28 (59%)	22 (74%)
Age (Median)	64 (33-87)	62 (19-88)	60 (35-97)	62 (21-87)	59 (31-80)
TET2					
MUT	26 (84%)	19 (65%)	0 (0%)	6 (32%)	0
WT	5 (16%)	10 (35%)	1 (100%)	13 (68%)	0
DNMT3A					
MUT	13 (42%)	9 (31%)	0	3 (16%)	0
WT	18 (58%)	20 (69%)	1 (100%)	16 (84%)	0
RHOA					
MUT	23 (79%)	15 (51%)	0	2 (10.5%)	0
WT	6 (21%)	14 (49%)	1 (100%)	17 (89.5%)	0
IDH2					
MUT	16 (45%)	1 (4%)	0	1 (33%)	0
WT	20 (55%)	22 (96%)	0	2 (67%)	0