Lung squamous cell carcinomas with basaloid histology represent a specific molecular entity.

Authors
*Christian Brambilla (1), *Julien Laffaire (2), Sylvie Lantuejoul (3), Denis Moro-Sibilot (1), Hélène Mignotte (1), François Arbib (1), Anne-Claire Toffart(1), Fabien Petel (2), Pierre Hainaut (4), Sophie Rousseaux (5), Saadi Khochbin (5), **Aurélien de Reyniès (2), **Elisabeth Brambilla (3)

*: co-first authors.
**: co-senior authors, jointly directed this work with their respective expertise.

Institutions
(1) Clinique de Pneumologie - Centre Hospitalier Universitaire de Grenoble; Institut Albert Bonniot, INserm U823 ; Université Joseph Fourrier, Grenoble, France
(2) Cartes d'Identité des Tumeurs (CIT) Program - Ligue Nationale Contre le Cancer, Paris, France
(3) Département d'Anatomie et Cytologie Pathologique – Centre Hospitalier Universitaire de Grenoble ; Institut Albert Bonniot, INserm U823 ; Université Joseph Fourrier, Grenoble, France
(4) International Prevention Research Institute, Lyon, France.
(5) INserm, U823; Université Joseph Fourier - Grenoble 1; Institut Albert Bonniot, Grenoble, France

Acknowledgements
This work is part of the program Cartes d'Identité des Tumeurs (CIT, http://cit.lique-cancer.net) funded and developed by the Ligue Nationale Contre le Cancer. This study was supported by Curelung European Project (2010 -2013), PNES cancers du poumon from Institut National Du Cancer (2010), Conseil Scientifique Agiradom, projet libre from Association pour la Recherche sur le Cancer (2010, Saadi Khochbin) and PHRC Biomarkscan (2003).

Corresponding authors
Professor Elisabeth Brambilla
Centre Hospitalier Universitaire de Grenoble
Département d'Anatomie et Cytologie Pathologiques
CS 10217 - 38043 Grenoble Cedex 09 - France
Tel. +33 (0)4 76 76 58 75 - Fax +33 (0)4 76 76 59 29
Email: EBrambilla@chu-genoble.fr

Aurélien de Reyniès
Programme Cartes d'Identité des Tumeurs
Ligue Nationale Contre le Cancer
14, rue Corvisart – 75013 Paris – France
Tel. +33 (0)1 53 55 25 12 – Fax +33 (0)1 53 55 25 34
Email: reyniesa@ligue-cancer.net

Running title
Lung basaloid carcinoma: a histo-molecular entity

Public presentation of the study
none

Disclaimers
none
Translational Relevance

The basaloid carcinoma (pure) and the (mixed) basaloid variant of lung squamous cell carcinoma (SCC) show a dismal prognosis compared to non-basaloid SCC. Their molecular characteristics are largely unknown. No validated molecular marker has yet been identified for the diagnosis of this aggressive entity. Here, we show for the first time that pure basaloid carcinoma constitute a distinct histo-molecular entity, with a distinct signature related to specific pathways that are highly coherent with the poorly differentiated status and aggressiveness of these tumors. These molecular characteristics highlight their intrinsic resistance to cytotoxic chemotherapy and could serve as a guide for targeted therapies. Finally, we propose an immunohistochemistry-based molecular predictor, based on two markers (SOX4, IVL), which reliably discriminates pure and mixed basaloid tumors from non-basaloid tumors.
Abstract

PURPOSE
The basaloid carcinoma (pure) and the (mixed) basaloid variant of lung squamous cell carcinoma (SCC) have a dismal prognosis but their underlying specific molecular characteristics remain obscure and no therapy has proven to be efficient.

MATERIALS AND METHODS
In order to assess their molecular specificity among other lung squamous cell carcinomas we analyzed DNA copy number aberrations and mRNA expression pangenomic profiles of 93 SCC, including 42 basaloid samples (24 pure, 18 mixed).

RESULTS
Supervised analyses reveal that pure basaloid tumors display a specific mRNA expression profile, encoding factors controlling the cell cycle, transcription, chromatin and splicing, with prevalent expression in germline and stem cells, while genes related to squamous differentiation are underexpressed. From this signature, we derived a 2-genes (SOX4, IVL) immunohistochemistry-based predictor which discriminated basaloid tumors (pure and mixed) from non-basaloid tumors with 94% accuracy in an independent series. The pure basaloid tumors are also distinguished through unsupervised analyses. Using a centroid-based predictor, the corresponding molecular subtype was found in 8 independent public datasets (n=58/533), and was shown associated to a very poor survival as compared to other SCC (adjusted HR =2.45, p =0.000001).

CONCLUSION
This study enlightens the heterogeneity of SCC that can be subclassified in mRNA expression subtypes. This study demonstrates, for the first time, that basaloid SCC constitute a distinct histo-molecular entity, which justifies its recognition and distinction from non-basaloid SCC. Additionally, their characteristic molecular profile highlights their intrinsic resistance to cytotoxic chemotherapy and could serve as a guide for targeted therapies.
Introduction

Recent advances in genetic profiling revealed characteristic molecular alterations in adenocarcinomas [1-5] and squamous cell carcinomas [6-7]. Small cell Lung Carcinomas [8] have also been revisited at molecular levels: genomic mutations, DNA copy number aberrations and mRNA expression, uncovering genetic alterations with potentially driver functions. However the genetic landscape of rare tumors entities remains largely obscure. Basaloid carcinoma account for about 5% of non-small cell lung carcinomas (NSCLC) and were identified as a histopathological entity with a dismal prognosis [9-10]. Among NSCLC and lung squamous cell carcinomas, the basaloid carcinoma (pure type) and the squamous variant of basaloid carcinoma (mixed basaloid SCC) (WHO 2004) show the poorest prognosis [11]. In the absence of specific molecular characteristics, basaloid carcinomas, in their pure form, were presented as a subtype of large cell carcinoma (WHO 2004), while they very likely present distinct biological characteristics, whose definition would be helpful in diagnosis and in the design of targeted treatment.

The cellular heterogeneity of mixed basaloid SCC hampers the identification of a specific molecular signature. Here, in order to assess whether basaloid SCC represent a distinct molecular entity as compared to non-basaloid SCC, we performed the transcriptomic and genomic analysis of a large series of SCC including pure and mixed basaloid tumors, as well as non-tumor tissues (lung alveoli, bronchi) and tumor controls (non-basaloid SCC). This approach allowed for the first time to uncover a very specific molecular profile of basaloid SCC revealing their biology and indicating possible therapeutic approaches.
Material and Methods

Patients and tumor samples

Ninety-three samples of squamous cell carcinoma (SCC) of the lung, referred as the CIT (Cartes d’Identité des Tumeurs) cohort, obtained from 93 distinct patients, and collected from 1988 to 2006, were included in the present study. All tumor specimens were collected, stored and used with the patients’ informed consent. All samples were independently reviewed according to the WHO criteria by two anatomopathologists: 100% concordance was obtained for the cases classified as basaloid carcinoma. For mixed basaloïd carcinomas (basaloid component ≥ 50%), the two pathologists independently assessed the percentage of the basaloid component: their evaluation was concordant for 80% of the cases. A consensus was reached after a common review of the remaining cases. In order to confirm the histopathological diagnosis, all basaloid carcinoma cases were immunostained with antibodies against P40, P63, CK34betaE12 (recognizing the cytokeratins 1, 5, 10, 14), TTF-1, Chromogranin A, Synaptophysin and CD56 (antibodies are described in Sup_Table_1A). A tumor cell rate over 70% was used as criterion to select samples. This cohort was deliberately enriched in basaloid samples (n=42, 24 pure basaloid and 18 mixed), the remaining samples (n=51) corresponding to non-basaloid SCC, including 36 well differentiated SCC and 15 poorly differentiated SCC. Endoalveolar features were observed in 28 non-basaloid SCC. Ninety-five percent of patients were males and the median age at diagnostic was 64 years (range: 44-82). All patients but 2 were current smokers (n=84) or former smokers (n=6). Forty non-basaloid SCC (78%) and ten basaloid SCC (24%) were classified as T1N0M0. The main clinical and phenotypic characteristics of the patients are summarized in Sup_Table_1B.
Microarrays and statistical analyses

All 93 tumor samples, and 16 samples collected from distant normal lung tissue (10 lung alveolar tissue, 6 bronchi), were studied on Affymetrix HG_U133_Plus_2.0 gene expression arrays. A subset of 64 tumor samples (27 basaloid, 37 non-basaloid) was also analyzed on Illumina HumanCNV370-Quad SNP arrays. Microarray data were deposited to ArrayExpress under accession number E-MTAB-2435.

Public gene expression datasets

Height public expression profiling datasets of lung SCC [1,6,12-17] were collected from GEO (http://www.ncbi.nlm.nih.gov/geo/) or supplemental data of the related article. Affymetrix series were normalized with RMA method; for other series the furnished normalized data were directly used. Probes (non Affymetrix chips) /probe sets (Affymetrix chips) were then averaged per HUGO Gene Symbol.

TP53 sequencing

Mutations were analysed as previously described [18] using genomic DNA isolated from paraffin-embedded archived tissue sections.

Immunohistochemistry-based predictor of basaloid histology

Four steps were followed to build and validate the predictor: (i) selection of discriminative markers using genome-wide expression data in a training set S1 (CIT series, n=93 SCC); (ii) measure by immunohistochemistry (IHC) of the selected markers (IVL, SOX4) in a subpart S1’ of the training set S1 (n=66 SCC), including 26 basaloid SCC (pure or mixed) and 40 non-basaloid SCC (with no or less than 50% basaloid component) and in an independent validation series S2 (n=15 basaloid SCC + 10 non-basaloid SCC + 10 lung adenocarcinomas); (iii) building of the predictor
using IHC measures of the selected markers in the training set S1'; (iv) validation of the predictor in the validation set S2. Of note, all cases included in the IHC series (S1'+ S2) were consensus cases (2 pathologists) and no borderline case was used, since our objective was to recognize basaloid carcinoma on the basis of the standard of morphological classification. The tested antibodies are described in Sup_Table_1A.

Detailed materials and methods are given in the supplemental information.

**Results**

**Pure and mixed basaloid SCC show the same dismal prognosis**

The ninety-three samples of primary squamous cell carcinoma (SCC) from the CIT cohort were classified according to the WHO criteria into 4 histological classes, illustrated on Figure_1_A-F: pure basaloid carcinoma (BAS_p, n=24), mixed basaloid SCC (BAS_m, n=18), non-basaloid well differentiated SCC (SCC wd, n=36) and non-basaloid poorly differentiated SCC (SCC pd, n=15). All basaloid cases expressed P40, P63 and Cytokeratins 1, 5, 10, 14 but did not express TTF-1 and neuroendocrine markers [19].

Pure and mixed basaloid tumors showed the same poor prognosis in our dataset (Figure_1_G), indicating that the presence of a basaloid cellular contingent is of most interest for prognosis. Misclassified mixed basaloid samples could explain the lack of prognostic significance of the basaloid feature in previous studies [20].

**Pure basaloid SCC show a specific transcriptome profile, related to the deregulation of specific pathways**
Expression profiles of 42 basaloid (24 BAS_p, 18 BAS_m), and 51 non-basaloid SCC (36 SCC_wd, 15 SCC_pd) were obtained. A high proportion of genes were found differentially expressed between BAS_p and non-basaloid SCC (H1 proportion=47 %, Sup_Table_2_A), indicating a specific mRNA profile for BAS_p. Significantly up-regulated genes in BAS_p (limma q-value < 0.05) included genes related to TP53 mutation signature (TMSB15A, PROM1), transcription factors (SOX4, SOX9, SOX11, MYB, E2F3, E2F5), embryonic development (FGF3, FGF19), methylation regulation (TET1, DNMT1, DNMT3A), cell cycle (MKI67, BUB1, DTL), splicing (SFRS1,2,3) and survival (BCL2), while the most down-regulated genes were related to epithelial cells and keratinocytes differentiation (KRT6, IVL, SPRR genes and SFN). Additionally, the recently defined ectopic male germ cell /placenta specific gene expression signature, as indicative of very aggressive lung tumours [21], also highly correlated with pure basaloid in our cohort (chi2 p < 0.0005, Sup_Information). An extensive pathways analysis also confirmed these findings (Sup_Table_3). The comparison between pure basaloid and non-basaloid SCC – either well differentiated or poorly differentiated (Figure_2), confirmed the up-regulation of pathways related to cell cycle, transcription factors, mRNA splicing and chromatin modifications, and the down-regulation of the squamous differentiation pathway. It also revealed the upregulation of signatures associated to testis and embryonic stem (ES) cells and poorly differentiated tumor markers (NANOG, OCT4, SOX2 and c-MYC targets) and the down-regulation of signatures related to the Polycomb gene silencing system (SUZ12, EED, H3K27ME3 and PRC2 gene sets) (Sup_Figure_1). Similarly mixed basaloid were compared to non-basaloid SCC (H1 proportion=20 %) and as in pure basaloid tumors, pathways related to cell cycle, spliceosome and ES cells were upregulated (Sup_Figure_1).
A 2-genes immunohistochemistry-based predictor discriminates pure and mixed basaloid tumors from non-basaloid tumors

To identify molecular markers of basaloid tumors, we selected overexpressed transcripts showing high AUC (> 75%) and specificity (> 95%), while optimizing sensitivity (cutoff yielding a Fisher test p < 1e-6, Sup_Table_2_A). This selection yielded 5 genes (SOX4, CBX5, PATZ1, RBMX, SFRS3), all showing a higher sensitivity in pure basaloid than in mixed basaloid tumors. The transcription factor SOX4 showed 100% specificity and 50% sensitivity to discriminate basaloid tumors in our cohort; it was thus selected for further analyses.

To identify negative markers of basaloid tumors, we selected underexpressed transcripts showing high AUC (>75%) and sensitivity (>90%), while optimizing specificity (cutoff yielding a Fisher test p < 1e-5, Sup_Table_2_A); this selection yielded 5 genes (IVL, KRT16, TOM1, C1ORF224, KCNK6). The squamous differentiation marker IVL (involucrin), showing a high fold change (>2) between non-basaloid and basaloid SCC, was selected for further analyses.

We measured by immunohistochemistry the protein expression of SOX4 and IVL in a training set of 66 tumors common to the CIT transcriptome series (26 basaloid + 40 non-basaloid SCC). Using this training set we built a predictor based on the quick scores (QS) related to these two genes, using the following formula: 

\[
\text{if } QS(SOX4) \geq 110 \text{ then Basaloid; if } QS(SOX4) < 50 \text{ then Non-Basaloid; if } QS(SOX4) \text{ in } [50;110[ \text{ and } QS(SOX4) - QS(IVL) \geq - 55 \text{ then Basaloid else Non-Basaloid (Figure 3, Sup_Table_2_B).}
\]

In the training set, the predictor correctly classified all basaloid tumors (pure and mixed) and correctly classified 90% of the SCC samples. The predictor was then applied to an independent validation series of 35 tumors (15
basaloid + 10 non-basaloid SCC + 10 adenocarcinomas), where it correctly classified all samples except 2 mixed basaloid tumors (classified as non-basaloid). The accuracy of this predictor was found to be 94% in the validation series (sensitivity:87%, specificity:100%, positive predictive value:100%), in line with the performances observed in the training series (accuracy:94%, sensitivity:100%, specificity:90%, positive predictive value:87%).

**SCC tumors mostly share a similar genomic aberration profile**

DNA from 27 basaloid (14 Bas_p, 13 Bas_m) and 37 non-basaloid SCC (25 SCC_wd, 12 SCC_pd) was hybridized on Illumina HumanCNV370 SNP arrays. The most frequent copy number aberrations (CNA) on the whole dataset were identified with GISTIC2.0. This analysis identified previously reported CNA [7] such as gains of 3q, 5p, 8q and losses of 1p, 3p, 4p, 4q, 5q, 8p, 9p. Similarly it pointed out known target genes, including gains of SOX2, MYC, CCND1, MDM1, FGFR1 and losses of CDKN2A, PCDH10, RB1, PTEN (Sup_Figure_2). Very few CNA were found in significantly different proportions among histological classes (Sup_Table_4). At the single gene level, gains of MYB, JUN, FGFR1, PIK3C3, DSC/DSG genes were found more frequent in pure basaloid tumors. However none of these differences reached significance after correction for multiple testing. From these data we were not able to identify CNA being specific of basaloid SCC, either pure or mixed.

**Consensus unsupervised clustering identifies a poor prognosis molecular subtype corresponding to pure basaloid SCC**

To assess whether basaloid tumors could correspond to a molecular subtype obtained without a priori knowledge we performed unsupervised analyses of the mRNA expression profiles. We identified consensus partitions in k= 2 to 8 clusters of
the 93 SCC expression profiles from the CIT cohort. The consensus partition in k=4 clusters was selected (Figure_4_A) based on the gap statistic method [22] (Sup_Figure_3_A). The underlying co-classification matrix showed a great level of agreement between the raw partitions in k=4 clusters obtained using the different experimental settings (Sup_Figure_3_B). Principal component analysis was consistent with these findings (Figure_4_B). This partition was significantly associated to histology (Fisher p= 1e-8). The 4 clusters were named Basaloid-Like (BL) (n=21), Peripheral EndoAlveolar (PEA)(n=29), Classical_1 (n=21) and Classical_2 (n=22).

The Basaloid-Like cluster contained almost only basaloid tumors (90%), and mostly contained pure basaloid tumors (72%), contrary to other clusters (<15%). It also showed enrichment in tumors with stage 2 or higher (62%) (Figure_4_A). Its expression pattern was the most singular compared to other clusters (Sup_Figure_3_C). Tumors from the PEA cluster mostly showed an alveolar contingent in their microenvironment (83%), most were stage 1 (66%) without basaloid features (79%). The PEA cluster was found enriched in non-basaloid poorly differentiated SCC (Fisher p=3e-4), tumors with a high (>75%) diploid tumor cells rate (Fisher p=1e-5), and TP53 wild type tumors (Fisher p=0.02). Non-basaloid well differentiated SCC were mostly found in the Classical_1 and Classical_2 clusters (67% and 45%) and absent in the Basaloid-Like cluster (0%). Tumors with exon 4 TP53 mutations were found enriched in the Classical_1 group (Fisher p=0.005). Mixed basaloid tumors were not associated to any particular cluster.

To assign independent SCC datasets to these four molecular subtypes, we built a 139-genes nearest-centroïd predictor (Sup_Table_5), and classified all SCC samples (n=533) of 8 public expression profiling datasets (Sup_Table_6_A).
Samples from each of the four subtypes were found in all datasets (Sup_Table_6_A/B). Overall 58 samples (11%) were classified in the Basaloid-Like subtype, 152 (28%) in the PEA subtype, 215 (41%) in the Classical_1 subtype and 108 (20%) in the Classical_2 subtype. A significantly poorer prognosis was observed in the Basaloid-Like subtype compared to other molecular subtypes in the validation datasets (Logrank p <6e-8, Figure_4_C) and in the discovery cohort (Sup_Figure_4). Among the 3 other subtypes, Classical_2 and PEA subtypes presented similar outcome and Classical_1 showed a significant better prognosis. These results were conserved in stage 1 tumors (Figure_4_D).

Pathways found consistently deregulated in the Basaloid-Like cluster across datasets (Figure_5, Sup_Table_7) were highly coherent with those identified in pure basaloid tumors compared to non-basaloid SCC (Figure_2, Sup_Table_3). As expected, normal bronchial basal cell signatures supposed to be the stem cell of bronchi [23] were found highly and significantly overexpressed in the Basaloid-Like subtype, whose expression profile was found clearly distinct from that of normal lung (alveoli) thus confirming the proposed derivation of basaloid carcinoma from a basal stem cell progenitor at the time of first description [9].

Finally we applied a SCC classification system, recently published by Wilkerson et al. [6], to all SCC profiles from the discovery and validation cohorts (n=626). Overall, the association between the CIT and Wilkerson classification systems was found to be very high (all series, Fisher test p ≈ 0). Wilkerson subtypes were also found very associated to histology (CIT series, Fisher test p =2e-9). To match subtypes of both systems we then used pairwise Cohen’s Kappa statistics. A fair and good agreement was respectively observed between Basaloid-Like and Primitive subtypes (κ=0.56) and PEA and Secretory subtypes (κ=0.72).
subtypes, related to non-basaloid well differentiated SCC, showed a poor agreement (κ≤0.40). Despite the fair agreement between the Basaloid-Like (n=78/626, 12%) and the Primitive (n=104/626, 17%) subtypes, the Basaloid-Like subtype was far more associated to overall survival on the validation cohort (dataset adjusted HR=2.45, 95%CI=[1.72; 3.5], Wald p < 1e-6; C-index= 0.67, 95%CI = [0.52 ; 0.79]) than the Primitive subtype (dataset adjusted HR=1.38; 95%CI=[1.00; 1.89]; Wald p =0.04; C-index= 0.58, 95%CI = [0.45 ; 0.70]). Concerning subtypes related to non-basaloid well differentiated SCC, contrary to that of Wilkerson, our classification reveals a subtype (Classical_1, n=240/626 (38%)) showing a higher overall survival rate than other subtypes in the validation cohort (dataset adjusted HR=0.61, 95%CI=[0.47; 0.79], Wald p = 2e-4; C-index=0.40, 95%CI = [0.30 ; 0.51]; Sup_Figure_5), even after excluding the Basaloid-Like subtype (dataset adjusted HR=0.69; 95% CI=[0.53; 0.95]; Wald p = 7e-3).

In conclusion, pure basaloid tumors correspond to a specific molecular subtype, either in our classification system or in that of Wilkerson. This finding further demonstrates that histologically pure basaloid tumors constitute a specific molecular entity.

**Discussion**

The worse prognosis of basaloid SCC as compared to non-basaloid SCC suggests the existence of underlying biological differences. Incidentally, we show here that
mixed basaloid SCC, representing 45% of the basaloid tumors in our series, share
the same poor prognosis as pure basaloid SCC, meaning that the presence of
basaloid contingents in SCC is informative for prognosis. To identify the molecular
characteristics of basaloid SCC we analyzed mRNA expression and DNA copy
number aberrations profiles of a large cohort of this rare entity (n=42) with tumor
controls (51 non-basaloid SCC). Mixed basaloid SCC are expected to be
heterogeneous at the molecular level, contrary to pure basaloid SCC. We thus mainly
concentrated on pure basaloid tumors to identify molecular specificities of this
histological group. Supervised analyses revealed that pure basaloid SCC display a
specific mRNA expression profile as compared to non-basaloid SCC. Enrichment in
TP53 mutation signature, upregulation of transcription, epigenetic, cell cycle, splicing
and survival factors, as well as that of male germ cells, embryonic development,
stemness/poor differentiation genes characterized pure basaloid SCC, which also
downregulated keratinocyte differentiation genes. Mixed basaloid SCC also showed
upregulation of proliferation and embryonic development related genes as compared
to non-basaloid SCC, but to a lesser extent than pure basaloid SCC, likely due to
their heterogeneous cellular content. These molecular observations are in line with
the poorly differentiated status and aggressiveness of the basaloid SCC.
DNA copy number aberrations (CNA) were found non-informative to distinguish
between SCC histological subgroups. All SCC, either basaloid or not, showed very
similar CNA profiles, pointing at regions previously reported [7], such as gain or
amplicon of the 3q region around SOX2, found in almost all SCC samples. However
CNA are potentially highly informative concerning targeted therapies [24]. In
particular, FGFR1 (8p12) and MYB (6q22-q23), found here in peak regions of gain,
are targetable: MYB could be targeted by several drugs under development [25-26].
and FGFR1 amplified Non-Small Cell Lung Cancer have been shown to be sensitive to FGFR1 inhibitors [27-28].

By unsupervised analyses of SCC mRNA expression profiles we identified four molecular subtypes: Basaloid-Like, PeriEndoAlveolar (PEA), Classical_1 and Classical_2. The Basaloid-Like cluster was shown highly associated to (pure) basaloid histology. Similarly, among Wilkerson molecular subtypes (Primitive, Secretory, Basal, Classical), the Primitive cluster was found associated to basaloid histology, a key characteristic that was not reported by Wilkerson et al. Altogether, analysis of the molecular subtypes obtained either from our series or the literature unambiguously shows that pure basaloid SCC do correspond to a specific molecular subtype.

Our classification system, applied to a large validation cohort via a centroid-based predictor, outperforms that of Wilkerson in predicting overall survival. The CIT Basaloid-Like subtype, while agreeing with Wilkerson Primitive subtype, is far more associated to a poor prognosis. A substantial proportion of the samples classified as Primitive were predicted as non Basaloid-like (44%). Expression-based analyses support that these samples could be more differentiated than Basaloid-like Primitive samples (data not shown). Moreover, contrary to Wilkerson, among well differentiated SCC we identify a subtype (Classical_1) showing a significantly better prognosis. These clinical observations suggest that our subtypes are more homogeneous at the molecular level than that of Wilkerson. Indeed, assuming that tumor molecular characteristics may greatly determine prognosis, then defining more homogeneous molecular subgroups may yield greater prognosis differences between subgroups. In particular, it supports that the Basaloid-Like subtype is more specifically related to...
basaloid SCC than the Primitive subtype of Wilkerson, even if not significantly different in the discovery cohort (NB: unavailable data in validation series).

Interestingly, most of the poorly differentiated non-basaloid SCC of our series (11/15) were found in the PEA subtype, shown to be very similar to Wilkerson Secretory subtype. Contrary to Wilkerson, suggesting secretory properties for this tumor subtype, our study reveals that these tumors show a periendoalveoloar microenvironment, which much more likely explains the “secreting” signatures of this group. Moreover comparison of PEA profiles with that of bronchi and lung controls unambiguously shows that the PEA subtype – contrary to all other subtypes- has much more in common with lung profiles (which include alveoli) than with bronchi profiles. From our data it seems excluded that PEA could be driven by non-tumor cell related artifacts, as a tumor cells rate higher than 70% criterion was strictly applied within our series, and given that the percentage of tumor cells was not found different across subtypes. Interestingly this subtype was characterized by a smaller fraction of tumors cells with genome instability and a smaller TP53 mutation rate. Accordingly, pathway analysis revealed upregulation of TP53 targets signatures and downregulation of cell cycle in this subtype (supplemental data). Upregulation of immune signatures was also found very significantly associated to the PEA subtype.

Our study also completely renews the molecular subtyping of non-basaloid well differentiated SCC. In the discovery cohort, the Classical_1 subtype is specifically mutated in the exon 4 of TP53 (NB: unavailable information in validation series). Pathway analysis across 5 series (supplemental data) revealed that TP53 target genes are more expressed in Classical_1 than in Classical_2 and Basaloid-Like subtypes, while having a similar TP53 mutation rate. These observations suggest that mutations of the 4th exon of TP53 could be less deleterious than mutations of
other exons observed within our series. Biological consequences of a more functional p53 protein are also supported by pathway analysis as we found a relative higher expression of pro-apoptotic genes and lower expression of cell cycle genes in Classical_1 tumors as compared to Basaloid-Like and Classical_2 subgroups. Finally, a clinical support of these biological hypotheses is given by the better prognostic observed in the Classical_1 subtype both in our cohort and in the validation series. Mixed basaloid SCC were not found associated to any particular molecular subtype, as expected due to their cellular heterogeneity. Accordingly, mRNA expression profiles yielded very specific but moderately sensitive markers for the identification of basaloid SCC as a whole, due to a lack of sensitivity concerning mixed basaloid SCC. Using immunohistochemistry-based markers we could overcome this limitation: we derived a 2-genes predictor based on a positive (SOX4) and a negative (IVL) marker of the basaloid SCC, which showed a great accuracy (94%) in an independent validation set (n=35). This simple predictor, used in addition to histologic examination, should considerably help the identification of basaloid tumors in clinical routine.

Of note, none of the present data on basaloid carcinoma (pure and mixed) is applicable to large cell carcinoma (NOS) of the WHO 2004 classification nor to the revised one (to be published in 2014), since a thorough revision of the concept of large cell carcinoma based on immunohistochemical differentiation markers and on the last genomic–based proposed classification of lung cancer restricted this class to a few cases with no clear differentiation phenotype, which is not the case for basaloid carcinoma.

In conclusion, our results establish that pure basaloid SCC correspond to a specific molecular entity, fully justifying its histologic recognition and distinction from non-
basaloid SCC. Related deregulated pathways enlighten its intrinsic resistance to cytotoxic chemotherapy and should serve as a guide to targeted therapies.

References


Legends

FIGURE 1: Pathology review

(A) Basaloid carcinoma: tumor cells are small, monomorphic, cuboidal or fusiform, with a scant cytoplasm and peripheral palisading. (B) Mixed basaloid carcinoma (SYN. Basaloid variant of squamous cell carcinoma): Basaloid component on the right field, squamous cell component on the left field. (C) Well differentiated Squamous cell Carcinoma: squamous differentiation and keratinisation are frequent and obvious. (D) Poorly differentiated squamous cell carcinoma: most of the tumor shows absence of definite differentiation with few intercellular bridges and keratinization. (E-F) Peripheral endoalveolar squamous cell carcinoma: the SCC shows predominant intra-alveolar topography entrapping numerous strands of Type 1 alveolar residual pneumocytes (E). (G) Kaplan Meier curves of Overall Survival in the CIT cohort, stratified by subhistology. P-values refer to logrank tests comparing overall survival between subhistologies.

FIGURE 2: Deregulated pathways in pure basaloid tumors

Illustration of 13 signatures / pathways found among the most deregulated ones between pure basaloid tumors (red) and non-basaloid tumors (poorly differentiated SCC (light gray); well differentiated SCC (dark gray)) in our dataset. Samples (n=75) are independently ordered for each signature, on the mean expression of the corresponding signature. References of the signatures and pathways are indicated in Sup_Table_3_B. The Wilcoxon test p-values for up-regulated (left) and down-regulated (right) mean expression in basaloid vs non-basaloid SCC, are shown for each pathway.

FIGURE 3: Immunohistochemistry-based predictor of the basaloid entity.

(A) To discriminate basaloid from non-basaloid tumors, the following predictor is used, based on the quick score (QS) values of SOX4 and IVL: if \( QS(SOX4) \geq 110 \) then Basaloid; if \( QS(SOX4) < 50 \) then Non-Basaloid; if \( QS(SOX4) \in [50;110[ \) and \( QS(SOX4) - QS(IVL) \geq -55 \) then Basaloid else Non-Basaloid. (B) Pie charts showing the results obtained (red: predicted basaloid; gray: predicted non-basaloid) when applying this predictor to the tumors from the training series (upper part) and the validation series (lower part), in basaloid (left) and non-basaloid (right) tumors. (C) Illustration of the immunohistochemical measures of SOX4 (upper part) and IVL (lower part) in a basaloid (left part) and a non-basaloid SCC (right part) tumor.
FIGURE 4: Molecular subtypes derived from mRNA expression profiles and related overall survival.

(A) Upper panel - Consensus dendrogram of the 93 tumor samples derived from 24 unsupervised partitions in k=4 clusters obtained using different experimental settings (see supplemental methods). Lower panel – Tumor samples annotations: Basaloid (black= basaloid SCC, white= NOS SCC) and Histology (pure basaloid in red, mixed basaloid in orange, poorly differentially SCC in black and well differentiated SCC in white); Alveolar contingent refers to endoalveolar feature (yes=black, mild=grey, none=white); Stage (black=stage 1, white=higher stage); TP53 mutation and TP53 exon4 mutation (yes =black, no= white); Diploidy refers to the rate of diploid tumor cells, according to a SNParray-based estimation (black= >75%, white= <75%). In grey are shown samples with a very high rate of diploid cells (> 90%) that cannot be precisely estimated due to a SNParray profile without detectable copy number aberration. Wilkerson refers to the predicted subtype according to Wilkerson classification system [6] (green: secretory, red: primitive, light blue: basal, dark blue: classical). (B) Principal Components Analysis of the CIT cohort: samples are projected in the plane of the two first principal components (PC1, PC2) and are colored according to their molecular subtype (red = Basaloid-Like, light blue = Classical_1, dark blue = Classical_2, green = PeriEndoAlveolar). (C) Kaplan-Meier curves of overall survival in 7 public datasets of lung SCC (n=457) stratified according to CIT molecular subtypes. (D) Kaplan-Meier curves of overall survival in 7 public datasets, restricted to stage 1 tumors (n=211), stratified according to CIT molecular subtypes. Stars refer to p-values of logrank tests comparing overall survival between two groups among the following groups: Basaloid-Like, (Classical_2+PeriEndoAlveolar), Classical_1 (*<5E-2, **<5E-4, ***<5E-6). The logrank test p-value for all 4 subtypes is shown below the survival curves.

Figure 5: Heatmap of most deregulated pathways in the 4 molecular subtypes, in the CIT cohort and 4 validation datasets.

Heatmap representing the mean gene expression level of 31 signatures and pathways (rows) found differentially deregulated (see Methods) across molecular subtypes in our data set and 4 validation datasets (columns) by pathway analysis (Sup_Table_7). Sample groups are ordered according to their molecular subtypes and dataset membership (upper annotation): CIT=discovery dataset, WI=Wilkerson et al. dataset, LEE=Lee et al. dataset, RA=Raponi et al. dataset and RO=Roepman et al. dataset. NTL_B and NTL_L refer to non tumoral bronchi and lung samples from the CIT cohort. References of signatures and pathways are indicated in Sup_Table_7B. Color scale: blue/white/red gradient corresponding to min/intermediate/max value for each row including non tumoral samples.
A. Pure basaloid carcinoma

B. Mixed basaloid carcinoma

C. Well differentiated SCC

D. Poorly differentiated SCC

E. Peripheral endoalveolar SCC

F. Peripheral endoalveolar SCC

G. OS rate (%)

- mixed basaloid (n=18)
- poorly diff SCC (n=15)
- pure basaloid (n=24)
- well diff SCC (n=36)

logrank test

p-value = 0.0275

* p=0.003

FIGURE 1
## FIGURE 2

<table>
<thead>
<tr>
<th>Signatures</th>
<th>Wilcoxon test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Up</td>
</tr>
<tr>
<td>METASTASIS</td>
<td>8.5E-9</td>
</tr>
<tr>
<td>CHROMATIN</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>SPlicing</td>
<td>5.1E-6</td>
</tr>
<tr>
<td>TESTIS</td>
<td>4.2E-5</td>
</tr>
<tr>
<td>CELL CYCLE</td>
<td>5.5E-5</td>
</tr>
<tr>
<td>TRANSCRIPTION FACTORS</td>
<td>2.4E-6</td>
</tr>
<tr>
<td>METHYLATION</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>NANOG-OCT4-SOX2 TARGETS</td>
<td>2.2E-3</td>
</tr>
<tr>
<td>SENESCENCE</td>
<td>ns</td>
</tr>
<tr>
<td>KERATINOCTYE</td>
<td>ns</td>
</tr>
<tr>
<td>TP53 TARGETS</td>
<td>ns</td>
</tr>
<tr>
<td>LUNG</td>
<td>ns</td>
</tr>
<tr>
<td>DRUG METABOLISM</td>
<td>ns</td>
</tr>
</tbody>
</table>

- **pure basaloid carcinoma (n=24)**
- **well differentiated SCC (n=36)**
- **poorly differentiated SCC (n=15)**
FIGURE 3

A. Predictor

- Basaloid SCC
- Non-Basaloid SCC

QS(SOX4)

<50 [50;110] ≥110

QS(SOX4) - QS(IVL) ≥ -55

Non-Basaloid

Basaloid

B. Application of the predictor

Training series (n=66)

- Basaloid tumors: 26
- Non-Basaloid tumors: 4

Validation series (n=35)

- Basaloid tumors: 13
- Non-Basaloid tumors: 2

C. Basaloid SCC vs Non-Basaloid SCC

SOX4

IVL

Downloaded from clincancerres.aacrjournals.org on April 20, 2017. © 2014 American Association for Cancer Research.
FIGURE 4

A

Consensus dendrogram

B

Principal Component Analysis

C

Public datasets – all stages

D

Public datasets – stage 1

Fisher test

Basaloid-like (n=47) p=5.2e-6
Histology (n=194) p=1.1e-8
Alveolar contingent (n=90) p=0.18
Stage (n=126) p=0.18
TP53 mutation (n=19) p=0.18
TP53 exon4 mutation (n=86) p=0.033
Tobacco (n=41) p=0.57
Diploidy (n=65) p=1.9e-4
Wilkerson (n=126) p=1.9e-20

PeriEndoAlveolar Basaloid-like Classical_1 Classical_2

Author Manuscript
Published OnlineFirst on September 4, 2014; DOI: 10.1158/1078-0432.CCR-14-0459

TP53 mutation

Stage

TP53 exon4 mutation

Tobacco

Diploidy

Wilkerson

Fisher test

Basaloid-like (n=47) p=5.2e-6
Histology (n=194) p=1.1e-8
Alveolar contingent (n=90) p=0.18
Stage (n=126) p=0.18
TP53 mutation (n=19) p=0.18
TP53 exon4 mutation (n=86) p=0.033
Tobacco (n=41) p=0.57
Diploidy (n=65) p=1.9e-4
Wilkerson (n=126) p=1.9e-20

PeriEndoAlveolar Basaloid-like Classical_1 Classical_2

C

Public datasets – all stages

D

Public datasets – stage 1

logrank test

p-value = 8.9e-08

logrank test

p-value = 3.37e-05
<table>
<thead>
<tr>
<th></th>
<th>CIT</th>
<th>WI</th>
<th>LEE</th>
<th>RA</th>
<th>RO</th>
<th>CIT</th>
<th>WI</th>
<th>LEE</th>
<th>RA</th>
<th>RO</th>
<th>CIT</th>
<th>WI</th>
<th>LEE</th>
<th>RA</th>
<th>RO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Periendoalveolar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basaloid-Like</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 5**

NRAS SIGNALING
T CELL
RTK SIGNALING
B CELL
NK CELL
LUNG
MAPK
RAS SIGNAL
LIQUID SURFACE TENSION
RESPIRATORY GASEOUS EXCHANGE
TP53 TARGETS
SEIQUENCE
CELL ADHESION
APOPTOSIS
CELL MOTILITY
KERATINOCYTE
GLYCOLYSIS / GLUCONEOGENESIS
DRUG METABOLISM
PROTEIN KINASE CK2
HISTONE EXCHANGE
HPV+ TUMOR
WNT SIGNALING
NANOG-Oct3-4x2 TARGETS
TESTIS
SPlicing
METASTASIS
CELL CYCLE
 METHYLATION
CHROMATIN
MITOCHONDRION
CREATIVE PHOSPHORYLATION
Lung squamous cell carcinomas with basaloid histology represent a specific molecular entity

Christian G Brambilla, Julien Laffaire, Sylvie Lantuejoul, et al.

Clin Cancer Res Published OnlineFirst September 4, 2014.