**Legends of supplemental figures**

**Supplemental Figure 1: Embryonic stem cell-like gene expression signatures in Basaloid tumors.**

Horizontal barplots indicate the proportion of genes significantly (ttest qvalue < 0.05) up- (in red) and down-regulated (in green) in each histological subtype for each of the 16 gene signatures published in [Ben-Porath 2008] and related to ES cells transcriptome, ES cell key regulators targets, MYC targets, proliferation and Polycomb complex, in the pure (respectively mixed) basaloid SCC as compared to non-basaloid SCC (respectively left and right panels).

**Supplemental Figure 2: DNA copy number aberrations in the CIT cohort.**

Statistics concerning DNA copy number gains (red, left side) and losses (blue, right side) are represented (x-axis) along the genome (y-axis) in the CIT cohort: Histograms **(A-E)** represent gains and losses frequencies in the whole cohort **(A)** and in each subhistology (pure basaloid carcinoma **(B)**, mixed basaloid SCC **(C)**, well differentiated SCC **(D)**, poorly differentiated SCC **(E)**). Q-values from GISTIC2.0 algorithm highlighting most significant DNA copy number gains and losses are shown on histogram **(F)**. P-values from Fisher tests assessing DNA copy number gains and losses more frequently observed in basaloid SCC than in non-basaloid SCC are shown on histograms **(G-J)**: for pure basaloid versus well differentiated SCC **(G)**, pure basaloid versus poorly differentiated SCC **(H)**, mixed basaloid versus well differentiated SCC **(I)**, mixed basaloid versus poorly differentiated SCC **(J)**.

**Supplementary Figure 3: Consensus classification of the expression profiles of tumor samples in the CIT cohort.**

**(A)** The gap statistics (see Methods) calculated on consensus partitions in k=2 to k=8 classes are shown (y-axis) for each value of k (x-axis). **(B)** Heatmaps of the *(samples x samples)* co-classification matrix, representing the number of times that two samples were co-classified (in the same cluster) among the 24 unsupervised partitions for each cut from k=2 to 7 clusters. The color scale corresponds to a gradient between the minimal value (white = 0) and the maximal value (blue = 24). **(C)** Hierarchical clustering (ward linkage, euclidian distance) of the 4 consensus clusters centroids. **(D-E)** Projection of the tumor samples from the CIT cohort (n=93) in the plane of the two first principal components of a Principal Component Analysis of mRNA expression profiles, samples being colored according to sub-histology **(D)** and CIT molecular subtype **(E)**.

**Supplemental Figure 4: Kaplan-Meier curves in the CIT cohort according to CIT molecular subtypes.**

Kaplan-Meier curves of overall survival (OS) **(A)** and relapse free survival (RFS) **(B)** in the CIT cohort according to CIT molecular subtypes (clusters). The all-way logrank test p-value is shown below the curves on the right side. Are also shown: (A) the logrank test p-value obtained by comparing OS in Basaloid-like *vs* non Basaloid-like samples, and (B) the logrank test p-values obtained by comparing RFS in Basaloid-like *vs* (Classical\_2+PeriEndoAlveolar), Basaloid-like *vs* Classical\_1, and (Classical\_2+PeriEndoAlveolar) *vs* Classical\_1 groups.

**Supplemental Figure 5: Kaplan-Meier curves of overall survival in 7 validation public datasets according to CIT molecular subtypes and Wilkerson et al. subtypes**

Overall survival (OS) information is given in 7 of the 8 public datasets (unavailable information in GSE8894 [Lee 2008] dataset). Here are shown Kaplan-Meier curves of overall survival in these 7 public datasets according to CIT molecular (predicted) subtypes **(A)** and Wilkerson molecular (predicted) subtypes **(B)**. The all-way logrank test p-value is shown below the curves on the right side. Are also shown: (A) the logrank test p-values obtained by comparing OS in Basaloid-like *vs* (Classical\_2+PeriEndoAlveolar), Basaloid-like *vs* Classical\_1, and (Classical\_2+PeriEndoAlveolar) *vs* Classical\_1 groups; and (B) the logrank test p-value obtained by comparing OS between primitive and (classical+secretory+basal) groups.