SUPPLEMENTARY MATERIAL AND METHODS AND FIGURE LEGENDS

SUPPLEMENTARY MATERIAL AND METHODS

Patient characteristics and tumor samples

All tumor specimens were collected, stored and used with the patients' informed consent. Most of these patients (N=84; see also Table 1) were previously included in Affymetrix GeneChip® and CGH analyses (1). Patients from the North-East region of France underwent initial surgical resection of their localized HNSCC between 1989 and 2002 at the St Barbe Clinic (Strasbourg, France), followed by post-operative radiotherapy (89/100 cases) or chemo-radiotherapy (10/100 cases) at the Paul Strauss Cancer Center (Strasbourg, France) or the Civil Hospitals of Colmar or Mulhouse. One patient (1/100) was treated exclusively by surgery. Hematoxylin-eosin slides of paraffin-embedded tumor specimens were examined by two pathologists. All of the tumors were squamous cell carcinomas. The median age of the patients was 58 years (35-82 years). The inclusion criteria were: tumor localization (hypopharynx, oropharynx, oral cavity or tongue), any size (Tx), any lymph node status (Nx), no clinically-evident distant metastases (M0) by conventional clinical and diagnostic radiological examinations (computed tomography). The patients did not have any previous or synchronous neoplasia.

101 additional samples were used to validate the methylation data (Supplementary Materials and Methods and Table S2). Patients included in this validation group underwent initial surgical resection of their HNSCC between 1990 and 2006, followed by post-operative radiotherapy (49/101 cases; 48.5%) or by post-operative chemo- radiotherapy (52/101 cases; 51.5%). The median age of the patients was 57 years (41-78 years). The inclusion criteria were: tumor localization (hypopharynx, oropharynx, oral cavity or tongue), any size (Tx), any lymph node status (Nx), no clinically-evident distant metastases (M0) by conventional clinical and diagnostic radiological examinations (computed tomography). Patients did not have any previous or synchronous neoplasia. The median follow up period was 32 months (3-187 months). 20/101 patients developed distant metastasis to lung, bone or liver as the first recurrence during the minimal 36-months follow up period (M); 81/101 patients developed no metastasis during the same period (NM). Tumor samples were collected at the time of surgery. A fragment was taken near the advancing edge of the primary tumor (avoiding its necrotic center), immediately frozen in liquid nitrogen and stored at -70°C. The rest of the tumor was fixed in 6% buffered formalin and embedded in paraffin for histopathological analysis. Examination of sections adjacent to each tumor fragment showed that the percentage of tumor cells was over 70%.

Data sets and preprocessing

Transcriptome data: HG-U133-plus-2.0 Affymetrix array data were obtained for 89 HSNCC samples (1):

- *RNA extraction and Quality Control:* tumor samples (10 to 50 mg) were powdered under liquid nitrogen. RNA were extracted using RNAble (Eurobio, Courtaboeuf, France), followed by a clean-up step on RNAeasy columns (Qiagen, Courtaboeuf, France). Aliquots of the RNA were analyzed by electrophoresis on a Bioanalyser 2100 (version A.02 S1292, Agilent Technologies, Waldbronn, Germany) and quantified using Nano Drop™ ND-1000 (Nyxor

Biotech). Stringent criteria for RNA quality were applied to rule out degradation, especially a 28s/18s ratio above 1.8 for microarray.

- *cRNA probe production and labeling:* 3 mg of total RNA were amplified and labeled according to the manufacturer's one-cycle target labeling protocol (http://www.affymetrix.com). 10 mg of cRNA were used per hybridization (GeneChip Fluidics Station 400; Affymetrix, Santa Clara, CA). The labeled cRNAs were hybridized to HG-U133 plus 2.0 Affymetrix GeneChip arrays (Affymetrix, Santa Clara, CA). Chips were scanned with a Affymetrix GeneChip Scanner 3000 and subsequent images analyzed using GCOS 1.4 (Affymetrix).
- *Affymetrix chips quality control:* we used the R package affyQCReport to generate a QC report for all chips (CEL files). All the chips that did not pass this QC filtering step were removed from further analysis.
- *Normalization:* raw feature data from Affymetrix HG-U133A Plus 2.0 GeneChip™ microarrays are normalized using Robust Multi-array Average (RMA) method (R package affy) (2).

Methylome data from the discovery cohort: Whole-genome DNA methylation was analyzed in 84 HNSCC samples using the Illumina Infinium HumanMethylation27 assay. In brief, genomic DNA was bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA). Bisulfite-converted DNA was whole-genome amplified, enzymatically fragmented, and hybridized to the BeadChip arrays according to the manufacturer's instructions. Bisulfite conversion and microarray experiments were performed by Integragen SA (Evry, France, http://www.integragen.com). The HumanMethylation27 BeadChip examines the DNA methylation status of 27,578 CpG sites at promoter regions of 14,495 protein-coding genes. Each CpG site is represented by two beads on the array, measuring the levels of methylated (*M*) and unmethylated (*U*) DNA. The beta value DNA methylation scores for each locus are then calculated as $M/(M+U)$. Detailed information on HumanMethylation27 BeadChip can be obtained at http://www.illumina.com. Data were exported as standard report files and normalized using le R lumi package (v2.15).

Methylome data from the validation cohort: Whole-genome DNA methylation was analyzed in 101 HNSCC samples using the Illumina Infinium HumanMethylation450 assay that examines the DNA methylation status of 485,000 CpG sites (covering 99% of RefSeq genes and 96% of CpG islands). Detailed information about this technology can be obtained at http://www.illumina.com.
In brief, genomic DNA was extracted using either the Manual-MagNa Pure LC DNA II kit (Roche) or the Manual-Gentra Puregene kit (Qiagen) and bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA). DNA was then whole-genome amplified, enzymatically fragmented and hybridized (500 ng) to the BeadChip arrays according to the manufacturer's instructions. Bisulfite conversion and microarray experiments were performed by Integragen SA (Evry, France, http://www.integragen.com). The BeadChips were scanned using the Illumina HiScan SQ scanner and raw images data were imported into the GenomeStudio (v2011.1) methylation module (v1.9.2.) which was used to extract and transform the fluorescent signal intensities into beta-value (range: 0 -unmethylated site- to 1 -fully methylated site-). Data were exported as standard report files and normalized using le R lumi package (v2.15).

miRNome data: MicroRNA sequencing were obtained for 64 HSNCC samples by the *Illumina sequencing process HiSeq2000.* miRNA-Seq libraries were performed from at least 1µg of extracted total RNA with a RIN greater than 7. Before starting, total RNAs were purified with miRNeasy kit

which allows the selection of the small RNA fraction less than 100 bases. From these samples enriched in small RNAs, libraries were performed according to the protocol described in (3).

Briefly, first, a 3' adenylated DNA adaptor was ligated to the enriched sample in the absence of ATP preventing the self-ligation of miRNAs. Then a 5' RNA adaptor was ligated in the presence of ATP at the other end of the miRNAs. The RT primer complementary of the 3' adaptor was added at this stage with which it will form a duplex thereby reducing the ligation between adaptors. A reverse transcription was therefore performed from the RT primer and finally these captured miRNAs were amplified by PCR with primers complementary to the 3' and 5' adaptors. During this PCR a specific barcode was incorporated allowing individualization of each library. Each PCR was loaded on the Fragment Analyzer (AATI) for a precise quantification of each miRNA peak of interest. Based on these results an equimolar pool of about ten of different samples were performed. Finally the pooled PCR product was loaded on PAGE in order to excise the band of miRNA that was extracted and purified on a Qiagen MinElute column. Finally the pooled PCR product was loaded on PAGE in order to excise the band of miRNA that was extracted and purified on a Qiagen MinElute column.

After the sequencing platform generated the sequencing images, the data were analyzed in three steps: image analysis, base calling and bcl conversion. CASAVA (4) demultiplexed multiplexed samples during the bcl conversion step. Convert *bcl files into compressed FASTQ files. To do some quality control checks on raw sequence data, fastqc software was used. Finally, the script "Trim_adapter", provided by mirExpress software, handled the sequence file which contained adapter or not according the input of adaptor sequence. The sequence adaptor was trimmed on sequence data.

For each sample, FASTA files were processed by miRanalyzer0.3 software (5) to obtain counts for each mir of mirBase v18 (6).

Genome data: 88 HSNCC samples were analyzed with Illumina HumanCNV370-Quad v3.0 chips, containing 373,397 probes. Hybridization was performed by IntegraGen (Evry, France), according to the instructions provided by the array manufacturer. Raw fluorescent signals were imported and normalized into Illumina BeadStudio software as previously described (7) to obtain the log R ratio (LRR) and B Allele Frequency (BAF) for each SNP. The tQN (8) normalization procedure was then applied to correct for the asymmetry in BAF signals due to the bias between the two dyes used in Illumina assays. Genomic profiles were segmented by applying the circular binary segmentation algorithm (DNAcopy package, Bioconductor) (9, 10) to the LRR and BAF data separately, as previously described (11, 12). The absolute copy numbers and genotype status of segments were then determined using the Genome Alteration Print (GAP) method (11). In brief, the GAP pattern of each sample (a sideview projection of segmented LRR and BAF) is built, and the best-fitting model GAP is used to determine the ploidy of the sample, the contamination by normal cells, and the absolute copy number and genotype corresponding to each cluster of segments. Segments with an absolute copy number above (resp. below) the ploidy of the sample were considered as gains (resp. losses).
The Genomic Identification of Significant Targets In Cancer (GISTIC) methodology (13) was used to identify significantly recurrent chromosome aberrations in our data set. In short, a G score is computed for each genomic marker on the array that is proportional to both the frequency and amplitude of copy number changes at each location. GISTIC then determines the distribution of the G scores that would be expected by chance, by permuting the locations of the markers in each tumor.

A significance threshold is determined from this distribution, above which aberrations are deemed to be significantly recurrent. For each significant aberration, a "peak region" (region with the greatest frequency and amplitude of aberration, most likely to contain the driver genes) is defined, and GISTIC determines whether the signal is primarily due to broad or focal events. GISTIC was run on the Gene Pattern platform of the Broad Institute (14). For more confidence, peak regions were extended such that each peak border was validated by at least two samples in the data set.

Omics Analysis

Except where indicated, all transcriptome and genome analysis are carried out using either an assortment of R system software (15) packages including those of Bioconductor (9) (V2.9) or original R codes. R packages and versions are indicated where appropriate.

1- Unsupervised classification

Three methods were used to find unsupervised clusters within the four omics.

- A model-based clustering method called RPMM for "Recursively Partitioned Mixture Model" (16). This method models beta distributions specifically for the beta-values of methylation data and it also models gaussian distributions for the other data types. RPMM estimates parameters of mixture models via recursive Expectation-Maximization algorithms and it determines the final clustering thanks to Bayesian Information Criteria (BIC). See (16) for more details.

- A consensus clustering method proposed by Monti *et al.* in (17). This method perturbs the original data *via* resampling techniques. For each perturbed data set, a clustering algorithm is performed. A consensus clustering among the multiple runs are then assessed by taking the average over the connectivity matrices of every perturbed dataset. See (17) for more details. The Bioconductor R package ConsensusClusterPlus was then used.

- In (18) we presented another method performing a consensus clustering. This method can be detailed in the six following steps. This method was used in a subgroup of our transcriptomic data in (1):
    - **Unsupervised probe set selection :** Probe set's unsupervised selection was based on the two following criteria: a p-value of a variance test (see below) less than 0.01, a "robust" coefficient of variation (rCV) less than 10 and superior to a given rCV percentile. Eight rCV percentile thresholds were used (60%; 70%; 80%; 90%; 95%; 97.5%; 99%; 99.5%) yielding 8 lists of probe sets.

    - **Variance test:** For each probe set (P) we tested whether its variance across samples was different from the median of the variances of all the probe sets. The statistic used was $((n-1)\times Var(P) / Var_{med})$, where n refers to the number of samples. This statistic was compared to a percentile of the Chi-square distribution with (n-1) degrees of freedom and yielded a p-value for each probe set. This criteria is the same used in the filtering tool of BRB ArrayTools software (linus.nci.nih.gov/BRB-ArrayTools.html), described in the User's Manual.

- **Robust coefficient of variation:** For each probe set, the rCV is calculated as follows: having ordered the intensity values of the n samples from min to max, we eliminate the minimum value and the maximum value and calculate the coefficient of variation (CV) for the rest of the values.

- **Obtaining a series of 24 dendrograms:** We performed hierarchical clustering of the samples, using samples profiles restricted to each of the 8 probe sets lists obtained via unsupervised selection (as described above), for 3 different linkage methods (average, complete and Ward's), using 1-Pearson correlation as a distance metric (package cluster V1.9.3). This analysis produced 24 dendrograms.

- **Similarity score** To compare two dendrograms, we compare the two partitions in k clusters (k = 2..18) being obtained from these two dendrograms. To compare a pair of partitions, we used a similarity measure corresponding to the symmetric difference distance (19) NB: A similarity matrix A can be obtained from distance a matrix B by posing Aij = X – Bij, where for any pair (i,j), Bij ≤ X.

- **Calculus of a consensus dendrogram and a consensus partition** To identify the groups of samples that consistently clustered together in the 24 dendrograms (that is robust clusters obtained independently from a given clustering method and/or threshold for unsupervised genes selection), we first calculated a consensus dendrogram using an algorithm derived from Diday (20), and similar to the approach used by (17). We proceeded as follows: each of the 24 dendrograms was cut in k clusters, thus yielding 24 partitions. We then calculated the (N,N) symmetrical matrix S of sample's co-classification (N = number of samples), giving for each pair of samples the number of times they were in the same cluster group in the 24 partitions (from 0 to 24). S is a similarity matrix, and a corresponding distance matrix D is obtained by posing D(i,j) = 24 – S(i,j) (i.e. A(i,j) = max(B) – B(i,j)). Finally to obtain the consensus dendrogram we use the distance D and complete linkage hierarchical clustering method. Then, cutting this consensus dendrogram in k clusters, we obtained a consensus partition.

Only the results obtained with the RPMM method were described in the paper. Nevertheless, we showed a strong association between the three methods for the four omics (Fisher exact pvalues from 2.10E-04 to 2.49E-19).

2- Differential Analysis

Since the different studied omics have different density distributions, several differential analysis methods were used.

Transcriptome data: Based on the RMA $\log_2$ single-intensity expression data, we used moderate T-tests to identify genes differentially expressed between groups of samples, using limma R package. The H1 proportion of T-tests over the set of measured transcripts was estimated using B Storey method : $(1 - 2 \times \text{mean}\{ \text{if}(p_i > 0.5) \text{ then } 1 \text{ else } 0\}_{\text{probe set i:1..55K}})$ (21). Anova models were used for multigroup comparison. To control for multiple testing we measured the local false discovery rate.

Methylome data: Because of the beta distribution of the beta values, Wilcoxon tests were used to identify genes differentially methylated between groups of samples. The H1 proportion of T-tests over the set of measured transcripts was estimated using B Storey method : $(1 - 2 \times \text{mean}\{ \text{if}(p_i > 0.5) \text{ then 1 else 0}\}_{\text{probe set i:1..55K}})$ (21). Anova models were used for multigroup comparison. To control for multiple testing we measured the local false discovery rate.

miRNome data: Differential expression (DE) analysis is performed using likelihood ratio tests (LRT) based on a negative binomial model for gene-level read counts. We used the Bioconductor R package edgeR (22) to fit a negative binomial model to gene-level read counts and perform likelihood ratio tests of DE. edgeR was used on the original counts by passing an offset to the generalized linear model. This method was also used on normalized data. Finally, a Wilcoxon test was performed. The results of these three methods were used to select differentially expressed microRNAs.

Genome data: we used moderate T-tests to identify differential aberrations between groups of samples, using limma R package. The H1 proportion of T-tests over the set of measured transcripts was estimated using B Storey method : $(1 - 2 \times \text{mean}\{ \text{if}(p_i > 0.5) \text{ then 1 else 0}\}_{\text{probe set i:1..55K}})$ (21). Because of the large number of probes, multiple test correction was not performed but we used the sensitivity and the specificity related to each probe to select the probes characterizing the R1 subgroup.

3- Signaling pathway analysis

To identify biological features associated with unsupervised samples partitions, 17306 pathways collected from KEGG, GO, MSigDB, SMD and Biocarta (and related genes) were tested. Four methods were used to compare gene sets with sample groups:

- GSA (23): R package GSA
- globaltest (24) R package globaltest
- SAM-GS (25): original R code (available at http://www.ualberta.ca/~yyasui/SAM-GS/SAM-GS%20code.txt)
- Tuckey approach (algorithm described in the table 4 from (26): original R code

Each method gave a p-value (based on Monte-Carlo simulations). The lower the p-value, the more the genes in the gene set that are differentially expressed between the sample groups. To aggregate the results of the four methods, given a list of gene sets (pathways/ GO terms/…), we first sorted the list of gene sets for each method according to p-value, and then we calculated for each gene set the geometric mean rank across the four methods. The final order is based on this mean rank.

For the miRNome data, genes used in the signaling pathway analysis are the targets predicted simultaneously by mirTarget2.0 (27) and miRanda (28).

4- Survival analysis

Survival time was calculated from the date of surgical resection. Patients who were lost to follow-up or alive at the time of the study were treated as censored events. Survival curves were calculated according to the Kaplan-Meier method (function *Surv*, R package *survival*, V2.29) and differences between curves were assessed using the log-rank test (function *survdiff*, R package *survival*).

To find clinical criteria related to metastasis-free survival, overall survival or disease-free survival, we built univariate and multivariate cox models (function *coxph*, R package *survival*) and we selected clinical criteria on the score test p-value of the cox models. The *coxph* function was also used to calculate adjusted models (using the parameter *strata*).

Events related to metastasis-free survival included the metastasis development whatever the body localization (Bone, CNS, kidney, lung…). Events related to disease-free survival included death events associated with the disease (metastasis and/or loco-regional relapse) contrarily to the overall survival that did not distinguish the death cause.

5- Classifier building

Methylome data: the training set was composed of our 84 samples analyzed by the Illumina HumanMethylation27K arrays and the validation set was our independent data set composed of 101 samples analyzed by the Illumina HumanMethylation450K arrays.

A centroid based predictor was build using the differentially methylated genes between the studied groups. A gene was defined differentially methylated when the wilcoxon pvalue was less than 1e-03. A manhattan distance metric was used to predict either R1 versus non-R1 or Me.1 vs Me.2+Me.3.

Transcriptome data: the training set was composed of our 89 samples with transcriptome data and the validation set was the 44 samples of Cohen et al. (29).

A centroid based predictor was build using the differentially expressed genes between the studied groups. A gene was defined differentially expressed when the moderatet-test pvalue was less than 1e-05. A dqda distance metric was used to predict either R1 versus non-R1 or T.1 vs T.2+T.3.

6- Controls for batch effects.

Transcriptome data. We used probes related to ubiquitously expressed genes (actin and GAPDH). For the 22 probes associated with these genes, there are no differences between the R1 and non-R1 groups (moderate t-test q-values from 2.78E-01 to 9.61E-01).

Methylome data. We used the background estimate as control. No difference in background intensity was found between R1 and non-R1 (moderate t-test p-value = 6.68E-01).

miRNome data. Only flow cell lane effects may influence the sequencing data. There is no association between the used flow cell lane and the R1 subgroup (Fisher-exact p-value = 7.12E-01).

Genomic data. Ee used probes that are well known to be CNVs in the literature. Three probes were chosen that have been cited in 22 papers to be CNV. No difference is observed in the groups R1 and non-R1 for these three probes (moderate t-test p-value = 8.25E-01).

Validation of omic data by alternative technologies

1. Gene expression assay

Total RNA were extracted from frozen tumor tissues using DNA/RNA allprep minikits (Qiagen, France), according to the manufacturer's instructions. The integrity of extracted RNA was verified on an Agilent 2100 Bioanalyser (Agilent Technologies, Palo Alto, CA). RNA concentrations were measured using a ND-1000 NanoDrop spectrophotometer (Labtech, Palaiseau, France). 0.5 µg of extracted RNA was used for cDNA synthesis using the iScript™ cDNA Synthesis Kit (Bio-Rad), according to the manufacturer's instructions. One µl of diluted cDNA corresponding to either 5 ng or 1.25 ng of reverse transcribed RNA, was analyzed with SyberGreen (Roche, Meylan, France), in duplicate, using the LightCycler 480 real-time PCR system c). qRT-PCR data were analyzed using LightCycler® 480 software. Ct levels were normalized to the average Ct values of 2 internal controls

(housekeeping genes): *UBB* (*Ubiquitine B*) and *RPLP0* (*Ribosomal Protein Large P0*). The following genes were evaluated: *AIM2* (*Absent in melanoma 2*); *COL9A3* (*Collagen, Type IX, alpha 3*); *DSG3* (*Desmoglein 3*); *KRT16* (*Keratin 16*); SFRP1 (Secreted Frizzled-related protein 1).

The primer pair sequences were as follow:
UBB forward: 5'-GCTTTGTTGGGTGAGCTTGT-3'
UBB reverse: 5'-CGAAGATCTGCATTTTGACCT-3'
RPLP0 forward: 5'-GAAGGCTGTGGTGCTGATGG-3'
RPLP0 reverse:5'- CCGGATATGAGGCAGCAGTT-3'
AIM2 forward: 5'- GCTGCACCAAAAGTCTCTCC-3'
AIM2 reverse: 5'- TGCCTTCTTGGGTCTCAAAC-3'
COL9A3 forward: 5'-CAACGTGAGGAAGCAAGTGA-3'
COL9A3 reverse: 5'-AGGGCCTTTTGAGGTATGCT-3'
DSG3 forward: 5'-GGGCTCTTCCCCAGAACTAC-3'
DSG3 reverse: 5'-CTCCTTCTCTGCAGGGTTTG-3'
KRT16 forward: 5'-TCCCCAGCTGCATATAAAGG-3'
KRT16 reverse: 5'-GAGCTGGAGGAGGTGAACTG-3'
SFRP1 forward: 5'-AAGGGAGGCTCTCTGTAGGC-3'
SFRP1 reverse: 5'-ACCTTGGCCCTAGCGATAAT-3

2. miRNA expression assay

Total RNA was extracted from frozen tumor tissues using the miRNeasy extraction kit (Qiagen, France), according to the manufacturer's instructions. The integrity of extracted RNA was verified on an Agilent 2100 Bioanalyser (Agilent Technologies, Palo Alto, CA). RNA concentrations were measured using a ND-1000 NanoDrop spectrophotometer (Labtech, Palaiseau, France). 0.1 µg of extracted RNA was used for cDNA synthesis using the miScript Reverse Transcription Kit (Qiagen, France) according to the manufacturer's instructions. qRT-PCR was performed using this cDNA as a template in a LC480 thermocycler (Roche, Meylan, France), using the miScript SYBR Green PCR kit (Qiagen, France). qRT-PCR data were analyzed using LightCycler® 480 software. Ct levels were normalized to the average Ct values of 2 internal controls: RNU44 and let-7a.
The primer sequences were:
Let-7a: 5'-TGAGGTAGTAGGTTGTATAGTT-3'
RNU44: 5'-TGCTGACTGAACATGAAGGTCT-3'
The expression of miR-1 and miR-345 was measured with Qiagen QuantiTect primers: Hs_miR-1_2 miScript Primer Assay, MS00008358, and Hs_miR-345_3 miScript Primer Assay, MS00031766, respectively.

3. Methylation assays

Genomic DNA was extracted from frozen tumor tissues using DNA/RNA allprep minikits (Qiagen, France), according to the manufacturer's instructions. 2 µg DNA was converted with sodium bisulfite using the EpiTect Bisulfite Kit (Qiagen, France), according to the manufacturer's instructions. GPR55 and IHH primers that surround the methylation sites to be probed (cg20287234 and cg 25908985) were designed with methprimer (http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi). The amplicons contain TaqI restriction sites (TCGA) that will only be cut by the enzyme if the CpG was methylated and not modified by bisulfite. The primers are complementary to the bisulfite-modified sequences (by converting Gs to As):
GPR55 forward: 5'-TTTGGTTTTTAGTAAGTATTTGTTTAGGG-3'
GPR55 reverse: 5'-TAAACTTTATACACACCTATCCCAACTC-3'
IHH forward: 5'-GTATATTGGGGTTGAATTGTTTGTAG-3'
IHH reverse: 5'-ACACTTCTACCTAATCCTATTACTACTACT-3'
PCR products were generated using the EpiTect HRM PCR Kit (Qiagen, France), according to the manufacturer's instructions, and touchdown PCR conditions, as follows:

- 95°C: 5 min
- 95°C: 30 sec
  62->52°C: 30 sec.
  72°C: 1 min
  ➔ 10 cycles with a gradual decrease of hybridization Tm: 1.0°C per cycle.
- 95°C: 30 sec
  52°C: 30 sec
  72°C: 1 min
  ➔ 30 cycles.
- 72°C: 5 min

PCR products were analyzed by High-Resolution Melting (HRM) in a LC480 thermocycler (Roche, Meylan, France). Alternatively, DNA was purified with the MinElute® PCR Purification Kit (Qiagen, France) digested with 10U of TaqI (Invitrogen, Life technologies, France) for one hour at 65°C. 10 µl samples were resolved by electrophoresis on 3% agarose gels.

REFERENCES

1. Rickman DS, Millon R, De Reynies A*, et al.* Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. Oncogene 2008.
2. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research 2003;31: e15.
3. Vigneault F, Ter-Ovanesyan D, Alon S*, et al.* High-throughput multiplex sequencing of miRNA. Current protocols in human genetics / editorial board, Jonathan L Haines [et al 2012;Chapter 11: Unit 11 2 1-0.
4. Illumina. Illumina iControlDB web page. http://www.illumina.com/science/icontroldb.ilmn.
5. Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic acids research 2009;37: W68-76.
6. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research 2011;39: D152-7.
7. Peiffer DA, Le JM, Steemers FJ*, et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome research 2006;16: 1136-48.
8. Staaf J, Vallon-Christersson J, Lindgren D*, et al.* Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. BMC bioinformatics 2008;9: 409.
9. Gentleman RC, Carey VJ, Bates DM*, et al.* Bioconductor: open software development for computational biology and bioinformatics. Genome biology 2004;5: R80.
10. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics (Oxford, England) 2007;23: 657-63.
11. Popova T, Manie E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome biology 2009;10: R128.
12. Staaf J, Lindgren D, Vallon-Christersson J*, et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome biology 2008;9: R136.

13. Beroukhim R, Getz G, Nghiemphu L*, et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences of the United States of America 2007;104: 20007-12.
14. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nature genetics 2006;38: 500-1.
15. R-project-V14.1. http://www.R-project.org.
16. Houseman EA, Christensen BC, Yeh RF*, et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC bioinformatics 2008;9: 365.
17. Monti S, Savage KJ, Kutok JL*, et al.* Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood 2005;105: 1851-61.
18. Boyault S, Rickman DS, de Reynies A*, et al.* Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology (Baltimore, Md 2007;45: 42-52.
19. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci 1981: 131-47.
20. Diday E. New concepts and new methods in automatic classification. Paris: University Paris VI; 1972.
21. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 2003;100: 9440-5.
22. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 2010;11: R25.
23. Efron B. On testing the significance of sets of genes (technical reports) http://www-stat.stanford.edu/~tibs/GSA. 2006.
24. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics (Oxford, England) 2004;20: 93-9.
25. Dinu I, Potter JD, Mueller T*, et al.* Improving gene set analysis of microarray data by SAM-GS. BMC bioinformatics 2007;8: 242.
26. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics (Oxford, England) 2007;23: 980-7.
27. Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics (Oxford, England) 2008;24: 325-32.
28. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome biology 2003;5: R1.
29. Cohen EE, Zhu H, Lingen MW*, et al.* A feed-forward loop involving protein kinase Calpha and microRNAs regulates tumor cell cycle. Cancer research 2009;69: 65-74.

SUPPLEMENTARY FIGURE LEGENDS.

Supplementary Figure S1. Venn diagram showing the overlap between patients analyzed at the level of the genome, transcriptome, methylome and miRNome. N = total number of samples for each type of analysis. 60 samples were analyzed by all four approaches.

Supplementary Figure S2. Metastasis free survival curves of the predicted R1 subgroup (A) and the predicted Me.1 and Me.2+3 (C) clusters for the methylation data for 101 independent patient samples and on the R1 subgroup (B) and the T.1-3 clusters (D) for gene expression data for 44 independent patient samples from Cohen *et al*.(1). Log-rank analysis was used to compare the survival distributions of the indicated groups.

Supplementary Figure S3. Metastasis free survival curves of the three unsupervised clusters vMe.1-3 generated by RPMM using the methylation validation data.

Supplementary Figure S4. Box-and-whisker plots of the expression of the *AIM2*, *KRT16*, *DSG3*, *SFRP1* and *Col9A3* mRNAs (A), and miR-345 and miR-1 (B), as measured by qRT-PCR on RNA extracts from R1 and non-R1 tumor samples.

Supplementary Figure S5. Evaluation of the methylation levels of candidate CpG islands. (A) Methylation scores for cg20287234 and cg 25908985, found in the regions of the *GPR55* and *IHH* genes, respectively, for six R1 and four non-R1 samples. (B, C) High-Resolution Melting (HRM)-PCR on bisulfite-converted DNA from the R1 and non-R1 samples using GRP55 (B) and IHH (C) specific primers.

Supplementary Figure S6. Combined Bisulfite Restriction Analysis (COBRA) of the DNA amplicons obtained with the GRP55 primers on bisulfite-converted DNA from the six R1 and four non-R1 samples (see Supplementary Figure S5A). PCR amplification products were digested with TaqI. Methylated DNA gives the lower bands that are not resolved on these gels. Non digested DNA from one R1 sample and one non-R1 sample were loaded on the gel as controls.

Supplementary Figure S7. Pangenomic profiles of the R1 subgroup (A) and of the other samples (B). Frequencies of gain (upper part) are shown in red, frequencies of loss (lower part) are shown in green.

Supplementary Figure S8. Chi2 p values (log10 scale) of the comparison of gains (red) and losses (green) between R1 and the other samples.