## Material and Methods

**Patient selection**

Tumor biopsies were obtained from either primary or metastatic sites and were immediately fresh-frozen. Tumor cellularity was assessed by a senior pathologist on a haematoxylin and eosin slide from the same biopsy core as the one used for nucleic acid extraction and further analysis by CGH and sequencing. Only samples presenting more than 30% of tumor cells were considered contributive. This study included the data from 92 consecutive patients with contributive biopsies analyzed with the four techniques (TGS, CGH, WES and RNAseq). Samples were originally analyzed between May and December 2014.

**Targeted gene panel sequencing and analysis**

Molecular analysis using TGS and CGH was carried out as previously described. Briefly, tumor DNA was extracted using DNeasy tissue kit (Qiagen, Hilden, Germany) according to manufacturer’s instructions. TGS was performed using Personal Genome Machine (Ion Torrent PGM, Thermo Fischer Scientific®) with a 74-gene panel, which included the Ion AmpliSeq™ Cancer Hotspot Panel v2 (CHP2) (Thermo Fischer Scientific, Darmstadt, Germany, see [www.ampliseq.com](http://www.ampliseq.com)) as well as custom made amplicons. The panel was designed in order to cover most known predictive markers. Analysis was performed as previously described[1]. All germline variants found in 1000 Genomes Project or ESP (Exome Sequencing Project database) with frequency >0.1% were removed. Filtered variants were reviewed, classified and annotated by an expert molecular medical geneticist according to available knowledge (COSMIC, TCGA and medical literature). Variants of interest were selected to generate clinical reports.

**Comparative Genomic Hybridization and analysis per patient**

CGH was performed following standard operating procedures from Agilent Technologies (Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis G4410-90010). Agilent SureTag DNA Labeling Kit (Agilent, 5190-3400) was used to label Tumor DNA and sex-matched control DNA (Human Genomic DNA Female, G1521, or Human Genomic DNA Male, G1471, Promega) with either cyanine 3 or cyanine 5. 0.5 µg of genomic DNA was used as for the labeling reaction. DNA was fragmented by a double enzymatic digestion (Alu I+ Rsa I) and checked with Bioanalyzer and DNA 7500 Kit (Agilent Technologies). Tumor DNA and control DNA were labelled by random priming with Cy5-dUTP and Cy3-dUTP, respectively. Cyanine 3 and cyanine 5 labeled samples were purified, mixed and hybridized to Agilent microarrays from the SurePrint G3 Human CGH Microarray Kit, 8x60K (Agilent, G4450A, Design ID 021924) 4x180K (Agilent, G4449A, Design ID 022060) at 65°C for 24h at 20 rpm.

Microarrays were scanned using an Agilent DNA Microarray Scanner (Agilent, G2505C). Intensities were extracted from the raw image files using Agilent's Feature Extraction software v10.7.3.1, with protocol versions CGH\_107\_Sep09\_1\_1, and grid versions 022060\_D\_F\_20111015. Resulting raw intensities were normalized according to their dye composition (Cy3 fitted over Cy5). Data were transformed to log2(Test/Ref) and normalized according to their local GC content through a LOWESS regression. Corresponding profiles were segmented with CBS algorithm[2] implemented in the DNAcopy package (v1.21 to v1.42) for the R environment (v2.6 to v3.1) using default parameters. Manual reprocessing using the *undo.sd* parameter to a maximum value of 0.75 was performed in case for a few noisier profiles. Profiles were centered using the most centered out of the three most populated peaks of the smoothed log2(Test/Ref) distribution. Aberration levels were called by setting a log2(Test/Ref) threshold automatically adapted to the internal noise for each profile, considered as one-fourth of the median value of the absolute differences between consecutive log2(Test/Ref) measures along the genome. All genomics coordinates were established using the human genome as defined by the UCSC build hg19 (GRCh37).

**Cohort analysis of CNA profiles**

For the cohort analysis, profile fragmentation was performed depending on the occurrence of a breakpoint in at least one profile, leading to a matrix of log2ratio values with the genome fragments (segments) in lines and the samples in columns. Using this log2ratio matrix, a hierarchical clustering was then performed according to Ward’s method by comparing the Euclidian distance between profiles, leading to detection of pairs. Similarly, a correlation matrix was generated by computing Pearson’s correlation two by two on all combinations of CGH vs. WES pairs amongst columns of the log2ratio matrix. This allows the creation of a new correlation matrix with CGH in lines and WES in columns, making it possible to determine whether correlation was higher between paired analysis (same sample) or if noise prevents proper pairing of CGH and WES copy number profiles.

For the paired analysis, pairs were compared two by two in an independent manner, by calculating the kappa coefficient, correlation, and regression slope of log2ratios and percentage of identity. These comparisons were computed using a fragmentation of profiles based on the occurrence of a breakpoint in any of the two profiles. Percentage of identity and kappa coefficient were computed on status values (gain, normal, loss) of segments, according to a log2ratio threshold of +/- 0.7. Correlation and regression slope of profile pairs were computed based on log2ratio values.

**Whole exome sequencing and analysis**

Library preparation, capture, and sequencing, were performed by IntegraGen, Evry, France. Genomic DNA was captured using Agilent in-solution enrichment methodology (SureSelect, Agilent) with their biotinylated oligonucleotides probes library (Clinical Research Exome - 54 Mb, Agilent), followed by paired-end 75 bases massively parallel sequencing on Illumina NextSeq 500. For detailed explanations of the process[3]. Sequence capture, enrichment and elution were performed according to manufacturer’s instruction and protocols (SureSelect, Agilent) without modification except for library preparation performed with NEBNext® Ultra II kit (New England Biolabs®). For library preparation, 150 ng of each genomic DNA were fragmented by sonication and purified to yield fragments of 150-200 bp. Paired-end adaptor oligonucleotides from the NEB kit were ligated on repaired, A tailed fragments then purified and enriched by 8 PCR cycles. 1200ng of these purified libraries were then hybridized to the SureSelect oligo probe capture library for 24 hr. After hybridization and washing, captured products were PCR-amplified with 9 cycles, purified and quantified by QPCR to obtain sufficient DNA template for downstream applications. Each eluted-enriched DNA sample was then sequenced on an Illumina NextSeq 500 as paired-end 75b reads.

Base calling was performed using the Real-Time Analysis software sequence pipeline (RTA2) with default parameters. Sequence reads were mapped to the human genome (build hg19) using bwa mem. PCR and optical duplicates were removed using PicardTools-SVNs and InDels were called using three methods: GATK for constitutional mutations, MuTecT for high specificity variant calling, and VarScan2 for broad-spectrum variant calling. Variants annotation was done using SnpEff and SnpSift with dbSNP (dbsnp132), Cosmic, and dbNSFP. CNA analysis was carried out by IntegraGen using exomeCNV and the Bioconductor DNACopy package (DNAcopy 1.32.0) by comparing tumor from each sample to the normal DNA exome data. Filtered variants were reviewed by an expert molecular medical geneticist to generate clinical reports.

Secondary analysis (e.g. comparison between WES and TGS) was conducted using BCFTools[4] and in-house scripts.

**Tumor mutation loads, mutational signatures, HLA genotyping and neoepitope prediction**

TML was calculated using MuTecT’s results based on missense variants only and the following thresholds were used: 14X for minimum depth of tumor samples, 8X for minimum depth or normal samples, 0.1 for minimum allelic frequency. Mutational signatures were identified using the Somatic Signatures R package[5] with default parameters. The association between mutational signatures and clinical and molecular data was assessed using Pearson correlation and Wilcoxon tests.

The HLA typing of 10 samples were analyzed with high resolution, using the gold standard technique (PCR-SSO reverse and/or PCR-SSP). Patients MHC Class I were determined using Optitype v1.0[6], HLA-VBSeq[7] and Phlat[54]. We used netMHC v3.4[8] to predict binding affinity between a patient-specific MHC Class I and every 9-amino acids string containing a mutant amino acid. Peptides with binding affinity lower than 500nM were considered as possible neoantigens.

**RNA sequencing**

Library preparation, capture, sequencing and bioinformatics analysis were performed by IntegraGen. Libraries were prepared using the TruSeq Stranded mRNA kit (Illumina) following manufacturer’s instruction. Briefly, the TruSeq stranded mRNA sample prep kit converts the poly-A containing mRNA in total RNA (1000ng engaged in the process) into a cDNA library using poly-T oligo-attached magnetic bead selection. Following mRNA purification, the RNA was chemically fragmented prior to reverse transcription and cDNA generation. The fragmentation step resulted in an RNAseq library that included inserts ranging in size from approximately 400mers. The cDNA fragments then went through an end repair process, the addition of a single ‘A’ base to the 3’ end and then ligation of the adapters. Finally, the products were purified and enriched with PCR to create the final double stranded cDNA library, which was then purified and quantified by QPCR. Each transcriptome library was sequenced on an Illumina NextSeq 500 as paired-end 75b reads. RNAseq base calling was performed using the Real-Time Analysis software sequence pipeline (RTA2) with default parameters.

**Identification of fusion transcripts**

A new analysis pipeline was developed in order to optimize detection of potential fusion transcript. Three different tools were used, namely TopHat v2.0.14[9], Defuse v0.6.2[10] and CRAC 1.11.4[11]. The GRCh37/hg19 build was used in all case as the human reference genome. Results obtained with CRAC were filtered using chimCT v0.12 (<http://cractools.gforge.inria.fr/softwares/chimct/>) and those obtained with Tophat were filtered using Tophat-fusion-post[12]. For all fusions detected, an oncogenic potential score was computed thanks to the Oncofuse algorithm[13]. Relevant fusions were selected according to several criteria: read support, number of tools that detected it, biological relevance of the two genes included in the fusion, and Oncofuse score. Fusions identified by the various tools were then combined in a single view using an in-house metacaller.

**RNA variant calling**

RNAseq rawdata were mapped against human genome (GRCh37/hg19) with STAR (v2.4.0) 2-pass method[14] and potential duplicates were marked using Picard tools (<http://picard.sourceforge.net/>). Remaining reads were split into exon segments and STAR mapping qualities were reassigned in order to fit GATK (v3.4-0) IndelRealignment requirements[15–17]. After local realignment around indels, a base quality score recalibration (BQSR) process was applied and the variant calling step was done with HaplotypeCaller in RNAseq mode. Finally, raw variants list obtained above was filtered on Phred-scaled p-value using Fisher's exact test to detect strand bias (FS > 30.0) and Variant Confidence/Quality by Depth (QD < 2.0) values.

**RNA gene expression analysis**

RNAseq rawdata were processed against human genome (hg19) and RefSeq annotations with RSEM (v1.2.21)[18] and using Bowtie2 mapping algorithm[19]. TPM (Transcripts Per Million) values per gene were used to compare gene expression between samples. A sample was described as outlier for a gene or positive for *ESR1* and *ERBB2* in the following cases:

TPM < (lower\_quartile - 1.5x(upper\_quartile - lower\_quartile))

or
TPM > (upper\_quartile + 1.5x(upper\_quartile - lower\_quartile))

**Detection of HPV DNA viral load using qPCR**

Human papillomavirus types were detected by PCR as previously described[20]. Briefly, the E6/E7 regions of the three oncogenic HPV strains (16, 18 and 33) were targeted using the TaqMan® PCR Core Reagents Kit (Ref. N8080228, Applied Biosystems Inc, CA, USA). We used the gACTB (β-actin) as a housekeeping gene and obtained a range of standard curves, which allowed us to standardize the amount of gDNA tested in each assay. We also obtained, in each assay, a range of standard curves for decremental plasmid concentrations. Primers and probes were synthesized by Sigma Aldrich Co. LLC. All samples were tested in duplicates, and the assay was performed in an ABI PRISM™7700 Sequence Detector (Applied Biosystems Inc).

RNAseq raw reads were mapped with BWA 0.7.10 using the following parameters "mem -M -t 8 -A 2 -E 1" against a database composed of the three major HPV types: gi|310698439:83-559 Human papillomavirus type 16, complete genome, gi|310698439:562-858 Human papillomavirus type 16, complete genome, gi|406856765|gb|JX412194.1| Human papillomavirus type 18 isolate PE56 E6 protein gene, complete cds, gi|406856783|gb|JX412203.1| Human papillomavirus type 18 isolate PE56 E7 protein gene, complete cds, gi|256260302|gb|GQ374552.1| Human papillomavirus type 33 isolate 5 E6 protein (E6) gene, complete cds, gi|257472285|gb|GQ465853.1| Human papillomavirus type 33 isolate 5 E7 protein (E7) gene, complete cds. Aligned reads were then counted for each type.

**Ethics**

Written informed consent for tumor and matched normal DNA sequencing analysis was obtained for all patients at Gustave Roussy for the MOSCATO v1 approved protocol IRB CSET 2011/1755 (NCT01566019 trial). TGS and CGH were performed at the Gustave Roussy genomic core (Villejuif, France). WES was performed at on the Gustave Roussy IntegraGen core (Villejuif, France). The study was conducted according to the Good Clinical Practice Guidelines of the Annals of Oncology International Conference on Harmonization and the Declaration of Helsinki recommendations.

[1] Massard C, Michiels S, Ferté C, Le Deley M-C, Lacroix L, Hollebecque A, et al. High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. Cancer Discov 2017:1–11. doi:10.1158/2159-8290.CD-16-1396.

[2] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557–72. doi:10.1093/biostatistics/kxh008.

[3] Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:182–9. doi:10.1038/nbt.1523.

[4] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27:2987–93. doi:10.1093/bioinformatics/btr509.

[5] Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. Bioinformatics 2015;31:3673–5. doi:10.1093/bioinformatics/btv408.

[6] Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 2014;30:3310–6. doi:10.1093/bioinformatics/btu548.

[7] Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, et al. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. BMC Genomics 2015;16 Suppl 2:S7. doi:10.1186/1471-2164-16-S2-S7.

[8] Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 2008;36. doi:10.1093/nar/gkn202.

[9] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14:R36. doi:10.1186/gb-2013-14-4-r36.

[10] McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MGF, et al. Defuse: An algorithm for gene fusion discovery in tumor rna-seq data. PLoS Comput Biol 2011;7. doi:10.1371/journal.pcbi.1001138.

[11] Philippe N, Salson M, Commes T, Rivals E. CRAC: an integrated approach to the analysis of RNA-seq reads. Genome Biol 2013;14:R30. doi:10.1186/gb-2013-14-3-r30.

[12] Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011;12:R72. doi:10.1186/gb-2011-12-8-r72.

[13] Shugay M, Ortiz de Mendíbil I, Vizmanos JL, Novo FJ. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics 2013;29:2539–46. doi:10.1093/bioinformatics/btt445.

[14] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 2013;10:1185–91. doi:10.1038/nmeth.2722.

[15] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. doi:10.1101/gr.107524.110.

[16] DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8. doi:10.1038/ng.806.

[17] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma 2013. doi:10.1002/0471250953.bi1110s43.

[18] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.

[19] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9. doi:10.1038/nmeth.1923.

[20] Melkane AE, Mirghani H, Aupérin A, Saulnier P, Lacroix L, Vielh P, et al. HPV-related oropharyngeal squamous cell carcinomas: a comparison between three diagnostic approaches. Am J Otolaryngol 2014;35:25–32. doi:10.1016/j.amjoto.2013.08.007.