Supplementary file 2

Bioinformatic methods

Genetic analysis approach was briefly described. The first step is quality control where different features were evaluated: the total number of raw reads, the number of reads correctly mapping to the reference genome, the percentage of reads classified as PCR duplicates, the enrichment, the coverage and the median depth.

Fastq files were aligned to the reference genome (human genome version 19) by BWA-mem algorithm, PCR duplicates were removed using RMDUP command of SAMtools and sequencing data were then filtered to delete known sequencing artifacts that could alter analysis results. Reads containing more than three mismatches and bases with Phred Score < 30 were filtered out.

Alignments from PBMC and tumor samples (tissue / plasma / urine) were compared to identify mutations and indels in tumor and matched normal samples. Mutations and indels common to both samples were classified as germline, alterations present in tumor but not in normal were classified as somatic. Since urine DNA presented many alterations which were not cancer related, we decided to perform a supervised analysis for the rest of the study considering only molecular alterations which were previously identified in the matched tumor tissue. All mutations that were supported by reads having a strand bias or by mismatch in first or last nucleotide of the read were discarded. Reads supporting MUC6 and MUC2 variations were filtered out because DNA sequence of the human mucin genes have not been completely resolved due to the repetitive nature of their central exon coding for Proline, Threonine and Serine rich sequence [1]

For each matched analysis (tissue, plasma, urine vs PBMC) a list of acquired mutations (target mutation) were identified. For each target mutation, two groups were created: the first group includes reads supporting mutations in the second group includes reads supporting wild type. For each group, the read length distribution was calculated using the nucleotide distance between the first position and the last position of each read aligned to the human genome (version 19, reference). Reads longer than 1000 nucleotides and reads with alteration (indel or mutation) in the nucleotide before or after the position of the somatic mutation were filtered out.

For each sample, the unique allelic profile (based on Single Nucleotide Polymorphism Identification) of each patients was identified using the dbSNP (version 147). Only alleles with

fractional abundance above 20% and depth above 10X were considered in order to verify that all sequenced samples corresponding to the correct patient.

Gene copy number (GCN) was obtained by calculating the ratio of median gene depth to the median depth of whole exome. For each gene, its GCN in the normal and tumor samples and the corresponding copy number variation (ratio tumor/matched GCN) were reported. The circular binary segmentation (CBS) algorithm, as implemented in the DNAcopy R module, was used to cluster all the gene copy-number alterations.

1. Svensson F, Lang T, Johansson MEV, Hansson GC: **The central exons of the human MUC2 and MUC6 mucins are highly repetitive and variable in sequence between individuals.** *Sci Rep* 2018, **8:**17503.