scientific reports

Check for updates

OPEN

Medium levels of transcription and replication related chromosomal instability are associated with poor clinical outcome

Ataaillah Benhaddou[1], Laetitia Gaston[2], Gaëlle Pérot[1,3], Nelly Desplat[4], Laura Leroy[1,5], Sophie Le Guellec[1,5], Mohamed Ben Haddou[6], Philippe Rochaix[1,5], Thibaud Valentin[1,7], Gwenaël Ferron[1,8], Christine Chevreau[7], Binh Bui[9], Eberhard Stoeckle[10], Axel Le Cesne[11], Sophie Piperno-Neumann[12], Françoise Collin[13], Nelly Firmin[14], Gonzague De Pinieux[15], Jean-Michel Coindre[16], Jean-Yves Blay[17,18] & Frédéric Chibon[1,5,19]✉

Genomic instability (GI) influences treatment efficacy and resistance, and an accurate measure of it is lacking. Current measures of GI are based on counts of specific structural variation (SV) and mutational signatures. Here, we present a holistic approach to measuring GI based on the quantification of the steady-state equilibrium between DNA damage and repair as assessed by the residual breakpoints (BP) remaining after repair, irrespective of SV type. We use the notion of Hscore, a BP "hotspotness" magnitude scale, to measure the propensity of genomic structural or functional DNA elements to break more than expected by chance. We then derived new measures of transcription- and replication-associated GI that we call iTRAC (transcription-associated chromosomal instability index) and iRACIN (replication-associated chromosomal instability index). We show that iTRAC and iRACIN are predictive of metastatic relapse in Leiomyosarcoma (LMS) and that they may be combined to form a new classifier called MAGIC (mixed transcription- and replication-associated genomic instability classifier). MAGIC outperforms the gold standards FNCLCC and CINSARC in stratifying metastatic risk in LMS. Furthermore, iTRAC stratifies chemotherapeutic response in LMS. We finally show that this approach is applicable to other cancers.

Genome instability (GI) is a hallmark of cancer[1] and may arise due to deleterious mutations in components of DNA repair pathways or to abnormally high levels of genotoxic stress from cellular processes such as transcription and replication that overwhelm high-fidelity DNA repair[2]. Replication stress is a threat to genome stability and has been implicated in tumorigenesis[3–5]. Notably, common fragile sites in cancer colocalize with chromosomal

[1]OncoSarc, INSERM U1037, Cancer Research Center in Toulouse (CRCT), 31000 Toulouse, France. [2]Department of Medical Genetics, CHU de Bordeaux, 33000 Bordeaux, France. [3]Centre Hospitalier Universitaire (CHU) de Toulouse, IUCT-Oncopole, 31000 Toulouse, France. [4]INSERM UMR1218, ACTION, Institut Bergonié, 33000 Bordeaux, France. [5]Department of Pathology, Institut Claudius Régaud, IUCT-Oncopole, 31000 Toulouse, France. [6]Mentis Consulting, 1000 Brussels, Belgique. [7]Department of Oncology, Institut Claudius Régaud, IUCT-Oncopole, 31000 Toulouse, France. [8]Department of Surgical Oncology, Institut Claudius Régaud, IUCT-Oncopole, 31000 Toulouse, France. [9]Department of Oncology, Institut Bergonié, 33000 Bordeaux, France. [10]Department of Surgery, Institut Bergonié, 33000 Bordeaux, France. [11]Department of Oncology, Institut Gustave Roussy, 94800 Villejuif, France. [12]Department of Medical Oncology, Institut Curie, 75005 Paris, France. [13]Department of Pathology, Centre Georges-François Leclerc, 21000 Dijon, France. [14]Department of Oncology, Institut Régional du Cancer de Montpellier, 34000 Montpellier, France. [15]Department of Pathology, Hôpital Universitaire Trousseau, 37170 Tours, France. [16]Department of Pathology, Institut Bergonié, 33000 Bordeaux, France. [17]Department of Medical Oncology, Centre Léon Bérard, 69000 Lyon, France. [18]Centre Léon Bérard, Université Claude Bernard Lyon 1, INSERM U1052, CNRS 5286, 69000 Lyon, France. [19]Cancer Research Center in Toulouse (CRCT), 2 Avenue Hubert Curien, 31037 Toulouse, France. ✉email: Frederic.chibon@inserm.fr

regions that are particularly prone to breakage following mild replication stress[6–8]. Transcription also creates conditions for mutations and recombination as well as DNA breaks, either by transcription-associated processes or by its ability to become a barrier to DNA replication[9–13]. Indeed, co-transcriptional R-loops constitute a barrier for replication fork progression and lead to fork stalling and collapse. This is thought to be a major mechanism of GI that involves transcription-replication collisions[14–17]. Finally, obstacles on the template DNA, such as non-B DNA (NBD) structures, DNA repeats, DNA-bound non-histone proteins and transcription complexes, can impede replication fork progression[18–21]. Cahill et al.[22] proposed that GI might contribute to oncogenesis only if it does not exceed a certain threshold above which it is likely to generate cells with unviable karyotypes. Interestingy, Radiation therapy as well as many of chemotherapeutic drugs have been shown to induce GI in vitro resulting in the acquisition of additional SV beyond a threshold that would be compatible with cell survival[23–25]. Indeed, the efficacy of some cancer treatments that induce GI, such as paclitaxel and radiation therapy, is improved in cells with a higher basal rate of GI[23, 26, 27]. Concordantly, high burden of somatic copy number alterations and greater levels of intratumor heterogeneity before treatment are both associated with better survival outcomes, while tumors with lower levels of tumor heterogeneity are associated with a poor clinical outcome[28, 29].

Leiomyosarcoma (LMS) with smooth muscle cell (SMC) differentiation is one of the most frequent soft tissue sarcomas (STS), a group of tumors that arise from connective tissue cells[30]. LMS develops due to frequent p53 and RB1 pathway alterations[31, 32] and a highly rearranged genome with a high number of chromosomal rearrangements. This leads to many copy number variations (CNV) and BP that are associated with poor outcome[33]. The stratification of LMS has long been based on histological measures like the Fédération Nationale de Centre de lutte contre le Cancer (FNCLCC) grading[34] and is currently challenged by expression-based signatures[31]. Next-generation sequencing (NGS) has recently demonstrated the ability to identify clinically actionable genetic variants across many genes[35]. Current approaches in cancer genomics often use exome-seq to build a catalogue of mutations in multiple cancer types by sequencing hundreds of tumor samples in order to find diagnostic, prognostic, and therapeutic targets[36]. While this approach remains relevant in most cancer types, it is rapidly becoming insufficient in highly rearranged cancer types like LMS, where recurrent driver gene mutations are very rare[37] and it is more likely that their rearranged genome is actually the driving force of oncogenesis[38]. Here we used whole genome sequencing (WGS) of LMS tumor samples to identify SV across all tumor genomes and to infer the mechanisms of GI at the genome level.

We therefore sought to evaluate whether transcription complexes, as well as NBD and DNA repeats, play any role in LMS GI. To do so, we used WGS data to compare LMS structural variations (SV) BP to known sites of transcription complexes, NBD and DNA repeats, and evaluated whether these BP are enriched more than expected. We present Hscore, a new Hotspotness magnitude scale. Hscore measures the propensity of a given DNA element to break more than expected with a random breakage model. By using this approach, we have developed a combination of both transcription-associated and replication-associated markers of GI and tested whether both measures are prognostic of metastatic risk in LMS. Furthermore, we assessed whether the transcription- and replication-associated markers are predictive of chemotherapeutic response in LMS.

Results

Since the LMS genome is highly rearranged, the question arises whether DNA breakages occur randomly or are due to mechanisms associated with structural and/or functional DNA elements and are thus specific to certain regions of the genome. To address this question, we tested the "hotspotness" of three different types of genomic DNA structure: regulatory DNA elements, and NBD and DNA repeats. We propose that Hscore may be used to assess the propensity of DNA elements to break more than expected by chance under the random breakage model (RBM) (Fig. 1A). Hscore for a given type of DNA element is defined as the − log10 of the probability of having more BP than those observed in those elements, given the total number of BP, the cumulative size of the DNA elements and the readable genome size ("Materials and methods" section). Hscore is very intuitive: the higher it is, the more significantly a given DNA element is more broken than expected. To determine whether a given genomic feature is a hotspot, we computed Hscore in sliding windows upstream and downstream and made profile plots (Fig. 1A; see "Materials and methods" section). Two significantly broken structures emerged: hotspots and hot regions. We define the characteristics of hotspots as follows: (1) DNA elements are significantly more broken than expected by chance (Hscore ≥ 3) while the immediate surrounding regions are not (Hscore < 3); (2) if the DNA elements and the immediate surrounding regions have a Hscore ≥ 3, a DNA element is considered as a hotspot if its Hscore is at least 1.5 times the Hscore of both surrounding regions. A hot region is inferred if an Hscore ≥ 3 is observed for both the DNA element and its surrounding regions, provided that Hscore of the DNA element is less than 1.5 times the Hscore of the surrounding regions.

Whole genome sequencing of 112 LMS (Supplemental Tables S1, S2) with our homemade SV detection pipeline (see "Materials and methods" section) allowed us to identify 24,870 BP forming 12,435 SV. Each SV is formed by 2 BP. Of all BP, 67.4% (16,764 BP) are implicated in intra-chromosomal SV (BPSVintra) and 32.6% (8106 BP) in inter-chromosomal SV (BPSVinter). While most BP occur inside regulatory DNA elements (13,377/24,870 = 53.79%), some affect NBD (4902/24,870 = 19.71%) and others (3574/24,870 = 14.37% of BP) arise in DNA repeats. Taken together, a total of 66.4% (16,510/24,870) of LMS BP were identified in the DNA elements in this study.

The total BP count per LMS (TBPc) was highly skewed (ranging from 26 to 1200; mean = 222.1, median = 181; Supplemental Fig. S1A) and did not fit a normal distribution (Shapiro–Wilk test $P = 1.4 \times 10^{-12}$; Supplemental Fig. S1A), but TBPc follows a log-normal distribution (Shapiro–Wilk test P = 0.81; Supplemental Fig. S1B). This is not surprising since log-normal distributions are common in nature and reflect forces acting independently and whose interactions result in multiplicative effects[39]. In cancer, these forces may include genetic, environmental, physiological, immune, sub-cellular and supra-cellular constraints.

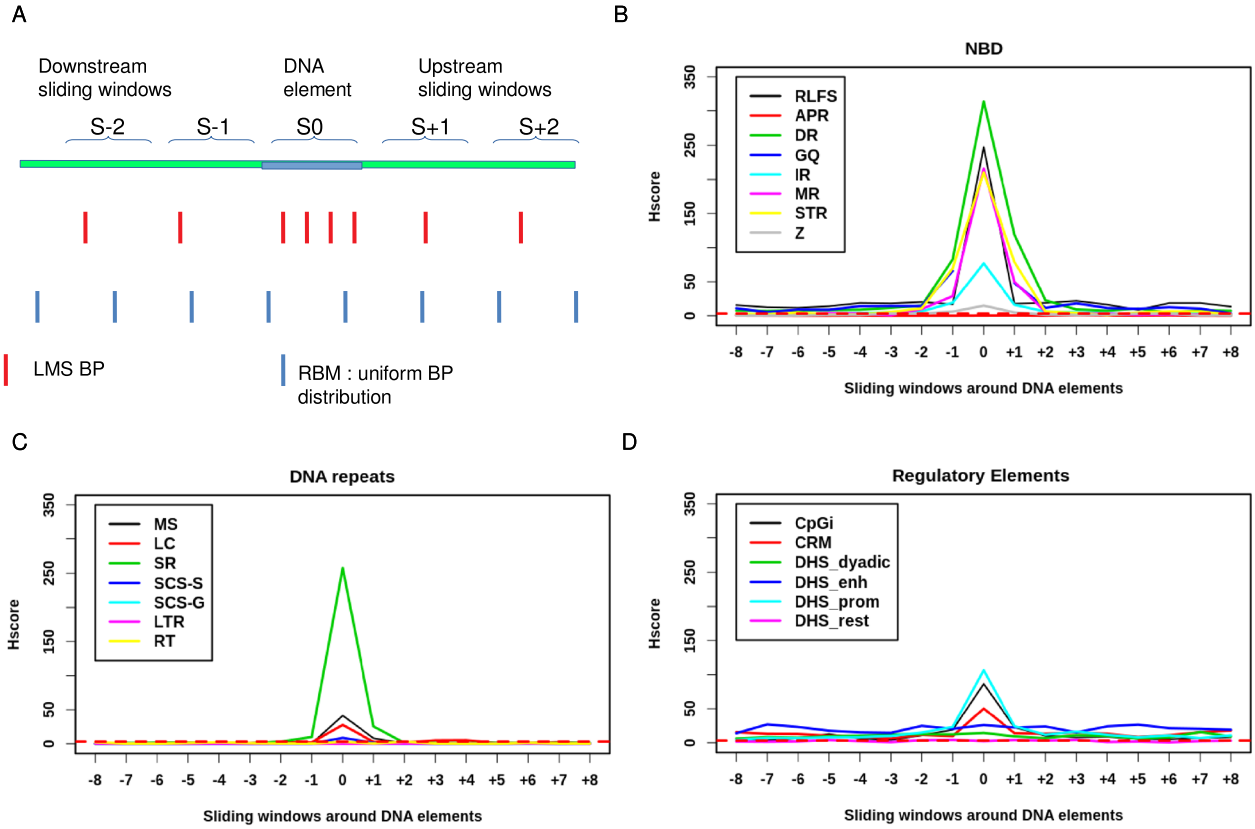

Figure 1. Regulatory elements, non-B DNA (NBD) and DNA repeats are hotspots. (**A**) Random Breakage Model: hotspots are defined as DNA elements containing more BP than expected under RBM and more than surrounding regions. Hscore profile for DNA elements and sliding windows upstream and downstream for (**B**) Non-B DNA: R-Loops Forming Sequences (RLFS), a-Phased Repeats (APR), Direct Repeats (DR), G-quadruplex (GQ), Inverted Repeats (IR), Mirror Repeats (MR), Short Tandem Repeats (STR), Z-DNA (Z) and (**C**) DNA repeats: MicroSatellites (MS), Low Complexity DNA (LC), Simple Repeats (SR), Self-Chains Segements-self-aligned (SCS-S) Self-Chains Segements-Gaped (SCS-G), Long terminal repeats (LTR), RetroTransposons (RT), (**D**) regulatory elements: CpG islands (CpGi), Cis-Regulatory Modules (CRM), Dnase Hyper sensitive sites (DHS) of type dyadic (DHS_dyadic), Enhancer (DHS_enh), Promoter (DHS_prom), other types (DHS_rest). Horizontal red dashed line corresponds to Hscore threshold of 3 for hotspotness. (**A**) was drawn using LibreOffice Impress 1:6.0.7ubuntu0.18.04.10.

NBD and DNA repeats are hotspots for DNA breakage in LMS. NBD are DNA elements that adopt non-canonical DNA structures[18]. Sequences prone to form NBD are widespread in the human genome and are associated with GI[40]. The formation of NBD requires unwinding of the DNA sequence, as occurs during replication and transcription[18]. NBD comprise A-Phased Repeats (APR), Direct Repeats (DR), G-Quadruplex (GQ), Inverted Repeats (IR), Mirror Repeats (MR), Short Tandem Repeats (STR), Z DNA (Z) and RLoops Forming Sequences (RLFS). We found that all NBD except APR are hotspots (Fig. 1B), especially GQ which are so highly broken so that Hscore tends to infinity ($P = 0$).

The DNA repeats investigated in this study comprise high-copy repeats: MicroSatellite (MS), Low Complexity (LC), simple repeats (SR); and low copy repeats like Self Chain Segments (SCS), which we split into self-aligned (SCS-S) and gapped (SCS-G), Long Terminal Repeats (LTR) and Retro Transposons (RT). While SR, MS, LC are hotspots, the viral origin repeats LTR and RT are not (Fig. 1C). Concordantly, no viral insertion was observed in LMS genomes when using HGT tools (Ref.[41] and data not shown). Furthermore, while SCS-S were hotspots, SCS-G were not.

Genes promoters are hotspots; gene enhancers are hot regions. The promoters of transcriptionally active genes have repeatedly been shown to recurrently harbor double-strand breaks (DSB)[13]. Furthermore, DNase Hyper sensitive (DHS) DNA elements as well as active chromatin marks have been shown to colocalize with DSB[42]. The regulatory DNA elements used in this study comprise CpG islands (CpGi), Cis regulatory Modules (CRM), and DHS of promoter type (DHS_prom), of enhancer type (DHS_enh), of dyadic type (both enhancer and promoter signatures) (DHS_dyadic), and of other types (DHS_rest)[43]. We found that promoter-associated DNA elements (i.e. DHS_prom and CpGi) are hotspots for DNA breakage (Fig. 1D) while DNA elements associated with enhancer activity (DHS_Enh) are hot regions (Fig. 1D). Furthermore, promoter-associated regulatory elements (CpGi, DHS_prom) break more significantly than Enhancers (DHS_enh), which break

more significantly than Dyadic regulatory elements (DHS_dyadic) and the rest of DHS (DHS_rest) (Fig. 1D). Interestingly, CRM, which comprises overlapping regulatory regions (promoters, enhancers, dyadic)[44], is a hotspot and has an intermediate Hscore. This further underlines the robustness of our BP hotspotness magnitude scale. Finally, DHS which are not labeled to have any gene regulatory function (DNS_rest), are not significantly broken more than random (Fig. 1D). Taken together, these results show that DNA breakages are more enriched in promoters than in enhancers and dyadic regions. Furthermore, BP are more significantly enriched in open chromatin structures (DHS) with attributable gene regulatory functions like enhancers and promoters than DHS, with no attributable gene regulatory function, like DHS_res. Therefore, because DHS has an open DNA structure both inside and outside genes, we hypothesize that they do not break merely due to the openness of their chromatin but rather due to their genic context. We thus refined our analysis by splitting each DNA element into two categories: located inside or outside genes.

Regulatory elements are almost exclusively "hot" inside genes and not outside them. To address the impact of the functional genomic context of a DNA element on the BP frequency, we split them into those located inside genes from the transcription start site (TSS) to the transcription end site (TES) (including DNA elements with at least 1 bp overlapping with genes) and those located outside genes, and then computed Hscore for each group in sliding windows. We found that all regulatory elements (Fig. 2) were more significantly broken when they were located inside genes than outside, independently of their structure (sequence type like CpGi vs DNA–protein complexes) or function (promoter type vs enhancer type). CRM were always broken more frequently than by chance when located inside genes, with an Hscore of 67.82 inside genes and less than 3 outside (Fig. 2A, Supplemental Table S3). Although CpGi located outside genes were slightly significantly broken (Hscore = 6.41), this significance was much higher when they were located inside the genes (Hscore = 83.30) (Fig. 2B), with a ratio of Hscores inside/outside genes (RHscore i/o) of 12.99 (Supplemental Table S3). Similarly, all types of DHS regulatory elements were broken more significantly inside genes than outside (Fig. 2C–F). These results suggest that the transcriptional and/or epigenetic context of genes strongly modulates the ability of DNA regulatory elements to harbor DNA BP. Since regulatory elements have an open DNA structure both inside and outside genes when they are active, we rule out the possibility that their propensity to harbor DNA BP is merely due to the openness of their chromatin.

NBD and DNA repeats are hotspots both inside and outside genes. We also split NBD and DNA repeats into those located inside genes and those located outside them (Fig. 3). MR, IR, DR, and STR were significantly broken to a similar extent inside and outside genes (Ratio Hscore i/o 1.02, 1.12, 1.02, 1.13 respectively) (Fig. 3A–D). While RLFS were highly significantly enriched in BP outside genes (Hscore = 66.24), they were even more significantly broken inside them (Fig. 3E) (Hscore = 185.41, RHscore i/o = 2.8) (Supplemental Table S3). GQ were also highly significantly enriched in BP outside genes (Hscore = 180.53) and tended to be more broken inside them (Hscore = 230.17, RHscore i/o = 1.27) (Fig. 3F). Z DNA tended to be more frequently broken outside genes than inside them (RHscore i/o = 0.73) (Fig. 4G). Hscores of DNA repeats located inside and outside genes were almost identical except for SCS-S, which were significantly more broken than random inside genes (Hscore = 8.94) but not outside them (Hscore = 1.18 < 3), while SCS-G were indistinguishable from random (Fig. 4). Together these results suggest that, except for SCS-S, GQ and RLFS, NBD and DNA repeats are insensitive to gene context and may cause GI by a mechanism largely independent from transcription and which is probably replication-dependent. Interestingly, GQ and RLFS share the properties of the three DNA element types: they are sensitive to their genic context as regulatory elements and are breakable when located outside genes, as are DNA repeats and NBD.

Not all LMS present BP distributed as hotspots. The abovementioned results obtained on the whole cohort were global so we still did not know what happens in each patient. To evaluate the DNA breakage mechanisms present in each patient, we computed Hscore for regulatory elements, NBD and DNA repeats in each LMS tumor sample and made a hierarchical clustering of LMS patients based on these Hscores (Fig. 5, see "Materials and methods" section). We found that not all LMS patients had BP distributed as hotspots and that there was a gradient of hotspotness. Interestingly, the level of hotspotness does not seem to correlate with total BP counts (Fig. 5) showing that Hscore is not a measure of tumor mutational burden, but rather a measure of the propensity of tumor DNA BP to be concentrated in given DNA elements. Furthermore, each LMS sample had its specific profile: while some patients had no detectable hotspots, others had BP hotspots mainly in the regulatory regions, and others still had them mainly in NBD and DNA repeats. There were even some patients who had them in both (Fig. 5). Interestingly, metastatic and non-metastatic patients were not evenly distributed over the gradient of hotspotness, the hotspot side of the heatmap having fewer metastatic events than the opposite side (Fig. 5). We therefore address the question of the relation between BP hotspotness and patient prognosis with different approaches in the next sections.

The LMS cohort can be stratified into clinically relevant groups. Because (a) regulatory elements, NBD and DNA repeats have been thoroughly documented to impede transcription and replication in both a transcription-associated[18] and replication-associated[3] manner and to cause GI, and (b) GI is predictive of poor prognosis[45], we hypothesized that there is a link between transcription- and replication-associated DNA-breakage mechanisms and metastatic clinical outcome in LMS. To test this hypothesis, we sought to quantify the overall transcription-associated and replication-associated GI. We used our BP hotspotness magnitude scale and derived genomic indexes for both transcription-dependent and transcription-independent DNA-breakage and genome instability. As DR, STR, MR, IR, Z DNA, SR, MS and LC are BP-enriched irrespective of their position

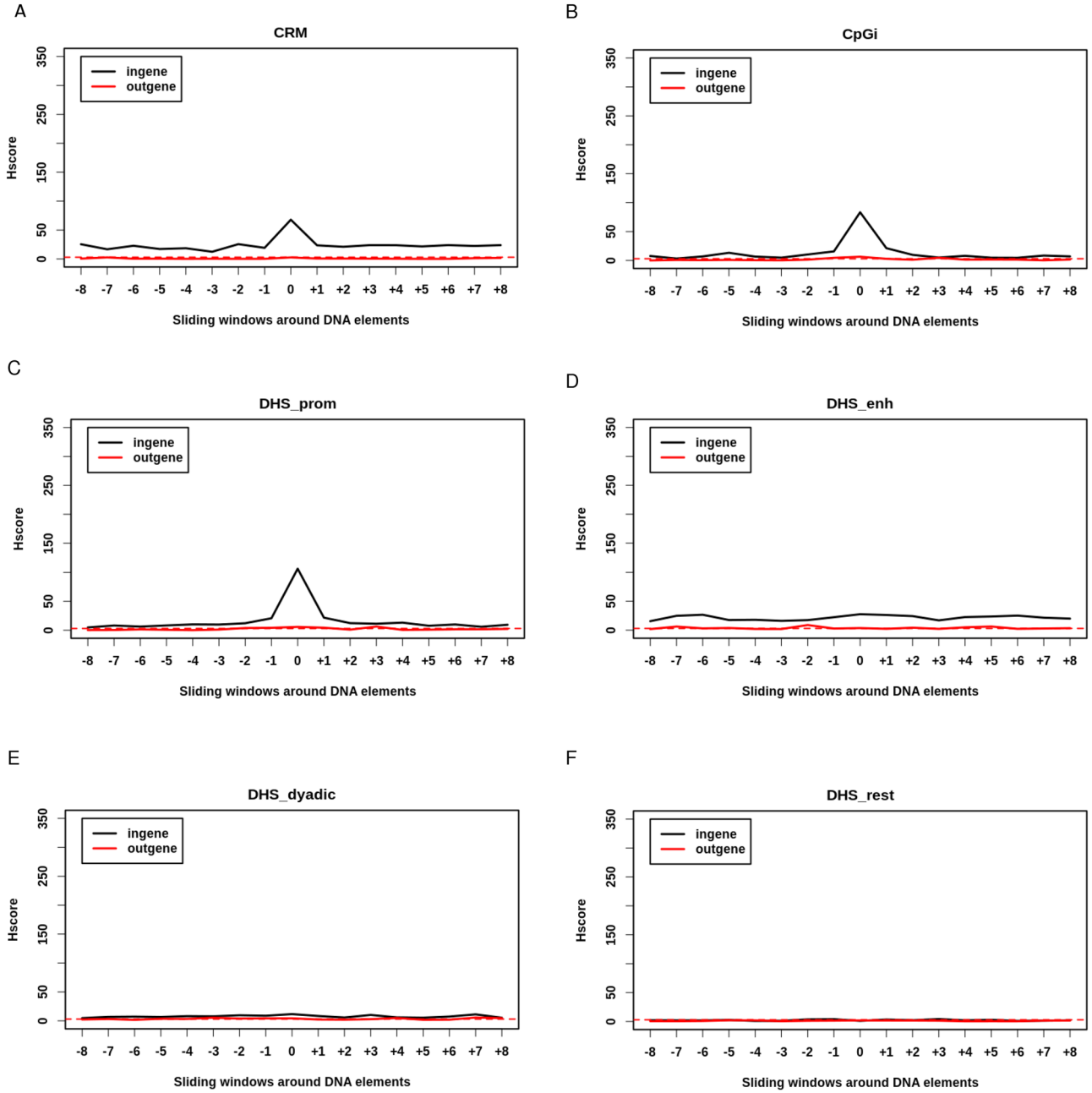

Figure 2. Regulatory elements are exclusively hot inside genes: Hscore for regulatory elements and sliding windows either inside (black) or outside genes (red). (**A**) Cis-Regulatory Modules (CRM). (**B**) CpG islands (CpGi). (**C**) DNase Hyper Sensitive sites (DHS) of type promoter. (**D**) DHS of type enhancer. (**E**) DHS of type dyadic. (**F**) DHS of other types. Horizontal red dashed line corresponds to Hscore threshold of 3 for hotspotness.

inside or outside genes, we considered these elements as transcription-independent and thus as replication-associated chromosomal instability elements (RACINe). Conversely, we considered RLFS, GQ, CpGi, CRM, SCS-S and DHS as transcription-associated chromosomal instability elements (TRACe), because they are more frequently broken than chance inside genes and not outside them. By consolidating TRACe and RACINe into one functional group and computing Hscores (see "Materials and methods" section), we derived a TRAC index (iTRAC) and a RACIN index (iRACIN), respectively. iTRAC and iRACIN allow the quantification of the overall contribution of RACINe and TRACe to DNA breakage and therefore to GI in each LMS patient. To address the relationship between iTRAC, iRACIN and metastatic clinical outcome, we developed a method called Iterative multi-threshold PARTioning (iPART) (see "Materials and methods" section).

First, we used iPART to establish a threshold for iTRAC and iRACIN that splits the LMS cohort into two groups with a maximum difference in MFS (Metastasis-Free Survival). Figure 6A,B left panels show that iTRAC has several thresholds giving P-values less than 0.05 (red horizontal dashed line), while iRACIN has none of

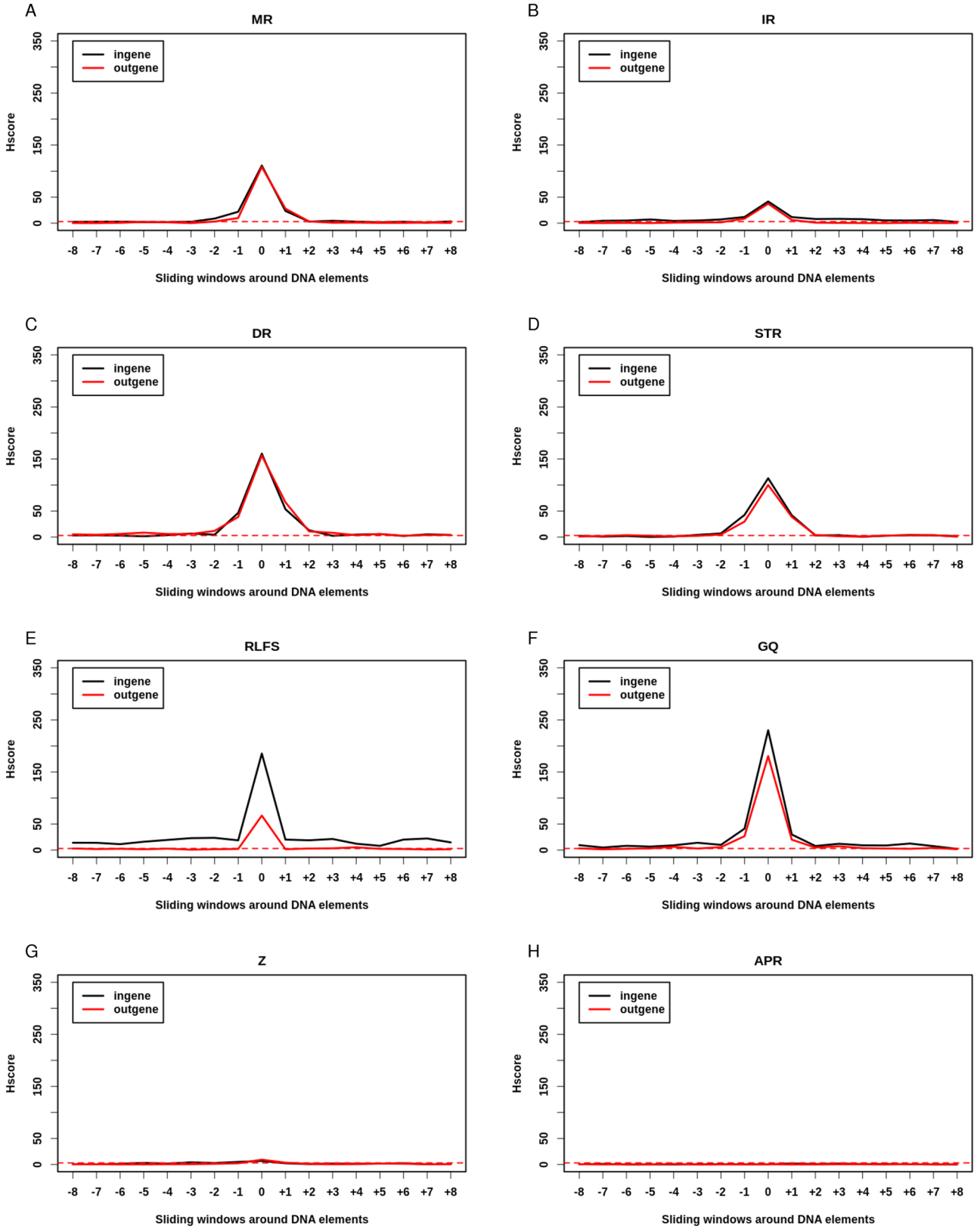

Figure 3. NBD are hotspots both inside and outside genes: Hscore for NBD and sliding windows either inside (black) or outside (red) genes. (**A**) Mirror Repeats (MR). (**B**) Inverted repeats (IR). (**C**) Direct Repeats (DR). (**D**) Short Tandem Repeats (STR). (**E**) R-loops forming sequences (RLFS). (**F**) G-quadruplex (GQ). (**G**) Z DNA (Z). (**H**) A-phased Repeats (APR). Horizontal red dashed line corresponds to Hscore threshold of 3 for hotspotness.

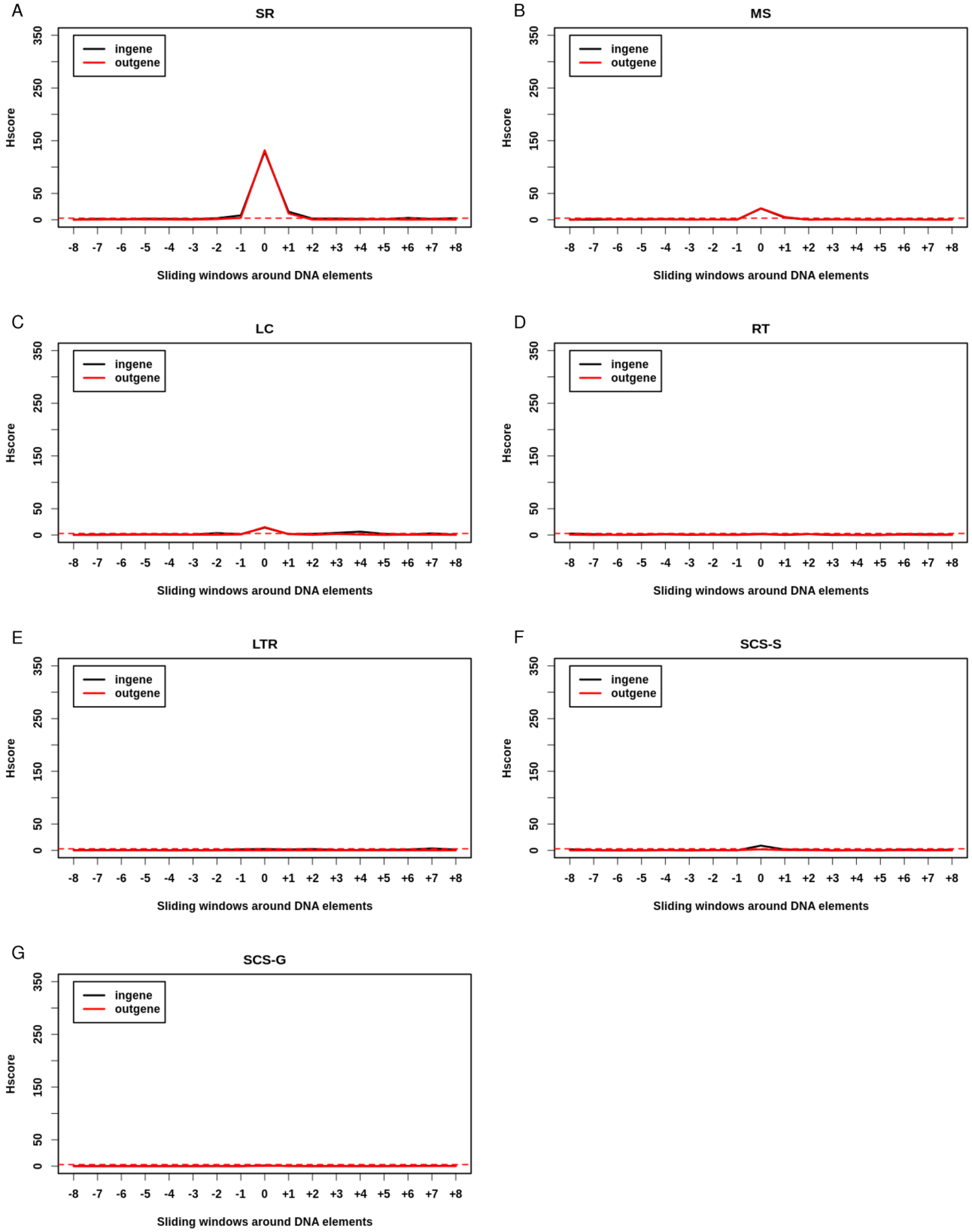

Figure 4. DNA repeats hotspotness relative to genes is dependent upon their type. High-copy DNA repeats: (**A–C**). (**A**) Simple Repeats (SR). (**B**) MiscoSatellites (MS). (**C**) Low Copy repeats (LCR). Viral orgin DNA repeats: (**D**) Retro-Transposons (RT). (**E**) Long Terminal Repeats (LTR). Low copy repeats: Selfchains segments (SCS) (**F,G**). (**F**) Selfchains segments of type self aligned (SCS-S). (**G**) Selfchains segments of type Gapped (SCS-G). Horizontal red dashed line corresponds to Hscore threshold of 3 for hotspotness.

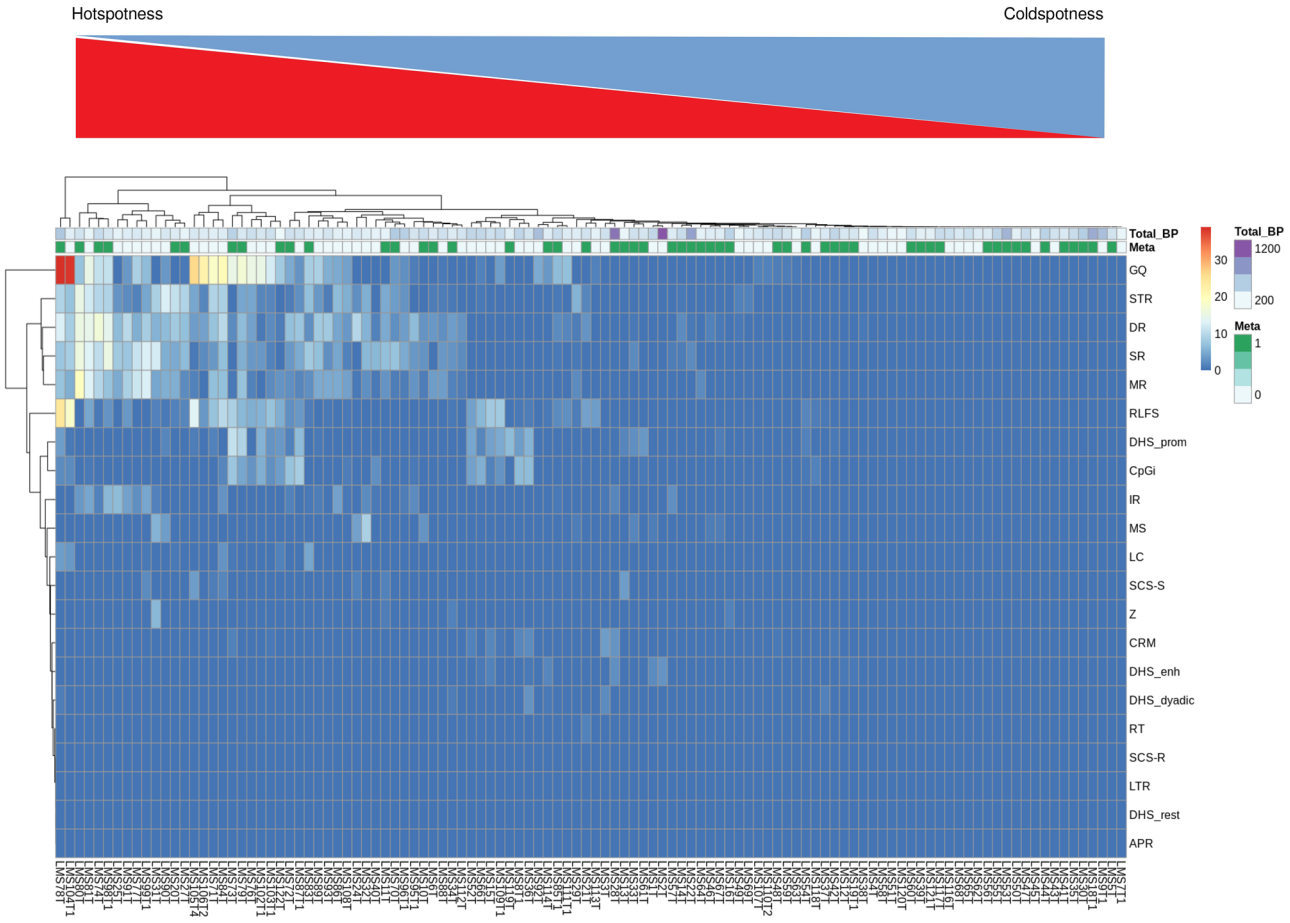

Figure 5. Not all LMS present BP distributed as hotspots. Unsupervised hierarchical clustering of LMS patients and DNA elements using Hscores. Used Euclidian distance on Hscores as a distance measure and the complete method as clustering method. The subtrees in the resulting dendrogram are sorted based on the average distance of subtrees at every merging point. Heatmap was generated using pheatmap_1.0.12 package of R software (http://www.r-project.org/index.html).

them. We also noted the presence of several local minima suggesting that there could be more than one threshold to split the LMS cohort into groups of different MFS. To test this idea, we applied iPART using all combinations of two thresholds from all local minima below 0.10 for iTRAC and 0.3 for iRACIN (blue horizontal dashed line) (materials and methods). Accordingly, both iTRAC and iRACIN significantly stratified the LMS cohort ($P = 3.08 \times 10^{-5}$ and $P = 4.13 \times 10^{-5}$, respectively) into three groups with distinct outcomes (Fig. 6A,B right panels). Strikingly, it was the groups of LMS with medium iTRAC or medium iRACIN which had the most unfavorable outcome (Fig. 6A,B right panels).

We also applied iPART on TBPc, nBPSVintra, nBPSVinter. Although TBPc, nBPSVintra and nBPSVinter were slightly significant (Supplemental Fig. S2), they were not comparable to the level of significance we obtained by iTRAC and iRACIN. Thus, iTRAC and iRACIN have far more added value in the stratification of metastatic risk in LMS than mere BP counts, further stressing the relevance our approach.

Mixed transcription and replication-associated genomic instability classifier (MAGIC). To translate these results into a clinically relevant stratification tool, we sought to integrate both iTRAC and iRACIN into one classifier called the Mixed transcription- and replication-associated genomic instability classifier (MAGIC). Given that both indexes have high and comparable statistical significance in stratifying LMS, we considered at high risk (MAGIC High-risk) any patient classified as medium level by either iTRAC or iRACIN, and the rest of patients as low risk (MAGIC Low-risk). MAGIC achieved a very high level of significance in stratifying LMS samples ($P = 8.75 \times 10^{-8}$; Fig. 7A), with a median MFS for the MAGIC High-risk group of 1.8 (CI = [1.52, 3.61]) which was 5 times lower than for MAGIC Low-risk group (10.5, CI = [5.78, NA]). Then Leave-one-out cross-validation procedure (Celisse 2014) validated this data demonstrating that both iTRAC and iRACIN as well as the resulting MAGIC predictions are strongly significant (Supplemental Fig. S4).

MAGIC outperforms histologic FNCLCC and molecular CINSARC gradings. The histological FNCLCC grading system predicting patient evolution is the current standard in sarcomas[34,46]. Complexity INdex in SARComa (CINSARC), which is the best molecular signature of sarcomas, is challenging the histological gold

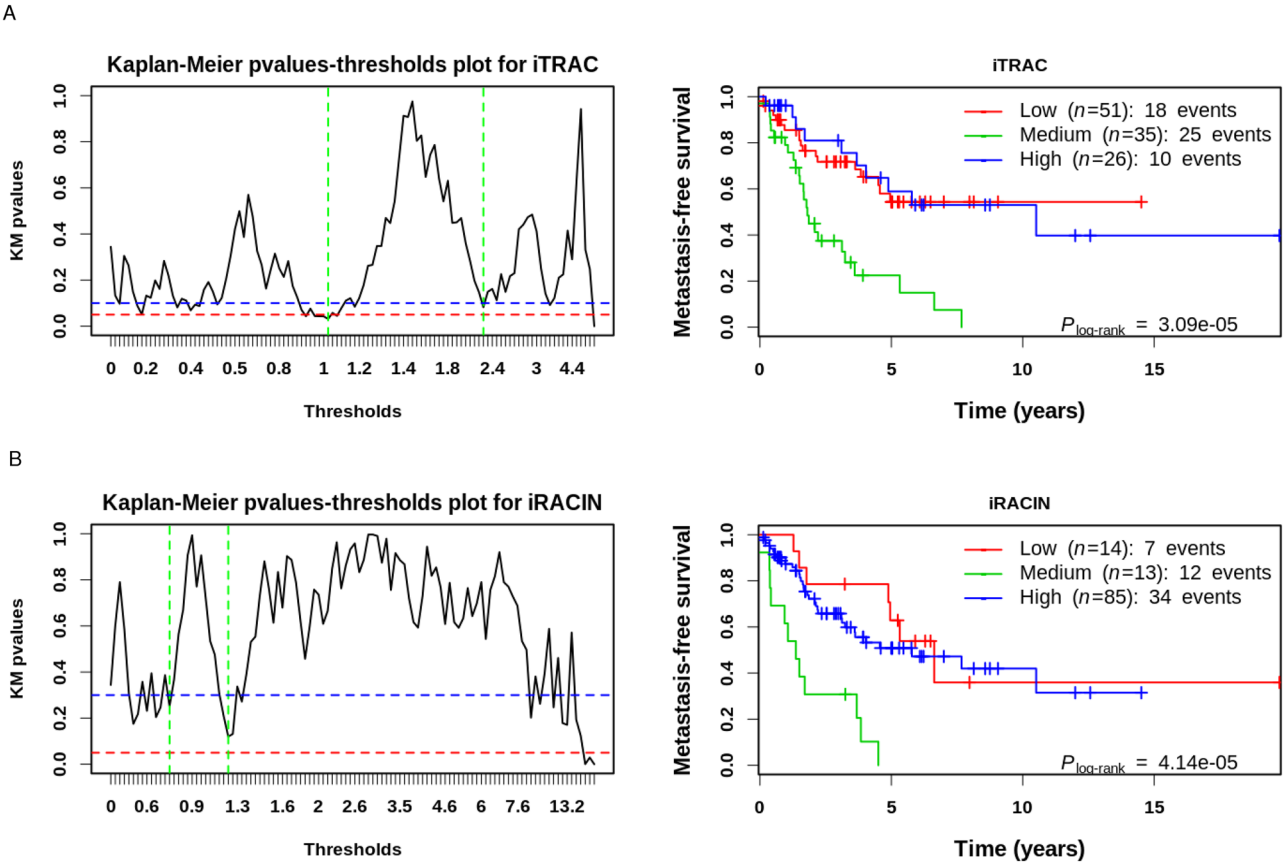

Figure 6. Stratification of the LMS cohort into Low, Medium and High levels of TRAC and RACIN using iPART. (**A**) iTRAC. Left, Kaplan–Meier P-values in function of iTRAC single threshold dividing the LMS cohort into Low and High groups (black solid line). Red dashed horizontal line corresponds to 0.05 arbitrary and commonly accepted significance P-value threshold. Blue dashed horizontal line corresponds to the P-value threshold we considered for P-values that will be included in combinatorial double threshold division of the LMS cohort into Low, Medium and High groups ("Materials and methods" section for more details). Vertical green dashed lines correspond to the best combination of thresholds (tl = 0.99, th = 2.29) spliting the LMS cohort into Low, High, Medium groups. Right, Kaplan–Meier plot for iTRAC stratified LMS cohort into Low, Medium and High based on the best combination of thresholds tl = 0.99 and th = 2.29. (**B**) iRACIN. Both plots correspond to the same procedure as A. tl and th for iRACIN are 0.74 and 1.30 respectively. *tl* threshold Low, *Th* threshold High.

standard and is currently under clinical investigation for stratification[31]. Both approaches individually did not significantly split the LMS cohort into groups with different metastatic evolution (Fig. 7B,C). This finding precludes the introduction of either of these grading systems in multivariate analysis, so we conclude that MAGIC strongly outperforms both the FNCLCC grading system and CINSARC in LMS metastasis risk stratification.

iPART stratifies a Pan-Cancer cohort of twelve cancer types significantly into clinically relevant groups. The intermediary level of GI resulting in poor clinical outcome compared to low and high levels was also reported in a Pan-Cancer study of 12 cancer types (TCGA cohort)[47]. The authors (Andor et al.) used CNV abundance as a measure of GI and found that CNV affecting between 25 and 75% of a tumor's meta-genome was predictive of poor survival. We thus hypothesized that what we observed in LMS might be a general mechanism associated with tumor aggressiveness. To tackle this question in a reasonable time scale with our computational resources, we used the iPART algorithm directly on CNV abundance from the Pan-Cancer study[47] as a proxy for iTRAC/iRACIN. Those authors used the arbitrary and commonly used method of splitting the data into quartiles based on thresholds of 25%, 50%, and 75% to segment the cohort into four groups based on CNV abundance in the tumor meta-genome. Using iPART, we split their cohort into 2, 3, 4, 5, 6, 7 and 8 groups and evaluated the influence of each split on the risk of mortality using the Log-rank test and hazard ratio (Fig. 8, Supplemental Fig. S4). The best data segmentation corresponded to 5 groups split with the log-rank test $P = 6.9 \times 10^{-10}$ and maximal HR 4.7 ($P = 3.42 \times 10^{-9}$) (Supplemental Fig. S4). On the other hand, the data segmentation applied by Andor et al. had a log-rank test $P = 5 \times 10^{-6}$ (Fig. 5b from Ref.[47]) and a maximum HR no exceeding 1.9 with $P < 0.005$ (Fig. 5c from Ref.[47]). Thus, as Andor et al. showed previously, it is the intermediate group (G.3) which is associated with the worst overall survival and presents the highest HR (Fig. 8; HR 4.75, P = 3.42e−9).

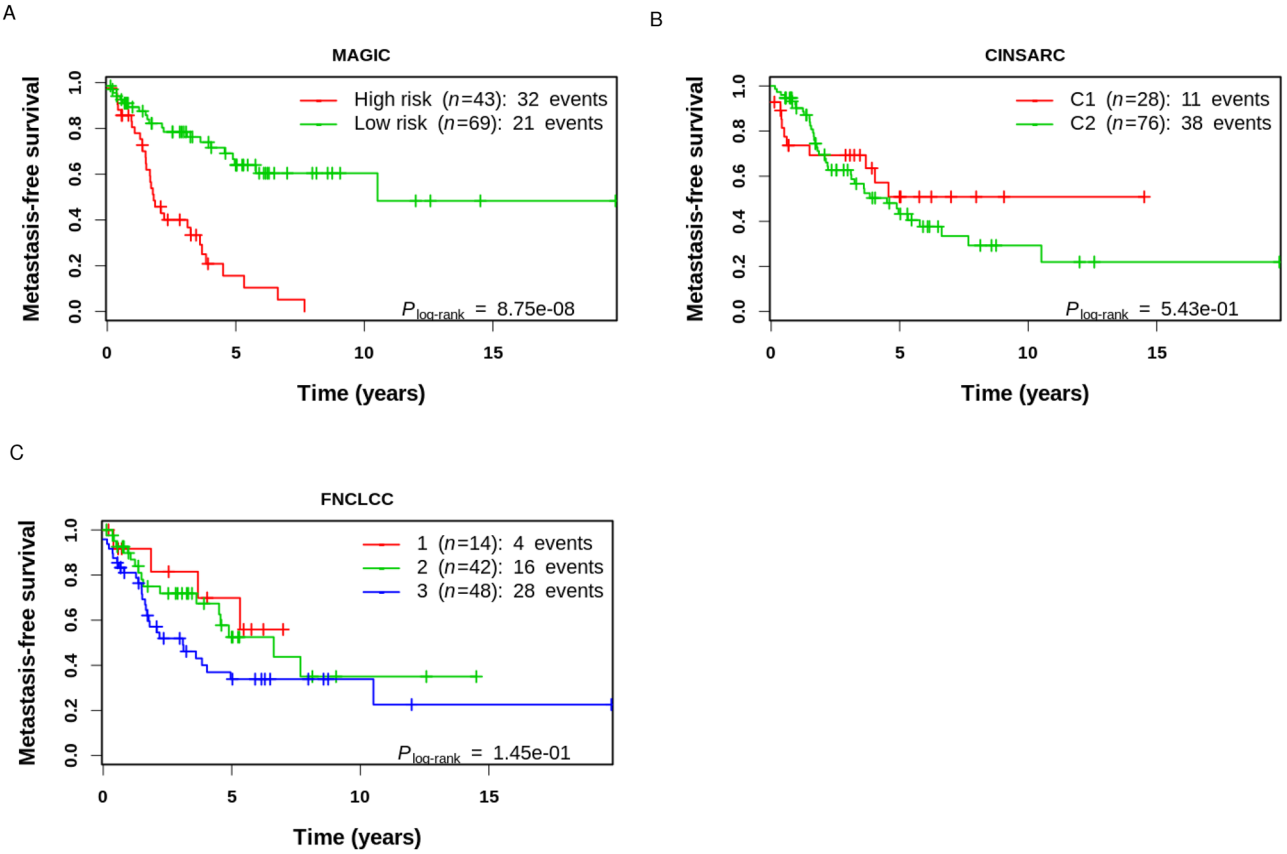

Figure 7. MAGIC outperforms CINSARC and FNCLCC in the stratification of metastatic risk in LMS cohort. Metastasis-free survival curves of LMS cohort stratified by (**A**) MAGIC, (**B**) CINSARC, (**C**) FNCLCC.

iTRAC stratifies chemotherapeutic response in LMS cohort. Chemotherapy remains controversial in LMS since no clinical trial has ever demonstrated its benefit. The criticism is that candidates, i.e. patients with a poor prognosis and responding to chemotherapy, are still not efficiently selected. We therefore sought to address this question in the 112 LMS in our cohort: 18 underwent chemotherapy (14 adjuvant, 1 neoadjuvant and 3 palliative), 18 patients were not annotated (NA) and 76 patients were not treated with chemotherapy. Because the level of GI influences the efficacy of several cancer treatments[23, 26, 27], we sought to quantify the contribution of TRAC and RACIN to the chemotherapeutic response in LMS. We first split the MAGIC risk groups according to their chemotherapeutic treatment status, i.e. Yes (18 pts) or No (76 pts). We found that MAGIC did not significantly stratify the chemotherapeutic response in LMS (Fig. 9A) as these patients, although similar in their metastatic risk, have rather different underlying GI mechanims. We then split each of the Low, Medium, and High groups of iTRAC, iRACIN according to chemotherapeutic treatment status. Interestingly, while patients in the Low risk iTRAC group receiving chemotherapy had a poorer prognosis (HR 4.47, CI [1.65, 12.08], P = 0.0032), no therapeutic benefit was found in the Medium and High risk iTRAC groups (Fig. 9B). In the High risk iTRAC group, there was a tendency for chemotherapy to be beneficial, but there was not enough statistical power (only three events in the chemotherapeutic arm) to test this hypothesis. Further patient inclusion would be needed to test the relevance of the prediction of the response to chemotherapy. Conversely, none of the iRACIN groups was relevant for stratifying chemotherapeutic response, again probably due to the small sample of treated patients (Fig. 9C). Interestingly, the ability of iTRAC to stratifiy chemotherapeutic response in LMS remains significant using overall survival (OS) as clinical endpoint (Supplemental Fig. S5) further stressing the relevance of our approach.

Discussion

This study describes new tools, measures and insights that address the question of the clinical outcome and its relationship with genomic rearrangement. By deciphering the mechanisms of GI, we have produced the iTRAC and iRACIN indexes which account for transcription- and replication-related GI. We also generated the MAGIC classifier and demonstrated its prognostic value for LMS in outperforming both the current gold standard histologic and molecular challenger grading systems. Moreover, our indexes are potentially applicable to the Pan-Cancer cohort, as shown by the significant prognosis we established in twelve different cancer types using iPART and CNV abundance as a proxy for iTRAC/iRACIN.

Application of Hscore on TRACe and RACINe as a holistic approach to measuring GI. Recent studies have used different GI scores and indexes to predict clinical outcome and to define homologous recombi-

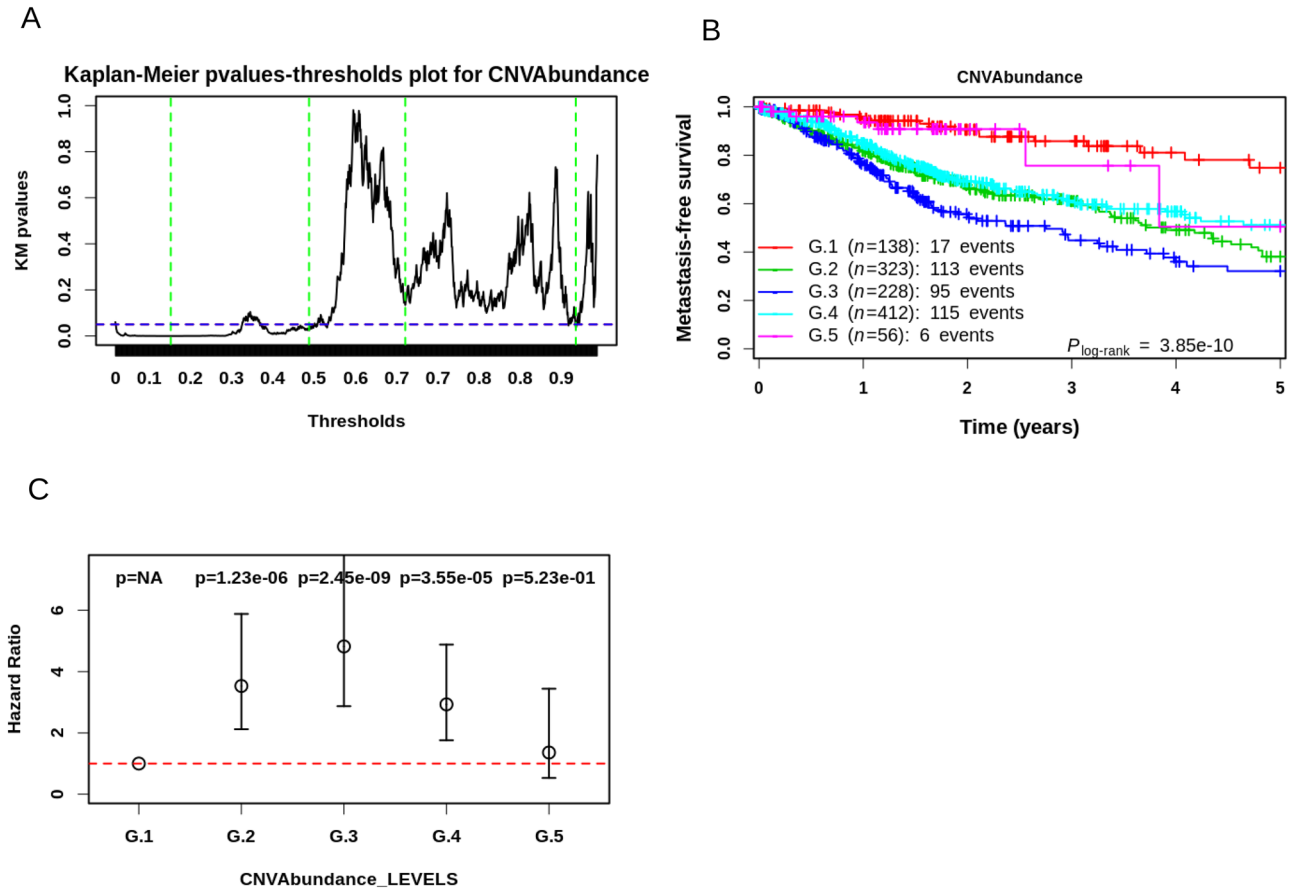

Figure 8. Application of iPART algorithm to the TCGA Pan-Cancer cohort of 12 cancers of Andor et al.[47]. (**A**) Kaplan–Meier P-values in function of CNVAbundance single threshold dividing the TCGA Pan-Cancer cohort into Low and High groups (black solid line). Green vertical lines correspond to the four thresholds giving the best data segmentation (see text and "Materials and methods" section for details) into 5 groups, which correspond from left to right to: 0.14, 0.45, 0.67, 0.86. (**B**) Kaplan–Meier plot showing the stratification of the TCGA cohort into 5 clinically relevant groups. (**C**) Hazard Ratio (HR) with 95% confidence interval of risk of mortality of each group relative the reference group (G.1). P-values of each HR are shown above each condition. Horizontal dashed line corresponds to a Hazard ratio of 1.

nation (HR)-deficient samples[48–54]. However, most of these studies were merely based on the per patient counts of *BRCA1/2* mutations, genome SNPs, loss of heterozygosity (LOH), or specific structural variations (CNV, Telomeric Allelic Imbalance, etc.). These studies assume that an a priori selection of chosen genomic alteration types can recapitulate the dynamics of GI in a tumor that is complex, heterogeneous and subject to selection pressure from the tumor micro-environment. Here we present a new measure of GI that is based on the quantification of all identified DNA BP in a tumor sample, irrespective of their SV type. We introduce the notion of Hscore, which is a BP hotspotness magnitude scale that measures the propensity of a given functional and/or structural genomic DNA element to harbor BP more than expected by chance. We also applied Hscore on TRACe and RACINe as a holistic approach to measure GI based on measuring the steady-state equilibrium between DNA damage and repair. The method quantifies the residual DNA BP remaining after unsuccessful repair, irrespective of SV type, and quantifies the relative contribution of two main contributors to GI: TRAC and RACIN. Hscore is measured on all tumor BPs without any biological-rational selection bias toward SV types. Values are comparable between different DNA elements in a patient and between patients. The higher the Hscore, the more unlikely it is that the observed BPs are due to random events. Therefore, the likelihood is greater that they are due to the structural and/or functional properties of those DNA elements.

MAGIC: towards a new standard for LMS grading? LMS prognostication is still challenging and mandatory to evaluate which therapeutic strategy can benefit which patient. Since some patients have a poor prognosis associated with a transcription stress and others with a replication stress, we produced a simple patient-stratification tool called MAGIC. While neither histologic FNCLCC nor molecular CINSARC grading are suitable predictive methods in an LMS cohort, MAGIC achieved a high level of significance, with the high-risk group having a median MFS = 1.8 years, i.e. fivefold lower than the low-risk group: 10.5 years. MAGIC could be used to spot patients with a high risk of metastasis. Prospective validation of this hypothesis is now mandatory.

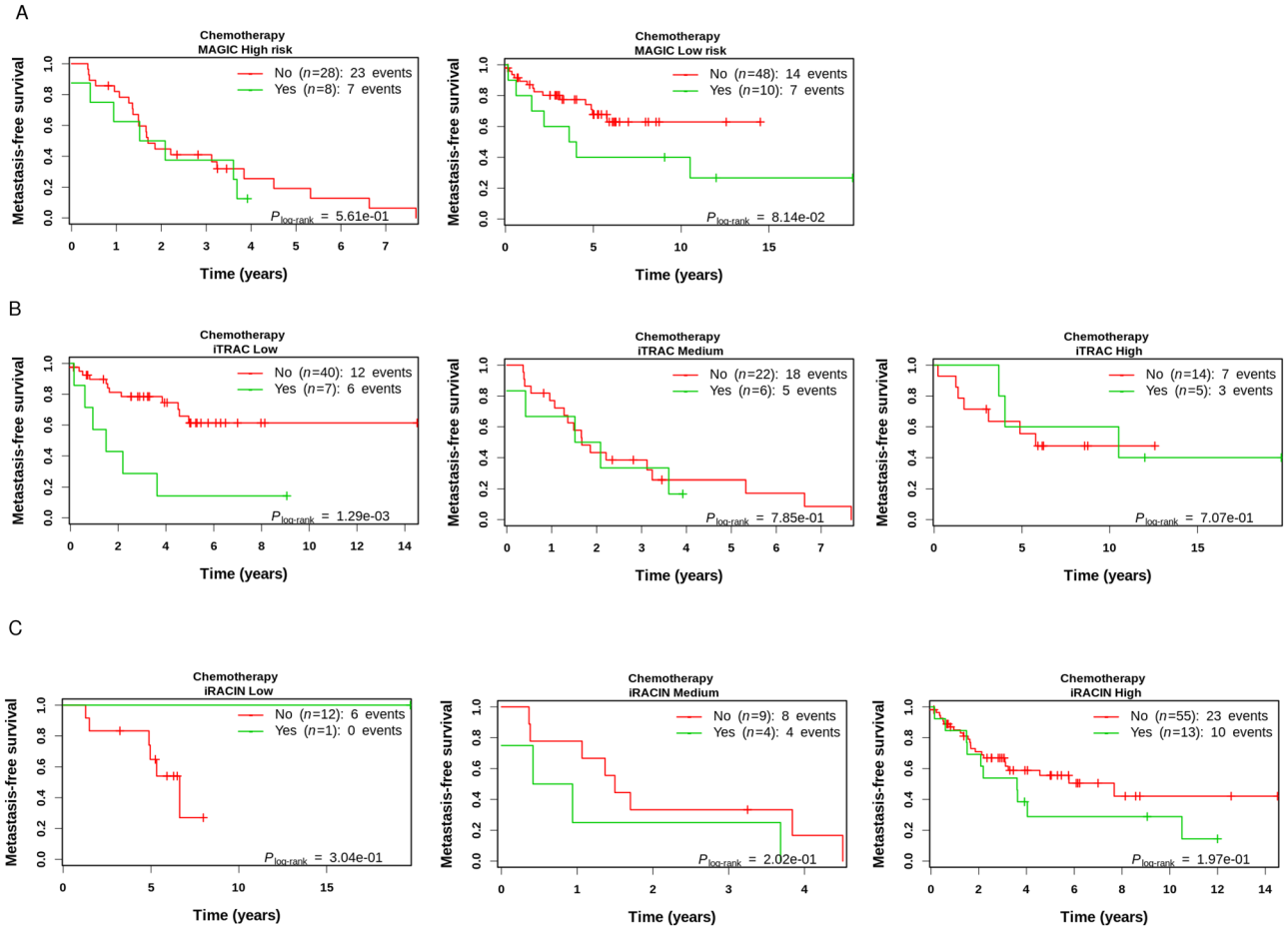

Figure 9. iTRAC stratifies chemotherapeutic response in LMS but not iRACIN. MFS curves in LMS groups of (**A**) MAGIC, (**B**) iTRAC and (**C**) iRACIN stratified by chemotherapeutic treatment.

Prognostic relevance of iTRAC and iRACIN for metastatic risk stratification of other cancers. Unexpectedly, the risk of metastasis was found to follow a lambda (Λ) shape with the increase in iTRAC and iRACIN: both low and high GI levels correspond to a lower metastatic risk, while a medium level indicates a higher metastatic risk. Interestingly, similar results were previously reported in a Pan-Cancer analysis of 12 cancer types[47] concerning the abundance of copy number variations (CNV). We improved Andor et al. results using iPART and we further highlighted the added value of quantifying transcription- and replication-associated GI and iPART algorithm versus a simple SV count. These results strongly suggest that our approach would work on different cancer types and and different GI measures on other highly remodeled cancer genomes.

Predictive relevance of iTRAC and iRACIN for chemotherapeutic response in LMS. The overall contribution of curative and adjuvant cytotoxic chemotherapy to 5-year survival in adults was estimated to be 2.3% in Australia and 2.1% in the USA[55]. Furthermore, it has been estimated that any class of cancer drugs is ineffective in 75% of patients (Personalized Medicine Coalition; The personalized medicine report Opportunity, Challenges, and the Future 2017; http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/The-Personalized-Medicine-Report1.pdf). Thus, predicting which patients are eligible for which treatment and those who are not is the holy grail of precision medicine. Here we show that chemotherapy for patients with a low iTRAC would be detrimental for their MFS and should therefore be prohibited. In addition, chemotherapy is likely to have no clinical benefit for patients with a medium iTRAC so another therapeutic strategy should be used for them. No conclusion can be drawn for patients with a high iTRAC because of the low statistical power of the log-rank test due to the low number of patients with a high iTRAC which underwent chemotherapy. Nevertheless, high-iTRAC patients would probably benefit from chemotherapy and more inclusions are needed to address this question. On the other hand, iRACIN was not predictively relevant for stratifying the chemotherapeutic response. Nevertheless, we expect it to be relevant for stratifying targeted therapies based on targeting replication and replication-associated repair. We do not consider using MAGIC in this setting as it is a combination of groups of patients albeit similar in their metastatic risk occurrence, but who have different GI mechanisms. Therefore, iPART/iTRAC/iRACIN could serve as a toolset to tackle the question of the relation of GI to therapeutic intervention based on the genomic processes undermining genomic integrity.

Targeted therapy as a therapeutic strategy in LMS. Our understanding of GI may serve both for prognosis and to orient the choice of therapeutic agent. In patients with germline BRCA1 or 2 mutations, PARP inhibitors have led to major therapeutic advances in patients with ovarian cancer[56] and to a lesser extent, breast cancer[57] over the past years. Furthermore, it has been proposed that most LMS tumors display hallmarks of "BRCAness", including alterations in homologous recombination DNA repair genes, enrichment of specific mutational signatures, and cultured LMS cells sensitive to olaparib and cisplatin[58]. Thus, PARP1 inhibitors are a candidate treatment in LMS. The use of iTRAC/iRACIN in clinical settings would allow the selection of LMS patients who potentially would have a significant therapeutic response to PARP1 inhibitors.

Future directions. Given the far-reaching consequences of GI for treatment success, therapeutic choices and clinical care, an accurate measure of GI and its dynamics is paramount in precision medicine. The development of robust biomarkers enabling GI dynamics to be captured is crucial if we are to leverage the potential of GI for patient stratification purposes and for exploiting this feature for making therapeutic choices. Deriving minimally invasive approaches that enable clinicians to assess whether GI is at play within a given tumor sample might be crucial for its efficient exploitation in clinical settings. Recent advances in whole genome/whole exome sequencing of formalin-fixed, paraffin-embedded (FFPE) tissues allows the application of this approach in routine clinical settings. Finally, prospective findings now need to be validated by measuring iTRAC/iRACIN in circulating tumor DNA (ctDNA) or in circulating tumor cells (CTC) and in subsequent clinical trials.

Materials and methods

Samples. LMS Samples (112) used in this study were collected as part of the ICGC program (International Cancer Genome Consortium; https://icgc.org/) with patient consent. Samples were frozen tissues provided by pathologists and a blood sample for each included patient provided by medical oncologists. All cases were systematically reviewed by expert pathologists of the French Sarcoma Group according to the World Health Organization recommendations[59].

DNA extraction. Genomic DNA from frozen samples was isolated using a standard phenol–chloroform extraction protocol[60]. DNA was quantified using a Nanodrop 1000 spectrophotometer according to manufacturer's recommendations (Thermo Scientific, Waltham, MA, USA). Blood material from included patients was also available. Genomic DNA from blood samples was extracted using customized automated purification of DNA from compromised blood samples on the Autopure LS protocol according to the manufacturer's recommendations (Qiagen, Hilden, Germany), with increased centrifugation of 10 min for DNA precipitation and DNA wash.

Whole genome sequencing and analysis. To construct short-insert paired-end libraries, a no-PCR protocol was used with the TruSeq™DNA Sample Preparation Kit v2 (Illumina Inc., San Diego, CA, USA) and the KAPA Library Preparation kit (Kapa Biosystems, Basel, Switzerland). Briefly, 2 µg of genomic DNA were sheared on a Covaris™ E220, size-selected and concentrated using AMPure XP beads (Agencourt, Beckman Coulter, Brea, CA, USA) in order to reach a fragment size of 220–480 bp. Fragmented DNA was end-repaired, adenylated and ligated to Illumina-specific indexed paired-end adapters.

DNA sequencing was performed in paired-end mode in lanes of HiSeq2000 Flowcell v3 (2 × 100 bp) or Flowcell v4 (2 × 125 bp) or v4 (2 × 125 bp) or in sequencing lanes of NovaSeq 6000 Flowcell S4 (2 × 150 bp) (Illumina Inc., San Diego, CA, USA) to analyze tumor or matched normal blood samples (from the same patient) and to reach a minimal yield of 145 or 85 Gb, respectively. Two tumor samples (LMS2T and LMS5T) were sequenced in 20 lanes of HiSeq2000 Flowcell v3 to reach a minimal yield of 560 Gb. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ sequence files by CASAVA (Illumina Inc., San Diego, CA, USA).

DNA reads were trimmed of the 5′ and 3′ low-quality bases (PHRED cut-off 20, maximum trimmed size: 30 nucleotides (nt)) and sequencing adapters were removed with Sickle2 (Joshi NA, Fass JN. 2011 available at https://github.com/najoshi/sickle). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle)and SeqPrep3 (J. St. John, SeqPrep. (2011) Available at https://github.com/jstjohn/SeqPrep), respectively. Then, DNA curated sequences were aligned using bwa v-0.7.15[61] with default parameters on the Human Genome version hg38[62] (http://genome.ucsc.edu/ or https://www.ncbi.nlm.nih.gov/grc/human). Thus, aligned reads were filtered out if their alignment score was less than 20 or if they were duplicated PCR reads, with SAMtools v1.3.1[63] and PicardTools v2.18.2 ("Picard Toolkit." 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute), respectively.

Random breakage model, Hscore, readable genome size. The readable genome is represented as a single interval of length L in base pairs (bp). The uniform probability Pu of any genomic position to carry a BP is the total number of BP (n) divided by L: Pu = n/L. For a given genomic interval of size (Li) and number of BP (ni), its probability to harbor ni BP under the random breakage model (RBM) is computed by the probability mass function of binomial distribution as: $f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$; where x = ni, n = Li, p = Pu. The probability of observing more than ni BP under RBM is defined as $P(X > ni) = 1 - P(X \leq ni) = 1 - \sum_{i=0}^{n} f(i)$; where n = ni, and X is the random variable accounting for the number of observed BP. For the bed file of each DNA element, overlapping intervals were merged (bedtools) and then BP number (ni) and interval size (Li) were

computed as follows: ni was computed as the sum of BP in all the merged intervals and Li was computed as the sum of all merged interval sizes. Hscore is then computed as the − log10 of P(X > ni) under RBM.

We define readable genome size as the total ungapped genome length as defined at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/. It is equal to 2,948,611,470 bp.

Breakpoint identification and detection of structural variants. Structural variants (SV) were detected from paired tumor/normal whole genome high-quality sequencing data. Paired-end reads were aligned using Bowtie v2.2.1.0[64], which is a sensitive local option allowing soft-clipped sequences. The algorithm has three main steps: (i) identification of potential breakpoints, (ii) characterization of the second side of the breakpoints, and (iii) selection of high-confidence breakpoints. All parameters were set to analyze 60× tumor and 30× normal sequencing depth. Very conservative filters were used to minimize false positive detection.

(i) *Identification* at this step, reads with at least one soft-clipped end were analyzed as singletons. A position was considered as a potential breakpoint if it was covered by at least 4 soft-clipped reads, 5 soft-clipped bases (with at least two occurrences of two different bases), and if they represented more than 5% of the total amount of reads at this position in the tumor sample. We selected potential somatic events by discarding positions covered by at least one read and one base in a surrounding 5-nucleotide window in the normal sample. We refer to them as the "first side" of the breakpoint.

(ii) *Characterization* to determine the genomic positions of the soft-clipped sequence from selected reads, we used the UCSC blat server[65]. If no match was returned, the reverse complement sequence was pulled to test. If there was still no match, the BAM file was investigated for some soft-clip somatic position around the discordant or oversized-insert read mate (hereafter named abnormal) location from the first side of the breakpoint. Because of the small size of the soft-clipped sequence, multiple matches can be found. We used soft-clipped abnormal read mates to select matches with the most coherent chromosomic locations. We refer to them as the "second side" of the breakpoint.

(iii) *Selection* Positions detected from both the first and second sides (in a 5-nucleotide window) were defined as the common pool. We considered as artifacts (due to repeat regions for instance) any couples of positions covered with reads and associated soft-clipped sequences separated by fewer than 15 nucleotides and discarded them. We classified the breakpoints in three groups: high-confidence breakpoints, breakpoints needing investigation, and unique position breakpoints. If a breakpoint was covered by reads and associated soft-clipped sequences having both positions belonging to the common pool, it was classified in the first group. If a breakpoint was covered by reads and associated soft-clipped sequences having only one of the positions belonging to the common pool, it was classified in the second group. Then the missing position was searched among the filtered positions. If it was present in the normal sample, the position was discarded and the breakpoint was completed otherwise. Finally, the third group corresponds to breakpoints with both sides outside the common pool and considered as unique: these were discarded. The sides of breakpoints were sorted according to their chromosomic positions to avoid duplicates.

Data collection. The following DNA elements were considered in the present analysis: DNA repeats comprising MicroSatellite (MS), Simple Repeats (SR), Low Complexity (LC), Self-Chain segments (SCS) which were classified into self-aligned inverted chains SCS (SCS-S) and gapped SCS (SCS-G), Long Terminal Repeats (LTR), and Retrotransposons (RT); Non-B DNA comprising A-Phased Repeats (APR), Direct Repeats (DR), G-quadruplex (GQ), Inverted Repeats (IR), Mirror Repeats (MR), Short Tandem Repeats (STR), Z-DNA (Z) and R-Loops Forming Sequences (RLFS); and Regulatory DNA elements comprising CpG islands (CpGi), *cis*-regulatory modules (CRM), DNase I hypersensitive site (DHS) of promoter type (DHS_prom), DHS of enhancer type (DHS_enh), DHS of dyadic type (both enhancer and promoter signatures) (DHS_dyadic), and DHS of other types (DHS_rest).

Data for CpG islands, microsatellites, simple repeats, low complexity, retrotransposons, long terminal repeats, self-chains and sequencing gaps were obtained from the UCSC Genome Browser website (http://genome.ucsc.edu/; genome assembly hg38). All Non-B DNA except RLFS were generated using the non-B DNA research tool from the non-B DNA database[66]. RLFS data were generated using QmRLFS-finder[67]. CRM data were obtained from Remap2018[44] and data were downloaded from (http://pedagogix-tagc.univ-mrs.fr/remap/). DNase I-accessible regulatory regions (with − log10(p) ≥ 2) were downloaded from the roadmap epigenomics project at https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2_release/ and coordinates were converted from genome assembly hg19 to hg38 using Liftover, the UCSC coordinates conversion tool[68].

TRAC index/RACIN index. The DNA elements were sorted in separate indexes depending on whether their enrichment in BP was dependent or not on their presence inside or outside the genes (see "Ingene/outgene split" section). DNA elements enriched in BP independently of their position inside or outside the genes were sorted as Replication-Associated Chromosomal INstability elements (RACINe). DNA elements enriched in BP according to their position inside the genes were sorted as TRanscription-Associated Chromosomal instability elements (TRACe). For each TRACe and RACINe element, we pooled all bed files of the corresponding DNA elements into one file and sorted them according to interval positions (bedtools) and merged (bedtools) all overlapping intervals to obtain the corresponding index iTRAC and iRACIN, each as a single bed file. BP counts and interval sizes were computed for each index. These were then used to compute Hscores under RBM.

Sliding windows. For each DNA element (sliding window 0 in Figs. 3, 4, 5), each genomic feature was shifted (bedtools) by 100% its length on the positive (+) DNA strand (sliding window + 1 in Figs. 3, 4, 5) and

on the negative (−) DNA strand (sliding window − 1 in Figs. 3, 4, 5) and Hscore was computed. This procedure was repeated by shifting each feature ± 2 × 100% (sliding window + 2, sliding window − 2), ± 3 × 100%, until ± 8 × 100%.

Heatmap. Hscore was computed with Holm's adjusted P-values procedure. Two patients (LMS78T and LMS131T) had maximal Hscores of 170.53 and 50.16 while all the other patients had a maximal Hscore less than 39. Therefore, the heatmap was unexploitable as only these two maximal data points were visible. For the sake of clarity, Hscores of patients LMS78T and LMS131T were normalized by dividing them by 170.53 and multiplying them by 39.

Ingene/outgene split. A DNA element was considered as inside a gene if overlapped by at least 1 bp the gene interval delimited by its Transcription Start Site (TSS) and Transcription End Site (TES). Gene coordinates were taken from curated RefSeq entries from the UCSC table browser page (https://genome.ucsc.edu/cgi-bin/hgTables; group = genes and genes prediction; track = NCBI RefSeq; table = RefSeq Curated). Only genes that had expression data in these tumors were considered (for list of genes, see Supplemental Table S4).

SCS-S/SCS-G. Self-chains (SC) were prepared as in Ref.[69] except that we split SC segments (SCS) into those are self-aligned (SCS-S) and those are gapped (SCS-G) that is having spacing intervals separting each pair of SC. SCS are defined as the segment of any paired SCs in the same chromosome and their spacing gap. The paired SCs located in different chromosomes and those in the same chromosome but having long spacing intervals (SCS size 30 kb) were filtered out to account only for local interactions. In addition, any SCS-S/SCS-G overlapping with the human genome gaps and segmental duplications was further filtered out.

Statistical analysis. All statistical tests and the heatmap were carried out using R (R Core Team (2020); R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria;http://www.r-project.org/index.html).

Leave-one-out cross-validation. Each time one patient was removed from the cohort of patients, thresholds of iTRAC and iRACIN was computed and used to classify the removed patient into Low, Medium, or High.

iPART. We consider iPART (Iterative multi-thresholds PARTitioning) to be an unsupervised decision tree (UDT). It is a method that combines the properties and objectives of both unsupervised clustering and decision trees (DT). Hence, iPART looks for thresholds that maximize the differences in groups instead of computing pairwise distances and constructing hierarchical clusters. It resembles DT and regression trees (RT) by using thresholds to split groups. It differs from DT in that it is unsupervised. It also differs from RT in that it does not attempt to predict quantitative variables. The fundamental difference with both RT and DT is its ability to use binary (splitting data into two groups) and ternary (splitting data into three groups) modes, i.e. a crucial feature in our method. It also differs from DT and RT in using the Kaplan–Meier (KM) estimate instead of the GINI purity index or information gain index and sum of squared residuals for DT and RT, respectively. It also resembles unsupervised machine learning like hierarchical clustering and k-means by aiming to find natural patterns/groups in data. On the other hand, it differs from them in that it does not compute pairwise distances or try to construct groups by minimizing their intra-group variance. Instead, it iterates all possible thresholds, find thresholds that maximize the difference between the split groups in terms of the speed at which metastatic events occur in the two groups by minimizing the P-value of the KM test. Briefly, it establishes natural frontiers that maximize the differences between groups instead of constructing groups that minimize intra-group and overall variance.

Ethics declaration. The ethics of this study was validated by French committee for protection of persons (CPP).

Data availability

The code to repreoduce the results presented in this paper is available at CODE OCEAN: https://codeocean.com/capsule/9703439/tree.

Received: 22 June 2021; Accepted: 8 November 2021
Published online: 06 December 2021

References

1. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability—An evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220–228 (2010).
2. Tubbs, A. & Nussenzweig, A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **168**, 644–656 (2017).
3. Gaillard, H., García-Muse, T. & Aguilera, A. Replication stress and cancer. *Nat. Rev. Cancer* **15**, 276–289 (2015).
4. Macheret, M. & Halazonetis, T. D. DNA Replication stress as a hallmark of cancer. *Annu. Rev. Pathol. Mech. Dis.* **10**, 425–448 (2015).
5. Técher, H., Koundrioukoff, S., Nicolas, A. & Debatisse, M. The impact of replication stress on replication dynamics and DNA damage in vertebrate cells. *Nat. Rev. Genet.* **18**, 535–550 (2017).
6. Debatisse, M., Le Tallec, B., Letessier, A., Dutrillaux, B. & Brison, O. Common fragile sites: Mechanisms of instability revisited. *Trends Genet.* **28**, 22–32 (2012).

7. Le Tallec, B. *et al.* Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Rep.* **4**, 420–428 (2013).
8. Blin, M. *et al.* Transcription-dependent regulation of replication dynamics modulates genome stability. *Nat. Struct. Mol. Biol.* **26**, 58–66 (2019).
9. Aguilera, A. The connection between transcription and genomic instability. *EMBO J.* **21**, 195–201 (2002).
10. Jinks-Robertson, S. & Bhagwat, A. S. Transcription-associated mutagenesis. *Annu. Rev. Genet.* **48**, 341–359 (2014).
11. Kim, N. & Jinks-Robertson, S. Transcription as a source of genome instability. *Nat. Rev. Genet.* **13**, 204–214 (2012).
12. Gaillard, H., Herrera-Moyano, E. & Aguilera, A. Transcription-associated genome instability. *Chem. Rev.* **113**, 8638–8661 (2013).
13. Marnef, A., Cohen, S. & Legube, G. Transcription-coupled DNA double-strand break repair: Active genes need special care. *J. Mol. Biol.* **429**, 1277–1288 (2017).
14. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44**, 966–977 (2011).
15. Wilson, T. E. *et al.* Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.* **25**, 189–200 (2015).
16. Pentzold, C. *et al.* FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells. *Nucleic Acids Res.* **46**, 1280–1294 (2018).
17. Madireddy, A. *et al.* FANCD2 facilitates replication through common fragile sites. *Mol. Cell* **64**, 388–404 (2016).
18. Gaillard, H. & Aguilera, A. Transcription as a threat to genome integrity. *Annu. Rev. Biochem.* **85**, 291–317 (2016).
19. Azvolinsky, A., Giresi, P. G., Lieb, J. D. & Zakian, V. A. Highly transcribed RNA polymerase II genes are impediments to replication fork progression in *Saccharomyces cerevisiae*. *Mol. Cell* **34**, 722–734 (2009).
20. French, S. Consequences of replication fork movement through transcription units in vivo. *Science* **258**, 1362–1365 (1992).
21. Deshpande, A. M. & Newlon, C. S. DNA replication fork pause sites dependent on transcription. *Science* **272**, 1030–1033 (1996).
22. Cahill, D. P., Kinzler, K. W., Vogelstein, B. & Lengauer, C. Genetic instability and darwinian selection in tumours. *Trends Genet.* **15**, M57–M60 (1999).
23. Bakhoum, S. F. *et al.* Numerical chromosomal instability mediates susceptibility to radiation treatment. *Nat. Commun.* **6**, 5990 (2015).
24. Lee, H.-S. *et al.* Effects of anticancer drugs on chromosome instability and new clinical implications for tumor-suppressing therapies. *Can. Res.* **76**, 902–911 (2016).
25. Kim, J.-H. *et al.* Development of a novel HAC-based "gain of signal" quantitative assay for measuring chromosome instability (CIN) in cancer cells. *Oncotarget* **7**, 14841–14856 (2016).
26. Janssen, A., Kops, G. J. P. L. & Medema, R. H. Elevating the frequency of chromosome mis-segregation as a strategy to kill tumor cells. *PNAS* **106**, 19108–19113 (2009).
27. Zasadil, L. M. *et al.* Cytotoxicity of paclitaxel in breast cancer is due to chromosome missegregation on multipolar spindles. *Sci. Transl. Med.* **6**, 229 (2014).
28. Birkbak, N. J. *et al.* Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Can. Res.* **71**, 3447–3452 (2011).
29. Roylance, R. *et al.* Relationship of extreme chromosomal instability with long-term survival in a retrospective analysis of primary breast cancer. *Cancer Epidemiol. Biomark. Prev.* **20**, 2183–2194 (2011).
30. Blay, J.-Y. *et al.* Improved survival using specialized multidisciplinary board in sarcoma patients. *Ann. Oncol.* **28**, 2852–2859 (2017).
31. Chibon, F. *et al.* Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* **16**, 781–787 (2010).
32. Derré, J. *et al.* Leiomyosarcomas and most malignant fibrous histiocytomas share very similar comparative genomic hybridization imbalances: An analysis of a series of 27 leiomyosarcomas. *Lab. Investig.* **81**, 211–215 (2001).
33. Pérot, G. *et al.* Constant p53 pathway inactivation in a large series of soft tissue sarcomas with complex genetics. *Am. J. Pathol.* **177**, 2080–2090 (2010).
34. Coindre, J. M. *et al.* Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: A study of 1240 patients from the French Federation of Cancer Centers Sarcoma Group. *Cancer* **91**, 1914–1926 (2001).
35. Spencer, D. H. *et al.* Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *J. Mol. Diagn.* **15**, 623–633 (2013).
36. Shabani Azim, F., Houri, H., Ghalavand, Z. & Nikmanesh, B. Next generation sequencing in clinical oncology: Applications, challenges and promises: A review article. *Iran. J. Public Health* **47**, 1453–1457 (2018).
37. Andersson, C., Fagman, H., Hansson, M. & Enlund, F. Profiling of potential driver mutations in sarcomas by targeted next generation sequencing. *Cancer Genet.* **209**, 154–160 (2016).
38. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, 8399 (2017).
39. Limpert, E., Stahel, W. A. & Abbt, M. Log-normal distributions across the sciences: Keys and clues. *Bioscience* **51**, 341 (2001).
40. Wang, G. & Vasquez, K. M. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair* **19**, 143–151 (2014).
41. Nguyen, M., Ekstrom, A., Li, X. & Yin, Y. HGT-Finder: A new tool for horizontal gene transfer finding and application to Aspergillus genomes. *Toxins* **7**, 4035–4053 (2015).
42. Mourad, R., Ginalski, K., Legube, G. & Cuvier, O. Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome Biol.* **19**, 34 (2018).
43. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
44. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).
45. Ahmad, S. S., Ahmed, K. & Venkitaraman, A. R. Science in focus: Genomic instability and its implications for clinical cancer Care. *Clin. Oncol.* **30**, 751–755 (2018).
46. Guillou, L. *et al.* Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. *JCO* **15**, 350–362 (1997).
47. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
48. Zhang, S., Yuan, Y. & Hao, D. A genomic instability score in discriminating nonequivalent outcomes of BRCA1/2 mutations and in predicting outcomes of ovarian cancer treated with platinum-based chemotherapy. *PLoS ONE* **9**, e113169 (2014).
49. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).
50. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
51. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Can. Res.* **72**, 5454–5462 (2012).
52. Stefansson, O. A. *et al.* Genomic profiling of breast tumours in relation to BRCAabnormalities and phenotypes. *Breast Cancer Res.* **11**, R47 (2009).

53. Baumbusch, L. O. *et al.* High levels of genomic aberrations in serous ovarian cancers are associated with better survival. *PLoS ONE* **8**, e54356 (2013).
54. Mirza, M. R. *et al.* Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *N. Engl. J. Med.* **375**, 2154–2164 (2016).
55. Morgan, G., Ward, R. & Barton, M. The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies. *Clin. Oncol.* **16**, 549–560 (2004).
56. Moore, K. *et al.* Maintenance olaparib in patients with newly diagnosed advanced ovarian cancer. *N. Engl. J. Med.* **379**, 2495–2505 (2018).
57. Litton, J. K. *et al.* Talazoparib in patients with advanced breast cancer and a germline BRCA mutation. *N. Engl. J. Med.* **379**, 753–763 (2018).
58. Chudasama, P. *et al.* Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nat. Commun.* **9**, 144 (2018).
59. Fletcher, C. D. M. (ed.) *Pathology and Genetics of Tumours of Soft Tissue and Bone* (IARC Press, 2002).
60. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
62. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
63. Li, H. *et al.* The sequence alignment/Map FORMAT and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
65. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
66. Cer, R. Z. *et al.* Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41**, D94–D100 (2012).
67. Jenjaroenpun, P., Wongsurawat, T., Yenamandra, S. P. & Kuznetsov, V. A. QmRLFS-finder: A model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res.* **43**, W527–W534 (2015).
68. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform.* **14**, 144–161 (2013).
69. Zhou, W. *et al.* Increased genome instability in human DNA segments with self-chains: Homology-induced structural variations via replicative mechanisms. *Hum. Mol. Genet.* **22**, 2642–2651 (2013).

Acknowledgements

We are grateful to the French Sarcoma Group for tumor banks and associated clinical annotations and to Jean-Baptiste Courrèges. The following French cancer centers also participated in this study: Centre Paul Papin (Angers), Centre Oscar Lambert (Lille), Institut Paoli Calmettes (Marseille), CHU La Timone (Marseille), CHU Bordeaux and hospital Ambroise Paré (Paris). This work was supported by grant from the "Instituts Thematiques Multiorganismes" (ITMO) Cancer (INSERM and INCa) and the Claudius Regaud Institute. The authors would like to thank the Centre Nacional d'Anàlisi Genòmica (CNAG, Barcelona, Spain) for WG sequencing services. We thank GENOTOUL for bioinformatics facilities and computer farm.

Author contributions

A.B. and F.Ch. wrote the main manuscript text. A.B. prepared figures, made bioinformatic analysis. L.L. and G.P. made tumor DNA extraction and library preparation. L.G. developed home made algorithm for breakpoints call. N.D. analyzed NGS data. M.B. has methodological and computational contribution. S.L., P.R., T.V., G.F., C.C., B.B., E.S., A.L., S.P.N., F.Co., N.F., G.D., J.M.C. and J.Y.B., provided tumor samples and clinical annotations. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no conflict of interest. Request for Grant of a European patent has been submitted under the application number: EP20306558.6. the applicants are: Institut National de la Santé et de la Recherche Médicale, UNIVERSITÉ PAUL SABATIER TOULOUSE III, Benhaddou Ataaillah, and Institut Claudius Regaud. The inventors are not yet designated. The Extended european search report is closed. The application covers the aspects of the manuscript relating to hotspots detection, iTRAC and iRACIN indexes, and a method for stratification of clinical outcome based on those indexes.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-02787-x.

Correspondence and requests for materials should be addressed to F.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.