The Hidden Genomic and Transcriptomic Plasticity of Giant Marker Chromosomes in Cancer

Gemma Macchia,^{*,1} Marco Severgnini,[†] Stefania Purgato,[‡] Doron Tolomeo,^{*} Hilen Casciaro,^{*}
Ingrid Cifola,[†] Alberto L'Abbate,^{*} Anna Loverro,^{*} Orazio Palumbo,[§] Massimo Carella,[§]
Laurence Bianchini,^{**} Giovanni Perini,[‡] Gianluca De Bellis,[†] Fredrik Mertens,^{††} Mariano Rocchi,^{*}
and Clelia Tiziana Storlazzi^{*}

^{*}Department of Biology, University of Bari Aldo Moro, 70125 Italy, [†]Institute for Biomedical Technologies, Consiglio Nazionale delle Ricerche, Segrate, 20090 Italy, [‡]Department of Pharmacy and Biotechnology, University of Bologna, 40126 Italy, [§]Laboratorio di Genetica Medica, Istituto di Ricovero e Cura a Carattere Scientifico, Casa Sollievo della Sofferenza, San Giovanni Rotondo, 71013 Italy, ^{**}Laboratory of Solid Tumor Genetics, Université Côte d'Azur, Centre National de la Recherche Scientifique, Institute of Research on Cancer and Aging in Nice, 06108 France, and ^{††}Department of Clinical Genetics, University and Regional Laboratories, Lund University, 22100 Sweden

ORCID IDs: 0000-0002-4323-6393 (H.C.); 0000-0001-6583-3482 (O.P.)

ABSTRACT Genome amplification in the form of rings or giant rod-shaped marker chromosomes (RGMs) is a common genetic alteration in soft tissue tumors. The mitotic stability of these structures is often rescued by perfectly functioning analphoid neocentromeres, which therefore significantly contribute to cancer progression. Here, we disentangled the genomic architecture of many neocentromeres stabilizing marker chromosomes in well-differentiated liposarcoma and lung sarcomatoid carcinoma samples. In cells carrying heavily rearranged RGMs, these structures were assembled as patchworks of multiple short amplified sequences, disclosing an extremely high level of complexity and definitely ruling out the existence of regions prone to neocentromere seeding. Moreover, by studying two well-differentiated liposarcoma samples derived from the onset and the recurrence of the same tumor, we documented an expansion of the neocentromeric domain that occurred during tumor progression, which reflects a strong selective pressure acting toward the improvement of the neocentromeric functionality in cancer. In lung sarcomatoid carcinoma cells we documented, extensive “centromere sliding” phenomena giving rise to multiple, closely mapping neocentromeric epialleles on separate coexisting markers occur, likely due to the instability of neocentromeres arising in cancer cells. Finally, by investigating the transcriptional activity of neocentromeres, we came across a burst of chimeric transcripts, both by extremely complex genomic rearrangements, and *cis/trans*-splicing events. Post-transcriptional editing events have been reported to expand and variegate the genetic repertoire of higher eukaryotes, so they might have a determining role in cancer. The increased incidence of fusion transcripts, might act as a driving force for the genomic amplification process, together with the increased transcription of oncogenes.

KEYWORDS neocentromere; fusion transcript; WDLPS; LSC; gene amplification

GENOME amplification is a frequent genetic alteration in cancer, with variable cytogenetic manifestations including double minutes, homogeneously-staining regions, and/or ring and giant rod-shaped marker chromosomes (RGMs) (Matsui *et al.* 2013; L'Abbate *et al.* 2014; Nord *et al.* 2014).

While double minutes and homogeneously-staining regions have been described in a variety of cancer types (Matsui *et al.* 2013), RGMs are particularly common in soft tissue tumors, notably in well-differentiated liposarcomas (WDLPS), and were shown to contain amplified sequences from several chromosomes (Nord *et al.* 2014). During tumor progression, the ring chromosomes are frequently broken and resealed, or transformed into rod-shaped markers, capturing the telomeres from other chromosomes (Nord *et al.* 2014). This instability results in a highly complex internal structure of these markers, as well as in extensive heterogeneity with respect to size and number per cell (Garsed *et al.* 2014; Nord *et al.*

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.117.300552>

Manuscript received September 1, 2017; accepted for publication December 11, 2017; published Early Online December 26, 2017.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300552/-/DC1.

¹Corresponding author: Department of Biology, University of Bari, 4 Via Orabona, 70125 Bari, Italy. E-mail: gemma.macchia@uniba.it

2014). RGMs frequently lack functional centromeric alphoid sequences and their mitotic stability is rescued by the emergence of perfectly functioning analphoid neocentromeres, which might indirectly contribute to cancer progression (Macchia *et al.* 2015). Nonetheless, there are few studies addressing neocentromeres in cancer, probably because most of the technologies employed to study the tumor genotypes are unable to unveil them. The occurrence of neocentromeres in cancer, therefore, could be more frequent than reported. Similarly, very little is known about the impact of neocentromeres on transcription, although centromeric satellite regions have been reported to produce noncoding transcripts actively involved in the centromere assembly (Chan *et al.* 2012; Rosic *et al.* 2014; Quenet and Dalal 2015; McNulty *et al.* 2017). Also, genes within neocentromeres are still actively transcribed (Amor and Choo 2002; Wong *et al.* 2006). In line with these notions, the occurrence of neocentromeres in colon cancer cell lines was reported to correlate with large DNase I hypersensitive sites, which are usually sites of active transcription or high nucleosome turnover (Athwal *et al.* 2015). By combining chromatin immunoprecipitation (IP) deep sequencing (ChIP-seq), whole-genome sequencing (WGS), immunofluorescence *in situ* hybridization (immuno-FISH), whole-transcriptome sequencing (total RNA-seq), and other molecular analyses, we investigated in detail the genomic architecture of neocentromeres arising on RGMs, as well as their contribution to transcription, in the lung sarcomatoid carcinoma (LSC) cell line 04T036 and in the three liposarcoma cell lines 93T449, 94T778, and 95T1000. Overall, our study uncovered the complex organization of neocentromeres in cancer, and shed light on the extraordinarily high genomic and transcriptomic plasticity associated with RGMs in solid tumors.

Materials and Methods

Tumor cell lines

Four tumor cell lines (04T036, 93T449, 94T778, and 95T1000), kindly provided by The Centre Hospitalier Universitaire de Nice (France), were included in the study. 04T036 was established from the LSC of a 50-year-old man. Cytogenetic and multicolor FISH analyses showed a near-triploid karyotype with numerous structural aberrations, four to six small RGMs containing chromosome 9 amplified sequences, and two RGMs containing chromosome 3 amplified sequences (Italiano *et al.* 2006). The 93T449 and 94T778 cell lines were obtained from a primary retroperitoneal WDLPS at onset and at relapse, respectively. These commercial cell lines showed complex karyotypes with multiple RGMs at G-banding and multicolor FISH analysis (Sirvent *et al.* 2000; Garsed *et al.* 2014). The 95T1000 cell line was generated from a WDLPS relapse; spectral karyotyping analysis revealed a hypertriploid karyotype with multiple chromosomal structural abnormalities (Pedeutour *et al.* 2012). All cells retained a giant marker chromosome, previously

identified in the primary cell cultures. This giant chromosome contained high-level amplification of chromosomal regions deriving from 10p to 12q and lacked α -satellite DNA (Pedeutour *et al.* 2012).

SNP array data

All cell lines were analyzed by the Affymetrix Genome-Wide Human SNP Array 6.0 platform (Affymetrix, Santa Clara, CA), as previously described (Storlazzi *et al.* 2010).

WGS

WGS was carried out to disentangle the genomic architecture of RGMs holding neocentromeres. Library preparations were performed using the TruSeqDNA Nano 350 bp protocol (Illumina, San Diego, CA). The sequencing data were acquired using the Illumina Xten at the New York Genome Center (New York, NY), in a paired-end 150-cycle run (mean coverage 40 $\times$ per sample). Reads were aligned to the human reference genome (GRCh37/hg19) using BWA-MEM (v.0.7.12) (<http://bio-bwa.sourceforge.net/>; Li and Durbin 2009) and PCR duplicates were removed using Picard (v.1.119) (<http://picard.sourceforge.net/>). Candidate structural variations (SVs) were identified using Delly (v. 0.5.9) and Crest (v. 1.0) with default parameters (Wang *et al.* 2011; Rausch *et al.* 2012). Copy number analysis was performed using Bayesian information criterion-seq (Xi *et al.* 2011), and genomic intervals showing a log2 copyRatio > 0.5 and > 2.5 were considered as amplified and highly amplified, respectively (Supplemental Material, File S1).

ChIP-seq

To determine the internal structure of the neocentromeres, native ChIP-seq was performed as described (Wade *et al.* 2009). IP was run using a polyclonal antibody against the CENP-A (Trazzi *et al.* 2009). Both input and IP DNA fragments were purified and processed using the TruSeq ChIP Library Preparation Kit (Illumina), and sequenced on the Illumina HiSeq 2500 at the Istituto di Genomica Applicata (IGA) Technology Services facility (Udine, Italy) (single-end 100-cycle run, 140 M reads per sample). Raw reads were aligned to the human reference genome (GRCh37/hg19) using BWA-MEM (v. 0.7.10). CENP-A-enriched regions corresponding to putative neocentromeres were identified using the CNV (copy number variant)-seq tool, merging all overlapping intervals (Xie and Tammi 2009). Selected regions were then filtered to exclude alphoid sequences, weak enrichments, and regions with read “spikes” piling-up in a single position. Next, putative neocentromeric fragments were ranked according to their Overall Evaluation Criteria (OEC) score (Severgnini *et al.* 2006). We then screened the first 40 hits looking at the ChIP-seq alignment data with the Integrative Genomic Viewer (IGV) software (Thorvaldsdottir *et al.* 2013) to exclude false positives, and validated the selected intervals by immuno-FISH as described below. Using this approach, multiple CENP-A-enriched peaks were detected in all samples. To define the neocentromere internal

structure, a local remapping of the IP reads was performed against the identified putative neocentromeric intervals (custom reference) by Blast (Altschul *et al.* 1990) or Blat (Kent 2002), searching for reads spanning over the junctions. The list of reads splitting between two neocentromeric regions was created and the OEC score was then used to prioritize the putative junctions. Full details are provided in [File S1](#).

Immuno-FISH assays on elongated chromosomes

Immuno-FISH on elongated chromosomes was performed to validate the putative structure of neocentromeres, as previously described (Earnshaw *et al.* 1989). As recommended by Beh *et al.* (2016), we used a rabbit anti-CENP-C polyclonal antibody and a goat anti-rabbit FITC-conjugate antibody to label functional centromeres (Beh *et al.* 2016). CENP-A and CENP-C were reported to localize exclusively to active centromeres, as part of the constitutive centromere-associated network (Klare *et al.* 2015; Shono *et al.* 2015; Beh *et al.* 2016). Alphoid and BAC probes spanning candidate neocentromeric regions were selected and labeled as reported (Trombetta *et al.* 2012; Macchia *et al.* 2015) ([Table S1](#)). Subsequent FISH experiments were conducted to verify the colocalization between CENP-C signals and the tested neocentromeric intervals as described (Storlazzi *et al.* 2006).

PCR and quantitative PCR (qPCR) assays

PCR and Sanger sequencing were performed to validate genomic SVs as described (Storlazzi *et al.* 2010). Primer sequences are available upon request. To test differential ChIP enrichments between the two related samples 93T449 and 94T778, qPCR experiments were performed starting from 10 ng of IP (target) and input DNA (negative control) with the ready-to-use hot start reaction mix for SYBR Green I-based real-time PCR assays using the LightCycler 96 System, according to the manufacturer's protocols (Roche, Basel, Switzerland). As positive control, we included the top ChIP-enriched region shared by the two cell lines, defined by immuno-FISH as the neocentromeric core. The results were analyzed with a custom approach based on the $\Delta\Delta C_t$ -method (Livak and Schmittgen 2001). In detail, the ΔC_t was first calculated for each neocentromeric peak as the difference in the C_t between the IP and the Input. Then, the $\Delta\Delta C_t$ was obtained using the positive control as calibrator. Finally, we compared the results for 93T449 and 94T778, looking at the $2^{-\Delta\Delta C_t}$ fold ratios between each region and the corresponding calibrator.

RNA-seq library preparation and analysis

We investigated the transcription activity of neocentromeres by total RNA-seq. Total RNA was extracted using an RNeasy Mini Kit (QIAGEN, Hilden, Germany), checked for integrity on a 2100 Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA), and stored at -80° until use. Libraries were prepared using the TruSeq Stranded Total RNA Library Prep Kit (Illumina), and sequenced on the HiSeq 2000 at the

IGA Technology Services facility (Udine, Italy) (paired-end 100-cycle run, 70 M reads per sample). After running FastQC for quality control (www.bioinformatics.babraham.ac.uk/projects/fastqc/), paired-end reads were mapped to the human reference genome (GRCh37/hg19) using STAR aligner (v.2.4) (Dobin *et al.* 2013), with Gencode (v. 19; Harrow *et al.* 2012) as the gene transcript model and Samtools (v. 0.1.19) (Li *et al.* 2009) for duplicate removal. Alignment tracks were checked by IGV to identify transcriptional activity of the neocentromeric sites.

Chimeric transcripts, RNA-seq, and WGS data integration

Looking at the RNA-seq alignment tracks by IGV, we found multiple truncated transcripts, likely derived from gene fusion events. To identify these chimeras, we analyzed RNA-seq data using ChimeraScan with default parameters (Iyer *et al.* 2011). The identified fusion transcripts were then mapped at SV positions with a custom approach combining WGS and RNA-seq data ([File S1](#)). RT-PCR and Sanger sequencing validations were performed on all transcripts, as previously described (Storlazzi *et al.* 2010), filtering out overlapping, adjacent, and readthrough transcripts, and all chimeras with a score <15 . For chimeras not supported by any SV, we considered for validation only those with split reads. All chimeras detected in 93T449 or 94T778 were tested in both cell lines.

Data availability

SNP array data are available at the European Molecular Biology Laboratory-European Bioinformatics Institute Array Express database (<https://www.ebi.ac.uk/arrayexpress>) under accession number E-MTAB-5625. WGS, RNA-seq, anti-CENP-A, and input ChIP-seq data are available at the National center for Biotechnology Information Short Read Archive (<https://www.ncbi.nlm.nih.gov/sra>) with accession number PRJNA378952. All validated SV sequences were submitted to the GenBank repository (<http://www.ncbi.nlm.nih.gov/genbank/>), under accession numbers KY966261–KY966313 and KY966314–KY966332 for chimeric transcripts and genomic fusions, respectively. Supplemental material is available for this study, and a description of each file is also provided ([File S2](#)).

Results

Seven distinct neocentromeres coexist in the cell line 04T036

The LSC cell line 04T036 was already described as carrying multiple markers stabilized by neocentromeres (Italiano *et al.* 2006). To study these neocentromeres, we first investigated the genomic context in which they occurred. Both SNP-array and WGS CNV analyses identified weak amplification of the 3q26.1-q29 and 9p24.3-p23 regions ([Table S1](#)). FISH with BAC probes mapping at these regions confirmed the presence of a variable number of RGMs, looking like isochromosomes, per cell: two derived from chromosome 3 and two to five

from chromosome 9 (Italiano *et al.* 2006). We mapped the fusion junctions of these isochromosomes using WGS (Figure 1), and validated the results by PCR and Sanger sequencing. Neocentromeres arising on the multiple copies of RGMs described by Italiano *et al.* (2006) were detected by looking at anti-CENP-C signals not colocalizing with alphoid probes at immuno-FISH on metaphase spreads (Figure 1A) (Italiano *et al.* 2006). To define the internal organization of these neocentromeres, we performed ChIP-seq with the anti-CENP-A antibody. This analysis disclosed seven separate peaks of enrichment spanning the amplified regions of the RGM of chromosome 3 (peaks 1 and 2) and 9 (peaks 3–7), ranging from 52 to 136 kb in size (Figure 1 and Table S2). The bell-shaped coverage profile of each candidate region, visualized in IGV by looking at the ChIP-seq alignment to the reference genome, as well as both ChIP-seq and WGS data (Table S3A and Table S3B), suggested that each of them was specific for a different marker chromosome (Figure 1, B and C). This hypothesis was verified by performing simultaneous immuno-FISH experiments on elongated chromosomes, with anti-CENP-C antibodies pinpointing centromeres, and different combinations of BAC probes spanning the ChIP-enriched peaks. The obtained results showed the colocalization of the neocentromeric CENP-C signal with a different BAC probe on a separate marker, confirming the existence of distinct neocentromeric epialleles arising on the multiple copies of RGMs (Figure 1).

Complex neocentromeres undergoing structural evolution in WDLPS cell lines

The combination of SNP-array and WGS approaches confirmed the already described highly complex internal structure of the marker chromosomes of the WDLPS cell lines (Italiano *et al.* 2009; Garsed *et al.* 2014) (Table S1, Table S3A, and Table S3B). In 95T1000, despite the internal arrangement of the multiple detected markers varying within and among cells, their amplified content remained highly conserved (Figure S1 and Table S1). The presence of neocentromeres arising on RGMs was confirmed by looking at anti-CENP-C signals not colocalizing with alphoid probes by immuno-FISH on metaphase spreads (Figure 2B). The anti-CENP-A ChIP-seq analysis, performed to characterize these neocentromeres, revealed four separate peaks of enrichment at 10p12.1, 10p12.33, 12q13.3, and 12q14.1, spanning 74, 9, 43, and 40 kb in size, respectively. The inspection of the ChIP-seq alignment data by IGV disclosed a coverage drop at one or both ends of each peak (Figure 2A), suggesting that the multiple noncontinuous, noncollinear enriched sequences might be juxtaposed by SVs to form a single neocentromere. Our custom ChIP-seq analysis (File S1) confirmed this hypothesis, revealing specific SVs joining the fragments (Figure 2A and Table S4), all validated by PCR and Sanger sequencing. Immuno-FISH experiments with anti-CENP-C antibodies and BAC probes spanning the core ChIP-enriched regions confirmed the occurrence of a single neocentromere stabilizing all the observed marker chromosomes (Figure 2C).

In 93T449 (primary tumor) and 94T778 (recurrence), we found 17 chromosomes involved in the amplification (Figure S2). By comparing the copy number profiles of the two related cell lines, we disclosed a perfect conservation of the overall RGM amplified content, with few differences in the copy number state of the amplicons (Figure S3 and Table S1). In metaphase spreads, a higher number of RGMs per cell was found in the recurrence vs. the onset tumor (average number of 2.63 vs. 1.14). Moreover, we detected 16 SVs specific for 93T449 and six for 94T778 (Table S5). In both cell lines, each observed RGM was mitotically stabilized by a neocentromere.

The CENP-A ChIP-seq analysis disclosed four strongly enriched regions shared by both cell lines, as well as five additional weakly enriched fragments, specific for 94T778, clearly suggesting an enlargement of the CENP-A centromeric domain in the recurrence of the tumor. As for 95T1000, the neocentromeric seeding occurred on a patchwork of amplified sequences (Figure 3). By juxtaposing each enrichment peak according to the identified SVs, we inferred two separate neocentromeric contiguous sequences (NEO1 and NEO2), sharing a 5-kb region at chr1: 188.377–188.382 Mb. Since no SV connecting these two contigs could be detected, we hypothesized the occurrence of two distinct neocentromeres arising at the tumor onset (Figure 3). To validate this hypothesis, we performed immuno-FISH experiments on elongated chromosomes with the anti-CENP-C antibody and BAC probes spanning the core regions of the two inferred neocentromeres. The results disclosed a mutually exclusive colocalization of the anti-CENP-C signals with either of the probes on each marker chromosome, confirming two separate neocentromeres occurring within the same cell in both samples (Figure 4). Finally, qPCR assays performed on the CENP-A IP and input DNA with primer pairs specific for (primers listed in Table S7) each fragment of NEO1 and NEO2 confirmed the differential enrichment ratio of shared vs. specific (94T778) ChIP-enriched sites, proving the size increase of both neocentromeres from the primary tumor to the recurrence (Figure S4). More specifically, NEO1 increased from 84 to 147 kb and NEO2 from 68 to 121 kb.

From neocentromeric transcription to the burst of chimeric transcripts

We investigated the transcription activity of neocentromeres by integrating RNA-seq, WGS, and ChIP-seq data. In line with previous studies (Amor and Choo 2002; Wong *et al.* 2006), we found that gene transcriptional activity was not hampered by the presence of neocentromeres. Surprisingly, we found two chimeric transcripts with one of the two partners mapping within the neocentromere domains of 93T449 and 94T778 (*i.e.*, *APP/HMCN1* and *LOC100507250/HMCN1*). Investigating their origin, we found that the SVs supporting these chimeras were not involved in the assembly of the neocentromere; therefore, they likely derived from additional rearrangements of the same amplified region, which did not acquire a centromeric function. Despite the fact that this analysis did not reveal any aberrant transcript specific for

Figure 1 The genomic architecture of the giant rod-shaped marker chromosomes (RGMs) and neocentromeres in the 04T036 cell line. (A) Metaphase spread showing immuno-FISH results on the 04T036 cell line. Probe names and signals are consistently colored: white arrowheads pinpoint neocentromeres on RGMs, detectable by CENP-C signals (green) not colocalizing with alphoid probe (red); BAC probes spanning the amplified regions are also hybridized (violet and blue) on the RGMs. (B and C) Schematic representation of the overall structures of RGM9 (B) and RGM3 (C) in 04T036; thin blue lines and breakpoint positions indicate the structural variation connections assembling the isochromosomes. Integrative Genomic Viewer partial auto-scaled plots of the CENP-A chromatin immunoprecipitation-enriched regions, with the corresponding input panels (negative controls); y- and x-axes represent read count averaged in windows (value shown at the left of each box) and reference human genome (GRCh37/hg19) coordinates (kb positions shown at the top), respectively; BAC probes spanning each neocentromeric peak are reported on the right. Immuno-FISH results are shown on the right (D); probe names and signals are consistently colored; CENP-C signals (green) mapping at RGMs pinpoint neocentromeres [as shown in (A)]. Left and middle boxes of Immuno-FISH panels show a two-way merge between each probe and CENP-C signals; right boxes show partial metaphases with three-way merged signals; in each box, pale blue and yellow arrowheads indicate neocentromeres carrying blue/green and red/green colocalizing signals, respectively. The results achieved with this combination of different neocentromeric probes confirmed the presence of multiple neocentromeric alleles [represented as red lines on the RGM ideograms (B and C)].

Figure 2 The genomic architecture of the neocentromeres in the 95T1000 cell line. (A) Integrative Genomic Viewer (IGV) partial plots of the CENP-A chromatin immunoprecipitation (IP)-enriched regions aligned to the human reference (GRCh37/hg19), juxtaposed in proper order and orientation to assemble the neocentromeric inferred structure of 95T1000. Top panel shows the normalized enrichment profile of each peak (IP vs. input; full details in File S1), and middle and lower panels show the IP and input raw enrichment profiles. y- and x-axes represent read count averaged in windows (value shown at the left of each box) and reference genome coordinates (kb positions shown at the top), respectively; we fixed the scale of the y-axis to the value of the peak with the highest read count average on IGV. BAC probes chosen for the immuno-FISH experiments are reported below each corresponding neocentromeric peak. The lower panel is a schematic reconstruction of the neocentromeric assembled contig. Each neocentromeric fragment is represented as an arrow and aligned to the upper panel. Thin blue lines represent structural variation connections, and flags indicate breakpoint positions. Immuno-FISH results with anti CENP-C antibodies (green) cohybridized with alphoid probe (B) and BAC probes spanning the neocentromeric core (C) are shown; probe names and signals are consistently colored; CENP-C signals (green), not colocalizing with the alphoid probe (red), pinpoint neocentromeres mapping at giant rod-shaped marker chromosomes in (B), while white arrowheads indicate neocentromeres carrying merged blue/red/green signals in (C).

neocentromeres, it shed light on the multiple chimeric transcripts originating through the amplification process. We detected thousands of putative chimeras, and validated the most abundant ones by RT-PCR and Sanger sequencing (Table 1 and Table S6). Looking at the Refseq gene function annotations, most of the chimeric partners were reported as cancer-associated genes (Table 1). To investigate the origin of these chimeras, we built a custom pipeline to integrate WGS and RNA-seq data (for details see File S1), and demonstrated that some transcripts were supported by perfectly matching SVs (class I), while others showed a much more complex origin. Class II transcripts were assembled by means

of several noncontiguous, noncollinear genomic fragments interposed between partner genes, which were actively transcribed, but subsequently spliced out from the mature mRNA (Figure S5 and Figure S6). On the contrary, class III chimeras, lacking any supporting SV, could possibly originate from post-transcriptional events (Table 1).

Discussion

In the present study, we unveiled the extremely complex molecular architecture of neocentromeres mitotically stabilizing RGMs in WDLPS and LSC cell lines, describing their

Figure 3 The genomic architecture of NEO1 and NEO2 in 93T449 and 94T778 cell lines. At the top of each box, Integrative Genomic Viewer (IGV) partial plots at the CENP-A chromatin immunoprecipitation (IP)-enriched regions of NEO1 (A) and NEO2 (B) are shown as juxtaposed in a proper order and orientation to assemble the neocentromeric domain of both 93T449 and 94T778 cell lines. The top panels of both (A) and (B) boxes show the normalized enrichment profile of the neocentromeric peaks for each cell line (IP vs. input; full details in File S1); the middle and lower panels show the IP and input raw enrichment profiles. y- and x-axes represent read count averaged in windows (the relative values are shown on the left of each box) and reference genome coordinates (kb positions shown at the top), respectively; we fixed the scale of the y-axis to the value of the peak with the highest

Figure 4 Metaphase spreads showing immuno-FISH results on the 93T449 (A and C) and 94T778 (B and D) cell lines. Probe names and signals are consistently colored; CENP-C signals (green), not colocalizing with alphoid probes (red), pinpoint neocentromeres on giant rod-shaped marker chromosomes (green arrowheads), while probes RP11-203K8 and RP11-30I11 are specific for NEO1 and NEO2 cores, respectively. (C and D) Left and middle boxes show two-way merged images of each probe and CENP-C signals; boxes to the right show the three-way merged signals on partial metaphase spreads; pale blue and yellow arrowheads indicate neocentromeres carrying blue/green and red/green merged signals, respectively. The results achieved with this combination of different neocentromeric probes demonstrate the presence of two neocentromeric alleles in both cell lines.

structure down to the nucleotide level, and disclosed the occurrence of a burst of chimeric transcripts associated with genomic amplification. Both these findings shed light on the extraordinary genomic and transcriptomic plasticity of RGMs, likely playing a role in cancer evolution.

The enhanced complexity of neocentromeres arising at RGMs

The genomic architecture of neocentromeres in LSC and WDLPS cancer types showed distinct features. Each of the multiple RGMs found in the 04T036 LSC cell line had a single neocentromere, embedded in a continuous, nonrearranged sequence. Four out of the five neocentromeres mitotically stabilizing the multiple copies of RGM9 mapped close to each other (within 1 Mb), while the fifth mapped ~10 Mb apart. Instead, the two neocentromeres stabilizing the RGM3 mapped ~5 Mb apart from each other. Very likely, only a single neocentromeric-seeding event occurred on each of the two ancestral types of RGM, which subsequently multiplied following mitotic errors. The different positions of the neocentromeres arising on RGM3 and RGM9 were likely due to extensive “sliding” processes occurred along the chromo-

somes during tumor evolution, which led to functional “epi-alleles.” This phenomenon was recently discovered by Purgato *et al.* (2015) while studying the satellite-free centromere of horse chromosome 11 (likely an evolutionary neocentromere) (Purgato *et al.* 2015). That centromere was sliding within a ~500-kb genomic segment, a smaller region compared to those described here in 04T036. When a neocentromere arises in natural populations, the meiotic process can be affected by large distance sliding events giving rise to epialleles. Indeed, in heterozygous individuals, crossing-over events involving regions delimited by neocentromeric epialleles would generate acentric/dicentric chromosomes, resulting in a reduced fitness. However, the mitotic process is not affected by distant neocentromeric epialleles, so, in cancer, it seems very likely that neocentromere fluctuations might involve several megabases, especially on RGMs. Moreover, the alternative hypothesis calling for the simultaneous and independent seeding of seven neocentromeres on separate acentric RGMs seems highly unlikely.

Different from what was observed in the LSC cell line, the neocentromeres of all the studied WDLPS cell lines showed a

read count average on IGV. BAC probes chosen for the immuno-FISH experiments are reported below the corresponding neocentromeric peak. At the bottom of each box, a schematic reconstruction of the neocentromeric assembled contig is represented. Each neocentromeric fragment is shown as an arrow and aligned to the upper panel; thin blue lines represent structural variation connections and flags indicate breakpoint positions. A clear differential enrichment between the neocentromeric tails is shown when comparing 93T449 and 94T778. chr, chromosome.

Table 1 List of all the validated chimeras detected by ChimeraScan; class I, II, and III refer to chimeras derived from canonical, complex, and transcription-induced fusions, respectively

Chimera	Score	Class	Frame	<i>In silico</i> translation
04T036				
PPARD /MYO1D	49		IF	N-terminally-truncated MYO1D
ZNF286A/SIAE-SPA17	18	Class I	IF	Promoter swapping
BC065763/abParts	19	Class III	IF	NA
95T1000				
FRS2 / <u>UPF2</u>	909	Class I	IF	Promoter swapping
<u>UP00462</u> / LEMD3	477	Class I	IF	N-terminally-truncated LEMD3 (loss of LEM domain)
ITGA5 / SLC9B1	294	Class I	IF	Chimeric protein
GNS/ANKRD26	262	Class I	IF	Chimeric protein
ANKRD26/ NCKAP1L	208	Class I	IF	Chimeric protein
CAMK1D /ANAPC5	205	Class I	IF	No conserved domain retained
LEMD3 /MKX	122	Class II	IF	C-terminally-truncated LEMD3
MLLT10 / CCNT1	85	Class II	IF	Chimeric protein
FAM107B/CDNF	72	Class I	IF	No conserved domain retained
HMGA2 /KRT121P	44	Class I	IF	C-terminally-truncated HMGA2
LEMD3 / <u>UPF2</u>	35	Class III	IF	Chimeric protein
KDM2B/ <u>CELF2</u>	22	Class I	IF	Promoter swapping
CDC123 /LOC100240734	92	Class I	IF	C-terminally-truncated CDC123
CRADD /LYZ	22	Class III	IF	Chimeric protein
93T449				
FURIN /ADAMTS1	328	Class I	OF	C-terminally-truncated FURIN
VEZT /CPZ	251	Class II	IF	N-terminally-truncated CPZ (loss of CRD domain)
B4GALNT1/ SLC28A3	175	Class I	IF	Promoter swapping
VAMP4 / TAF3	170	Class I	IF	N-terminally-truncated TAF3 (loss of BTP domain)
NDIFP2/ <i>MYO16-AS1</i>	113	Class II	IF	No conserved domain retained
<i>LOC100507250</i> /HMCN1	103	Class II	IF	N-terminally-truncated HMCN1 (loss of IG domains)
SUCO/ SLC12A7	90	Class II	IF	No conserved domain retained
<i>QKI</i> / <i>LOC100132735</i>	78	Class I	IF	C-term truncated QKI
APP/HMCN1	76	Class I	IF	N-terminally-truncated HMCN1 (loss of IG domains)
SH3TC1/ TRIP13	73	Class I	IF	C-term truncated SHETC1
SLC3A1 /BACE2	51	Class III	IF	N-terminally-truncated BACE2 (no domain loss)
MARCH9/ SLC26A10	37	Class I	OF	No conserved domain retained (uc001spy.3 isoform)
MARCH9/BC073932	36	Class I	OF	No conserved domain retained (uc001spx.2 isoform)
IGF1R / <i>LOC101926960</i>	36	Class II	IF	No conserved domain retained
RASAL2 / ACVR1	19	Class I	IF	Promoter swapping
94T778				
NDIFP2/ <i>MYO16-AS1</i>	190	Class II	IF	No conserved domain retained
FURIN /ADAMTS1	180	Class I	OF	C-terminally-truncated FURIN
SUCO/ SLC12A7	140	Class II	IF	No conserved domain retained
MARCH9/ SLC26A10	23	Class I	OF	No conserved domain retained (uc001spy.3 isoform)
VEZT /CPZ	119	Class II	IF	N-terminally-truncated CPZ (loss of CRD domain)
APP/HMCN1	101	Class I	IF	N-terminally-truncated HMCN1 (loss of Ig domains)
VAMP4 / TAF3	91	Class I	IF	N-terminally-truncated TAF3 (loss of BTP domain)
B4GALNT1/ SLC28A3	78	Class I	IF	Promoter swapping
IGF1R / <i>LOC101926960</i>	77	Class II	IF	No conserved domain retained
SH3TC1/ TRIP13	69	Class I	IF	C-terminally-truncated SHETC1
<i>LOC100507250</i> /HMCN1	65	Class II	IF	N-terminally-truncated HMCN1 (loss of Ig domains)
<i>LIMS3-LOC440895</i> / MALL	28	Class III	IF	N-terminally-truncated MALL (MARVEL domain retained)
ASB7/FGD6	18	Class III	IF	Chimeric protein
SLC3A1 /BACE2	15	Class III	IF	N-terminally-truncated BACE2 (no domain loss)
MARCH9/BC073932	8	Class I	OF	No conserved domain retained (uc001spx.2 isoform)
RASAL2 / ACVR1	11	Class I	IF	Promoter swapping

Gene names styles are consistent with their main function (bold = cancer genes; underlined = ubiquitination; italic and underlined = RNA-splicing and surveillance; italic = long noncoding; and bold = trafficking and transport). IF, in frame; NA, not analyzed; OF, out of frame.

very complex structure, consisting of a patchwork of multiple short sequences amplified from several chromosomes. The structural complexity of neocentromeres in WDLPS strongly indicates that they arose secondarily to the amplification process, very likely as a consequence of the massive recruitment of CENP-A on highly rearranged RGMs to enable double-strand-

break repair (Zeitlin *et al.* 2009). Moreover, by comparing the 93T449 and 94T778 cell lines (primary and relapse of the same tumor, respectively), we disclosed additional interesting features. The perfect match between the amplified RGM content of the two related cell lines indicates that these structures arose as early events in the tumor's evolution and were maintained

under a strong selective pressure. The higher number of SVs specific for the onset cell line (93T449) also suggested that the recurrence was established from a minor subclone of the primary tumor. Moreover, the neocentromeres of 94T778 were 50–60 kb larger than those of 93T449, suggesting an evolution of these structures during tumor progression. As the size of centromeres tends to be uniform within species and the expansion of the neocentromeric size has been suggested to constitute a key factor in the survival of neocentric chromosomes (Wang *et al.* 2014), we speculate that the size increase of neocentromeres during tumorigenesis might be correlated with acquired mitotic stability of the RGMs harboring them. As for 04T036, the presence of two separate neocentromeres, even sharing a 5-kb domain in the 94T778 cell line, might also be the result of centromeric sliding phenomena.

Literature on human neocentromeres arising in acentric chromosomal fragments indicates a variable functional efficiency of these structures, which often leads to mosaicism for the marker (Marshall *et al.* 2008). In cancer, an incomplete functional efficiency of neocentromeres, causing nondisjunction, might lead to the segregation of multiple copies of RGMs to daughter cells, potentially increasing their selective advantage. This consideration could apply to the finding of a higher number of RGMs in a relapse (94T778) than in a primary tumor (93T449). It might be also conjectured that the instability of the neocentromeres in the primary tumor led to the selection of a higher number of RGMs per cell, subsequently stabilized by a neocentromere expansion in the recurrence. Combined, our results shed light on the extraordinary evolutionary plasticity of cancer-associated neocentromeres, and provide strong support for the strictly epigenetic nature of the neocentromeric seeding process.

Transcriptomic plasticity: the hidden side of the complex rearrangements in RGMs

Centromeric transcription has been reported as actively participating in kinetochore assembly and function (Muller and Almouzni 2017). However, our RNA-seq analyses did not reveal any unusual transcriptional activity specific for neocentromeres, but disclosed the occurrence of several chimeric transcripts derived from the whole RGM, which we considered worthwhile to investigate, as very few previous studies focused on this topic. The combined analysis of genomic and transcriptomic sequencing data allowed us to detect chimeric transcripts originating from extremely complex rearrangements, such as those resulting in the fusion of three partner genes (among class I transcripts), or those assembled by means of multiple interposed noncontiguous, noncollinear genomic fragments (class II). To the best of our knowledge, such complex fusions have not been described before in cancer. The molecular mechanism underlying these genomic chimeras might resemble the “exon-shuffling” process creating functional genetic novelties and multi-domain proteins over natural evolution (Franca *et al.* 2012). Indeed, a similar but accelerated mechanism might enrich the genetic repertoire of cancer cells, likely providing a selective advantage. The splic-

ing machinery also seems to be actively involved in shaping chimeras and creating variability in cancer. In assembling class II chimeras, for instance, it actively removes the interposed fragments from the mature mRNA. Moreover, it creates the multiple examples of transcription-induced chimeras (class III) here described, although our analysis likely just uncovered the tip of the iceberg (Lu *et al.* 2016).

The burst of chimeric transcripts documented in this study brings a focus on the extreme transcriptome plasticity of highly rearranged RGMs, which might be crucial for cancer initiation and progression. Indeed, the shuffling of coding sequences might strongly contribute to the proliferative success of the cancer cells by enriching and variegating the genetic substrate for selection to act on. We propose that the increased incidence of gene fusion events, caused by the extremely rearranged nature of RGMs, might act as a driving force for genome amplification, together with the well-established increased expression of oncogenes (Taylor *et al.* 2011; Panagopoulos *et al.* 2014).

Acknowledgments

This work was supported by a grant from the Italian Association on Cancer Research (AIRC Investigator grant 15413) to C.T.S., a grant from the Progetti di Interesse Nazionale and EPIGEN (Consiglio Nazionale delle Ricerche) to M.R., and Fondazione cassa di Risparmio di Puglia and accordo di programma quadro (APQ) Ricerca Regione Puglia (FutureInResearch) to G.M. The authors have no competing interests to declare.

Literature Cited

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.
- Amor, D. J., and K. H. Choo, 2002 Neocentromeres: role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* 71: 695–714.
- Athwal, R. K., M. P. Walkiewicz, S. Baek, S. Fu, M. Bui *et al.*, 2015 CENP-A nucleosomes localize to transcription factor hot-spots and subtelomeric sites in human cancer cells. *Epigenetics Chromatin* 8: 2.
- Beh, T. T., R. N. MacKinnon, and P. Kalitsis, 2016 Active centromere and chromosome identification in fixed cell lines. *Mol. Cytogenet.* 9: 28.
- Chan, F. L., O. J. Marshall, R. Saffery, B. W. Kim, E. Earle *et al.*, 2012 Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proc. Natl. Acad. Sci. USA* 109: 1979–1984.
- Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.
- Earnshaw, W. C., H. Ratrie, III, and G. Stetten, 1989 Visualization of centromere proteins CENP-B and CENP-C on a stable dicentric chromosome in cytological spreads. *Chromosoma* 98: 1–12.
- Franca, G. S., D. V. Cancherini, and S. J. de Souza, 2012 Evolutionary history of exon shuffling. *Genetica* 140: 249–257.
- Garsed, D. W., O. J. Marshall, V. D. Corbin, A. Hsu, L. Di Stefano *et al.*, 2014 The architecture and evolution of cancer neochromosomes. *Cancer Cell* 26: 653–667.

- Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans *et al.*, 2012 GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22: 1760–1774.
- Italiano, A., R. Attias, A. Aurias, G. Perot, F. Burel-Vandenbos *et al.*, 2006 Molecular cytogenetic characterization of a metastatic lung sarcomatoid carcinoma: 9p23 neocentromere and 9p23-p24 amplification including JAK2 and JMJD2C. *Cancer Genet. Cytogenet.* 167: 122–130.
- Italiano, A., G. Maire, N. Sirvent, P. A. Nuin, F. Keslair *et al.*, 2009 Variability of origin for the neocentromeric sequences in aneuploid supernumerary marker chromosomes of well-differentiated liposarcomas. *Cancer Lett.* 273: 323–330.
- Iyer, M. K., A. M. Chinnaiyan, and C. A. Maher, 2011 ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* 27: 2903–2904.
- Kent, W. J., 2002 BLAT—the BLAST-like alignment tool. *Genome Res.* 12: 656–664.
- Klare, K., J. R. Weir, F. Basilico, T. Zimniak, L. Massimiliano *et al.*, 2015 CENP-C is a blueprint for constitutive centromere-associated network assembly within human kinetochores. *J. Cell Biol.* 210: 11–22.
- L'Abbate, A., G. Macchia, P. D'Addabbo, A. Lonoce, D. Tolomeo *et al.*, 2014 Genomic organization and evolution of double minutes/homogeneously staining regions with MYC amplification in human cancer. *Nucleic Acids Res.* 42: 9131–9145.
- Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Livak, K. J., and T. D. Schmittgen, 2001 Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)). *Method. Methods* 25: 402–408.
- Lu, G., J. Wu, G. Zhao, Z. Wang, W. Chen *et al.*, 2016 Abundant and broad expression of transcription-induced chimeras and protein products in mammalian genomes. *Biochem. Biophys. Res. Commun.* 470: 759–765.
- Macchia, G., K. H. Nord, M. Zoli, S. Purgato, P. D'Addabbo *et al.*, 2015 Ring chromosomes, breakpoint clusters, and neocentromeres in sarcomas. *Genes Chromosomes Cancer* 54: 156–167.
- Marshall, O. J., A. C. Chueh, L. H. Wong, and K. H. Choo, 2008 Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. *Am. J. Hum. Genet.* 82: 261–282.
- Matsui, A., T. Ihara, H. Suda, H. Mikami, and K. Semba, 2013 Gene amplification: mechanisms and involvement in cancer. *Biomol. Concepts* 4: 567–582.
- McNulty, S. M., L. L. Sullivan, and B. A. Sullivan, 2017 Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. *Dev. Cell* 42: 226–240.e6.
- Muller, S., and G. Almouzni, 2017 Chromatin dynamics during the cell cycle at centromeres. *Nat. Rev. Genet.* 18: 192–208.
- Nord, K. H., G. Macchia, J. Tayebwa, J. Nilsson, F. Vult von Steyern *et al.*, 2014 Integrative genome and transcriptome analyses reveal two distinct types of ring chromosome in soft tissue sarcomas. *Hum. Mol. Genet.* 23: 878–888.
- Panagopoulos, I., B. Bjerkehaugen, L. Gorunova, J. M. Berner, K. Boye *et al.*, 2014 Several fusion genes identified by whole transcriptome sequencing in a spindle cell sarcoma with rearrangements of chromosome arm 12q and MDM2 amplification. *Int. J. Oncol.* 45: 1829–1836.
- Pedeutour, F., G. Maire, A. Pierron, D. M. Thomas, D. W. Garsed *et al.*, 2012 A newly characterized human well-differentiated liposarcoma cell line contains amplifications of the 12q12–21 and 10p11–14 regions. *Virchows Arch.* 461: 67–78.
- Purgato, S., E. Belloni, F. M. Piras, M. Zoli, C. Badiale *et al.*, 2015 Erratum to: centromere sliding on a mammalian chromosome. *Chromosoma* 124: 289.
- Quenet, D., and Y. Dalal, 2015 Correction: a long non-coding RNA is required for targeting centromeric protein A to the human centromere. *Elife* 4: e07239.
- Rausch, T., T. Zichner, A. Schlattl, A. M. Stutz, V. Benes *et al.*, 2012 DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28: i333–i339.
- Rosic, S., F. Kohler, and S. Erhardt, 2014 Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *J. Cell Biol.* 207: 335–349.
- Severgnini, M., L. Pattini, C. Consolandi, E. Rizzi, C. Battaglia *et al.*, 2006 Application of the Taguchi method to the analysis of the deposition step in microarray production. *IEEE Trans. Nanobioscience* 5: 164–172.
- Shono, N., J. Ohzeki, K. Otake, N. M. Martins, T. Nagase *et al.*, 2015 CENP-C and CENP-I are key connecting factors for kinetochore and CENP-A assembly. *J. Cell Sci.* 128: 4572–4587.
- Sirvent, N., A. Forus, W. Lescaut, F. Burel, S. Benzaken *et al.*, 2000 Characterization of centromere alterations in liposarcomas. *Genes Chromosomes Cancer* 29: 117–129.
- Storlazzi, C. T., T. Fioretos, C. Surace, A. Lonoce, A. Mastorilli *et al.*, 2006 MYC-containing double minutes in hematologic malignancies: evidence in favor of the episome model and exclusion of MYC as the target gene. *Hum. Mol. Genet.* 15: 933–942.
- Storlazzi, C. T., A. Lonoce, M. C. Guastadisegni, D. Trombetta, P. D'Addabbo *et al.*, 2010 Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res.* 20: 1198–1206.
- Taylor, B. S., P. L. DeCarolis, C. V. Angeles, F. Brenet, N. Schultz *et al.*, 2011 Frequent alterations and epigenetic silencing of differentiation pathway genes in structurally rearranged liposarcomas. *Cancer Discov.* 1: 587–597.
- Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14: 178–192.
- Trazzi, S., G. Perini, R. Bernardoni, M. Zoli, J. C. Reese *et al.*, 2009 The C-terminal domain of CENP-C displays multiple and critical functions for mammalian centromere formation. *PLoS One* 4: e5832.
- Trombetta, D., G. Macchia, N. Mandahl, K. H. Nord, and F. Mertens, 2012 Molecular genetic characterization of the 11q13 breakpoint in a desmoplastic fibroma of bone. *Cancer Genet.* 205: 410–413.
- Wade, C. M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre *et al.*, 2009 Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326: 865–867.
- Wang, J., C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley *et al.*, 2011 CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* 8: 652–654.
- Wang, K., Y. Wu, W. Zhang, R. K. Dawe, and J. Jiang, 2014 Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res.* 24: 107–116.
- Wong, N. C., L. H. Wong, J. M. Quach, P. Canham, J. M. Craig *et al.*, 2006 Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS Genet.* 2: e17.
- Xi, R., A. G. Hadjipanayis, L. J. Luquette, T. M. Kim, E. Lee *et al.*, 2011 Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. USA* 108: E1128–E1136.
- Xie, C., and M. T. Tammi, 2009 CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10: 80.
- Zeitlin, S. G., N. M. Baker, B. R. Chapados, E. Soutoglou, J. Y. Wang *et al.*, 2009 Double-strand DNA breaks recruit the centromeric histone CENP-A. *Proc. Natl. Acad. Sci. USA* 106: 15762–15767.

Communicating editor: S. Sharan