Somatic structural variation and cancer

Jose M. C. Tubio

Corresponding author. J. M. C. Tubio, Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. Tel.: +44 (0) 1223 494951; E-mail:jt14@sanger.ac.uk

Abstract

With the advent of next-generation sequencing technologies, we have witnessed a rapid pace of discovery of new patterns of somatic structural variation in cancer genomes, and an attempt to figure out their underlying mechanisms. Some of these mechanisms are associated with particular cancer types, and in some cases are the main cause of the structural mutations that drive the oncogenic process. This review provides an overview of the patterns of somatic structural variation and chromosomal structures that characterize cancer genomes, their causal mechanisms and their impact in oncogenesis.

Key words: somatic structural variation; cancer genome; driver mutation; mutational process

Introduction

In cancer genomics, somatic structural variation is defined as the entire repertoire of DNA rearrangements harbored by the genome of the cancer cell, and absent in the genome of the ‘normal’ cells from the same patient. Hence, somatic structural variation could be acquired during somatic evolution before or after the cell became cancerous. Somatic structural variation includes rearrangements involving copy number changes (i.e. deletions and amplifications, usually above 1 kb), intrachromosomal inversions, interchromosomal rearrangements and the insertion of transposable elements (TEs) and other exogenous sequences (such as viral DNA and mitochondrial DNA). Although most structural variants generated somatically are neutral for the propagation of the cancer clone, some reorganizations may result in the alteration of key genes that drive oncogenesis [1].

For the past years, the pace of discovery in cancer genomics has been greatly accelerated due to the advent of next-generation sequencing technology. The development of algorithms for the analysis of paired-end read sequencing data [2] has allowed the identification of genomic structural variants in cancer genomes [3]. Thousands of cancer genomes from tens of different cancer types have been sequenced and analyzed to date, showing a variable number of genomic structural rearrangements acquired somatically, ranging from zero to thousands, and a variable relative proportion of rearrangement types [4]. Within this variability, it is possible to identify patterns of somatic structural variation that characterize particular tumor types. These patterns reflect the existence of a wide range of mutational processes, sometimes only restricted to specific cancer types or, at least, showing different prevalence across cancer types.

This review intends to illustrate the patterns of somatic genomic structural rearrangements and chromosomal structures that characterize cancer genomes. The structural configurations described here (summarized in Table 1) include: chromothripsis – the patterns of massive, highly localized rearrangement breakpoints found in 3% of all cancer genomes; double-minute chromosomes (DMs) and neochromosomes – episomally replicating structures found in a wide variety of solid tumors and leukemias; chromoplexy – the complex pattern of chained rearrangements found in prostate cancer; the secondary deletions caused by an off-target activity of RAG enzymes found in ETV6/RUNX1 acute lymphoblastic leukemia (ALL); the pattern of diverse structural variant types leading to enhancer hijacking found in some medulloblastomas; Igh translocations and the long-range impact of regulatory regions in the activation of oncogenes in lymphoma; the amplification of coding and regulatory regions, and the mobilization of processed pseudogenes, mediated by L1 retrotransposon activity in lung and colorectal cancers; and the focal amplification of genomic DNA mediated by human papillomavirus (HPV) in cervical cancer.

Chromothripsis

The traditional view of oncogenesis postulates that cancer is the consequence of a gradual process involving the continual acquisition of genetic mutations, leading to the activation of oncogenes and/or the inactivation of tumor-suppressor genes.

Jose Tubio is a ‘Marie Curie’ postdoctoral fellow at the Wellcome Trust Sanger Institute, and a member of the PanCancer initiative.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com
Confined to just one chromosomal arm (Figure 1B), where the DNA breakpoints typically show pronounced clustering (i.e. at least 5–10 breaks are usually observed within 50 kb, followed by long tracks of intact chromosomal sequence). Rearrangements are frequently interspersed with deletions, which might be a consequence of DNA loss during the repair process after shattering. The chromosomal regions with chromothripsis show a copy number state oscillating between one and three copies (usually one or two copies), whereas statistical simulations demonstrated that, under a progressive acquisition of somatic rearrangements, the number of different states observed increases as the number of breakpoints rises (Figure 1C).

Overall, it is estimated that chromothripsis is present in 2–3% of all cancers [5], but the prevalence of chromothripsis shows extensive variation across cancer types [13]: 30% of osteosarcoma samples, 18% of chondromas, 10% of chronic lymphocytic leukemia samples, 11% of neuroblastomas and 1.3% in multiple myeloma. The analysis of structural variation in medulloblastoma samples carried out by Raush et al. [8] suggests that chromothripsis is linked to deficiencies in TP53 activity. The researchers found that in a type of medulloblastoma, called Sonic Hedgehog-pathway (SHH) medulloblastoma, bearing TP53 mutations, chromothripsis frequency varies from 30 to 100%, whereas there is no evidence of chromothripsis in SHH medulloblastoma bearing the wild-type TP53/++ genotype. Similarly, their analysis of acute myeloid leukemia (AML) samples with somatically acquired TP53 mutations versus wild-type TP53 revealed chromothripsis patterns in about half of the samples with the mutated TP53 genotype, but just 1% of AMLs with the standard TP53 genotype.

### Breakage-fusion-bridge cycles and micronuclei formation trigger chromothripsis

Barbara McClintock described the mechanism of breakage-fusion-bridge (BFB) in the late 1930s studying maize [14], pointing out the structural impact of such events in generating inversion and amplification of DNA segments. Almost 80 years
later, BFB is considered a relevant mutational mechanism involved in oncogenesis, because it may lead to the duplication of oncogenes and the generation of a complex pattern of massive genomic rearrangements that characterizes chromothripsis [5, 15].

The mechanism of BFB is triggered by telomere dysfunction (Figure 2A). When telomeres are shortened, or uncapped by loss of shelterin (the complex of proteins that protects them), p53-signaling triggers programmed cell death (apoptosis). But in cells with gene mutations that diminish p53, apoptosis may not take place [16]. Under this last scenario, replication of a telomere-free chromosome results in fusion at the position of breakage between the two split halves, generating an end-to-end chromosomal fusion. This fusion is promoted by the end-joining DNA repair mechanisms of the cell, and creates a dicentric chromosome in which the two centromeres are pulled to opposite poles during mitosis, forming an anaphase bridge. The tension on the bridge configuration, following the poleward migration of the centromeres, results in rupture of the dicentric chromosome. Once again, a chromatid with a new broken end can repeat the process in the next cell cycle.

The breakage of the dicentric chromosome after the anaphase bridge has major consequences for the structural configuration of the derived chromosomes. First, the chromosomes in anaphase conform a rigid structure that, when broken, may generate hundreds of shards of genomic DNA that circulate in the nucleus. The DNA repair machinery of the cell collects some of these fragments of DNA and pulls them together at the broken extreme of the derivative chromosome, creating structural variability that follows the pattern of chromothripsis. Second,
one of the two derivative chromosomes bears the duplication of a region of variable length that includes the subtelomeric region of the arm that missed the telomere. The length of the amplified region depends on where the breakage has occurred along the homologous arm in the dicentric chromosome. After breakage, some chromosomal regions are incorporated to the derived chromosome as shown in Figure 1A, others are fused to create episomal structures or are retained in micronuclei after mitosis. (B) In a later cell cycle, chromothripsis occurs in micronuclei after defective DNA replication. Only two chromosomes are shown. According to Crasta et al. [9], in S phase, DNA replication within the micronucleus occurs slower than in the nucleus, and it is still incomplete when nuclear chromosomes start condensation. Premature condensation in the micronucleus triggers DNA shattering. DNA fragments are joined again leading to a new chromosomal configuration with the pattern of chromothripsis.

The structural, and potential functional, consequences of BFB are well-represented by a subtype of ALL characterized by defective DNA replication. These structures have incomplete nuclear pore complex formation and nuclear import defects. Crasta et al. [9] propose that chromothripsis occurs in micronuclei. The authors promoted the formation of micronuclei by inducing chromosome mis-segregation, and identified DNA damage foci that occurred inside these structures due to defective DNA replication (Figure 2B). DNA damage was characterized by extensive fragmentation. Moreover, Crasta et al. observed that DNA from micronuclei frequently rejoins the main nucleus following nuclear envelope breakdown at mitosis, providing an opportunity for the advent of oncogenic fusions typically found in chromothripsis. The segregation of chromosome fragments in micronuclei and subsequent shattering represent a complementary explanation for the localized nature of chromothripsis.

The structural, and potential functional, consequences of BFB are well-represented by a subtype of ALL characterized by defective DNA replication. These structures have incomplete nuclear pore complex formation and nuclear import defects. Crasta et al. [9] propose that chromothripsis occurs in micronuclei. The authors promoted the formation of micronuclei by inducing chromosome mis-segregation, and identified DNA damage foci that occurred inside these structures due to defective DNA replication (Figure 2B). DNA damage was characterized by extensive fragmentation. Moreover, Crasta et al. observed that DNA from micronuclei frequently rejoins the main nucleus following nuclear envelope breakdown at mitosis, providing an opportunity for the advent of oncogenic fusions typically found in chromothripsis. The segregation of chromosome fragments in micronuclei and subsequent shattering represent a complementary explanation for the localized nature of chromothripsis.
by the intrachromosomal amplification of chromosome 21 (iAMP21) [7], where BFBs are typically initiation events.

The role of breakage-fusion-bridge cycles in iAMP21-acute lymphoblastic leukemia

ALL is the most common childhood cancer, affecting children between 0 and 14 years old. Approximately 2% of the cases show the recurrent (but size variable) amplification of a mega-base region of chromosome 21, which contains the cancer gene RUNX1 [20]. This cytogenetic abnormality characterizes a subtype of ALL called intrachromosomal amplification of chromosome 21 (iAMP21) [21].

Traditionally, studies based on cytogenetics and array-based comparative genomic hybridization (aCGH) attempted the molecular annotation of the iAMP21 region, finding complexity inside it [20, 21]. In addition to the variable length of the relevant region at chromosome 21 across affected individuals, fluorescence in-situ hybridization (FISH) and aCGH analyses revealed that each particular region belonging to a given patient shows different levels of amplification along this region of the chromosome, and also losses of genetic material. Accordingly, RUNX1 was also present in a variable number of multiple copies. FISH probes used to hybridize known regions often binned to unexpected locations, indicating that a series of complex intrachromosomal rearrangements have occurred. Chromosomal inversions were also identified surrounding regions of complex rearrangements. These preliminary approaches provided convincing evidence that iAMP21 abnormality in ALL is extraordinarily complex [20].

Whole-genome sequencing has recently provided a high-resolution perspective of the underlying mechanism to iAMP21. Li et al. [7] characterized the structural rearrangements in patients with sporadic iAMP21-ALL using paired-end mapping data. The analysis revealed fold-back inversions demarcating the breakpoints of the amplified region. As mentioned, fold-back inversion is a rearrangement type that characterizes DNA repair mechanisms involved in BFB DNA cycles [22]. In agreement with this finding, previous studies using cytogenetic and aCGH approaches suggested BFB as the process triggering iAMP21 in ALL [20]. According to the BFB model, telomere dysfunction promotes the emergence of a dicentric chromosome 21 in which the two centromeres are pulled to opposite poles during mitosis, conforming an anaphase bridge that is broken in cytokinesis. After this first BFB cycle, one of the two derived chromosomes bears a duplication of chromosome 21. Li et al. [7] designed FISH probes to match RUNX1 sequences, demonstrating the amplification of this gene along chromosome 21 in all the iAMP21-ALL samples, indicating that RUNX1 would be duplicated after the BFB cycle. In addition, the analysis of rearrangements from the region between both fold-back inversions in the iAMP21 samples revealed clusters of back-and-forth rearrangements, combined with copy-number profiles that oscillate among three states. These features suggested to the authors that chromothripsis occurred after a BFB cycle. According to this model, the multiple copies (4–14) of RUNX1 revealed by FISH in previous cytogenetic studies [20] could be explained by multiple chains of BFB cycles.

Double-minute chromosomes

DMs are small, circular, chromatin bodies that lack centromeres and telomeres and can duplicate autonomously, representing a mechanism for extrachromosomal gene amplification [23]. In cancer, DMs have been identified in different tumor types, including solid [24, 25] and blood-derived [23]. DMs are associated with oncogene amplification. In glioblastoma, these small circular chromosomes are present in about half of all cases, and the epidermal growth factor receptor gene (EGFR) is the most commonly amplified gene [24, 26]. In AML and myelodysplastic syndromes (MDS), DMs are observed in ~1% of cases, most of them consisting of an amplified segment from chromosome 8q24 that always includes C-MYC [27]. In a remarkable example of a chronic myeloid leukemia, Morel et al. [28] identified DMs containing the oncogenic gene fusion BCR-ABL, leading to the amplification (25 copies) of the fused rearrangement.

Several mechanisms could explain the genesis of DMs. The episomal model proposes the excision of a DNA segment followed by its circularization and head-to-tail amplification to generate DMs [29]. Storlazzi et al. [27] obtained evidence in favor of the episome model in hematologic malignancies. They performed cytogenetic analyses on 32 AML and 2 MDS samples with C-MYC-containing DMs. FISH revealed that 68% of the samples with DMs involved C-MYC. In addition, they confirmed that the region of chromosome 8p24 that normally contains C-MYC was deleted in one of the chromosome 8 homologs, suggesting the excision of a DNA segment from the original chromosomal location according to the episome model. They obtained definitive evidence of this model’s validity in one case, where capillary sequencing revealed that the junction of the deletion at 8q24 was in agreement with the sequence junction of the DM amplicon. Similarly, although paired-end sequencing technology is limited in its power to determine whether amplifications are tandem duplications on the same chromosome or circularized as DMs, Yang et al. [4] identified one cancer sample with a deletion whose breakpoints matched the tandem duplication, suggesting an excision of the DNA followed by circularization of that fragment, similar to the excisions of amplified DM fragments described above.

Stephens et al. [5] proposed that chromothripsis could underlie the excision of DNA that would later acquire a circular configuration (Figure 2). In a whole-genome sequencing analysis of a small cell lung cancer cell line, they found evidence of chromothripsis at chromosome 8. As mentioned before, the regions affected by chromothripsis are characterized by low copy number states, but the authors found that within the region with the expected profiles, there were a handful of discrete regions with an intriguing pattern showing massive amplification, with 50–200 copies per cell. Each amplified region was demarcated by rearrangements linking it to other highly amplified segments. They probed by FISH three of the amplified regions from chromosome 8, revealing huge numbers of extrachromosomal copies of the segments. In addition, the probes showed homogeneously staining regions, consistent with chromosomal integration of the DMs. Similarly, Rausch et al. [8] analyzed whole-genome sequencing data from medulloblastoma, finding oncogene amplification among highly amplified segments rearranged by chromothripsis that could be explained by DMs obtained under a catastrophic model. Interestingly, some of the DMs they detected included fragments from different chromosomes. Their analyses of the rearrangement breakpoints are in support of massive DNA double-strand breaks (DSBs), followed by non-homologous end-joining-mediated repair.

Alternatively, Yang et al. [4] argued that DMs observed in cancer may be the consequence of the circularization of DNA fragments that may result from replication errors (by fork stall-ing and template switching [FoSTes]) or by non-homology-based replicative errors. On the other hand, Rausch et al. [8] have...
some objections to the FoSTeS mechanism. Although they do agree that this replication-associated mechanism can also generate alterations with multiple breakpoints, the lack of template insertions at the breakpoint junctions, which are thought to result from abortive attempts to use another template during replication, does not substantiate the involvement of a replication-associated rearrangement mechanism in the complex alterations observed.

Cancer neochromosomes

Contrary to DMs, neochromosomes are circular or linear DNA structures with a centromere, and with telomeres if they display a linear configuration. Neochromosomes occur in about 3% of cancers, being characteristic of some liposarcoma and osteosarcoma. Garsed et al. [6] isolated linear neochromosomes from five liposarcoma cell lines, pooled those belonging to the same cell line and performed DNA sequencing to precisely define the size and structure of the neochromosomes. Extreme structural rearrangements and copy number variation were apparent in the sequencing data. The amount of unique DNA contributed by neochromosomes from the five cell lines varied from 14 to 53 Mb, and derived from different donor chromosomes. However, the estimated size of the chromosomes varied from 150 to 412 Mb (from 279 to 640 Mb including telomeres), indicating that neochromosomes are largely composed by the amplification of donor segments.

Analysis of the rearrangement data carried out by Garsed et al. [6] confirmed that the neochromosomes from liposarcoma cell lines comprise highly amplified and rearranged fragments of DNA from every chromosome in the genome. High-level copy number changes were extremely focal and showed nonlinear distribution of copy number states. Overall, the neochromosomes from the five cell lines showed between 256 and 586 fusions, involving 6–18 different chromosomal partners. Neochromosomes comprise hundreds of spatially distinct, contiguous genomic regions (CGRs). These CGRs are defined as islands in the reference genome, flanked by regions effectively absent in the flow-isolated neochromosome sequence data. CGRs ranged from 88 to 1.15 Mb (median, 26 kb), and were highly rearranged and amplified.

Garsed et al. [6] propose a model for the formation of a neochromosome that consists of three phases. It starts with the initial formation of an episomal, self-replicating, structure. These DM-like chromosomes may be created by chromothripsis. Some features observed in specific regions of the neochromosomes support the idea of chromothripsis: low copy number, oscillating copy number states in the low copy number regions, and clustering of breakpoints. A primordial centromere of alphoid (repetitive) nature is likely to be added at this initial step. By definition, chromothripsis cannot account for the high-level copy number observed along neochromosomes. However, as mentioned above, BFB cycles may account for such amplification. So, the second phase in neochromosome formation involves BFB cycles. The standard model for BFB assumes a linear form of the chromosome but, alternatively, Garsed et al. [6] propose a circular form of BFB and chromothripsis acting over thousands of cell cycles. This alternative BFB mechanism would explain some of the structural features found in neochromosomes that could not have been explained under a linear-BFB model, including the absence of fold-back inversions, a rearrangement type that characterizes linear BFB cycles. During this stage of BFB cycles, new DNA material from other chromosomes may be added to the neochromosomes in circular conformation. BFB cycles lead to the corrosion of the native centromere, which is replaced by a non-alphoid centromere that is formed by sequences already harbored by the neochromosomes, or by new alphoid sequence from other chromosomes in the nucleus. The final phase in neochromosomes formation involves their linearization by the acquisition of telomeres from other chromothriptic chromosomes of the genome.

The neochromosomes analyzed by Garsed et al. [6] contain genes that are amplified to more than 10 copies. The transcriptomic analyses they carried out confirmed amplification of the expression in the amplified genes. However, some amplified genes were found to be transcriptionally silent. Fusion transcripts were also found during analyses of the transcriptomic data. As expected, known liposarcoma oncogenes were within the most amplified (copy number > 20) and overexpressed (Reads Per Kilobase per Million mapped reads > 100) genes, consistent with the principle that genes recurrently amplified confer a selective advantage to the clone.

Chromoplexy in prostate cancer

Chromosomal rearrangements appear to be important in prostate carcinogenesis. About half of all prostate cancers over-express an ETS (E26 transformation specific) transcription factor gene (usually ERG) as a result of a somatic fusion with a constitutively active or androgen-regulated promoter [30, 31]. In addition, disruptive rearrangements frequently inactivate tumor-suppressor genes in prostate tumors, such as the Phosphatase and Tensin Homolog (PTEN) [32].

Chromoplexy is a term used to describe an alternative pattern of complex genomic rearrangements, where rearrangements occur in the context of a ‘chain’ (Figure 3A), in which the two breakpoints from the same rearrangement map to the reference genome near breakpoints from other rearrangements. This phenomenon, discovered in prostate cancer [33], is able to disrupt multiple cancer genes in a coordinated fashion.

Although some patterns of chromoplexy are similar to those in chromothripsis, Baca et al. [33] identified some key hallmarks that define this as a different pattern of structural complexity in cancer genomes. In chromoplexy, the pattern consists of chained structural rearrangements where deletions are often evident near the breakpoints of the regions involved. When these deletions are overlaid with the location of the remaining somatic rearrangements on the reference genome, they create ‘bridges’ that bypass the sequence between two consecutive breakpoints located in the same chromosome. This observation indicates that genomic rearrangements showing DNA loss near a breakpoint may be linked by deletion-bridges to additional rearrangements from the same chain. A probabilistic model for the independent generation of rearrangements across the genome revealed that rearrangements from the same chain are unlikely to arise from independent events, supporting the hypothesis that they occur by a coordinated process. Baca et al. [33] point out that in chromoplexy it is frequent that one single genome bears multiple rearrangement chains, each of them involving a variable number of rearrangements, from less than 10 up to 40, and most of the times weaving more than two chromosomes together (Figure 3B). These features differ from most of the chromothripsis patterns, where typically one-single event generates tens to hundreds of independent (non-chained) rearrangements involving normally just one, sometimes two, chromosomal arms [5].
Baca et al. [33] noted remarkable phenotypic differences in chromoplexy between different prostate cancer types. Tumors bearing oncogenic ETS fusions (ETS⁺) show significantly more interchromosomal rearrangements than ETS⁻, and the analysis of transcriptomic data showed that ETS⁺ chromoplexy breakpoints are enriched near DNA that is highly expressed. In contrast, chromoplexy in ETS⁻ contains 7-fold more rearrangements than ETS⁺, and in addition, are focused in a lower number of chromosomal regions (usually affecting just one or two chromosomes), leading to an excess of intrachromosomal chained-rearrangements with breakpoints enriched in late-replicating regions.

Baca et al. [33] found that in 88% of the tumors, five or more chromosomes are affected by chromoplexy, and 63% of tumors contain two or more chains of rearrangements. They analyzed whole-genome sequencing data from 57 prostate cancer samples and found ~5,600 rearrangements, of which 40% of the rearrangements were components of chains. About 50% of prostate tumors have either deletion or rearrangement of at least one gene in a chain of three or more rearrangements, where PTEN and NKK3-1 are frequently disrupted. In this context, chromoplexy can concomitantly inactivate tumor-suppressor genes that are distant from each other. The authors showed a remarkable example where a chain of 27 rearrangements, involving six different chromosomes, promoted the disruption of the tumor-suppressor gene SMAD4 and the TMPRSS2-ERG fusion, affecting chromosomes 18q and 21q, respectively.

In chromoplexy, the segments involved in the chained rearrangements originate from DNA DSBs. As a result, Baca et al. [33] propose that a progressive acquisition of the different rearrangements over multiple cellular generations seems unlikely to explain the observed pattern of chained rearrangements, because for a chain to occur in a sequential-dependent manner, it is necessary that multiple junctions from ancestral somatic fusions would have to be rebroken precisely and fused to each other to complete a chain. Therefore, chained rearrangements may have occurred simultaneously within the same cell cycle, in an event causing multiple DNA breaks. The mechanism underlying the events causing chromoplexy remains unknown, but Baca et al. [33] propose hypotheses for the two different ETS-type tumors analyzed: In ETS⁺ prostate cancer, where chromoplexy joins DNA from dispersed regions from a higher number of chromosomes and tend to occur near highly expressed loci, the locations of the breakpoints suggest a transcription-associated mechanisms as a cause of DNA damage. The complex chains of rearrangements here could be the consequence of DNA injury at transcriptional hubs occupied by loci from multiple chromosomes. Regarding ETS⁻ tumors, it is important to note that the prostate tumors of this type analyzed by Baca et al. harbored deletions or disruptive rearrangements involving the chromatin-modifying gene CHD1, a putative tumor-suppressor gene that may regulate genome stability. So, Baca et al. propose that the specific patterns of chromoplexy displayed in ETS⁻ tumors may be influenced in some way by the deletion of CHD1.

**RAG-mediated deletions in acute lymphoblastic leukemia**

About one-quarter of childhood ALL cases are characterized by a translocation that creates an ETV6/RUNX1 fusion gene (also known as TEL/AML1) with oncogenic properties [34]. However, the fusion gene is not sufficient on its own to cause this leukemia, making necessary the cooperation of secondary mutations affecting key genes [35]. Recent work revealed an aberrant activity of RAG endonucleases in the promotion of genomic rearrangements that affect critical genes in leukemia transformation [36, 37].

Recombination-activating genes (RAG) encode the enzymes in charge of the recombination that promotes the rearrangements of V(D)J genes during lymphocyte development. These enzymes, called RAG1 and RAG2, are endonucleases that recognize recombination signal sequence (RSS) motifs, consisting of a highly conserved heptamer (CACAGTG) and a less conserved nonamer (ACAAAAACC), separated by a sequence spacer of 12 bp or 23 bp [38]. After binding to these RSS motifs, RAG1/RAG2 form a complex in which the two DNA strands immediately adjacent to each RSS are cleaved and processed by a
Different genome configurations that activate GFI1B variant types. These variable rearrangement types lead to quite the contrary, it is a consequence of diverse structural consequence of a recurrent-single structural variation type; GFI1B activation of the two growth factor independent 1 oncogenes, about one-third of the group 3 and 5–10% of the group 4 states. The enrichment was particularly pronounced for promoter and enhancer regions relative to other chromatin states. The enrichment was particularly pronounced for rearrangements that had an RSS-like motif. For example, deletions with RSS motifs showed 33-fold enrichment of rearrangements that had an RSS-like motif. Notably, the activation of GFI1 variants (including focal deletion, tandem duplication, translocation and complex variants exhibiting inversions and deletions). The analysis of the expression data of the seven genes located in that region revealed a significant upregulation of one of the genes, GFI1B, in samples harboring 9q34 rearrangements relative to the non-rearranged ones. Irrespective of the underlying structural variant type, in about three-quarters of the cases where GFI1B showed upregulation together with rearrangement, the somatic rearrangement repositioned GFI1B close to a region normally located ~370 kb upstream of the GFI1B transcriptional start site. Chromatin immunoprecipitation coupled with sequencing showed that GFI1B was juxtaposed to highly active enhancers.

Northcott et al. [40] had the hypothesis that genes belonging to the same gene family could be playing a similar role in medulloblastoma. Thus, they carried out a detailed analysis of structural rearrangements associated with GFI1, a paralog of GFI1 located in chromosome 1p22, finding an activation of GFI1 in medulloblastoma attributable to similar structural variant mechanisms targeting GFI1B. In this case, interchromosomal translocations represented half of the rearrangements. The rearrangements they found promoted the juxtaposition of the normally repressed GFI1 locus to regions harboring highly active enhancers. Finally, by means of medulloblastoma mouse models, the authors showed that neither the activation of GFI1 nor GFI1B alone is enough to promote tumorigenesis, but they are sufficient to promote tumorigenesis when combined with MYC activity.

Enhancer hijacking in medulloblastoma

Medulloblastoma is a malignant brain tumor originating in the cerebellum, which is classified in four molecular subgroups, namely, Wingless (WT1), SHH, Group 3 and Group 4. An analysis of sequencing data from 137 medulloblastoma genomes carried out by Northcott et al. [40] revealed that, in about one-third of the group 3 and 5–10% of the group 4 medulloblastomas, the disease is driven by mutually exclusive activation of the two growth factor independent 1 oncogenes, GFI1B and GFI1, by means of structural rearrangements. Notably, the activation of GFI1B and GFI1 oncogenes is not the consequence of a recurrent-single structural variation type; quite the contrary, it is a consequence of diverse structural variant types. These variable rearrangement types lead to different genome configurations that activate GFI1B, or its paralog GFI1, by juxtaposing one of both genes to active enhancers located thousands of nucleotides apart, sometimes located even in different chromosomes.

In their study, Northcott et al. [40] observed that a relevant number of cases showed focal somatic rearrangements spanning a region of ~425 kb of chromosome 9q34. The rearrangements observed corresponded to different types of structural variants (including focal deletion, tandem duplication, translocation and complex variants exhibiting inversions and deletions). The analysis of the expression data of the seven genes located in that region revealed a significant upregulation of one of the genes, GFI1B, in samples harboring 9q34 rearrangements relative to the non-rearranged ones. Irrespective of the underlying structural variant type, in about three-quarters of the cases where GFI1B showed upregulation together with rearrangement, the somatic rearrangement repositioned GFI1B close to a region normally located ~370 kb upstream of the GFI1B transcriptional start site. Chromatin immunoprecipitation coupled with sequencing showed that GFI1B was juxtaposed to highly active enhancers.

Proto-oncogenes activated by Igh regulatory elements in lymphoma

B-cell malignancies often contain translocations that link proto-oncogenes to the Immunoglobulin Heavy Chain (Igh) locus located at chromosome 14. Remarkable examples are the BCL-1/Igh translocation in mantle cell lymphoma, the BCL-2/Igh translocation in follicular lymphoma and the C-MYC/Igh translocation in Burkitt’s lymphoma [41]. The frequent involvement of immunoglobulin loci in chromosomal translocations reflects the remodeling of Igh genes that occur in B cell development [41, 42]. Igh-translocations involve the formation of DSBs in the DNA, which are initiated either by RAG1/RAG2 enzymes during V(D)J recombination in early stages of B-cell development, or by the enzyme cytidine deaminase during class-switch recombination, a process that switches the constant (C) genes of the Igh locus in mature B-cells. In most of the interchromosomal rearrangements, the translocated partner gene becomes deregulated by the action of regulatory elements in Igh. For example, in Burkitt’s lymphoma, about 80% of the patients show the translocation t(8;14)(q24;q32) [43], which juxtapose the proto-oncogene C-MYC, a major cellular factor for cellular growth located at human chromosome 8, and the Igh locus [44]. Several studies on transgenic mice evidence that the new rearranged configuration places C-MYC under the control of 3’ regulatory elements of Igh (the immunoglobulin heavy chain gene intronic enhancer, Eenh, and a set of regulatory elements lying downstream of the Igh locus, namely, Igh 3’ regulatory regions or Igh3’RR [45]), deregulating C-MYC, which promotes cellular transformation [46-48].

By constructing transgenic mice bearing different forms of C-MYC, Adams et al. [47] established that subjugation of C-MYC to immunoglobulin enhancers promotes leukemogenesis in B-lymphoid cells. One of the constructions, where the Eenh enhancer was located upstream of the C-MYC exon 1 (simulating the structure previously observed in a murine plasmacytoma), showed potent oncogenic activity, as it induced lymphoma development in 86% of the cases. Similarly, Truffinet et al. [46] generated mice harboring a single copy of a transgene with c-myc followed by 3’ Igh locus regulatory region (Igh3’RR). Interestingly,
mice developed lymphomas within a few weeks. In 75% of the cases, proliferations were Burkitt lymphoma-like, and the remaining 25% cases were characterized as diffuse anaplastic lymphoma. The transcriptomics analyses showed elevated levels of C-MYC expression in premalignant transgenic mice and Burkitt-like lymphomas relative to non-transgenic controls. The results of Truffinet et al. [46] provided evidence that the Igh3'RR acting alone strongly deregulates C-MYC and mimics the critical features promoting human Burkitt’s lymphoma. Further investigation in mice suggests that Igh3'RR may confer long-range oncogenic activation of Igh-C-MYC translocations in human B-cell lymphomas (for instance, Burkitt’s lymphoma) where Igh-C-MYC translocations delete E4 enhancer and place C-MYC up to 200 kb upstream of the Igh3'RR [47, 48]. Given the similar organization of the mouse and human Igh locus, these findings suggest that Igh regulatory regions also support activated oncogene expression in human B-cell tumors with Igh translocations.

Somatic retrotransposition

Broadly, TEs are defined as DNA sequences that have the ability to move in a host genome. Some of them are autonomous, meaning that they encode for all the protein machinery necessary for their own mobilization, whereas others are non-autonomous, requiring the protein machinery encoded by other TEs that allow their mobilization. The mechanism of TEs mobilization is called transposition and, depending on the mode of transposition, TEs are generally classified in two categories: DNA transposons, which mobilize by a ‘cut and paste’ mechanism, and retrotransposons, which are mobilized by a ‘copy and paste’ mechanism that involves an intermediary mRNA, which is retrotranscribed prior to the integration of the new TE copy into a new location of the genome [49]. TEs represent at least 45% of the human genome and, although our genome bears millions of TE copies, most of them are remnants of ancient activity no longer capable of transposition [50]. Only a small set of copies, from the group of retrotransposons, are still active in our genome. These active loci belong to the families LINE1 (L1), Alu, SVA and some endogenous retroviruses (ERVs). While L1 and ERVs are autonomous elements, Alu and SVA require the enzymes encoded by L1 for their retrotransposition.

In cancer, L1 is the most frequently mobilized TE type [51–53]. Although our genome contains about half a million of L1 elements, only a handful of these L1 loci have demonstrated somatic activity in cancer genomes [51]. Among cancer types, retrotransposons showed higher transposition rates in lung and colorectal cancers. In some cancer genomes, somatic retrotransposition accounts for hundreds of events, being the most frequent rearrangement type of the structural variation landscape.

Despite their high activity in certain tumor types, there is some controversy on the effective impact of TE insertions in cancer genome function. Most somatic integrations of TEs occur in late replicating regions of the cancer genome (mostly heterochromatic) and, when integration occurs within the boundaries of euchromatic genes, these genes are usually not expressed in the relevant tissue. In general, these data would suggest that retrotransposons may not have a significant impact on gene expression in cancer [51, 53]. On the other hand, there are some studies showing that somatic TE integration alters the expression of target genes in cancer [52, 53], and that some somatic mobilizations may have a relevant impact in oncogenesis [54].

Interestingly, apart from their own mobilization, L1 elements are capable of promoting the mobilization of genomic sequences in cis and trans [55, 56], increasing the number of ways by which these TEs can impact the structure and function of the cancer genome.

Cis-mobilization of DNA sequences by L1

Although the standard transcription of L1 is restricted to the length of the repeat, being initiated at the 5’-UTR and finishing at a 3’-poly(A) tail, sometimes transcription can overpass the poly(A) tail of the element, finishing in an alternative poly(A) sequence located downstream. When this abnormal transcription occurs, unique (non-L1) DNA sequences located downstream to L1 are picked up by the mRNA, retrotransposed together with the L1 element and integrated elsewhere into the genome (Figure 4). This mechanism, called 3’-transduction mediated by L1 [56], has been recently rediscovered as a new mutational process that occurs naturally in cancer genomes [51, 53, 57]. Tubio et al. [51] analyzed the mechanism of L1 3’-transductions in whole-genome sequencing data from 244 cancer patients from 12 cancer types. They found that 3’-transductions represent about one quarter of the total mobilizations of L1 and, as expected, transductions are particularly frequent in some lung cancer genomes. Although transductions usually mobilize genomic regions lower than 1 kb long, the mobilized regions can be located as far as 12 kb from the original L1 copy. The process has proven the ability to promote exon shuffling, duplication of small genes and the mobilization of DNA sequences with regulatory potential in primary tumors [51].

Figure 4. The mechanism of L1-mediated transductions. (A) Transcription of L1 can overpass the poly(A) tail of the element, finishing in an alternative poly(A) sequence located downstream. When this abnormal transcription occurs, unique (non-L1) DNA sequences located downstream to L1 are picked up by the mRNA, retrotransposed together with the L1 element and integrated elsewhere into the genome. (B) The Circos plot shows tens of transductions mediated by a L1 germline master copy located at chromosome 6p24 (see reference [51]).
Trans-mobilization of pseudogenes by L1

Genes located far away from active L1 elements, even in different chromosomes, can be mobilized by the action of L1 in trans [55]. The enzymes encoded by active L1 loci may mediate the retrotransposition of other genes of the genome. This process, which was well-documented in the germline, has also demonstrated to occur somatically during cancer evolution [58, 59]. When a gene is retrotransposed by this mechanism, the final product after integration shows distinctive features of pseudogenes, including the presence of poly(A) tails, target-site duplications at the integration point and absence of intronic sequence. Cooke et al. [58] have analyzed the mechanism of somatic acquisition of processed pseudogenes using next-generation sequencing data from 660 cancer genomes from multiple cancer types. They found that somatically acquired pseudogenes are present in 2.6% of all cancers. As expected, the formation of processed pseudogenes is found in tumor types where L1 shows high activity rates, essentially lung and colorectal cancers. Their analysis of RNA sequencing data revealed evidence of expression of pseudogenes when these are inserted in the 3'UTRs of other genes. They also demonstrated that pseudogene integration could alter the expression of target genes.

Focal amplification of host DNA mediated by human papillomavirus

Although the link between viruses and certain cancer types remains controversial [60], viral infection has been recognized as one of the multiple causes driving oncogenesis [61]. During infection, full or partial sequences of the viral DNA could be integrated into the nucleus DNA of the infected cell, a mechanism of mutation called insertional mutagenesis. This is a necessary stage of the ‘life’ cycle of some viruses, such as retroviruses. However, the host nuclear DNA also contains DNA from other viruses that, in principle, do not need to integrate their nucleic acid to complete their cycle. The group of DNA viruses called papillomaviruses illustrates one such example.

HPV is the principal cause of human cervical cancer, and some anogenital and head-and-neck cancers. Tang et al. [62] examined RNA-sequencing data encompassing 4,433 cancers and 19 tumor types. They confirmed that HPV is present in 97% of 87 cervical carcinoma samples, identifying 12 different HPV types, with HPV-16 and HPV-18 being predominant (65 and 13% of cases, respectively). Their study also revealed a 14% HPV association in head-and-neck cancers, with 87% and 14% of positive tumors attributed to HPV-16 and HPV-33, respectively. Although the ability of HPV to cause these cancers is mainly via the expression of the viral oncoproteins E6 and E7, it seems that other secondary genetic events are necessary for oncogenic transformation in vivo [63]. Some studies have shown that the insertion of HPV DNA has a deep impact in the structure and function of the host genome, suggesting that HPV integration may be behind the critical events that help to transform the cell [64, 65].

Tang et al. [62] identified HPV integration events into cancer genomes by means of the detection of host-virus fusion transcripts. They found a high integration frequency for HPV-18 (100%) and a lower frequency for HPV-16 (58.5%), the two HPV types recognized as cancer-causing agents [66]. According to the authors, although HPV integration seems to be widespread across the genome, it is observed recurrent integration in genomic regions involving key genes.

From the point of view of the structural variation, the integration of HPV may have higher-level implications in remodeling the structure of the host genome, other than the simple structural modification at the integration site. In cervical and head-and-neck cancers, there is striking association between HPV insertion events with adjacent host genomic structural variants, especially DNA amplifications. Previous studies, based on single nucleotide polymorphism arrays, pointed out a significant association between HPV insertions and regional copy number variants in cervical cancers [65]. The analysis of whole-genome sequencing data by Akagi et al. [64] has confirmed this pattern, showing that in about half of the HPV insertions, the DNA of the virus flanks and bypasses amplifications of neighboring host DNA.

Akagi et al. [64] found that, in some samples, HPV integration sites detected by whole-genome sequencing were less numerous than the estimated number of viral copies. The analysis of the read-coverage at the chromosomal regions surrounding the HPV integration sites showed focal amplification of neighboring DNA, ranging from ~1.5 to ~58-fold increase, indicating that the discrepancy they previously observed, relative to viral copy numbers and number of integration breakpoints, was due to the amplification of viral inserts together with the flanking genomic sequences, leading to redundant-identical breakpoints. They proposed a looping model to explain the formation of these patterns. This model suggests that the transient formation of looped-circular structures [Figure 5], composed of integrated HPV DNA and their adjacent host DNA sequences, would allow the replication of the enclosed sequences as a rolling circle. Under this model, the replication is mediated by host DNA polymerases starting their activity at the viral origin of replication. In addition, statistical analyses showed a strong association of HPV integration sites with other intra- and inter-chromosomal structural rearrangements, especially at these highly amplified regions. The resulting genomic alterations frequently disrupt the expression and structure of neighboring genes, some of them involved in oncogenesis, and result in the amplification and expression of viral oncoproteins E6 and E7 as viral–host fusion transcripts [64, 67, 68].

The recent analysis of next-generation sequencing data in bladder carcinoma by The Cancer Genome Atlas network [69] provided further evidence on a potential role of HPV in the amplification of cancer genes. The researchers found one integration event of HPV-16 in the apoptosis-regulating gene BCL2L1. The gene was amplified (~6X) in the tumor and, consequently, overexpressed. Overall, these findings indicate that HPV integration may have a role in the development of other cancers.

It is remarkable that similar structural patterns are found in the vicinity of hepatitis B virus (HBV) integration sites in hepatocellular carcinoma genomes. In HBV-infected genomes, there is a significant association between viral integration sites and boundaries of copy number alterations, which suggests an underlying mechanistic connection between HBV integration and genome instability [70].

Future directions and conclusions

Although next-generation sequencing technology has provided an unprecedented understanding of the structural processes involved in cancer, current genomic studies have important limitations. Some structural variation patterns
described in this review provide evidence that somatic genomic rearrangements may have unexpected effects on the evolution of a cancer clone, as rearrangements sometimes involve regulatory regions of uncertain or unknown function. These examples demonstrate that genomic approaches must be coupled with transcriptomics, epigenomics and chromatin immunoprecipitation sequencing, to have a complete understanding of the potential effects that somatic genomic configurations may have on genome function. The main limitation of the current next-generation sequencing technology is the short length of the sequencing reads used to detect structural variation. Thus, sequencing technologies are now moving toward an increase in sequencing read length without a significant increase in sequencing cost. Another limitation of current structural variation studies in cancer genomes is the relatively small number of samples, and limited number of different tumor types that have been analyzed. In the near future, we will need to survey structural variation on a considerably larger scale, across thousands of cancer genomes from each cancer type. Our understanding of the patterns of structural variation in cancer and their underlying mechanisms is also limited by the fact that somatic rearrangements are detected in bulk tumor tissue, thus overlooking subclonal structural variation. Single-cell sequencing will provide us with a new level of knowledge of the structural variation mutational processes acting in cancer.

Key points

• There is variation across cancer genomes in number of somatically acquired structural rearrangements, and relative proportion of different rearrangement types.
• Different mutational processes underlying somatic structural variation show different prevalence across cancer types.
• Although most rearrangements generated somatically are neutral, some are able to drive cancer origin and/or development.
• The recent discovery of new patterns of somatically acquired structural variation in cancer provides new insight into the mutational processes underlying specific cancer types.
• Although next-generation sequencing technology has provided an unprecedented understanding of the mutational processes involved in cancer genome structural variation, current genomic studies have important limitations that require future improvements.

Acknowledgements

JMCT wants to thank the following colleagues: Peter J. Campbell, Elizabeth P. Murchison, Anthony Fullam, Yilong
Li, Gunes Gundem, Eli Papaemmanuil, Nicola Roberts, Peter J. Park, Lixing Yang, Young Seok Ju and Jan Korbel.

Funding

J.M.C.T. is supported by a Marie Curie Fellowship FP7-PEOPLE-2012-IEF (project number 328264).

References