Review

The role of microhomology in genomic structural variation

Diego Ottaviani*, Magdalena LeCain*, and Denise Sheer

Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK

Genomic structural variation, which can be defined as differences in the copy number, orientation, or location of relatively large DNA segments, is not only crucial in evolution, but also gives rise to genomic disorders. Whereas the major mechanisms that generate structural variation have been well characterised, insights into additional mechanisms are emerging from the identification of short regions of DNA sequence homology, also known as microhomology, at chromosomal breakpoints. In addition, functional studies are elucidating the characteristics of microhomology-mediated pathways, which are mutagenic. Here, we describe the features and mechanistic models of microhomology-mediated events, discuss their physiological and pathological significance, and highlight recent advances in this rapidly evolving field of research.

Microhomology as a mutational signature

Large-scale population studies, such as the '1000 genomes project', indicate that genomic structural variation is a major source of genetic diversity among individuals and populations [1,2]. Structural variation typically involves genomic segments over 100 bp in length and includes tandem duplications, insertions, and inversions, which can generate DNA copy number variants (CNVs), as well as translocations and complex rearrangements [3]. These ambitious research efforts have revealed that structural variants are abundant, and should be considered as important as single nucleotide polymorphisms (SNPs) and single nucleotide variation [1,4].

Germline structural variation can be phenotypically neutral, having no effect on the organism. However, if it affects gene expression, structural variation can have a significant impact on the fitness of an individual [5] by conferring disease susceptibility, giving rise to human disorders [6] or leading to traits that can be selected for if beneficial [5,7]. For example, the copy number of the human salivary amylase gene, *AMY1*, is higher in populations with high starch diets, where the increased amylase protein levels are likely to improve the digestion of starchy foods [8]. In somatic cells, genomic structural variation is also significant because it is a key mediator of neoplastic transformation and progression of cancer [9]. Furthermore, somatic structural variation has a normal physiological role at immunoglobulin gene loci, where it is essential for generating antibody diversity [10].

Recent high-resolution sequencing studies of germline and somatic rearrangement breakpoints have revealed molecular signatures that enable reconstruction of mutational mechanisms [11–13]. For example, blunt joins, or small insertions or deletions at the breakpoint junction, are characteristic of DNA double-strand break (DSB) repair through direct ligation by nonhomologous end joining (NHEJ), whereas long stretches of sequence homology at or near the breakpoint can be attributed to homologous recombination (HR). HR repairs DSBs using template sequences, and relies on the presence of DNA segments sharing extremely high similarity or identity.

Glossary

Class switch recombination: recombination event in mature B lymphocytes that generates immunoglobulin isotypes with different effector functions, switching from IgM or IgD to IgG, IgE, or IgA following an immune response.
Complex genomic rearrangements: rearrangements with two or more breakpoint junctions.
Flanking microhomology: microhomology adjacent to the junction of a genomic rearrangement but not overlapping it.
Genomic disorder: pathological phenotype resulting from structural rearrangements in genomic loci where architectural features render the genome unstable.
Junctional microhomology: microhomology occurring directly at a breakpoint junction of a genomic rearrangement. Given that the sequence is identical in each of the genomic segments that contribute to the rearrangement, it is not possible to identify the exact breakpoint, because the microhomology cannot be assigned to either of the respective segments.
Low processivity polymerase: a polymerase that incorporates a relatively low number of nucleotides before it dissociates.
Microhomology: two short DNA sequences that are identical.
Nonrecurrent genomic rearrangements: rearrangements with variable breakpoints at sites lacking extensive sequence homology.
Recurrent genomic rearrangements: rearrangements of the same genomic interval occurring repeatedly in multiple unrelated individuals, found at sites of extensive sequence homology.
Replication fork collapse: breakage of the replication fork and detachment of one arm.
Replication fork stalling: an abnormality arising during DNA replication, where DNA synthesis at the replication fork pauses. This can arise from low levels of DNA polymerase or nucleotides, or from the fork encountering a barrier, such as complex DNA architecture.
Single-ended DSB: a DSB with only one end, which can arise from telomere erosion, separation of partner strands of a DSB, or when progression of a replication fork is interrupted.
Structural variation: genomic insertions, duplications, or deletions, which are collectively termed 'CNVs', or translocation or inversion of segments of the genome.

Corresponding author: Sheer, D. (d.sheer@qmul.ac.uk).
Keywords: microhomology; double-strand breaks; structural variation; genomic disorders; microhomology-mediated end joining; microhomology-mediated break-induced replication.
*These authors contributed equally to this article.

0168-9525/$ – see front matter
 http://dx.doi.org/10.1016/j.tig.2014.01.001

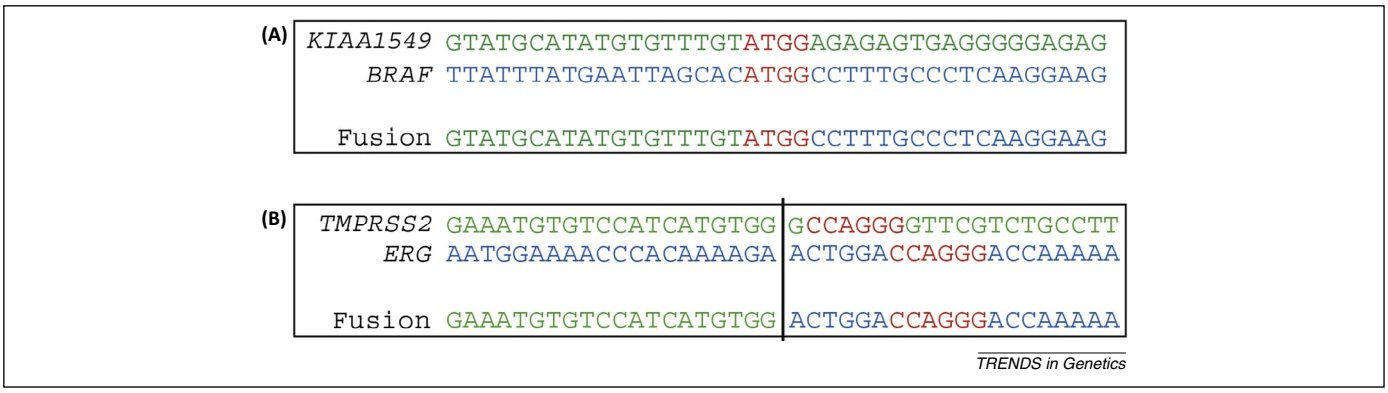

Figure 1. Microhomology at breakpoint junctions and flanking regions of simple gene fusions. **(A)** Junctional microhomology (red) at a *KIAA1549–BRAF* gene fusion in a paediatric low-grade astrocytoma. The exact breakpoint in each of the partner genes cannot be determined at a nucleotide level because the microhomology is present in both segments. **(B)** Flanking microhomology (red) at a *TMPRSS2–ERG* gene fusion in prostate cancer. The breakpoint is indicated by the black vertical line. Abbreviations: *BRAF,* v-raf murine sarcoma viral oncogene homolog B; *ERG,* serine 2- v-ets erythroblastosis virus E26 oncogene homolog; *TMPRSS2,* transmembrane protease. Adapted from [16] (A) and [18] (B).

These sequencing studies have also revealed short regions of DNA sequence homology, called 'microhomology' (see Glossary), at certain germline and somatic breakpoint junctions (e.g., [11,14,15]). Although definitions of breakpoint microhomology vary with respect to the length of the homologous region, it can be defined as a series of nucleotides (<70) that are identical at the junctions of the two genomic segments that contribute to the rearrangement (Figure 1A, Figure S1 in the supplementary material online). Microhomology has also been reported in DNA sequences that are adjacent to, but do not overlap, breakpoint junctions [16–18] (Figure 1B).

There is now evidence for additional repair mechanisms, besides the prevalent NHEJ and HR, that result in structural variation through the use of sequence microhomology. Whereas junctional microhomology of 1–4 bp can be a feature of NHEJ [19], as discussed below, one of these alternative mechanisms, termed 'microhomology-mediated end joining' (MMEJ), is independent of key proteins involved in NHEJ (Figure 2) [20,21]. MMEJ is error prone and frequently produces genomic rearrangements [22]. Further alternative mechanisms, termed 'fork stalling and template switching' (FoSTeS) and 'microhomology-mediated break-induced replication' (MMBIR), involve erroneous DNA replication, and template switching facilitated through annealing of microhomologous sequences [23,24]. These replicative mechanisms have been proposed to account for complex rearrangements that have multiple breakpoint junctions, insertions of DNA segments mapped to different genomic regions, as well as breakpoint microhomology, which together form a molecular signature inconsistent with NHEJ and HR. In this review, we examine the proposed molecular basis and regulation of these microhomology-mediated DNA repair mechanisms, and discuss their biological significance.

Microhomology-mediated end joining

DNA DSBs, which can be caused by a variety of agents, including reactive oxygen species, ionising radiation and UV light, are important mediators of structural variation [25]. The major repair mechanisms for DNA DSBs are NHEJ and HR [25]. NHEJ directly ligates broken DNA strands and is active throughout the cell cycle, although it predominates during the G0 and G1 phases. This repair pathway can lead to blunt joins, or small insertions or deletions at the breakpoint junction (reviewed in [19]). HR can lead to faithful DNA repair of DSBs during the S and G2 phases of the cell cycle, when a sister chromatid is available to serve as a template, but is often mutagenic during G1, when it relies on alternative homologous sequences, such as repetitive elements.

More recently, a further DSB repair pathway has been described that is thought to serve as a back-up repair process. MMEJ, which is sometimes referred to as an alternative NHEJ pathway, relies on the recombination of short stretches of microhomology for repair of DSBs [22]. Although understanding of MMEJ is still incomplete, it is emerging that this pathway can support DNA repair throughout the cell cycle [22], and shares elements with both HR and NHEJ.

MMEJ has been studied most extensively in yeast cells, where the mechanism was originally characterised, although mammalian functional orthologues have since been identified for most proteins (Table 1) [22]. As in HR, the essential initial step for MMEJ repair in mammals is digestion of the 5′ DNA strand by the Mre11–Rad50–Nbs1 (MRN) complex in association with retinoblastoma binding protein 8 (CtIP) to obtain a 3′ single-stranded DNA tail [26–28] (Figure 2). This occurs on each of the DNA strands beside the break. The exposed microhomologous sequences on the complementary 3′ ends then anneal to form a complex with gaps that need to be filled and ligated. The overhanging noncomplementary 3′ flaps are then trimmed by the endonuclease excision repair cross-complementing rodent repair deficiency, complementation group 4–excision repair cross-complementing rodent repair deficiency, complementation group 1 (Xpf–Ercc1) complex, whereas the gaps created on both strands through resection are filled in by a DNA polymerase, which has been proposed to be polymerase lambda [29,30]. DNA ligase I and ligase IIIa/X-ray complementing defective repair in Chinese hamster cells 1 (Xrcc1) are responsible for the subsequent joining of the DNA segments [31,32]. Given that MMEJ results in deletions of the DNA regions flanking the original break, it is an error-prone repair pathway. The mechanistic model of MMEJ, as it is