Seminars in Cancer Biology 20 (2010) 222–233

Contents lists available at ScienceDirect

Seminars in Cancer Biology

journal homepage: www.elsevier.com/locate/semcancer

Review

Genomic rearrangements in inherited disease and cancer

Jian-Min Chen[a,b,c,*], David N. Cooper[d], Claude Férec[a,b,c,e], Hildegard Kehrer-Sawatzki[f], George P. Patrinos[g]

[a] Etablissement Français du Sang (EFS) – Bretagne, Brest, France
[b] Institut National de la Santé et de la Recherche Médicale (INSERM), U613, Brest, France
[c] Faculté de Médecine et des Sciences de la Santé, Université de Bretagne Occidentale (UBO), Brest, France
[d] Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK
[e] Laboratoire de Génétique Moléculaire et d'Histocompatibilité, Centre Hospitalier Universitaire (CHU), Hôpital Morvan, Brest, France
[f] Institute of Human Genetics, University of Ulm, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[g] University of Patras, School of Health Sciences, Department of Pharmacy, University Campus, Rion GR-26504, Patras, Greece

ARTICLE INFO

Keywords:
Cancer
Copy number variation
Genomic rearrangements
Homologous recombination
Non-homologous end joining

ABSTRACT

Genomic rearrangements in inherited disease and cancer involve gross alterations of chromosomes or large chromosomal regions and can take the form of deletions, duplications, insertions, inversions or translocations. The characterization of a considerable number of rearrangement breakpoints has now been accomplished at the nucleotide sequence level, thereby providing an invaluable resource for the detailed study of the mutational mechanisms which underlie genomic recombination events. A better understanding of these mutational mechanisms is vital for improving the design of mutation detection strategies. At least five categories of mutational mechanism are known to give rise to genomic rearrangements: (i) homologous recombination including non-allelic homologous recombination (NAHR), gene conversion, single strand annealing (SSA) and break-induced replication (BIR), (ii) non-homologous end joining (NHEJ), (iii) microhomology-mediated replication-dependent recombination (MMRDR), (iv) long interspersed element-1 (LINE-1 or L1)-mediated retrotransposition and (v) telomere healing. Focussing on the first three of these general mechanisms, we compare and contrast their hallmark characteristics, and discuss the role of various local DNA sequence features (e.g. recombination-promoting motifs, repetitive sequences and sequences capable of non-B DNA formation) in mediating the recombination events that underlie gross genomic rearrangements. Finally, we explore how studies both at the level of the gene (using the neurofibromatosis type-1 gene as an example) and the whole genome (using data derived from cancer genome sequencing studies) are shaping our understanding of the impact of genomic rearrangements as a cause of human genetic disease.

1. Introduction

Genomic rearrangements constitute changes in the genetic linkage relationship of discrete chromosomal fragments and can involve deletions, duplications, insertions, inversions or translocations. Historically, genomic rearrangements have been extensively studied by means of either classical cytogenetic or molecular biological techniques. Only fairly recently has the resolution gap between these techniques been bridged by technological advances. Since two landmark studies six years ago [1,2], genomic rearrangements of intermediate scale—now commonly known as copy number variation (CNV; a ≥1 kb DNA segment that differs in terms of its copy number with respect to a reference genome sequence [3])—have been found in increasing numbers to cause or predispose to human inherited disease and cancer. An increasing number of rearrangement breakpoints have been characterized at the nucleotide sequence level, thereby providing an invaluable resource for the detailed study of mutational mechanisms underlying genomic recombination events. A better understanding of these mutational mechanisms is vital for improving the design of mutation detection strategies. In this article, we shall provide an overview of the mutational mechanisms put forward to account for the diverse range of known genomic rearrangements, with an emphasis on new insights generated from recent

Abbreviations: aCGH, array comparative genomic hybridization; BIR, break-induced replication; CNV, copy number variation; CNM, copy number mutation; D-loop, displacement loop; DSB, double-strand break; FoSTes, fork stalling and template switching; HJ, Holliday junction; LCRs, low copy repeats; LINE-1 or L1, long interspersed element-1; NAHR, non-allelic homologous recombination; NHEJ, non-homologous end joining; MMRDR, microhomology-mediated replication-dependent recombination; ROHs, runs of homozygosity; SRS, serial replication slippage; SSA, single-strand annealing; ssDNA, single-stranded DNA.

* Corresponding author at: Etablissement Français du Sang (EFS) – Bretagne, 46 rue Félix Le Dantec, 29218 Brest, France. Tel.: +33 2 9801 8102; fax: +33 2 9843 0555.
E-mail address: jian-min.chen@univ-brest.fr (J.-M. Chen).

1044-579X/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.semcancer.2010.05.007

studies of both inherited disease and cancer, and highlight the most significant findings obtained from cancer genome sequencing studies.

2. Mutational mechanisms of genomic rearrangement

At least five categories of mutational mechanism can give rise to genomic rearrangements: homologous recombination, non-homologous end joining (NHEJ), microhomology-mediated replication-dependent recombination (MMRDR), long interspersed element-1 (LINE-1 or L1)-mediated retrotransposition, and telomere healing. The latter two can perhaps be described as specialized mechanisms as compared with the first three. L1-dependent retrotransposition is thought to occur by target site-primed reverse transcription. Besides simple self-insertion, L1 elements can mobilize their 5′- and 3′-flanking DNA sequences *in cis* and non-autonomous sequences *in trans* (e.g. *Alu* sequences) to new genomic locations. Moreover, L1 retrotransposition can also give rise to large genomic deletions (for reviews, see [4–6]). Telomere healing refers to a process during which the end of a broken chromosome is stabilized by the telomerase-dependent addition of telomeres at non-telomeric sites (reviewed in [7]). In this section, we shall focus on the first three general mechanisms. We shall attempt to compare and contrast their characteristic hallmarks, emphasize new developments, and discuss the role of various local DNA sequence features in mediating gross genomic rearrangements.

2.1. Homologous recombination

Homologous recombination is one of the major pathways for the repair of double-strand breaks (DSBs). As the term implies, it is mediated through sequences which exhibit considerable homology (generally >200 bp) that presumably serves to stabilize chromosomal mispairing. Homologous recombination is upregulated in the S and G2 phases of the cell cycle, when sister chromatids are readily available. It can be further sub-divided into four pathways, namely, non-allelic homologous recombination (NAHR), gene conversion, break-induced replication (BIR) and single-strand annealing (SSA) (Fig. 1). These pathways share similar initiating events: the DSB generated within one of the duplicated or repeated sequences undergoes extensive 5′-end resection to form 3′ single-stranded DNA (ssDNA) tails; these tails, once coated with the Rad51 recombi-

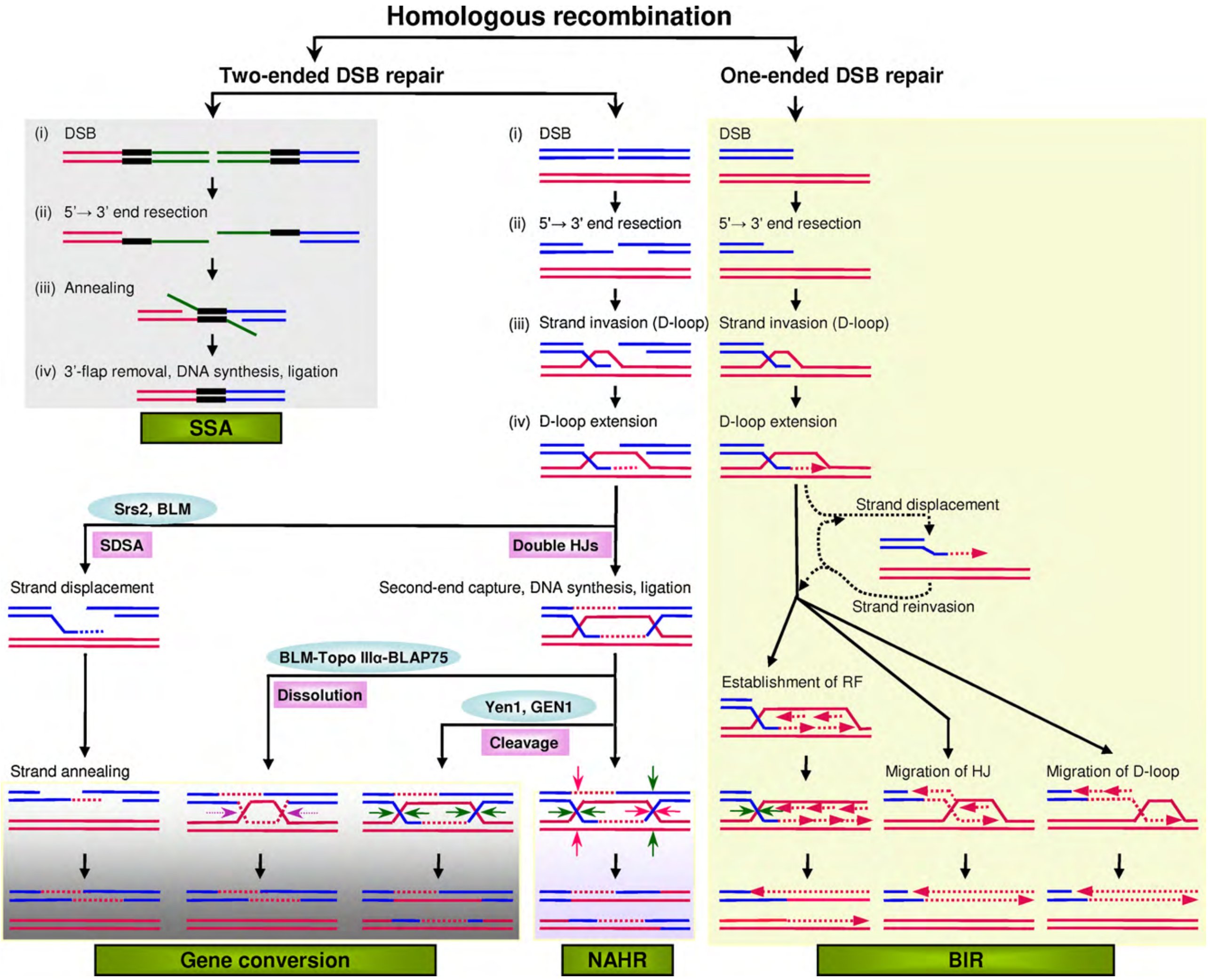

Fig. 1. Mutational models of homologous recombination. In the models of gene conversion, NAHR (non-allelic homologous recombination) and BIR (break-induced replication), the invading strand invariably binds to a homologous sequence. In the model of SSA (single-strand annealing), the black bars indicate the direct repeats that flank a DSB (double-strand break). In the dissolution model of gene conversion, the two facing horizontal purple arrows indicate convergent branch migration. In the double HJs (Holliday junctions) cleavage model of gene conversion, the four horizontal green arrows indicate the orientation of resolution. In the double HJ cleavage model of NAHR, the double HJs can be cleaved as indicated by the green arrows or by the red arrows. In the first pathway of BIR, the HJ is resolved as indicated by the facing horizontal green arrows. See text for details. D-loop, displacement loop; RF, replication fork; SDSA, synthesis-dependent strand annealing.

nase protein, search for and base-pair with a homologous sequence [8].

2.1.1. NAHR

2.1.1.1. Mechanistic model of NAHR. NAHR is a type of two-ended DSB repair. One of the ssDNA tails invades the non-allelic homologous DNA duplex forming a displacement (D)-loop, which is then extended by DNA synthesis. The other 3′ ssDNA tail of the DSB is then captured, with DNA synthesis and ligation of nicks leading to the formation of double Holliday junctions (HJs). (HJ refers to a mobile junction between four strands of DNA. It is named after Robin Holliday, who originally proposed it back in 1964.) Finally, cleavage of the double HJs by a HJ resolvase, GEN1 in humans and Yen1 in yeast [9], gives rise to either a crossover (i.e. NAHR) or a non-crossover (i.e. gene conversion) event, depending on the orientation of HJ cleavage (Fig. 1).

2.1.1.2. Types of genomic rearrangements caused by NAHR. NAHR can result in deletion, inversion, duplication or translocation, depending upon the location and orientation of the interacting homologous sequences (Fig. 2).

2.1.1.3. Meiotic NAHR vs mitotic NAHR: genomic rearrangements at the NF1 locus as a model system. NAHR is the major mechanism leading to recurrent pathogenic CNVs and occurs in both meiotic and mitotic cells (for a review, see [10]). Whereas both meiotic and mitotic NAHR events could be mediated through the same pairs of low copy repeats (LCRs) [11], the study of the different types of gross rearrangement at the type 1 neurofibromatosis (*NF1*) locus has yielded some unexpected findings. Large deletions in 17q11.2 that encompass the *NF1* gene and its flanking regions constitute the most frequently recurring mutations causing NF1, an inherited tumour predisposition syndrome. Three recurrent subtypes of these gross *NF1* deletions have been noted that differ both in terms of deletion size and the positions of their respective breakpoints: type-1, type-2 and type-3 *NF1* deletions (Fig. 3). The most common of these are type-1 deletions which encompass 1.4 Mb and lead to the loss of 14 genes including the *NF1* gene. Type-1 deletions are mediated by NAHR between LCRs flanking the *NF1* gene region, specifically NF1-REPs A and C. Two preferred regions of NAHR have been noted within the NF1-REPs: the paralogous recombination sites PRS2 and PRS1 [12,13]. Of 60 type-1 deletions investigated, 40 had breakpoints within a 3.4 kb region spanning PRS2 whereas 13 had breakpoints within PRS1, a region encompassing 1.8 kb [13]. Thus, hotspots of NAHR clearly occur within NF1-REPs A and C. Notably, NAHR underlying type-1 deletions occur preferentially during maternal meiosis [14,15], contrasting with the overwhelming occurrence of CMT1A duplications at 17p11.2 in spermatogenesis [16,17]. Type-3 *NF1* deletions are mediated by NAHR between NF1-REPs B and C; they encompass only 1 Mb and are much less frequent than type-1 deletions since only three patients with germline type-3 deletions have so far been identified [18]. Nevertheless, both type-1 and type-3 deletions serve to demonstrate the high recombinogenic potential of the LCRs in the *NF1* gene region.

Type-2 deletions constitute the second most common type of recurrent gross *NF1* deletion. They span 1.2 Mb and are characterized by breakpoints within the *SUZ12* gene and its pseudogene (*SUZ12P*) respectively which flank the NF1-REPs (Fig. 3). Type-2 deletions lead to the loss of only 13 genes since, in contrast to type-1 deletions, the functional *LRRC37B* gene within the distal NF1-REP C is retained in type-2 deletions. Steinmann et al. [19] performed a comprehensive breakpoint analysis of 13 type-2 deletions but did not detect any obvious hotspots of NAHR that would be confined to only a few kilobase pairs. An over-representation of polypyrimidine/polypurine tracts and triplex-forming sequences

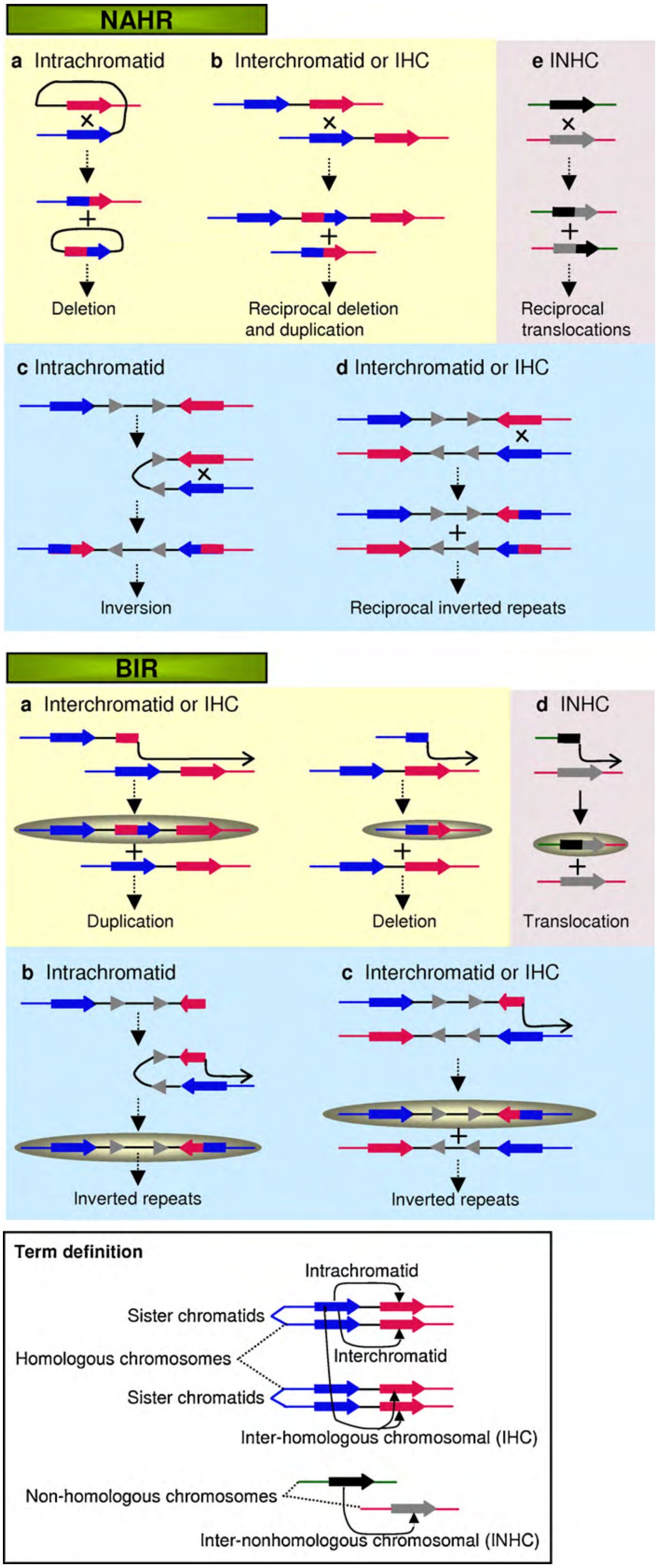

Fig. 2. Types of genomic rearrangements resulting from NAHR (non-allelic homologous recombination) and BIR (break-induced replication). Arrowed bars indicate duplicated sequences or LCRs and their relative orientations. The direction of BIR is indicated by a curved arrow. In BIR, the resulting rearranged chromosomes are within ovals.

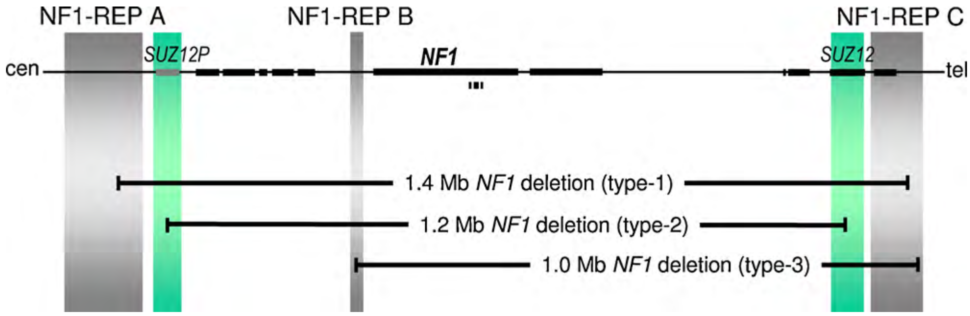
NF1-REP A
NF1-REP B
NF1-REP C
cen
SUZ12P
NF1
SUZ12
tel
1.4 Mb NF1 deletion (type-1)
1.2 Mb NF1 deletion (type-2)
1.0 Mb NF1 deletion (type-3)

Fig. 3. Structure of the *NF1* gene region at 17q11.2 showing the relative locations of the low copy repeats and the three known types of gross deletion at this locus. The genomic positions of NF1-REPs A, B and C are indicated, as are the positions of functional genes and the *SUZ12* pseudogene (*SUZ12P*). The horizontal black bars represent the genes located in this region, with the *NF1* gene, *SUZ12* and *SUZ12P* specifically indicated. The extent of the common 1.4 Mb type-1 *NF1* deletions, the 1.2 Mb type-2 deletions (with breakpoints located within *SUZ12*/*SUZ12P* sequences) and the 1.0 Mb type-3 deletions are given in relation to the positions of the genes that map to the extended *NF1* gene region.

was however noted in the breakpoint regions that could have facilitated NAHR.

Intriguingly, Steinmann et al. [19] demonstrated that all 13 type-2 deletions so far identified are characterized by somatic mosaicism, indicating a positional preference for mitotic NAHR within the *NF1* gene region. Thus, whereas meiotic NAHR occurs between the NF1-REPs giving rise to type-1 deletions, NAHR during mitosis appears to occur between the *SUZ12* gene and its pseudogene, thereby generating type-2 deletions. Such a clear distinction between the preferred sites of mitotic versus meiotic NAHR is unprecedented in any other genomic disorder induced by the local genomic architecture. This notwithstanding, 12 of the 13 mosaic type-2 deletions were found in females, a finding consistent with the observation that type-1 deletions occurred preferentially during maternal meiosis. Although an influence of chromatin structure was strongly suspected, no gender-specific differences in the methylation pattern exhibited by the *SUZ12* gene were apparent that could explain the higher rate of mitotic recombination in females.

More recently, Roehl et al. [20] tested the hypothesis that regions of high allelic similarity [also termed 'runs of homozygosity' (ROHs)] in regions flanking the type-2 *NF1* deletions might facilitate their occurrence. ROHs are quite common in the human genome; they are characterized by multiple contiguous homozygous single nucleotide polymorphisms (SNPs) and range in size from 200 kb to several Mb [21]. ROHs originate neither by deletions nor uniparental disomy but rather from the inheritance of identical-by-descent haplotypes ('autozygosity') in outbred human populations. Evidence linking consanguinity to higher rates of cancer suggests that autozygosity could influence cancer predisposition [22]. Further, since genomic regions with a high degree of germline homozygosity have been reported to constitute hotspots for deletions or mitotic recombination in various human solid tumours [23], it may be that regions of extended homozygosity could promote somatic deletion. In support of this postulate, a significant increase in the prevalence of cancer has been noted in populations/groups characterized by high rates of consanguinity and/or extended homozygosity (autozygosity) [22,24]. Based primarily on the observations made by Assie et al. [23], Roehl et al. [20] postulated that germline homozygosity might, under certain circumstances, predispose to somatic deletions by increasing the rate of mitotic allelic homologous recombination (AHR). Increased mitotic AHR may be tightly linked to, or could even trigger, NAHR in genomic regions harbouring LCRs, such as the *NF1* gene region. To test this hypothesis, Roehl et al. [20] used Affymetrix SNP 6.0 arrays to reinvestigate 12 previously described NF1 patients with type-2 deletions and precisely identified breakpoints. They observed that in 6 of these 12 investigated deletions (50%), the *NF1* deletions were flanked by extended regions of homozygosity without copy number loss. These regions of homozygosity surrounding the deletions differed in size between different patients, but in all cases extended beyond the bounds of the deletions themselves, spanning several hundred kb in length. However, ROHs >500 kb directly flanking the *NF1* deletion region within 17q11.2 on both sides were not found to occur disproportionately in NF1 patients harbouring type-2 deletions as compared to controls. Hence, low allelic diversity in 17q11.2 is unlikely to be a key factor in promoting NAHR-mediated somatic type-2 deletions. Nevertheless, Roehl et al. [20] identified a specific ROH of 588 kb (roh1), located some 525 kb proximal to the deletion interval, which was found to occur significantly more frequently in the type-2 deletion patients as compared with controls. A potential role for roh1 in increasing the frequency of somatic NAHR between the duplicated *SUZ12* sequences remains to be investigated.

2.1.2. Gene conversion

Gene conversion refers to the unidirectional transfer of genetic material from a 'donor' sequence to a highly homologous 'acceptor' (for a review, see [25]). Mechanistically, gene conversion and NAHR represent alternative outcomes of a common two-ended DSB repair process, with divergence occurring at two time-points. The synthesis-dependent strand annealing (SDSA) model predicts the divergence after D-loop extension; the invading strand and newly synthesized DNA is displaced from the template and anneal to the other 3′ ssDNA tail of the DSB, followed by DNA synthesis and ligation of nicks. The other divergence timepoint is after the formation of the double HJs. The double HJs can be dissolved by the BLM-Topo IIIα-BLAP75 complex through convergent branch migration. Alternatively, cleavage of the double HJs (by GEN1 in humans) can also result in gene conversion (Fig. 1). Note that in both SDSA and double HJ dissolution, DNA synthesis occurs in the receiving strand. In double HJ cleavage, gene conversion is thought to be derived from the mismatch repair of the heteroduplex DNA that is formed between the donor and acceptor DNA sequences. The mismatch correction probably occurs before the resolution of the double HJs; and it is the broken strand that is usually corrected using the intact strand as a template (reviewed in [26]).

As NAHR, gene conversion can occur between homologous sequences located within the same chromatid, sister chromatids, homologous chromosomes and non-homologous chromosomes (see Fig. 2). However, gene conversion constitutes a unique type of genomic rearrangement since, in all possible scenarios, it ends in the substitution of a sequence tract by a 'copy' of a homologous sequence. Moreover, in mammals, gene conversion tracts are usually short, of the order of <1-kb in length [25]. Nonetheless, gene conversion is a well-established cause of both inherited disease and cancer [25]. Gene conversion also plays an important role in sequence homogenization of segmental duplications or LCRs [25]. Such an effect could significantly modify the activity of homologous recombination hotspots, potentially affecting the rate of *de novo* occurrence of NAHR-derived human disorders.

Finally, it is important to emphasize that there is a qualitative difference between gene conversion events and transient hypermutability-mediated multiple mutations. Whereas gene conversion represents a template-switching event through which a highly homologous template is faithfully copied by a normal replicative DNA polymerase, transient hypermutability-derived multiple mutations are thought to arise from (a) the deregulated expression of, or conformational change in, either a replicative DNA polymerase or another protein involved in the maintenance of replication fidelity, (b) the disruption of the balance of the nucleotide pool or (c) the recruitment of error-prone DNA polymerases in DNA replication or repair [27].

2.1.3. *BIR*

Both NAHR and gene conversion can repair a DSB with two ends. However, under some circumstances, e.g. when a replication fork is collapsed or broken, only one end of a DSB is available. Such one-ended DSBs can still be repaired by homologous recombination, via a mechanism known as BIR or recombination-dependent DNA replication (Fig. 1). Unlike NAHR and gene conversion, what happens after D-loop extension in BIR remains unclear. One possibility is that the invading strand succeeds in establishing a unidirectional replication fork that is capable of proceeding until the end of the template chromosome; cleavage of the HJ junction then gives rise to two semiconservative replication products. Alternatively, the strand invasion sets up a replication fork but both the newly synthesized leading and lagging strands are constantly displaced through the action of a branch-migration enzyme(s). A third possibility is that the D-loop migrates down the template chromosome with the lagging strand being synthesized on the displaced nascent strand (for a review, see [28]). In the latter two models, the newly synthesized DNA is invariably associated with the broken chromosome (Fig. 1). Before a stable replication structure is established, the invading strand may undergo multiple rounds of displacement and annealing, probably reflecting repeated attempts to find the other end of the DSB [29]. Apart from reinitiating stalled and broken replication forks, BIR has been considered to play an important role in maintaining telomere length through a “roll and spread” mechanism, in the case of excessive telomere shortening or disruptions in the function of telomere-binding proteins [28].

Compared with NAHR, BIR yields only non-reciprocal duplications, deletions, inverted repeats and translocations (Fig. 2). These BIR events are indistinguishable from their NAHR counterparts (Fig. 2). Thus, some of the pathogenic events that have been accounted for by NAHR may in fact result from BIR.

2.1.4. *SSA*

Direct repeats flanking the DSB, which are rendered single-stranded by 5′ to 3′ end resection, may simply anneal with each other before one of the 3′ ssDNA tails can find and base-pair with a homologous sequence. 3′-flaps are then removed and gaps filled by DNA synthesis, resulting in simple deletions (Fig. 1). Obviously, the success rate of this pathway, SSA, is inversely related to the distance separating the two direct repeats. Thus, SSA may only account for some small-scale deletions [30].

2.2. *NHEJ*

NHEJ involves simple ligation of any two broken DNA ends together. It is the most prominent DNA repair mechanism because it can occur at any time during the cell cycle (although preferentially during G_0, G_1, and early S phase) and does not require a homologous sequence. NHEJ is divided into two sub-pathways, classical and non-classical.

2.2.1. *Classical NHEJ*

Classical NHEJ is controlled by the Ku heterodimer (Ku70/Ku80) and DNA-PKcs; and the ends are joined by ligase IV, XLF and XRCC4. The detailed molecular mechanism of classical NHEJ has been reviewed elsewhere [31]. Several issues are however noteworthy: (a) The presence of terminal microhomologies (typically 1–4 bp) facilitates classical NHEJ but this is not absolutely necessary [31], (b) NHEJ of two compatible or blunt ends of a same DSB is of high fidelity. By contrast, the NHEJ junctions of two incompatible ends of the same DSB are often characterized by small (typically 1–4 bp) deletions, insertions or indels (for reviews, see [32,33]). The main reason is that end resection in classical NHEJ is very limited since this pathway only efficiently joins DSBs with overhangs of fewer than four bases [33], (c) NHEJ of ends from simultaneous DSBs has the potential to account for a diverse range of genomic rearrangements, with some possible outcomes illustrated in Fig. 4. In this regard, ‘telomere capture’ (e.g. [34,35]) and ‘addition of preformed oligonucleotides into broken ends’ [36] require the presence of simultaneously generated multiple DSBs whilst the ‘breakage–fusion-bridge cycle’ [37] involves the ligation of successively generated DSBs during different cell cycles (see also Section 3.1). Lastly, NHEJ-compatible events involving microhomologies can be alternatively explained by the model of MMRDR (see Section 2.3).

Obviously, NHEJ of two ends from different DSBs requires such ends to be physically located in the immediate vicinity. In mammalian cells, high-precision tracking of tagged broken chromosome ends indicates that these ends can only partially separate and, consequently, DSBs preferentially undergo translocations with neighbouring chromosomes [38]. This provides strong support to the ‘contact-first’ hypothesis, which proposes that interactions between different DSBs can only take place when they colocalize at the time of DNA damage [39]. Consistent with this hypothesis, close spatial proximity has been observed in several frequent translocation partners (for a review, see [40]). Recently, the mechanism underlying the recurrent fusion of the 5′ end of the untranslated region of the androgen receptor (*AR*) target gene *TMPRSS2* (located on chromosome 21) with members of the *ETS* family of genes (either *ERG* or *ETV1* on chromosomes 21 and 7, respectively) in prostate cancer [41] has been elucidated. Intronic binding of liganded AR first brings the loci involved in the translocation into close proximity. Subsequently, the *AR* gene promotes site-specific DNA DSBs at translocation loci by recruiting activation-induced cytidine deaminase and the LINE-1 repeat-encoded ORF2 endonuclease. These enzymes then synergistically generate site-selective DSBs at juxtaposed translocation loci that are ligated by the NHEJ pathway for specific translocations [42,43]. According to this model, non-random tumour translocation arising from NHEJ would actually be promoted by transcription factor binding and chromatin remodelling [44]. Correlations between the targeted spatial proximity of chromosomes and DNA repair mechanisms such as NHEJ are increasingly being recognized as underlying many recurrent tumour translocations [45].

2.2.2. *Non-classical NHEJ*

The identification of rare Ku-independent end-joining events using longer microhomology (5–25 bases) revealed a new DSB repair pathway, originally termed microhomology-mediated end joining (MMEJ) [46]. As noted by Lieber [31], the term ‘MMEJ’ is confusing because a subset of classical NHEJ events also uses 1–4 bp of terminal microhomology. Taking this and the description of Yan et al. [47] into consideration, we tentatively term this pathway ‘non-

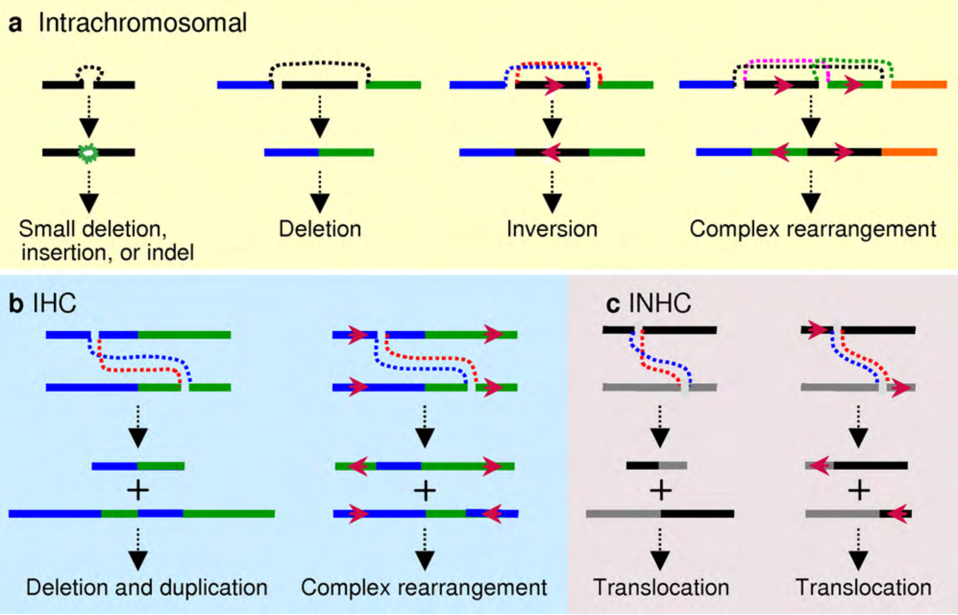

Fig. 4. Examples of genomic rearrangements resulting from non-homologous end joining (NHEJ). Ends ligated are indicated by dotted lines. In b and c, the final outcome, unlike non-allelic homologous recombination (NAHR), is not necessarily reciprocal. In theory, the flexibility of NHEJ implies an unlimited number of different types of genomic rearrangement. IHC, inter-homologous chromosomes. INHC, inter-nonhomologous chromosomes.

classical NHEJ'. Non-classical NHEJ can repair ends of a single DSB in a way similar to that described in SSA (refer to Fig. 1), resulting in the generation of small-scale deletions. This pathway can also repair ends of different DSBs, evidenced by the longer microhomologies found at some translocation breakpoints [32].

2.3. MMRDR

Replication slippage or template switching during replication has long been used to explain the generation of small deletions and duplications with terminal microhomologies. The key feature of this canonical model is that the newly synthesized strand can dissociate with its template and then reassociate in a misaligned configuration using microhomology. Hence, if the newly synthesized strand misaligns at a downstream short direct repeat, continued DNA synthesis will lead to the deletion of one of the direct repeats and the intervening sequence between the two direct repeats. On the other hand, if the newly synthesized strand misaligns at an upstream direct repeat, continued DNA synthesis will lead to the insertion of one of two direct repeats plus the intervening sequence [48].

Recently, two similar models known as serial replication slippage (SRS) [48–50] and fork stalling and template switching (FoSTes) [51] were proposed to account for the generation of complex genomic rearrangements. As noted by Gu et al. [10], "both models assume serial replication slippage, and both stress the importance of the genomic architectural elements such as palindromic DNA, stem–loop structures, repeats and so on, which may facilitate the initial stalling of the replication fork." Irrespective of the canonical replication slippage model or SRS/FosTes, the newly synthesized strand is invariably predicted to realign to its original template strand that remains unbroken during the process. Taking duplication for example, a single-stranded/loop structure will form in the newly synthesized mutant strand. Two distinct means were postulated for the conversion of the single-stranded mutant sequence to double-stranded. Firstly, the synthesis of, and replication against, the nascent mutant strand could have occurred within the same cell cycle; a process that would have required cleavage of the original template strand followed by DNA gap filling and ligation. Alternatively, the single-stranded/loop structure could have escaped the host repair system; DNA replication against the nascent mutant strand would then have occurred in the next cell cycle (see Fig. 2A in [52]).

More recently, a new model, termed break-induced SRS (BISRS; [53]) has been proposed; this model successfully integrated the key features of SRS with those of the earlier microhomology-dependent BIR model [54] to account for the generation of a double complex copy number mutation (CNM) involving the *F8* and *FUNDC2* genes. The microhomology-dependent BIR model could also account for large simple deletions and duplications associated with short direct repeats [50]. This model, alternatively known as microhomology-mediated BIR (MMBIR), has now been increasingly recognized as a plausible mechanism for generating human CNVs (e.g. [52,55–58]). In BISRS or MMBIR, replication ends with the engagement of a misaligned template instead of reannealing to its original template; the synthesis of the second strand follows the synthesis of the first strand (see Fig. 2A in [52]). Given that replication is a frequent source of one-ended DSBs, the break-induced models may have greater explanatory potential than the simple replication slippage and SRS/FosTes models [52].

All the aforementioned replication-based models are predicated upon the use of microhomology for strand misaligning. The term 'MMRDR', which stands for microhomology-mediated replication-dependent recombination, appears to best define the hallmarks characteristic of these replication-based mutational mechanisms as compared with homologous recombination and NHEJ.

2.4. Local sequence features predisposing to genomic recombination

Most of the abovementioned mutational models invoke the formation of DSB. It has been repeatedly demonstrated that the occurrence of DSB in the human genome is not random but rather strongly influenced by the local DNA sequence environment. Insights generated from representative meta-analysis of pathogenic breakpoint sequences will be highlighted below.

2.4.1. Multigene study of the nature of chromosomal breakpoint junctions

Early analyses of the sequence context of chromosomal rearrangements were confined to the examination of small numbers of gross deletion and translocation breakpoint junctions at specific gene loci. The construction of the *Gross Rearrangement Breakpoint Database* (GRaBD), containing 397 germ-line and somatic DNA breakpoint junction sequences derived from a total of 219 different rearrangements underlying human inherited disease and cancer [59], however allowed the first methodical examination of the local DNA sequence environment of translocation and deletion breakpoints across a wide variety of different gene loci. Using GRaBD, Abeysinghe et al. [60] analyzed the sequence context of translocation and deletion breakpoints in a search for general characteristics that might have rendered these sequences prone to rearrangement. The oligonucleotide composition of breakpoint junctions and a set of reference sequences, matched for length and genomic location, were compared with respect to their nucleotide composition. Deletion breakpoints were found to be AT-rich whereas by comparison, translocation breakpoints tended to be GC-rich. Alternating purine–pyrimidine sequences were found to be significantly over-represented in the vicinity of deletion breakpoints while polypyrimidine tracts were over-represented at translocation breakpoints. A number of recombination-associated motifs were found to be over-represented at translocation breakpoints (including DNA polymerase pause sites/frameshift hotspots, immunoglobulin heavy chain class switch sites, heptamer/nonamer V(D)J recombination signal sequences, translin binding sites, and the χ-like element) but, with the exception of the translin-binding site and immunoglobulin heavy chain class switch sites, none of these motifs were over-represented at deletion breakpoints. *Alu* sequences were found to span both breakpoints in seven cases of gross deletion that may thus be inferred to have arisen by homologous recombination. Re-analysis of some of these questions should be performed using much larger datasets of DNA sequences from gross deletion and translocation breakpoint junctions now available courtesy of the cancer genome sequencing projects.

2.4.2. Formation of DNA secondary structures between DNA ends at recombination breakpoints

Early studies of gross rearrangement at specific gene loci served to document the occurrence of various different types of repetitive sequence element in the vicinity of breakpoint junctions. Chuzhanova et al. [61] studied the potential involvement of various types of repetitive sequence element in the formation of secondary structure intermediates between the single-stranded DNA ends that recombine during gene rearrangements. Complexity analysis was then used to assess the potential of these ends to form secondary structures, the maximum decrease in complexity consequent to a gross rearrangement being used as an indicator of the type of repeat and the specific DNA ends involved. A total of 175 pairs of deletion/translocation breakpoint junction sequences available from GRaBD [59] were analyzed. Potential secondary structure was noted between the 5′ flanking sequence of the first breakpoint and the 3′ flanking sequence of the second breakpoint in 49% of rearrangements and between the 5′ flanking sequence of the second breakpoint and the 3′ flanking sequence of the first breakpoint in 36% of rearrangements. Inverted repeats, mirror repeats and symmetric elements were found in association with gross rearrangements at approximately the same frequency. However, inverted repeats and inversions of inverted repeats accounted for the vast majority (83%) of deletions plus small insertions, symmetric elements for one-half of all antigen receptor-mediated translocations, while direct repeats appear only to be involved in mediating simple deletions. These findings served to extend our understanding of illegitimate recombination by highlighting the importance of secondary structure formation between single-stranded DNA ends at breakpoint junctions.

2.4.3. Formation of non-B DNA structures at chromosomal breakpoints

Specific DNA sequence motifs such as alternating purine–pyrimidines, polypurines, polypyrimidines and G-rich tetrad direct repeats undergo structural transitions from the orthodox right-handed *B*-helical duplex to high energy state non-*B* DNA structures under negative superhelical stress. Over the last few years, it has become apparent that non-*B* DNA conformations often coincide with chromosomal breakpoints in both inherited disease and cancer [62–64]. These structures are thought to initiate genomic rearrangements by increasing the rate of single-strand lesion formation at these sites. Consistent with these results, Abeysinghe et al. [60] found that alternating purine–pyrimidine sequences of between 2 and 74 bp were found to be significantly over-represented in the vicinity of deletion breakpoint junctions. Such sequences are prone to form Z-DNA particularly under conditions of negative superhelical stress [65]. Z-DNA is a left-handed helix with only a single minor groove that forms during transcription *in vivo* as a result of torsional strain generated by the moving RNA polymerase. Z-DNA may be recombinogenic in a number of different ways. Firstly, it may facilitate recombination between homologous chromosomal regions by relieving the topological stress that arises when intact duplexes are intertwined. Secondly, Z-DNA regions may exclude histones and other architectural proteins, thereby influencing both the location of nucleosomes and the organisation of chromosomal domains, as well as increasing accessibility to recombinases.

Abeysinghe et al. [60] found that polypurine runs of 2–23 bp and polypyrimidine tracts of 2–44 bp were significantly over-represented at translocation breakpoint junctions while polypurine tracts of 25–39 bp were over-represented at deletion breakpoint junctions. Such sequences have been reported before to occur at both translocation [66] and gross deletion breakpoints [67] and appear to stimulate homologous recombination *in vivo*. Both polypurine and polypyrimidine sequences are capable of adopting the triple helical H-form of DNA particularly when exposed to an acidic environment and negative superhelical stress [68]. Since H-DNA is partially single-stranded, it may be susceptible to nuclease attack that could then facilitate recombination, but it may also promote recombination by blocking DNA replication. The ability of polypurine and polypyrimidine sequences to form H-DNA may thus render these sequences prone to illegitimate recombination.

2.4.4. The influence of local sequence features on gene conversion

Chuzhanova et al. [69] found that gene conversion events tend to occur within (C + G)- and CpG-rich regions and that sequences with the potential to form non-*B*-DNA structures occur disproportionately in the immediate vicinity of the converted tracts. They also showed that gene conversion events tend to occur in genomic regions that have the potential to fold into stable hairpin conformations. These findings support the concept that recombination-inducing motifs, in association with alternative DNA conformations, can promote recombination in the human genome.

2.5. What is the ratio of deletions to duplications in vivo?

Based on a survey of the known mutational mechanisms, one might intuitively conclude that deletions would be generated in a higher ratio than duplications. The logic behind this assertion is that some mechanisms only give rise to deletions. Can a general ratio of *de novo* deletions vs duplications *in vivo* be derived? Findings from three studies appeared to provide a first answer to this question.

By means of sperm analysis, Turner et al. [70] measured the numbers of *de novo* meiotic deletions and duplications in three autosomal NAHR hotspots (i.e. CMT1A-REP at 17p11.2, WBS-LCR at 7q11.23 and LCR17p at 17p11.2) and found quite similar ratios of deletions to duplications (i.e. 2.43, 2.10 and 2.14, respectively). The higher rate of deletion over duplication is readily explicable in terms of intrachromatid NAHR occurring more frequently than either interchromatid or IHC NAHR; the former type of NAHR generates only deletions whereas the latter two types generate reciprocal deletions and duplications (see Fig. 2). Consistent with this explanation, the chromosome Y-located AZFa-HERV hotspot, which cannot (on the basis of its chromosomal location) undergo IHC NAHR, exhibits the highest reported ratio (4.11) of deletions:duplications [70].

Unlike *in vitro* assays, clinically observed findings are often confounded by many diverse factors, most notably clinical selection. This is perhaps best exemplified by the recurrent reciprocal 1.4 Mb deletions and duplications at the CMT1A-REP NAHR hotspot associated with quite distinct clinical phenotypes [viz. deletion/hereditary neuropathy with liability to pressure palsies (HNPP); duplication/Charcot-Marie-Tooth disease type 1A (CMT1A), respectively]. Although the sperm-based assay has demonstrated that the ratio of deletions to duplications at this hotspot should be of the order of 2:1, the actual ratio of HNPP to CMT1A coming to clinical attention is ~1:4. The under-ascertainment of HNPP is clearly attributable to the relatively mild and variable clinical phenotype associated with the CMT1A-REP deletion [70].

Using array comparative genomic hybridization (aCGH) with an average of one interrogating probe every 95 bp, we analyzed the entire length of the *CFTR* gene (189 kb) in a large number ($n = 233$) of cystic fibrosis chromosomes lacking conventional mutations and identified a total of 15 intragenic deletion CNMs and 5 intragenic duplication CNMs [71]. Clinical selection is unlikely to be a major concern in this case because it is intragenic events that were studied. Another potential confounding factor may be that large duplications, once they have arisen, could be more unstable than deletions. This possibility is not easily refuted, but duplicated (or even triplicated) sequences appear to be stably transmissible from one generation to another, not only in the context of evolution [72] but also in the context of inherited disease (see [73] and references therein). Taking this into account, we considered that the ratio of disease-causing intragenic *CFTR* deletions to duplications (3:1) would approximate to the actual relative occurrence of *de novo* deletion CNMs to duplication CNMs at this locus [71]. A similar ratio (i.e. 2.4) of pathogenic intragenic events was reported in the 91 kb *LIS1* gene; analysis of 53 patients with isolated lissencephaly (all patients were previously found to be negative for microdeletions in the 17p13.3 region by FISH and were also negative for conventional mutations upon sequencing the *LIS1* gene) by MLPA, identified 12 intragenic deletion CNMs and 5 intragenic duplication CNMs [74].

The similarity of these ratios in the three studies, despite the widely different contexts and means of detection, were thought to imply the operation of a common biological mechanism underlying the generation of deletion and duplication CNVs [71]. Indeed, intrachromatidal events, irrespective of whether they originate via homologous recombination, NHEJ or MMRDR mechanisms, should occur more frequently than either interchromatidal or IHC events; the former type of event can only generate deletions whereas the latter two types of event should generate deletions and duplications in equal proportions. We therefore proposed that a deletion:duplication ratio of between 2 and 3 is likely to represent the best estimate of the relative occurrence of deletion and duplication CNMs in the human autosomal genome [71].

Such a ratio may not however not be confined to CNMs. As of February 4, 2010, the Human Gene Mutation Database (HGMD; Professional Release; http://www.hgmd.org) registered 15,231 microdeletions and 6273 microinsertions, a ratio of 2.4. Since these events are all ≤20 bp, significant bias due to either the aforementioned confounding factors or mutation detection efficiency is unlikely to be a major concern. By contrast, HGMD Professional registered 5912 gross deletions and 1210 gross deletions (all events >20 bp), a ratio of 4.9. This rather higher ratio may be largely attributable to the detection bias operating against duplication CNMs [71].

3. Gross genomic rearrangements in cancer

3.1. Differences between human inherited disease and cancer

In principle, the mutational mechanisms described above are applicable in the context of both inherited disease and cancer. However, as compared with inherited disease, cancer demonstrates a unique feature that becomes understood in the light of the 'two-hit' theory of carcinogenesis; the genome of all cancer cells harbours somatic mutations.

Some of the somatic mutations, known as 'driver' mutations, confer a growth advantage upon the cell in which they occur (and have therefore been positively selected for in the emerging tumour mass) and may thus be deemed to be causally implicated in tumorigenesis. By contrast, those mutations which do not confer any growth advantage and which have not been subject to selection during tumorigenesis are termed 'passenger' mutations [75]. Such passenger mutations may arise at high frequency as a consequence either of increased genomic instability or simply due to the considerable number of cell divisions between a single transformed cell and the clinically detectable tumour. A variety of *in silico* methodologies are being developed with the aim of distinguishing driver mutations from passenger mutations [75]. The catalogue of Somatic Mutations in Cancer (COSMIC; http://www.sanger.ac.uk/cosmic) represents the largest public resource for information on somatically acquired mutations in human cancer. Currently (v46, March 2010), COSMIC contains details of over 108,000 mutations in 18,478 genes from almost 450,000 tumours [76]. Clearly, many of these genes are likely to harbour passenger mutations. Those genes for which mutations have been causally implicated in cancer are catalogued in the Cancer Gene Census [77] which currently (April 2010) lists a total of 427 different genes; cancer-causing translocations have been reported in some 301 of these genes, large deletions in 31 of these genes, and amplification in 12 genes (http://www.sanger.ac.uk/genetics/CGP/Census).

Two types of somatic genomic rearrangements in cancer are worthy of particular note. Translocations, which generally confer a growth advantage upon the affected cells or tissue through the creation of a hybrid gene encoding a tumour-specific fusion protein [78–80] and gene amplification, a fairly frequent type of somatic rearrangement. Somatic gene amplification usually involves the copy number increase of a specific region of a chromosome in a specific tumour tissue and is often associated with the over-expression of amplified gene(s) located within the amplified region [81]. In principle, gene amplification can be accounted for by the breakage-fusion-bridge cycle: (a) either DSB or telomere attrition generates an uncapped chromosomal end, (b) replication results in two identical sister chromatids lacking telomeres, (c) the two free ends are directly ligated (e.g. by NHEJ), resulting in the formation of a palindromic dicentric chromosome, (d) the palindromic dicentric chromosome promotes further chromosome breaks during anaphase separation, and (e) these breaks initiate another round of the breakage-fusion-bridge cycle (for a review, see [30]). Recent developments have potentiated the role of inverted repeats in generating the first palindromic dicentric chromosome. The widely accepted model is that inverted repeats located near the DSB can 'snap-back' to form a hairpin at the chromosome end. Following fill-in and gap ligation, this will generate a capped end that resolves into a large DNA palindrome after DNA replication [82]. A recent review of the literature identified a total of 77 genes that represent good candidates for involvement in tumorigenesis through gene amplification [83]. Gene amplification has also been described in other genes (e.g. globin genes) as a cause of human inherited disease [84].

3.2. Gross rearrangements detected in cancer genome sequencing studies

Entire genome sequences have now been determined for a variety of tumour types including colorectal, breast, lung adenocarcinoma, melanoma, glioblastoma, myeloid leukemia and lymphoma [75,85]. Thus, Stephens et al. [86] reported a total of 2166 gross somatic rearrangements detected in 24 breast cancer genomes comprising intra-chromosomal deletions (16.5%), tandem duplications (34%), inversions (10%) and amplifications (28.5%) and inter-chromosomal events (11%). By contrast, the sequencing of the genome of a malignant melanoma yielded only 37 gross somatic rearrangements; of these, 3 were inter-chromosomal and 34 intra-chromosomal, including 25 deletions, 6 inversions, 2 duplications and one other large intra-chromosomal event [87]. Nineteen of these 37 somatic rearrangements were found to have interrupted protein-coding genes.

The above notwithstanding, one of the best understood cancer genomes, in terms of the gross somatic rearrangements associated with the process of tumorigenesis, is lung cancer. Some of the key studies on the lung cancer genome are therefore described below. The lessons learned during the course of these studies are however very likely to be applicable to other types of cancer.

3.2.1. Sequence analysis of the lung cancer genome reveals the nature of somatically acquired rearrangements

Campbell et al. [88] employed genome-wide massively parallel sequencing to generate sequence reads from both ends of short DNA fragments derived from the genomes of two individuals with lung cancer [specifically, one SCLC cell line (NCI-H2171) and one neuroendocrine cell lung cancer cell line (NCI-H1770)]. Coverage was 2.4 gigabases (Gb) for NCI-H2171 and 1.8 Gb for NCI-H1770. Some 325 rearrangements were identified in NCI-H2171 of which 81 were somatic and 244 were germline, whereas some 84 rearrangements were identified in NCI-H1770 of which 22 were somatic and 62 were germline. These rearrangements were all characterized at the base-pair level. The patterns of germline and somatic rearrangement were markedly different in the two lung cancer cell lines examined. The vast majority of the germline rearrangements involved *Alu* sequences or LINE elements and appeared to represent insertions by comparison to the reference genome. A few inversions and tandem duplications were noted but only one interchromosomal germline rearrangement was identified. Of the 103 somatic rearrangements identified in the two cell lines, most (79%) were intra-chromosomal but only two were deletions. The vast majority of the intra-chromosomal somatic rearrangements (63/81) were confined to 'amplicons' within already heavily amplified regions, although 11 tandem duplications were observed. Of the 22 somatic interchromosomal rearrangements, 15 were between amplicons whereas 7 involved the transfer of an amplicon to a non-amplified region. One of the interchromosomal rearrangements [t(2;12)] was found to lead to the generation of a hybrid *CACNA2D4-WDR43* gene yielding out-of-frame fusion transcripts. Two of the eleven intra-chromosomal tandem duplications of internal exons also served to generate out-of-frame transcripts in two genes (*GRID2* and *CNTNAP5*). In addition, a t(8;8)(q12;q24) intra-chromosomal translocation was identified which was predicted to result in the fusion of the *PVT1* and *CHD7* genes, a hybrid fragment which was itself subsequently subject to amplification. Finally, Campbell et al. [88] noted that (i) non-templated sequence of 1-57 bp in length was present at the breakpoints of the somatic rearrangements and (ii) that some 53% of these acquired rearrangements exhibited short (1–10 bp) stretches of homology. These observations were consistent with the postulate that NHEJ is the predominant mechanism of mutagenesis in the soma.

3.2.2. CNVs in the lung adenocarcinoma genome

The genomes of a collection of 371 resected lung adenocarcinomas and matched normal DNAs have also been screened for CNVs using high density microarrays [89]; a total of 26 'large-scale events' (10 significant gains and 16 significant losses) and 31 'focal events' were detected which were distributed between most of the chromosomal arms. Similar patterns of copy number gain and loss were noted in most of the lung adenocarcinoma samples but these samples exhibited marked differences in the amplitude of CNV. Some attenuation of the signal was evident in all samples and this was held to be due to admixture with non-tumour DNA. The most common genomic alteration in lung adenocarcinoma was a copy number gain on chromosome 5p (60% of all samples) with the remaining 15 large-scale events being evident in at least 33% of all samples. Together, the regions of common copy number gain (650 Mb) and copy number loss (1010 Mb) comprise more than half the human genome.

The most significant of the focal deletions encompassed a region on 9p21 containing the *CDKN2A* and *CDKN2B* tumour suppressor genes and was detected in 3% of samples. Focal deletions on 10q23.31 and 13q14.2, encompassing the *PTEN* and *RB1* tumour suppressor genes, respectively were both found in 0.5% of samples. Three other genes (*PTPRD*, *PDE4D* and *AUTS2*) were also found to be individually deleted. When these genes were subsequently screened in all lung adenocarcinoma samples for subtle mutations, somatic mutations in the *PTPRD* gene, encoding a tyrosine phosphatase, were detected in 11/188 samples (all missense mutations, six of which were predicted to be deleterious to function).

With respect to focal amplification events, some 24 recurrent regions were identified in 1–7% of samples, with levels of copy number amplification varying from 4- to 16-fold. A number of these regions contain proto-oncogenes which have been previously reported as having been amplified in lung tumour material (e.g. *MDM2*, *MYC*, *EGFR*, *CDK4*, *KRAS*, *ERBB2*, *CCND1* and *TERT*). The most common focal amplification (that of 14q13.3, amplified in 6–12% of lung adenocarcinoma samples) contained the *NKX2-1* (NK2 homeobox 1/TITF1) gene; studies of RNA interference indicated that the *NKX2-1* gene is essential for the survival and maintenance of lung adenocarcinoma cells.

In a collection of colorectal cancer specimens and cell lines, Martin et al. [90] identified 50 'minimal common regions' of CNV including 28 amplifications and 22 deletions. Of the 28 amplification events, 11 were also found in lung adenocarcinoma. These regions contain gene loci that have already been implicated in lung carcinogenesis (e.g. *EGFR*, *MYC*, *CCND2* and *KRAS*) as well as a number of other genes that represent potential candidates for involvement in a wide range of cancers.

The question naturally arises as to whether CNVs are accompanied by corresponding changes in gene expression. Lockwood et al. [91] studied a total of 24,892 genomic loci in each of 53 lung cancer cell lines and identified a considerable number of genes residing in amplification 'hotspots'. A total of 1690 amplicons were identified in the lung cancer genome involving a total of 106 oncogenes. These amplicons occurred at a frequency of 31.9 per lung tumour and, on average, involved 0.68 Mb of DNA. To address the consequences of amplification for gene expression, Lockwood et al. [91] integrated parallel gene expression profiles with aCGH data for genes within the amplification hotspots in 27 NSCLC cell lines. Some 221/442 (50%) of the amplified genes were expressed at a significantly higher level as a consequence of the increased dosage. Several genes of the EGFR-family signalling pathway (including *CDK5*, *AKT1*, *EGFR*, *MYC* and *SHC1*) were found to be overexpressed as a direct consequence of gene amplification in lung cancer. Indeed, one or more components of the EGFR family pathway were over-expressed as a consequence of amplification in ~70% of the lung adenocarcinoma cell lines analysed. These findings indicate not only that gene amplification is a much more common mechanism of oncogene activation in lung cancer than has been previously realised [77,81] but also that specific regions of the genome represent hotspots of gene amplification.

3.2.3. EML4-ALK fusion genes in lung cancer

Recurring rearrangements of the *ALK* (anaplastic lymphoma kinase) gene have recently been described in non-small cell lung cancer (NSCLC) in Japanese NSCLC tumours [92]. The most common of these is due to an inversion of the short arm of chromosome 2 that creates a fusion between the 5′ portion of the *EML4* (echinoderm microtubule-associated protein-like 4) gene and the 3′ portion of the *ALK* gene. The hybrid *EML4–ALK* gene is formed by disruption of the *ALK* gene at a position 297 bp upstream of exon 21 followed by fusion with an inverted segment of the *EML4* gene disrupted ~3.6 kb downstream of exon 13. This generates a transforming fusion kinase with the N-terminal of EML4 and the C-terminal of ALK. The *EML4–ALK* fusion transcript was initially detected in 6.7% of NSCLC tumours in the Japanese population [92] but has subsequently been detected at similar frequencies in European populations [93]. *EML4–ALK* fusion gene mutations appear to occur in mutual exclusion with *EGFR* and *KRAS* mutations and have been observed disproportionately in lung adenocarcinomas and other tumours removed from never/light smokers [94–96].

4. Concluding remarks

The diverse mutational mechanisms reviewed in this article provide a glimpse of a complex emerging story. While new models will certainly be put forward as more experimental and mutation data become available, many questions still remain to be clarified regarding the known models. How is the choice of appropriate DSB repair pathway made *in vivo*? Why do some NAHR events occur preferentially in females whilst others occur preferentially in males? Can two distinct repair pathways cooperate to generate a specific mutation? What are the relative contributions of recombination-prone motifs, inverted repeats and non-*B* structure-forming sequences to DSB formation? The answers to these questions should significantly enhance our understanding of the mechanisms that ensure the fidelity of DNA repair and the maintenance of genome integrity.

A better understanding of the mutational mechanisms underlying genomic rearrangements should also help to improve the design of mutation detection strategies. For example, the prediction of regions prone to genomic instability based upon the greater understanding of NAHR mechanisms led to the discovery of new genomic disorders (reviewed in [10]). In addition, the proposition that deletions and duplications in the human genome are likely to be generated in the proportion of approximately 2–3:1 has suggested that large intragenic gene duplications in many disease loci have almost certainly been routinely under-ascertained. Finally, a more complete understanding of the mutational processes may provide new therapeutic targets for human genetic disease, particularly cancer [32].

Conflict of interest statement

None declared.

Acknowledgements

We regret that, owing to space limitations, we could not cite many of the original papers that underpin the now widely accepted mutational models. This work was partly supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), Brest, France (J-MC, CF), the Deutsche Forschungsgemeinschaft and Deutsche Krebshilfe (HK-S) and the University of Patras, Department of Pharmacy, Patras, Greece (GPP).

References

[1] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science 2004;305:525–8.

[2] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–51.

[3] Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet 2007;39:S7–15.

[4] Chen JM, Stenson PD, Cooper DN, Férec C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet 2005;117:411–27.

[5] Babushok DV, Kazazian Jr HH. Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat 2007;28:527–39.

[6] Belancio VP, Deininger PL, Roy-Engel AM. LINE dancing in the human genome: transposable elements and disease. Genome Med 2009;1:97.

[7] Pennaneach V, Putnam CD, Kolodner RD. Chromosome healing by *de novo* telomere addition in *Saccharomyces cerevisiae*. Mol Microbiol 2006;59:1357–68.

[8] Jain S, Sugawara N, Lydeard J, Vaze M, Tanguy Le Gac N, Haber JE. A recombination execution checkpoint regulates the choice of homologous recombination pathway during DNA double-strand break repair. Genes Dev 2009;23:291–303.

[9] Ip SC, Rass U, Blanco MG, Flynn HR, Skehel JM, West SC. Identification of Holliday junction resolvases from humans and yeast. Nature 2008;456:357–61.

[10] Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics 2008;1:4.

[11] Carvalho CM, Lupski JR. Copy number variation at the breakpoint region of isochromosome 17q. Genome Res 2008;18:1724–32.

[12] Forbes SH, Dorschner MO, Le R, Stephens K. Genomic context of paralogous recombination hotspots mediating recurrent NF1 region microdeletion. Genes Chrom Cancer 2004;41:12–25.

[13] Raedt TD, Stephens M, Heyns I, Brems H, Thijs D, Messiaen L, et al. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. Nat Genet 2006;38:1419–23.

[14] Upadhyaya M, Ruggieri M, Maynard J, Osborn M, Hartog C, Mudd S, et al. Gross deletions of the neurofibromatosis type 1 (*NF1*) gene are predominantly of maternal origin and commonly associated with a learning disability, dysmorphic features and developmental delay. Hum Genet 1998;102:591–7.

[15] Lopez Correa C, Brems H, Lazaro C, Marynen P, Legius E. Unequal meiotic crossover: a frequent cause of NF1 microdeletions. Am J Hum Genet 2000;66:1969–74.

[16] Palau F, Lofgren A, De Jonghe P, Bort S, Nelis E, Sevilla T, et al. Origin of the *de novo* duplication in Charcot-Marie-Tooth disease type 1A: unequal nonsister chromatid exchange during spermatogenesis. Hum Mol Genet 1993;2:2031–5.

[17] Lopes J, Vandenberghe A, Tardieu S, Ionasescu V, Levy N, Wood N, et al. Sex-dependent rearrangements resulting in CMT1A and HNPP. Nat Genet 1997;17:136–7.

[18] Bengesser KCD, Steinmann K, Kluwe L, Chuzhanova NA, Wimmer K, Tatagiba M, et al. A novel third type of recurrent NF1 microdeletion mediated by non-allelic homologous recombination between LRRC37B-containing low-copy repeats in 17q11.2. Hum Mutat 2010;31:742–51.

[19] Steinmann K, Cooper DN, Kluwe L, Chuzhanova NA, Senger C, Serra E, et al. Type 2 *NF1* deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. Am J Hum Genet 2007;81:1201–20.

[20] Roehl AC, Cooper DN, Kluwe L, Helbrich A, Wimmer K, Hogel J, et al. Extended runs of homozygosity at 17q11. 2: an association with type-2 NF1 deletions? Hum Mutat 2010;31:325–34.

[21] McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet 2008;83:359–72.

[22] Bacolod MD, Schemmann GS, Giardina SF, Paty P, Notterman DA, Barany F. Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies. Cancer Res 2009;69:723–7.

[23] Assie G, LaFramboise T, Platzer P, Eng C. Frequency of germline genomic homozygosity associated with cancer cases. JAMA 2008;299:1437–45.

[24] Bacolod MD, Schemmann GS, Wang S, Shattock R, Giardina SF, Zeng Z, et al. The signatures of autozygosity among patients with colorectal cancer. Cancer Res 2008;68:2610–21.

[25] Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 2007;8: 762–75.

[26] Haber JE, Ira G, Malkova A, Sugawara N. Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday's model. Philos Trans R Soc Lond B Biol Sci 2004;359:79–86.

[27] Chen JM, Férec C, Cooper DN. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. Hum Mutat 2009;30:1435–48.

[28] McEachern MJ, Haber JE. Break-induced replication and recombinational telomere elongation in yeast. Annu Rev Biochem 2006;75:111–35.

[29] Smith CE, Llorente B, Symington LS. Template switching during break-induced replication. Nature 2007;447:102–5.

[30] Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet 2009;10:551–64.

[31] Lieber MR. The mechanism of human nonhomologous DNA end joining. J Biol Chem 2008;283:1–5.

[32] McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet 2008;24:529–38.

[33] Pardo B, Gomez-Gonzalez B, Aguilera A. DNA repair in mammalian cells: DNA double-strand break repair: how to fix a broken relationship. Cell Mol Life Sci 2009;66:1039–56.

[34] Buysse K, Antonacci F, Callewaert B, Loeys B, Frankel U, Siu V, et al. Unusual 8p inverted duplication deletion with telomere capture from 8q. Eur J Med Genet 2009;52:31–6.

[35] Yatsenko SA, Brundage EK, Roney EK, Cheung SW, Chinault AC, Lupski JR. Molecular mechanisms for subtelomeric rearrangements associated with the 9q34.3 microdeletion syndrome. Hum Mol Genet 2009;18:1924–36.

[36] Roth DB, Chang XB, Wilson JH. Comparison of filler DNA at immune, nonimmune, and oncogenic rearrangements suggests multiple mechanisms of formation. Mol Cell Biol 1989;9:3049–57.

[37] McClintock B. Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 1951;16:13–47.

[38] Soutoglou E, Dorn JF, Sengupta K, Jasin M, Nussenzweig A, Ried T, et al. Positional stability of single double-strand breaks in mammalian cells. Nat Cell Biol 2007;9:675–82.

[39] Nikiforova MN, Stringer JR, Blough R, Medvedovic M, Fagin JA, Nikiforov YE. Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. Science 2000;290:138–41.

[40] Meaburn KJ, Misteli T, Soutoglou E. Spatial genome organization in the formation of chromosomal translocations. Semin Cancer Biol 2007;17:80–90.

[41] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer. Science 2005;310:644–8.

[42] Lin C, Yang L, Tanasa B, Hutt K, Ju BG, Ohgi K, et al. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. Cell 2009;139:1069–83.

[43] Mani RS, Tomlins SA, Callahan K, Ghosh A, Nyati MK, Varambally S, et al. Induced chromosomal proximity and gene fusions in prostate cancer. Science 2009;326:1230.

[44] Mathas S, Misteli T. The dangers of transcription. Cell 2009;139:1047–9.

[45] Misteli T, Soutoglou E. The emerging role of nuclear architecture in DNA repair and genome maintenance. Nat Rev Mol Cell Biol 2009;10:243–54.

[46] Ma JL, Kim EM, Haber JE, Lee SE. Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. Mol Cell Biol 2003;23:8820–8.

[47] Yan CT, Boboila C, Souza EK, Franco S, Hickernell TR, Murphy M, et al. IgH class switching and translocations use a robust non-classical end-joining pathway. Nature 2007;449:478–82.

[48] Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat 2005;25:207–21.

[49] Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN. Complex gene rearrangements caused by serial replication slippage. Hum Mutat 2005;26:125–34.

[50] Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN. Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. Hum Mutat 2005;26:362–73.

[51] Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 2007;131:1235–47.

[52] Chauvin A, Chen JM, Quemener S, Masson E, Kehrer-Sawatzki H, Ohmle B, et al. Elucidation of the complex structure and origin of the human trypsinogen locus triplication. Hum Mol Genet 2009;18:3605–14.

[53] Sheen CR, Jewell UR, Morris CM, Brennan SO, Férec C, George PM, et al. Double complex mutations involving *F8* and *FUNDC2* caused by distinct break-induced replication. Hum Mutat 2007;28:1198–206.

[54] Koszul R, Caburet S, Dujon B, Fischer G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J 2004;23:234–43.

[55] Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, et al. Nonrecurrent *MECP2* duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. Genome Res 2008;18:847–58.

[56] Collie A, Landsverk M, Ruzzo E, Mefford H, Buysse K, Adkins J, et al. Non-recurrent *SEPT9* duplications cause hereditary neuralgic amyotrophy. J Med Genet 2009. doi:10.1136/jmg.2009.072348.

[57] Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 2009;5:e1000327.

[58] Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 2009;41:849–53.

[59] Abeysinghe SS, Stenson PD, Krawczak M, Cooper DN. Gross Rearrangement Breakpoint Database (GRaBD). Hum Mutat 2004;23:219–21.

[60] Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat 2003;22:229–44.

[61] Chuzhanova N, Abeysinghe SS, Krawczak M, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Hum Mutat 2003;22:245–51.

[62] Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, et al. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci U S A 2004;101:14162–7.

[63] Raghavan SC, Lieber MR. DNA structures at chromosomal translocation sites. Bioessays 2006;28:480–94.

[64] Bacolla A, Wells RD. Non-B DNA conformations as determinants of mutagenesis and human disease. Mol Carcinog 2009;48:273–85.

[65] Herbert A, Rich A. Left-handed Z-DNA: structure and function. Genetica 1999;106:37–47.

[66] Hirai H, Ogawa S, Kurokawa M, Yazaki Y, Mitani K. Molecular characterization of the genomic breakpoints in a case of t(3;21)(q26;q22). Genes Chrom Cancer 1999;26:92–6.

[67] Cao X, Eu KW, Seow-Choen F, Zhao Y, Cheah PY. Topoisomerase-I- and *Alu*-mediated genomic deletions of the *APC* gene in familial adenomatous polyposis. Hum Genet 2001;108:436–42.

[68] Frank-Kamenetskii MD, Mirkin SM. Triplex DNA structures. Annu Rev Biochem 1995;64:65–95.

[69] Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Férec C, Wells RD, et al. Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. Hum Mutat 2009;30:1189–98.

[70] Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, et al. Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. Nat Genet 2008;40:90–5.

[71] Quemener S, Chen JM, Chuzhanova N, Benech C, Casals T, Macek Jr M, et al. Complete ascertainment of intragenic copy number mutations (CNMs) in the *CFTR* gene and its implications for CNM formation at other autosomal loci. Hum Mutat 2010;31:421–8.

[72] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature 2009.

[73] Chen JM, Masson E, Le Marechal C, Férec C. Copy number variations in chronic pancreatitis. Cytogenet Genome Res 2008;123:102–7.

[74] Haverfield EV, Whited AJ, Petras KS, Dobyns WB, Das S. Intragenic deletions and duplications of the *LIS1* and *DCX* genes: a major disease-causing mechanism in lissencephaly and subcortical band heterotopia. Eur J Hum Genet 2009;17:911–8.

[75] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24.

[76] Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 2010;38:D652–7.

[77] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.

[78] Aman P. Fusion oncogenes in tumor development. Semin Cancer Biol 2005;15:236–43.

[79] Teixeira MR. Recurrent fusion oncogenes in carcinomas. Crit Rev Oncog 2006;12:257–71.

[80] Ortiz de Mendibil I, Vizmanos JL, Novo FJ. Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. PLoS One 2009;4:e4805.

[81] Albertson DG. Gene amplification in cancer. Trends Genet 2006;22:447–55.

[82] Tanaka H, Cao Y, Bergstrom DA, Kooperberg C, Tapscott SJ, Yao MC. Intrastrand annealing leads to the formation of a large DNA palindrome and determines the boundaries of genomic amplification in human cancer. Mol Cell Biol 2007;27:1993–2002.

[83] Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 2010;10:59–64.

[84] Borg J, Georgitsi M, Aleporou-Marinou V, Kollia P, Patrinos GP. Genetic recombination as a major cause of mutagenesis in the human globin gene clusters. Clin Biochem 2009;42:1839–50.

[85] Velculescu VE. Defining the blueprint of the cancer genome. Carcinogenesis 2008;29:1087–91.

[86] Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 2009;462:1005–10.

[87] Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010;463:191–6.

[88] Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40: 722–9.

[89] Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007;450:893–8.

[90] Martin ES, Tonon G, Sinha R, Xiao Y, Feng B, Kimmelman AC, et al. Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. Cancer Res 2007;67:10736–43.

[91] Lockwood WW, Chari R, Coe BP, Girard L, Macaulay C, Lam S, et al. DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. Oncogene 2008;27:4615–24.

[92] Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming *EML4-ALK* fusion gene in non-small-cell lung cancer. Nature 2007;448:561–6.

[93] Martelli MP, Sozzi G, Hernandez L, Pettirossi V, Navarro A, Conte D, et al. *EML4-ALK* rearrangement in non-small cell lung cancer and non-tumor lung tissues. Am J Pathol 2009;174:661–70.

[94] Wong DW, Leung EL, So KK, Tam IY, Sihoe AD, Cheng LC, et al. The *EML4–ALK* fusion gene is involved in various histologic types of lung cancers from nonsmokers with wild-type *EGFR* and *KRAS*. Cancer 2009;115:1723–33.

[95] Shaw AT, Yeap BY, Mino-Kenudson M, Digumarthy SR, Costa DB, Heist RS, et al. Clinical features and outcome of patients with non-small-cell lung cancer who harbor *EML4-ALK*. J Clin Oncol 2009;27:4247–53.

[96] Rodig SJ, Mino-Kenudson M, Dacic S, Yeap BY, Shaw A, Barletta JA, et al. Unique clinicopathologic features characterize *ALK*-rearranged lung adenocarcinoma in the western population. Clin Cancer Res 2009;15:5216–23.