Characterizing complex structural variation in germline and somatic genomes

Aaron R. Quinlan[1,2,3] and Ira M. Hall[1,3]

[1] Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22908, USA
[2] Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22908, USA
[3] Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA

Genome structural variation (SV) is a major source of genetic diversity in mammals and a hallmark of cancer. Although SV is typically defined by its canonical forms (duplication, deletion, insertion, inversion and translocation), recent breakpoint mapping studies have revealed a surprising number of 'complex' variants that evade simple classification. Complex SVs are defined by clustered breakpoints that arose through a single mutation but cannot be explained by one simple end-joining or recombination event. Some complex variants exhibit profoundly complicated rearrangements between distinct loci from multiple chromosomes, whereas others involve more subtle alterations at a single locus. These diverse and unpredictable features present a challenge for SV mapping experiments. Here, we review current knowledge of complex SV in mammals, and outline techniques for identifying and characterizing complex variants using next-generation DNA sequencing.

The genomic landscape of structural variation

Structural variation (SV) is defined as differences in the copy number, orientation or location of relatively large genomic segments (typically >100 bp). The canonical forms include deletions, tandem duplications, insertions, inversions and translocations. Large-scale, microscopically visible genomic rearrangements have long been recognized for their role in evolution and disease, but the remarkable prevalence of submicroscopic SVs only became apparent during the past decade, with the development of high-resolution methods such as array-comparative genomic hybridization (array-CGH) and next-generation DNA sequencing. Current data [1,2] suggest that two humans differ by 5000–10,000 inherited SVs and that both inherited and *de novo* SVs contribute to a variety of normal and disease phenotypes [3]. Similar levels are apparent in other mammalian species, including chimpanzee [4], mouse [5,6], rat [7], dog [8,9] and cattle [10].

Although most tumor genomes harbor somatically acquired SV, the landscape is extremely diverse. Some tumors have tens or hundreds [11–16], whereas others have very few [12,15,17,18], and the abundance of different SV classes varies considerably within and among tumor types [12,15,19]. A subset of cancer-associated SV appears to be functional and under strong selection, such as amplification of oncogenes, deletion of tumor suppressors and translocations that produce fusion genes, but many appear benign. Tumor genome instability may be caused by mutations in DNA maintenance machinery, widespread telomere erosion and/or unstable chromosome architectures acquired during tumorigenesis (e.g. dicentrics) [20].

In this context, the recent discovery of many complex variants that defy simple classification into the typical SV classes has led to a re-examination of the mechanisms and impact of SV. Here, we review recent findings regarding complex variants and discuss techniques for their identification and characterization. We limit our discussion to mammals, mainly human, but we note that these same issues are relevant to other species.

Complex SV defined

Structural variants are defined by their breakpoints, which are the novel sequence junctions generated by structural mutation (Figure 1a–c). Structural variants arise through four general mechanisms (reviewed in [21]): (i) ligation of double-strand DNA breaks (DSBs) through non-homologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ); (ii) exchange between sequences sharing significant stretches of homology, as can occur either by non-allelic homologous recombination (NAHR) during DSB repair or meiosis, or by single-strand annealing (SSA) at DSBs; (iii) DNA replication errors, such as strand slippage or template switching; and (iv) transposition of mobile elements.

Breakpoints are usually identified by comparing the structure of an experimental genome to that of the reference genome, and breakpoint positions are reported based on the coordinate system of the reference (Figure 2). This can cause some confusion because the number of sites may be different depending on which genome one is referring to. For example, a deletion produces a single junction in the experimental genome, but this junction is defined by two coordinates in the reference; the term 'breakpoint' has been used in various studies to describe either or both points of view. The Variant Call Format (VCF) definition resolves this ambiguity by using the terms 'novel adjacency' and 'breakend' to refer to sites in the experimental and reference genomes, respectively [22]. For simplicity, we define

Corresponding author: Hall, I.M. (irahall@virginia.edu).

0168-9525/$ – see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2011.10.002

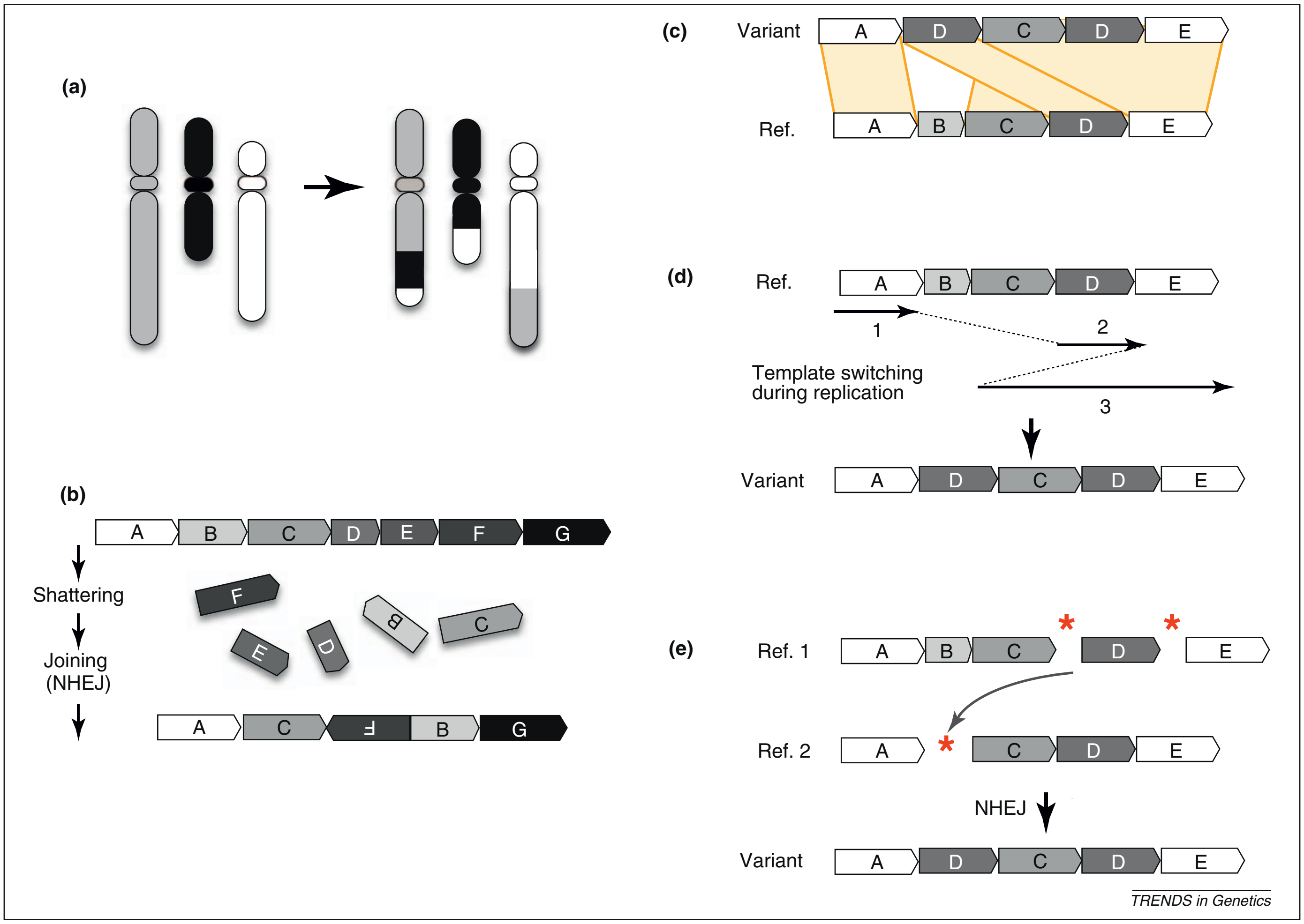


Figure 1. High-level view of complex rearrangements. **(a)** A depiction of a complex chromosomal rearrangement exhibiting four breakpoints from three differently shaded chromosomes. **(b)** A complex rearrangement formed by chromothripsis, where a large section of a chromosome (segments B–F) is shattered and imprecisely stitched back together, leading to multiple intrachromosomal rearrangements and loss of 'D' and 'E' segments. Chromosomal segments are indicated as blocks with different letters and shading, and the orientation by an arrow in each block. This diagram is loosely based on a figure in [94]. **(c)** A complex genomic rearrangement. The variant structure is shown at the top and the reference genome (Ref.) below. Alignments are shown as orange blocks connecting the two sequences. This variant involves a deletion of 'B' and a duplication of 'D', with 'C' unaffected. **(d)** A model for how the variant structure in part **(c)** could arise through template switching during replication [e.g. fork stalling and template switching (FoSTeS) [26], and microhomology-mediated break-induced replication (MMBIR)], where continuously replicated segments are indicated as solid lines and template switches as dotted lines. **(e)** A model for how the variant structure in **(c)** could arise through non-homologous end-joining (NHEJ)-mediated repair of three double-strand DNA breaks. DNA breakages are indicated by red asterisks. In this example, a single DNA break destroys segment 'B', as can occur owing to resection. The two copies of the reference locus ('Ref. 1' and 'Ref. 2') are present as either homologous chromosomes or sister chromatids.

breakpoints based upon their number and position in the experimental genome. This facilitates technical discussion because it is the genome for which experimental data are generated and interpreted, and usually the genome that harbors the derived SV allele produced by recent mutation. However, we note that the reference genome harbors a finite number of derived alleles that can only be reliably discerned by comparison to related species [2,23].

By definition, complex structural variants comprise multiple breakpoints whose origin cannot be explained by a single end-joining or DNA exchange event. Complex SVs vary considerably in their architecture. The most extreme forms exhibit multiple rearrangements between distinct loci and/or different chromosomes, sometimes involving complex patterns of copy number alteration at or near rearrangement breakpoints [16,24,25]. Many comprise multiple deletions, duplications and/or rearrangements at a single locus [6,26–29]. The most subtle forms contain one or more small-scale insertions, deletions or rearrangements at the breakpoint of a larger SV [6,30–33]. As one might expect, the most extreme forms of complex SV are generally associated with cancer or sporadic disorders, and the majority of complex SVs identified in healthy individuals are ostensibly benign.

By definition, complex SVs arise through a single mutational event. A central caveat is that this fact can be difficult to establish. Any complex variant structure can, in theory, also be produced by independent temporally distinct mutations (Figure 1c–e), and repeated mutation is known to occur at localized regions within dicentric chromosomes subjected to breakage–fusion–bridge cycles [34], or at unstable loci such as fragile sites [35], recombination hotspots [36], palindromes [37] and 'core duplicons' [38]. Artificially complex breakpoint patterns can also be produced by one simple mutation at an otherwise complex locus in the reference genome, such as those formed by repeated segmental duplication during evolution [39]. Thus, it can be difficult to distinguish accurately between simple and complex forms of structural mutation.

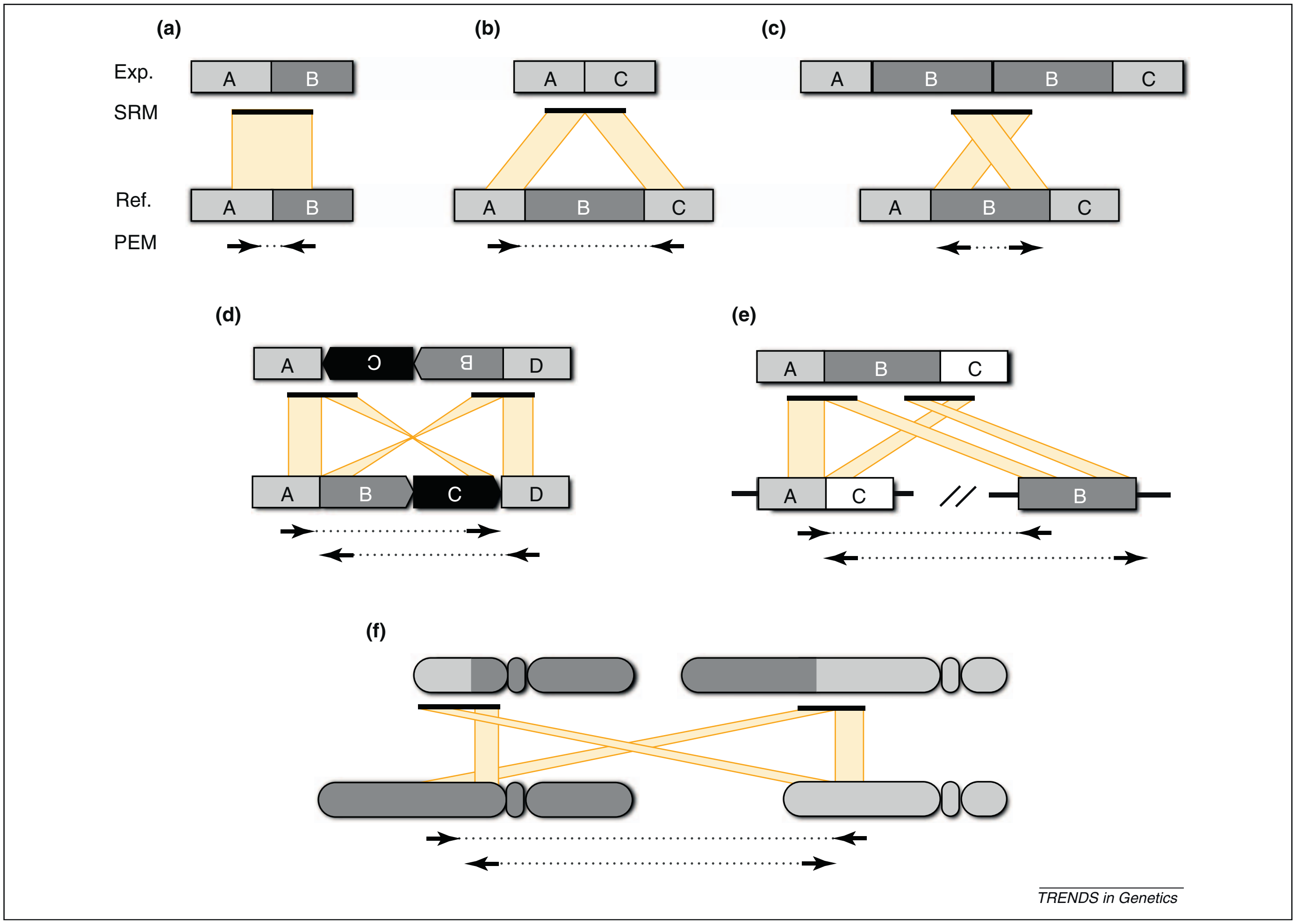


Figure 2. Detecting canonical structural variation (SV) breakpoints through sequencing. When DNA sequences are collected from an experimental (Exp.) genome and aligned to a reference (Ref.) genome, each structural variant class generates a distinct alignment pattern. The patterns observed for paired-end mapping (PEM) and split-read mapping (SRM) are illustrated when both genomes have identical structure **(a)**, and cases where the experimental genome contains a deletion **(b)**, a tandem duplication **(c)**, an inversion **(d)**, a transposon insertion **(e)** or a reciprocal translocation **(f)**. PEM relies upon readpairs whose unsequenced portion (dotted lines) spans a SV breakpoint. When aligned to the reference genome, the alignment distance and orientation of such readpairs indicate the type of rearrangement that has occurred. Reads that map to the plus strand are shown as right-facing arrows, those that map to the negative strand as leftward-facing arrows. All examples depict Illumina paired-end sequence data, where in the absence of SV the normal concordant orientation is plus for the leftmost read and minus for the rightmost read. Note that the expected orientation is different for Illumina mate-pair libraries and for other sequencing platforms, such as SOLiD. In the case of a deletion **(b)**, the readpairs ends will align much farther apart than expected for the DNA library. In contrast to PEM, SRM depends on contiguous sequences that contain an SV breakpoint. Consequently, the sequences before and after the breakpoint will align to disjoint regions of the reference genome. In contrast to PEM, breakpoints are identified at single-base resolution.

Although the methods for detecting complex breakpoint patterns formed by sequential versus complex mutation are essentially the same, their origins and consequences are different. We focus mainly on variants formed by complex mutation and attempt to distinguish between these classes when possible.

Complex SV in the germline

The observation of complex structural mutation is not new. Using standard cytogenetic methods, such as G-banded karyotyping and fluorescent *in situ* hybridization (FISH), numerous complex chromosome rearrangements (CCRs) have been identified in patients suffering from sporadic disorders or infertility (reviewed in [24,40]). CCRs involve at least three breakpoints from two or more chromosomes (Figure 1a), and are estimated to comprise approximately 3% of spontaneous rearrangements detected in prenatal diagnoses [41]. Some events are remarkably complex: approximately 26% of 251 well-characterized CCRs have more than five breakpoints [24,40], and two contain 15–17 breakpoints [42,43]. Similarly complex intrachromosomal rearrangements have also been reported [44]. Interestingly, when rearrangements are fine-mapped, many apparently simple rearrangements are found to be complex, the number of detected breakpoints tends to increase and additional copy number mutations and local rearrangements are often found near breakpoints [43,45,46] (prior work reviewed in [24,40]). The fact that most CCRs are identified as spontaneous events strongly argues that they arise through a single complex mutation rather than through multiple independent simple mutations.

More recently, array-CGH has revealed smaller-scale submicroscopic complex genomic rearrangements associated with sporadic disease. These mutations are generally ‘non-recurrent’ in that they exhibit novel breakpoints, as opposed to recurrent mutations formed by NAHR. The most detailed studies characterized a series of non-recurrent pathogenic *de novo* SVs at the proteolipid protein 1 (*PLP1*) [26] and methyl CpG binding protein 2 (*MECP2*) genes [28], and at a 3-Mb locus associated with

Potocki–Lupski and Smith–Magenis syndromes [27]. Remarkably, the authors found complex structures in 41% of 61 non-recurrent mutations. Taking into account previous (reviewed in [24]) and subsequent [47–50] reports of complex SV at disease-associated loci, these data indicate that complex mutations account for a significant fraction of *de novo* SVs. Reported patterns include adjacent copy number alterations separated by unaltered intervening sequence, deletions or duplications embedded within larger duplications, and triplications.

These observations, as well as previous data from bacterial studies [51], led to two related models for the generation of complex SVs: fork stalling and template switching (FoSTeS) [26], and microhomology-mediated break-induced replication (MMBIR) [29]. In these models, a stalled or broken replication fork undergoes template switching events using microhomology (e.g. 2–5 bp) between the 3′ end of the newly synthesized strand and non-allelic loci (Figure 1d). Complex SVs are produced when multiple switches occur at a single broken and/or stalled fork. Importantly, template switches may occur between distant loci spanning entire chromosome arms [52], presumably owing to proximity in the nucleus, which implies that they may also be involved in many complex chromosomal rearrangements. Fine-scale mapping of CCRs using modern sequencing technologies [46] will help resolve this question.

One might predict that many inherited germline SVs, most of which are probably benign, might also exhibit these features. An early study found that five of 24 deletion breakpoints showed small-scale insertions or rearrangements, or multiple deletions separated by non-deleted sequence [30]. Another clue came from sequencing breakpoints in synteny between the human and gibbon genomes [31]. Of 24 rearrangement breakpoints, 11 contained insertions ranging from 9 bp to 20 kb, and some insertions were mosaic structures comprising common repeats and segmental duplications originating from nearby genomic regions.

Three recent genome-wide DNA sequencing-based studies have assessed the prevalence of complex SVs by characterizing inherited SV breakpoints at single base resolution. The first [6] examined 1171 breakpoints in the mouse genome and found approximately 16% of variants to be complex. Of these, 84% comprised multiple breakpoints in close proximity (<1 kb), often with intertwined breakpoint patterns caused by one or more adjacent deletion and/or duplication events plus local rearrangement; the remainder contained small breakpoint insertions or rearrangements. Common patterns included duplications separated by small non-duplicated segments, deletions adjacent to larger duplications, and deletions with an internal sequence transposed to edge of the breakpoint, often in inverted orientation (Figure 3). Two subsequent studies in humans focused mainly on breakpoint insertions. One [32] used DNA capture technology to sequence 324 breakpoints predicted by array-CGH [53], and found that 5.2% contained breakpoint insertions, most of which were derived from nearby loci and inserted in inverted orientation. Another [33] sequenced 1054 SV breakpoints identified by fosmid paired-end mapping [54], and found that 5.5% contained insertions of DNA larger than 20 bp, and 73% of the breakpoint insertions were derived from a locus less than 250 kb away. Thus, three studies, using distinct methods and definitions of variant complexity, have converged on a similar estimate for inherited complex SV: 5–16%. Given the technical difficulties associated with high-throughput mapping, assembly and interpretation of breakpoint sequences, as well as the apparently higher incidence of *de novo* complex variants (discussed above), we suspect that the true number is somewhat higher.

Complex SV in tumor genomes

The architecture of a somatic genome is less constrained than that of a germline genome, which must complete meiosis and development to survive, and tumors evolve under diverse selective pressures and mutational forces. As a result, the types and numbers of *de novo* SV in different tumors vary widely, and diverse karyotypic configurations have been observed. Many tumors show complex patterns of gene amplification [55], presumably owing to repeated mutation and strong selection. In some breast tumors, ‘firestorms’ of amplification and deletion have been observed [56] on chromosome arms, probably resulting from breakage–fusion–bridge. These complex patterns have historically been explained by a gradual accumulation of mutations during tumorigenesis [20].

The field has been upended by the discovery of extraordinarily complex intra- and interchromosomal rearrangements in certain tumor genomes. In the initial finding, sequencing of a single chronic lymphocytic leukemia (CLL) genome revealed 42 somatically acquired SV breakpoints in several clusters on the long arm of chromosome 4 (4q) [25]. These included deletions, intrachromosomal rearrangements and interchromosomal rearrangements to a single site each on chromosomes 1, 12 and 15. Remarkably, only one additional somatic SV was discovered in the rest of the genome. The 4q region exhibited numerous hemizygous deletions (one copy) separated from each other by unaltered segments (two copies), and the boundaries of deleted segments corresponded to intra- and interchromosomal rearrangement breakpoints. This pattern differs markedly from previously described tumors, but is not rare; the authors mined single nucleotide polymorphism (SNP) array data and found similar patterns in 18 of 746 (2.4%) diverse cancers and/or cell lines, four of which were confirmed by whole-genome sequencing, and in five of 20 (25%) unselected bone cancers also analyzed by genome sequencing.

The authors presented three lines of evidence that these unprecedented rearrangements are generated through a single catastrophic event [25]. First, simulations revealed that breakpoints are clustered in a highly nonrandom manner. Second, the copy number profiles associated with complex events only exhibit two states (either losses or gains but not both) interdigitated with unaltered segments, whereas sequential mutation should produce many states. Third, within breakpoint clusters harboring intertwined deletions and rearrangements, losses derive from the same parental chromosome and heterozygosity is preserved at unaltered segments, which constrains the order of events under a model of sequential mutation. The

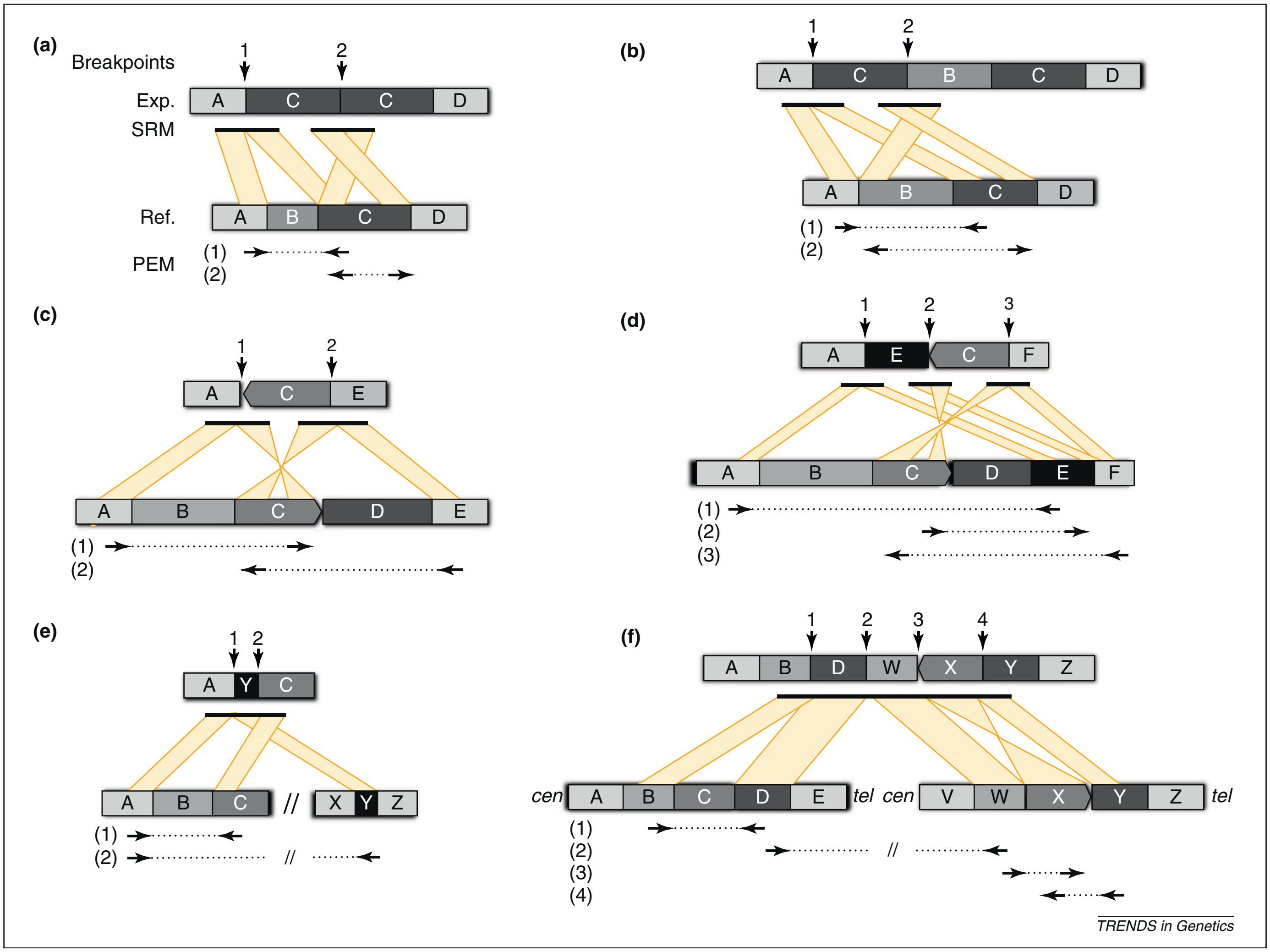


Figure 3. Some common complex structural variation (SV) architectures. In each example, the structure of the experimentally sequenced genome (Exp.) is shown above the reference genome (Ref.), with genomic segments represented as shaded blocks with letters. Distinct loci in the reference genome, either from a nearby or distant location, are separated by '//'. From the top, the positions of 'breakpoints' in the experimental genome are shown with arrows. 'SRM' shows the alignment patterns generated by split-read mapping (SRM) of long reads or assembled contigs that span the breakpoint(s). Shown beneath the reference genome are the discordant alignment patterns generated by paired-end mapping (PEM) using Illumina paired-end sequencing, following the conventions outlined in Figure 2 (main text). **(a)** A deletion of segment 'B' and tandem duplication of 'C'. **(b)** A duplication of 'C' with an intervening unaffected segment ('B'). **(c)** A deletion of 'B' and 'D', with an inversion of the intervening 'C' segment. **(d)** A highly complex variant involving deletion of 'B' and 'D', inversion of 'C', and rearrangement of 'E' between 'A' and 'C'. **(e)** A deletion of 'B' where a segment from a different locus 'Y', either nearby or from an entirely different chromosome, has inserted directly into the deletion breakpoint. **(f)** A large-scale rearrangement between distant loci that is associated with additional local alterations. In this example, there is a rearrangement between 'D' and 'W', segment 'C' has been deleted and segment 'X' has been inverted.

authors refer to this mutational process as 'chromothripsis', and propose that a chromosome is shattered in a one-off event, perhaps by ionizing radiation or one dramatic cycle of breakage–fusion–bridge, and is then stitched back together again in imprecise fashion (Figure 1b). Interestingly, a recent study [43] reported an inherited complex rearrangement with a similar structure, which indicates that chromothripsis-like mechanisms also operate in the germline.

More recently, a single complex rearrangement was identified in three out of seven prostate cancer genomes analyzed by whole-genome sequencing [16]. One involved four loci on a single chromosome, another involved four loci on two chromosomes, and the third involved nine loci on four chromosomes. Strikingly, two involved a novel 'closed chain' breakpoint pattern, such that each locus was connected to two other distinct loci. Although the precise structure of 'closed chain' rearrangements is unclear (Figure 4), there are two key differences between them and those attributed to chromothripsis: (i) there is no obvious clustering of breakpoints on a single chromosome; and (ii) the breakpoint regions do not exhibit copy number mutations. It is an open question whether these rearrangements are caused by chromothripsis or a distinct mechanism, such as FoSTeS/MMBIR. Perhaps indicating the latter, the data shown for one rearrangement are more consistent with three small insertions into a single locus rather than a series of translocations.

Identification and interpretation of complex variation

Advances in DNA sequencing technologies have enabled the exploration of genome structure with exquisite detail. Unlike conventional cytogenetic methods or array-CGH, sequencing permits genome-wide characterization of breakpoints from all classes of SV with high precision. The general algorithmic approaches and available tools

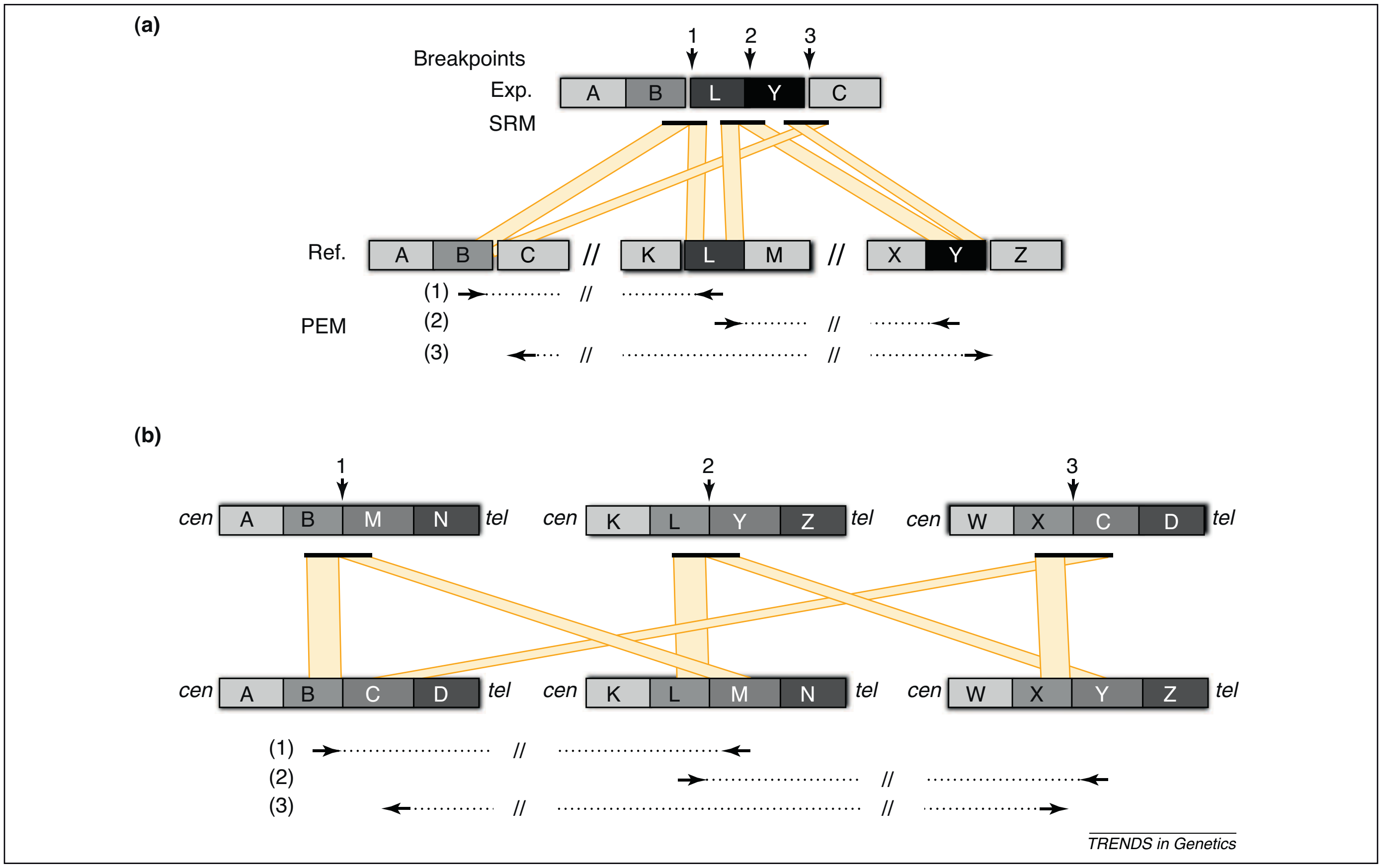


Figure 4. Two potential explanations for ‘closed-chain’ rearrangement patterns. The breakpoint calls and the experimental (‘Exp.’) and reference (‘Ref.’) genomes are shown as in Figure 3 (main text). **(a)** DNA segments from two donor loci ‘L’ and ‘Y’ each insert the recipient locus between ‘B’ and ‘C’ at adjacent positions This generates a closed chain of paired-end mapping (PEM) calls, whereby each locus connects to exactly two other loci. **(b)** A series of translocations involving three chromosomes can also generate a closed-chain rearrangement pattern. The direction of the centromere and telomere are indicated by *‘cen’ and ‘tel’*, respectively. Note that these two forms of closed-chain rearrangement can be distinguished because insertion donor loci have a distinct pattern of two outward facing PEM calls [see ‘L’ and ‘Y’ in part **(a)**], whereas the PEM calls at translocation breakpoints are inward facing [e.g. see breakpoint between ‘L’ and ‘M’ in part **(b)**].

for detecting SV breakpoints from DNA sequence data have been reviewed elsewhere [57,58]. In essence, the identification and interpretation of complex SV involves three steps: (i) genome-wide breakpoint detection using one or more of the techniques, discussed in Box 1; (ii) screening for clusters or interconnected chains of breakpoints that comprise a single complex variant; and (iii) reconstructing the architecture of the variant locus to infer the causal mechanism and potential functional impact.

Screening for complex SV

Once raw breakpoints have been mapped, the primary goal is to distinguish clusters of breakpoints delineating complex variants from nearby, yet potentially simple SV breakpoints caused by independent mutations. The development of robust tools for identifying complex events is a difficult and unsolved problem because there are currently no defined rules for constraining the expected breakpoint patterns. It is not clear whether such rules exist. Nevertheless, discerning complex mutations can be relatively straightforward when analyzing human families or minimally mutated cancer genomes, because spontaneous events can be readily distinguished from inherited variants by analyzing related samples. However, detecting complex variants in a ‘sea’ of simple variants, as in studies of inherited SV or highly rearranged cancer genomes, is problematic because breakpoints may lie in close proximity owing to chance alone. This may not be a concern for functional studies but is crucial for inferring mechanism. There is no simple solution to this conundrum and, thus, most studies have focused on the most obvious examples of complex SV.

Simple and flexible approaches are therefore preferable. Screens must begin by accounting for simple multi-breakpoint variants, such as inversions, retrotranspositions and reciprocal translocations (Figure 2e,f). Merging these breakpoint calls is conceptually simple, but we are not aware of any available software that does so comprehensively. Breakpoint clusters can then be identified by simple sliding window schemes that compare local breakpoint density to a null model. Ideally, this screening method should take into account the non-uniform distribution of simple SV in normal and tumor genomes, as well as commonly observed complex variant architectures. It may be possible to use homology profiles to tease apart nearby or overlapping clusters that arose through distinct mechanisms, but because breakpoints formed by template switching and end-joining can display similar levels of microhomology, this will be difficult in practice. Complex SVs that do not involve obvious breakpoint clusters at a single locus can be identified by computationally searching for chains of interconnected breakpoints that share at least

Box 1. Mapping SV breakpoints with modern sequencing technologies

Depth of coverage

When DNA sequences are aligned to the reference genome, copy number variations (CNVs) are evident as significant increases or decreases in the depth of aligned sequence [89]. Inferring CNV via DOC analysis is conceptually similar to array-CGH and typically yields similar or moderately better resolution (1–15 kb). DOC is inadequate for mapping fine-scale locus complexity, but permits initial identification of complex CNVs and helps determine whether complex breakpoint patterns involve copy number mutations. DOC can also detect NAHR-derived CNVs whose breakpoints lie within large repeats (which can confound the methods below).

Paired-end mapping

PEM strategies identify SV breakpoints by examining the alignments of relatively short sequences from the ends of larger DNA molecules [90]. Sequencing libraries are created with fragments of known length (generally 200–500 bp for paired-end libraries and 1–10 kb for mate-pair libraries). Paired-end sequences (readpairs) that are 'concordant' with the reference genome align with the expected distance and orientation, whereas readpairs spanning an SV breakpoint will produce 'discordant' alignments with an unexpected alignment distance and/or orientation. Each SV class produces a distinct mapping signature (Figure 2, main text). However, current fragment sizes limit sensitivity, discordant mapping patterns can be difficult to interpret at complex SVs (Figure 4, main text), and PEM cannot map breakpoints to single-base resolution.

Split-read mapping

SRM identifies sequences that actually contain a breakpoint [2,91] (Figure 2, main text). The alignments for such sequences are 'split' because DNA segments flanking the breakpoint align to disjoint locations in the reference genome. SRM inherently maps breakpoints to single base resolution and thus provides mechanistic insight. Owing to the repetitive structure of mammalian genomes, genome-wide SRM requires reads longer than approximately 200 bp. Long-read (>500 bp) SRM is a particularly powerful approach for studying complex SV because multiple breakpoints can potentially be captured by a single read, aiding in variant locus reconstruction (Figure 3e,f, main text).

Assembly

Local *de novo* sequence assembly [92] can be used to reconstruct a variant locus. Once an assembled sequence is aligned to the reference genome, breakpoint(s) are discerned following the same principles as SRM. Unlike SRM, assembly typically generates substantially larger 'contigs' that are more amenable to characterizing complex SV. Although currently infeasible for most laboratories, whole-genome assemblies promise the most comprehensive description of SV, as large portions of entire chromosomes can be aligned to identify precisely both canonical and complex rearrangements [1,93].

one locus in common. Tools in the BEDTools software suite [59] can be adapted for this purpose [6]. By integrating results from clustering and chaining approaches, most classes of complex SV can be discerned. We stress, however, that these higher-order clustering steps can produce falsely complex SVs at repetitive or poorly assembled loci in the reference genome that generate abundant breakpoint calls, as often occurs at or near centromeres, telomeres, simple tandem repeats and regions laden with segmental duplications. Thus, subsequent annotation and characterization steps are crucial.

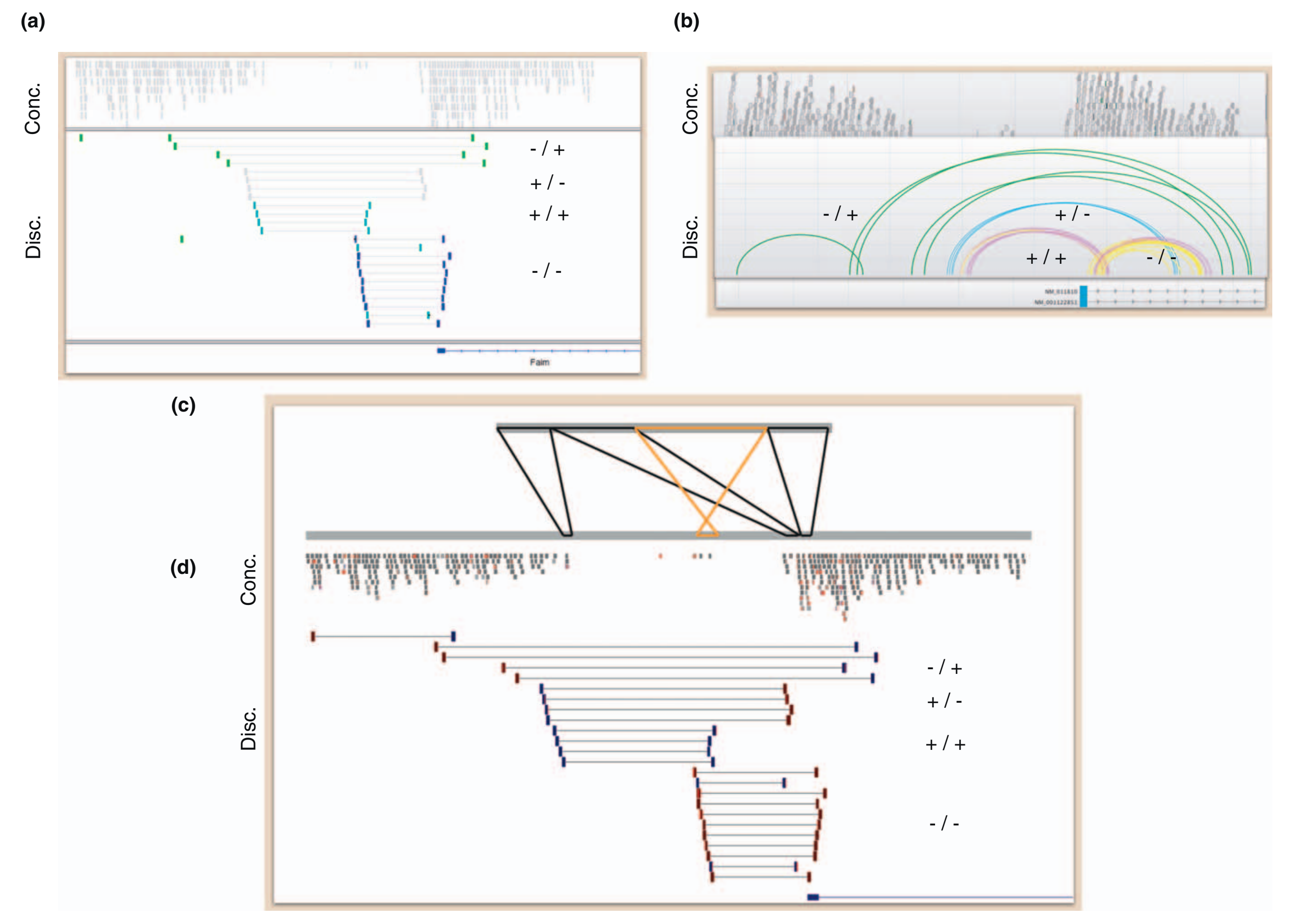
The above methods may fail to detect complex SVs that possess neither clustered nor chained breakpoints, but rather comprise nested or overlapping variant calls that affect a common genomic interval. This pattern is trivial to detect, but is also commonly produced by sequential mutation and should be interpreted with caution. These methods may also miss cryptic complex variants that contain small-scale insertions or rearrangements at the breakpoint itself. For these, it is necessary to inspect carefully breakpoints at single-base resolution and to align the breakpoint sequence to the reference genome. Sensitive alignment is crucial because small breakpoint alterations can masquerade as non-templated addition of nucleotides during NHEJ, merely owing to the inability of aligners to find significant matches.

Interpreting complex variants

A key question for any complex variant is: what exactly does it look like? Integration of breakpoints identified by paired-end mapping (PEM), split-read mapping (SRM) and/or local assembly (Box 1), combined with depth of coverage (DOC) analysis to distinguish between balanced rearrangements and copy number mutations, is theoretically sufficient to infer the architecture of most variants (Figures 3 and 4). However, this remains a major challenge for two reasons. First, neither reconstructing nor visualizing complex variant structures are trivial problems and there is a notable dearth of suitable computational tools. Thus, to our knowledge, all DNA sequencing-based studies to date have relied heavily on manual curation and human expertise to interpret complex breakpoint patterns. This laborious approach has proven effective and resulted in detailed architectural information for over 250 complex SVs [6,16,25,32,33,43], but is unsustainable given the scale of current genome sequencing projects. Second, the accuracy of interpretation depends entirely on the accuracy of the underlying breakpoint calls, and current breakpoint mapping strategies suffer from either high false positive or high false negative rates, and sometimes both. It is therefore likely that complex SVs are more prevalent, and more architecturally diverse, than currently recognized owing to underascertainment and misinterpretation.

Manual variant reconstruction is greatly aided by data visualization software (Figure 5). The UCSC Genome Browser [60], Integrative Genomics Viewer (IGV) [61] and Savant [62] excel at displaying raw sequence data aligned to the reference genome and can also display annotation tracks, but are only practical for visualizing small genomic regions (<100 kb). A current advantage of IGV is the ability to visualize two distinct loci in 'split-screen' mode, but Savant offers superior visualization of readpair connectivity. At the other end of the spectrum, visualization tools such as CIRCOS [63] or GREMLIN [64] provide aesthetically pleasing rearrangement depictions, but are mainly useful for summarizing results, not interpreting data. A major limitation of the above tools is that they display data solely with respect to the reference genome, which does not allow one to infer variant architecture easily.

Rapid interpretation requires a direct comparison of the structure of assembled breakpoint sequences, or entire variant loci, to the structure of the reference genome. In some cases, a simple dotplot may suffice. The PARASIGHT software (J. Bailey et al., unpublished: http://eichlerlab.gs.washington.edu/jeff/parasight) is ideally suited to this task



TRENDS in Genetics

Figure 5. Visualizing complex loci. Snapshots of aligned paired-end sequence data from a complex locus (chr9: 98,880,333–98,889,602; NCBI37/mm9) in the DBA/2J mouse strain are depicted with **(a)** the Integrated Genomics Viewer (IGV), **(b)** SAVANT and **(d)** the UCSC Genome Browser. In **(c)**, alignment of a partial *de novo* assembly of this locus is displayed with PARASIGHT. In each panel, concordant readpairs (Conc.) are displayed above the discordant (Disc.) readpairs that collectively indicate the complex rearrangement at this locus. The orientation of the 'left' and 'right' ends of each type of discordant readpair are displayed next to the alignments: '-/+' for tandem duplication; '+/-' for deletion; '-/-' and '+/+' for distinct inversion breakpoints. Also in **(c)**, black alignments between the assembled locus and the reference genome are in the same orientation, whereas orange alignments are in opposite orientation. Note that this assembled contig does not represent the entire locus, as the outermost readpairs suggesting a tandem duplication (-/+) were not incorporated into the assembly. However, the assembled sequence suggests that, at a minimum, this complex locus involves two adjacent deletions of 2.5 kb and 0.9 kb, which are separated by an intervening 300-bp segment that was not deleted, but instead inverted.

because it shows pairwise alignments in an informative format that preserves the structure of both variant and reference sequences (Figure 5c), and can display annotation tracks. For example, an automated PARASIGHT pipeline enabled visualization and interpretation of several thousand assembled breakpoints in several days [6]. Unfortunately, although PARASIGHT is flexible, it is difficult to use and often requires substantial customization for informative viewing. Other tools, such as MIROPEATS [65] and BARAVI (R. Ophoff et al., unpublished: http://www.genetics.ucla.edu/labs/ophoff/BARAVI/), support pairwise alignment and visualization but cannot display tracks. The paucity of user-friendly breakpoint visualization software presents a major bottleneck for interpreting complex variants and underscores the need for improved tools.

Manual curation is the most accurate approach for variant reconstruction, but as the study of complex SV expands to thousands of genomes, it is neither practical nor reproducible. In theory, it should be possible to develop software that infers variant architecture from breakpoint predictions and DOC profiles, but we are unaware of any that explicitly attempts to do so. Moreover, we suspect that automated reconstruction of complex SVs would require impeccable input data. For example, sophisticated algorithms have proven necessary merely to integrate breakpoint calls and DOC profiles for simple deletions [2,66]. As sequencing methods continue to improve, automated approaches will eventually be feasible through increased read lengths, emerging technologies such as 'strobe' sequencing [67] and, ultimately, routine generation of high-quality diploid genome assemblies.

If a complex SV can be assembled into a single contig, variant reconstruction becomes a tractable problem of describing the relative structure of two DNA sequences. The first step is to align the variant sequence to the reference genome. A complication is that portions of the variant 'query' sequence containing repeats will align to multiple loci. This problem is trivial for variants that involve a single well-defined locus, but for rearrangements that involve repetitive regions or multiple loci, resolving these ambiguities can be difficult. This is also a significant problem for the initial detection of complex SVs from long-reads or draft assemblies. Most suitable aligners report all significant alignments, including irrelevant 'sub-alignments' contained within larger aligned sections of the query [68–70], which necessitates subsequent selection of the 'best' minimal set for locus reconstruction. The BWA-SW aligner uses a greedy heuristic strategy to discard subalignments that are subsumed by larger alignments [71]; we have found that this, or similar, heuristic strategies are adequate for moderately complex variants comprising mainly unique sequence. Otherwise, it is preferable to pursue a more optimal alignment selection strategy.

Once alignments are defined, reconstructing variant architecture is a semantic problem of describing the relationship between alignment blocks based upon their relative positions and orientations in the variant and reference sequences. The VCF 4.1 specification offers a sensible solution for this practical problem [22].

Mechanistically minded studies might seek to reconstruct the mutational events that generated each complex variant. Similar problems has been studied in the context of ancestral genome reconstruction using breakpoint graphs [72–75], and for inferring the mutational history of segmental duplications using modified A-Bruijn graphs [76] or DAWGs [77]. Genome-scale models are subjected to various simplifying assumptions to prevent intractable computational complexity, but for any given complex variant, optimal solutions are possible. An unsolved problem is how to define optimal solutions that take into account current models of mutation.

Concluding remarks

Studies of complex SV have provided new insights into the processes that generate genome variation, and this has clear implications for conventional models of species and cancer evolution that generally assume progressive, step-wise mutations. In both contexts, complex mutations represent a form of punctuated genome evolution. Resulting variants may have more subtle, unpredictable and multi-faceted phenotypic impacts compared with simple variants. For example, complex mutations can rearrange exons to create novel proteins, shuffle promoters, enhancers and/or repressors into a novel regulatory configuration, or simultaneously disrupt multiple genes and pathways. In the context of a developing tumor, simultaneous formation of multiple fusion genes, amplified oncogenes or deleted tumor suppressors may lead to rapid expansion of a clone with very different characteristics than neighboring cells.

A major unresolved question in the field is how complex variants arise. The two general models for complex SV formation [template switching during DNA replication (FoSTeS/MMBIR) [26,29] and chromosome shattering (chromothripsis) [25]] each have eminently sensible features, but it is worth remembering that neither has been directly implicated. This begs the question of whether these mechanisms indeed account, either alone or through collusion, for the architecturally diverse rearrangements that have been observed. Or is another as-yet undescribed mechanism at work? At present, there are not sufficient data to answer these questions. However, we speculate that most complex variants arise through a common mechanism. The rearrangements thus far attributed to chromothripsis differ from those explained by FoSTeS/MMBIR mainly in their greater size and complexity; the patterns are ostensibly similar. We further note that a recent study of germline rearrangements [78] has proposed that FoSTeS/MMBIR may explain complex breakpoint clusters that resemble those attributed to chromothripsis [25,43,79]. These clusters contain three copy number states, including duplications and triplications, and small breakpoint insertions derived from nearby loci. These features are easier to explain by replication than by chromosome shattering. By contrast, shattering is a more simple explanation for the staggeringly complex variants that exhibit frequent oscillation between two copy number states (deleted and unaltered), as observed in tumor genomes. We expect future breakpoint sequencing studies to yield additional clues, but we are not confident that the true mechanism(s) can be resolved by sequencing alone, given that neither variant architectures nor breakpoint homology profiles appear sufficient to distinguish the two models. Direct experimental studies may be necessary to yield clarity.

The likelihood that complex mutations primarily arise through processes that are active in somatic cells, and not concentrated in meiosis, also implies that many other simple mutations do as well, and thus each individual may be a mosaic composition of cells with different genome structures. Indeed, evidence of somatic variation is growing [80–86], and this may potentially account for certain phenotypes that emerge during development and aging. The potential link to replication also implies that environmental conditions or *trans*-acting mutations that affect replication fidelity can modulate mutation rates. It has been proposed that replication stress may lead to flurries of structural mutation [21,29], and there is direct evidence for this in *Escherichia coli* [51] and cultured human cells [87,88]. Further work is necessary to prove this theory, but the potential existence of genetic and environmental modulators of complex mutation is intriguing.

In most cases, the functional consequences of complex SVs are unclear, and their true contribution to natural variation remains an open question. Whether these variants turn out to be a curious sideshow of mutational complexity or a driving force of functional innovation can only be answered by ongoing and future whole-genome sequencing of well-phenotyped samples. Rapidly improving DNA sequencing technologies will aid this effort, but perhaps the greater challenge lies in bioinformatic interpretation. At present, there is a notable paucity of high-throughput methods for complex SV identification, visualization, reconstruction or interpretation. We expect this challenge to be met in coming years, and we look forward to a more complete understanding of the mechanisms and functional ramifications of complex SV.

Acknowledgments

Our work has been sponsored by the National Institutes of Health (DP2OD006493-01 to IMH; 1F32HG005197-01 to ARQ), the Burroughs Wellcome Fund (IMH) and the March of Dimes (IMH). We thank R.A. Clark for implementing our SV visualization pipeline.

References

1 Pang, A.W. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11, R52

2 Mills, R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59–65

3 Zhang, F. *et al.* (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481

4 Perry, G.H. *et al.* (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8006–8011

5 Graubert, T.A. *et al.* (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet.* 3, e3

6 Quinlan, A.R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* 20, 623–635

7 Guryev, V. *et al.* (2008) Distribution and functional impact of DNA copy number variation in the rat. *Nat. Genet.* 40, 538–545

8 Chen, W.K. *et al.* (2009) Mapping DNA structural variation in dogs. *Genome Res.* 19, 500–509

9 Nicholas, T.J. *et al.* (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res.* 19, 491–499

10 Liu, G.E. *et al.* (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Res.* 20, 693–703

11 Campbell, P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* 40, 722–729

12 Stephens, P.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 462, 1005–1010

13 Ding, L. *et al.* (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464, 999–1005

14 Pleasance, E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191–196

15 Campbell, P.J. *et al.* (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* 467, 1109–1113

16 Berger, M.F. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature* 470, 214–220

17 Welch, J.S. *et al.* (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* 305, 1577–1584

18 Puente, X.S. *et al.* (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475, 101–105

19 Stratton, M.R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science* 331, 1553–1558

20 Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell* 144, 646–674

21 Hastings, P.J. *et al.* (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 10, 551–564

22 Danecek, P. *et al.* (2011) The Variant Call Format and VCF tools. *Bioinformatics* 27, 2156–2158

23 Lam, H.Y. *et al.* (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* 28, 47–55

24 Zhang, F. *et al.* (2009) Complex human chromosomal and genomic rearrangements. *Trends Genet.* 25, 298–307

25 Stephens, P.J. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27–40

26 Lee, J.A. *et al.* (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131, 1235–1247

27 Zhang, F. *et al.* (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* 41, 849–853

28 Carvalho, C.M. *et al.* (2009) Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum. Mol. Genet.* 18, 2188–2203

29 Hastings, P.J. *et al.* (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* 5, e1000327

30 Perry, G.H. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* 82, 685–695

31 Girirajan, S. *et al.* (2009) Sequencing human–gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. *Genome Res.* 19, 178–190

32 Conrad, D.F. *et al.* (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.* 42, 385–391

33 Kidd, J.M. *et al.* (2010) A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 143, 837–847

34 Artandi, S.E. and DePinho, R.A. (2010) Telomeres and telomerase in cancer. *Carcinogenesis* 31, 9–18

35 Durkin, S.G. and Glover, T.W. (2007) Chromosome fragile sites. *Annu. Rev. Genet.* 41, 169–192

36 Myers, S. *et al.* (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* 40, 1124–1129

37 Inagaki, H. *et al.* (2009) Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res.* 19, 191–198

38 Marques-Bonet, T. and Eichler, E.E. (2009) The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harb. Symp. Quant. Biol.* 74, 355–362

39 Bailey, J.A. *et al.* (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* 11, 1005–1017

40 Pellestor, F. *et al.* (2011) Complex chromosomal rearrangements: origin and meiotic behavior. *Hum. Reprod. Update* 17, 476–494

41 Giardino, D. *et al.* (2009) *De novo* balanced chromosome rearrangements in prenatal diagnosis. *Prenat. Diagn.* 29, 257–265

42 Tupler, R. *et al.* (1992) A complex chromosome rearrangement with 10 breakpoints: tentative assignment of the locus for Williams syndrome to 4q33——q35.1. *J. Med. Genet.* 29, 253–255

43 Kloosterman, W.P. *et al.* (2011) Chromothripsis as a mechanism driving complex *de novo* structural rearrangements in the germline. *Hum. Mol. Genet.* 20, 1916–1924

44 Lindstrand, A. *et al.* (2008) Molecular cytogenetic characterization of a constitutional, highly complex intrachromosomal rearrangement of chromosome 1, with 14 breakpoints and a 0.5 Mb submicroscopic deletion. *Am. J. Med. Genet. Part A* 146A, 3217–3222

45 Feenstra, I. *et al.* (2011) Balanced into array: genome-wide array analysis in 54 patients with an apparently balanced *de novo* chromosome rearrangement and a meta-analysis. *Eur. J. Hum. Genet.* DOI: 10.1038/ejhg.2011.120

46 Talkowski, M.E. *et al.* (2011) Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am. J. Hum. Genet.* 88, 469–481

47 Zhang, F. *et al.* (2010) Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. *Am. J. Hum. Genet.* 86, 892–903

48 Liu, P. *et al.* (2011) Copy number gain at Xp22.31 includes complex duplication rearrangements and recurrent triplications. *Hum. Mol. Genet.* 20, 1975–1988

49 Zhang, F. *et al.* (2010) Identification of uncommon recurrent Potocki-Lupski syndrome-associated duplications and the distribution of rearrangement types and mechanisms in PTLS. *Am. J. Hum. Genet.* 86, 462–470

50 Choi, B.O. *et al.* (2011) Inheritance of Charcot-Marie-Tooth disease 1A with rare nonrecurrent genomic rearrangement. *Neurogenetics* 12, 51–58

51 Slack, A. *et al.* (2006) On the mechanism of gene amplification induced under stress in *Escherichia coli*. *PLoS Genet.* 2, e48

52 Koumbaris, G. *et al.* (2011) FoSTeS, MMBIR and NAHR at the human proximal Xp region and the mechanisms of human Xq isochromosome formation. *Hum. Mol. Genet.* 20, 1925–1936

53 Conrad, D.F. *et al.* (2009) Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712

54 Kidd, J.M. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56–64

55 Korkola, J. and Gray, J.W. (2010) Breast cancer genomes—form and function. *Curr. Opin. Genet. Dev.* 20, 4–14

56 Hicks, J. *et al.* (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* 16, 1465–1479

57 Medvedev, P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6, S13–S20

58 Alkan, C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376

59 Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842

60 Kent, W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.* 12, 996–1006

61 Robinson, J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26

62 Fiume, M. *et al.* (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics* 26, 1938–1944

63 Krzywinski, M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645

64 O'Brien, T.M. *et al.* (2010) Gremlin: an interactive visualization model for analyzing genomic rearrangements. *IEEE Trans. Visual. Comput. Graph.* 16, 918–926

65 Parsons, J.D. (1995) Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* 11, 615–619

66 Handsaker, R.E. *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* 43, 269–276

67 Ritz, A. *et al.* (2010) Structural variation analysis with strobe reads. *Bioinformatics* 26, 1291–1298

68 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

69 Ning, Z. *et al.* (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.* 11, 1725–1729

70 Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664

71 Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595

72 Pevzner, P. and Tesler, G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7672–7677

73 Bourque, G. *et al.* (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* 14, 507–516

74 Murphy, W.J. *et al.* (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309, 613–617

75 Alekseyev, M.A. and Pevzner, P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res.* 19, 943–957

76 Jiang, Z. *et al.* (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* 39, 1361–1368

77 Kahn, C.L. and Raphael, B.J. (2009) A parsimony approach to analysis of human segmental duplications. *Pac. Symp. Biocomput.* 126–137

78 Liu, P. *et al.* (2011) Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* 146, 889–903

79 Magrangeas, F. *et al.* (2011) Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood* 118, 675–678

80 Liang, Q. *et al.* (2008) Extensive genomic copy number variation in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 105, 17453–17456

81 Piotrowski, A. *et al.* (2008) Somatic mosaicism for copy number variation in differentiated human tissues. *Hum. Mutat.* 29, 1118–1124

82 Bruder, C.E. *et al.* (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* 82, 763–771

83 Lam, K.W. and Jeffreys, A.J. (2007) Processes of *de novo* duplication of human alpha-globin genes. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10950–10955

84 Flores, M. *et al.* (2007) Recurrent DNA inversion rearrangements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6099–6106

85 Muotri, A.R. *et al.* (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 435, 903–910

86 Coufal, N.G. *et al.* (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460, 1127–1131

87 Arlt, M.F. *et al.* (2009) Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *Am. J. Hum. Genet.* 84, 339–350

88 Arlt, M.F. *et al.* (2011) Comparison of constitutional and replication stress-induced genome structural variation by SNP array and mate-pair sequencing. *Genetics* 187, 675–683

89 Chiang, D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103

90 Raphael, B.J. *et al.* (2003) Reconstructing tumor genome architectures. *Bioinformatics* 19 (Suppl. 2), ii162–ii171

91 Mills, R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16, 1182–1190

92 Miller, J.R. *et al.* (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95, 315–327

93 Li, Y. *et al.* (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat. Biotechnol.* 29, 723–730

94 Meyerson, M. and Pellman, D. (2011) Cancer genomes evolve by pulverizing single chromosomes. *Cell* 144, 9–10